The Annals of Statistics
2024, Vol. 52, No. 6, 2931–2955
https://doi.org/10.1214/24-AOS2452

STATISTICAL INFERENCE FOR DECENTRALIZED FEDERATED LEARNING

BY JIA GU[1,a] AND SONG XI CHEN[2,b]

[1]*Center for Data Science, Zhejiang University,* [a]*gujia@zju.edu.cn*
[2]*Department of Statistics and Data Science, Tsinghua University,* [b]*sxchen@tsinghua.edu.cn*

This paper considers decentralized Federated Learning (FL) under heterogeneous distributions among distributed clients or data blocks for the M-estimation. The mean squared error and consensus error across the estimators from different clients via the decentralized stochastic gradient descent algorithm are derived. The asymptotic normality of the Polyak–Ruppert (PR) averaged estimator in the decentralized distributed setting is attained, which shows that its statistical efficiency comes at a cost as it is more restrictive on the number of clients than that in the distributed M-estimation. To overcome the restriction, a one-step estimator is proposed which permits a much larger number of clients while still achieving the same efficiency as the original PR-averaged estimator in the nondistributed setting. The confidence regions based on both the PR-averaged estimator and the proposed one-step estimator are constructed to facilitate statistical inference for decentralized FL.

1. Introduction. The stochastic gradient descent (SGD) algorithm is a commonly used method for M-estimation in applications of statistical learning, which is known for more efficient computation than the traditional gradient descent (GD) algorithm. The SGD was proposed by Robbins and Monro (1951) in the context of the sequential estimation, which avoids calculating gradients using the entire dataset, but only one observation at a time. The consistency and the asymptotic normality of the SGD-based estimators were established in Chung (1954), Ruppert (1988), and Polyak and Juditsky (1992). The rationale behind the SGD procedure is that at each time the calculated gradient is an unbiased estimator of the gradient of the population objective function. The difference between the SGD iteration and the minimizer of the population function can be viewed as a weighted averaging of past gradient noises, and thus the consistency of the SGD iteration follows. The averaging when the step size $\eta_t = O(t^{-\alpha})$ at time t with $1/2 < \alpha < 1$ is called the Polyak–Ruppert (PR) averaging. It is noted that for $\alpha = 1$, averaging all past estimates will lead to less efficient estimation due to the introduced strong serial correlation with a smaller step size.

To facilitate SGD-based statistical inference in a full sample, Fang, Xu and Yang (2018) proposed a bootstrap SGD iteration for online inference of the true parameter in the M-estimation. To estimate the asymptotic covariance matrix of the PR-averaged estimator, Chen et al. (2020a) constructed two covariance estimators. One was a sample covariance type estimator called the batch-means estimator and the other was an online plug-in estimator. The batch-means estimator was later extended to a fully online version by Zhu, Chen and Wu (2023). Lee et al. (2022) studied an online random scaling algorithm that led to confidence intervals with more accurate coverage than the batch-means approach.

However, it is often the case that communication is restricted among the data samples due to divided ownership, as the data are often collected and stored by different clients (Gu and

Received May 2024; revised September 2024.
MSC2020 subject classifications. Primary 62-08; secondary 62L12.
Key words and phrases. Decentralized estimation, decentralized stochastic gradient descent, federated learning, heterogeneity, one-step estimation.

Chen, 2023). The difference among clients can create heterogeneity among the data distributions of the clients. These realities have motivated a new M-estimation framework called Federated Learning (FL) (McMahan et al. (2017)) which is gaining popularity, where a weighted average of local risk functions identifies the parameter of interest. The clients are required to collaboratively solve this heterogeneous M-estimation problem, while keeping the non-i.i.d. training data stored locally (Li et al. (2020), Kairouz et al. (2021)).

The local SGD algorithm (Stich (2019)) is proposed to solve the FL problem, which allows the clients to run their respective SGD in parallel and synchronizes the local parameter estimates every τ ($\tau \geq 1$) steps via a central server. To alleviate the communication burden of the central host in the FL, the local SGD algorithm is further generalized to the DFL algorithm (Wang and Joshi (2021)), which synchronizes the local parameter estimates in a decentralized style. That is, the clients only share gradient information with their neighbors according to a network during the optimization. This algorithm extends the decentralized SGD ($\tau = 1$) algorithm (Lian et al. (2017)), which, in turn, is an extension of both the nondistributed SGD algorithm and the decentralized GD algorithm (Yuan, Ling and Yin (2016)).

Most of the existing analysis focused on the constant step size scenario for the SGD-based algorithms (Wang and Joshi (2021), Alghunaim and Yuan (2022)) with only a few exceptions (Li et al. (2022)), which typically leads to asymptotically biased estimators. Moreover, the existing studies tended to treat K, the number of clients in the FL network as fixed, not reflecting the reality that the number of clients can increase along with the local sample size. One applicable setting of the double asymptotic is the large mobile networks, where the mobile keyword prediction is performed based on users' historical text data as in the GBOARD project of Google (Hard et al. (2018)) and the QUICKTYPE KEYBOARD by Apple (Apple (2019)). Another setting is the modern Internet-of-Things (IoT) networks, where wearable devices are used for health event prediction (Pantelopoulos and Bourbakis (2010), Chen et al. (2020b)) or the autonomous vehicles control (Chen and Cui (2024)). In both settings, the FL is used to train a model, where the clients are the device users and the local sample size refers to the amount of data produced by each device. Given the large number of mobile devices, it is reasonable to consider the double asymptotic setting where the number of clients increases with the local sample size.

This paper considers statistical inference for the heterogeneous M-estimation based on the most general DFL algorithm, which includes many distributed SGD-based algorithms mentioned as special cases. We derive the mean squared error (MSE) bound for the spatially averaged trajectory and the consensus error bound of the local estimates across clients. For each fixed number of clients K, the almost sure convergence of the averaged trajectory is derived. We also establish the asymptotic normality of a DFL version of the Polyak–Ruppert (PR) averaged estimator. Our study reveals that the PR-averaged estimator in the context of DFL is efficient in the sense that its asymptotic variance is the same as that of the full-sample M-estimator if K is either finite or diverges at the rate $o(T^{2\alpha-1})$, where T is the common local sample size and $\alpha \in (1/2, 1)$ is the diminishing rate in the SGD step size $\eta_t = O(t^{-\alpha})$.

To allow a higher divergence rate for the number of clients K in the PR-averaged estimator, we propose a computation-efficient one-step estimator that is also statistically efficient but permits $K = o(T)$. The proposed one-step estimator utilizes the PR-averaged estimator with a smaller step size ($\alpha = 1$) as an initial estimator and a correction term to improve its statistical efficiency. Regardless the objective functions being the second-order differentiable or not, confidence regions based on both the PR-estimator and the one-step estimator are constructed in the DFL context, respectively, with asymptotically correct coverages. We also discuss the impacts of the sparse connectedness of the connection network.

The paper is organized as follows. The framework for the DFL estimation is outlined in Section 2 with a review of the statistical properties of the SGD estimators. The mean squared

error of the averaged estimator, the consensus error across clients and the asymptotic normality of the PR-averaged estimator are derived in Sections 3 and 4 to motivate the construction of the efficient one-step DFL estimator. The construction of the confidence region for the estimators are shown in Sections 4 and 5. The statistical properties of the proposed one-step estimator are revealed in Section 6. Section 7 reports simulation results to verify the theoretical results. Section 8 concludes with a discussion. Extra technical details are reported in the Supplementary Material (SM) (Gu and Chen (2024)).

2. Preliminaries. A typical federated learning (FL) setting involves K clients, and we define $f_k(\cdot;\boldsymbol{\xi}^k)$ as the loss function specific to the kth client and $F_k(\boldsymbol{\theta}) = \mathrm{E}_{\mathcal{P}_k}(f_k(\boldsymbol{\theta};\boldsymbol{\xi}^k))$ as the corresponding risk function, where $\boldsymbol{\xi}^k$ is drawn from an unknown distribution $\mathcal{P}_k$. We do not assume $\{\mathcal{P}_k\}_{k=1}^K$ being identical to accommodate heterogeneity across the clients' local data distributions. We write $\mathrm{E}_{\mathcal{P}_k}(\cdot)$ as $\mathrm{E}(\cdot)$ for simplicity. Define the federated risk function

$$F(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k F_k(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter of interest, $\{w_k\}_{k=1}^K$ is a set of positive pre-specified weights such that $\sum_{k=1}^K w_k = 1$. The purpose of the FL is to estimate the parameter $\boldsymbol{\theta}_K^*$ defined as

$$\boldsymbol{\theta}_K^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Phi}} F(\boldsymbol{\theta}), \tag{2}$$

where the subscript K reflects the dependence on the number of clients and $\boldsymbol{\Phi}$ is the parameter space. For each client k, the observations $\mathcal{D}_k = \{\boldsymbol{\xi}_t^k\}_{t=1}^{n_k}$ are independent and identically distributed (i.i.d.), drawn from $\mathcal{P}_k$, and n_k is the local sample size. The full dataset of all clients is $\mathcal{D} = \bigcup_{k=1}^K \mathcal{D}_k$, leading to the overall sample size $N = \sum_{k=1}^K n_k$.

In the conventional setting where one has dataset $\mathcal{D}$ and the full data communication among the local datasets is available, one can minimize the empirical version of (2), namely, $\sum_{k=1}^K w_k n_k^{-1} \sum_{t=1}^{n_k} f_k(\boldsymbol{\theta};\boldsymbol{\xi}_t^k)$ to obtain the full sample M-estimator of $\boldsymbol{\theta}_K^*$

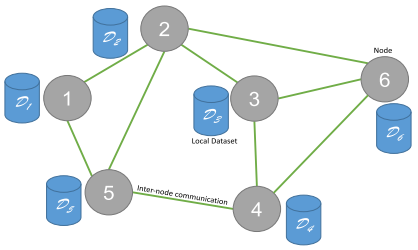
$$\hat{\boldsymbol{\theta}}_K = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Phi}} \sum_{k=1}^{K} w_k \left(\frac{1}{n_k} \sum_{t=1}^{n_k} f_k(\boldsymbol{\theta};\boldsymbol{\xi}_t^k) \right), \tag{3}$$

which is usually solved using the gradient-based methods.

However, the above setting of the M-estimation is infeasible for the FL scenario considered in this work since the pre-given datasets $\mathcal{D}_k$ are not available. Instead, each client's local dataset are incrementally gathered, making the setting more aligned with the sequential estimation as outlined in Section 2.2, where each client or cluster contributes one datum at a time. Hence, it is more appropriate to use the number of step sizes T to represent the sample sizes of the local datasets $\mathcal{D}_k$ such that $n_k = T$ for all $1 \le k \le K$. Throughout this paper, when we refer to an estimator of $\boldsymbol{\theta}_K^*$ as asymptotically efficient, we imply that the estimator possesses the same asymptotic variance as the estimator

$$\hat{\boldsymbol{\theta}}_K = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Phi}} \sum_{k=1}^{K} w_k \left(\frac{1}{T} \sum_{t=1}^{T} f_k(\boldsymbol{\theta};\boldsymbol{\xi}_t^k) \right). \tag{4}$$

This work assumes that there exist positive constants b_1 and b_2 such that $b_1 \le w_k K \le b_2$ for $1 \le k \le K$. The most popular choice of the weights is the equal weights $w_k = K^{-1}$ (Kairouz et al. (2021)), that treats all the clients equally. If we know the sampling distribution of the clients, the sampling weights may be used as the $\{w_k\}_{k=1}^K$ (Wang et al. (2021)). We consider statistical optimization and inference for $\boldsymbol{\theta}_K^*$ in the DFL setting, which generalizes



$$C = \begin{pmatrix} 11/20 & 1/5 & 0 & 0 & 1/4 & 0 \\ 1/5 & 1/5 & 1/5 & 0 & 1/5 & 1/5 \\ 0 & 1/5 & 3/10 & 1/4 & 0 & 1/4 \\ 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/5 & 0 & 1/4 & 3/10 & 0 \\ 0 & 1/5 & 1/4 & 1/4 & 0 & 3/10 \end{pmatrix}$$

FIG. 1. *A connection network with* 6 *nodes (left) and its connection matrix* C *(right) according to the Metropolis–Hastings rule.*

the nondistributed SGD (Robbins and Monro (1951)), based on a connection network where the clients are seen as nodes, where only connected notes are permitted to communicate.

Throughout the paper, we use $\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \|\boldsymbol{A}\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{A}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})}$ to denote the spectral and Frobenius norms of the matrix $\boldsymbol{A} \in \mathbb{R}^{K\times K}$, respectively, where tr is the trace operator. We use $\boldsymbol{A} \succeq \boldsymbol{B}$ to show that for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} - \boldsymbol{B}$ is semi-positive definite. We denote the d-dimensional vector of ones as $\mathbf{1}_d$, and the identity matrix and $d^{-1}\mathbf{1}_d\mathbf{1}_d^T$ by $\boldsymbol{I}$ and $\boldsymbol{J}$, respectively. We assume that the parameter space $\boldsymbol{\Phi}$ is convex and closed, and denote $R_d := \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 < \infty$ as the diameter of $\boldsymbol{\Phi}$ when it is bounded. By $\mathrm{M}_{\boldsymbol{\Phi}}$, we denote the metric projection operator onto the parameter space $\boldsymbol{\Phi}$, that is, $\mathrm{M}_{\boldsymbol{\Phi}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}'\in\boldsymbol{\Phi}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$. Note that the metric projection operator is a nonexpanding operator, that is,

$$\left\|\mathrm{M}_{\boldsymbol{\Phi}}(\boldsymbol{x}) - \mathrm{M}_{\boldsymbol{\Phi}}(\boldsymbol{x}')\right\|_2 \leq \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_2 \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d.$$

We define similarily the matrix form of the metric projection operator $\mathrm{M}_{\boldsymbol{\Phi}}$, that is, for $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K) \in \mathbb{R}^{d\times K}$, $\mathrm{M}_{\boldsymbol{\Phi}}(\boldsymbol{X}) = (\mathrm{M}_{\boldsymbol{\Phi}}(\boldsymbol{x}_1), \mathrm{M}_{\boldsymbol{\Phi}}(\boldsymbol{x}_2), \ldots, \mathrm{M}_{\boldsymbol{\Phi}}(\boldsymbol{x}_K))$.

In the following, we first define the connection network, upon which the DFL algorithm is formally stated. Then, we review the statistical properties of the classical nondistributed SGD algorithm to prepare for the theoretical results of the DFL algorithm.

2.1. *The connection network.* The connection network of the participating clients in the FL system is defined by an undirected graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ where $\mathrm{V} = \{v_k\}_{k=1}^K$ represents the set of clients and E specifies the edge set such that $(i, j) \in \mathrm{E}$ if and only if clients i and j are connected. We assume that there is a self-loop for each client (node) such that $(i, i) \in \mathrm{E}$ for $1 \leq i \leq K$. Let $\boldsymbol{C} = (c_{ij}) \in \mathbb{R}^{K\times K}$ be a symmetric connection matrix defined on $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, where c_{ij} is a nonnegative constant that specifies the contribution of the jth data block to the estimation at node i. It is required that $c_{ij} > 0$ if and only if $(i, j) \in \mathrm{E}$ and $\sum_{j=1}^K c_{ij} = 1$ for all i. An example of the connection matrix is the Metropolis–Hastings (MH) rule (Boyd et al. (2006)), which has

$$c_{ij} = \begin{cases} 0 & \text{if } (i, j) \notin E, \\ \left(\max\{d_i, d_j\}\right)^{-1} & \text{if } (i, j) \in \mathrm{E} \text{ and } i \neq j, \\ 1 - \displaystyle\sum_{s=1, s\neq i}^{K} c_{i,s} & \text{if } i = j, \end{cases} \tag{5}$$

where d_i is the number of connected neighbors of node i (out-degree). Figure 1 illustrates a decentralized FL system with 6 nodes and the corresponding connection matrix $\boldsymbol{C}$.

2.2. *The DFL algorithm.* Given a connection matrix $\boldsymbol{C}$, we are to solve the FL problem (4) by a DFL algorithm designed for the decentralized FL that extends the classical SGD by allowing multiple local SGD steps between two rounds of communication among the neighboring clients as specified by $\boldsymbol{C}$. Let the local parameter estimate conducted on the kth data block at the tth step of the algorithm be $\hat{\boldsymbol{\theta}}_t^k$, the corresponding matrix of estimates of all clients be $\widehat{\boldsymbol{\Theta}}_t = (\hat{\boldsymbol{\theta}}_t^1, \hat{\boldsymbol{\theta}}_t^2, \ldots, \hat{\boldsymbol{\theta}}_t^K) \in \mathbb{R}^{d\times K}$, the step size be η_t and the weighted stochastic gradient matrix be

$$\widehat{\boldsymbol{G}}_t = K\big(w_1\nabla f_1(\hat{\boldsymbol{\theta}}_{t-1}^1; \boldsymbol{\xi}_t^1), w_2\nabla f_2(\hat{\boldsymbol{\theta}}_{t-1}^2; \boldsymbol{\xi}_t^2), \ldots, w_K\nabla f_K(\hat{\boldsymbol{\theta}}_{t-1}^K; \boldsymbol{\xi}_t^K)\big). \tag{6}$$

Here for each k, $\{\boldsymbol{\xi}_t^k\}_{t\geq 1}$ are drawn independently and sequentially at each step t from the distribution $\mathcal{P}_k$. The DFL algorithm (summarized as Algorithm 1 in the SM) proceeds as follows. At $t = 0$, all the local estimates are initialized as $\hat{\boldsymbol{\theta}}_0 \in \mathbb{R}^d$. For $t = 1, 2, \ldots, T$ and some positive integer τ, if t is divisible by τ, there is a synchronization of the computation results among neighboring nodes according to $\boldsymbol{C}$, and the parameter estimates are updated as $\widehat{\boldsymbol{\Theta}}_t = \mathrm{M}_{\boldsymbol{\Phi}}((\widehat{\boldsymbol{\Theta}}_{t-1} - \eta_t\widehat{\boldsymbol{G}}_t)\boldsymbol{C})$; otherwise we update the parameter estimates locally and in parallel by $\widehat{\boldsymbol{\Theta}}_t = \mathrm{M}_{\boldsymbol{\Phi}}(\widehat{\boldsymbol{\Theta}}_{t-1} - \eta_t\widehat{\boldsymbol{G}}_t)$. By allowing $\tau > 1$, the DFL algorithm reduces the total communication cost by $(1-1/\tau)\times 100\%$ as compared with that of the classical SGD ($\tau = 1$). The communication among the neighboring clients happens at $\mathcal{I} = \{t \in \mathbb{N}_+ | t = s\tau, s \in \mathbb{N}_+\}$ where $\mathbb{N}_+$ denotes the set of positive integers.

REMARK. There are three types of averaging in the DFL. The first one is the spatial averaging across the K clients at each step:

$$\bar{\hat{\boldsymbol{\theta}}}_t = K^{-1}\sum_{k=1}^K \hat{\boldsymbol{\theta}}_t^k;$$

the second type is the temporal averaging within a client k for $1 \leq t \leq T$:

$$\bar{\hat{\boldsymbol{\theta}}}_T^k = T^{-1}\sum_{t=1}^T \hat{\boldsymbol{\theta}}_t^k;$$

and the last type is the spatial-temporal averaging

$$\bar{\bar{\hat{\boldsymbol{\theta}}}}_T = (TK)^{-1}\sum_{t=1}^T\sum_{k=1}^K \hat{\boldsymbol{\theta}}_t^k.$$

The first one is infeasible in the DFL due to a lack of full communication. Starting from $\bar{\hat{\boldsymbol{\theta}}}_1^k := \hat{\boldsymbol{\theta}}_1^k$, the second one can be updated locally and recursively by $\bar{\hat{\boldsymbol{\theta}}}_{t+1}^k = \{t\bar{\hat{\boldsymbol{\theta}}}_t^k + \hat{\boldsymbol{\theta}}_{t+1}^k\}/(t+1)$ for $1 \leq t \leq T-1$, and the last one can be obtained via one round of full synchronization since $\bar{\bar{\hat{\boldsymbol{\theta}}}}_T = K^{-1}\sum_{k=1}^K \bar{\hat{\boldsymbol{\theta}}}_T^k$. In fact, $\bar{\bar{\hat{\boldsymbol{\theta}}}}_T$ is the PR-averaged estimator in the DFL algorithm which is a focus of Section 4.

The difficulty in analyzing the estimates trajectory of the DFL algorithm is largely due to the decentralized structure, since at each step t of the algorithm, (i) the starting values $\{\hat{\boldsymbol{\theta}}_{t-1}^k\}_{k=1}^K$ are different, and (ii) the updating directions $\{\nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^k; \boldsymbol{\xi}_t^k)\}_{k=1}^K$ are different. The following assumption on the connection matrix $\boldsymbol{C}$ is needed for consistent estimation of $\boldsymbol{\theta}_K^*$ by the DFL algorithm, and was proposed in Boyd et al. (2006).

ASSUMPTION 2.1. The K-dimensional connection matrix $\boldsymbol{C}$ satisfies $\boldsymbol{C}\mathbf{1} = \mathbf{1}$ and $\boldsymbol{C}^T = \boldsymbol{C}$ whose largest eigenvalue is 1 and the absolute values of other eigenvalues are strictly less than 1, namely $\max\{|\lambda_k(\boldsymbol{C})| | k = 2, 3, \ldots, K\} \le \rho < \lambda_1(\boldsymbol{C}) = 1$ for some $0 \le \rho < 1$, where $\lambda_k(\boldsymbol{C})$ denotes the kth largest eigenvalue of $\boldsymbol{C}$.

REMARK. This condition is sufficient and necessary to ensure $\lim_{s\to\infty} \boldsymbol{C}^s = K^{-1}\mathbf{1}_K\mathbf{1}_K^T$. Specifically, the case of $\rho = 0$ corresponds to the centralized FL scenario with $\boldsymbol{C} = K^{-1}\mathbf{1}_K\mathbf{1}_K^T$. Applying this result to the DFL algorithm,

$$\lim_{s\to\infty} \widehat{\boldsymbol{G}}_t \boldsymbol{C}^s = \left(\sum_{k=1}^K w_k \nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^k; \boldsymbol{\xi}_t^k)\right)\mathbf{1}_K^T.$$

This implies that the local updates made by $\widehat{\boldsymbol{G}}_t$ to the K local estimates at step t, after sufficient rounds of local averaging, are asymptotically equal to $\sum_{k=1}^K w_k \nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^k; \boldsymbol{\xi}_t^k)$. If one can properly control the consensus error at the tth step $K^{-1}\sum_{k=1}^K \mathrm{E}(\|\hat{\boldsymbol{\theta}}_{t-1}^k - \hat{\bar{\boldsymbol{\theta}}}_t\|_2^2)$, where $\hat{\bar{\boldsymbol{\theta}}}_{t-1} = K^{-1}\sum_{k=1}^K \hat{\boldsymbol{\theta}}_{t-1}^k$, then $\sum_{k=1}^K w_k \nabla F_k(\hat{\boldsymbol{\theta}}_{t-1}^k)$ can approximate $\nabla F(\hat{\bar{\boldsymbol{\theta}}}_{t-1})$, the gradient of the FL risk function (1) evaluated at $\boldsymbol{\theta} = \hat{\bar{\boldsymbol{\theta}}}_{t-1}$. Hence, the tth local update $-\eta_t \widehat{\boldsymbol{G}}_t$ can be viewed as performing a gradient descent step with $-\eta_t \nabla F(\hat{\bar{\boldsymbol{\theta}}}_{t-1})$ starting from $\hat{\bar{\boldsymbol{\theta}}}_{t-1}$ for all of the local estimates. Thus, according to the standard theory of the gradient descent algorithm, the sequence $\{\hat{\bar{\boldsymbol{\theta}}}_t\}_{t\ge1}$ will converge to $\boldsymbol{\theta}_K^*$ (Bottou, Curtis and Nocedal (2018), Choi and Kim (2022)). In the next section, we will rigorously establish the above argument.

2.3. *Properties of conventional SGD.* The SGD was introduced by Robbins and Monro (1951) as a method of stochastic approximation. The task was to find the root θ^* of the equation $\nabla F_1(\theta) = 0$ with $\theta \in \mathbb{R}$, which is equivalent to finding the minimizer of $F_1(\theta)$. For each θ, instead of knowing the value of $\nabla F_1(\theta)$, suppose we can perform a statistical experiment at θ with a response $Y_\theta = \nabla f_1(\theta; \boldsymbol{\xi}^1)$, where $\boldsymbol{\xi}^1$ is sampled from distribution $\mathcal{P}_1$ such that $\mathrm{E}(Y_\theta) = \nabla F_1(\theta)$. The Robbins–Monro (RM) procedure for estimating θ^* starts from an initial estimate $\hat{\theta}_0$ of θ^* and updates recursively

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \eta_t Y_{\hat{\theta}_t}, \tag{7}$$

where η_t is the step size and $Y_{\hat{\theta}_t} = \nabla f_1(\hat{\theta}_t; \boldsymbol{\xi}_{t+1}^1)$, and $\{\boldsymbol{\xi}_t^1\}_{t\ge1}$ are sequentially from $\mathcal{P}_1$.

A proper choice of the step size $\{\eta_t\}_{t\ge1}$ is critical to the convergence of the RM procedure. The step size η_t should diminish sufficiently quickly so that the variance of the stochastic gradients $Y_{\hat{\theta}_t}$ does not affect the convergence. At the same time, it should not diminish too quickly so that $\sum_{t=1}^\infty \eta_t < \infty$, since under such scenario the SGD iterates are not guaranteed to converge to the true θ^*. Specifically, it is required that

$$\sum_{t=1}^\infty \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^\infty \eta_t^2 < \infty. \tag{8}$$

When η_t is of the order $O(t^{-\alpha})$, the above requirement corresponds to $1/2 < \alpha \le 1$, under which Chung (1954) established the following asymptotic properties of the RM procedure. Let $\eta_t = Dt^{-\alpha}$ for some $D > 0$ and T be the total number of sequentially sampled observations, then when either $1/2 < \alpha < 1$ or $\alpha = 1$ with $D > 1/(2\nabla^2 F_1(\theta^*))$,

$$T^{\alpha/2}\big(\hat{\theta}_T - \theta^*\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2(\alpha, D)\big) \quad \text{as } T \to \infty, \quad \text{where}$$
$$\sigma^2(\alpha, D) = \begin{cases} D\sigma_{\theta^*}^2/\big(2\nabla^2 F_1(\theta^*)\big) & \text{if } 1/2 < \alpha < 1, \\ D^2\sigma_{\theta^*}^2/\big(2\nabla^2 F_1(\theta^*)D - 1\big) & \text{if } \alpha = 1, \end{cases} \tag{9}$$

and $\sigma_\theta^2 = \mathrm{Var}(\nabla f_1(\theta; \boldsymbol{\xi}^1))$. These suggest that to achieve the same $\sqrt{T}$-convergence rate for the estimator $\hat{\theta}_T$ as the regular M-estimator based on T i.i.d. observations, we require a small step size namely $\alpha = 1$. To further achieve statistical efficiency for $\hat{\theta}_T$ when $\alpha = 1$, (9) suggests choosing $D = 1/\nabla^2 F_1(\theta^*)$ as it minimizes $\sigma^2(1, D)$. This optimal D requires estimating $\nabla^2 F(\theta^*)$, which is the focus of Lai (2003) via the adaptive estimation. It is noted that $\sigma^2(1, 1/\nabla^2 F_1(\theta^*))$ is the same as the asymptotic variance of the full sample M-estimator.

Instead of using the last iteration $\hat{\theta}_T$, Ruppert (1988) and Polyak and Juditsky (1992) suggested to use the average of the SGD trajectory $\hat{\bar{\boldsymbol{\theta}}}_T = T^{-1} \sum_{t=1}^T \hat{\boldsymbol{\theta}}_t$ (the so-called Polyak–Ruppert (PR) averaging) to estimate θ^*. The intuition is that when the step size of the SGD trajectory is large, the temporal correlation among the sequence $\{\hat{\theta}_t\}_{t=0}^{T-1}$ is weak, and averaging the trajectory of the estimates may improve the statistical efficiency. Both studies proved that when $1/2 < \alpha < 1$

$$\sqrt{T}(\hat{\bar{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(0, \nabla^2 F_1(\boldsymbol{\theta}^*)^{-1} \mathrm{Cov}(\nabla f_1(\boldsymbol{\theta}^*; \boldsymbol{\xi}^1)) \nabla^2 F_1(\boldsymbol{\theta}^*)^{-1}) \quad \text{as } T \to \infty, \tag{10}$$

for multivariate $\boldsymbol{\theta}^*$. This means that the PR-averaged estimator $\hat{\bar{\boldsymbol{\theta}}}_T$ can achieve the statistical efficiency without the second-order derivative (Hessian) information. However, when a small step size ($\alpha = 1$) is applied in the SGD algorithm, although $\hat{\bar{\boldsymbol{\theta}}}_T$ is still $\sqrt{T}$-consistent, it is in general no longer efficient due to the strong serial correlation.

There are works focusing on inference for the parameter of interest based on the nondistributed SGD (Fang, Xu and Yang (2018), Chen et al. (2020a), Zhu, Chen and Wu (2023), Lee et al. (2022)). However, these studies are not applicable to the decentralized FL problem. We will investigate both the finite and asymptotic properties of the SGD iteration in DFL, and establish the corresponding statistical inference procedure in the current new context.

Following the literature on the SGD-based estimation, we impose the following assumption on the step sizes in the DFL algorithm.

ASSUMPTION 2.2. The step sizes $\{\eta_t\}_{t \geq 1}$ in the DFL algorithm satisfy $\eta_t = D(t + \gamma)^{-\alpha}$ for some positive constants D, γ, and $1/2 < \alpha \leq 1$.

This is a standard condition on the decaying rate of the step sizes, which satisfies the classical constraint (8), and has been assumed in the existing SGD-based inference literature (Chen et al. (2020a, 2024)).

3. Nonasymptotic analysis of SGD iteration in DFL. First, we establish an upper bound of the consensus error $K^{-1} \sum_{k=1}^K \mathrm{E}(\|\hat{\boldsymbol{\theta}}_t^k - \hat{\bar{\boldsymbol{\theta}}}_t\|_2^2)$ of the DFL algorithm, characterizing the deviation of the local estimators $\{\hat{\boldsymbol{\theta}}_t^k\}_{k=1}^K$ to their average $\hat{\bar{\boldsymbol{\theta}}}_t$. Based on the bound, we derive an upper bound of the mean squared error (MSE) $\mathrm{E}(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2)$ of $\hat{\bar{\boldsymbol{\theta}}}_t$, generalizing the upper bound for the nondistributed SGD (Bottou, Curtis and Nocedal (2018)).

3.1. *Consensus error bound.* The following assumptions are needed to establish the upper bound of the consensus error.

ASSUMPTION 3.1. There exists nonnegative constants $L_{\boldsymbol{\xi}}$ and σ^2, and a positive integer v such that the gradient noise $\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k) = \nabla f_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k) - \nabla F_k(\boldsymbol{\theta})$ satisfies $\mathrm{E}(\|\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)\|_2^{2s}) \leq \sigma^{2s} + L_{\boldsymbol{\xi}} \|\nabla F_k(\boldsymbol{\theta})\|_2^{2s}$ for all positive integers $s \leq v$, $\boldsymbol{\theta} \in \boldsymbol{\Phi}$, and $k = 1, 2, \ldots, K$.

ASSUMPTION 3.2. For $k = 1, 2, \ldots, K$, the objective function $F_k(\cdot)$ is differentiable, convex and L-smooth with a positive constant L such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Phi}$,

$$0 \leq F_k(\boldsymbol{\theta}_1) - F_k(\boldsymbol{\theta}_2) - \langle \nabla F_k(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \leq \frac{L}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2. \tag{11}$$

ASSUMPTION 3.3. There exist nonnegative constants κ and B such that $\sum_{k=1}^K w_k \|\nabla F(\boldsymbol{\theta}) - \nabla F_k(\boldsymbol{\theta})\|_2^2 \le \kappa^2$ for any $\boldsymbol{\theta} \in \boldsymbol{\Phi}$, and $\|\nabla F_k(\boldsymbol{\theta}_K^*)\|_2 \le B$ for all $1 \le k \le K$.

Assumption 3.1 controls the variability of the gradient noise $\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)$. It is noted that $v = 1$ is enough for establishing the nonasymptotic bounds and asymptotic normality of the DFL estimators. However, we need $v = 2$ to attain the consistency of the asymptotic covariance matrix estimator for the construction of the confidence region of the parameter $\boldsymbol{\theta}_K^*$ when the loss function f_k is not second-order differentiable as established in Theorem 5. When $v = 1$, the σ^2 term allows the variance of $\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)$ to be nonzero at any stationary point of $F_k(\cdot)$. Combined with Assumption 3.2, the $L_\xi \|\nabla F_k(\boldsymbol{\theta})\|_2^2$ term scales quadratically with the Euclidean distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_K^*$, which relaxes the bounded variance condition required in Lian et al. (2017) and Koloskova, Stich and Jaggi (2019) for the decentralized estimation. Besides, Nguyen et al. (2019) proved that Assumption 3.1 holds if $F_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)$ is convex with respect to $\boldsymbol{\theta}$ given $\boldsymbol{\xi}^k$. Assumption 3.2 is a standard assumption in stochastic optimization, and is satisfied for many statistical estimation tasks including linear regression and logistic regression. See Bottou, Curtis and Nocedal (2018) for more discussions.

A challenge for analysing the statistical properties of the decentralized FL is how to control the consensus error $K^{-1}\mathrm{E}(\|\widehat{\boldsymbol{\Theta}}_t(\boldsymbol{I} - \boldsymbol{J})\|_F^2)$, since only the local averaging is allowed at every $\tau \ge 1$ steps. To tackle the challenge, we first establish the following lemma.

LEMMA 1. *Under Assumptions 2.1–2.2, 3.1 with $v = 1$, and Assumptions 3.2–3.3, there exist two positive constants B_{MSE} and B_{CE} such that for all $1 \le t \le T$ and $K \ge 1$, the following hold regardless of the parameter space $\boldsymbol{\Phi}$ is bounded by a diameter $R_d < \infty$ or is $\mathbb{R}^d$:*

$$\mathrm{E}\big(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2\big) < B_{\mathrm{MSE}} \quad \text{and} \quad K^{-1}\mathrm{E}\big(\|\widehat{\boldsymbol{\Theta}}_t(\boldsymbol{I} - \boldsymbol{J})\|_F^2\big) < B_{\mathrm{CE}}. \tag{12}$$

REMARK. The boundedness of the above terms is trivial when the parameter space $\boldsymbol{\Phi}$ is also bounded. The case when $\boldsymbol{\Phi} = \mathbb{R}^d$ is more complicated as the cumulative step size $\sum_{s=1}^t \eta_s$ diverges to infinity. This lemma implies that the expected norms of the stochastic gradient matrix scaled by $K^{-\frac{1}{2}}$ is bounded, that is, $K^{-1}\mathrm{E}(\|\widehat{\boldsymbol{G}}_t\|_F^2) < C_3$ for some positive constant $C_3 > 0$. Thus, the whole process driven by the DFL algorithm is bounded.

THEOREM 1. *Under Assumptions required in Lemma 1 and for all $1 \le \tau \le t \le T$ and $K \ge 1$, the consensus error in the DFL algorithm satisfies*

$$\begin{aligned}\frac{1}{K}\mathrm{E}\big(\|\widehat{\boldsymbol{\Theta}}_t(\boldsymbol{I} - \boldsymbol{J})\|_F^2\big) \le{}& 3b_2^2 Q\left(2(L_\xi + 1)\delta(t, \tilde{\rho}^2, \tau) + \left((\tau - 1) + \frac{\mathbb{I}_{\{\tilde{\rho} > 0\}}}{1 - \tilde{\rho}}\right)\delta(t, \tilde{\rho}, \tau)\right) \\ &+ 2\sigma^2 b_2^2 \delta\big(t, \tilde{\rho}^2, \tau\big),\end{aligned} \tag{13}$$

where b_2 is the constant such that $w_k K \le b_2$ for all k, $\tilde{\rho} = \rho^{\frac{1}{\tau}}$,

$$Q = \begin{cases} \kappa^2 + 2L^2 R_d^2 & \text{if } \boldsymbol{\Phi} \text{ is bounded by diameter} \\ & R_d = \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Phi} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 < \infty, \\ \kappa^2 + L^2(B_{\mathrm{MSE}} + B_{\mathrm{CE}}) & \text{if } \boldsymbol{\Phi} = \mathbb{R}^d, \end{cases}$$

$$\delta(t, a, \tau) := \left(\sum_{s=t-\tau+2}^{t} \eta_s^2\right) + \frac{\mathbb{I}_{\{a>0\}}}{1 - a}\big(\eta_1^2 a^{(t-\tau+1)/2} + \eta_{\lfloor (t-\tau+1)/2 \rfloor}^2\big)$$

for any $a \in (0, 1)$, and B_{MSE} and B_{CE} are constants defined in Lemma 1.

REMARK. It is noted that the diminishing rate of the consensus error bound (13) in Theorem 1 is determined by the order of $\delta(t,\tau,a)$ (14) with $a \in \{\tilde{\rho}, \tilde{\rho}^2\}$, which consists of three terms. First, it is straightforward to see that both the first term $\sum_{s=t-\tau+2}^{t} \eta_s^2$ and the third term $\eta^2_{\lfloor (t-\tau+1)/2 \rfloor}$ are $O(\eta_t^2)$. Besides, since $\tilde{\rho} < 1$, it is easy to show that the second term $\tilde{\rho}^{t/2} = o(\eta^2_{\lfloor t/2 \rfloor})$ as t increases. Thus, this theorem informs that there exists a universal constant c_0 such that the consensus error in the DFL algorithm $K^{-1}\mathrm{E}(\|\widehat{\boldsymbol{\Theta}}_t(\boldsymbol{I}-\boldsymbol{J})\|_F^2) \le c_0\eta_t^2$ for all $1 \le t \le T$. In contrast, the existing studies on DFL are usually based on a constant step size (Wang and Joshi (2021)), under which scenario the SGD iteration is biased.

The consensus error upper bound (13) is affected by hyperparameters including η_t, τ and ρ. It becomes larger for larger step size $\eta_t = D(t+\gamma)^{-\alpha}$ with $\alpha < 1$. To see the roles of ρ and τ, we note that ρ is mainly related to the last two terms of $\delta(t,\rho,\tau)$ through $\mathbb{I}_{\{\rho>0\}}(1-\rho)^{-1}$, which becomes larger as $\rho \to 1$. On the other hand, the network becomes denser as $\rho \to 0$. When $\rho = 0$, the DFL algorithm reduces to the centralized version, and the last two terms vanish. The local update parameter τ mainly contributes to the first term of $\delta(t,\rho,\tau)$, and all the past $\tau-1$ step sizes contributes to the current consensus error. When $\tau = 1$, the first term vanishes. The local update procedure also enlarges the upper bound when $\tau > 1$ through transforming ρ into $\tilde{\rho} = \rho^{1/\tau} > \rho$. From this perspective, the per τ-steps local update procedure via the connection matrix $\boldsymbol{C}$ is approximately equivalent to applying the classical decentralized SGD (without local update) (Lian et al. (2017)) under a sparser network with the connection matrix $\tilde{\boldsymbol{C}} = \boldsymbol{C}^{1/\tau}$, as a larger ρ means a sparser connection network (Nedić, Olshevsky and Rabbat (2018)). Finally, when $\tau = 1$ and $\rho = 0$, the upper bound in (13) equals zero, which perfectly matches the definition of the consensus error.

3.2. *MSE bounds of the DFL sequence*. In addition to the consensus error, the mean squared error of the averaged sequence $\{\hat{\bar{\boldsymbol{\theta}}}_t\}_{t\ge1}$ starting from an initial value $\hat{\boldsymbol{\theta}}_0$ is also of interest. We have shown in Theorem 1 that the local estimators $\{\hat{\boldsymbol{\theta}}_t^k\}_{k=1}^K$ gradually concentrate to their spatial average $\hat{\bar{\boldsymbol{\theta}}}_t$ at the rate of the step size η_t. However, the theorem only reflects the variability of the local estimators across the clients due to the decentralized structure of the network and the local update procedure applied in the DFL algorithm, and does not imply the consistency of either the local estimators $\{\hat{\boldsymbol{\theta}}_T^k\}_{k=1}^K$ or their spatial average $\hat{\bar{\boldsymbol{\theta}}}_T$ as the local sample size T increases. In fact, one can only establish the boundedness of $\sum_{t=0}^{T}\eta_{t+1}\mathrm{E}(\|\nabla F(\hat{\bar{\boldsymbol{\theta}}}_t)\|_2^2)$ (see Lemma S4 in the SM), which suggests that the expected gradient norms can not be bounded away from zero, matching results established for the nondistributed SGD (Theorems 4.9 and 4.10 in Bottou, Curtis and Nocedal (2018)).

To establish an upper bound of $\mathrm{E}(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2)$ for $1 \le t \le T$ and $K > 1$, a condition stronger than convexity is necessary. We impose the following condition which deals with both the unbounded ($\boldsymbol{\Phi} = \mathbb{R}^d$) and bounded parameter space ($R_d < \infty$), respectively.

ASSUMPTION 3.4. For all $k = 1, 2, \ldots, K$, $F_k(\cdot)$ is differentiable and (strongly-)convex, and let μ be the corresponding largest nonnegative constant such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$

$$\frac{\mu}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 \le F_k(\boldsymbol{\theta}_1) - F_k(\boldsymbol{\theta}_2) - \langle \nabla F_k(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\rangle. \tag{14}$$

If $\mu > 0$, the parameter space $\boldsymbol{\Phi}$ is allowed to be unbounded, for instance, $\mathbb{R}^d$. If $\mu = 0$, $\boldsymbol{\Phi}$ is required to be bounded with $R_d = \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\Phi}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 < \infty$, and the federated objective function $F(\cdot) = \sum_{k=1}^K w_k F_k(\cdot)$ satisfies a generalized self-concordance property. The latter means that there exist positive constants μ_* and B_{G} such that:

(i) $F(\cdot)$ is three-times differentiable, and $\nabla^2 F(\boldsymbol{\theta}_K^*) \succeq \mu_* \boldsymbol{I}$;

(ii) for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Phi}$, $\varphi'''(t) \le B_{\mathrm{G}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \varphi''(u)$, where $\varphi : u \mapsto F(\boldsymbol{\theta}_1 + u(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1))$ for $u \in \mathbb{R}$;

(iii) $\|\nabla F(\boldsymbol{\theta})\|_2 \le B_{\mathrm{G}}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Phi}$.

REMARK. The global strong convexity property assumed in Assumption 3.4 with $\mu > 0$ is critical for establishing an upper bound of the following type in nondistributed SGD $(K, \tau = 1)$:

$$\mathrm{E}(\|\hat{\boldsymbol{\theta}}_t^1 - \boldsymbol{\theta}_1^*\|_2^2) \le C\eta_t \tag{15}$$

for all $1 \le t \le T$, where C is a positive constant. The upper bound is obtained by iterating back the following inequality to the initial estimates:

$$\mathrm{E}(\|\hat{\boldsymbol{\theta}}_t^1 - \boldsymbol{\theta}_1^*\|_2^2) \le (1 - \mu\eta_t)\mathrm{E}(\|\hat{\boldsymbol{\theta}}_{t-1}^1 - \boldsymbol{\theta}_1^*\|_2^2) + C'\eta_t^2, \tag{16}$$

where $\mu > 0$ is the corresponding global convexity parameter and C' is a positive constant. While $\mu > 0$ has been assumed in existing works (Bottou, Curtis and Nocedal (2018), Chen et al. (2020a, 2024)), and is satisfied in many estimation problems including the linear regression, it fails to hold for tasks such as the logistic regression. To fix this issue, we restrict the parameter space to be bounded with $R_d = \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Phi} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 < \infty$. Moreover, we introduce the generalized self-concordance property in Assumption 3.4 for $\mu = 0$, which has been assumed in Bach (2010, 2014) in the analysis of nondistributed SGD estimation for the logistic regression with a constant step size. In particular, the condition (ii) is quite nonstandard, requiring that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Phi}$, $\varphi'''(t) \le B_{\mathrm{G}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \varphi''(u)$, where $\varphi : u \mapsto F(\boldsymbol{\theta}_1 + u(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1))$ for $u \in \mathbb{R}$. This condition properly controls the third-order derivative of the objective function with its second-order derivative. The next theorem 2 first establishes an iterative bound (18) of the mean squared error $\mathrm{E}(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2)$, similar to (16). The theorem needs us to define

$$\mu_{R_d} = \begin{cases} \left(\dfrac{\mu}{2L} + \dfrac{1}{2}\right)\mu & \text{if } \mu > 0 \text{ and} \\ \mu_*^2 \left(2L\left(4 + \dfrac{16 B_{\mathrm{G}}^2 R_d^2}{9}\right)\right)^{-1} & \text{if } \mu = 0, \end{cases} \tag{17}$$

where μ and L are defined in (14) and Assumption 3.2, respectively. It also need constants μ^* and B_{G} defined in Assumption 3.4 (i) and (ii), respectively. When the global strong convexity parameter $\mu = 0$, μ_{R_d} decays quadratically with respect to the diameter R_d of the parameter space $\boldsymbol{\Phi}$. Besides, it only depends on the Hessian of the population objective function at the true parameter instead of the whole parameter space $\boldsymbol{\Phi}$. We will see explicitly the benefit of the generalized self-concordance property in (i) of Theorem 2.

THEOREM 2. *Under Assumptions 2.1–2.2, 3.1 with $v = 1$, and Assumptions 3.2–3.4, let $\boldsymbol{\Delta}_t = \hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*$, then the following results hold.*

(i) *For all $0 \le t < T$ and $K \ge 1$,*

$$\mathrm{E}(\|\boldsymbol{\Delta}_{t+1}\|_2^2) \le (1 - \mu_{R_d}\eta_{t+1})\mathrm{E}(\|\boldsymbol{\Delta}_t\|_2^2) + \eta_{t+1}\left(c_1\frac{\eta_{t+1}}{K} + c_2\frac{1}{K}\mathrm{E}(\|\widehat{\boldsymbol{\Theta}}_t(\boldsymbol{I} - \boldsymbol{J})\|_F^2)\right), \tag{18}$$

where $c_1 = b_2\sigma^2 + 3b_2^3 L_\xi \kappa^2$ and $c_2 = 3b_2(L + \mu)$.

(ii) *If $D > 2/\mu_{R_d}$ and $\gamma > 0$ such that $\eta_1 \le D/(D\mu_{R_d} - 1)$, and let c_0 be a positive constant such that $K^{-1}\mathrm{E}(\|\widehat{\boldsymbol{\Theta}}_t(\boldsymbol{I} - \boldsymbol{J})\|_F^2) \le c_0\eta_t^2$ as implied in Theorem 1, then*

$$\mathrm{E}(\|\boldsymbol{\Delta}_t\|_2^2) \le v_1\frac{\eta_t}{K} + v_2\eta_t^2, \tag{19}$$

where $v_1 = c_1 D/(D\mu_{R_d} - 1)$ and $v_2 = \max\{c_2 c_0 D/(D\mu_{R_d} - 2), (\gamma + 1)^2\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_K^*\|_2^2 D^{-2}\}$.

(iii) *For each fixed K, $\hat{\bar{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}_K^* \to \mathbf{0}_d$ almost surely as $T \to \infty$.*

It is noted that (i) of Theorem 2 establishes an iterative bound similar to (16) but for the DFL. First, if we only restrict the parameter space to be bounded when $\mu = 0$ without assuming the generalized self-concordance property, then μ_{R_d} in (18) will be replaced by $\tilde{\mu}_{R_d}$, which is

$$\tilde{\mu}_{R_d} = \begin{cases} \left(\dfrac{\mu}{2L} + \dfrac{1}{2}\right)\mu & \text{if } \mu > 0 \text{ and} \\ \sup\{c > 0 | \nabla^2 F(\boldsymbol{\theta}) \succeq c\boldsymbol{I} \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Phi}\} & \text{if } \mu = 0. \end{cases}$$

When $\mu = 0$, it can be shown that $\tilde{\mu}_{R_d}$ decays exponentially fast with respect to the diameter R_d in the logistic regression problem. In comparison, the bound (18) is much tighter with μ_{R_d}, as μ_{R_d} has an explicit quadratic dependence on the inverse of R_d and only depends on the eigenvalue of $\nabla^2 F(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_K^*$. It can also be seen that due to the consensus error in the DFL, there is an extra term in (18) compared with the nondistributed counterpart (15). This leads to the extra $v_2\eta_T^2$ term in (19) of the (ii) part of the theorem compared with the MSE bound of the nondistributed SGD iteration (Bottou, Curtis and Nocedal (2018)). The extra term is asymptotically negligible compared with the leading $v_1\eta_T/K$ term as the local sample size T increases to infinity if K is finite. However, as the FL is developed largely for model training for large distributed systems, it is appropriate to allow K to increase with T at some rate. In this case, $v_2\eta_T^2$ can dominate the upper bound in (19) when K increases faster than T^α. We will show how the consensus error affect the asymptotic properties of the PR-averaged estimator $\bar{\bar{\boldsymbol{\theta}}}_T$ in Theorem 3 in the next section.

The decentralized structure $\boldsymbol{C}$ affects the MSE bound only through c_0, and is thus a second-order effect for moderate K. Besides, it can be seen that the initialization error is also of second-order, as it only appears in the definition of v_2. Moreover, the heterogeneity factor κ^2 enlarges the v_1 term when $L_{\boldsymbol{\xi}} > 0$. However, if all the gradient noise has bounded variance σ^2 so that $L_{\boldsymbol{\xi}} = 0$, the heterogeneity effect is confined to the second-order.

In part (iii), the almost sure convergence of the averaged estimator $\bar{\bar{\boldsymbol{\theta}}}_T$ for fixed K can be viewed as a generalization of the Robbins and Siegmund (1971) result for the nondistributed SGD, which is based on the martingale convergence theorem. However, for K diverging with T, the almost sure convergence property no longer holds in general since the Doob's upcrossing inequality (Hall and Heyde (1980)) does not have a triangular array version.

4. Online statistical inference for the Polyak–Ruppert (PR) procedure in DFL. Recall that in the nondistributed SGD (7), the PR procedure achieves efficiency by first adopting a larger step size $\eta_t = O(t^{-\alpha})$ with $1/2 < \alpha < 1$ and then averaging the SGD trajectory, as presented in (10). For the DFL, the corresponding PR-averaged estimator is the spatial-temporal averaged estimator $\bar{\bar{\boldsymbol{\theta}}}_T = (TK)^{-1}\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_t^k$ as defined in Section 2. In this section, we will establish the asymptotic normality of the estimator $\bar{\bar{\boldsymbol{\theta}}}_T$ by allowing both K and T to diverge to infinity, and provide one-pass algorithms for statistical inference tasks.

4.1. *Asymptotic normality.* We first outline some assumptions.

ASSUMPTION 4.1. For $k = 1, 2, \ldots, K$, the objective function $f_k(\cdot;\cdot)$ is L-average smooth with a positive constant L_a such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Phi}$,

$$\mathrm{E}\big(\|\nabla f_k(\boldsymbol{\theta}_1; \boldsymbol{\xi}^k) - \nabla f_k(\boldsymbol{\theta}_2; \boldsymbol{\xi}^k)\|_2^2\big) \le L_a\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2. \tag{20}$$

This assumption is stronger than the smoothness condition in Assumption 3.2, and holds for common objective functions such as those for the linear regression, the ridge regression,

and the logistic regression if $\boldsymbol{\xi}^k$ has certain bounded moments; see the Supplementary Material of Su and Zhu (2018) for more details.

ASSUMPTION 4.2. There exist positive constants ℓ_{cov}, δ and C_e such that for all $k = 1, 2, \ldots, K$, $\boldsymbol{S}_k = \mathrm{E}(\boldsymbol{\epsilon}_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\boldsymbol{\epsilon}_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)^T)$ satisfies $\boldsymbol{S}_k \succeq \ell_{\mathrm{cov}}\boldsymbol{I}$ and $\mathrm{E}(\|\boldsymbol{\epsilon}(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\|_2^{2+\delta}) < C_e$ for all $K \geq 1$, where $\boldsymbol{\epsilon}(\boldsymbol{\theta}) = \sqrt{K}\sum_{k=1}^K w_k\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)$.

ASSUMPTION 4.3. The federated risk function $F(\boldsymbol{\theta}) = \sum_{k=1}^K w_k F_k(\boldsymbol{\theta})$ is second-order differentiable with respect to $\boldsymbol{\theta} \in \boldsymbol{\Phi}$, and the Hessian matrix $\nabla^2 F(\boldsymbol{\theta})$ is Lipschitz continuous at $\boldsymbol{\theta}_K^*$ in the sense that there exists a positive constant L_H such that $\|\nabla^2 F(\boldsymbol{\theta}) - \nabla^2 F(\boldsymbol{\theta}_K^*)\|_2 \leq L_H\|\boldsymbol{\theta} - \boldsymbol{\theta}_K^*\|_2$ for all $\boldsymbol{\theta} \in \boldsymbol{\Phi}$ and $K \geq 1$.

The above two assumptions are needed to establish the asymptotic normality of the PR-averaged estimator $\bar{\hat{\boldsymbol{\theta}}}_T$. In fact, the Lipschitz continuity of the Hessian matirx $\nabla^2 F(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_K^*$ is only needed when K increases to infinity with T, under which circumstance the averaged estimator $\bar{\hat{\boldsymbol{\theta}}}_T = K^{-1}\sum_{k=1}^K \hat{\boldsymbol{\theta}}_T$ in general does not possess the almost sure convergence property as discussed after Theorem 2.

THEOREM 3. *Under assumptions required in Theorem 2 and Assumptions 4.1, 4.2 and 4.3, if K is either finite or diverges at the rate $o(T^{2\alpha-1})$ with $\alpha < 1$ and $\sup_{K\geq 1}\|\boldsymbol{\theta}_K^*\|_2 < \infty$, we have*

$$\sqrt{TK}\boldsymbol{S}^{-1/2}\boldsymbol{H}(\bar{\hat{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}_K^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{as } T \to \infty, \tag{21}$$

where $\boldsymbol{H} = \nabla^2 F(\boldsymbol{\theta}_K^*)$ is the population Hessian matrix, $\boldsymbol{S} = \mathrm{E}(\boldsymbol{\epsilon}(\boldsymbol{\theta}_K^*)\boldsymbol{\epsilon}(\boldsymbol{\theta}_K^*)^T)$ is the covariance matrix of the aggregated gradient noise, and $\boldsymbol{\epsilon}(\boldsymbol{\theta})$ is defined in Assumption 4.2.

The $\sqrt{TK}$ convergence of $\bar{\hat{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}_K^*$ is the same as that of a regular M-estimator based on TK-order independent observations. Besides, the asymptotic variance of $\bar{\hat{\boldsymbol{\theta}}}_T$ is $\boldsymbol{H}^{-1}\boldsymbol{S}\boldsymbol{H}^{-1}$, which is the same as that of the full sample M-estimator (4).

While Theorem 3 shows that the Polyak–Ruppert averaging procedure can still achieve statistical efficiency when $1/2 < \alpha < 1$ in the DFL, this gain in statistical efficiency comes at a price in terms of a slower divergence rate for K. In the DFL scenario, a larger step size will lead to a larger bias (consensus error) among the local estimators $\{\hat{\boldsymbol{\theta}}_t^k\}_{k=1}^K$, which can not be eliminated by the averaging. As a consequence, the bias term could potentially dominate the would-be leading term that facilitates the asymptotic normality of the estimator $\bar{\hat{\boldsymbol{\theta}}}_T$, which in turn restricts the allowable increasing rate of K relative to T. We will discuss how to relax this $K = o(T^{2\alpha-1})$ restriction in Section 6. Before that, to facilitate statistical inference based on the asymptotic normality of the PR-averaged estimator $\bar{\hat{\boldsymbol{\theta}}}_T$, we need to construct asymptotically valid confidence regions. To this end, we need to consistently estimate both the covariance matrix of noise $\boldsymbol{S}$ and the aggregated Hessian matrix matrix $\boldsymbol{H}$.

4.2. *One-pass estimation of covariance matrix.* We begin by noting that

$$\boldsymbol{S} = K\sum_{k=1}^K w_k^2 \mathrm{E}\big(\boldsymbol{\epsilon}_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\boldsymbol{\epsilon}_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)^T\big),$$

which motivates us to estimate each of those $\boldsymbol{S}_k$ locally and combine them with only one round of global synchronization. Define

$$\widehat{\boldsymbol{S}}_k = \frac{1}{T}\sum_{t=0}^{T-1} \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k) \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)^T - \frac{1}{T^2}\left(\sum_{t=0}^{T-1} \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\right)\left(\sum_{t=0}^{T-1} \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\right)^T. \tag{22}$$

Then, an estimator of $\boldsymbol{S}$ can be constructed as

$$\widehat{\boldsymbol{S}} = K\sum_{k=1}^{K} w_k^2 \widehat{\boldsymbol{S}}_k. \tag{23}$$

The second term in (22) is necessary in the expression of $\widehat{\boldsymbol{S}}_k$ since generally $\nabla F_k(\boldsymbol{\theta}_K^*) \neq \mathbf{0}_{d\times 1}$ due to the heterogeneity across the clients.

4.3. *One-pass estimation of the Hessian matrix for smooth loss functions.* When the local objective functions $\{f_k(\cdot;\cdot)\}$ are smooth namely second-order differentiable with respect to $\boldsymbol{\theta}$, we propose a plug-in estimator as follows. Let $a(\cdot): \mathbb{N}_+ \to \mathbb{N}_+$ be a nondecreasing function, then the estimators of $\boldsymbol{H}_k = \nabla^2 F_k(\boldsymbol{\theta}_K^*)$ and $\boldsymbol{H}$ can be defined as

$$\widehat{\boldsymbol{H}}_k = \frac{1}{a(T)}\sum_{s=0}^{a(T)-1} \nabla^2 f_k(\hat{\hat{\boldsymbol{\theta}}}_{T-s-1}^k; \boldsymbol{\xi}_{T-s}^k) \quad \text{and} \quad \widehat{\boldsymbol{H}} = \sum_{k=1}^{K} w_k \widehat{\boldsymbol{H}}_k, \tag{24}$$

respectively. Direct estimation of the Hessian matrix as given in (24) is often considered to be computationally expensive since each $\nabla^2 f_k(\hat{\hat{\boldsymbol{\theta}}}_{T-s-1}^k; \boldsymbol{\xi}_{T-s}^k)$ consists of d^2 elements which is large compared with the d elements of the gradient $\nabla f_k(\hat{\hat{\boldsymbol{\theta}}}_{T-s-1}^k; \boldsymbol{\xi}_{T-s}^k)$. Besides, typically we need $a(T) \to \infty$ as $T \to \infty$ to ensure the consistency of $\widehat{\boldsymbol{H}}_k$ under the nondistributed scenario. It is noted that Fang, Xu and Yang (2018) proposed an offline plug-in estimator for the Hessian matrix, and Chen et al. (2020a) generalized the procedure to an online version. Later, Li et al. (2022) proposed a corresponding centralized FL version under finite K scenario. In all these works, the condition $a(T) \to \infty$ is required. However, given the fact that the number of clients K is usually large in the FL scenario, any single Hessian matrix $\boldsymbol{H}_k$ contributes little to the final aggregated matrix $\boldsymbol{H} = \sum_{k=1}^K w_k \boldsymbol{H}_k$, which suggests that we may not need to consistently estimate each $\boldsymbol{H}_k$. Instead, the law of large numbers takes effect as $K \to \infty$ and thus we can derive the consistency of $\sum_{k=1}^K w_k \widehat{\boldsymbol{H}}_k$ as a whole. Hence, $a(T)$ can be some finite number only to stabilize the numerical performance of $\widehat{\boldsymbol{H}}$. In this sense, $\widehat{\boldsymbol{H}}_k$ can be computed efficiently for each client k.

To provide a theoretical guarantee on the consistency of $\widehat{\boldsymbol{H}}$, we need the following assumption, which is a stronger version of Assumption 4.3.

ASSUMPTION 4.4. For all $k = 1, 2, \ldots, K$, we assume that the objective function $f_k(\boldsymbol{\theta}; \boldsymbol{\xi})$ is second-order differentiable with respect to $\boldsymbol{\theta} \in \boldsymbol{\Phi}$, and there exists positive constants ℓ_H and H, such that

$$\sqrt{\mathrm{E}(\|\nabla^2 f_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k) - \nabla^2 f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\|_2^2)} \leq \ell_H \|\boldsymbol{\theta} - \boldsymbol{\theta}_K^*\|_2$$

and $\mathrm{E}(\|\nabla^2 f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k) - \nabla^2 F_k(\boldsymbol{\theta}_K^*)\|_2^2) \leq H^2$, where $\boldsymbol{\theta} \in \boldsymbol{\Phi}$ and $\boldsymbol{\theta}_K^*$ is the true value defined in (2).

THEOREM 4. *Under assumptions required in Theorem 2 and Assumptions 4.1, 4.2 and 4.4, if* $K = o(T^{2\alpha-1})$ *and* $\mathrm{Ka}(T) \to \infty$, $\alpha < 1$, $\sup_{K\geq 1} \|\boldsymbol{\theta}_K^*\|_2 < \infty$ *and*

$\sup_{K\geq 1}\max_{1\leq k\leq K}\|\nabla F_k(\boldsymbol{\theta}_K^*)\|_2<\infty$, *then* $\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{H}^{-1}\boldsymbol{S}\boldsymbol{H}^{-1}\|_2=o_p(1)$ *and for any* $\beta\in(0,1)$,

$$\mathrm{P}\big(TK(\hat{\bar{\boldsymbol{\theta}}}_T-\boldsymbol{\theta}_K^*)^T\widehat{\boldsymbol{\Sigma}}^{-1}(\hat{\bar{\boldsymbol{\theta}}}_T-\boldsymbol{\theta}_K^*)\leq\chi^2_{d,\beta}\big)\to 1-\beta\quad \text{as } T\to\infty, \tag{25}$$

where $\chi^2_{d,\beta}$ *is the upper* β *quantile of the* χ^2_d *distribution*, $\widehat{\boldsymbol{\Sigma}}=\widehat{\boldsymbol{H}}^{-1}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{H}}^{-1}$.

This theorem is readily useful for the construction of the $1-\beta$ confidence region for $\boldsymbol{\theta}_K^*$

$$\big\{\boldsymbol{\theta}\,|\,TK(\hat{\bar{\boldsymbol{\theta}}}_T-\boldsymbol{\theta})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\hat{\bar{\boldsymbol{\theta}}}_T-\boldsymbol{\theta})\leq\chi^2_{d,\beta}\big\}. \tag{26}$$

5. One-pass estimation of the Hessian matrix for nonsmooth loss functions. In some statistical applications including the Huber regression and quantile regression, the local objective functions $f_k(\cdot;\cdot)$ are not second-order differentiable, which means that the plug-in estimator (24) for the aggregated Hessian matrix is not longer applicable. It is also of practical value to further reduce the computational complexity of the statistical inference procedure by developing the first-order method, even if the loss functions $f_k(\cdot;\cdot)$ are second-order differentiable. To this end, in the nondistributed SGD or the centralized FL scenarios, (Chen et al. (2020a), Zhu, Chen and Wu (2023)) proposed a batch-means estimator which directly estimated the asymptotic covariance matrix of the PR-averaged estimator, and Lee et al. (2022) and Li et al. (2022) proposed a random scaling estimator which facilitates the statistical inference justified by a functional central limit theorem. However, both estimators required frequent spatial averaging among the local estimators, which was unrealistic in the DFL setting.

Motivated by Ruppert (1988), we propose a regression-based estimator. This estimator only uses $\{\hat{\boldsymbol{\theta}}_t^k\}$ and $\{\nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^k;\boldsymbol{\xi}_t^k)\}$, which have already been calculated during the optimization, and is thus a first-order method. For simplicity, we first present the idea via the nondistributed ($K=1$) SGD. It is noted that despite the loss function $f_k(\cdot;\cdot)$ may not be second-order differentiable, the risk function $F_k(\cdot)$ is twice-differentiable. Hence, at each step t, by Taylor's expansion of $\nabla F(\hat{\boldsymbol{\theta}}_{t-1}^1)$ at $\boldsymbol{\theta}_1^*$,

$$\nabla f(\hat{\boldsymbol{\theta}}_{t-1}^1;\boldsymbol{\xi}_t^1)\approx-\nabla^2F(\boldsymbol{\theta}_1^*)\boldsymbol{\theta}_1^*+\nabla^2F(\boldsymbol{\theta}_1^*)\hat{\boldsymbol{\theta}}_{t-1}^1+\boldsymbol{\epsilon}_1(\hat{\boldsymbol{\theta}}_{t-1}^1;\boldsymbol{\xi}_t^1), \tag{27}$$

where $\boldsymbol{\epsilon}_1(\hat{\boldsymbol{\theta}}_{t-1}^1;\boldsymbol{\xi}_t^1)=\nabla f(\hat{\boldsymbol{\theta}}_{t-1}^1;\boldsymbol{\xi}_t^1)-\nabla F(\hat{\boldsymbol{\theta}}_{t-1}^1)$. For the SGD-based optimization, one naturally have a sequence of estimators $\{\hat{\boldsymbol{\theta}}_{t-1}^1\}_{t=1}^T$ and the corresponding stochastic gradients $\{\nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^1;\boldsymbol{\xi}_t^1)\}_{t=1}^T$. This motivates us to construct a multiple-response linear regression to estimate the Hessian matrix $\nabla^2F(\boldsymbol{\theta}_1^*)$, that is, to regress $\nabla f(\hat{\boldsymbol{\theta}}_{t-1}^1;\boldsymbol{\xi}_t^1)$ on $\hat{\boldsymbol{\theta}}_{t-1}^1$ to obtain $\widehat{\boldsymbol{H}}_1^{\mathrm{reg}}:=\boldsymbol{Z}_1\boldsymbol{V}_1^{-1}$, where

$$\begin{aligned}\boldsymbol{Z}_1&=C(T,\alpha)\sum_{t=2}^{T}\nabla f(\hat{\boldsymbol{\theta}}_{t-1}^1;\boldsymbol{\xi}_t^1)(\hat{\boldsymbol{\theta}}_{t-1}^1-\hat{\bar{\boldsymbol{\theta}}}_{T-1})^T,\\ \boldsymbol{V}_1&=C(T,\alpha)\sum_{t=2}^{T}(\hat{\boldsymbol{\theta}}_{t-1}^1-\hat{\bar{\boldsymbol{\theta}}}_{T-1})(\hat{\boldsymbol{\theta}}_{t-1}^1-\hat{\bar{\boldsymbol{\theta}}}_{T-1})^T\end{aligned} \tag{28}$$

and

$$C(T,\alpha)=\begin{cases}T^{\alpha-1} & \text{if } \dfrac{1}{2}<\alpha<1\\ (\log(T))^{-1} & \text{if } \alpha=1.\end{cases} \tag{29}$$

To avoid possible singularity of $\boldsymbol{V}_1$, we use the thresholding version (Chen et al. (2020a)). Specifically, for a small $\delta > 0$, let the $\boldsymbol{\Psi}\mathbf{D}\boldsymbol{\Psi}^T$ be the eigenvalue decomposition of $\boldsymbol{V}_1$, where $\mathbf{D} = (D_{ij})$ is a nonnegative diagonal matrix. The thresholding version $\tilde{\boldsymbol{V}}_1$ is

$$\tilde{\boldsymbol{V}}_1 = \boldsymbol{\Psi}\tilde{\mathbf{D}}\boldsymbol{\Psi}^T, \ \left(\tilde{\mathbf{D}}_{jj}\right) = \min\left\{\max\{(\mathbf{D}_{jj}), \delta\}, \frac{1}{\delta}\right\}, \tag{30}$$

and the thresholding version of $\widehat{\boldsymbol{H}}_1^{\mathrm{reg}}$ is defined as $\widetilde{\boldsymbol{H}}_1^{\mathrm{reg}} = \boldsymbol{Z}_1(\tilde{\boldsymbol{V}}_1)^{-1}$. Section S3 in the SM provides the theoretical property of the estimator $\widetilde{\boldsymbol{H}}_1^{\mathrm{reg}}$ in the nondistributed SGD setting.

The estimator $\widetilde{\boldsymbol{H}}_1^{\mathrm{reg}}$ can be extended to the decentralized FL setting with $K > 1$. First, given a positive constant $\zeta \in (0, 1/2)$, we define $r_i(T)$ for $i = 1, 2, 3$ such that $\sum_{i=1}^3 r_i(T) = T$. Specifically,

$$r_1(T) = \begin{cases} \dfrac{T}{C_1} & \text{for } \alpha < 1 \\ \dfrac{T^{1-\zeta}}{C_2} & \text{for } \alpha = 1 \end{cases}, \qquad r_2(T) = \frac{T}{C_3}, \quad \text{and} \quad r_3(T) = \frac{T}{C_4}, \tag{31}$$

where $C_j > 1$ for $1 \le j \le 4$ are constants. Then, the regression-based estimator $\widehat{\boldsymbol{H}}^{\mathrm{reg}}$ for the aggregated Hessian matrix $\boldsymbol{H}$ is $\widehat{\boldsymbol{H}}^{\mathrm{reg}} = \boldsymbol{Z}\boldsymbol{V}^{-1}$, where

$$\boldsymbol{Z} = B\big(T - r_3(T), r_1(T), \alpha\big)K \sum_{k=1}^K w_k \boldsymbol{Z}_k, \quad \boldsymbol{V} = B\big(T, r_1(T), \alpha\big) \sum_{k=1}^K \boldsymbol{V}_k.$$

$$\boldsymbol{Z}_k = \sum_{t=r_1(T)+1}^{T-r_3(T)} \left(\nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^k; \boldsymbol{\xi}_t^k) - \left(\frac{1}{r_3(T)} \sum_{s=T-r_3(T)+1}^{T} \nabla f_k(\hat{\boldsymbol{\theta}}_{s-1}^k; \boldsymbol{\xi}_s^k)\right)\right)(\hat{\boldsymbol{\theta}}_{t-1}^k - \hat{\bar{\boldsymbol{\theta}}}_{T-1})^T,$$

$$\boldsymbol{V}_k = \sum_{t=r_1(T)+1}^{T} (\hat{\boldsymbol{\theta}}_{t-1}^k - \hat{\bar{\boldsymbol{\theta}}}_{T-1})(\hat{\boldsymbol{\theta}}_{t-1}^k - \hat{\bar{\boldsymbol{\theta}}}_{T-1})^T$$

and $B(T_1, T_2, \alpha) = (C(T_1, \alpha)^{-1} - C(T_2, \alpha)^{-1})^{-1}$, where $C(T, \alpha)$ is defined in (29).

REMARK. In the decentralized FL, there tends to be a large consensus error among the local estimates during the initial stage as reflected in Theorem 1. Thus, in the above formulation, we have removed the first $r_1(T)$ local estimates by treating them as the warming-up estimates in the $\boldsymbol{Z}_k$ and $\boldsymbol{V}_k$ to mitigate the effect of the error. Doing so preserves the orders of both $\boldsymbol{Z}$ and $\boldsymbol{V}$ estimates, which are shown by their respective normalizing constants. Besides, since $\nabla F_k(\boldsymbol{\theta}_K^*) \neq \mathbf{0}_d$ when $K > 1$, compared with $\boldsymbol{Z}_1$ in (28), the centering in $\boldsymbol{Z}_k$ is necessary. Furthermore, the use of the estimates over different time segments in $\boldsymbol{Z}_k$ is to remove the conditional dependence between the terms $r_3(T)^{-1} \sum_{s=T-r_3(T)+1}^T \nabla f_k(\hat{\boldsymbol{\theta}}_{s-1}^k; \boldsymbol{\xi}_s^k)$ and $\{\hat{\boldsymbol{\theta}}_t^k\}_{t<T-r_3(T)}$, which helps to improve the estimation accuracy of $\widehat{\boldsymbol{H}}^{\mathrm{reg}}$ when the number of clients K is large. See Section S2.5 in the SM for more discussions.

Denote $\widetilde{\boldsymbol{H}}^{\mathrm{reg}} = \boldsymbol{Z}(\widetilde{\boldsymbol{V}})^{-1}$, where $\widetilde{\boldsymbol{V}}$ is the thresholding version of $\boldsymbol{V}$ according to (30). The following theorem establishes the consistency of $\widetilde{\boldsymbol{H}}^{\mathrm{reg}}$.

THEOREM 5. *Under Assumptions 2.1–2.2, 3.1 with $v = 2$, 3.3–3.4, 4.1 and 4.3, if the parameter space $\boldsymbol{\Phi}$ is bounded with $R_d < \infty$, $K = o(T^\alpha)$ if $\frac{1}{2} < \alpha < 1$ and $K = o(\log(T)T^{1-\zeta})$ if $\alpha = 1$, where the constant $\zeta \in (0, 1/2)$ is defined in (31), then as $T \to \infty$*

$$\mathrm{E}(\|\widetilde{\boldsymbol{H}}^{\mathrm{reg}} - \boldsymbol{H}\|_F) \to 0.$$

The estimator $\widetilde{\boldsymbol{H}}^{\mathrm{reg}}$ for $\boldsymbol{H}$ can be used to replace the estimator (24) when the local objective functions are not second-order differentiable with respect to $\boldsymbol{\theta}$, which facilitates a $1-\beta$ confidence region for $\boldsymbol{\theta}_K^*$

$$\{\boldsymbol{\theta} | TK(\bar{\hat{\boldsymbol{\theta}}}_T - \boldsymbol{\theta})^T \widetilde{\boldsymbol{\Sigma}}^{-1}(\bar{\hat{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}) \le \chi^2_{d,\beta}\}, \tag{32}$$

where $\widetilde{\boldsymbol{\Sigma}} = (\widetilde{\boldsymbol{H}}^{\mathrm{reg}})^{-1}\widehat{\boldsymbol{S}}(\widetilde{\boldsymbol{H}}^{\mathrm{reg}})^{-1}$, and $\widehat{\boldsymbol{S}}$ is defined in (23).

6. Efficient one-step estimator. We have shown in Theorem 3 that the PR-averaged estimator $\bar{\hat{\boldsymbol{\theta}}}_T$ is statistically efficient for $\alpha \in (1/2, 1)$. However, the efficiency comes at a price in the decentralized FL setting reflected in the restriction that K is either finite or diverging at the rate $K = o(T^{2\alpha-1})$. This is more restrictive than $K = o(T)$ required in the "split-and-conquer" paradigm for the distributed inference under both the homogeneous (Zhang, Duchi and Wainwright (2013)) and the heterogeneous (Gu and Chen (2023)) M-estimation.

The need to restrict on K can be found by examining (19) in Theorem 2. The $v_1\eta_T K^{-1}$ term is the leading order term in the expansion of $\bar{\hat{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}_K^*$, and the second $v_2\eta_T^2$ term can be viewed as the squared bias (the consensus error) of the local estimates $\{\hat{\boldsymbol{\theta}}_t^k\}_{k=1}^K$. It can be seen that a large step size with $\alpha \in (1/2, 1)$, as compared with $\alpha = 1$, introduces a large ratio $\eta_T^2/(\eta_T K^{-1}) = O(K/T^\alpha)$. For a fixed K including the nondistributed case ($K = 1$), the bias is not a concern. For the distributed scenarios with large K, however, the larger step size significantly restricts the allowable number K of clients. In particular, if we choose a large step size with $\alpha = 1/2 + \epsilon$ as suggested in Ruppert (1988) for a small positive constant ϵ, then only $K = o(T^{2\epsilon})$ clients are allowed to join the FL to preserve the validity of Theorem 3. This is too restrictive for the FL with large number of clients.

The restriction on K encourages us to choose $\alpha = 1$ in the estimator $\bar{\hat{\boldsymbol{\theta}}}_T$ when K diverges with T. Although $\bar{\hat{\boldsymbol{\theta}}}_T$ is inefficient (Sacks (1958), Polyak and Juditsky (1992)) with $\alpha = 1$ as in the nondistributed scenario, it is $\sqrt{TK}$-consistent as long as $K = o(T)$. This motivates using the strategy of the one-step estimation (Bickel (1975)) to improve the statistical efficiency. Given the preliminary estimator $\bar{\hat{\boldsymbol{\theta}}}_T$ with $\alpha = 1$, the one-step estimator is

$$\bar{\hat{\boldsymbol{\theta}}}_T^{\mathrm{os}} = \bar{\hat{\boldsymbol{\theta}}}_T - (\widehat{\boldsymbol{H}})^{-1}\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} w_k \nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^k; \boldsymbol{\xi}_t^k), \tag{33}$$

where $\widehat{\boldsymbol{H}}$ is an estimator of $\boldsymbol{H}$ given in (24).Note that the computation of $\widehat{\boldsymbol{H}}$ of $\boldsymbol{H}$ and its inverse $\widehat{\boldsymbol{H}}^{-1}$ of $d \times d$ dimension is necessary for statistical inference, and those gradients $\{\nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k) | t \ge 1, 1 \le k \le K\}$ are already calculated during the optimization process as given in the DFL algorithm. So taking average of those gradients is only extra computation to obtain the proposed one-step estimator $\bar{\hat{\boldsymbol{\theta}}}_T^{\mathrm{os}}$. We summarize the procedure of the one-step estimator in Algorithm 2 of the SM. We will see in Theorem 6, to establish the asymptotic normality of the one-step estimator, we need both K and T increase to infinity to ensure the validity of the following first-order expansion of the estimator $\bar{\hat{\boldsymbol{\theta}}}_T$ when $\alpha = 1$ such that

$$\left\| (\bar{\hat{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}_K^*) - \boldsymbol{H}^{-1}\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} w_k(\nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}; \boldsymbol{\xi}_t^k) - \nabla f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}_t^k)) \right\|_2 = o_p\left(\frac{1}{\sqrt{TK}}\right).$$

It is also noted that the number of clients K is allowed to be finite in Theorem 3. The condition on K for the one-step estimator is natural, since we are considering a large-scale decentralized FL problem where many clients conduct the optimization collaboratively. For moderate K, it suffices to use the PR-averaged estimator $\bar{\hat{\boldsymbol{\theta}}}_T$ for statistical inference purposes.

THEOREM 6. *Under assumptions of Theorem 2 and Assumptions 4.1, 4.2 and 4.4, if* $\alpha = 1$ *and* $\sup_{K\geq 1} \|\boldsymbol{\theta}_K^*\|_2 < \infty$, *then the one-step estimator* $\hat{\bar{\boldsymbol{\theta}}}_T^{\mathrm{os}}$ *defined in* (33) *admits the following expansion*:

$$\sqrt{TK}\boldsymbol{S}^{-1/2}\boldsymbol{H}\big(\hat{\bar{\boldsymbol{\theta}}}_T^{\mathrm{os}} - \boldsymbol{\theta}_K^*\big) = \boldsymbol{S}^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{k=1}^{K}\sqrt{K}w_k\nabla F_k\big(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}_t^k\big) + O_p\bigg(\sqrt{\frac{K}{T}} + \frac{1}{\sqrt{K}}\bigg).$$

Consequently, if $K = o(T)$, $\sqrt{TK}\boldsymbol{S}^{-1/2}\boldsymbol{H}(\hat{\bar{\boldsymbol{\theta}}}_T^{\mathrm{os}} - \boldsymbol{\theta}_K^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ *as* T *and* $K \to \infty$.

Theorem 6 establishes the asymptotic efficiency of the proposed one-step estimator with a relaxed constraint on the number K of data nodes. From the data storage perspective, this procedure only requires saving extra $2d$ numbers in each data node, since both $\hat{\bar{\boldsymbol{\theta}}}_T^k$ and $\sum_{t=1}^{T}\nabla f_k(\hat{\boldsymbol{\theta}}_{t-1}^k; \boldsymbol{\xi}_t^k)$ are d-dimensional vectors that can be obtained recursively. Besides, different from the PR-averaged estimator, the one-step estimator $\hat{\bar{\boldsymbol{\theta}}}_T^{\mathrm{os}}$ can be decomposed into two parts: $\hat{\bar{\boldsymbol{\theta}}}_T$ for fast statistical convergence rate and $(\widehat{\boldsymbol{H}})^{-1}\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}w_k\nabla F_k(\hat{\boldsymbol{\theta}}_{t-1}^k; \boldsymbol{\xi}_t^k)$ as a correction term to improve statistical efficiency. During the optimization stage, we only require the decentralized gradients sharing. So the proposed estimator is both communication and statistically efficient. The construction of the confidence regions based on the one-step estimator $\hat{\bar{\boldsymbol{\theta}}}_T^{\mathrm{os}}$ is similar to that of the PR-averaged estimator $\hat{\bar{\boldsymbol{\theta}}}_T$ and is implied by the following corollary.

COROLLARY 1. *Under the assumptions of Theorem 6,*

$$\mathrm{P}\big(TK(\hat{\bar{\boldsymbol{\theta}}}_T^{\mathrm{os}} - \boldsymbol{\theta}_K^*)^T\widehat{\boldsymbol{\Sigma}}^{-1}(\hat{\bar{\boldsymbol{\theta}}}_T^{\mathrm{os}} - \boldsymbol{\theta}_K^*) \leq \chi_{d,\beta}^2\big) \to 1-\beta \quad \text{as } T \text{ and } K \to \infty.$$

for any constant $\beta \in (0,1)$, *where* $\chi_{d,\beta}^2$ *is the upper* β*-quantile of the* χ_d^2 *distribution,* $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{H}}^{-1}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{H}}^{-1}$, *and* $\widehat{\boldsymbol{H}}$ *and* $\widehat{\boldsymbol{S}}$ *are defined in* (24) *and* (23), *respectively.*

Section S4 of the SM discusses the theoretical properties of the one-step estimator for sparsely-connected networks, such that Assumption 2.1 fails to hold: If $\max\{|\lambda_k(\boldsymbol{C})||k = 2,3,\ldots,K\} \leq 1-\rho' K^{-q} < \lambda_1(\boldsymbol{C}) = 1$ for some $0 < \rho' < 1$ and $q \geq 0$, then the one-step estimator achieves the same asymptotic distribution as in Theorem 6 for $K = o(T^{\frac{1}{6q+1}})$.

7. Simulation results. We report results from three sets of simulation experiments designed to verify the theoretical findings in the previous sections. In all simulation experiments, the decentralized connection network was constructed according to the Metropolis–Hastings Rule in (5). Given a network size K, the nodes were denoted by their labels $1,2,\ldots,K$, and a number K_{neigh} was used to denote the number of neighbors each node has, which controlled the connectivity of the network. Clients k and k' are connected if and only if $|k-k'| \leq \frac{K_{\mathrm{neigh}}}{2}$ or $|k-k'| \geq K - \frac{K_{\mathrm{neigh}}}{2}$. Thus, for a given K, a larger (small) K_{neigh} means a tightly (loosely) connected network.

The local data of the clients or the nodes of the network were generated according to a linear regression model. For each client k, $\{(\boldsymbol{X}_{k,t}; Y_{k,t})\}_{t=1}^T$ were independently sampled from the following model:

$$\boldsymbol{X}_{k,t} \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}_{(d-1)\times 1}, \mathbf{I}_{(d-1)\times(d-1)}),\ \varepsilon_{k,t} \overset{i.i.d}{\sim} \Gamma(1,1) - 1 \quad \text{and} \quad Y_{k,t} = (1, \boldsymbol{X}_{k,t}^T)\boldsymbol{\phi}_k^* + \varepsilon_{k,t},$$

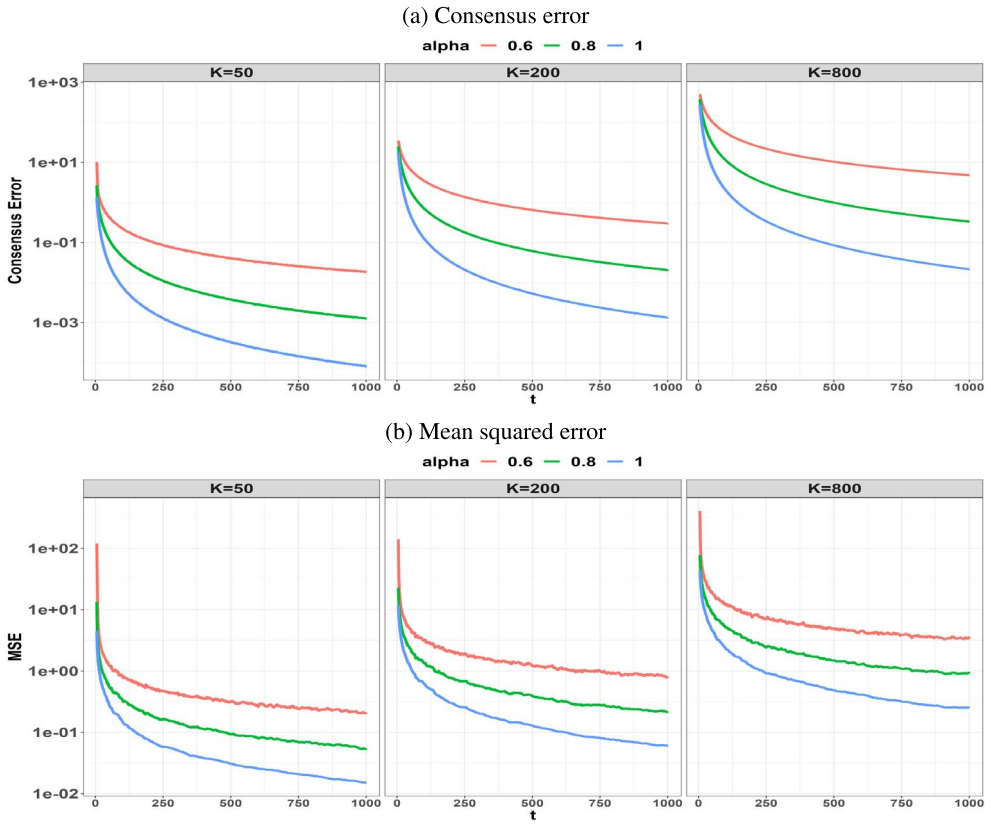


FIG. 2. *Average consensus error* (*a*) *and the mean squared error of the averaged estimator* $\hat{\bar{\boldsymbol{\theta}}}_T$ (*b*) *under different numbers of block size* K ($K = 50, 200$ *and* 800) *with respect to the number of SGD steps* t ($t \le T$, *where* T *was the local sample size*) *when the rate* α *of the step size was* 0.6, 0.8 *and* 1, *respectively, and the local update parameter* τ *was* 1. *The gap parameter* δ_{gap} *was* 0.2 *corresponding to a stronger case of heterogeneity.*

where $\boldsymbol{\phi}_k^* = (\phi_{k,1}^*, \phi_{k,2}^*, \ldots, \phi_{k,d}^*)^T$, $\Gamma(1,1)$ denotes the $\Gamma(1,1)$ random variables and the parameter's dimension $d = 6$. The true parameter $\boldsymbol{\theta}_K^*$ was $\boldsymbol{\theta}_K^* = \sum_{k=1}^K w_k \boldsymbol{\phi}_k^*$ where $w_k \equiv 1/K$ and $\phi_{k,j}^* = \delta_{\text{gap}}((k-1) - (K-1)/2)$ for a $\delta_{\text{gap}} > 0$. This made the true parameter $\boldsymbol{\theta}_K^* = \mathbf{0}_d$ while δ_{gap} quantifies the heterogeneity across the data blocks. The results of each simulation setting were based on $B = 500$ replications.

In the first experiment, we evaluated the trajectory of the local estimates $\{\hat{\boldsymbol{\theta}}_t^k\}$ by assessing the averaged estimate $\hat{\bar{\boldsymbol{\theta}}}_t (1 \le t \le T)$ by averaging the local estimates $\{\hat{\boldsymbol{\theta}}_t^k\}_{1 \le k \le K}$ among the clients, and their variability. We set the local sample size $T = 1000$ and the number of neighbors $K_{\text{neigh}} = (3K)/5$, $K \in \{50, 100, 200, 400, 800\}$, the update frequency $\tau \in \{1, 3, 5\}$, the diminishing rate $\alpha \in \{0.6, 0.8, 1.0\}$, and the delta gap $\delta_{\text{gap}} \in \{0.06, 0.2\}$.

Figure 2 reports the estimated consensus error $K^{-1} \sum_{k=1}^K \mathrm{E}(\|\hat{\boldsymbol{\theta}}_t^k - \hat{\bar{\boldsymbol{\theta}}}_t\|_2^2)$ of the local estimates $\{\hat{\boldsymbol{\theta}}_t^k\}_{1 \le k \le K}$ and the mean squared error $\mathrm{E}(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2)$ of the averaged estimate $\hat{\bar{\boldsymbol{\theta}}}_t$ for $K = 50, 200$ and 800, $\alpha \in \{0.6, 0.8, 1.0\}$, $\delta_{\text{gap}} = 0.2$ and $\tau = 1$. It is observed that as t increased, both the MSE and the consensus error decreased for all choices of α. However, the decrease was faster for $\alpha = 1$ than those of $\alpha = 0.6$ and 0.8, confirming Theorems 1 and 2, which suggest that a larger step size (smaller α) incurs a larger consensus error.

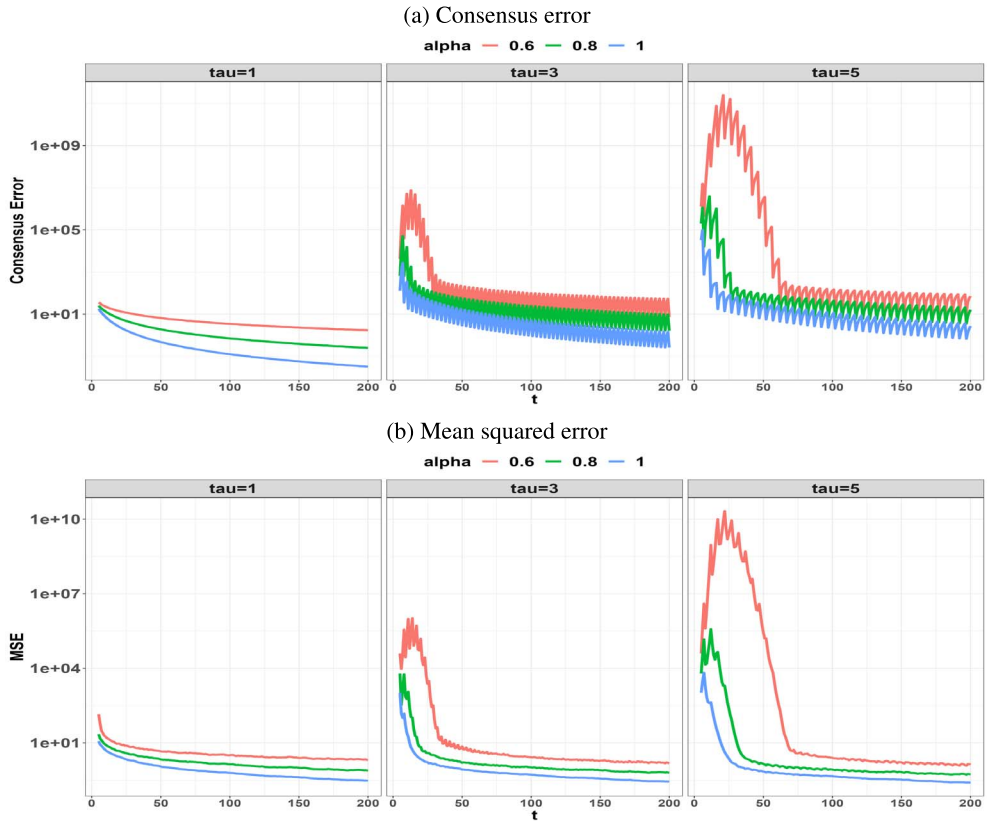


FIG. 3. *Average consensus error* (*a*) *and the mean squared error of the averaged estimator* $\hat{\bar{\boldsymbol{\theta}}}_T$ (*b*) *under different numbers of local* τ ($\tau = 1, 3$ *and* 5) *with respect to the number of SGD steps* t ($t \le T$, *where* T *was the local sample size*) *when the rate* α *of the step size was* 0.6, 0.8 *and* 1.0, *respectively, and the number of clients* K *was* 200. *The gap parameter* δ_{gap} *was* 0.2 *corresponding to a stronger case of heterogeneity.*

Both Theorems 1 and 2 also suggest that

$$\frac{\sum_{k=1}^{K} \mathrm{E}(\|\hat{\boldsymbol{\theta}}_t^k - \hat{\bar{\boldsymbol{\theta}}}_t\|_2^2)}{K\mathrm{E}(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2)} = O\Big(\frac{K}{t^\alpha + K}\Big),$$

which means that the consensus error should be much smaller than the corresponding MSE especially when t was large and K was small. Indeed, comparing Panels (a) and (b) of Figure 2, it is observed that for each given t, the ratio increased as K increased for all α, which also verified numerically that the consensus error was no longer an ignorable term compared with the leading term of the upper bound (19) of the averaged estimate $\hat{\bar{\boldsymbol{\theta}}}_t$ for a network with large K, especially when α is small.

Figure 3 reports the estimated consensus error $K^{-1}\sum_{k=1}^{K} \mathrm{E}(\|\hat{\boldsymbol{\theta}}_t^k - \hat{\bar{\boldsymbol{\theta}}}_t\|_2^2)$ and the mean squared error $\mathrm{E}(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2)$ for $\tau = 1, 3$ and 5, $t \le 200$, $\alpha \in \{0.6, 0.8, 1.0\}$, $\delta_{\text{gap}} = 0.2$ and $K = 200$, respectively. Here we only report the first 200 steps to better capture effects of τ. It shows that at the initial stage for $t \le 100$, a larger τ led to both a larger consensus error and a larger MSE, and this effect was more pronounced for a smaller α. As t increased, while the consensus error was still larger for a larger τ, there was no significant difference among different τ for the MSEs, justifying the role of the consensus error as a second-order effect as revealed in Theorems 1 and 2. Besides, the local update nature in the DFL algorithm was reflected by the volatility in both the consensus error and the MSE sequences when $\tau > 1$ along with the decreasing trend as t increased.

TABLE 1

Empirical absolute coverage errors and the volumes (in parentheses, multiplied by 10^7) of the $1-\beta$ confidence regions for the parameter $\boldsymbol{\theta}_K^*$ in the linear regression model based on the asymptotic normality of the Polyak–Ruppert averaged estimator in decentralized FL with respect to the diminishing rate α of the step size, the number of clients K and the local sample sizes T (the number of SGD steps). The total sample size $N = KT$. The gap parameter δ_{gap}, the local update parameter τ and $a(T)$ were 0.05, 1 and 100, respectively

				$\alpha = 0.6$		$\alpha = 0.8$	
K	T	N	$1-\beta$	0.95	0.9	0.95	0.9
100	500	5×10^4		0.008 (1.540)	0.014 (0.931)	0.006 (1.954)	0.012 (1.181)
	1000	1×10^5		0.010 (0.216)	0.018 (1.307)	0.010 (0.262)	0.018 (0.158)
	2000	2×10^5		0.006 (0.030)	0.000 (0.018)	0.010 (0.034)	0.010 (0.021)
200	500	1×10^5		0.018 (9.810)	0.022 (5.926)	0.020 (12.89)	0.022 (7.787)
	1000	2×10^5		0.008 (1.404)	0.006 (0.848)	0.016 (1.747)	0.020 (1.056)
	2000	4×10^5		0.004 (0.194)	0.002 (0.118)	0.000 (0.230)	0.006 (0.118)
400	500	2×10^5		0.012 (73.49)	0.010 (44.40)	0.010 (97.92)	0.010 (59.16)
	1000	4×10^5		0.006 (10.58)	0.006 (6.393)	0.014 (13.31)	0.016 (6.751)
	2000	8×10^5		0.004 (1.473)	0.004 (0.890)	0.008 (1.756)	0.008 (1.061)

Both figures suggest that the consensus error and the MSE can be quite sensitive to α. In particular, with $\tau > 1$ applied to a large FL network, a larger α took less steps to stabilize and converge as shown in Figures 2 and 3. Hence, it is suggested to use a larger α, especially for larger FL networks. Figures S1–S4 report more results of this experiment.

In the second experiment, we assessed the asymptotic normality of the PR-averaged estimator by calculating the empirical absolute coverage errors and volumes of the 95% and 90% "confidence regions" (CRs), constructed according to (26) based on Theorem 4 with $\delta_{\text{gap}} = 0.05$, $K \in \{100, 200, 400\}$, $K_{\text{neigh}} = (3K)/5$, $T \in \{500, 1000, 2000\}$, $\tau \in \{1, 2\}$ and $a(T) \in \{25, 100\}$. The empirical coverage error is calculated as the difference between the nominal confidence level and the simulated coverage probability based on $B = 500$ replications of each simulation setting. It is noted that as the "confidence regions" were d-dimensional ellipses, their volumes can be calculated numerically. Table 1 reports the empirical absolute coverage errors and the average volume of the CRs when the number of local update $\tau = 1$ and $a(T) = 100$, which conveyed quite good coverage. Indeed, the coverage accuracy was quite uniformly maintained with respect to the step-size α, the number of clients K and the local sample size T. The effects of K and T were shown in the volume of the CRs. For each given K, the volume of the CRs decreased as the minimum local sample size T increased, and on the other hand, for a given T, the volume increased as K increased. This reflected the underlying variance of the PR-averaged estimator. Besides, the volumes of the constructed CRs when $\alpha = 0.8$ were larger than those when $\alpha = 0.6$, which means that for each dimension, the width of the CRs when $\alpha = 0.8$ was in average $3\% - 5\%$ wider than that of the CRs when $\alpha = 0.6$, reflecting the slightly under-estimation of the asymptotic covariance matrix when $\alpha = 0.6$ as shown in Table S4 in the SM. In comparison, when τ increased to 2 as shown in Table 2, the coverage errors of the CRs when $\alpha = 0.6$ were much larger than those when $\alpha = 0.8$ due to a larger consensus error according to Theorem 1. For both choices of the step sizes, the coverage errors decreased as T and K increased due to the larger total sample size. Besides, in most cases except $K = 100$ and $\alpha = 0.6$, the volumes of the CRs when $\tau = 2$ were slightly smaller than those when $\tau = 1$. Similar results were obtained when $a(T) = 25$ (see Tables S2–S3 in the SM), which shows the robustness of the procedure against different choices of $a(T)$. In practice, $a(T)$ can be chosen so that

TABLE 2

Empirical absolute coverage errors and the volumes (in parentheses, multiplied by 10^7) of the $1-\beta$ confidence regions for the parameter $\boldsymbol{\theta}_K^*$ in the linear regression model based on the asymptotic normality of the Polyak–Ruppert averaged estimator in decentralized FL with respect to the diminishing rate α of the step size, the number of clients K and the local sample sizes T (the number of SGD steps). The total sample size $N = KT$. The gap parameter δ_{gap}, the local update parameter τ and $a(T)$ were 0.05, 2 and 100, respectively

				$\alpha = 0.6$		$\alpha = 0.8$	
K	T	N	$1-\beta$	0.95	0.9	0.95	0.9
100	500	5×10^4		0.484 (3133)	0.543 (1893)	0.116 (1.624)	0.158 (0.981)
	1000	1×10^5		0.326 (48.95)	0.404 (29.57)	0.072 (0.227)	0.096 (0.137)
	2000	2×10^5		0.218 (0.891)	0.284 (0.539)	0.036 (0.031)	0.068 (0.019)
200	500	1×10^5		0.228 (12.73)	0.290 (7.670)	0.058 (10.34)	0.082 (6.248)
	1000	2×10^5		0.114 (1.310)	0.150 (0.792)	0.028 (1.475)	0.038 (0.891)
	2000	4×10^5		0.082 (0.165)	0.096 (0.099)	0.018 (0.204)	0.026 (0.123)
400	500	2×10^5		0.106 (69.59)	0.140 (39.79)	0.026 (77.60)	0.044 (46.88)
	1000	4×10^5		0.048 (8.377)	0.064 (5.061)	0.016 (11.16)	0.028 (6.742)
	2000	8×10^5		0.040 (1.150)	0.042 (0.695)	0.010 (1.554)	0.008 (0.939)

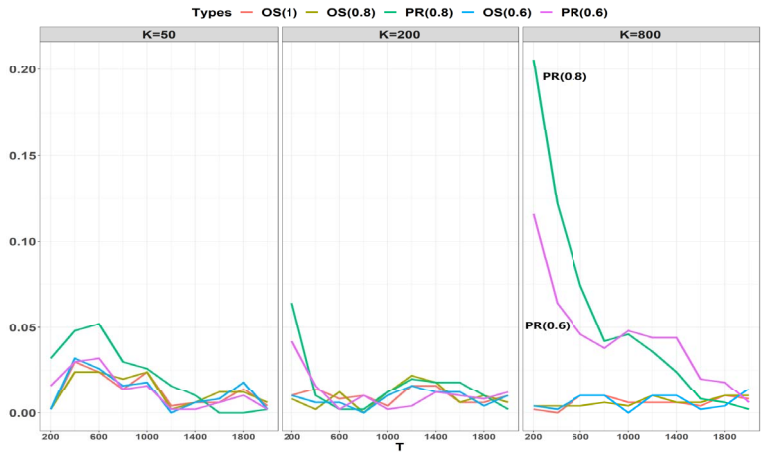
$a(T) \geq \max\{T^{2\alpha-1}/K^2, C_0\}$, based on the proof of Theorems 3 and 4 in the SM, where C_0 is a positive integer, for instance 25.

In the third experiment, we evaluated and compared the coverage accuracy, the bias and variance of the PR-averaged (PR) estimator $\bar{\bar{\boldsymbol{\theta}}}_T$ and the one-step (OS) estimator $\bar{\bar{\boldsymbol{\theta}}}_T^{\text{os}}$ for $\boldsymbol{\theta}_K^*$ with the number of clients or nodes $K \in \{500, 200, 800\}$, the local sample size $T \in \{200h | h = 1, 2, \ldots, 10\}$, the connection network $K_{\text{neigh}} = 10$, and the heterogeneity parameter $\delta_{\text{gap}} \in \{0.2, 0.01\}$ and $a(T) = 100$. Figures 4 and 5 displays the absolute coverage errors of the 95% confidence regions for $\boldsymbol{\theta}_K^*$ when $\delta_{\text{gap}} = 0.2$ and 0.01, respectively. In both figures, the advantages of the proposed one-step estimator over the PR-averaged estimator in term of having more accurate CRs were very visible in Panel (a) when K was larger ($K = 800$), which was readily seen for smaller T for $K = 200$. This suggested that the proposed one-step estimator is more suitable for the decentralized FL as it often has large K and small T. The figures also show that the one-step estimator had smaller bias and variance than the PR-averaged estimator for the large K but small T scenario, although the PR-averaged estimators gradually outperformed the one-step estimators in the bias and variance as T was increased. Moreover, the effect of heterogeneity was shown by a comparison between Panel (b) of Figures 4 and 5. Specifically, when the heterogeneity was large ($\delta_{\text{gap}} = 0.2$), for each given local sample size T, the bias of the estimators did not decrease as K increased, although a larger K means a larger total sample size $N = KT$. In contrast, when the heterogeneity was weak ($\delta_{\text{gap}} = 0.01$), the bias of the estimator did decrease as K increased.

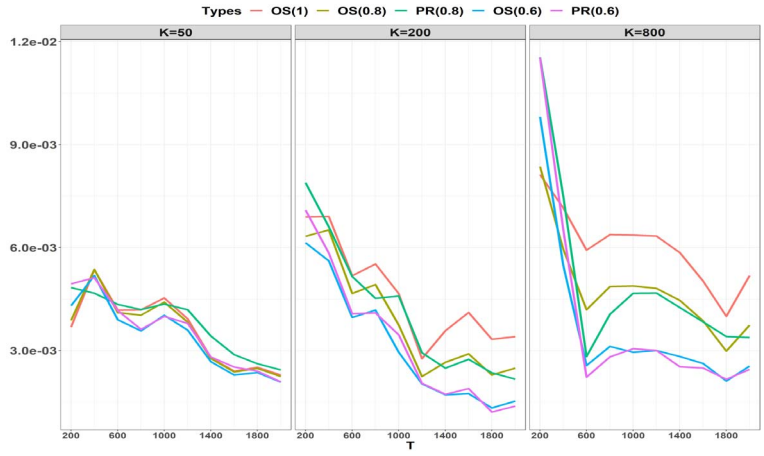
8. Discussion. This study investigates the decentralized FL in the context of client heterogeneity, where the clients can only share gradient information with their neighbors defined via a connection network. The Polyak–Ruppert (PR) averaged estimator are analysed in the decentralized FL setting that allows diverging network size, and the corresponding confidence regions are constructed. The one-step estimator is shown to permits larger network size than the Polyak–Ruppert (PR) averaged estimator, without sacrificing statistical efficiency.

It is noted that the convergence of the local estimators in the decentralized FL can be further improved by implementing two techniques. One is to involve acceleration (Qian (1999), Johnson and Zhang (2013)), which was initially designed for the nondistributed SGD estimator. The second technique is the bias correction technique, say *Exact Diffusion* (Yuan et al.

(a) Empirical absolute coverage error



(b) ℓ_2 norm of the bias



(c) Frobenius norm of the sample covariance matrix

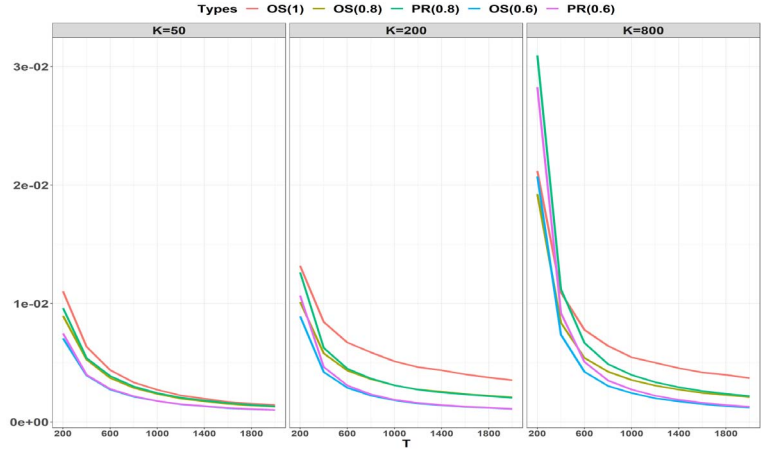
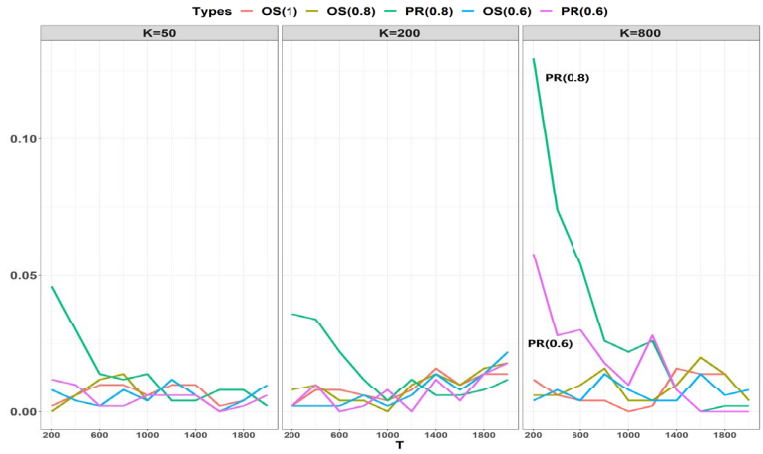
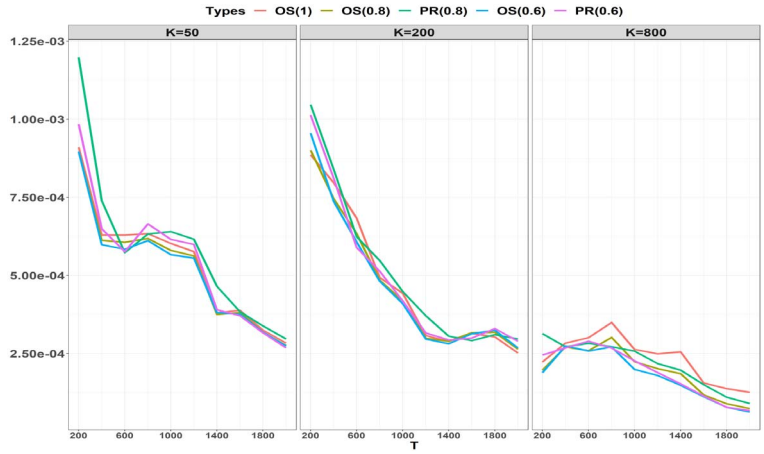


FIG. 4. *Empirical absolute coverage errors of the* 95% *confidence regions* (*a*) *based on the asymptotic normality of the one-step estimator* (*OS*, $\alpha = 1, 0.8, 0.6$) *and the Polyak–Ruppert averaged estimators* (*PR*, $\alpha = 0.8, 0.6$), *the* ℓ_2 *norm of the bias* (*b*) *and the Frobenius norm of the sample covariance matrices* (*c*) *with respect to* K *and* T. *The gap parameter* δ_{gap} *was* 0.2 *corresponding to a stronger case of heterogeneity.*

(a) Empirical absolute coverage error



(b) ℓ_2 norm of the bias



(c) Frobenius norm of the sample covariance matrix

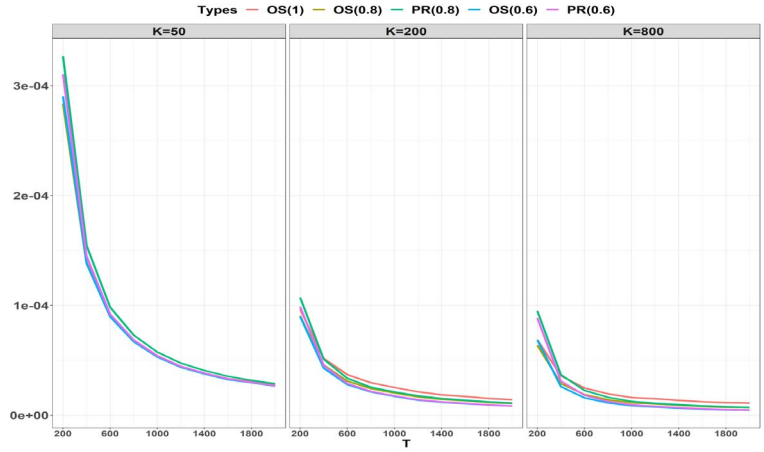


FIG. 5. *Empirical absolute coverage errors of the* 95% *confidence regions* (*a*) *based on the asymptotic normality of the one-step estimator* (*OS*, $\alpha = 1, 0.8, 0.6$) *and the Polyak–Ruppert averaged estimators* (*PR*, $\alpha = 0.8, 0.6$), *the* ℓ_2 *norm of the bias* (*b*) *and the Frobenius norm of the sample covariance matrices* (*c*) *with respect to* K *and* T. *The gap parameter* δ_{gap} *was* 0.01 *corresponding to a weaker case of heterogeneity.*

(2020)) or *Gradient Tracking* methods (Nedić, Olshevsky and Shi (2017)), which is designed specifically for decentralized optimization. Both the acceleration and bias correction techniques may be used to further relax the constraints on the number of blocks K relative to the local sample size T for both the PR-averaged estimator and the proposed one-step estimator, in both dense and sparse networks. These are interesting topics for future research.

Funding. This research is supported by National Natural Science Foundation of China grants 12292980, 12292983 and 92358303.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical inference for decentralized federated learning." (DOI: 10.1214/24-AOS2452SUPP; .pdf). In the SM, we present technical details, proofs of main theorems and additional numerical results.

REFERENCES

ALGHUNAIM, S. A. and YUAN, K. (2022). A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Trans. Signal Process.* **70** 3264–3279. MR4449095 https://doi.org/10.1109/tsp.2022.3184770

APPLE (2019). Private federated learning. NeurIPS 2019 Expo Talk.

BACH, F. (2010). Self-concordant analysis for logistic regression. *Electron. J. Stat.* **4** 384–414. MR2645490 https://doi.org/10.1214/09-EJS521

BACH, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.* **15** 595–627. MR3190851

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

BOTTOU, L., CURTIS, F. E. and NOCEDAL, J. (2018). Optimization methods for large-scale machine learning. *SIAM Rev.* **60** 223–311. MR3797719 https://doi.org/10.1137/16M1080173

BOYD, S., GHOSH, A., PRABHAKAR, B. and SHAH, D. (2006). Randomized gossip algorithms. *IEEE Trans. Inf. Theory* **52** 2508–2530. MR2238556 https://doi.org/10.1109/TIT.2006.874516

CHEN, M. and CUI, S. (2024). Federated learning for autonomous vehicles control. In *Communication Efficient Federated Learning for Wireless Networks* 129–150. Springer, Switzerland. https://doi.org/10.1007/978-3-031-51266-7_6

CHEN, X., LAI, Z., LI, H. et al. (2024). Online statistical inference for stochastic optimization via Kiefer–Wolfowitz methods. *J. Amer. Statist. Assoc.* 1–24. https://doi.org/10.1080/01621459.2023.2296703

CHEN, X., LEE, J. D., TONG, X. T. and ZHANG, Y. (2020a). Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.* **48** 251–273. MR4065161 https://doi.org/10.1214/18-AOS1801

CHEN, Y., QIN, X., WANG, J. et al. (2020b). FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intell. Syst.* **35** 83–93. https://doi.org/10.1109/MIS.2020.2988604

CHOI, W. and KIM, J. (2022). On the convergence of decentralized gradient descent with diminishing stepsize. Revisited.

CHUNG, K. L. (1954). On a stochastic approximation method. *Ann. Math. Stat.* **25** 463–483. MR0064365 https://doi.org/10.1214/aoms/1177728716

FANG, Y., XU, J. and YANG, L. (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *J. Mach. Learn. Res.* **19** 1–21. MR3899780

GU, J. and CHEN, S. X. (2023). Distributed statistical inference under heterogeneity. *J. Mach. Learn. Res.* **24** 1–56. MR4720843

GU, J. and CHEN, S. X. (2024). Supplement to "Statistical Inference for Decentralized Federated Learning." https://doi.org/10.1214/24-AOS2452SUPP

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application. Probability and Mathematical Statistics*. Academic Press, San Diego. MR0624435

HARD, A., RAO, K., MATHEWS, R. et al. (2018). Federated learning for mobile keyboard prediction.

JOHNSON, R. and ZHANG, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Adv. Neural Inf. Process. Syst.* **1** 315–323.

KAIROUZ, P., MCMAHAN, H. B., AVENT, B. et al. (2021). Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **14** 1–210. https://doi.org/10.1561/2200000083

KOLOSKOVA, A., STICH, S. and JAGGI, M. (2019). Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the* 36*th International Conference on Machine Learning* **97** 3478–3487.

LAI, T. L. (2003). Stochastic approximation. *Ann. Statist.* **31** 391–406. MR1983535 https://doi.org/10.1214/aos/1051027873

LEE, S., LIAO, Y., SEO, M. H. et al. (2022). Fast and robust online inference with stochastic gradient descent via random scaling. In *The AAAI Conference on Artificial Intelligence*.

LI, T., SAHU, A., TALWALKAR, A. et al. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37** 50–60. https://doi.org/10.1109/MSP.2020.2975749

LI, X., LIANG, J., CHANG, X. et al. (2022). Statistical estimation and online inference via local SGD. In *Conference on Learning Theory* (*COLT*) **178** 1613–1661.

LIAN, X., ZHANG, C., ZHANG, H. et al. (2017). Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems* **30**.

MCMAHAN, B., MOORE, E., RAMAGE, D. et al. (2017). Communication-efficient learning of deep networks from decentralized data. *Int. Conf. Artif. Intell. Stat.* **54** 1273–1282.

NEDIĆ, A., OLSHEVSKY, A. and RABBAT, M. G. (2018). Network topology and communication-computation tradeoffs in decentralized optimization. *Proc. IEEE* **106** 953–976. https://doi.org/10.1109/JPROC.2018.2817461

NEDIĆ, A., OLSHEVSKY, A. and SHI, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM J. Optim.* **27** 2597–2633. MR3738851 https://doi.org/10.1137/16M1084316

NGUYEN, L. M., NGUYEN, P. H., RICHTÁRIK, P., SCHEINBERG, K., TAKÁČ, M. and VAN DIJK, M. (2019). New convergence aspects of stochastic gradient algorithms. *J. Mach. Learn. Res.* **20** Paper No. 176, 49. MR4048987

PANTELOPOULOS, A. and BOURBAKIS, N. G. (2010). A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.* **40** 1–12. https://doi.org/10.1109/TSMCC.2009.2032660

POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814 https://doi.org/10.1137/0330046

QIAN, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Netw.* **12** 145–151. https://doi.org/10.1016/S0893-6080(98)00116-6

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (*Proc. Sympos., Ohio State Univ., Columbus, Ohio*, 1971) 233–257. Academic Press, New York. MR0343355

RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins–Monro process Technical report, Cornell Univ. Operations Research and Industrial Engineering.

SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Stat.* **29** 373–405. MR0098427 https://doi.org/10.1214/aoms/1177706619

STICH, S. U. (2019). Local SGD converges fast and communicates little. In *International Conference on Learning Representations*.

SU, W. J. and ZHU, Y. (2018). Uncertainty Quantification for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent. arXiv. Available at arXiv:10.48550/ARXIV.1802.04876.

WANG, J., CHARLES, Z., XU, Z. et al. (2021). A field guide to federated optimization.

WANG, J. and JOSHI, G. (2021). Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *J. Mach. Learn. Res.* **22** 1–50. MR4329792

YUAN, K., ALGHUNAIM, S. A., YING, B. and SAYED, A. H. (2020). On the influence of bias-correction on distributed stochastic optimization. *IEEE Trans. Signal Process.* **68** 4352–4367. MR4144907 https://doi.org/10.1109/TSP.2020.3008605

YUAN, K., LING, Q. and YIN, W. (2016). On the convergence of decentralized gradient descent. *SIAM J. Optim.* **26** 1835–1854. MR3544854 https://doi.org/10.1137/130943170

ZHANG, Y., DUCHI, J. C. and WAINWRIGHT, M. J. (2013). Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.* **14** 3321–3363. MR3144464

ZHU, W., CHEN, X. and WU, W. B. (2023). Online covariance matrix estimation in stochastic gradient descent. *J. Amer. Statist. Assoc.* **118** 393–404. MR4571129 https://doi.org/10.1080/01621459.2021.1933498