FedDec: Peer-to-peer Aided Federated Learning

Marina Costantini, Giovanni Neglia, Thrasyvoulos Spyropoulos

FedDec: Peer-to-peer Aided Federated Learning

Marina Costantini*, Giovanni Neglia†, Thrasyvoulos Spyropoulos*‡
*Eurecom, Sophia Antipolis, France
†Inria, Université Côte d'Azur, Sophia Antipolis, France
‡Technical University of Crete, Crete, Greece
marina.costantini@eurecom.fr, giovanni.neglia@inria.fr, spyropoulos@tuc.gr

***Abstract*—Federated learning (FL) has enabled training machine learning models that exploit the data of multiple agents without compromising privacy. However, FL is known to be vulnerable to data heterogeneity, partial device participation, and infrequent communication with the server, which are nonetheless distinctive characteristics of this framework. While much of the literature has tackled these weaknesses using different tools, only a few works have considered inter-agent communication to improve FL's performance. In this work, we present FedDec, an algorithm that interleaves peer-to-peer communication and parameter averaging between the local gradient updates of FL. We analyze the convergence of FedDec and show that inter-agent communication alleviates the negative impact of infrequent communication rounds with the server by reducing the dependence on the number of local updates H from $O(H^2)$ to $O(H)$. Furthermore, our analysis reveals that the term improved in the bound vanishes quickly the more connected the network is. We confirm the predictions of our theory in numerical simulations, where we show that FedDec converges faster than FedAvg, and that the gains are greater as either H or the connectivity of the network increase.**

***Index Terms*—Federated Learning, Decentralized Optimization, Multi-agent Optimization, Distributed Machine Learning**

I. Introduction

Federated learning (FL) is a recent machine learning framework that allows multiple agents, each with their own dataset, to train a model collaboratively without sharing their data [1]–[3]. The *federated* setting assumes that all agents are connected to a server that can communicate with each of them and that is in charge of aggregating the agents' updates to obtain the global model. In FL (i) multiple SGD updates (called *local updates*) can happen before a new server communication round takes place, and (ii) not all devices need to engage in the server communication round (usually called *partial participation*).

In contrast, the *decentralized* setting assumes that the agents are interconnected in a network and each of them can exchange optimization values (either parameters or gradients, depending on the algorithm) with its direct neighbors [4]–[7]. In this setting, every node averages all of its neighbors' values before taking a new gradient step. Algorithms for this setting are designed such that the local parameters of all nodes converge to the global minimizer, while in FL it is the central server who keeps track of the most recent parameter value and broadcasts it to all agents every once in a while.

The attractive feature of FL of allowing to have server communication rounds every once in a while comes at a cost: the more infrequent the server communication rounds are (i.e., the more local updates are performed at the agents), the slower is the convergence [8], [9]. For this reason, in this paper we propose to exploit inter-agent communication to reduce the negative impact of infrequent server communication rounds. Given that (i) each agent is expected to have much fewer neighbors than the total number of agents, and (ii) short-range inter-agent communications allow for spectrum reuse, agents can communicate much more often between them than with the server [10].

We propose FedDec, an FL algorithm where the agents can exchange and average their parameters with those of their neighbors in between the local SGD steps. We show that for smooth and strongly convex functions, this reduces the dependence of the convergence bound on the number of local SGD steps H from $O(H^2)$ [8], [11] to $O(H)$. Furthermore, we show that, in our analysis, the extra H factor is replaced by a value α that depends on the spectrum of the graph defining the inter-agent communication. Since the value of α quickly decreases as the network becomes more connected, our result indicates that for mildly connected networks H can be increased without severely hurting convergence speed. Moreover, since the term improved is also the one affected by partial device participation, denser connectivity also allows for sampling fewer devices without affecting convergence.

Peer-to-peer communication within FL has been considered a few times in the past. Works [12], [13] analyze very general settings that have FL with inter-agent communication as a particular case, although [13] requires full device participation. Works [14]–[17] show that FL with inter-agent exchange converges at the same rate as standard FL and outperforms it in simulations. However, none of these works has shown *analytically* how inter-agent communication *reduces the impact of local updates and partial participation* on convergence, and in particular, how this reduction depends on inter-agent connectivity. The exceptions are [18], [19], but unlike this work, they do not consider failures in the inter-agent links and their setting leads to $O(1/\sqrt{T})$ convergence.

Our contributions can be summarized as follows:

- We introduce FedDec, an FL algorithm where the agents can average their parameters with those of their neighbors in between the SGD steps and allows for failures in the inter-agent communication links.
- We prove that, for non-iid data, partial device participation, and smooth and strongly convex objectives, FedDec converges at the $O(1/T)$ rate (where T is the total number of

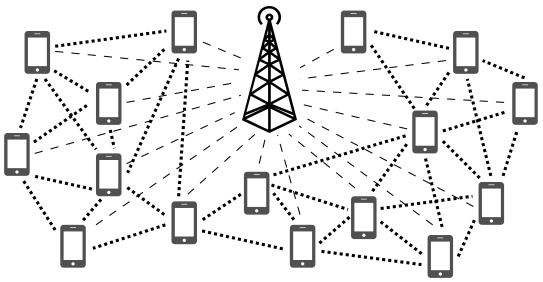

Fig. 1: The FedDec setting.

iterations) of FL algorithms with no inter-agent communication [8], [11], but improves the dependence on the number of local updates H from $O(H^2)$ to $O(H)$.

- Furthermore, we show that the improved term is multiplied by a quantity that depends on the spectrum of the inter-agent communication network, and which quickly vanishes the more connected the network is.
- We support our theoretical findings with numerical simulations, where we confirm that the performance of FedDec with respect to FedAvg [1], [8] (its counterpart without inter-agent exchange) increases with both H and the connectivity of the network.

II. System Model and FedDec

We consider a system where n agents can exchange messages with a central server and also with some other nearby agents. We assume that the inter-agent communication links may fail at some iterations (e.g. due to outage), but when all links are active the agents form a connected network (see Figure 1). Each node $i \in [n]$ has a local cost $F_i : \mathbb{R}^d \to \mathbb{R}$,

$$F_i(\mathbf{z}) := \mathbb{E}_{\psi_i \sim \mathcal{D}_i} F_i(\mathbf{z}, \psi_i),$$

where $\mathcal{D}_i$ can be an underlying local data distribution from which new samples (or mini-batches) are drawn each time an SGD step is taken, or the uniform distribution over a static dataset. Note that the $\mathcal{D}_i$ can be different at each node. The objective of the nodes and the server is to find the minimizer

$$\mathbf{z}^* := \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^d} f(\mathbf{z}), \qquad f(\mathbf{z}) := \frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{z})$$

under the constraints that nodes can only communicate with their direct neighbors (high-bandwidth links) and every once in a while they can get a request from the server to send their current parameter values (low-bandwidth links). In wireless settings, these capacities are imposed by the shared nature of the cellular medium. While the communication with the server is constrained by the bandwidth available, device-to-device communications in the short-range allow for spectrum reuse, and thus for higher throughput [10].

At each server communication round, the server samples the devices uniformly[1] at random with replacement to form an index pool $\mathcal{S}_t$ of devices that it will poll during that round. We assume $|\mathcal{S}_t| = K \ \forall t$. It then averages the parameters of all $j \in \mathcal{S}_t$ and broadcasts the new value to *all* nodes in the network. Due to the limited bandwidth, we assume partial participation, i.e. $K \ll n$. We assume that the server aggregation rounds happen every H local updates, and we call $\mathcal{H} = \{t : t \text{ modulo } H = 0\}$ the set of those times.

[1]Our analysis is readily extendable to the case where the server samples with non-uniform probabilities $\{p_i\}_{i=1}^n$, in which case the cost becomes $f(\mathbf{z})=\sum_{i=1}^n p_i F_i(\mathbf{z})$ and the term $\frac{\sigma^2}{n}$ in Theorem 1 becomes $\sum_{i=1}^n p_i^2\sigma_i^2$.

Algorithm 1 Peer-to-peer aided FL (FedDec)

Initialize $\mathbf{z}_i^1 = \mathbf{z}^1 \ \forall i \in [n]$ and let $\eta_t = \frac{2}{\mu(t+\gamma)}$.
for $t = 1, \dots, T$ all agents $i \in [n]$ **do**
  Sample mixing matrix $W^t \sim \mathcal{W}$
  Sample mini-batch ξ_i^t and compute $\nabla F_i(\mathbf{z}_i^t, \xi_i^t)$
  Update local parameter $\mathbf{x}_i^{t+\frac{1}{2}} = \mathbf{z}_i^t - \eta_t \nabla F_i(\mathbf{z}_i^t, \xi_i^t)$
  Average with neighbors $\mathbf{x}_i^{t+1} = \sum_{j=1}^n W_{ij}^t \mathbf{x}_j^{t+\frac{1}{2}}$
  If $t+1 \in \mathcal{H}$ **then**
    Server samples $\mathcal{S}_t = \{j_\ell : j_\ell \sim \mathcal{U}([n])\}_{\ell=1}^K$
    it computes $\mathbf{z}^{t+1} = \frac{1}{K}\sum_{\ell=1}^K \mathbf{x}_{j_\ell}^{t+1}$
    and broadcasts so that $\mathbf{z}_i^{t+1} = \mathbf{z}^{t+1} \ \forall i \in [n]$
  otherwise
    $\mathbf{z}_i^{t+1} = \mathbf{x}_i^{t+1}$
end for
Output $\mathbf{z}^{T+1}$

One local update of FedDec for a node i consists on (i) taking an SGD step, (ii) for all active links (i.e. $\forall j\colon W_{ij}^t > 0$), exchanging the new parameter value with its neighbors, and (iii) combining all new values (including its own) with weights W_{ij}^t to form the new iterate. We call this algorithm FedDec, and the precise steps are shown in Algorithm 1.

With slight abuse of notation, the stochastic gradient of a node i computed on a mini-batch $\xi_i = \{\psi_i^j : \psi_i^j \sim \mathcal{D}_i\}_{j=1}^m$ of size m is given by $\nabla F_i(\mathbf{z}_i, \xi_i) = \frac{1}{m}\sum_{j=1}^m \nabla F_i(\mathbf{z}_i, \psi_i^j)$. We will also take the following assumptions, which are standard in the literature [6], [8], [11], [14]–[16].

Assumption 1. *We assume the following $\forall F_i(\mathbf{z}), i \in [n]$:*

*1) **L-smoothness and μ-strong convexity:***

$$F_i(\mathbf{y}) \le F_i(\mathbf{x}) + \langle \nabla F_i(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + (L/2)\|\mathbf{y}-\mathbf{x}\|_2^2, \quad (1)$$

$$F_i(\mathbf{y}) \ge F_i(\mathbf{x}) + \langle \nabla F_i(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + (\mu/2)\|\mathbf{y}-\mathbf{x}\|_2^2. \quad (2)$$

*2) **Bounded variance of the local gradients:***

$$\mathbb{E}\|\nabla F_i(\mathbf{x}, \xi_i) - \nabla F_i(\mathbf{x})\|_2^2 \le \sigma_i^2.$$

*3) **Bounded energy of the local gradients:***

$$\mathbb{E}\|\nabla F_i(\mathbf{x}, \xi_i)\|_2^2 \le G^2 \text{ for } i \in [n]. \quad (3)$$

We remark that (2) implies $\|\nabla F_i(\mathbf{z}_i)\|_2 \ge \mu\|\mathbf{z}_i - \mathbf{z}_i^*\|_2$ (see definition of $\mathbf{z}_i^*$ below). Therefore, to satisfy (3) we must also assume that the parameter iterates $\mathbf{z}_i$ belong to a bounded set throughout the iterations.

Note that L-smoothness implies

$$\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|_2, \quad (4)$$

$$\|\nabla f(\mathbf{z})\|_2^2 = \|\nabla f(\mathbf{z}) - \nabla f(\mathbf{z}^*)\|_2^2 \le 2L(f(\mathbf{z}) - f(\mathbf{z}^*)), \quad (5)$$

and the local gradient's bounded variance implies

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\big(\nabla F_i(\mathbf{z}_i^t)-\nabla F_i(\mathbf{z}_i^t,\xi_i^t)\big)\right\|_2^2 \le \frac{\overline{\sigma}^2}{n},\quad \overline{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2. \tag{6}$$

We quantify the degree of heterogeneity through

$$\Gamma=\frac{1}{n}\sum_{i=1}^{n}(F_i(\mathbf{z}^*)-F_i(\mathbf{z}_i^*)),\ \text{where}\ \ \mathbf{z}_i^*=\operatorname*{argmin}_{\mathbf{z}} F_i(\mathbf{z}).$$

For each of FedDec's update parameters $\mathbf{z}_i$ and $\mathbf{x}_i$ we define

$$\overline{\mathbf{x}}^t=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^t,\qquad \overline{\mathbf{z}}^t=\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i^t,$$

which will be useful in the analysis. Note that $\overline{\mathbf{z}}^t=\overline{\mathbf{x}}^t$ only when $t\notin\mathcal{H}$. Otherwise, if $t\in\mathcal{H}$, we have that the equality holds only in expectation:

$$\mathbb{E}_{\mathcal{S}_t}\overline{\mathbf{z}}^t=\mathbb{E}_{\mathcal{S}_t}\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i^t=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{\ell=1}^{K}\mathbb{E}_{\mathcal{S}_t}\mathbf{x}_{j_\ell}^t=\overline{\mathbf{x}}^t. \tag{7}$$

Lastly, we assume the following about the $W^t=\{W_{ij}^t\}$.

Assumption 2. *The averaging matrices $W^t\in\mathbb{R}^{n\times n}$ are iid random variables drawn from a distribution $\mathcal{W}$ of matrices that (i) are symmetric, (ii) are doubly stochastic, and (iii) have $W_{ij}^t\ge 0$ if agents i and j are connected and $W_{ij}^t=0$ otherwise. Note that this implies that $\forall W\in\mathcal{W}: W\mathbf{1}=\mathbf{1}$, $\mathbf{1}^TW=\mathbf{1}^T$. Additionally, we require that the eigenvalues of $\mathbb{E}_W\big[WW^T\big]$ satisfy $1=\lambda_1>|\lambda_2|\ge\ldots\ge|\lambda_n|$.*

In the next section we prove that FedDec converges as $O(1/T)$, similarly to other FL algorithms taking the same assumptions, but it reduces the negative impact of local updates and partial device participation by replacing an H^2 factor [8], [11], with $H\alpha$, where α decreases quickly as the inter-agent communication network becomes more connected.

III. Convergence Analysis

The following theorem establishes the convergence rate of FedDec and constitutes our main result.

Theorem 1. *Under Assumptions 1, 2, and for diminishing stepsize $\eta_t=\frac{2}{\mu(\gamma+t)}$, FedDec in Algorithm 1 converges as*

$$\mathbb{E}[f(\overline{\mathbf{z}}^t)]-f(\mathbf{z}^*)\le\frac{L}{\gamma+t}\left(\frac{2B}{\mu^2}+\frac{(\gamma+1)}{2}\big\|\mathbf{z}^1-\mathbf{z}^*\big\|_2^2\right)$$

where

$$\gamma=\max\{8(L/\mu)-1,H\},$$
$$B=(4/K+8)\alpha HG^2+6L\Gamma+\overline{\sigma}^2/n,$$
$$\alpha=|\widehat{\lambda}_2|\,/\,(1-|\widehat{\lambda}_2|),\quad \textit{and}\quad |\widehat{\lambda}_2|=\Big|\lambda_2\big(\mathbb{E}_W\big[WW^T\big]\big)\Big|.$$

The theorem shows that factors like gradient variance and function heterogeneity slow down convergence, which is known. However, it also shows how inter-agent communication partially mitigates the negative impact of local updates and partial participation: the term where H and K appear decreases with α, and thus decreases very fast with $|\widehat{\lambda}_2|$.

If the inter-agent links are assumed to be always active, then $W^t=W$ is fixed and $|\widehat{\lambda}_2|=|\lambda_2|^2$. For any given heuristic to construct W (e.g. graph's Laplacian), the value of $|\lambda_2|$ is lower the more connected the network is. Therefore, the more dense is the network, the faster FedDec is expected to converge.

Comparing the bound of Theorem 1 with that of Theorem 2 in [8] (obtained for the same setting but without inter-agent communication) we note that the dependence of the first term in B on the number of local iterations H drops from $O(H^2)$ in [8] to $O(H)$ in our theorem. This suggests that *the peer-to-peer communication of FedDec reduces the impact of the infrequent communication rounds with the server*, and thus its convergence should be less affected than that of FedAvg as H increases. We verify this behavior in our simulations.

To prove the theorem, we will bound $\big\|\overline{\mathbf{z}}^t-\mathbf{z}^*\big\|_2^2$ by decomposing it into $\big\|\overline{\mathbf{z}}^t-\overline{\mathbf{x}}^t\big\|_2^2$ and $\big\|\overline{\mathbf{x}}^t-\mathbf{z}^*\big\|_2^2$ and bounding each separately. The following lemmas present these intermediate results, and we prove Theorem 1 at the end of the section. Due to space limitations, we refer the reader to the Arxiv version[2] or this work for the missing proofs. We have kept in the appendix the proof of Lemma 3, which provides the key steps to analyze FL with decentralized updates and shows how the connectivity of the network appears in the analysis.

Lemma 2. *For FedDec with stepsize $\eta_t\le\frac{1}{4L}$ it holds*

$$\mathbb{E}\big\|\overline{\mathbf{x}}^{t+1}-\mathbf{z}^*\big\|_2^2\le(1-\mu\eta_t)\mathbb{E}\big\|\overline{\mathbf{z}}^t-\mathbf{z}^*\big\|_2^2 + \frac{2}{n}\mathbb{E}\sum_{i=1}^{n}\big\|\mathbf{z}_i^t-\overline{\mathbf{z}}^t\big\|_2^2+6L\eta_t^2\Gamma+\eta_t^2\frac{\overline{\sigma}^2}{n}.$$

Lemma 2 bounds the one-step progress of the algorithm *before* a potential server aggregation round.

Lemma 3. *For stepsizes satisfying $\eta_t\le 2\eta^{t+H}$, it holds that*

$$\mathbb{E}\sum_{i=1}^{n}\big\|\mathbf{z}_i^t-\overline{\mathbf{z}}^t\big\|_2^2\le\eta_t^2 4\alpha HnG^2\quad\textit{with}\quad\alpha=\frac{|\widehat{\lambda}_2|}{1-|\widehat{\lambda}_2|}.$$

Lemma 3 bounds the divergence of the local parameters to their average, which increases through the H iterations. It is here where we see the impact of the graph connectivity.

As remarked in Section II, $\overline{\mathbf{z}}^t=\overline{\mathbf{x}}^t\,\forall\, t\notin\mathcal{H}$. Otherwise, the equality holds only in expectation (eq. (7)). Lemma 4 bounds the variance of $\overline{\mathbf{z}}$ in the latter case.

Lemma 4. *For $t\in\mathcal{H}$ and stepsizes satisfying $\eta_t\le 2\eta_{t+H}$*

$$\mathbb{E}\big\|\overline{\mathbf{x}}^t-\overline{\mathbf{z}}^t\big\|_2^2\le\frac{1}{K}\eta_t^2 4\alpha HG^2,$$

with α given in Lemma 3.

Lastly, we have the following lemma proved in [8].

Lemma 5. *Let a sequence Δ^t satisfy*

$$\Delta^{t+1}\le(1-\mu\eta_t)\Delta^t+\eta_t^2B \tag{8}$$

with $\mu,B>0$. Then, for diminishing stepsize $\eta_t=\frac{2}{\mu(\gamma+t)}$ with $\gamma>0$, it holds that $\Delta_t\le\frac{v}{\gamma+t}$, where $v=\max\{\frac{4B}{\mu^2},(\gamma+1)\Delta_1\}$.

[2]For the complete proofs refer to https://doi.org/10.48550/arXiv.2306.06715

We can now proceed to prove the main theorem.

Proof of Theorem 1. We start by noting that

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{z}}^{t+1}-\mathbf{z}^*\right\|_2^2 &= \mathbb{E}\left\|\overline{\mathbf{z}}^{t+1}-\overline{\mathbf{x}}^{t+1}+\overline{\mathbf{x}}^{t+1}-\mathbf{z}^*\right\|_2^2 \\ &= \mathbb{E}\left\|\overline{\mathbf{z}}^{t+1}-\overline{\mathbf{x}}^{t+1}\right\|_2^2+\mathbb{E}\left\|\overline{\mathbf{x}}^{t+1}-\mathbf{z}^*\right\|_2^2 \\ &+ 2\mathbb{E}\left\langle\overline{\mathbf{z}}^{t+1}-\overline{\mathbf{x}}^{t+1},\overline{\mathbf{x}}^{t+1}-\mathbf{z}^*\right\rangle.\end{aligned}$$

The last term is zero in expectation, since $\mathbb{E}_{\mathcal{S}_t}\overline{\mathbf{z}}^t=\overline{\mathbf{x}}^t$ (eq. (7)). The first term is zero when $t+1\notin\mathcal{H}$, and for all other iterations we can bound it using Lemma 4 (and $\eta_{t+1}^2<\eta_t^2$). We bound the second term using Lemma 2. We have then

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{z}}^{t+1}-\mathbf{z}^*\right\|_2^2 \leq \frac{1}{K}\eta_t^2 4\alpha HG^2+(1-\mu\eta_t)\mathbb{E}\left\|\overline{\mathbf{z}}^t-\mathbf{z}^*\right\|_2^2 \\ +\frac{2}{n}\mathbb{E}\sum_{i=1}^n\left\|\mathbf{z}_i^t-\overline{\mathbf{z}}^t\right\|_2^2+6L\eta_t^2\Gamma+\eta_t^2\frac{\overline{\sigma}^2}{n}.\end{aligned}$$

Using Lemma 3 to bound $\mathbb{E}\sum_{i=1}^n\left\|\mathbf{z}_i^t-\overline{\mathbf{z}}^t\right\|_2^2$ we get

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{z}}^{t+1}-\mathbf{z}^*\right\|_2^2 \leq (1-\mu\eta_t)\mathbb{E}\left\|\overline{\mathbf{z}}^t-\mathbf{z}^*\right\|_2^2 \\ +\eta_t^2\left[\left(\frac{4}{K}+8\right)\alpha HG^2+6L\Gamma+\frac{\overline{\sigma}^2}{n}\right].\end{aligned}$$

This has the form of (8) with $\Delta^t=\left\|\overline{\mathbf{z}}^t-\mathbf{z}^*\right\|_2^2$ and $B=\left(\frac{4}{K}+8\right)\alpha HG^2+6L\Gamma+\overline{\sigma}^2/n$, so applying Lemma 5:

$$\mathbb{E}\left\|\overline{\mathbf{z}}^t-\mathbf{z}^*\right\|_2^2\leq\frac{v}{\gamma+t}\leq\frac{1}{\gamma+t}\left(\frac{4B}{\mu^2}+(\gamma+1)\Delta_1\right). \quad (9)$$

Note that in order to ensure $\eta_t\leq\frac{1}{4L}$ (Lemma 2) and $\eta_t\leq 2\eta_{t+H}$ (Lemmas 3 and 4) we need to set $\gamma=\max\{8\frac{L}{\mu}-1,H\}$. Finally, using L-smoothness and $\nabla f(\mathbf{z}^*)=0$,

$$\mathbb{E}[f(\overline{\mathbf{z}}^t)]-f(\mathbf{z}^*)\leq\frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{z}}^t-\mathbf{z}^*\right\|_2^2.$$

Using (9) in the inequality above gives the result. □

IV. Numerical Results

In this section we compare the performance of FedDec with that of FedAvg [1]. We consider the linear regression

$$F_i(\mathbf{z})=\frac{1}{M}\|X_i\,\mathbf{z}-Y_i\|_2^2,\qquad i\in[n],$$

with $X_i,Y_i\in\mathbb{R}^{M\times d}$, $M=10$, and $d=25$. To generate the data we follow [5]: we set $[X_i]_j\sim\mathcal{N}(0,0.25^2)$, $j\in[d]$ and $Y_i=c_i(v+\cos(v))$, where $v=X_i\mathbf{1}$ and $c_i=2^i, i\in[n]$, which makes the data significantly different between nodes.

For the inter-agent links, we generate geographic graphs of $n=20$ nodes by distributing points uniformly at random in a 1×1 square and linking all pairs of points whose L2 distance is smaller than a radius r. We test our algorithms in two graphs with $r=0.35$ and 0.5, respectively. We run $T=5000$ iterations with $K=2$, $m=1$, and $H=10,100$.

Figure 2 shows the convergence of FedDec and FedAvg for each graph and the two values of H. The stepsize was set according to Theorem 1. The lines shown are the average of ten independent runs on the same problem instance.

Comparing the plots in Fig. 2 vertically (i.e., for the same H), we confirm that higher connectivity (i.e. smaller $|\lambda_2|$) leads to larger gains of FedDec over FedAvg. We remark, however, that $|\lambda_2|$ seems to predict the convergence speed of decentralized algorithms only when the nodes have sufficiently different data [20], as in our simulations.

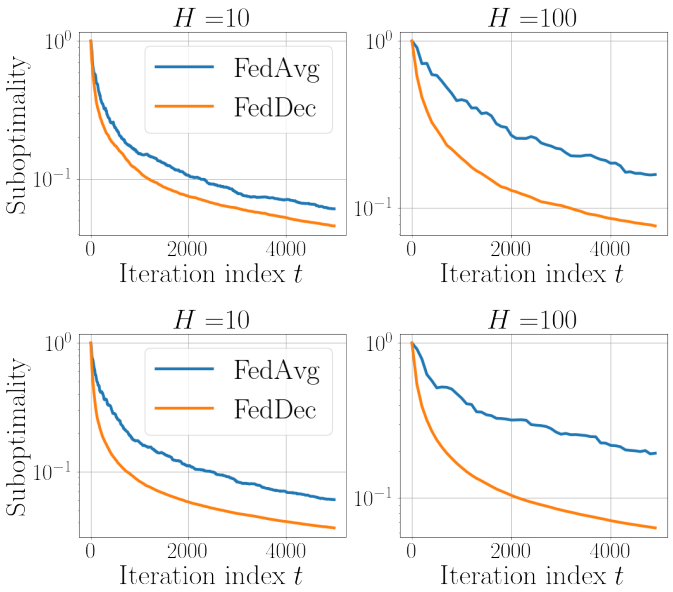

Fig. 2: FedDec versus FedAvg for a sparse graph (top) and a dense graph (bottom).

Comparing the plots in Fig. 2 horizontally (i.e., for a given graph), we verify that as H increases the convergence speed of FedAvg decreases more than that of FedDec. Therefore, FedDec allows for sparser server communication rounds without significantly sacrificing convergence speed, which is in accordance with Theorem 1.

V. Conclusion

We presented FedDec, an FL algorithm where the agents average their parameters with those of their neighbors before each local SGD update. We proved that this modification reduces the negative impact of local updates and partial device participation on convergence, and that it can make them negligible for sufficiently connected networks.

This insight suggests that there exists a connectivity threshold where the server does not help convergence anymore. Furthermore, we conjecture that for sufficiently dense networks, server communication rounds might even *hurt* convergence. Future directions include studying this threshold and other trade-offs of peer-to-peer aided FL.

Overall, exploiting inter-agent communication in FL is a promising way to reduce the amount of communication with the server without compromising convergence speed.

Appendix

Proof of Lemma 3. We define the arrays in $\mathbb{R}^{d\times n}$:

$$Z^t:=\left[\mathbf{z}_1^t\cdots\mathbf{z}_n^t\right],\quad \overline{Z}^t:=Z^t\frac{1}{n}\mathbf{1}\mathbf{1}^T=\left[\overline{\mathbf{z}}^t\cdots\overline{\mathbf{z}}^t\right]$$
$$\partial F(Z^t,\xi^t):=\left[\nabla F_1(\mathbf{z}_1^t,\xi_1^t)\cdots\nabla F_n(\mathbf{z}_n^t,\xi_n^t)\right]$$

and $X^t, \overline{X}^t$ similarly. Note that $\sum_{i=1}^n \left\|\mathbf{z}_i^t - \overline{\mathbf{z}}^t\right\|_2^2 = \left\|Z^t - \overline{Z}^t\right\|_F^2$.

We denote $t_b \in \mathcal{H}$ the last time the central server broadcasted the sample average to all nodes so that $\mathbf{z}_i^{t_b} = \mathbf{z}^{t_b}\ \forall i \in [n]$, and define $h := t - t_b \leq (H-1)$. Note that if $t = t_b$ then $\sum_{i=1}^n \|\mathbf{z}_i^t - \overline{\mathbf{z}}\|_2^2 = 0$. Thus, below we assume $h \geq 1$. We have

$$\begin{aligned}
&\mathbb{E}\left\|Z^t - \overline{Z}^t\right\|_F^2 \overset{t\notin\mathcal{H}}{=} \mathbb{E}\left\|X^t - \overline{X}^t\right\|_F^2 \\
&= \mathbb{E}\Big\| \underbrace{(X^{t-\frac{1}{2}} - \overline{X}^{t-\frac{1}{2}})}_{Y} W^{t-1}\Big\|_F^2 \\
&= \mathbb{E}\sum_{i=1}^d \left\|Y_{[i,:]}W^{t-1}\right\|_2^2 = \mathbb{E}\sum_{i=1}^d Y_{[i,:]}\mathbb{E}_W\left[WW^T\right][Y_{[i,:]}]^T \\
&\leq \mathbb{E}\sum_{i=1}^d Y_{[i,:]}\left|\lambda_2\big(\mathbb{E}_W\left[WW^T\right]\big)\right|[Y_{[i,:]}]^T
\end{aligned}$$

where in the first line we used that $\mathbf{1}^T W^t = \mathbf{1}^T$, and in the third line, that the matrices W^t are identically distributed and independent of the time t. Notation $Y_{[i,:]}$ indicates i-th row of matrix Y. Denoting $|\widehat{\lambda}_2| := \left|\lambda_2\big(\mathbb{E}_W\left[WW^T\right]\big)\right|$ we have

$$\begin{aligned}
&\mathbb{E}\left\|Z^t - \overline{Z}^t\right\|_F^2 \leq |\widehat{\lambda}_2|\ \mathbb{E}\left\|X^{t-\frac{1}{2}} - \overline{X}^{t-\frac{1}{2}}\right\|_F^2 \\
&= |\widehat{\lambda}_2|\ \mathbb{E}\left\|Z^{t-1} - \overline{Z}^{t-1} - \eta_{t-1}\partial F(Z^{t-1},\xi^{t-1})(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\right\|_F^2 \\
&\leq |\widehat{\lambda}_2|\left(1+\frac{1}{\alpha}\right)\mathbb{E}\left\|Z^{t-1} - \overline{Z}^{t-1}\right\|_F^2 \\
&\quad + |\widehat{\lambda}_2|(1+\alpha)\eta_{t-1}^2\mathbb{E}\left\|\partial F(Z^{t-1},\xi^{t-1})(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\right\|_F^2 \\
&\leq |\widehat{\lambda}_2|\left(1+\frac{1}{\alpha}\right)\mathbb{E}\left\|Z^{t-1} - \overline{Z}^{t-1}\right\|_F^2 + \\
&\quad + |\widehat{\lambda}_2|\ (1+\alpha)\eta_{t-1}^2\mathbb{E}\left\|\partial F(Z^{t-1},\xi^{t-1})\right\|_F^2 \\
&= |\widehat{\lambda}_2|\left(1+\frac{1}{\alpha}\right)\mathbb{E}\left\|Z^{t-1} - \overline{Z}^{t-1}\right\|_F^2 + |\widehat{\lambda}_2|(1+\alpha)\eta_{t-1}^2 nG^2,
\end{aligned}$$

We can now apply the inequality recursively to get

$$\begin{aligned}
&\mathbb{E}\left\|Z^t - \overline{Z}^t\right\|_F^2 \leq \left(1+\frac{1}{\alpha}\right)|\widehat{\lambda}_2|\ \mathbb{E}\left\|Z^{t-1} - \overline{Z}^{t-1}\right\|_F^2 \\
&+ (1+\alpha)|\widehat{\lambda}_2|\ \eta_{t-1}^2 nG^2 \ \leq\ \left[\left(1+\frac{1}{\alpha}\right)|\widehat{\lambda}_2|\ \right]^h \mathbb{E}\left\|Z^{t_b} - \overline{Z}^{t_b}\right\|_F^2 \\
&\quad + (1+\alpha)|\widehat{\lambda}_2|\ nG^2\sum_{i=1}^h\left[\left(1+\frac{1}{\alpha}\right)|\widehat{\lambda}_2|\ \right]^{i-1}\eta_{t-i}^2.
\end{aligned}$$

We note that the first term is zero, since at broadcasting time $Z^{t_b} = \overline{Z}^{t_b}$. We now set $\alpha = \frac{|\widehat{\lambda}_2|}{1-|\widehat{\lambda}_2|}$ so that the expression between square brackets takes value 1. Therefore,

$$\mathbb{E}\left\|Z^t - \overline{Z}^t\right\|_F^2 \ \leq\ \alpha nG^2 H\eta_{t_b}^2 \ \leq\ \alpha nG^2 H 4\eta_t^2,$$

where we have used that for the choice of α given above it holds $(1+\alpha)|\widehat{\lambda}_2| \ = \alpha$, and that the stepsizes η_t are monotonically decreasing and satisfy $\eta_t \leq 2\eta_{t+H}$. □

REFERENCES

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

[2] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*, pp. 5132–5143, PMLR, 2020.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[4] 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.

[5] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *international conference on machine learning*, pp. 3027–3036, PMLR, 2017.

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

[7] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "A dual approach for optimal algorithms in distributed optimization over networks," in *2020 Information Theory and Applications Workshop (ITA)*, pp. 1–37, IEEE, 2020.

[8] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-iid data," in *International Conference on Learning Representations*, 2020.

[9] J. Zhang, C. De Sa, I. Mitliagkas, and C. Ré, "Parallel SGD: When does averaging help?," *arXiv preprint arXiv:1606.07365*, 2016.

[10] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[11] S. U. Stich, "Local SGD converges fast and communicates little," in *ICLR 2019-International Conference on Learning Representations*, no. CONF, 2019.

[12] S. Hosseinalipour, S. S. Azam, C. G. Brinton, N. Michelusi, V. Aggarwal, D. J. Love, and H. Dai, "Multi-stage hybrid federated learning over large-scale d2d-enabled fog networks," *IEEE/ACM Transactions on Networking*, vol. 30, no. 4, pp. 1569–1584, 2022.

[13] 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*, pp. 5381–5393, PMLR, 2020.

[14] M. Yemini, R. Saha, E. Ozfatura, D. Gündüz, and A. J. Goldsmith, "Semi-decentralized federated learning with collaborative relaying," in *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 1471–1476, IEEE, 2022.

[15] F. P.-C. Lin, S. Hosseinalipour, S. S. Azam, C. G. Brinton, and N. Michelusi, "Semi-decentralized federated learning with cooperative D2D local model aggregations," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3851–3869, 2021.

[16] L. Chou, Z. Liu, Z. Wang, and A. Shrivastava, "Efficient and less centralized federated learning," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, pp. 772–787, Springer, 2021.

[17] R. Parasnis, S. Hosseinalipour, Y.-W. Chu, M. Chiang, and C. G. Brinton, "Connectivity-aware semi-decentralized federated learning over time-varying d2d networks," in *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 31–40, 2023.

[18] Y. Chen, K. Yuan, Y. Zhang, P. Pan, Y. Xu, and W. Yin, "Accelerating gossip SGD with periodic global averaging," in *International Conference on Machine Learning*, pp. 1791–1802, PMLR, 2021.

[19] Y. Guo, Y. Sun, R. Hu, and Y. Gong, "Hybrid local SGD for federated learning with heterogeneous communications," in *International Conference on Learning Representations*, 2022.

[20] G. Neglia, C. Xu, D. Towsley, and G. Calbi, "Decentralized gradient methods: does topology matter?," in *International Conference on Artificial Intelligence and Statistics*, pp. 2348–2358, PMLR, 2020.