Semi-Decentralized Federated Learning with Collaborative Relaying

Michal Yemini Princeton University
Rajarshi Saha Stanford University
Emre Ozfatura Imperial College London
Deniz Gündüz Imperial College London
Andrea J. Goldsmith Princeton University

***Abstract*—We present a semi-decentralized federated learning algorithm wherein clients collaborate by relaying their neighbors' local updates to a central parameter server (PS). At every communication round to the PS, each client computes a local consensus of the updates from its neighboring clients and eventually transmits a weighted average of its own update and those of its neighbors to the PS. We appropriately optimize these averaging weights to ensure that the global update at the PS is unbiased and to reduce the variance of the global update at the PS, consequently improving the rate of convergence. Numerical simulations substantiate our theoretical claims and demonstrate settings with intermittent connectivity between the clients and the PS, where our proposed algorithm shows an improved convergence rate and accuracy in comparison with the federated averaging algorithm.**

I. Introduction

Federated learning (FL) algorithms iteratively optimize a common objective function to learn a shared model over data samples that are localized over multiple distributed clients [1]. FL approaches aim to reduce the required communication overhead and improve clients' privacy by training local models of private dataset at the clients and forwarding them periodically to a centralized parameter server (PS). In practical FL setups, some clients are stragglers and cannot send their updates regularly, either because: *(i)* they cannot finish their computation within a prescribed deadline, or *(ii)* due to communication limitations [2], where they suffer from intermittent connectivity to the PS since their wireless channel is temporarily blocked [3]–[8]. Stragglers deteriorate the convergence of FL as the computed local updates become stale. This can even result in bias in the final model in the case of persistent stragglers. On the other hand, *Communication stragglers* (type *(ii)*) are inherently different from *computation-limited stragglers* (type *(i)*), since it can be solved by relaying the updates to the PS via neighboring clients.

Communication quality at the wireless edge as a key design principle is considered in the federated edge learning (FEEL) framework [9], which takes into account the wireless channel characteristics from the clients to the PS to optimize the convergence and final model performance at the PS. So far the FEEL paradigm has mainly focused on direct communication from the clients to the PS, and aimed at improving the performance by resource allocation across clients [9]–[18]; these approaches have ignored possible cooperation between clients in the case of intermittent communication blockages.

Motivated by our prior works [19]–[21], where client cooperation is used to improve the connectivity to the cloud and to reduce the latency and scheduling overhead, this work proposes a new FEEL paradigm, where the clients cooperate to mitigate the detrimental effects of communication stragglers. In our proposed method, clients share their local updates with neighbors so that each client sends to the PS a weighted average of its current update and those of its neighbors. Using this approach, the PS receives new updates from disconnected clients, which would otherwise become stale and be discarded. We demonstrate the success of our relaying scheme through both theoretical analysis and numerical simulations.

Related Works: Decentralized collaborative learning frameworks have been introduced as an alternative to centralized FL, in which the PS is removed to mitigate a potential communication bottleneck and a single point of failure [22]–[33]. In decentralized learning, each client shares its local model with its neighbors through device-to-device (D2D) communications, and model aggregation is executed at each client in parallel. This aggregation strategy is determined at each client according to the network topology, i.e., the connection pattern between the clients.

An alternative approach to both centralized and decentralized schemes is *hierarchical FL (HFL)* [21], [34]–[36], where multiple PSs are employed for the aggregation to prevent a communication bottleneck. In HFL, clients are divided into clusters and a PS is assigned to each cluster to perform local aggregation. The aggregated models at the clusters are later aggregated at the main PS in a subsequent step to obtain the global model. HFL has significant advantages over centralized and decentralized schemes, particularly when the communication takes place over wireless channels since it allows spatial reuse of available resources [21]. Nonetheless, HFL requires employing multiple PSs that may not be practical in certain scenarios. Instead, the idea of hierarchical collaborative learning can be redesigned to combine hierarchical and decentralized learning, which is referred as *semi-decentralized FL*, where the local consensus follows decentralized learning with D2D communications, whereas the global consensus is orchestrated by the PS [37], [38]. One of the major challenges in FL that is not considered in [37], [38] is the partial client connectivity [39], [40]. Unequal client participation due

M. Yemini, R. Saha, and A.J. Goldsmith are partially supported by the AFOSR award #002484665 and a Huawei Intelligent Spectrum grant.
E. Ozfatura and D. Gündüz received funding from the European Research Council (ERC) through Starting Grant BEACON (no. 677854) and the UK EPSRC (grant no. EP/T023600/1) under the CHIST-ERA program.

to intermittent connectivity exacerbates the impact of data heterogeneity [41]–[44], and increases the generalization gap.

Most existing works on FL assume error-free rate-limited orthogonal communication links, with an underlying communication protocol that takes care of wireless imperfections. However, this separation between the communication medium and the learning protocol can be suboptimal [9]. An alternative approach treats the communication of the model updates to the PS as an uplink communication problem and jointly optimizes the learning algorithm and the communication scheme [9]. Within this framework is an original and promising approach known as *over-the-air computation (OAC)* [14]–[16], which exploits the superposition property of wireless signals to convey the sum of the model updates that are transmitted by each client in an uncoded fashion. In addition to bandwidth efficiency, the OAC framework provides a certain level of anonymity to clients due to its superposition nature; hence, it can enhance the privacy of the participating clients [17], [18]. We emphasize that in OAC, PS receives the aggregate model, and it is not possible to disentangle the individual model updates. Therefore, any strategy that utilizes a PS side aggregation mechanism with individual model updates to address unequal client participation is not compatible with the OAC framework. One of the major advantages of our proposed scheme is that it mitigates the drawbacks of unequal client participation without requiring the identity of transmitting clients or their individual updates at the PS. Therefore, our solution is compatible with OAC.

Client connectivity is a particularly significant challenge in FEEL, where the clients and the PS communicate over unreliable wireless channels. Due to their different physical environments and distances to the PS, clients can have different connectivity to each other and the PS. This problem has been recently addressed in [10]–[13], [45]–[48] by considering customized client selection mechanisms to balance the participation of the clients and the latency for the model aggregation in order to speed up the learning process. We adopt a different approach to this problem, where instead of designing a client selection mechanism, or optimizing resource allocation to balance client participation, we introduce a *relaying* mechanism that takes into account the nature of individual clients' connectivity to the PS and ensures that, in case of poor connectivity, their local updates are conveyed to the PS with the help of their neighboring clients.

Paper Organization: Sec. II presents the FL system model and the proposed FL collaborative relaying scheme. Sec. III presents conditions for the unbiasedness of our proposed scheme and an analysis of the convergence rate. Sec. IV optimizes the convergence rate of our proposed scheme, while Sec. V presents numerical results that validate our theoretical analysis and highlight the performance improvement in terms of training accuracy. Finally, Sec. VI concludes this paper.

Remark. *Due to space limitations, all proofs are omitted, and can be found in an online extended version of this paper [49].*

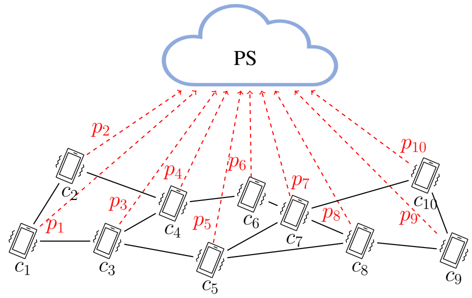

Fig. 1: System model with intermittent uplink communication between clients and PS (dotted lines) and reliable communication between neighboring clients (solid lines).

II. System Model for Collaborative Relaying

Consider n clients communicating periodically with a PS that trains a global model $\mathbf{x} \in \mathbb{R}^d$. Let $\mathcal{L}(\mathbf{x}, \boldsymbol{\zeta})$ be the loss evaluated for a model $\mathbf{x}$ at data point $\boldsymbol{\zeta}$. Denote the local loss at client i by $f_i : \mathbb{R}^d \times \mathcal{Z}_i \to \mathbb{R}$, where $f(\mathbf{x}; \mathcal{Z}_i) = \frac{1}{|\mathcal{Z}_i|} \sum_{\boldsymbol{\zeta} \in \mathcal{Z}_i} \mathcal{L}(\mathbf{x}; \boldsymbol{\zeta})$. Here, $\mathcal{Z}_i$ is the local dataset of client i. The goal of PS is to solve the following empirical risk minimization (ERM) problem:[1]

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{arg min.}}\, f(\mathbf{x}) \triangleq \underset{\mathbf{x}\in\mathbb{R}^d}{\text{arg min.}}\, \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}; \mathcal{Z}_i).$$

A. FL with Local SGD at Clients

Denote the local gradient as $\nabla f_i(\mathbf{x}) \triangleq \nabla_{\mathbf{x}} f(\mathbf{x}; \mathcal{Z}_i)$, and let $g_i(\mathbf{x})$ be an unbiased estimate of it. In the r^{th} *round* of FL, the PS broadcasts the global model $\mathbf{x}^{(r)}$ to the clients. For a local averaging *period* of $\mathcal{T}$, each client performs $\mathcal{T}$ iterations of local training, after which the local models are sent to the PS for aggregation. For local iteration $k \in [0 : \mathcal{T}]$ of the r^{th} round, client i applies the local update rule:

$$\mathbf{x}_i^{(r,k+1)} = \mathbf{x}_i^{(r,k)} - \eta_r \mathbf{g}_i\left(\mathbf{x}_i^{(r,k)}\right), \tag{1}$$

where η_r is the learning rate for round r and $\mathbf{x}_i^{(r,0)} = \mathbf{x}^{(r)}$.

B. Communication Model

Communication between clients and PS. We consider a setting where the uplink connections between the clients and the PS are intermittent. As shown in Fig. 1, we model the connectivity of client i to the PS at round r by the Bernoulli random variable $\tau_i(r) \sim \text{Bern}(p_i)$, where $\tau_i = 1$ indicates the presence of an uplink communication opportunity, whereas $\tau_i(r) = 0$ indicates a blocked uplink. For simplicity of exposition, we consider the uplink channel to be either completely blocked or perfectly available without any noise, and the downlink from PS to clients does not suffer from intermittent dropouts.

Remark 1. *The connectivity probabilities $\{p_i\}_{i\in[n]}$ can be easily estimated using pilot signals. Moreover, clients can share their p_i with each other using local links in a pre-training phase. On the other hand, we do not assume that*

[1] For simplicity, we assume $|\mathcal{Z}_i| = |\mathcal{Z}_j|$ for all $i, j \in [n]$. Our method can be extended to the setting of imbalanced local dataset sizes as well.

Algorithm 1 COLREL-CLIENT: Collaborative Relaying

```
Input: Round index $r$, Step-size $\eta_r$, Local avg. period $\mathcal{T}$, Neighborhood of client $i$ $\mathcal{N}_i$, $\alpha_{ij}$ for every $j \in \mathcal{N}_i \cup \{i\}$.
Output: $\Delta\widetilde{\mathbf{x}}_i^{r+1}$.
Receive $\mathbf{x}^{(r)}$ from PS.
Set $\mathbf{x}_i^{(r,0)} = \mathbf{x}^{(r)}$.
for $k \leftarrow 0$ to $\mathcal{T} - 1$ do
    Compute (stochastic) gradient $g_i(\mathbf{x}_i^{(r,k)}t)$.
    $\mathbf{x}_i^{(r,k+1)} = \mathbf{x}_i^{(r,k)} - \eta_r g_i\left(\mathbf{x}_i^{(r,k)}\right)$.
end for
Set $\Delta\mathbf{x}_i^{r+1} = \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}$.
Send $\Delta\mathbf{x}_i$ to every $j \in \mathcal{N}_i$.
Receive $\Delta\mathbf{x}_j$ from every $j \in \mathcal{N}_i$.
Compute $\Delta\widetilde{\mathbf{x}}_i^{r+1} = \sum_{j\in\mathcal{N}_i\cup\{i\}} \alpha_{ij} \cdot \Delta\mathbf{x}_j^{r+1}$.
Transmit (relay) $\Delta\widetilde{\mathbf{x}}_i^{r+1}$ to the PS.
```

the instantaneous connectivity information, i.e., $\tau_i(r), r \in [n]$ is available to any of the clients.

Communication between clients. The connectivity between clients is modeled by an undirected graph $G = (V, E)$ where $V = [n]$ and $(i, j) \in E \iff$ client i can communicate with client j. Let $\mathcal{N}_i \triangleq \{j \in V : \{i, j\} \in E\}$. We do not assume that the graph G is connected. Instead, it can be composed of multiple connected subgraphs.

C. Collaborative Relaying of Local Updates

Let $\Delta\mathbf{x}_i^{(r+1)}$ denote client i's update at the end of $\mathcal{T}^{th}$ local iteration of round r, i.e., $\Delta\mathbf{x}_j^{(r+1)} = \mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)}$. We assume that client i's update $\Delta\mathbf{x}_j^{(r+1)}$ is readily available to its neighbors. Then client i computes a weighted average of its own update and those of its neighbors in $\mathcal{N}_i$, i.e.,

$$\Delta\widetilde{\mathbf{x}}_i^{(r+1)} = \sum_{j\in\mathcal{N}_i\cup\{i\}} \alpha_{ij}\Delta\mathbf{x}_j^{(r+1)} = \sum_{j\in\mathcal{N}_i\cup\{i\}} \alpha_{ij}\left(\mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right),$$

where α_{ij} is a non-negative importance weight assigned by client i while relaying the client j's update. Note that weighted averaging entails a complexity of $O\left(\max_{i\in[n]} |\mathcal{N}_i| + 1\right)$.

D. PS Aggregation

In our setting, the PS does not explicitly select the subset of clients from which it wants to receive information, rather it receives updates from all *communicating* clients at the beginning of every round. The PS uses the following re-scaled sum of received updates:

$$\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + w \sum_{i\in[n]} \tau_i(r+1)\Delta\widetilde{\mathbf{x}}_i^{(r+1)}. \qquad (2)$$

This update can be computed over-the-air and does not require the PS to know the identities of the communicating clients. We set $w{=}1/n$ to preserve the unbiasedness of the objective function at the PS, as discussed in Sec. III. Our Collaborative Relaying (ColRel) algorithm is presented in Algs. 1 and 2.

Algorithm 2 COLREL-PS: PS Aggregation

```
Input: Number of rounds $R$, a set of clients $[n]$.
Output: Global model $\mathbf{x}^{(R)}$.
Set $\mathbf{x}^{(0)} = \mathbf{0}$
for $k \leftarrow 0$ to $\mathcal{T} - 1$ do
    Broadcast $\mathbf{x}^{(r)}$ to all clients.
    Set $\tau_i(r+1) = 1$ or 0 depending on connectivity.
    Update $\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \frac{1}{n}\sum_{i\in[n]} \tau_i(r+1)\Delta\widetilde{\mathbf{x}}_i^{r+1}$
end for
```

III. CONVERGENCE ANALYSIS

A. Unbiasedness of COLREL *FL*

In our collaborative relaying scheme, the local update of a particular client i can be transmitted to the PS by itself, or by one or more of its neighbors $j \in \mathcal{N}_i$. Since the PS may be blind to the identities of the clients, the clients collaborate among themselves to ensure that this redundancy is mitigated. This is done by appropriately choosing the weights α_{ij}. In particular, Lemma 1 gives a sufficient condition on the values of $\{\alpha_{ij}\}_{i,j\in[n]}$ that ensures that the aggregated global update at the PS is an unbiased estimate of $\frac{1}{n}\sum_{i\in[n]} \Delta\mathbf{x}_i^{(r)}$, the true aggregate in the case of perfect channel connectivity.

Lemma 1. *Let $w = 1/n$ and $\{\alpha_{ij}\}_{i,j\in[n]}$ be such that*

$$\mathbb{E}\left[\sum_{j\in\mathcal{N}_i\cup\{i\}} \tau_j(r+1)\alpha_{ji}\right] = p_i\alpha_{ii} + \sum_{j\in\mathcal{N}_i} p_j\alpha_{ji} = 1. \qquad (3)$$

Then, for every $i \in [n]$,

$$w \cdot \mathbb{E}\left[\sum_{j\in\mathcal{N}_i\cup\{i\}} \tau_j(r+1)\alpha_{ji}\Delta\mathbf{x}_i^{r+1}\middle|\Delta\mathbf{x}_i^{r+1}\right] = \frac{1}{n}\Delta\mathbf{x}_i^{r+1}.$$

Note that the standard FL setting under intermittent client connectivity to the PS but with no collaboration between the clients is captured by the choice $w = 1/n$, $\mathcal{N}_i = \emptyset, p_i = p, \alpha_{ii} = 1, \alpha_{ij} = 0$ for all $i, j \in [n]$ and $j \neq i$.

B. Expected Suboptimality Gap

Next, Thm. 1 presents the convergence rate of COLREL as a function of $\{\alpha_{ij}\}$, under the following assumptions.

Assumption 1. *For any i, the local loss f_i is L-smooth w.r.t. $\mathbf{x}$, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$.*

Assumption 2. *The stochastic gradients $\mathbf{g}_i(\mathbf{x})$ are unbiased and have bounded variance, i.e., $\forall\, i \in [n]$:*
1) $\mathbb{E}[\mathbf{g}_i(\mathbf{x})] = \nabla f_i(\mathbf{x})$, and
2) $\mathbb{E}\|\mathbf{g}_i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|_2^2 \leq \sigma^2$ for some finite σ^2.

Assumption 3. *For any i, the loss f_i is μ-strongly convex, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $(\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}))^\top(\mathbf{x} - \mathbf{y}) \geq \mu\|\mathbf{x} - \mathbf{y}\|_2^2$.*

Algorithm 3 OPT-α: Optimization of relay weight matrix $\boldsymbol{A}$

Input: Connectivity graph G, Transmission probability vector $\mathbf{p}$, Maximum number of iterations L.

Output: Relay weight matrix $\mathbf{A}^{(L)}$ that approximately solves (6).

```
1: Set $\mathbf{A}_{ji}^{(0)} = \frac{1}{(|\mathcal{N}_i|+1)\cdot p_j} \cdot \mathbb{1}_{\{j \in \mathcal{N}_i \cup \{i\} : p_j > 0\}}$.
2: for $\ell \leftarrow 0$ to $L-1$ do
       Set $\ell \leftarrow \ell + 1$.
       Set $i = \ell \mod n + n \cdot \mathbb{1}_{\{\ell \mod n = 0\}}$.
       Compute $\widehat{\mathbf{A}}_i^{(\ell)}$ according to (9).
       Set $\mathbf{A}_k^{(\ell)}$ according to (7) for every $k \in [n]$.
3: end for
```

Let $\mathbf{A} = (\alpha_{ij})_{i,j \in [n]}$ denote the $n \times n$ matrix of relay weights, and let $\mathcal{N}_{il} = (\mathcal{N}_i \cup \{i\}) \cap (\mathcal{N}_l \cup \{l\})$ denote the common neighborhood of nodes i and j. Suppose,

$$S(\boldsymbol{p}, \mathbf{A}) = \sum_{i,l \in [n]} \sum_{j : j \in \mathcal{N}_{il}} p_j(1-p_j)\alpha_{ji}\alpha_{jl}. \tag{4}$$

Theorem 1. *Under Asms. 1-3 and condition* (3), ColRel, *as specified by Algs. 1 and 2, with* $\eta_r = \frac{4\mu^{-1}}{r\mathcal{T}+1}$, *satisfies for every* $r \geq r_0(\boldsymbol{p}, \mathbf{A})$,

$$\mathbb{E}\|\mathbf{x}^{(r+1)} - x^\star\|^2 \leq \frac{(r_0\mathcal{T}+1)}{(r\mathcal{T}+1)^2}\|\mathbf{x}^{(0)} - x^\star\|^2 + \frac{C_1(\boldsymbol{p}, \mathbf{A})\mathcal{T}}{k\mathcal{T}+1} + C_2\frac{(\mathcal{T}-1)^2}{k\mathcal{T}+1} + C_3(\boldsymbol{p}, \mathbf{A})\frac{\mathcal{T}}{(k\mathcal{T}+1)^2},$$

where $B(\boldsymbol{p}, \mathbf{A}) = \frac{2L^2}{n^2}S(\boldsymbol{p}, \mathbf{A})$, $C_1(\boldsymbol{p}, \mathbf{A}) = \frac{4^2}{\mu^2} \cdot \frac{2\sigma^2}{n^2}S(\boldsymbol{p}, \mathbf{A})$, $C_2 = \frac{4^2}{\mu^2} \cdot L^2\frac{\sigma^2}{n}e$, $C_3(\boldsymbol{p}, \mathbf{A}) = \frac{4^4}{\mu^4} \cdot \left(L^2\sigma^2 e + \frac{2L^2\sigma^2 e}{n^2}S(\boldsymbol{p}, \mathbf{A})\right)$, *and* $r_0(\boldsymbol{p}, \mathbf{A}) = \max\left\{\frac{L}{\mu}, 4\left(\frac{B(\boldsymbol{p},\mathbf{A})}{\mu^2} + 1\right), \frac{1}{\mathcal{T}}, \frac{4n}{\mu^2\mathcal{T}}\right\}$.

As a consequence of Thm. 1, it follows that,

$$\mathbb{E}\left\|\mathbf{x}^{(r+1)} - x^\star\right\|^2 = O\left(\frac{\left\|\mathbf{x}^{(0)} - x^\star\right\|^2}{r^2} + \frac{S(\boldsymbol{p}, \mathbf{A})}{r}\right). \tag{5}$$

Therefore, the convergence rate can be improved by minimizing the term $S(\mathbf{p}, \mathbf{A})$ subject to the unbiasedness condition in Lemma 1. Minimizing $S(\boldsymbol{p}, \mathbf{A})$ can also reduce $r_0(\boldsymbol{p}, \mathbf{A})$.

IV. Optimizing the Relaying Weights

We choose the relay weight matrix $\mathbf{A}$ to minimize the upper bound on the expected distance to optimality as given by Thm. 1. Thus, we solve the following optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{A}} S(\boldsymbol{p}, \mathbf{A}), \\ &\text{s.t.: } \sum_{j : j \in \mathcal{N}_i} p_j\alpha_{ji} = 1, \quad \alpha_{ji} \geq 0 \quad \forall i, j \in [n]. \end{aligned} \tag{6}$$

The function $S(\mathbf{p}, \mathbf{A})$ is convex with respect to (w.r.t.) $\mathbf{A}$ for $\mathbf{p} \in [0,1]^n$. It can be shown that the domain of (6) is separable w.r.t. $\mathbf{A}_i$, the i^{th} column of $\mathbf{A}$, and we can use the Gauss-Seidel method [50, Prop. 2.7.1] to iteratively solve (6). At every iteration ℓ, we compute the estimate $\mathbf{A}^\ell$ as

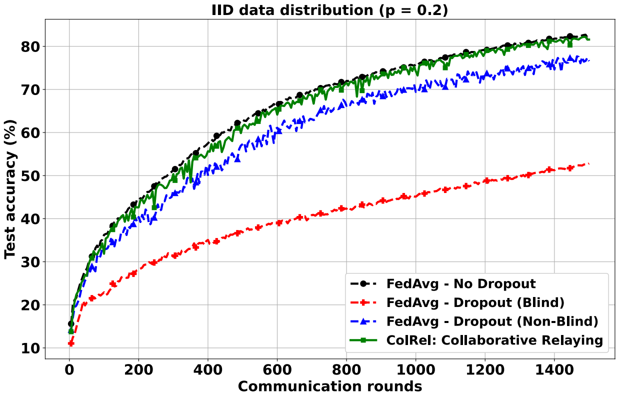

Fig. 2: Homogeneous connectivity with $p_i = 0.2, \forall i \in [n]$ and FCT.

$$\mathbf{A}_i^{(\ell)} = \begin{cases} \widehat{\mathbf{A}}_i^{(\ell)} & \text{if } \ell \bmod n + n \cdot \mathbb{1}_{\{\ell \bmod n = 0\}} = i, \\ \mathbf{A}_i^{(\ell-1)} & \text{otherwise} \end{cases}. \tag{7}$$

Here, $\widehat{\mathbf{A}}_i^{(\ell)}$ is given by

$$\begin{aligned} \widehat{\mathbf{A}}_i^{(\ell)} &= \arg\min \sum_{j \in \mathcal{N}_i \cup \{i\}} p_j(1-p_j)\alpha_{ji}^2 \\ &+ 2\sum_{l \in [n], l \neq i} \sum_{j \in \mathcal{N}_{il}} p_j(1-p_j)\alpha_{ji}\alpha_{jl}^{(\ell-1)}, \\ &\text{s.t.: } \sum_{j \in \mathcal{N}_i \cup \{i\}} p_j\alpha_{ji} = 1, \quad \alpha_{ji} \geq 0 \quad \forall j \in [n]. \end{aligned} \tag{8}$$

Let $L_{ji} = \{l : l \in [n], l \neq i, j \in \mathcal{N}_{il}\}$, that is, the set of all clients that have j as a mutual neighbor with i, and let $\beta_{ji} = \sum_{l \in L_{ji}} \alpha_{jl}^{(\ell-1)}$. Let $\overline{p}(i) = \max_{k \in \mathcal{N}_i \cup \{i\}} p_k$. Using Lagrange multipliers we solve (8) for $j \in \mathcal{N}_i \cup \{i\}$ as follows:

$$\widehat{\alpha}_{ji}^{(\ell)} = \begin{cases} \left(-\beta_{ji} + \frac{\lambda_i}{2(1-p_j)}\right)^+ & \text{if } p_j \in (0,1), \overline{p}(i) < 1, \\ \frac{1}{\sum_{k \in [n]} \mathbb{1}_{\{p_k = 1, k \in \mathcal{N}_i \cup \{i\}\}}} & \text{if } p_j = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Here, λ_i satisfies $\sum_{j \in \mathcal{N}_i \cup \{i\}} p_j\left(-\beta_{ji} + \frac{\lambda_i}{2(1-p_j)}\right)^+ = 1$, and $(\cdot)^+ \triangleq \max\{\cdot, 0\}$. We can find λ_i using the bisection method. The complete algorithm is detailed in Alg. 3. Its overall *complexity* is $O(L \cdot (n^2 + K))$, where K is the number of iterations used in the bisection method for optimizing λ_i.

Remark 2. *The optimization* (6) *only requires client* i *to know the weight values for its neighbors of distance 2. Thus, we can exploit the communication links between clients, and optimize* (6) *distributively. We present the distributed algorithm in [49].*

V. Numerical Simulations

We consider training a ResNet-20 model for image classification on CIFAR-10 dataset over 10 clients; each executes 8 local training steps of local-SGD. All plots have been averaged over 5 different realizations. We used a learning rate of 0.1 for SGD, a coefficient of $1e-4$ for ℓ_2-regularization to prevent overfitting, and a batch-size of 64.

In Figs. 2 and 3, the dataset is distributed across the clients in an IID fashion. As benchmarks, we consider *Federated*

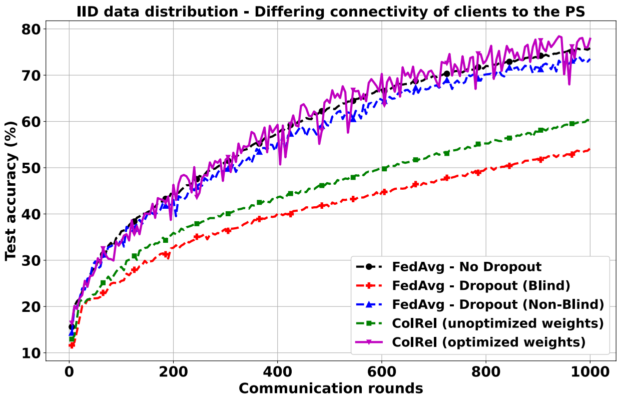

Fig. 3: Different connectivity across clients with a ring topology.

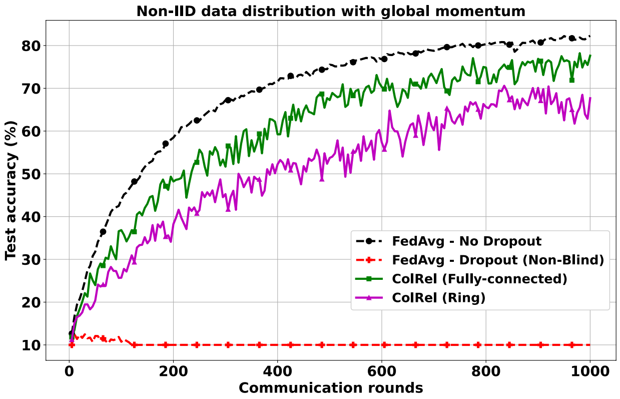

Fig. 4: Non-IID data + global momentum.

Averaging (FedAvg) - No Dropout, in which all clients are able to successfully transmit their local updates to the PS at every communication round. We also consider *FedAvg - Dropout*, in which the PS is unaware of the identity of the clients, and simply assumes that the update for any client unable to successfully transmit is zero. These benchmarks serve as natural upper and lower bounds to the performance of the proposed algorithm.

In Fig. 2, we have a homogeneous connectivity setup with equal probability $p_i = 0.2$ that client $i \in [n]$ successfully transmits its local updates to the PS. Furthermore, we assume a fully-connected topology (FCT) where each clients is connected to all the other clients in the system. ColRel achieves a performance on par with *FedAvg - No Dropout*. We also consider a non-blind strategy, *FedAvg - Dropout (Non-Blind)* where the PS is aware of the identity of the clients, and knows exactly which clients have been successful in sending their local updates to the PS. This is common in point-to-point learning settings. In this case the PS simply ignores the clients that have been unable to send their updates, and averages the successful updates by dividing the global aggregate at the PS by the number of successful transmissions.

In Fig. 3 (and also in Fig. 4), we consider every client has a different probability of successful transmission to the PS according to $\mathbf{p} = [0.1, 0.2, 0.3, 0.1, 0.1, 0.5, 0.8, 0.1, 0.2, 0.9]$. We have deliberately chosen some clients to have a very low connectivity, some others moderate, and others very high. We consider a ring topology where client i is connected to clients $(i-1) \mod n$ and $(i+1) \mod n$. For this setting, we distinguish the cases with and without optimized weights. The weights are optimized in order to minimize the term $S(\mathbf{p}, \mathbf{A})$, which consequently minimizes the variance of the iterates, subject to ensuring that the updates are unbiased according to Alg. 3. Note that explicitly optimizing the consensus weights that the clients use for their neighbors was not essential in Fig. 2 because the initial weights of Alg. 3 are optimal for a FCT with homogeneous connectivity to the PS, i.e., $p_i = p \forall i \in [n]$.

Finally, in Fig. 4, we consider the setting in which the training data is distributed across the clients in a non-IID fashion. To emulate non-IID-ness, we consider the *sort-and-partition* approach in which the training data is initially sorted based on labels, and then divided into blocks and distributed among clients in a skewed fashion so that each client has data from only a few classes. For the ring topology in this plot, we have considered each client to be connected to 4 of its nearest neighbors. We also use global momentum at the PS to update the global model. Remarkably, FedAvg (even with non-blind averaging) fails to converge in this setting. This is because in the absence of collaboration, clients that have important training samples that are critical for training a good model with high accuracy, may have a low probability of successful transmission and thus are rarely able to convey their updates to the PS. Therefore, when these clients are unable to convey their updates to the PS, the resulting test accuracy of the global model is $\sim 10\%$, as good as a random classifier for 10 classes. Collaborative relaying ensures that the information from these critical datapoints are also conveyed to the PS even when the data owner does not have connectivity to the PS.

VI. Conclusions

Our goal in this paper is to mitigate the detrimental effect of clients' intermittent connectivity on the training accuracy of FL systems. For this purpose, we proposed a collaborative relaying strategy, which exploits the connections between clients to relay potentially missing model updates to the PS due to blocked clients. Our algorithm allows the PS to receive an unbiased estimate of the model update, which would not be possible without relaying. We optimized the consensus weights at each client to improve the rate of convergence. Our proposed approach can be implemented even when the PS is blind to the identities of clients which successfully communicate with it at each round. Numerical results showed the improvement in training accuracy and convergence time that our approach provides under various setting, including IID and non-IID data distributions, different communication graph topologies, as well as blind and non-blind PSs.

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 *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54, Apr 2017, pp. 1273–1282.

[2] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," arxiv:2104.02151, 2021.

[3] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[4] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, M. R. Akdeniz, E. Aryafar, N. Himayat, S. Andreev, and Y. Koucheryavy, "On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10 124–10 138, Nov 2017.

[5] Y. Yan and Y. Mostofi, "Co-optimization of communication and motion planning of a robotic operation under resource constraints and in fading environments," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1562–1572, April 2013.

[6] M. M. Zavlanos, M. B. Egerstedt, and G. J. Pappas, "Graph-theoretic connectivity control of mobile robot networks," *Proc. IEEE*, vol. 99, no. 9, pp. 1525–1540, Sep. 2011.

[7] N. Michael, M. M. Zavlanos, V. Kumar, and G. J. Pappas, "Maintaining connectivity in mobile robot networks," in *Experimental Robotics*, 2009.

[8] S. Gil, S. Kumar, D. Katabi, and D. Rus, "Adaptive communication in multi-robot systems using directionality of signal strength," *Int. J. Rob. Res.*, vol. 34, no. 7, pp. 946–968, 2015.

[9] D. Gündüz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to learn at the edge," *IEEE Comm. Magazine*, vol. 58, no. 12, pp. 14–19, 2020.

[10] M. E. Ozfatura, J. Zhao, and D. Gündüz, "Fast federated edge learning with overlapped communication and computation and channel-aware fair client scheduling," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2021, pp. 311–315.

[11] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Data-importance aware user scheduling for communication-efficient edge machine learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 265–278, 2021.

[12] W. Xia, T. Q. S. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu, "Multi-armed bandit based client scheduling for federated learning," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2020.

[13] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. Vincent Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in *Proc. - ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 8743–8747.

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

[15] E. Ozfatura, S. Rini, and D. Gündüz, "Decentralized sgd with over-the-air computation," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[16] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning (extended version)," arxiv:1812.11494, 2019.

[17] B. Hasircioglu and D. Gunduz, "Private wireless federated learning with anonymous over-the-air computation," arxiv:2011.08579, 2021.

[18] M. Seif, W.-T. Chang, and R. Tandon, "Privacy amplification for federated learning via user sampling and wireless aggregation," arxiv:2103.01953, 2021.

[19] M. Yemini, S. Gil, and A. J. Goldsmith, "Exploiting local and cloud sensor fusion in intermittently connected sensor networks," in *2020 IEEE Global Communications Conference (Globecom)*, December 2020.

[20] ——, "Cloud-cluster architecture for detection in intermittently connected sensor networks," arXiv:2110.01119, 2021.

[21] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," in *Proc. - ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 8866–8870.

[22] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *NIPS*, Dec. 2017.

[23] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "d^2: Decentralized training over decentralized data," in *35th Int. Conf. Mach. Learn.*, vol. 80. PMLR, Jul 2018, pp. 4848–4856.

[24] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative deep learning in fixed topology networks," in *NIPS*, Dec. 2017.

[25] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, 2016.

[26] M. Kamp, L. Adilova, J. Sicking, F. Hüger, P. Schlicht, T. Wirtz, and S. Wrobel, "Efficient decentralized deep learning by dynamic model averaging," in *Conf. Mach. Learn. Knowl. Discovery in Databases*, 2019, pp. 393–409.

[27] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2834–2848, June 2018.

[28] L. Kong, T. Lin, A. Koloskova, M. Jaggi, and S. U. Stich, "Consensus control for decentralized deep learning," arXiv:2102.04828, 2021.

[29] T. Vogels, L. He, A. Koloskova, S. P. Karimireddy, T. Lin, S. U. Stich, and M. Jaggi, "Relaysum for decentralized deep learning on heterogeneous data," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[30] R. Saha, S. Rini, M. Rao, and A. J. Goldsmith, "Decentralized optimization over noisy, rate-constrained networks: Achieving consensus by communicating differences," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 2, pp. 449–467, 2022.

[31] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *37th Int. Conf. Mach. Learn.*, vol. 119. PMLR, Jul 2020, pp. 5381–5393.

[32] J. Wang, A. K. Sahu, Z. Yang, G. Joshi, and S. Kar, "MATCHA: speeding up decentralized SGD via matching decomposition sampling," arxiv:1905.09435, 2019.

[33] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *36th Int. Conf. Mach. Learn.* PMLR, Jun 2019, pp. 344–353.

[34] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[35] W. Y. B. Lim, J. S. Ng, Z. Xiong, J. Jin, Y. Zhang, D. Niyato, C. Leung, and C. Miao, "Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 3, pp. 536–550, 2022.

[36] T. Castiglia, A. Das, and S. Patterson, "Multi-level local {sgd}: Distributed {sgd} for heterogeneous hierarchical networks," in *International Conference on Learning Representations*, 2021.

[37] 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.

[38] Anonymous, "Hybrid local SGD for federated learning with heterogeneous communications," in *Submitted to The Tenth International Conference on Learning Representations*, 2022, under review. [Online]. Available: https://openreview.net/forum?id=H0oaWl6THa

[39] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-IID federated learning," in *International Conference on Learning Representations*, 2021.

[40] X. Gu, K. Huang, J. Zhang, and L. Huang, "Fast federated learning in the presence of arbitrary device unavailability," arxiv:2106.04159, 2021.

[41] T. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *CoRR*, vol. abs/1909.06335, 2019.

[42] T.-M. H. Hsu, H. Qi, and M. Brown, "Federated visual classification with real-world data distribution," *CoRR*, vol. abs/2003.08082, 2020.

[43] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-IID data quagmire of decentralized machine learning," in *37th Int. Conf. Mach. Learn.*, vol. 119. PMLR, Jul 2020, pp. 4387–4398.

[44] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *CoRR*, vol. abs/1806.00582, 2018.

[45] E. Ozfatura, D. Gunduz, and H. V. Poor, "Collaborative learning over wireless networks: An introductory overview," arxiv:2112.05559, 2021.

[46] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, 2020.

[47] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," in *IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[48] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," *IEEE Trans. Wireless Comm.*, vol. 20, no. 6, pp. 3643–3658, 2021.

[49] M. Yemini, R. Saha, E. Ozfatura, D. Gündüz, and A. J. Goldsmith, "Robust federated learning with connectivity failures: A semi-decentralized framework with collaborative relaying," arXiv:2202.11850, 2022.

[50] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.