Robust Federated Learning with Connectivity Failures: A Semi-Decentralized Framework with Collaborative Relaying

Michal Yemini, Rajarshi Saha, Emre Ozfatura,

Deniz Gündüz and Andrea J. Goldsmith

Abstract

Intermittent connectivity of clients to the parameter server (PS) is a major bottleneck in federated edge learning frameworks. The lack of constant connectivity induces a large generalization gap, especially when the local data distribution amongst clients exhibits heterogeneity. To overcome intermittent communication outages between clients and the central PS, we introduce the concept of collaborative relaying wherein the participating clients relay their neighbors' local updates to the PS in order to boost the participation of clients with poor connectivity to the PS. We propose a semi-decentralized federated learning framework in which at every communication round, each client initially computes a local consensus of a subset of its neighboring clients' updates, and eventually transmits to the PS a weighted average of its own update and those of its neighbors'. We appropriately optimize these local consensus weights to ensure that the global update at the PS is unbiased with minimal variance – consequently improving the convergence rate. Numerical evaluations on the CIFAR-10 dataset demonstrate that our collaborative relaying approach outperforms federated averaging-based benchmarks for learning over intermittently-connected networks such as when the clients communicate over millimeter wave channels with intermittent blockages.

M. Yemini and A. J. Goldsmith are with the Faculty of Electrical and Computer Engineering, Princeton University. (Emails: myemini@princeton.edu, goldsmith@princeton.edu). R. Saha is with the Department of Electrical Engineering, Stanford University. (Email: rajsaha@stanford.edu). E. Ozfatura and D. Gündüz are with the Department of Electrical and Electronic Engineering, Imperial College London. (Emails: m.ozfatura@imperial.ac.uk, d.gunduz@imperial.ac.uk).

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 [2]. FL approaches aim to reduce the required communication overhead and improve clients' privacy by letting each client train a local model on its local (private) dataset and forwarding them periodically to a centralized parameter server (PS).

In practical FL scenarios, some clients are stragglers and cannot send their estimates regularly, either because: (i) they cannot finish their computation within a prescribed deadline, or (ii) they cannot transmit their estimate to the PS successfully due to communication limitations [3]. Specifically, clients can suffer from intermittent connectivity to the PS, where their wireless communication channel is temporary blocked [4]–[10]. Stragglers deteriorate the convergence of FL as the computed local estimates become stale which reduced accuracy and can even result in bias in the final model in the case of persistent stragglers. However, the case of communication stragglers that are limited due to loss of direct communication opportunities to the PS is inherently different from those that result from limited computation resources at the client, since the former can be solved by relaying the updates via neighboring clients.

Communication channel quality is considered a key design principle in the federated edge learning (FEEL) framework [11], 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 [11]–[23]; this model ignores possible cooperation among clients in the case of intermittent communication blockages. Robustness to channel blockages is critical for the reliable operations of mmWave communication systems and robotic systems. While mmWave systems have many advantages in terms of utilizing a previously unused section of the radio spectrum for communication purposes, it is known to suffer from temporary channel blockage due to moving bodies [4]–[6]. Additionally, overcoming channel blockage is highly relevant for robotic networks where mobile robots explore remote parts of the area of interests and thus can be disconnected from a central server [7]–[10].

Relaying, also known as cooperative communications, has been widely studied in the literature in order to improve the coverage [24], diversity [25], end-to-end signal quality [26], [27] as

well as connectivity [28], [29] of wireless terminals. In FEEL, relaying has been considered previously in the context of hierarchical FL [30] to bring the clients closer to the PS with the aim of reducing the latency and interference. In this paper, motivated by our prior works [28], [29], where client cooperation is used to improve the connectivity to the cloud and to reduce the latency and scheduling overhead, we propose and analyze a new FEEL paradigm, where the clients cooperate to mitigate the detrimental effects of communication stragglers. In the proposed method, clients share with each other their current updates so that each client can send to the PS a weighted average of its own current update and those of its neighbors. Thanks to relaying, the PS receives new updates from clients with intermittently failing uplink connections, which would otherwise become stale and be discarded. Moreover, we optimize the averaging weights in order to ensure that the transmitted updates to the PS (i) achieve weak unbiasedness that preserves the objective function at the PS, and (ii) minimize the convergence time of the learning algorithm. We provide theoretical guarantees for the improvement in convergence rate of our scheme.

Most existing works on FL assume error-free rate-limited orthogonal communication links, assuming that the wireless imperfections are taken care of by an underlying communication protocol. However, such a separation between the communication medium and the learning protocol can be strictly suboptimal [11]. An alternative approach is to treat the communication of the model updates to the PS as an uplink communication problem and jointly optimize the learning algorithm and the communication scheme, taking wireless channel imperfections into account [11], [31]. Within this framework, an original and promising approach is the *over-the-air computation (OAC)* [16]–[20], which exploits the signal superposition property of the wireless medium to convey the sum of the model updates in an uncoded fashion. In addition to bandwidth efficiency, the OAC framework also provides a certain level of anonymity to the clients due to its superposition nature; and hence, enhances the privacy of the participating clients [21], [22]. We emphasize here that, in the OAC framework, the 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 the transmitting clients or their individual updates at the PS. Therefore, our solution is compatible with OAC, and can retain its bandwidth efficiency and anonymity benefits.

Related works: The conventional FL framework [2] is orchestrated by a centralized entity called PS, which helps participating clients to reach a consensus on the model parameters by aggregating their locally trained models. However, such a consensus mechanism requires the exchange of a large number of model parameters between the PS and clients, thereby creating significant communication overhead. A decentralized learning framework has been introduced as an alternative to FL, in which the PS is removed to mitigate a potential communication bottleneck and a single point of failure. In decentralized learning, each client shares its local model with the neighboring clients through device-to-device (D2D) communications, and model aggregation is executed at each client in parallel. In a sense, in decentralized learning, each client becomes a PS. The aggregation strategy at each client is determined according to the network topology, that is the connection pattern between the clients, and often a fixed topology is considered [32]–[41]. These results can be further extended to scenarios with time varying topologies [42]–[44].

An alternative approach to both centralized and decentralized schemes is the *hierarchical FL (HFL)* framework [30], [45]–[47], 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, while the aggregated models at the clusters are later aggregated at the main PS in a subsequent step to obtain the global model. This framework 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 [30].

Although HFL has certain advantages, this framework 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 concepts, which is referred to as *semi-decentralized FL*, where the local consensus follows decentralized learning with D2D communication, whereas the global consensus is orchestrated by the PS [48], [49]. One of the major challenges in FL that is not considered in the aforementioned works on semi-decentralized FL is the partial client connectivity [50], [51]. Unequal client participation due to intermittent connectivity exacerbates the impact of data heterogeneity [52]–[55], and increases the generalization gap.

The connectivity of the clients is a particularly significant challenge in FEEL, where the clients and the PS communicate over unreliable wireless channels. Due to their diverse physical environments and distances to the PS, clients may have different connectivity to each other and

the PS. This problem has been recently addressed in [12]–[15], [23], [56]–[59] by considering customized client selection mechanisms to seek a balance between the participation of the clients and the latency for the model aggregation in order to speed up the learning process. In this work, we adopt a different approach to the connectivity problem, and 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 information is conveyed to the PS with the help of their neighboring clients.

Finally, another related body of work considers coded computation as a possible solution to mitigate stragglers in distributed learning [60]–[65]. However, in the classical approach [60]–[64], data is strategically allocated to clients to create redundant computations that can be exploited by the PS using additional information about the stragglers' identities. The work [65] incorporates coded computing into FL by sharing additional information about client's features with the PS, this can cause additional privacy leakage. Furthermore, the PS cannot be blind to the client's identities and thus this scheme cannot be implemented OAC straightforwardly. Our approach does not require sharing additional client's information beside the client's updates. Finally, we note that conventional coded computation schemes spread the data among clients without constraints on the allowed combinations. In contrast, in our scheme the clients cannot choose their neighbors as they depend on the channel characteristics, thus coded computing techniques cannot be readily adopted.

A. Main Contributions

The main contributions of our paper can be summarized as follows:

- We propose a new semi-decentralized FL framework, which exploits local connections between clients to relay each other's estimates to the PS, to mitigate the negative impacts of intermittent client-PS connections on the learning performance. In the proposed framework. clients cooperate not only to learn a common model by exchanging model updates with the PS, but also to improve their connectivity.

- We optimize our proposed collaborative relaying approach to preserve the unbiasedness of the local updates and to minimize the expected convergence time to an optimal model.

- Our approach can be applied to PSs that are blind to the identities of the transmitting clients. Therefore, it is suitable to be used in conjunction with OAC.

- By conducting extensive simulations, we numerically show the superiority of the proposed framework compared to federated averaging (FedAvg), particularly when the data heterogeneity is taken into account to mimic realistic scenarios.

B. Paper Organization

The rest of the paper is organized as follows: Section II presents the FL system model and the proposed collaborative relaying scheme. Section III derives conditions for the unbiasedness of our collaborative relaying scheme and presents an upper bound on its expected distance to optimality. Using these analytical guarantees, Section IV optimizes our collaborative relaying scheme to reduce the upper bound on the expected suboptimality gap. Section V presents numerical results that validate our theoretical analysis and highlight the performance improvement in terms of the training accuracy collaborative relaying provides. Finally, Section VI concludes this paper.

II. SYSTEM MODEL FOR COLLABORATIVE RELAYING

Consider n clients each with its own local dataset collaborating with the help of a central PS to train a common model with d parameters represented by the d-dimensional vector $\mathbf{x} \in \mathbb{R}^d$. The clients communicate periodically with the PS over intermittently connected links to minimize an empirical loss function we define below.

Let $\mathcal{L}(\mathbf{x}, \boldsymbol{\zeta})$ be the loss evaluated for a model $\mathbf{x}$ at data point $\boldsymbol{\zeta}$. Denote the local dataset of client i by $\mathcal{Z}_i$, and its local loss function by $f_i : \mathbb{R}^d \times \mathcal{Z}_i \to \mathbb{R}$, where $f_i(\mathbf{x}; \mathcal{Z}_i) = \frac{1}{|\mathcal{Z}_i|} \sum_{\boldsymbol{\zeta} \in \mathcal{Z}_i} \mathcal{L}(\mathbf{x}; \boldsymbol{\zeta})$. The PS aims to solve the following empirical risk minimization (ERM) problem:

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

We assume uniform sizes for local datasets, i.e., $|\mathcal{Z}_i| = |\mathcal{Z}_j|$, $\forall i, j \in [n]$.

A. FL with Local SGD at Clients

We denote by $\nabla f_i(\mathbf{x})$ the gradient of the loss function at client i, i.e., $\nabla f_i(\mathbf{x}) \triangleq \nabla f_i(\mathbf{x}; \mathcal{Z}_i)$. Additionally, let $g_i(\mathbf{x})$ denote a stochastic gradient of $f_i(\mathbf{x}; \mathcal{Z}_i)$.

FL consists of two stages, a local optimization stage where clients update their local models iteratively, and an aggregation stage where the PS aggregates the local updates it receives from the clients. Let $\mathcal{T}$ be the *period* of local averaging, i.e., the number of local iterations *after* which the PS receives clients' estimates for the model parameters. At the beginning of the rth *round*

of FL, the PS broadcasts the global model $\mathbf{x}^{(r)}$ to the clients. For local iteration $k \in [0 : \mathcal{T}]$ of the rth round, client i applies the following local update rule:

$$\mathbf{x}_i^{(r,k+1)} = \mathbf{x}_i^{(r,k)} - \eta_r g_i\left(\mathbf{x}_i^{(r,k)}\right), \tag{2}$$

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

B. Communication Model

For simplicity of the exposition, we consider communication links from the clients to the PS, where each link is either unavailable or perfect; that is, when it is available it does not suffer from any channel impairments. We depict the considered communication model in Fig. 1.

Communication between the clients and PS: We model the intermittent connectivity of client i to the PS at round r by the Bernoulli random variable $\tau_i(r) \sim \text{Bernoulli}(p_i)$, where $\tau_i(r) = 1$ denotes a communication opportunity between client i and the PS at round r, whereas $\tau_i(r) = 0$ means the connection is blocked. We assume that the link connectivity to the PS is independent across rounds and across clients.

Communication between clients: We consider intermittent connectivity for the links between the clients as well. We capture the intermittent connectivity of a transmission from client i to client j by the random variable $\tau_{ij}(r) \sim \text{Bernoulli}(p_{ij})$, where $p_{ii} = 1$ for every $i \in [n]$. For simplicity of exposition, we assume that the inter-client links and the links from the clients to the PS are statistically independent.

Additionally, we assume that $\tau_{ij}(r)$, $\tau_{lm}(r')$, are statistically independent for every $i, j, l, m \in [n]$ and $r, r' \in \mathbb{N}$ such that $(i,j) \neq (l,m)$ and $(j,i) \neq (l,m)$, or $r \neq r'$.

We use the notation $E_{\{i,j\}} = \mathbb{E}[\tau_{ij}(r)\tau_{ji}(r)]$ for every $i, j \in [n]$ and $r \in \mathbb{N}$ to denote the correlation due to channel reciprocity. Finally, we assume that $\tau_i(r+1), \tau_{ij}(r+1)$ are statistically independent of $\Delta \mathbf{x}_l^{r+1}$ for every $i, k, l \in [n]$

Remark: We assume that the connectivity probabilities p_i, p_{ij}, $i, j \in [n]$, are known. In practice, they can be easily estimated. 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 the instantaneous connectivity information $\tau_i(r)$, $r \in [n]$, is available to any of the clients.

We assume that the communicating clients can send their updates to one another over the inter-client links. Therefore, each client can send to the PS a weighted average of its own update and those of its neighbors. In this way, the PS can receive estimates from clients with failing uplink connections. Based on this observation, we next present our collaborative relaying procedure.
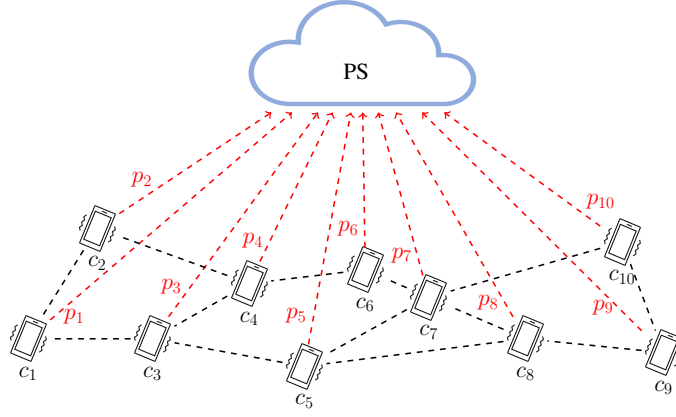
Fig. 1: Illustration of the system model. Red arrows denote intermittent uplink connections from the clients to the PS. Additionally, black dashed lines denote intermittent connections between the clients. Here, c_i denotes client i, while p_i is the connectivity probability of the link from c_i to the PS.

C. Collaborative Relaying of Local Updates

Let $\Delta\mathbf{x}_j^{r+1}$ denote the update of client i at the end of the $\mathcal{T}$th local iteration in round r; i.e.,

$$\Delta\mathbf{x}_j^{r+1} = \mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)}.$$

We assume that the model update of client i at the end of the $\mathcal{T}$th iteration is readily available to every client j such that $\tau_{ij}(r+1) = 1$. Then, client i forms a linear combination of its own update and those of its neighbors

$$\Delta\widetilde{\mathbf{x}}_i^{r+1} = \sum_{j\in[n]} \tau_{ji}(r+1)\alpha_{ij} \cdot \Delta\mathbf{x}_j^{r+1} = \sum_{j\in[n]} \alpha_{ij}\tau_{ji}(r+1)\left(\mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right), \tag{3}$$

where α_{ij} is the weight used by client i to relay the update from client j. The complexity of this averaging is low since it involves a weighted averaging of $O(\max_{i\in[n]} |\{j : p_{ji} > 0\}| + 1)$.

D. PS Aggregation

We assume that the PS does not explicitly select which subset of clients it will receive updates from, but rather receives information from all the *communicating* clients in each round.

The PS can use the following rescaled sum of the received updates:

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

$$= \mathbf{x}^{(r)} + \frac{1}{n}\sum_{i\in[n]} \tau_i(r+1)\sum_{j\in[n]} \tau_{ji}(r+1)\alpha_{ij}\left(\mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right). \tag{4}$$

Note that this is a simple sum of the updates from all the connected devices, thus it can be computed over-the-air and does not require the PS to know the identities of the communicating

Algorithm 1: COLREL-CLIENT (Client Collaborative Relaying)

Input: Round number r, step-size η_r, round duration $\mathcal{T}$, neighbors of client i $\mathcal{N}_i$, α_{ij} for
   every $j \in \mathcal{N}_i \cup \{i\}$.

Output: $\Delta \widetilde{\mathbf{x}}_i^{r+1}$.

1 Receive $\mathbf{x}^{(r)}$ from PS

2 Set $\mathbf{x}_i^{(r,0)} = \mathbf{x}^{(r)}$

3 for $k \leftarrow 0$ **to** $\mathcal{T} - 1$ **do**

4   Generate a random gradient $g_i\left(\mathbf{x}_i^{(r,k)}\right)$

5 $\mathbf{x}_i^{(r,k+1)} = \mathbf{x}_i^{(r,k)} - \eta_r g_i\left(\mathbf{x}_i^{(r,k)}\right)$

6 end

7 Set $\Delta \mathbf{x}_i^{r+1} = \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}$

8 Send $\Delta \mathbf{x}_i$ to every $j \in \mathcal{N}_i$

9 Receive $\Delta \mathbf{x}_j$ from every $j \in [n]$ such that $\tau_{ji}(r+1) = 1$

10 Compute $\Delta \widetilde{\mathbf{x}}_i^{r+1} = \sum_{j \in [n]} \tau_{ji}(r+1) \alpha_{ij} \cdot \Delta \mathbf{x}_j^{r+1}$

11 Transmit $\Delta \widetilde{\mathbf{x}}_i^{r+1}$ to the PS

Algorithm 2: COLREL-PS (PS Aggregation)

Input: Number of rounds R, a set of clients $[n]$, scalar w.

Output: An estimate $\mathbf{x}^{(R)}$ of optimal value $\mathbf{x}^\star$

1 Set $\mathbf{x}^0 = \mathbf{0}$

2 for $r \leftarrow 0$ **to** $R - 1$ **do**

3   Send $\mathbf{x}^{(r)}$ to all clients

4   Let $\tau_i(r+1) = 1$ if PS successfully receives $\Delta \widetilde{\mathbf{x}}_i^{r+1}$ from client i and set
      $\tau_i(r+1) = 0$ otherwise.

5   Compute $\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \frac{1}{n} \sum_{i \in [n]} \tau_i(r+1) \Delta \widetilde{\mathbf{x}}_i^{r+1}$

6 end

---

clients at each round. We depict the overall collaborative relaying (ColRel) FL procedure in Algorithm 1 and Algorithm 2. Algorithm 1 describes the local updates at each client $i \in [n]$, and its collaborative relaying mechanism. Algorithm 2 describes the PS aggregation.

III. UNBIASEDNESS AND EXPECTED SUBOPTIMALITY GAP

This section presents sufficient conditions for the unbiasedness of the proposed collaborative relaying method. Under the unbiasedness conditions, we derive an upper bound on the expected suboptimality gap of the proposed method.

A. Sufficient Conditions for the Unbiasedness of Collaborative Relaying

Recall that α_{ji} is the weight client j assigns to the estimate it receives from client i. Client i and each client j that receives the transmission from client i successfully, i.e. $\{j \in [n] : \tau_{ij}(r+1) = 1\}$, try to send to the PS $\alpha_{ji}\Delta\mathbf{x}_i^{r+1}$ on behalf of client i.

Then, the total received update of client i at the PS is given by

$$\sum_{j\in[n]} \tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\Delta\mathbf{x}_i^{r+1}.$$

We are interested in observing the objective function at the PS, thus, we aim at deriving a condition on the weights α_{ji} such that the function at the PS, which we want to minimize, remains unbiased. The following lemma presents a sufficient condition on α_{ji} for unbiasedness.

Lemma 1 (Sufficient condition for unbiasedness)**.** *Let α_{ij} be such that*

$$\mathbb{E}\left[\sum_{j\in[n]} \tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\right] = p_i\alpha_{ii} + \sum_{j\neq i} p_j p_{ij}\alpha_{ji} = 1. \tag{5}$$

Then, for every $i \in [n]$ and $r \geq 0$

$$\frac{1}{n}\cdot\mathbb{E}\left[\sum_{j\in[n]} \tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\Delta\mathbf{x}_i^r \Big| \Delta\mathbf{x}_i^{r+1}\right] = \frac{1}{n}\cdot\Delta\mathbf{x}_i^r.$$

Proof. Since $\tau_j(r+1)$ and $\tau_{ij}(r+1)$ are statistically independent of $\Delta\mathbf{x}_i^r$ for all $i,j \in [n]$, we have that

$$\mathbb{E}\left[\sum_{j\in[n]} \tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\Delta\mathbf{x}_i^{r+1} \Big| \Delta\mathbf{x}_i^{r+1}\right] = \mathbb{E}\left[\sum_{j\in[n]} \tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\right]\Delta\mathbf{x}_i^{r+1}.$$

Substituting (5) concludes the proof. □

We note that if condition (5) is fulfilled, then we also have

$$\frac{1}{n}\cdot\mathbb{E}\left[\sum_{i\in[n]} \tau_i(r+1)\sum_{j\in[n]}\tau_{ji}(r+1)\alpha_{ij}\right] = \frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\right] = \frac{1}{n}\cdot n = 1.$$

However, since $\tau_j(r+1)$ and $\tau_{ji}(r+1)$ are random variables Lemma 1 does *not* imply that

$$\frac{1}{n} \cdot \sum_{j \in [n]} \tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\Delta \mathbf{x}_i^{r+1} = \frac{1}{n} \cdot \Delta \mathbf{x}_i^{r+1}.$$

Finally, the standard model of FL with random client sampling and no connectivity among clients is captured by substituting $p_{ij} = 0, p_i = p$, $\alpha_{ii} = 1$, and $\alpha_{ij} = 0$ for all $i, j \in [n]$ such that $j \neq i$.

B. Expected Suboptimality Gap

Next, we present an upper bound on the expected distance to optimality as a function of the weights α_{ij}, $i, j \in [n]$.

Assumption 1. *The loss functions f_i are L-smooth with respect to $\mathbf{x}$. That is, for every $i \in [n]$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.*

Assumption 2. *The stochastic gradients $g_i(\mathbf{x})$ are unbiased and have bounded variance, i.e.:*

1) $\mathbb{E}[g_i(\mathbf{x})] = \nabla f(\mathbf{x})$ *and*
2) *there exists σ^2 such that $\mathbb{E}[\|g_i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$ for every $i \in [n]$, $\mathbf{x} \in \mathbb{R}^d$.*

The main focus of this paper is mitigating the harmful effect of intermittent connectivity for the convergence of the global model at the PS. Thus, for simplicity of exposition, we present analytical results for homogeneous data distribution setting, where $E[g_i(\mathbf{x})] = \nabla f(\mathbf{x})$. Our numerical results show that the collaborative relaying approach and the proposed weight optimization algorithms we present in this paper lead to a significant reduction in the training loss in heterogeneous data settings as long as $\mathbb{E}[g_i(\mathbf{x})] = \nabla f_i(\mathbf{x})$.

Assumption 3. *The loss functions f_i are μ strongly convex. That is, for every $i \in [n]$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have $\langle \nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$.*

Assumption 4. *The weights α_{ij} are nonnegative, i.e., $\alpha_{ij} \geq 0, \forall i, j \in [n]$.*

Before stating the main result, we introduce some definitions. We let $\mathbf{A} \triangleq (\alpha_{ij})_{i,j \in [n]}$ and $\boldsymbol{P} \triangleq (p_{ij})_{i,j \in [n]}$. Additionally, we define

$$S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \sum_{i,j,l \in [n]} p_j(1 - p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl} + \sum_{i,j \in [n]} p_{ij}p_j(1 - p_{ij})\alpha_{ji}^2 + \sum_{i,l \in [n]} p_i p_l(E_{\{i,l\}} - p_{il}p_{li})\alpha_{il}\alpha_{li}.$$

Theorem 1 (Expected Distance to Optimality). *Let*

$$B(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \frac{2L^2}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \quad and \quad r_0(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \max\left\{\frac{L}{\mu}, 4\left(\frac{B(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})}{\mu^2} + 1\right), \frac{1}{\mathcal{T}}, \frac{4n}{\mu^2 \mathcal{T}}\right\}.$$

Additionally, let $\eta_r \triangleq \frac{4\mu^{-1}}{r\mathcal{T}+1}$, $C_1(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \frac{4^2}{\mu^2} \cdot \frac{2\sigma^2}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$, $C_2 \triangleq \frac{4^2}{\mu^2} \cdot L^2 \frac{\sigma^2}{n} e$, *and* $C_3(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \frac{4^4}{\mu^4} \cdot \left(L^2\sigma^2 e + \frac{2L^2\sigma^2 e}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})\right)$. *Then, under Assumptions 1-4 and condition* (5)*, for every* $r \geq r_0(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ *we have*

$$\mathbb{E}\left\|\mathbf{x}^{(r+1)} - x^\star\right\|^2 \leq$$

$$\frac{(r_0\mathcal{T}+1)}{(r\mathcal{T}+1)^2}\left\|\mathbf{x}^{(0)} - x^\star\right\|^2 + C_1(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})\frac{\mathcal{T}}{r\mathcal{T}+1} + C_2\frac{(\mathcal{T}-1)^2}{r\mathcal{T}+1} + C_3(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})\frac{\mathcal{T}-1}{(r\mathcal{T}+1)^2}, \quad (6)$$

given the update dynamic captured by (2)-(4).

We present the proof of Theorem 1 in Appendix A.

Theorem 1 implies that

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

Consequently, we can reduce the upper bound on the distance to optimality (6) by minimizing the term $S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ under the unbiasedness condition (5).

IV. OPTIMIZING THE WEIGHTS α_{ij}

Based on Theorem 1 we can minimize the upper bound on the expected distance to optimality by solving the following optimization problem

$$\min_{\mathbf{A}} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$$

$$\text{s.t.:} \sum_{j \in [n]} p_j p_{ij} \alpha_{ji} = 1, \qquad \alpha_{ji} \geq 0 \quad \forall i, j \in [n]. \quad (7)$$

Due to the term $\sum_{i,l \in [n]} p_i p_l (E_{\{i,l\}} - p_{il} p_{li}) \alpha_{il} \alpha_{li}$, the function $S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ is not necessarily convex with respect to $\mathbf{A}$. Therefore, we will instead consider the convex upper bound $\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ where,

$$\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \sum_{i,j,l \in [n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl} + \sum_{i,j \in [n]} p_{ij}p_j(1-p_{ij})\alpha_{ji}^2 + \sum_{i,l \in [n]} p_i p_l (E_{\{i,l\}} - p_{il}p_{li})\alpha_{li}^2.$$

Recalling that $E_{\{i,j\}} \geq p_{ij} p_{ji}, \forall i, j \in [n]$, we show in Lemma 2 that minimizing $\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ is a convex relaxation of (7). Thus, we initially use the Gauss-Seidel method to iteratively minimize $\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ with respect to $\mathbf{A}$, we denote this solution by $\mathbf{A}^*$. Then, we use the final iterate as

a warm-start initialization for Gauss-Seidel method to iteratively converge to a stationary point of $S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ in the vicinity of $\mathbf{A}^*$.

Lemma 2. *For every $\mathbf{A}$ such that $\alpha_{ij} \geq 0$, $\forall\, i, j \in [n]$, $S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \leq \overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$. Additionally, $\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ is convex with respect $\mathbf{A}$ for $\boldsymbol{p} \in [0, 1]^n$.*

We present the proof of this lemma in Appendix D. By Lemma 2, the following optimization problem is convex:

$$\min_{\mathbf{A}} \overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$$

$$\text{s.t.:} \sum_{j \in [n]} p_j p_{ij} \alpha_{ji} = 1, \qquad \alpha_{ji} \geq 0 \quad \forall i, j \in [n]. \tag{8}$$

Let $\mathbf{A}_i$ denote the ith column of $\mathbf{A}$, that is, $\mathbf{A}_i = (\mathbf{A}_{1i}, \ldots, \mathbf{A}_{ni})^T$. Since the domain of the problems (7) and (8) is separable with respect to $\mathbf{A}_i$, we can use the Gauss-Seidel method to iteratively solve (8) and converge to an optimal solution [66, Proposition 2.7.1] for the upper bound (8). We can then converge to a stationary point of (7) in the vicinity of this solution by utilizing again the Gauss-Seidel method.

Remark: We note that when $p_{ij} \in \{0, 1\}$, $\forall i, j \in [n]$ the problem (7) is convex. Additionally, in this case the problems (7) and (8) coincide since $E_{\{i,l\}} - p_{il} p_{li} = 0$ for every $i, j \in [n]$.

a) Optimizing the convex relaxation (8): Let $\mathbf{A}_i^{(\ell)}$ denote the approximated value for $\mathbf{A}_i$ in the ℓth iteration of the Gauss-Seidel method of (8). We choose the initial solution $\mathbf{A}_{ji}^{(0)} = \frac{1}{\left(\sum_{k \in [n]} \mathbb{1}_{\{p_k > 0, p_{ik} > 0\}}\right)} p_j p_{ij} \cdot \mathbb{1}_{\{p_j > 0, p_{ij} > 0\}}$. We can improve our solution iteratively by the Gauss-Seidel method until convergence to an optimal point of (8). That is, at every iteration ℓ we compute $\mathbf{A}^{(\ell)}$ as follows

$$\mathbf{A}_i^{(\ell)} = \begin{cases} \widehat{\mathbf{A}}_i^{(\ell)} & \text{if } \ell \mod n + n \cdot \mathbb{1}_{\{\ell \mod n = 0\}} = i \\ \mathbf{A}_i^{(\ell-1)} & \text{otherwise,} \end{cases} \tag{9}$$

where

$$\widehat{\mathbf{A}}_i^{(\ell)} = \arg\min \left[\sum_{j \in [n]} p_j p_{ij} \left(1 - p_j p_{ij}\right) \alpha_{ji}^2 + 2 \sum_{l \in [n], l \neq i} \sum_{j \in [n]} p_j (1 - p_j) p_{ij} p_{lj} \alpha_{ji} \alpha_{jl}^{(\ell-1)} \right.$$

$$\left. + \sum_{j \in [n]} p_i p_j (E_{\{i,j\}} - p_{ij} p_{ji}) \alpha_{ji}^2 \right],$$

$$\text{s.t.:} \sum_{j \in [n]} p_j p_{ij} \alpha_{ji} = 1, \ \alpha_{ji} \geq 0, \quad \forall j \in [n]. \tag{10}$$

Using Lagrange multipliers, we show in Appendix E that the optimal value for $\widehat{\mathbf{A}}_i^{(\ell)}$ is:

$$\widehat{\alpha}_{ji}^{(\ell)}(\lambda_i) = \begin{cases} \left(\dfrac{-2(1-p_j)\sum_{l\in[n]:l\neq i} p_{lj}\alpha_{jl}^{(\ell-1)}+\lambda_i}{2[(1-p_jp_{ij})+p_i(E_{\{i,j\}}/p_{ij}-p_{ji})]} \right)^+ & \text{if } p_jp_{ij}\cdot\max_{k\in[n]} p_kp_{ik} \in (0,1), \\[2ex] \dfrac{1}{\sum_{k\in[n]} \mathbb{1}_{\{p_kp_{ik}=1\}}} & \text{if } p_jp_{ij}=1, \\[2ex] 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $(a)^+ \triangleq \max\{a,0\}$ and λ_i is set such that $\sum_{j\in[n]} p_jp_{ij}\widehat{\alpha}_{ji}^{(\ell)}(\lambda_i) = 1$. We can find λ_i using the bisection method over the interval $\left[0, \max_{j:p_jp_{ij}\in(0,1)} \left\{ \frac{2[(1-p_jp_{ij})+p_i(E_{\{i,j\}}/p_{ij}-p_{ji})]}{p_jp_{ij}} + 2(1-p_j)\sum_{l\in[n]:l\neq i} p_{lj}\alpha_{jl}^{(\ell-1)}\right\}\right]$

b) Fine tuning (7): Let $\mathbf{A}_i^{(\ell)}$ denote the approximated value for $\mathbf{A}_i$ in the ℓth iteration of the Gauss-Seidel method of (7), and assume a given initialization $\mathbf{A}_i^{(0)}$. We can improve our solution iteratively by the Gauss-Seidel method until convergence to a stationary point of (7).

At every iteration ℓ we compute $\mathbf{A}^{(\ell)}$ as follows

$$\mathbf{A}_i^{(\ell)} = \begin{cases} \widehat{\mathbf{A}}_i^{(\ell)} & \text{if } \ell \mod n + n\cdot\mathbb{1}_{\{\ell \mod n=0\}} = i \\[1.5ex] \mathbf{A}_i^{(\ell-1)} & \text{otherwise,} \end{cases} \tag{12}$$

where $\mathbb{1}_{\{.\}}$ denotes the indicator function, and

$$\widehat{\mathbf{A}}_i^{(\ell)} = \arg\min \left[\sum_{j\in[n]} p_jp_{ij}(1-p_jp_{ij})\alpha_{ji}^2 + 2\sum_{l\in[n],l\neq i}\sum_{j\in[n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl}^{(\ell-1)} \right.$$
$$\left. + 2\sum_{j\in[n]} p_ip_j(E_{\{i,j\}}-p_{ij}p_{ji})\alpha_{ij}\alpha_{ji} \right],$$

$$\text{s.t.: } \sum_{j\in[n]} p_jp_{ij}\alpha_{ji}=1, \quad \alpha_{ji}\geq 0, \quad \forall j\in[n]. \tag{13}$$

Using Lagrange multipliers, we show in Appendix E that the optimal value for $\widehat{\mathbf{A}}_i^{(\ell)}$ is as follows:

$$\widehat{\alpha}_{ji}^{(\ell)}(\lambda_i) = \begin{cases} \left(\dfrac{-2(1-p_j)\sum_{l\in[n]:l\neq i} p_{lj}\alpha_{jl}^{(\ell-1)}-2p_i\left(\frac{E_{\{i,j\}}}{p_{ij}}-p_{ji}\right)\alpha_{ij}^{(\ell-1)}+\lambda_i}{2(1-p_jp_{ij})} \right)^+ & \text{if } p_jp_{ij}\max_{k\in[n]} p_kp_{ik} \in (0,1), \\[2ex] \dfrac{1}{\sum_{k\in[n]} \mathbb{1}_{\{p_kp_{ik}=1\}}} & \text{if } p_jp_{ij}=1, \\[2ex] 0 & \text{otherwise,} \end{cases}$$
$$\tag{14}$$

where λ_i is set such that $\sum_{j\in[n]} p_jp_{ij}\widehat{\alpha}_{ji}^{(\ell)}(\lambda_i)=1$. We can find λ_i using the bisection method over the interval $\left[0, \max_{j:p_jp_{ij}\in(0,1)} \left\{ \frac{2(1-p_jp_{ij})}{p_jp_{ij}} + 2(1-p_j)\sum_{l\in[n]:l\neq i} p_{lj}\alpha_{jl}^{(\ell-1)} + 2p_i\left(\frac{E_{\{i,j\}}}{p_{ij}}-p_{ji}\right)\alpha_{ij}^{(\ell-1)}\right\}\right]$.

We summarize the centralized optimization procedure for $\mathbf{A}$ in Algorithm 3. Additionally, when the connectivity between two communicating clients is reliable, i.e., two clients are either connected with high probability or disconnected, Algorithm 3 can be implemented in a distributed fashion by the clients. In this case each client does not need to fully know $\boldsymbol{p}, \boldsymbol{P}$ and $\mathbf{A}$, but only the weights and transmission probabilities of all its direct neighbors and its second degree neighbors (i.e., neighbors of its neighbors). Such a distributed algorithm can be used to optimize the weights when the PS is blind to the identities of the transmitting clients at all times and cannot use a training period for learning the transmission probabilities. Finally, we note the weights resulting from Algorithm 3 can be used as long as the probabilities $\boldsymbol{p}$ and $\boldsymbol{P}$ are fixed, and are not needed to be calculated in every communication round.

Computation complexity: The overall computational complexity of Algorithm 3 is $O(I \cdot (n^2 + K))$, where K is the number of iterations used in the bisection method for optimizing λ_i.

V. NUMERICAL SIMULATIONS

Simulation setup: We train a ResNet-20 model for image classification on the CIFAR-10 [67] dataset. We distributed the training set of $50,000$ images across $n = 10$ clients in both independent and identically distributed (IID) and non-IID fashions. Non-IID data distribution amongst clients is prevalent in FL applications, to emulate it, we consider the *sort-and-partition* approach wherein the training data is initially sorted based on the labels, and then they are divided into blocks and distributed among the clients randomly based on a parameter s, which measures the skewness of the data distribution. More precisely, s defines the maximum number of different labels present in the local dataset of each user, and therefore, smaller s implies a more skewed data distribution. We use $s = 3$, i.e. each client has images from at most 3 classes.

The plotted results for all the simulations are obtained by averaging over 5 independent realizations. In between every communication round to the PS, the clients execute 8 local training steps of local-SGD. We utilize the SGD optimizer at the clients with a global momentum ($\beta = 0.9$) at the PS. We used a learning rate of 0.05 for SGD, a coefficient of 10^{-4} for ℓ_2-regularization to prevent overfitting, and a batch-size of 64. All simulations were carried out on NVIDIA GeForce GTX 1080 Ti with a CUDA Version 11.4. We also note that we have ensured all the simulations to have the same step-size for a fair comparison. In other words, we have *not* optimized the learning-rate for individual simulations.

Algorithm 3: COPT-α Centralized optimization of the weight matrix $\mathbf{A}$

Input: A set of clients $[n]$, a connectivity graph G, the functions $\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ and $S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$, vector of connectivity probabilities $\boldsymbol{p}$, matrix of inter-client connectivity probabilities $\boldsymbol{P}$, maximal number of iteration I.

Output: A matrix $\mathbf{A}^{(L)}$ that approximately minimizes (7)

1 **Initialize (convex relaxation)**: $\mathbf{A}_{ji}^{(0)} = \frac{1}{\left(\sum_{j\in[n]} \mathbb{1}_{\{p_j>0, p_{ij}>0\}}\right)p_j p_{ij}} \cdot \mathbb{1}_{\{p_j>0, p_{ij}>0\}}$, and $\ell = 0$.

2 **while** $\ell \leq I - 1$ **do**

3     $\ell \leftarrow \ell + 1$

4     $i \leftarrow \ell \mod n + n \cdot \mathbb{1}_{\{\ell \mod n = 0\}}$

5   Compute $\widehat{\mathbf{A}}_i^{(\ell)}$ according to (11)

6   Set $\mathbf{A}_k^{(\ell)}$ according to (9) for every $k \in [n]$

7 **end**

8 **Warm-initialize (fine tuning)**: $\mathbf{A}^{(0)} = \mathbf{A}^{(L)}$, and $\ell = 0$

9 **while** $\ell \leq I - 1$ **do**

10     $\ell \leftarrow \ell + 1$

11     $i \leftarrow \ell \mod n + n \cdot \mathbb{1}_{\{\ell \mod n = 0\}}$

12   Compute $\widehat{\mathbf{A}}_i^{(\ell)}$ according to (14)

13   Set $\mathbf{A}_k^{(\ell)}$ according to (12) for every $k \in [n]$

14 **end**

To illustrate the advantages of COLREL we compare it with three benchmark strategies:

FedAvg – perfect connectivity. We consider FedAvg when all the clients are able to successfully transmit their local updates to the PS at every communication round. This serves as a natural upper bound to the performance of any algorithm proposed in the presence of intermittent connectivity.

FedAvg – Blind. As a natural performance lower bound in the presence of intermittent client connectivity, we consider a naïve FedAvg strategy wherein the PS is unaware of the identity of clients. In this strategy, for the clients that are unable to send their updates to the PS due to a communication failure, the PS simply assumes that their update is zero. Essentially, the PS adds all the local updates it receives at any communication round, and divides it by the total number of clients irrespective of the knowledge of the number of actual successful transmissions. Such
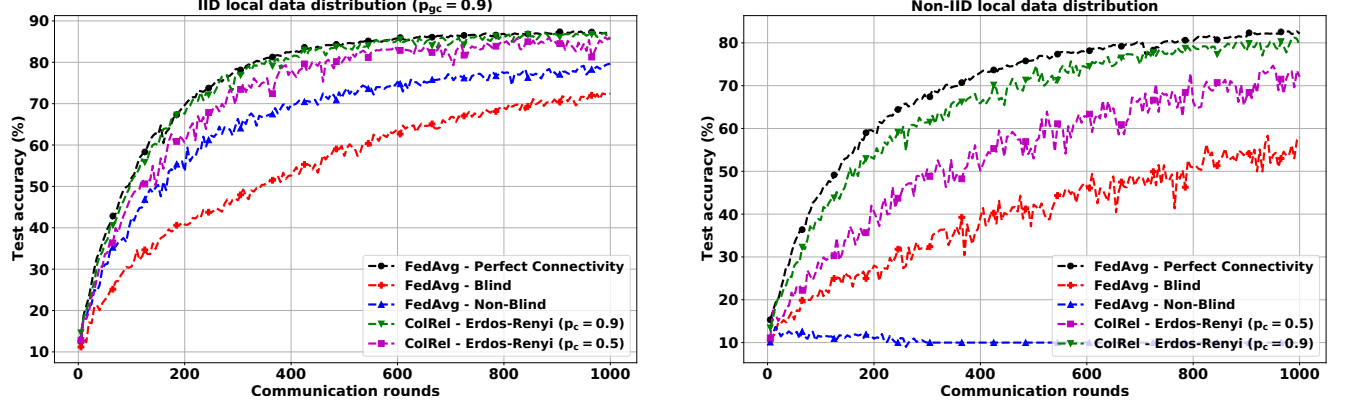
(a) Comparison of federated learning strategies with only one client with good connectivity to the PS

(b) Comparison of federated learning strategies with heterogenous connectivity of clients to the PS

Fig. 2: Network topologies for perfect vs. intermittent client-client connectivities

blind averaging strategies are often the norm for FEEL employing OAC.

FedAvg – Non-Blind. As another benchmark, we consider a non-blind strategy, where the PS is aware of the identity of the clients, and knows exactly, how many and which clients have successfully been able to send their local update 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.

We compare the above-mentioned strategies with our proposed collaborative relaying strategy in the presence of intermittent connectivity of clients to the PS, as well as amongst themselves. In order to demonstrate the improved performance of our proposed strategy with respect to the above-mentioned benchmarks, we consider the following setups:

Effect of topology. The decentralized topology according to which the clients collaborate in sharing their updates plays an important role in the performance of COLREL. In particular, we investigate a setting in which intermittent connectivity between clients might be preferred over perfect connectivity because it leads to increased collaboration on average.

Non-IID data distribution. We evaluate the performance of COLREL in the presence of non-IID local data distribution. For instance, COLREL outperforms other strategies when some clients with important local data have persistently poor connectivity.

Heterogeneity in client connectivity. Different clients may have different connectivity to the PS, and we show that as long as there is one client with reasonably good connection to the PS, its neighbors can relay their updates to it – yielding improved performance for COLREL.

1) IID local data distribution across clients: In Fig. 2a, we consider the setup wherein only one client has good connectivity to the PS with probability $p_{gc} = 0.9$, while all the remaining clients have a probability of only 0.1 for successfully transmitting their updates to the PS. For collaboration, we consider an Erdős-Rényi topology wherein a connection between any two clients is present with a certain probability p_c, with an additional constraint that $\tau_{ij} = 0 \iff \tau_{ji} = 0$. The cases $p_c = 0.9$ and $p_c = 0.5$ correspond to different degrees of intermittent decentralized collaboration. The performance of our proposed collaborative strategy, COLREL is comparable to FedAvg with perfect connectivity. We also note that we have ensured all the simulations to have the same step-size, i.e. the learning rate for different simulations has not been tuned individually.

2) Non-IID local data distribution across clients: In Fig. 2b, we consider different clients have different connection probabilities to the PS. In particular, some clients have a low probability of transmission, namely $p_1 = p_4 = p_5 = p_8 = 0.1$, some others moderate, and a couple of them high, i.e., $p_7 = 0.8$ and $p_{10} = 0.9$. Once again, COLREL for both cases when $p_c = 0.9$ and $p_c = 0.5$ outperform both blind and non-blind FedAvg algorithms, even with the relatively low local connectivity with $p_c = 0.5$. we also observe that the convergence of COLREL is faster and more stable when the inter-client connectivity probability p_c increases.

3) Network topology with mmWave links: In our preliminary work [1], we allow cooperation between clients only if they are connected via permanent communication links, that is, two clients i and j are connected so as to allow collaboration if they are within each other's coverage area. Thus, [1] assumes that p_{ij} and p_{ji} are either 0 or 1. In this work, we allow for client-client cooperation even when the communication links between them are intermittent.

Let d_i be the distance between client i an the PS and let d_{ij} be the distance between client i and client j, in meters. We consider a network topology with mmWave links, as in [4], in which $p_i = \min\left(1, \mathrm{e}^{-d_i/30+5.2}\right)$ and $p_{ij} = \min\left(1, \mathrm{e}^{-d_{ij}/30+5.2}\right)$.

In Fig. 3a we allow cooperation only among clients that are connected through permanent links. We use a threshold-based connectivity relation where $p_{ij} = 1$ if $d_{ij} > d_{\text{th}}$, and $p_{ij} = 0$ otherwise, where d_{th} is the distance that satisfies $\mathrm{e}^{-d_{\text{th}}/30+5.2} = 0.99$. The PS is located at the origin and the clients are distributed in a way such that only three of them can communicate with the PS. Similarly, in Fig. 3b, the connectivities between clients are intermittent. To avoid links that are too unreliable, if $p_{ij} < 0.5$, we consider those clients not to collaborate. Note that Fig. 3b has a few additional links compared to Fig. 3a. In Fig. 4, we observe that allowing intermittent

(a) Perfect client-client collaboration

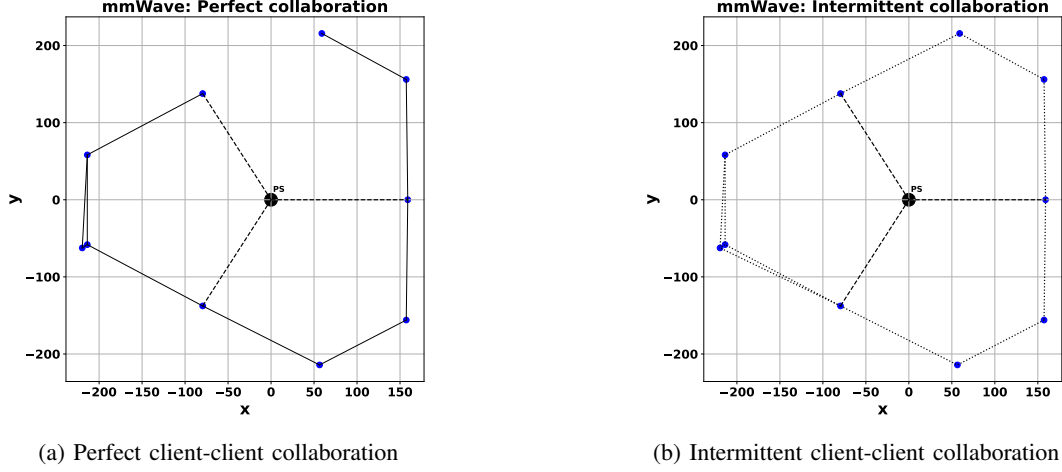(b) Intermittent client-client collaboration

Fig. 3: Network topologies for perfect vs. intermittent client-client connectivities
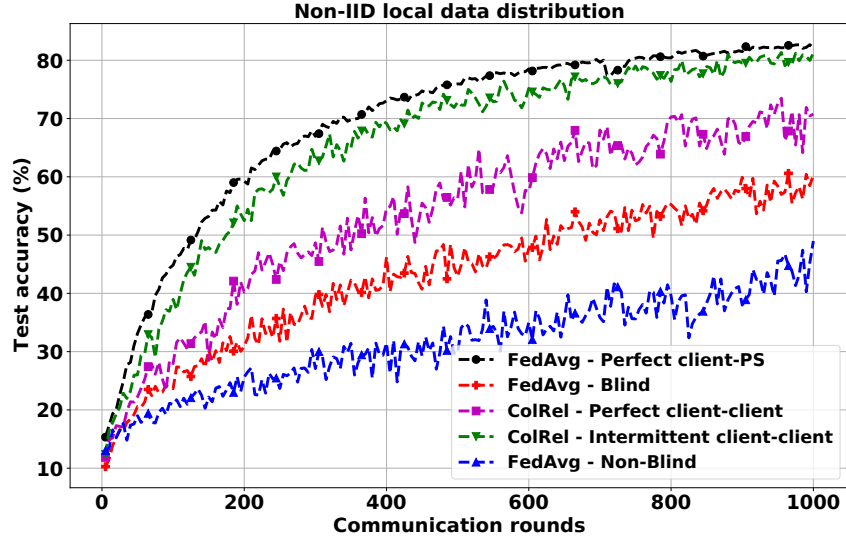

Fig. 4: Improved convergence due to increased collaboration with intermittent client-client connectivity

collaboration amongst clients results in improved convergence rate of training. Moreover, both perfect as well as intermittent connectivity outperform FedAvg without collaboration.

VI. CONCLUSION

FL has been proposed as an effective distributed learning strategy for mobile devices to train a common model with localized data. However, mobile devices may suffer from weak connectivity for extended durations, especially at the network edge. Our goal in this paper is to mitigate the detrimental effect of clients' intermittent connectivity on the training accuracy of FEEL systems. For this purpose, we proposed a collaborative relaying strategy, called COLREL, 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.

APPENDIX A

Before proving Theorem 1, we first introduce the following notation and lemmas. Denote:

$$\overline{\mathbf{x}}^{(r+1)} \triangleq \frac{1}{n} \sum_{i \in [n]} \mathbf{x}_i^{(r,\mathcal{T})}. \tag{15}$$

Lemma 3. *Under condition* (5), *for every round* $r \geq 0$ *we have*

$$\mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - x^\star\right\|^2\right] = \mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)}\right\|^2\right] + \mathbb{E}\left[\left\|\overline{\mathbf{x}}^{(r+1)} - \mathbf{x}^\star\right\|^2\right]. \tag{16}$$

We prove Lemma 3 in Appendix B.

Lemma 4. *Under Assumptions 1-4 and condition* (5), *for every* $r \geq 0$ *we have*

$$\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{(r+1)} - \mathbf{x}^\star\right\|^2\right] \leq$$

$$(1 + n\eta_r^2)(1 - \mu\eta_r)^{\mathcal{T}} \mathbb{E}\left[\|\mathbf{x}^{(r)} - \mathbf{x}^\star\|^2\right] + \mathcal{T}(\mathcal{T}-1)^2 L^2 \frac{\sigma^2}{n} e\eta_r^2 + \mathcal{T}^2 \frac{\sigma^2}{n}\eta_r^2 + \mathcal{T}^2(\mathcal{T}-1)^2 L^2 \sigma^2 e\eta_r^4,$$

for every positive η_r *such that* $\eta_r \leq \min\left\{\frac{\mu}{L^2}, \frac{1}{L\mathcal{T}}\right\}$.

Proof. The proof of this lemma follows directly from Lemma 2 in [68]. $\square$

Lemma 5. *Under Assumptions 1-4 and condition* (5), *for every* $r \geq 0$ *we have that*

$$\mathbb{E}\left[\left\|\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\|^2\right] \leq 2\mathcal{T}^2 L^2 \eta_r^2 \mathbb{E}\left[\|\mathbf{x}^{(r)} - \mathbf{x}^\star\|^2\right] + 2\mathcal{T}^2 \sigma^2 \eta_r^2 + 2(\mathcal{T}-1)\mathcal{T}^2 L^2 \sigma^2 e\eta_r^4,$$

for every positive η_r *such that* $\eta_r \leq \min\left\{\frac{\mu}{L^2}, \frac{1}{L\mathcal{T}}\right\}$.

Proof. The proof of this lemma follows directly from (57) in the proof of [68, Lemma 3]. $\square$

Lemma 6. *Under Assumptions 1-4 and condition* (5), *for every* $r \geq 0$ *we have that*

$$\mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)}\right\|^2\right] \leq \frac{2\mathcal{T}^2 L^2 \eta_r^2 \mathbb{E}\left[\|\mathbf{x}^{(r)} - \mathbf{x}^\star\|^2\right] + 2\mathcal{T}^2 \sigma^2 \eta_r^2 + 2(\mathcal{T}-1)\mathcal{T}^2 L^2 \sigma^2 e\eta_r^4}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}),$$

for every positive η_r *such that* $\eta_r \leq \min\left\{\frac{\mu}{L^2}, \frac{1}{L\mathcal{T}}\right\}$.

We prove Lemma 6 in Appendix C.

Proof of Theorem 1. By Lemmas 3-6, for every positive η_r that satisfies $\eta_r \leq \min\left\{\frac{\mu}{L^2}, \frac{1}{L\mathcal{T}}\right\}$, the following holds

$$\mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - x^\star\right\|^2\right] \leq (1 + n\eta_r^2)(1 - \mu\eta_r)^{\mathcal{T}}\mathbb{E}\left[\|\mathbf{x}^{(r)} - \mathbf{x}^\star\|^2\right] + \mathcal{T}(\mathcal{T}-1)^2 L^2 \frac{\sigma^2}{n} e\eta_r^2 + \mathcal{T}^2 \frac{\sigma^2}{n}\eta_r^2$$

$$+ \mathcal{T}^2(\mathcal{T}-1)^2 L^2\sigma^2 e\eta_r^4 + \frac{2\mathcal{T}^2 L^2\eta_r^2\mathbb{E}\left[\|\mathbf{x}^{(r)} - \mathbf{x}^\star\|^2\right] + 2\mathcal{T}^2\sigma^2\eta_r^2 + 2(\mathcal{T}-1)\mathcal{T}^2 L^2\sigma^2 e\eta_r^4}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}).$$

Recall the notation $B(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) = \frac{2L^2}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ and denote,

$$C(\eta_r, \boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq (1 + n\eta_r^2)(1 - \mu\eta_r)^{\mathcal{T}} + B(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})\mathcal{T}^2\eta_r^2, \quad C_1(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \frac{4^2}{\mu^2} \cdot \frac{2\sigma^2}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$$

$$C_2 \triangleq \frac{4^2}{\mu^2} \cdot L^2\frac{\sigma^2}{n}e, \quad C_3(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \triangleq \frac{4^4}{\mu^4} \cdot \left(L^2\sigma^2 e + \frac{2L^2\sigma^2 e}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})\right).$$

Therefore,

$$\mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - x^\star\right\|^2\right] \leq C(\eta_r, \boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \cdot \mathbb{E}\left[\left\|\mathbf{x}^{(r)} - x^\star\right\|^2\right]$$

$$+ \left[\frac{\mu^2}{4^2}C_1(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \cdot \mathcal{T}^2 + \frac{\mu^2}{4^2}C_2 \cdot \mathcal{T}(\mathcal{T}-1)^2\right]\eta_r^2 + \frac{\mu^4}{4^4}C_3(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \cdot \mathcal{T}^2(\mathcal{T}-1)\eta_r^4.$$

$$(17)$$

We can conclude the proof using a similar approach to that presented in [68]. First, we upper bound the term $C(\eta_r, \boldsymbol{p}, \mathbf{A})$. Since $\eta_r = \frac{4\mu^{-1}}{r\mathcal{T}+1} \leq \frac{1}{\mu}$ for every $r \geq r_0$, we can use the upper bound $\left(1 - \frac{x}{m}\right)^m \leq e^{-x}$ for every $x \leq m$:

$$(1 - \mu\eta_k)^{\mathcal{T}} = \left(1 - \frac{\mathcal{T}\mu\eta_k}{\mathcal{T}}\right) \leq e^{-\mu\mathcal{T}\eta_r}.$$

Now, using the bound $e^{-x} \leq 1 - x + x^2$ for every $x \geq 0$, we can conclude that $(1 - \mu\eta_k)^{\mathcal{T}} \leq 1 - \mu\mathcal{T}\eta_r + \mu^2\mathcal{T}^2\eta_r^2$. It follows that

$$C(\eta_r, \boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \leq (1 + n\eta_r^2)(1 - \mu\mathcal{T}\eta_r + \mu^2\mathcal{T}^2\eta_r^2) + B(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})\mathcal{T}^2\eta_r^2$$

$$= n\eta_r^2(1 - \mu\mathcal{T}\eta_r + \mu^2\mathcal{T}^2\eta_r^2) + 1 - \mu\mathcal{T}\eta_r + \mathcal{T}^2\eta_r^2\left(\mu^2 + B(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})\right).$$

Now, since $\eta_r = \frac{4\mu^{-1}}{r\mathcal{T}+1} \leq \frac{1}{\mu}$ and $r \geq r_0$ we have

$$\mu^2 + B(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \leq \frac{\mu}{4\eta_r}, \quad \eta_r \leq \frac{\mu\mathcal{T}}{4n}, \quad \text{and} \quad 0 \leq \mu\mathcal{T}\eta_r \leq 1.$$

The maximal value of the function $1 - x + x^2$ in the interval $x \in [0, 1]$ is 1, therefore, $1 - \mu\mathcal{T}\eta_r + \mu^2\mathcal{T}^2\eta_r^2 \leq 1$. It follows that $C(\eta_r, \boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \leq 1 - \frac{1}{2}\mu\mathcal{T}\eta_r$ and

$$\mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - x^\star\right\|^2\right] \leq \left(1 - \frac{1}{2}\mu\mathcal{T}\eta_r\right) \cdot \mathbb{E}\left[\left\|\mathbf{x}^{(r)} - x^\star\right\|^2\right]$$

$$+ \left[\frac{\mu^2}{4^2} C_1(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \cdot \mathcal{T}^2 + \frac{\mu^2}{4^2} C_2 \cdot \mathcal{T}(\mathcal{T} - 1)^2 \right] \eta_r^2 + \frac{\mu^4}{4^4} C_3(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \cdot \mathcal{T}^2(\mathcal{T} - 1) \eta_r^4. \quad (18)$$

Substituting $\eta_r = \frac{4\mu^{-1}}{r\mathcal{T} + 1}$ we can conclude the proof by using [68, Lemma 5] when replacing k with r and using the constants $k_1 = \frac{1}{\mathcal{T}}$, $a = C_1(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) + C_2 \cdot \frac{(\mathcal{T}-1)^2}{\mathcal{T}}$ and $b = C_3 \cdot \frac{(\mathcal{T}-1)}{\mathcal{T}^2}$. $\quad \square$

APPENDIX B

Proof of Lemma 3. Since $x \in \mathbb{R}^d$, for every $r \geq 0$ we have

$$\mathbb{E}\left[\left\| \mathbf{x}^{(r+1)} - x^\star \right\|^2 \right] = \mathbb{E}\left[\left\| \mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)} \right\|^2 \right] + \mathbb{E}\left[\left\| \overline{\mathbf{x}}^{(r+1)} - \mathbf{x}^\star \right\|^2 \right]$$
$$+ 2\mathbb{E}\left(\left\langle \mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)}, \overline{\mathbf{x}}^{(r+1)} - \mathbf{x}^\star \right\rangle \right).$$

Since $\overline{\mathbf{x}}^{(r+1)}$ and $\mathbf{x}^\star$ are deterministic functions of $(\mathbf{x}_i^{(r,\mathcal{T})})_{i \in [n]}$, by the law of total expectation

$$\mathbb{E}\left(\left\langle \mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)}, \overline{\mathbf{x}}^{(r+1)} - \mathbf{x}^\star \right\rangle \right) = \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathbf{x}^{(r+1)} \Big| \left(\mathbf{x}_i^{(r,\mathcal{T})} \right)_{i \in [n]}, \mathbf{x}^{(r)} \right] - \overline{\mathbf{x}}^{(r+1)}, \overline{x}^{(r+1)} - \mathbf{x}^\star \right\rangle \right].$$

Now, we calculate the conditional expectation

$$\mathbb{E}\left[\mathbf{x}^{(r+1)} \Big| \left(\mathbf{x}_i^{(r,\mathcal{T})} \right)_{i \in [n]}, \mathbf{x}^{(r)} \right] = \mathbb{E}\left[\mathbf{x}^{(r+1)} \Big| \left(\mathbf{x}_i^{(r,\mathcal{T})} \right)_{i \in [n]}, \mathbf{x}^{(r)} \right]$$
$$= \mathbb{E}\left[\mathbf{x}_i^{(r)} + \frac{1}{n} \sum_{i \in [n]} \tau_i(r+1) \sum_{j \in [n]} \tau_{ji}(r+1) \alpha_{ij} \left(\mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)} \right) \Big| \left(\mathbf{x}_i^{(r,\mathcal{T})}, \right)_{i \in [n]}, \mathbf{x}^{(r)} \right]$$
$$\overset{(a)}{=} \mathbf{x}^{(r)} - \frac{1}{n} \sum_{i \in [n]} p_i \sum_{j \in [n]} p_{ji} \alpha_{ij} \mathbf{x}_j^{(r)} + \frac{1}{n} \sum_{i \in [n]} p_i \sum_{j \in [n]} p_{ji} \alpha_{ij} \mathbf{x}_j^{(r,\mathcal{T})}$$
$$\overset{(b)}{=} \frac{1}{n} \sum_{i \in [n]} \mathbf{x}_i^{(r,\mathcal{T})} \triangleq \overline{\mathbf{x}}^{(r+1)},$$

where (a) follows since $\mathbb{E}[\tau_i(r+1)] = p_i$ and $\mathbb{E}[\tau_{ij}(r+1)] = p_{ij}$, and (b) follows from (5). $\quad \square$

APPENDIX C

Proof of Lemma 6. First, we observe that

$$\mathbb{E}\left[\left\| \mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)} \right\|^2 \right] = \mathbb{E}\left[\left\| \mathbf{x}^{(r)} + \frac{1}{n} \sum_{i \in [n]} \tau_i(r+1) \sum_{j \in [n]} \tau_{ji}(r+1) \alpha_{ij} \left(\mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)} \right) - \frac{1}{n} \sum_{i \in [n]} \mathbf{x}_j^{(r,\mathcal{T})} \right\|^2 \right]$$
$$= \mathbb{E}\left[\left\| \mathbf{x}^{(r)} - \frac{1}{n} \sum_{i \in [n]} \tau_i(r+1) \sum_{j \in [n]} \tau_{ji}(r+1) \alpha_{ij} \mathbf{x}^{(r)} \right. \right.$$
$$\left. \left. + \frac{1}{n} \sum_{i \in [n]} \tau_i(r+1) \sum_{j \in [n]} \tau_{ji}(r+1) \alpha_{ij} \mathbf{x}_j^{(r,\mathcal{T})} - \frac{1}{n} \sum_{i \in [n]} \mathbf{x}_i^{(r,\mathcal{T})} \right\|^2 \right]$$

$$= \mathbb{E}\left[\left\|\mathbf{x}^{(r)} - \frac{1}{n}\sum_{i,j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\mathbf{x}^{(r)} + \frac{1}{n}\sum_{i,j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\mathbf{x}_i^{(r,\mathcal{T})} - \frac{1}{n}\sum_{i\in[n]}\mathbf{x}_i^{(r,\mathcal{T})}\right\|^2\right]$$

$$= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji} - 1\right)\mathbf{x}_i^{(r,\mathcal{T})} - \frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji} - 1\right)\mathbf{x}^{(r)}\right\|^2\right]$$

$$= \frac{1}{n^2}\cdot\mathbb{E}\left[\left\|\sum_{i\in[n]}\left(\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji} - 1\right)\left(\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right)\right\|^2\right]$$

$$= \frac{1}{n^2}\sum_{i\in[n]}\mathbb{E}\left[\left(\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji} - 1\right)^2\right]\mathbb{E}\left[\left\|\left(\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right)\right\|^2\right]$$

$$+ \frac{1}{n^2}\sum_{i,l\in[n]:i\neq l}\mathbb{E}\left[\left(\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji} - 1\right)\left(\sum_{m\in[n]}\tau_m(r+1)\tau_{lm}(r+1)\alpha_{ml} - 1\right)\right]$$

$$\cdot\,\mathbb{E}\left[\left\langle\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}, \mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\rangle\right],$$

where the last equality follows since the random variables $\tau_i(r+1), i\in[n]$ and $\tau_{ij}(r+1), i,j\in[n]$ are statistically independent of the random vectors $\mathbf{x}_i^{(r,\mathcal{T})}, i\in[n]$. Now,

$$\mathbb{E}\left[\left(\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji} - 1\right)^2\right]$$

$$= \sum_{j\in[n]}\mathbb{E}\left[\tau_j^2(r+1)\tau_{ij}^2(r+1)\alpha_{ji}^2\right] + \sum_{j_1,j_2\in[n]:j_1\neq j_2}\mathbb{E}\left[\tau_{j_1}(r+1)\tau_{j_2}(r+1)\tau_{ij_1}(r+1)\tau_{ij_2}(r+1)\alpha_{j_1 i}\alpha_{j_2 i}\right]$$

$$- 2\sum_{j\in[n]}\mathbb{E}\left[\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\right] + 1$$

$$= \sum_{j\in[n]}p_j p_{ij}\alpha_{ji}^2 + \sum_{j_1,j_2\in[n]:j_1\neq j_2}p_{j_1}p_{j_2}p_{ij_1}p_{ij_2}\alpha_{j_1 i}\alpha_{j_2 i} - 2\underbrace{\sum_{j\in[n]}p_j p_{ij}\alpha_{ji}}_{=1} + 1$$

$$= \sum_{j\in[n]}p_j p_{ij}\alpha_{ji}^2 + \underbrace{\left(\sum_{j\in[n]}p_j p_{ij}\alpha_{ji}\right)^2}_{=1} - \sum_{j\in[n]}p_j^2 p_{ij}^2\alpha_{ji}^2 - 1 = \sum_{j\in[n]}p_j p_{ij}(1 - p_j p_{ij})\alpha_{ji}^2.$$

Similarly, for every $l\in[n]$ such that $l\neq i$

$$\mathbb{E}\left[\left(\sum_{j\in[n]}\tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji} - 1\right)\left(\sum_{m\in[n]}\tau_m(r+1)\tau_{lm}(r+1)\alpha_{ml} - 1\right)\right]$$

$$= \mathbb{E}\left[\sum_{j,m\in[n]} \tau_j(r+1)\tau_m(r+1)\tau_{ij}(r+1)\tau_{lm}(r+1)\alpha_{ji}\alpha_{ml}\right]$$

$$- \underbrace{\mathbb{E}\left[\sum_{j\in[n]} \tau_j(r+1)\tau_{ij}(r+1)\alpha_{ji}\right]}_{=1} - \underbrace{\mathbb{E}\left[\sum_{m\in[n]} \tau_m(r+1)\tau_{lm}(r+1)\alpha_{ml}\right]}_{=1} + 1$$

$$= \mathbb{E}\left[\sum_{j\in[n]} \tau_j^2(r+1)\tau_{ij}(r+1)\tau_{lj}(r+1)\alpha_{ji}\alpha_{jl}\right] + \mathbb{E}\left[\tau_l(r+1)\tau_i(r+1)\tau_{il}(r+1)\tau_{li}(r+1)\alpha_{li}\alpha_{il}\right]$$

$$+ \mathbb{E}\left[\sum_{m\in[n]:m\neq l,i} \tau_l(r+1)\tau_m(r+1)\tau_{il}(r+1)\tau_{lm}(r+1)\alpha_{li}\alpha_{ml}\right]$$

$$+ \mathbb{E}\left[\sum_{j\in[n]:j\neq l}\sum_{m\in[n]:m\neq j} \tau_j(r+1)\tau_m(r+1)\tau_{ij}(r+1)\tau_{lm}(r+1)\alpha_{ji}\alpha_{ml}\right] - 1$$

$$= \sum_{j\in[n]} p_j p_{ij} p_{lj}\alpha_{ji}\alpha_{jl} + p_i p_l E_{\{i,l\}}\alpha_{li}\alpha_{il} + \sum_{m\in[n]:m\neq l,i} p_l p_m p_{il} p_{lm}\alpha_{li}\alpha_{ml} + \sum_{\substack{j\in[n]:\\j\neq l}}\sum_{\substack{m\in[n]:\\m\neq j}} p_j p_m p_{ij} p_{lm}\alpha_{ji}\alpha_{ml} - 1$$

$$= \left(\sum_{j\in[n]} p_j p_{ij}\alpha_{ji}\right)\left(\sum_{m\in[n]} p_m p_{lm}\alpha_{ml}\right) - 1$$

$$- \sum_{j\in[n]} p_j^2 p_{ij}^2 p_{lj}\alpha_{ji}\alpha_{jl} + \sum_{j\in[n]} p_j p_{ij}^2 p_{lj}\alpha_{ji}\alpha_{jl} - p_i p_l p_{il} p_{li}\alpha_{il}\alpha_{li} + p_i p_l E_{\{i,l\}}\alpha_{il}\alpha_{li}$$

$$= \sum_{j\in[n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl} + p_i p_l(E_{\{i,l\}} - p_{il}p_{li})\alpha_{il}\alpha_{li}.$$

Therefore,

$$\mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)}\right\|^2\right]$$

$$= \frac{1}{n^2}\sum_{i,j\in[n]} p_j p_{ij}(1 - p_j p_{ij})\alpha_{ji}^2 \mathbb{E}\left[\left\|\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\|^2\right]$$

$$+ \frac{1}{n^2}\sum_{i,l\in[n]:i\neq l}\left(\sum_{j\in[n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl} + p_i p_l(E_{\{i,l\}} - p_{il}p_{li})\alpha_{il}\alpha_{li}\right) \mathbb{E}\left[\left\langle \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}, \mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\rangle\right]$$

$$= \frac{1}{n^2}\sum_{i,j,l\in[n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl}\cdot \mathbb{E}\left[\left\langle \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}, \mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\rangle\right]$$

$$+ \frac{1}{n^2}\sum_{i,j\in[n]} [p_{ij}p_j(1-p_{ij}p_j) - p_j(1-p_j)p_{ij}^2]\alpha_{ji}^2 \mathbb{E}\left[\left\|\left(\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right)\right\|^2\right]$$

$$+ \frac{1}{n^2} \sum_{i,l \in [n]: i \neq l} p_i p_l (E_{\{i,l\}} - p_{il}p_{li})\alpha_{il}\alpha_{li} \cdot \mathbb{E}\left[\left\langle \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}, \mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\rangle\right]$$

$$= \frac{1}{n^2} \sum_{i,j,l \in [n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl} \cdot \mathbb{E}\left[\left\langle \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}, \mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\rangle\right]$$

$$+ \frac{1}{n^2} \sum_{i,j \in [n]} p_{ij}p_j(1-p_{ij})\alpha_{ji}^2 \mathbb{E}\left[\left\|\left(\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right)\right\|^2\right]$$

$$+ \frac{1}{n^2} \sum_{i,l \in [n]} p_i p_l (E_{\{i,l\}} - p_{il}p_{li})\alpha_{il}\alpha_{li} \cdot \mathbb{E}\left[\left\langle \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}, \mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\rangle\right],$$

where the last equality follows since $E_{\{i,i\}} - p_{ii}p_{ii} = 1 - 1 \cdot 1 = 0$.

By Lemma 5 and the Cauchy–Schwarz inequality we have for each $i \in [n]$

$$\left|\mathbb{E}\left[\left\langle \mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}, \mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\rangle\right]\right| \leq \sqrt{\mathbb{E}\left[\left\|\mathbf{x}_i^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\|^2\right]} \cdot \sqrt{\mathbb{E}\left[\left\|\mathbf{x}_l^{(r,\mathcal{T})} - \mathbf{x}^{(r)}\right\|^2\right]}$$

$$\leq 2\mathcal{T}^2 L^2 \eta_r^2 E\|\mathbf{x}^{(r)} - \mathbf{x}^\star\|^2 + 2\mathcal{T}^2\sigma^2\eta_r^2 + 2(\mathcal{T}-1)\mathcal{T}^2 L^2\sigma^2 e\eta_r^4.$$

Consequently, it follows from Assumption 4 and the definition of $S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ that

$$\mathbb{E}\left[\left\|\mathbf{x}^{(r+1)} - \overline{\mathbf{x}}^{(r+1)}\right\|^2\right] \leq \frac{2\mathcal{T}^2 L^2 \eta_r^2 \mathbb{E}\left[\|\mathbf{x}^{(r)} - \mathbf{x}^\star\|^2\right] + 2\mathcal{T}^2\sigma^2\eta_r^2 + 2(\mathcal{T}-1)\mathcal{T}^2 L^2\sigma^2 e\eta_r^4}{n^2} S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}).$$

$\square$

APPENDIX D

To prove Lemma 2 we first present the following auxiliary result.

Lemma 7. *Let* $\mathbf{y}, \mathbf{c} \in \mathbb{R}^{\tilde{d}}$ *where* $\tilde{d} \in \mathbb{N}_+$, *and* $h_{\mathbf{c}}(\mathbf{y}) \triangleq \left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{y}_i\right)^2$. *Function* $h_{\mathbf{c}}(\mathbf{y})$ *is convex.*

Proof of Lemma 7. We prove this lemma by using the definition of a convex function, namely showing that $h_{\mathbf{c}}(\lambda\mathbf{y} + (1-\lambda)\mathbf{z}) \leq \lambda h_{\mathbf{c}}(\mathbf{y}) + (1-\lambda)h_{\mathbf{c}}(\mathbf{z})$, for every $\mathbf{y}, \mathbf{z}, \mathbf{c} \in \mathbb{R}^{\tilde{d}}$ and $\lambda \in [0,1]$:

$$\lambda h_{\mathbf{c}}(\mathbf{y}) + (1-\lambda)h_{\mathbf{c}}(\mathbf{z}) - h_{\mathbf{c}}(\lambda\mathbf{y} + (1-\lambda)\mathbf{z})$$

$$= \lambda \left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{y}_i\right)^2 + (1-\lambda)\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{z}_i\right)^2 - \lambda^2\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{y}_i\right)^2$$

$$- 2\lambda(1-\lambda)\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{y}_i\right)\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{z}_i\right) - (1-\lambda)^2\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{z}_i\right)^2$$

$$= \lambda(1-\lambda)\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{y}_i\right)^2 + \lambda(1-\lambda)\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{z}_i\right)^2 - 2\lambda(1-\lambda)\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{y}_i\right)\left(\sum_{i=1}^{\tilde{d}} \mathbf{c}_i\mathbf{z}_i\right)$$

$$= \lambda(1-\lambda) \left[\sum_{i=1}^{\tilde{d}} \mathbf{c}_i \mathbf{y}_i - \sum_{i=1}^{\tilde{d}} \mathbf{c}_i \mathbf{z}_i \right]^2 \geq 0,$$

where the last inequality follows since $\mathbf{y}, \mathbf{z} \in \mathbb{R}^{\tilde{d}}$ and $\lambda \in [0,1]$. $\qquad\square$

Proof of Lemma 2. First, note that $S(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A}) \leq \overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ since $xy \leq \frac{1}{2}(x^2 + y^2)$, $\forall x, y \in \mathbb{R}$.

We prove that $\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ is convex with respect to $\mathbf{A}$ by rewriting its first summation as a nonnegative weighted sum of convex quadratic functions:

$$\sum_{i,l \in [n]} \sum_{j \in [n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl} = \sum_{j \in [n]} p_j(1-p_j) \sum_{i,l \in [n]} p_{ij}p_{lj}\alpha_{ji}\alpha_{jl} = \sum_{j \in [n]} p_j(1-p_j) \left(\sum_{i \in [n]} p_{ij}\alpha_{ji} \right)^2,$$

where the equality follows since if $j \in \mathcal{N}_i \cup \{i\}$ and $j \in \mathcal{N}_l \cup \{l\}$, then $j \in \mathcal{N}_{il}$. Now, since the function $\left(\sum_{i \in [n]} p_{ij}\alpha_{ji} \right)^2$ is convex in $\mathbf{A}$ for every $\boldsymbol{P}$ (see Lemma 7), and since $p_j, p_{ij} \in [0,1]$ for every $i, j \in [n]$, and $E_{\{i,j\}} \geq p_{ij}p_{ji}$ for every $i, j \in [n]$, the function $\overline{S}(\boldsymbol{p}, \boldsymbol{P}, \mathbf{A})$ is convex. $\qquad\square$

APPENDIX E

First, observe that we can set $\alpha_{ji} = 0$ for every j such that $p_{ij}p_j = 0$. Additionally, if $\max_{k \in [n]}\{p_k p_{ik}\} = 1$, then we can set $\alpha_{ji} = \mathbb{1}_{\{p_{ij}p_j=1\}} \cdot (\sum_{k \in [n]} \mathbb{1}_{\{p_k p_{ik}=1\}})^{-1}$. Therefore, hereafter we assume that i is such that $\max_{k \in [n]}\{p_k p_{ik}\} < 1$ and that j is such that $p_{ij}p_j \in (0,1)$. We proceed to solve each of the problems (10) and (13), respectively.

A. Solving the Convex Optimization Problem (10)

The Lagrangian of (10) is

$$\overline{L}(\mathbf{A}_i^{(\ell)}, \lambda_i) = \sum_{j \in [n]} p_j p_{ij} (1 - p_j p_{ij}) \alpha_{ji}^2 + 2 \sum_{l \in [n], l \neq i} \sum_{j \in [n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl}^{(\ell-1)}$$

$$+ \sum_{j \in [n]} p_i p_j (E_{\{i,j\}} - p_{ij}p_{ji})\alpha_{ji}^2 - \lambda_i \left(\sum_{j \in [n]} p_j p_{ij}\alpha_{ji} - 1 \right) - \mu_{ji}(\alpha_{ji}).$$

Additionally,

$$\frac{\partial \overline{L}(\mathbf{A}_i^{(\ell)}, \lambda_i)}{\partial \alpha_{ji}} = 2p_j[p_{ij}(1 - p_j p_{ij}) + p_i(E_{\{i,j\}} - p_{ij}p_{ji})]\alpha_{ji}$$

$$+ 2p_j(1-p_j)p_{ij} \sum_{l \in [n]:l \neq i} p_{lj}\alpha_{jl}^{(\ell-1)} - \lambda_i p_{ij}p_j + \mu_{ji},$$

$$\frac{\partial \overline{L}(\mathbf{A}_i^{(\ell)}, \lambda_i)}{\partial \lambda_i} = 1 - \sum_{j \in [n]} p_j p_{ij}\alpha_{ji}, \qquad \frac{\partial \overline{L}(\mathbf{A}_i^{(\ell)}, \lambda_i)}{\partial \mu_{ji}} = -\alpha_{ji}.$$

It follows from the Karush–Kuhn–Tucker conditions that

$$\alpha_{ji}(\lambda_i) = \left(\frac{-2(1-p_j)\sum_{l\in[n]:l\neq i} p_{lj}\alpha_{jl}^{(\ell-1)} + \lambda_i}{2[(1-p_jp_{ij}) + p_i(E_{\{i,j\}}/p_{ij} - p_{ji})]} \right)^+,$$

and $\lambda_i \geq 0$ is set such that $\sum_{j\in[n]} p_jp_{ij}\alpha_{ji}(\lambda_i) = 1$.

B. Solving the Convex Optimization Problem (13)

The Lagrangian of (13) is

$$L(\mathbf{A}_i^{(\ell)}, \lambda_i) = \sum_{j\in[n]} p_jp_{ij}\left(1 - p_jp_{ij}\right)\alpha_{ji}^2 + 2\sum_{l\in[n], l\neq i}\sum_{j\in[n]} p_j(1-p_j)p_{ij}p_{lj}\alpha_{ji}\alpha_{jl}^{(\ell-1)}$$

$$+ 2\sum_{j\in[n]} p_ip_j(E_{\{i,j\}} - p_{ij}p_{ji})\alpha_{ji}\alpha_{ij}^{(\ell-1)} - \lambda_i\left(\sum_{j\in[n]} p_jp_{ij}\alpha_{ji} - 1\right) - \mu_{ji}(\alpha_{ji}).$$

Additionally,

$$\frac{\partial L(\mathbf{A}_i^{(\ell)}, \lambda_i)}{\partial \alpha_{ji}} = 2p_jp_{ij}\left(1 - p_jp_{ij}\right) + 2p_j(1-p_j)p_{ij}\sum_{l\in[n]:l\neq i} p_{lj}\alpha_{jl}^{(\ell-1)}$$

$$+ 2p_ip_j(E_{\{i,j\}} - p_{ij}p_{ji})\alpha_{ij}^{(\ell-1)} - \lambda_ip_{ij}p_j + \mu_{ji},$$

$$\frac{\partial L(\mathbf{A}_i^{(\ell)}, \lambda_i)}{\partial \lambda_i} = 1 - \sum_{j\in[n]} p_jp_{ij}\alpha_{ji}, \qquad \frac{\partial L(\mathbf{A}_i^{(\ell)}, \lambda_i)}{\partial \mu_{ji}} = -\alpha_{ji}.$$

It follows from the Karush–Kuhn–Tucker conditions that

$$\alpha_{ji}(\lambda_i) = \left(\frac{-2(1-p_j)\sum_{l\in[n]:l\neq i} p_{lj}\alpha_{jl}^{(\ell-1)} - 2p_i(E_{\{i,j\}}/p_{ij} - p_{ji})\alpha_{ij}^{(\ell-1)} + \lambda_i}{2\left(1 - p_jp_{ij}\right)} \right)^+,$$

and $\lambda_i \geq 0$ is set such that $\sum_{j\in[n]} p_jp_{ij}\alpha_{ji}(\lambda_i) = 1$.

REFERENCES

[1] M. Yemini, R. Saha, E. Ozfatura, D. Gündüz, and A. J. Goldsmith, "Semi-decentralized federated learning with collaborative relaying," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2022.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 54, Apr 2017, pp. 1273–1282.

[3] 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.

[4] 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.

[5] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, E. Aryafar, S. Yeh, N. Himayat, S. Andreev, and Y. Koucheryavy, "Analysis of human-body blockage in urban millimeter-wave cellular communications," in *IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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 Workshop Signal Process. Adv. Wirel. Commun*, 2021, pp. 311–315.

[13] 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.

[14] 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.

[15] 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.

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

[17] E. Ozfatura, S. Rini, and D. Gündüz, "Decentralized sgd with over-the-air computation," in *IEEE Glob. Commun. Conf*, 2020.

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

[19] 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, July 2021.

[20] K. Cohen, T. Gafni, and Y. C. Eldar, "CoBAAF: controlled bayesian air aggregation federated learning from heterogeneous data," in *58th Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2022.

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

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

[23] 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.

[24] G. Kramer, I. Marić, and R. D. Yates, "Cooperative communications," *Found. Trends Netw.*, vol. 1, no. 3, p. 271–425, Aug. 2006.

[25] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. part i. system description," *EEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, 2003.

[26] H. Y. Shutoy, D. Gunduz, E. Erkip, and Y. Wang, "Cooperative source and channel coding for wireless multimedia

communications," *IEEE J. Sel. Top. Signal Process.*, vol. 1, no. 2, pp. 295–307, 2007.

[27] D. Gunduz, E. Erkip, A. Goldsmith, and H. V. Poor, "Reliable joint source-channel cooperative transmission over relay networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2442–2458, 2013.

[28] M. Yemini, S. Gil, and A. J. Goldsmith, "Exploiting local and cloud sensor fusion in intermittently connected sensor networks," in *IEEE Glob. Commun. Conf (Globecom)*, December 2020.

[29] ——, "Cloud-cluster architecture for detection in intermittently connected sensor networks," accepted for publication in the IEEE Trans. Wirel. Commun., 2022.

[30] 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.

[31] 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," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, 2021.

[32] 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.

[33] 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.

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

[35] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optim.*, 2016.

[36] 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.

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

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

[39] 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.

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

[41] M. Yemini, A. Nedić, S. Gil, and A. J. Goldsmith, "Resilience to malicious activity in distributed optimization for cyberphysical systems," in *accepted to the 2022 IEEE Conference on Decision and Control (CDC)*, 2022.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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 J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3851–3869, 2021.

[49] 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.

[50] 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.

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

[52] 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.

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

[54] 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.

[55] 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.

[56] E. Ozfatura, D. Gündüz, and H. V. Poor, "Collaborative learning over wireless networks: An introductory overview," Machine Learning and Wireless Communications, Cambridge University Press, 2022.

[57] 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.

[58] 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.

[59] 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 Int. Workshop Signal Process. Adv. in Wireless Commun. (SPAWC)*, 2021, pp. 311–315.

[60] S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental tradeoff between computation and communication in distributed computing," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2016, pp. 1814–1818.

[61] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient coding: Avoiding stragglers in distributed learning," in *34th Int. Conf. Mach. Learn.*, ser. Proceedings of Machine Learning Research, vol. 70.   PMLR, Aug 2017, pp. 3368–3376.

[62] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1514–1529, 2018.

[63] E. Ozfatura, S. Ulukus, and D. Gündüz, "Straggler-aware distributed learning: Communication–computation latency trade-off," *Entropy*, vol. 22, no. 5, 2020.

[64] M. Mohammadi Amiri and D. Gündüz, "Computation scheduling for distributed machine learning with straggling workers," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6270–6284, 2019.

[65] S. Prakash, S. Dhakal, M. R. Akdeniz, Y. Yona, S. Talwar, S. Avestimehr, and N. Himayat, "Coded computing for low-latency federated learning over wireless edge networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 233–250, 2021.

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

[67] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[68] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Int. Conf. Artif. Intell. Stat. (AISTATS)*, August 2020.