IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

On the Convergence of Decentralized Federated Learning Under Imperfect Information Sharing

Vishnu Pandi Chellapandi, Antesh Upadhyay, Abolfazl Hashemi, and Stanislaw H Żak

Abstract— **Most of the current literature focused on centralized learning is centered around the celebrated average-consensus paradigm and less attention is devoted to scenarios where the communication between the agents may be imperfect. This paper presents three different algorithms of Decentralized Federated Learning (DFL) in the presence of imperfect information sharing modeled as noisy communication channels. The first algorithm, Federated Noisy Decentralized Learning (FedNDL1) comes from the literature, where the noise is added to the algorithm parameters to simulate the scenario of the presence of noisy communication channels. This algorithm shares parameters to form a consensus with the clients based on a communication graph topology through a noisy communication channel. The proposed second algorithm (FedNDL2) is similar to the first algorithm but with added noise to the parameters and it performs the gossip averaging before the gradient optimization. The proposed third algorithm (FedNDL3), on the other hand, shares the gradients through noisy communication channels instead of the parameters. Theoretical and experimental results show that under imperfect information sharing, the third scheme that mixes gradients is more robust in the presence of a noisy channel compared with the algorithms from the literature that mix the parameters.**

I. Introduction

In many applications, massive amounts of data are being generated from devices which are collected in centralized data centers and subsequently used for training machine learning models. However, challenges such as limited communication bandwidth, and privacy concerns make centralized learning unreliable and non-scalable. This led to an advancement in decentralized optimization (DFL) algorithms [1], [2] such as the Decentralized Federated Learning [3] and Federated Learning (FL) [4] that has applications in several domains like hospitals, smart cities, and connected vehicles [5], [6].

A. Related work

Common approach to decentralized optimization is consensus-based gradient descent methods [1], [7], [8], which share the computed parameter with other clients. The parameters are then averaged from all the clients based on a network topology that dictates the communication structure of the learning paradigm. These decentralized topologies minimize critical bottlenecks of centralized methods, such as communication latency, and bandwidth, as well as improve scalability and efficiency in large-scale settings [8], [9].

While communication efficiency is one of the critical elements and challenges for distributed learning and efforts, including communication compression, have been made in this regard [3], [10]–[14], these methods generally assume that the communication channels are noiseless. The performance of the trained model in the presence of noise should be one of the critical criteria in choosing the machine learning framework to ensure the robustness and safety of emerging applications that rely on distributed learning.

The effect of imperfect information sharing such as noisy communication or quantization noise in an average consensus algorithm in a distributed framework was studied in [15], [16]. However, the impact of various levels of noise has not been studied. Additionally, the study in [15] is limited to consensus problems only and does not encompass unique challenges that arise in modern decentralized optimization and learning, e.g., the inherent non-convexity of the learning objective. Other works including [17]–[21] study the impact of noise in server-assisted FL. These works require a server and have restrictive assumptions that typically are not satisfied in practical settings or are hard to verify.

In this paper, our primary focus is on DFL in the presence of noise in communication channels. Recently, [22]–[25] studied the performance of a two-time scale method [26] for DFL with channel noise while requiring the convexity of the objective function, uniformly bounded gradients, and access to the deterministic gradients. Note that these three considerations are very restrictive assumptions, especially in emerging settings in large-scale learning.

B. Contribution

Motivated by the existing gap between perfect information and noisy decentralized learning, in this paper, we model the presence of noise in the communication channels as a random vector with zero mean and different variances and study the performance of three decentralized FL algorithms by adding the noise to the parameters. The first algorithm, Federated Noisy Decentralized Learning (FedNDL1) was recently considered in [2], where the parameters were not subjected to any communication noise. In our analysis of this algorithm, we added noise to the parameters after the local SGD update. The new parameter with the noise is then exchanged with other clients through the gossip/mixing matrix and the global parameters are updated. This iteration, also known as communication rounds, continues throughout the training. The mixing matrix can be defined as a weighted adjacency matrix of a given communication graph.In the second algorithm (FedNDL2), which is related to [1], the

V. P. Chellapandi, A. Upadhyay, A. Hashemi, and S. H. Żak are with Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA. Emails: {cvp,aantesh,abolfazl,zak}@purdue.edu

Copyright ©2023 IEEE

noise is added before the consensus and local SGD update. In the third algorithm (FedNDL3), considered in the noiseless case in [27], the noise is added to the gradients instead of the parameters, and the result is exchanged with the clients.

We demonstrate, theoretically and empirically, that there are benefits in using FedNDL3 in the imperfect information setting that communicates the gradients. The intuition which is formalized theoretically is that the parameters are sensitive to the added noise while the gradients, which are already imperfect, are resilient. Therefore, the error stemming from weaker consensus in FedNDL3 is not as severe as the detrimental impact of noise on FedNDL1 and FedNDL2.

II. Problem Statement

In this section, we describe the problem structure, assumptions, and the proposed algorithms that we analyzed in this paper. We start with a standard DFL setup in which n clients have their own local datasets and collaborate with each other to update the global parameters. Formally, the problem can be represented as

$$\min_{x\in\mathbb{R}^d}\left[f(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x_i)\right], \tag{1}$$

where $f_i:\mathbb{R}^d\to\mathbb{R}$ for $i\in\{1,\ldots,n\}$ is the local objective function of the i^{th} client node. The stochastic formulation of the local objective function can be written

$$f_i(x_i)=\mathbb{E}_{\xi_i\sim\mathcal{D}_i}[\ell(x_i,\xi_i)], \tag{2}$$

where ξ_i is the data that has been sampled from the data distribution $\mathcal{D}_i$ for the i^{th} client. The function $\ell(x_i,\xi_i)$ is the loss function evaluated for each client and for each data sample ξ_i. Here $x_i\in\mathbb{R}^d$ is the parameter vector of client i, and $X\in\mathbb{R}^{d\times n}$ is the matrix formed using these parameter vectors. The primary objective of the clients is to achieve optimality through collaboration i.e., $x_i=x^*=\arg\min_{x\in\mathbb{R}^d}f(x)$, which is a global minimizer.

Definition 1 **(Mixing matrix).** *The mixing/gossip matrix, $W=[w_{ij}]\in[0,1]^{n\times n}$, is a non-negative, symmetric $(W=W^\top)$ and doubly stochastic $(W\mathbb{1}=\mathbb{1},\mathbb{1}^\top W=\mathbb{1}^\top)$ matrix, where $\mathbb{1}$ is the column vector of unit elements of size n*

We next describe three different scenarios of noise injection, resulting in three different algorithms.

FedNDL1: In this algorithm, each client in parallel performs updates first, see—lines 4–6, and then communicates the updated parameters to their neighbors. The communication depends on the topology of the communication graph, i.e., the mixing matrix, W, through a noisy channel (line 7).

$$x_i^{(t+1)} = \sum_{j=1}^{n} w_{ij}\ (x_j^{(t+\frac{1}{2})}+\delta_j^{(t)}), \tag{3}$$

where $\delta_j^{(t)}\in\mathbb{R}^d$, is a zero mean random noise and $x_j^{(t+\frac{1}{2})}$ is the vector of parameters sent by client j. Since we assume the noise to have a zero mean, the noise variance is

$$D_{t,j}^2=\mathbb{E}[\|\delta_j^{(t)}\|^2]. \tag{4}$$

FedNDL2: In this algorithm, we perform the consensus step (line 9) before computing the individual gradients and parameters(lines 10–12),

$$x_i^{(t+\frac{1}{2})}=\sum_{j=1}^{n}w_{ij}(x_j^{(t)}+\delta_j^{(t)}), \tag{5}$$

FedNDL3: , In this algorithm,the clients share their gradients over a noisy communication channel instead of the weights followed by the SGD update. This idea comes from our Noisy-FL motivation and the fact that SGD is inherently a noisy process. So, pursuing this scenario gives more flexibility to handle the noise as a part of the SGD process.

$$x_i^{(t+1)} = x_i^{(t)}-\eta_t\sum_{j=1}^{n}w_{ij}\ (g_j^{(t)}+\delta_j^{(t)}), \tag{6}$$

where $g_j^{(t)}$ refers to the gradient of client j at iteration t

Algorithm FedNDL1, FedNDL2, and FedNDL3

```
1: Input: For each node $i$ initialize: $x_i^{(0)} \in \mathbb{R}^d$, step size $\{\eta_t\}_{t=0}^{T-1}$, mixing matrix $W$, noise from the communication channel $\delta^{(t)}$
2: for $t = 0,\ldots,T$ do
3:     FedNDL1:
4:     Run in parallel for each client $i$
5:     Sample $\xi_i^{(t)}$, compute $g_i^{(t)} = \widetilde{\nabla} f_i(x_i^{(t)},\xi_i^{(t)})$
6:     $x_i^{(t+\frac{1}{2})} = x_i^{(t)} - \eta_t g_i^{(t)}$
7:     $x_i^{(t+1)} = \sum_{j=1}^{n} w_{ij}\ (x_j^{(t+\frac{1}{2})} + \delta_j^{(t)})$
8:     FedNDL2:
9:     $x_i^{(t+\frac{1}{2})} = \sum_{j=1}^{n} w_{ij}\ (x_j^{(t)} + \delta_j^{(t)})$
10:    Run in parallel for each clients $i$
11:    Sample $\xi_i^{(t)}$ , $g_i^{(t+\frac{1}{2})} = \widetilde{\nabla} f_i(x_i^{(t+\frac{1}{2})},\xi_i^{(t)})$
12:    $x_i^{(t+1)} = x_i^{(t+\frac{1}{2})} - \eta_t g_i^{(t+\frac{1}{2})}$
13:    FedNDL3:
14:    Run in parallel for each client $i$
15:    Sample $\xi_i^{(t)}$, compute $g_i^{(t)} = \widetilde{\nabla} f_i(x_i^{(t)},\xi_i^{(t)})$
16:    $x_i^{(t+1)} = x_i^{(t)} - \eta_t \sum_{j=1}^{n} w_{ij}\ (g_j^{(t)} + \delta_j^{(t)})$
17: end for
```

A. Assumptions

We now discuss the assumptions made in our analysis of the algorithms. They are standard assumptions used in the analysis of decentralized algorithms, see [2], [3], [28].

Assumption 1 **(Smoothness).** *The objective function $\ell(x,\xi)$ is L-smooth with respect to x, for all ξ. Each $f_i(x)$ is L-smooth, that is,*

$$\|\nabla f_i(x)-\nabla f_i(y)\|\le L\|x-y\|,\qquad \textit{for all x,y}. \tag{7}$$

Hence the function f is also L-smooth.

Assumption 2 **(Bounded Variance).** *The variance of the stochastic gradient of each client i is bounded,*

$$\mathbb{E}[\|\widetilde{\nabla} f_i(x_i^t,\xi_i^t)-\nabla f_i(x_i^t)\|^2]\le\sigma^2,$$

where ξ_i^t denotes random batch of samples in client node i for t^{th} round, and $\widetilde{\nabla} f_i(x_i^t,\xi_i^t)$ denotes the stochastic gradient. In addition, we also assume that the stochastic gradient is unbiased, i.e., $\mathbb{E}[\widetilde{\nabla}(f_i(x_i^t,\xi_i^t))]=\nabla f_i(x_i^t)$.

Assumption 3 **(Mixing matrix).** *The mixing matrix W satisfies for $\rho\in(0,1]$,*

$$\|(\bar{X}-X)W\|_F^2\le(1-\rho)\|\bar{X}-X\|_F^2,$$

which means that the gossip averaging step brings the columns of $X \in \mathbb{R}^{d\times n}$ closer to the row-wise average, that is, $\bar{X} = X\frac{\mathbb{1}\mathbb{1}^\top}{n}$.

Note that standard topologies such as ring, torus, and fully-connected satisfy the above assumption.

Assumption 4 (Bounded Client Dissimilarity (BCD)). *For all $x \in \mathbb{R}^d$, where B is a constant.*

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f(x)\|^2 \le B^2,$$

The above assumption is made to limit the extent of client heterogeneity. While gradient tracking methods [29] don't require this assumption, they suffer from increased communication cost and the variance, which limits their practicality [30]. Note that this assumption is only used in the analysis of FedNDL1 and FedNDL2.

Assumption 5 (Noise model). *The noise present due to contamination of communication channel $\delta_i^{(t)}$ is independent, has zero mean and bounded variance, that is, $\mathbb{E}[\delta_i^{(t)}] = 0$ and $\mathbb{E}[||\delta_i^{(t)}||^2] = D_{t,i}^2 < \infty$.*

This assumption is specific to the imperfect information sharing setup and is considered recently in [24], [25], [31].

Assumption 6 (Bounded Recursive Consensus Error). *Let the consensus error be defined as $(C.E)_t = \frac{1}{n}\|\bar{X}_t - X_t\|_F^2$. We assume that the consensus error is upper bounded,*

$$\mathbb{E}[(C.E)_{t+1}] \le \alpha_t \, \mathbb{E}[(C.E)_t] + \gamma_t,$$

where $\alpha_t \in (0,1)$ and $\gamma_t \ge 0$.

Remark 1. We use the above assumption in the convergence analysis of FedNDL3. Theoretically speaking Assumption 6 can be viewed as a general formulation of the recursive upper-bound on the evolution of the consensus error in the analysis of decentralized SGD—refer, e.g., [2]. Assumption 6 is trivially satisfied for the first two algorithms; FedNDL1 and FedNDL2. Please refer to the proof in [32]. This assumption is also satisfied for FedNDL3 if the network topology is fully connected or the union of a finite collection of consecutive communication graphs is fully connected; see Figure 2. We further note that using multi-round gossiping [3] or acceleration methods such as Chebyshev acceleration [33], [34] this assumption may be satisfied by FedNDL3 as well and α_t and γ_t can be significantly reduced.

III. Convergence Analysis

In this section, we state the main theorem providing an upper bound on the convergence errors of FedNDL1, FedNDL2, and FedNDL3. The convergence results are for non-convex L-smooth loss functions and noisy channels.

Theorem 1 (Smooth non-convex cases for Noisy-DFL).

$$\text{Let } LHS = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{x}_t)\|^2] + \frac{L^2}{T}\sum_{t=1}^{T}\mathbb{E}[(C.E)_t].$$

Suppose Assumptions 1–5, and 6 (only for FedNDL3) hold. Let $\eta L < \frac{1}{12}$, $\eta = \mathcal{O}(\frac{1}{\sqrt{T}})$ and $\bar{D}^2 = \frac{1}{nT}\sum_{t,i=1,1}^{T,n} D_{t,i}^2$, then

- ***FedNDL1**: For $\eta L < \frac{\rho}{2\sqrt{6}}$,*

$$LHS = \mathcal{O}\Big(\frac{\rho}{n\sqrt{T}}\sigma^2 + \frac{\rho^2}{T}B^2 + \frac{T^{\frac{3}{2}}}{\rho}\bar{D}^2\Big), \quad (8)$$

- ***FedNDL2**: For $\eta L < \frac{\rho}{4\sqrt{3}}$,*

$$LHS = \mathcal{O}\Big(\frac{\rho}{n\sqrt{T}}\sigma^2 + \frac{\rho^2}{T}B^2 + \frac{T^{\frac{3}{2}}}{\rho}\bar{D}^2\Big), \quad (9)$$

- ***FedNDL3**:*

$$LHS = \mathcal{O}\Big(\frac{1}{n\sqrt{T}}\sigma^2 + \frac{1}{T}\sum_{t=1}^{T}\frac{\gamma_t}{\alpha_t} + \frac{1}{\sqrt{T}}\bar{D}^2\Big), \quad (10)$$

where all expectations are w.r.t. the data and the noise.

Proof. We prove the theorem for FedNDL3. The proofs for FedNDL1 and FedNDL2 are similar [32]. We start our proof by upper bounding the second moment of the gradient on the average of iterates by using the L-the smoothness of the loss function. The second moment here is bounded by an inaccurate initialization, the variance of the stochastic gradients, noise present due to imperfect channels, and the consensus error function, $(C.E)_t$. Recall that $\bar{X}_{t+1} = \bar{X}_t - \frac{\eta}{n}\sum_{i=1}^{n}(\widetilde{\nabla}_{t,i} + \delta_{t,i})$. Hence,

$$f(\bar{X}_{t+1}) \le f(\bar{X}_t) + \langle \nabla f(\bar{X}_t), \bar{X}_{t+1} - \bar{X}_t\rangle + \frac{L}{2}\|\bar{X}_{t+1} - \bar{X}_t\|^2$$

$$\le f(\bar{X}_t) \underbrace{-\eta\langle \nabla f(\bar{X}_t), \frac{1}{n}\sum_{i=1}^{n}\widetilde{\nabla}_{t,i}\rangle}_{Term\,(A)} \underbrace{-\eta\langle \nabla f(\bar{X}_t), \frac{1}{n}\sum_{i=1}^{n}\delta_{t,i}\rangle}_{\text{Expectation wrt noise=0}}$$

$$+ \underbrace{\frac{L\eta^2}{2}\|\frac{1}{n}\sum_{i=1}^{n}(\widetilde{\nabla}_{t,i} + \delta_{t,i})\|^2}_{Term\,(B)} \quad \text{(Using } L\text{-smoothness).}$$

Taking the expectation of Term (A) w.r.t noise gives, $\mathbb{E}[A] = -\frac{\eta}{2}\Big[\mathbb{E}[\|\bar{\nabla}_t\|^2] + \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}\|^2] - \mathbb{E}[\|\bar{\nabla}_t - \frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}\|^2]\Big]$.

We represent the Term (B) as, $B = \frac{L\eta^2}{2}\Big[\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\nabla}_{t,i}\|^2 + \|\frac{1}{n}\sum_{i=1}^{n}\delta_{t,i}\|^2 + 2\langle\frac{1}{n}\sum_{i=1}^{n}\widetilde{\nabla}_{t,i}, \frac{1}{n}\sum_{i=1}^{n}\delta_{t,i}\rangle\Big]$.

Taking the expectation of Term (B) w.r.t noise, we get.

$$\mathbb{E}[B] \le \frac{L\eta^2}{2}\Big[\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\nabla}_{t,i}\|^2 + \frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2\Big].$$

Taking in the above the expectation w.r.t data gives

$$\mathbb{E}[B] \le \frac{L\eta^2}{2}\Big[\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\nabla}_{t,i}\|^2] + \frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2\Big]$$

$$\le \frac{L\eta^2}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\nabla}_{t,i} - \nabla_{t,i} + \nabla_{t,i}\|^2] + \frac{L\eta^2}{2}\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2$$

$$\le \frac{L\eta^2}{2}\Big(\frac{\sigma^2}{n} + \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}\|^2] + \frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2\Big).$$

Taking the above into account, we obtain

$$\mathbb{E}[f(\bar{X}_{t+1})] \le \mathbb{E}[f(\bar{X}_t)] - \frac{\eta}{2}\Big[\mathbb{E}[\|\bar{\nabla}_t\|^2] + \underbrace{\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}\|^2]}_{Term(M)}$$

$$- \mathbb{E}[\|\bar{\nabla}_t - \frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}\|^2]\Big]$$

$$+\frac{L\eta^2}{2}\Big[\frac{\sigma^2}{n}+\underbrace{\mathbb{E}\,\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}\|^2]}_{Term(N)}+\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2\Big].$$

Dropping Term (M) increases the right-hand side. In the equation above reducing N using Young's Inequality gives

$$N=L\eta^2\,\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}-\bar{\nabla}_t+\bar{\nabla}_t\|^2]$$

$$\leq 2L\eta^2\,\mathbb{E}[\|\bar{\nabla}_t\|^2]+2L\eta^2\,\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}-\bar{\nabla}_t\|^2].$$

Hence,

$$\mathbb{E}[f(\bar{X}_{t+1})]\leq\mathbb{E}[f(\bar{X}_t)]-\frac{\eta}{2}\,\mathbb{E}[\|\bar{\nabla}_t\|^2]+\frac{\eta}{2}\,\mathbb{E}[\|\bar{\nabla}_t-\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}\|^2]+\frac{L\eta^2}{2}\frac{\sigma^2}{n}+L\eta^2\,\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}-\bar{\nabla}_t\|^2]$$

$$+L\eta^2\,\mathbb{E}[\|\bar{\nabla}_t\|^2]+\frac{L\eta^2}{2}\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2$$

$$\leq\mathbb{E}[f(\bar{X}_t)]-\frac{\eta}{2}(1-2L\eta)\,\mathbb{E}[\|\bar{\nabla}_t\|^2]+\frac{L\eta^2}{2}\frac{\sigma^2}{n}$$

$$+\frac{\eta}{2}(1+2L\eta)\underbrace{\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}-\bar{\nabla}_t\|^2]}_{Term(O)}+\frac{L\eta^2}{2}\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2.$$

Bounding Term (O), we obtain,

$$O=\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{t,i}-\bar{\nabla}_t\|^2]\leq\frac{L^2}{n}\,\mathbb{E}[\|X^t-\bar{X}^t\|_F^2].$$

Substituting back and denoting $\frac{1}{n}\|\bar{X}^t-X^t\|_F^2=(C.E)_t$,

$$\mathbb{E}[f(\bar{X}_{t+1})]\leq\mathbb{E}[f(\bar{X}_t)]-\frac{\eta}{2}(1-2L\eta)\,\mathbb{E}[\|\bar{\nabla}_t\|^2]+\frac{L\eta^2}{2}\frac{\sigma^2}{n}$$

$$+\frac{L\eta^2}{2}\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2+\frac{L^2\eta}{2}(1+2L\eta)\,\mathbb{E}[(C.E)_t].$$

In the next step, we proceed with upper bounding the $(C.E)_{t+1}$ followed by defining a potential function to jointly bound the expected gradient norm and the consensus error without requiring restrictive and impractical assumptions such as the bounded gradient norm assumption. Using the assumption on the consensus error at $t+1$ gives,

$$\mathbb{E}[(C.E)_{t+1}]\leq\alpha_t\,\mathbb{E}[(C.E)_t]]+\gamma_t.$$

Let ψ_t denote a potential function defined as

$$\psi_t=\mathbb{E}[f(\bar{X}_t)]+\phi_t\,\mathbb{E}[(C.E)_t],\ \text{where } \phi_t>0.$$

Now, we use the potential function to complete the proof.

$$\psi_{t+1}-\psi_t=\big\{\,\mathbb{E}[f(\bar{X}_{t+1})]-\mathbb{E}[f(\bar{X}_t)]\big\}$$

$$+\phi_{t+1}\,\mathbb{E}[(C.E)_{t+1}]-\phi_t\,\mathbb{E}[(C.E)_t]$$

$$\leq-\frac{\eta}{2}(1-2L\eta)\,\mathbb{E}[\|\bar{\nabla}_t\|^2]+\frac{L\eta^2}{2}\Big[\frac{\sigma^2}{n}+\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2$$

$$+(1+2L\eta)\,\mathbb{E}[(C.E)_t]\Big]+\phi_{t+1}\alpha_t\,\mathbb{E}[(C.E)_t]+\phi_{t+1}\gamma_t$$

$$-\phi_t\,\mathbb{E}[(C.E)_t]$$

$$\leq-\frac{\eta}{2}(1-2L\eta)\,\mathbb{E}[\|\bar{\nabla}_t\|^2]+\frac{L\eta^2}{2}\Big[\frac{\sigma^2}{n}+\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2+\phi_{t+1}\gamma_t\Big]$$

$$+\Big(\frac{L^2\eta}{2}(1+2L\eta)+\phi_{t+1}\alpha_t-\phi_t\Big)\,\mathbb{E}[(C.E)_t].$$

Pick ϕ_t such that, $\phi_t>\frac{L^2\eta}{2}(1+2L\eta)+\phi_{t+1}\alpha_t$.

Let $\phi_t=L^2\eta(1+2L\eta)+2\phi_{t+1}\alpha_t$ and $\eta L<\frac{1}{2}$, then

$$\mathbb{E}[\|\bar{\nabla}_t\|^2]+\frac{\Big(L^2\eta(1+2L\eta)+2\phi_{t+1}\alpha_t\Big)}{\eta(1-2L\eta)}\,\mathbb{E}[(C.E)_t]$$

$$\leq\frac{2(\psi_t-\psi_{t+1})}{\eta(1-2L\eta)}+\frac{L}{(1-2L\eta)}\frac{\sigma^2}{n}$$

$$+\frac{L\eta}{(1-2L\eta)}\frac{1}{n}\sum_{i=1}^{n}D_{t,i}^2+\frac{2\phi_{t+1}\gamma_t}{\eta(1-2L\eta)}$$

Let $C=\frac{L^2\eta(1+2L\eta)+2\phi_{t+1}\alpha_t}{\eta(1-2L\eta)}$. Solving for ϕ_{t+1} gives

$$\phi_{t+1}=\frac{C\eta(1-2\eta L)-L^2\eta(1+2\eta L)}{2\alpha_t}.$$

Hence, for $C=2L^2$ and $\eta L<\frac{1}{6}$, we have $\phi_{t+1}=\frac{L^2\eta(1-6\eta L)}{2\alpha_t}$. Now, summing the above wrt to T and dividing by T gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\bar{\nabla}_t\|^2]$$

$$+\frac{1}{T}\sum_{t=1}^{T}\frac{\Big(L^2\eta(1+2L\eta)+2\phi_{t+1}\alpha_t\Big)}{\eta(1-2L\eta)}\,\mathbb{E}[(C.E)_t]$$

$$\leq\frac{1}{1-2L\eta}\Big[\frac{2\sum_{t=1}^{T}(\psi_t-\psi_{t+1})}{T\eta}+\frac{L\sigma^2}{n}+\frac{2\sum_{t=1}^{T}\phi_{t+1}\gamma_t}{T\eta}$$

$$+\frac{L\eta}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{t,i}^2\Big].\quad(11)$$

Now, we telescope over the potential function for $t=\{1,\ldots,T\}$ and dividing it by T, we get

$$\frac{\psi_{T+1}-\psi_1}{T}\geq\frac{f^*-f(\bar{X}_1)-\phi_1(C.E)_1}{T}.$$

Using the above inequality along with specific choices of ϕ and ϕ_t yields :

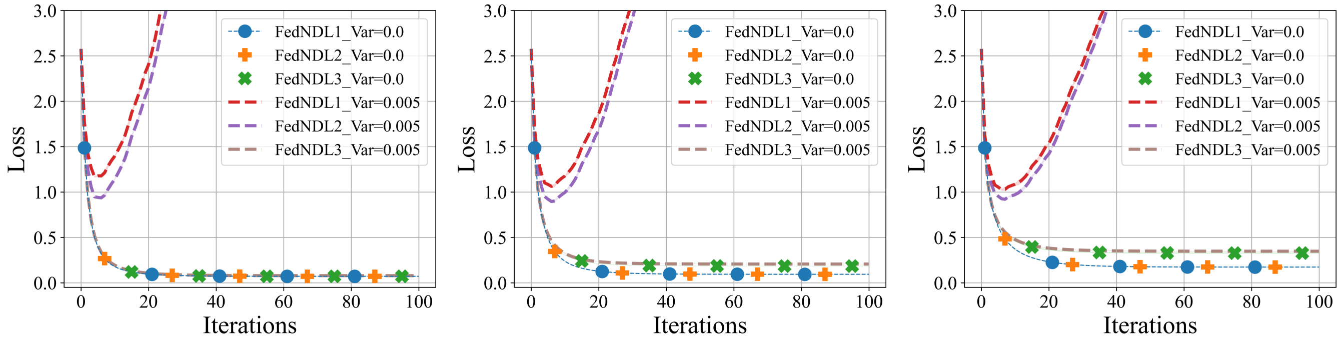
$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{X}_t)\|^2]+\frac{2L^2}{T}\sum_{t=1}^{T}\mathbb{E}[(C.E)_t]\leq$$

$$\frac{2(f(\bar{X}_1)-f^*+\phi_1(C.E)_1)}{\eta(1-2L\eta)T}+\frac{L\eta}{n(1-2L\eta)}\sigma^2$$

$$+\frac{L^2(1-6L\eta)}{T(1-2L\eta)}\sum_{t=1}^{T}\frac{\gamma_t}{\alpha_t}+\frac{L\eta}{(1-2L\eta)}\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{t,i}^2.$$

■

Theorem 1 establishes a worst-case upper bound on the convergence of the three algorithms studied in the paper. In particular, the theorem jointly bounds the expected gradient norm, which is a notion of approximate first-order stationarity of the average iterate $\bar{x}_t$, and the consensus error. The convergence bounds consist of three terms: the first term effectively captures the error arising from inaccurate initialization and stochasticity of the first-order oracle, which matches the error of centralized SGD. The second term captures the effect of data heterogeneity, and the last term captures the adverse effect of imperfect communication modeled as communication noise.

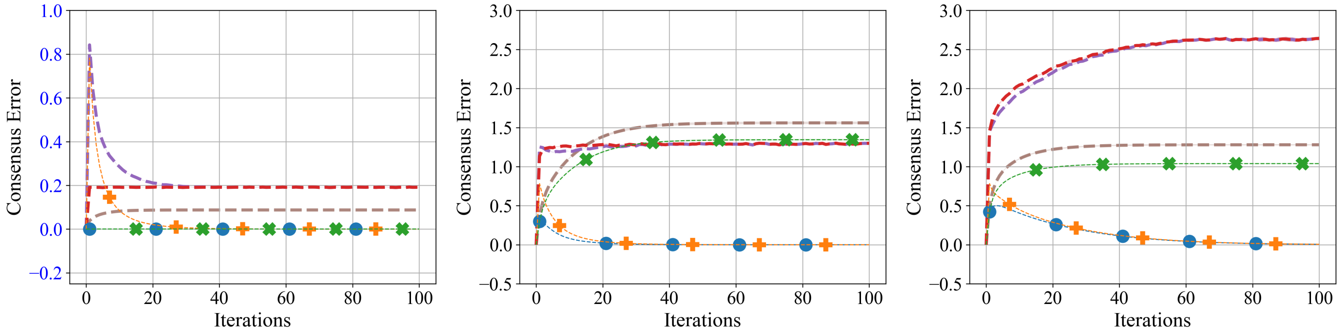


(a) Fully-connected topology

(b) Torus topology

(c) Ring topology

Fig. 1: Loss vs. iterations with and without noise for different communication topologies.



(a) Fully-connected topology

(b) Torus topology

(c) Ring topology

Fig. 2: Consensus error vs. iterations with and without noise for different communication topologies. Note the different Y-axis scale in Fig (a) as compared with Fig (b) and (c) for better readability.

Theorem 1 also captures the impact of presence channel noise on the convergence of studied algorithms. Specifically, (8) and (9) indicate that FedNDL1 and FedNDL2 suffer from a severe impact of noise on the worst-case convergence: T increases the guarantee on finding a stationary solution and consensus error weakens. In fact, the error increases with T. We verify these results numerically in Section IV. Furthermore, as the connectivity of the communication graphs decreases (corresponding to a smaller ρ), the impact of noise increases. This point is further confirmed in numerical simulations in Section IV. With regard to FedNDL3, Theorem 1 establishes that the algorithm is resilient to the presence of noise. In particular, in contrast with the convergence bounds of FedNDL1 and FedNDL2, the last term in (10) decreases with T. This theoretically-grounded property is linked to SGD which inherently is a noisy process and thus is more resilient with respect to noise. Theorem 1 further shows that, different from FedNDL1 and FedNDL2, the impact of noise is independent of the communication topology as the last term in (10) is independent of α_t and γ_t. In Section IV, we numerically verify these two properties of FedNDL3.

IV. Experiments

In this section, we perform several experiments on regression problems to verify the impact of noise on the convergence of the three proposed algorithms as established in Theorem 1. We consider the case when the number of clients, n = 16. The experiments are repeated three times, and the results (loss/consensus error) are averaged. We use the mean-squared error loss function with L_2 regularization. The learning rate of the model is set as 0.2 with a decay of 0.9 with every iteration. We generate data samples (m = 10000) $(x_i; y_i)_{i=1}^m$ according to $y_i = \langle w, x_i \rangle + \epsilon_i$, where $w \in \mathbb{R}^{2000}$, $x_i \sim \mathcal{N}(0; I_{2000})$ and noise, $\epsilon_i \sim \mathcal{N}(0, 0.05)$.

The experiments are performed with various levels of noise variance, $D_{t,i}^2$ for all t, i, as described in the algorithms, for various communication topologies, namely the ring, torus, and fully connected network. The nonzero weights in the mixing matrix for ring topology are equal to $\frac{1}{3}$, in the torus topology $\frac{1}{5}$, and fully connected topology, $\frac{1}{n}$.

We first perform the experiments with no noise as a baseline and then gradually increase the noise variance to study the robustness of the algorithms. For the purpose of consistency, we have shown the results of the experiments with noise variance $D_{t,i}^2 = 0.005$ in Figures 1 and 2 along with no noise scenario. We observed, see Figure 1, that the algorithms FedNDL1 and FedNDL2 perform poorly in terms of convergence due to noise presence which is consistent with Theorem 1. In addition, as seen in Figure 2, the consensus error also increases with the noise consistent with our theoretical analysis presented in Theorem 1.

On the other hand, the algorithm FedNDL3 is observed to be the most robust as it does not diverge in the presence of added channel noise. The noise term for the FedNDL3 in the upper bound given in Theorem 1 is of order $\mathcal{O}(T^{-\frac{1}{2}})$, whereas it is of order $\mathcal{O}(T^{\frac{3}{2}})$ for FedNDL1 and FedNDL2. This effect of the noise can also be observed in Figure 1.

The consensus error depends on the topology of the communication network. We observed that the consensus error is low for the fully connected network and high for the ring topology for the same algorithm in the presence of noise which is also consistent with Theorem 1.

V. Conclusion

We studied the impact of noisy communication channels on the convergence of DFL. We proposed multiple scenarios for establishing consensus in the presence of noise and provided experimental results on all the algorithms. Additionally, we provided theoretical results for FedNDL1, FedNDL2, and FedNDL3, under the assumption of smooth non-convex function, and we observed that in FedNDL3, the noise term in the upper bound given by Theorem 1 is of order $\mathcal{O}(T^{-\frac{1}{2}})$ and is independent of communication topology. In contrast, the impact of noise on the convergence of FedNDL1 and FedNDL2 increases with T and weaker communication structures. We conducted numerical experiments and observed that FedNDL3 is more robust against the added noise than the other two algorithms analyzed in this paper.

References

[1] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[2] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5381–5393.

[3] A. Hashemi, A. Acharya, R. Das, H. Vikalo, S. Sanghavi, and I. Dhillon, "On the benefits of multiple gossip steps in communication-constrained decentralized federated learning," *IEEE Trans. Parallel and Distributed Systems*, vol. 33, no. 11, pp. 2727–2739, 2021.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*. PMLR, 2017, pp. 1273–1282.

[5] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Zak, and Z. Wang, "Federated learning for connected and automated vehicles: A survey of existing approaches and challenges," *arXiv preprint arXiv:2308.10407*, 2023.

[6] S. Pandya, G. Srivastava, R. Jhaveri, M. R. Babu *et al.*, "Federated learning for smart cities: A comprehensive survey," *Sustainable Energy Technologies and Assessments*, vol. 55, p. 102987, 2023.

[7] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[8] J. N. Tsitsiklis, "Problems in decentralized decision making and computation." Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, Tech. Rep., 1984.

[9] J. M. Hendrickx and M. G. Rabbat, "Stability of decentralized gradient descent in open multi-agent systems," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 4885–4890.

[10] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE transactions on signal processing*, vol. 68, pp. 2128–2142, 2020.

[11] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2150–2167, 2020.

[12] Y. Chen, A. Hashemi, and H. Vikalo, "Communication-efficient variance-reduced decentralized stochastic optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 67, no. 12, pp. 6583–6594, 2021.

[13] ——, "Decentralized optimization on time-varying directed graphs under communication constraints," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3670–3674.

[14] Y. Li, X. Wei, Y. Li, Z. Dong, and M. Shahidehpour, "Detection of false data injection attacks in smart grid: A secure federated deep learning approach," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4862–4872, 2022.

[15] R. Carli, F. Fagnani, P. Frasca, T. Taylor, and S. Zampieri, "Average consensus on networks with transmission noise or quantization," in *European Control Conference*. IEEE, 2007, pp. 1852–1857.

[16] T. Qin, S. R. Etesami, and C. A. Uribe, "Communication-efficient decentralized local SGD over undirected networks," in *IEEE Conference on Decision and Control*. IEEE, 2021, pp. 3361–3366.

[17] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[18] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2019.

[19] S. Xia, J. Zhu, Y. Yang *et al.*, "Fast convergence algorithm for analog federated learning," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[20] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3796–3811, 2021.

[21] H. Guo, A. Liu, and V. K. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 197–210, 2020.

[22] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.

[23] M. M. Vasconcelos, T. T. Doan, and U. Mitra, "Improved convergence rate for a distributed two-time-scale gradient method under random quantization," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 3117–3122.

[24] H. Reisizadeh, B. Touri, and S. Mohajer, "Distributed optimization over time-varying graphs with imperfect sharing of information," *IEEE Transactions on Automatic Control*, 2022.

[25] H. Reisizadeh, A. Gokhale, B. Touri, and S. Mohajer, "Almost sure convergence of distributed optimization with imperfect information sharing," *arXiv preprint arXiv:2210.05897*, 2022.

[26] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE journal of selected topics in signal processing*, vol. 5, no. 4, pp. 772–790, 2011.

[27] M. Rabbat, "Multi-agent mirror descent for decentralized stochastic optimization," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2015, pp. 517–520.

[28] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *International Conference on Machine Learning*, 2019, pp. 3478–3487.

[29] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[30] K. Yuan, W. Xu, and Q. Ling, "Can primal methods outperform primal-dual methods in decentralized dynamic optimization?" *IEEE Transactions on Signal Processing*, vol. 68, pp. 4466–4480, 2020.

[31] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Transactions on Cognitive Communications and Networking*, 2022.

[32] V. P. Chellapandi, A. Upadhyay, A. Hashemi, and S. H. Żak, "On the convergence of decentralized federated learning under imperfect information sharing," *arXiv preprint arXiv:2303.10695*, 2023.

[33] K. Scaman, F. Bach, S. Bubeck *et al.*, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *international conference on machine learning*. PMLR, 2017, pp. 3027–3036.

[34] M. Arioli and J. Scott, "Chebyshev acceleration of iterative refinement," *Numerical Algorithms*, vol. 66, no. 3, pp. 591–608, 2014.