

OCD-FL: A Novel Communication-Efficient Peer Selection-based Decentralized Federated Learning

Nizar Masmoudi and Wael Jaafar, *Senior Member, IEEE*

Abstract—The conjunction of edge intelligence and the ever-growing Internet-of-Things (IoT) network heralds a new era of collaborative machine learning, with federated learning (FL) emerging as the most prominent paradigm. With the growing interest in these learning schemes, researchers started addressing some of their most fundamental limitations. Indeed, conventional FL with a central aggregator presents a single point of failure and a network bottleneck. To bypass this issue, decentralized FL where nodes collaborate in a peer-to-peer network has been proposed. Despite the latter’s efficiency, communication costs and data heterogeneity remain key challenges in decentralized FL. In this context, we propose a novel scheme, called opportunistic communication-efficient decentralized federated learning, a.k.a., OCD-FL, consisting of a systematic FL peer selection for collaboration, aiming to achieve maximum FL knowledge gain while reducing energy consumption. Experimental results demonstrate the capability of OCD-FL to achieve similar or better performances than the fully collaborative FL, while significantly reducing consumed energy by at least 30% and up to 80%.

I. INTRODUCTION

With the increasing concerns around data privacy and continuous efforts to enhance the quality and speed of data processing, edge intelligence is becoming the new standard [1]. An ever-growing Internet-of-Things (IoT) network is laying the groundwork for a massive edge environment that will revolutionize smart devices and networks. Thus, interest in collaborative machine learning (ML) has massively increased with Google’s Federated Learning (FL) presented as one of the most promising paradigms [2]. Surveys [3] focused on investigating the FL model while highlighting its key challenges, e.g., costly communication, resource heterogeneity, and data imbalance. Others attempted to tackle these problems. For instance, Yang *et al.* proposed in [4] a resource allocation model to minimize the energy consumption of clients under a latency constraint. Zhang *et al.* presented in [5] a relay-based topology where each client serves as a relay to assist distant clients in sharing their FL models. Their approach focused on maximizing each client’s utility according to its serving role as a computational node and a relay. Furthermore, Wang *et al.* focused on counterbalancing the bias introduced by non-IID data through a deep reinforcement learning algorithm that systematically selects clients to participate in each round [6], while Han *et al.* proposed an adaptive heterogeneity-aware scheduling to mitigate resource and data heterogeneity [7].

As Beltrán *et al.* highlighted in their survey [8], the aggregation server in a centralized Federated Learning (FL) setting serves as a single point of failure and a network

bottleneck. Additionally, this centralized approach introduces security vulnerabilities, making it susceptible to network breaches and model corruption due to its status as a single point of attack. Furthermore, the aggregation server is tasked with combining the model parameters from all participating nodes, resulting in significant computational overhead. Finally, centralized FL may not be suitable for systems where components are dispersed with limited connectivity such as IoT networks, vehicular networks, and drone swarm networks. These limitations pushed the proposal of a decentralized topology where clients communicate with each other inside a peer-to-peer network. Nevertheless, communication costs and data heterogeneity persist as key constraints in decentralized federated learning (D-FL). In this context, Zheng *et al.* proposed an algorithm to balance between energy consumption and learning accuracy [9]. Li *et al.* focused on designing a robust solution in non-identically and distributed (non-IID) environments by achieving an effective clustered topology using client similarity and implementing a neighbor matching algorithm [10]. Liu *et al.* aimed to achieve a balance between communication efficiency and model consensus using multiple periodic local updates and inter-node communications [11]. Du *et al.* introduced in [12] a dynamic device scheduling mechanism that optimizes the peer selection strategy and power allocation to improve the federated edge learning model accuracy. Their approach leverages the superposition characteristics of wireless channels to enhance model training at the server and proposes a method to measure local data importance based on the gradient of local model parameters, channel conditions, and energy consumption. Simulation results show that the proposed scheduling mechanism achieves high test accuracy, fast convergence rates, and robustness against different channel conditions. In [13], Zhang *et al.* proposed a blockchain and AI-based secure cloud-edge-end collaboration scheme coupled with a blockchain-empowered federated deep actor-critic-based task offloading algorithm to tackle the secure and low-latency computation offloading problem. Finally, Xiao *et al.* addressed the time-varying dynamic network behavior by proposing a D-FL framework based on an inexact stochastic parallel random walk alternating direction method of multipliers, called ISPW-ADMM [14].

Despite the compelling results of D-FL, most published works establish their algorithms on a dense network of devices that do not consider mobility constraints. In a real-world setting, smart mobile devices, for instance, UAVs, tend to constitute a sparse graph where vertices are in continuous movement. This setting hinders model consensus across the

N. Masmoudi is with Université Paris Dauphine-PSL, Tunis Campus, Tunisia, e-mail: nizarasmoudi@outlook.fr. W. Jaafar is with the department of Software and IT Engineering, École de technologie supérieure (ÉTS) Montreal, QC, Canada, e-mail: wael.jaafar@etsmtl.ca.

entire network as each client performs a federated averaging procedure with a random fragment of the network in each FL round. Furthermore, the majority of the aforementioned papers rely on a bidirectional communication protocol between clients. This bidirectional exchange of knowledge can potentially improve the performances of one model at the expense of another, in addition to increasing communication costs.

Motivated by the aforementioned observations, we propose an opportunistic communication-efficient decentralized federated learning (OCD-FL) scheme established on a sparse network of clients who follow different motion patterns. The main contributions are summarized as follows:

- 1) Unlike previous works, we conduct our study for a sparse network of clients where each node can communicate only with its neighbors, i.e., nodes within its range of communication. Also, nodes' locations vary over time such that the neighbors of a given node change from one FL round to another.
- 2) We design a novel D-FL framework where each client makes a systematic decision to share its model with a neighbor aiming to enhance the latter's FL knowledge gain. Our approach is designed to get the maximum benefit from aggregation while saving as much energy as possible per FL client.
- 3) Using benchmark datasets under IID and non-IID scenarios, we implement our algorithm and baseline schemes, then run extensive simulations to demonstrate the efficiency of our solution compared to others, in terms of accuracy, loss, and energy. The obtained results demonstrate the high potential of OCD-FL.

The paper is organized as follows. Section II describes the system model. Section III describes D-FL. In section IV, we expose our proposed method, while section V presents the simulation results. Finally, Section VI concludes the paper.

II. SYSTEM MODEL

In this section, we present an overview of the adopted D-FL scheme by describing the network layout, the allocation strategy of data chunks across nodes, the local learning scheme, and the collaboration algorithm used in our design. Several notions are also introduced to help pave the path towards the proposed OCD-FL. Finally, a summary of the entire framework is described in Algorithm 1.

A. Network Layout

We assume an ad hoc network of N nodes. The network is represented using an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N} = \{1, 2, \dots, N\}$ is the set of nodes and \mathcal{E} is the set of edges. Note that $N \triangleq |\mathcal{N}|$ denotes the number of nodes with $|\cdot|$ being the cardinality of the set. Two nodes $(i, j) \in \mathcal{N}^2$ are connected if they are within each other's range of communication. This connection is denoted by $(i, j) \in \mathcal{E}$. We define by \mathcal{K}_i the neighborhood of node i , i.e., the set of nodes within its range of communication, particularly $\mathcal{K}_i = \{j \in \mathcal{N} \text{ s.t. } (i, j) \in \mathcal{E}\}$. We also define $K_i \triangleq |\mathcal{K}_i|$ the number of neighbors of node i . To emulate a real-world setting where nodes are mobile, the graph configuration changes at each FL round. Particularly, nodes change locations following several

patterns, and therefore each one can connect to a different set of neighbors at each FL round.

B. Communication Model

We assume that each node is equipped with a single antenna used in half-duplex mode. Moreover, the random way-point mobility model is used to represent the change in clients' locations over time [15]. At any given FL round, the wireless channel between each pair of nodes (i, k) is dominated by the Line-of-Sight (LoS) component, i.e., the channel path loss $G_{i,k}$ between node i and a neighbor k is written as (in dB)

$$G_{i,k} = 10 \log_{10} (P_{i,k}^r / P_i^t), \quad \forall i \in \mathcal{N}, \forall k \in \mathcal{K}_i, \quad (1)$$

where $P_{i,k}^r$ and P_i^t are the received power at node k and transmitted power by node i , respectively. Based on the Friis formula, the received power $P_{i,k}^r$ can be expressed by [16]

$$P_{i,k}^r = P_i^t G_i^t G_k^r (c/4\pi f)^2 (d_{i,k})^{-n}, \quad \forall i \in \mathcal{N}, \forall k \in \mathcal{K}_i \quad (2)$$

where G_i^t and G_k^r are the antenna gains of the transmitter and receiver, respectively. c denotes the speed of light, f is the signal frequency, $d_{i,k}$ is the Euclidean distance between the transmitter and receiver, and n is an environment variable.

Using the Shannon-Hartley channel capacity formula, the achievable data rate can be given by (in bits/sec) [17]

$$r_{i,k} = B_i \log_2 \left(1 + \frac{P_{i,k}^r}{N_0 B_i} \right), \quad \forall i \in \mathcal{N}, \forall k \in \mathcal{K}_i \quad (3)$$

where B_i is the allocated bandwidth and N_0 is the power of the unitary additive white Gaussian noise (AWGN) in dBm/Hz.

III. DISTRIBUTED FL: BACKGROUND

In this section, we describe the D-FL scheme, where each node may peer with its neighbors for FL aggregation.

A. Dataset Distribution

Let \mathcal{D} be the global dataset distributed across all nodes, i.e., each node $i \in \mathcal{N}$ owns a chunk of data denoted by \mathcal{D}_i such that $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_N$ and $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset \quad \forall i \neq j$. We define $D \triangleq |\mathcal{D}|$ and $D_i \triangleq |\mathcal{D}_i|, \forall i \in \mathcal{N}$. The allocation of data chunks follows a Dirichlet distribution $\text{Dir}(\alpha)$ where $\alpha > 0$ is a parameter that determines the distribution and concentration of the Dirichlet. Dirichlet distributions are commonly used as prior distributions in Bayesian statistics and constitute an appropriate choice to simulate real-world data imbalance. It allows tuning distribution imbalance levels by varying α from low values (highly unbalanced) to high values (balanced).

B. Local Update

All nodes carry the same FL model architecture. We define, by \mathbf{W}_i the model weight matrix of node i . To learn the intrinsic features of its local dataset, a node performs a local training operation. Assuming \mathbf{X}_i as the input matrix of the learning model and \mathbf{Y}_i is its target matrix, then the local optimization problem of node i is defined as follows:

$$\mathbf{W}^* = \arg \min_{\mathbf{W}_i} F(\mathbf{W}_i, \mathbf{X}_i, \mathbf{Y}_i), \quad (4)$$

where \mathbf{W}^* denotes the optimal (model weights) solution and F denotes the loss function. The complexity of machine learning models and modern datasets translates into a complex

Algorithm 1: D-FL scheme.

Input : Graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, dataset \mathcal{D} , Dirichlet parameter α , number of rounds Q .

for $i \in \mathcal{N}$ **do**

- Allocate data chunk \mathcal{D}_i to node i following Dir(α);
- Initialize model weight \mathbf{W}_i of node i ;

for $r \leftarrow 1$ **to** Q **do**

- for** $i \in \mathcal{N}$ **do**
- Move node i to a different location;
- Simultaneously perform local training (satisfy (4)) and receive models from other nodes;
- Update \mathcal{K}_i and select set of peers $\mathcal{R}_i \subseteq \mathcal{K}_i$;
- Transmit local model to peers;
- Execute federated averaging using (5);

Return $\mathbf{W}_i, \forall i \in \mathcal{N}$

shape of the loss function. With no guarantee of convexity, finding a closed-form solution to the problem (4) is usually intractable. As a result, gradient-based algorithms are used to solve it, namely Stochastic Gradient Descent (SGD) [18] and Adaptive Moment Estimation (Adam) [19].

C. Federated Averaging

Collaboration among nodes is achieved through inter-node communications. Each node i transmits its FL model to a set of neighbors, called peers, and denoted \mathcal{R}_i , such that $\mathcal{R}_i \subseteq \mathcal{K}_i$. Upon reception, each node carries out a federated averaging operation. Specifically, assuming $R_i \triangleq |\mathcal{R}_i|$ is the number of peers, federated averaging is performed as follows:

$$\mathbf{W}_i^{\text{agg}} = \frac{1}{R_i + 1} \left(\sum_{j \in \mathcal{R}_i} \mathbf{W}_j + \mathbf{W}_i \right). \quad (5)$$

OCD-FL is built on an asynchronous wireless network that leverages orthogonal frequency-division multiple access (OFDMA) to avoid the interference of concurrent transmissions [20]. This allows clients to send their updates as soon as they are ready without waiting for stragglers [21]. Nevertheless, the local aggregation has to wait for the end of the training round before aggregating the local model with the received ones. In Algo. 1, we summarize the proposed D-FL.

IV. PROPOSED OCD-FL SCHEME

The proposed OCD-FL is based on Algo. 1. However, it designs a specific peer selection mechanism that maximizes the benefit of peer-to-peer aggregation, while saving communication energy. To formulate the peer selection problem, we preliminarily define the energy consumption and knowledge gain expressions, needed for the objective design.

A. Energy Consumption

Assuming that S is the size of data a node transmits to peers, the transmission energy can be expressed by (Joules)

$$E_{i,k} = \frac{P_i^t S}{r_{i,k}} = \frac{P_i^t S}{B_i \log_2 \left(1 + \frac{P_{i,k}^r}{N_0 B_i} \right)}, \quad \forall i \in \mathcal{N}, \forall k \in \mathcal{K}_i \quad (6)$$

Energy is positive and increases with distance. Thus, assuming that the communication range is d_i^{max} ¹, the energy consumed by node i with a node located at its range edge, E_i^{max} , is

$$E_i^{\text{max}} = P_i^t S / \left(B_i \log_2 \left(1 + \frac{P_{i,\text{max}}^r}{N_0 B_i} \right) \right), \quad \forall i \in \mathcal{N}, \quad (7)$$

where $P_{i,\text{max}}^r = P_i^t G_i^t G^r \left(\frac{c}{4\pi f} \right)^2 (d_i^{\text{max}})^{-n}$, and G^r is the antenna gain of the neighbor located at distance d_i^{max} from node i . Accordingly, energy can be scaled with min-max normalization as $\tilde{E}_{i,k} = E_{i,k} / E_i^{\text{max}}, \forall i \in \mathcal{N}$, i.e., $\tilde{E}_{i,k} \in [0, 1]$.

B. Knowledge Gain

Although federated averaging remains an efficient FL collaboration method, low-performing models may negatively influence their peers and thus degrade the results of high-performing models. The latter, however, offer a good opportunity for low-performing models to progress further and improve their efficiency. This parasitic exchange between models may hinder model consensus. The following propositions highlight this phenomenon:

Proposition 1. *Let \mathbf{W}^* be the optimal solution of problem (4), while \mathbf{W}_1 and \mathbf{W}_2 are the weight matrices of two different models. For convenience, model efficiency is assumed analogous to its similarity with the optimal solution. Also, the model defined by \mathbf{W}_1 outperforms the one of \mathbf{W}_2 . Hence,*

$$\|\mathbf{W}^* - \mathbf{W}_1\| \leq \|\mathbf{W}^* - \mathbf{W}_2\|. \quad (8)$$

Now, we can deduce the following statements:

$$\|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\| \leq \|\mathbf{W}^* - \mathbf{W}_2\|, \quad (9a)$$

$$\|\mathbf{W}^* - \mathbf{W}_1\| - \frac{1}{2} \|\mathbf{W}_2 - \mathbf{W}_1\| \leq \|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\|, \quad (9b)$$

$$\|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\| \leq \|\mathbf{W}^* - \mathbf{W}_1\| + \frac{1}{2} \|\mathbf{W}_2 - \mathbf{W}_1\|, \quad (9c)$$

where $\mathbf{W}^{\text{agg}} = (\mathbf{W}_1 + \mathbf{W}_2)/2$.

Proof: Using Cauchy-Schwarz inequality and (8), the proof of (9a) is as follows:

$$\begin{aligned} \|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\| &= \left\| \mathbf{W}^* - \frac{\mathbf{W}_1 + \mathbf{W}_2}{2} \right\| = \frac{1}{2} \|2\mathbf{W}^* - \mathbf{W}_1 - \mathbf{W}_2\| \\ &\leq \frac{1}{2} (\|\mathbf{W}^* - \mathbf{W}_1\| + \|\mathbf{W}^* - \mathbf{W}_2\|) \\ &\leq \frac{1}{2} (\|\mathbf{W}^* - \mathbf{W}_2\| + \|\mathbf{W}^* - \mathbf{W}_2\|) \leq \|\mathbf{W}^* - \mathbf{W}_2\|. \end{aligned} \quad (10)$$

Since (9b) and (9c) derive from the reverse Cauchy-Schwarz inequality, their joint proof is as follows:

$$\begin{aligned} \left| \|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\| - \|\mathbf{W}^* - \mathbf{W}_1\| \right| &\leq \|\mathbf{W}^{\text{agg}} - \mathbf{W}_1\| \\ &\leq \left\| \frac{\mathbf{W}_1 + \mathbf{W}_2}{2} - \mathbf{W}_1 \right\| \leq \frac{1}{2} \|\mathbf{W}_2 - \mathbf{W}_1\|. \end{aligned} \quad (11)$$

¹The communication range ensures that a transmission from node i to node j occurs only if the distance between them $d_{i,j} \leq d_i^{\text{max}}$.

By extending (11), we obtain,

$$\begin{aligned} -\frac{1}{2}\|\mathbf{W}_2 - \mathbf{W}_1\| &\leq \|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\| - \|\mathbf{W}^* - \mathbf{W}_1\| \\ \Leftrightarrow -\frac{1}{2}\|\mathbf{W}_2 - \mathbf{W}_1\| + \|\mathbf{W}^* - \mathbf{W}_1\| &\leq \|\mathbf{W}^* - \mathbf{W}_{\text{agg}}\|, \\ \text{and} \\ \frac{1}{2}\|\mathbf{W}_2 - \mathbf{W}_1\| &\geq \|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\| - \|\mathbf{W}^* - \mathbf{W}_1\| \\ \Leftrightarrow \frac{1}{2}\|\mathbf{W}_2 - \mathbf{W}_1\| + \|\mathbf{W}^* - \mathbf{W}_1\| &\geq \|\mathbf{W}^* - \mathbf{W}^{\text{agg}}\|. \end{aligned}$$

Thus, statements (9b) and (9c) are obtained. ■

Proposition 1 confirms that low-performing models always benefit from high-performing models, while the opposite is not always true. Indeed, a low-performing model may hinder a high-performing one especially when models' dissimilarity is significant. Accordingly, we introduce a *knowledge gain* measure to identify peers with low-performing and high-performing models. The knowledge gained by k when receiving the model of node i is defined as

$$\gamma_{i,k} = \max(l_k - l_i, 0), \quad \forall i \in \mathcal{N}, \quad \forall k \in \mathcal{K}_i, \quad (12)$$

where l_k and l_i are the loss measures of k and i , respectively. $(l_k - l_i)$ is an underlying component that measures the model performance disparity between i and k . $l_k < l_i$ indicates that neighbor k 's performance outperforms that of node i , thus $\gamma_{i,k} = 0$, and no benefit is gained from peering².

Since $\gamma_{i,k}$ is computed using the loss functions, its values are unbounded. To fit within our objective, we propose an exponential normalization such that the scaled knowledge gain is $\tilde{\gamma}_{i,k} = \Gamma(\max(l_k - l_i, 0)) = 1 - \exp(-\mu \cdot \max(l_k - l_i, 0))$, where $\mu > 0$ determines the slope of exponential scaling.

C. Problem Formulation

We formulate our problem as a node-specific multi-objective optimization problem. The goal is to efficiently select neighbors for collaboration taking into account the amount of energy required for transmission as well as the knowledge gained by neighbors as a result of the collaboration. Hence, for a given node i , we state the related problem as follows:

$$\begin{aligned} \max_{w_k} \quad & \frac{\sum_{k \in \mathcal{K}_i} \sigma(w_k) \tilde{\gamma}_{i,k}}{\sum_{k \in \mathcal{K}_i} \sigma(w_k) \tilde{E}_{i,k}} + \theta \|w_k\|_2 \\ \text{s.t.} \quad & 1 \leq \sum_{k \in \mathcal{K}_i} \sigma(w_k) \leq K_i \end{aligned} \quad (13)$$

$$\text{s.t.} \quad 1 \leq \sum_{k \in \mathcal{K}_i} \sigma(w_k) \leq K_i \quad (13a)$$

where $(w_k)_{k \in \mathcal{K}_i}$ are the selection model's trainable parameters, $\sigma(\cdot)$ denotes the sigmoid function, θ is a regularization parameter and $\|\cdot\|_2$ is the Euclidean norm. In contrast to L2-regularization in ML that aims to minimize the number of parameters in a system [22], we use $\theta \|w_k\|_2$ with the maximization function to promote the selection of a higher number of D-FL neighbors by increasing the magnitude of the k^{th} model's parameters $(w_k)_{k \in \mathcal{K}_i}$. θ is a parameter that controls the strength of the regularization effect. This component is necessary as empirical studies have shown that

²While loss and accuracy are important metrics to evaluate local model training, knowledge gain, defined with the loss metric, is used with energy consumption as criteria to dictate the peer selection and aggregation strategy.

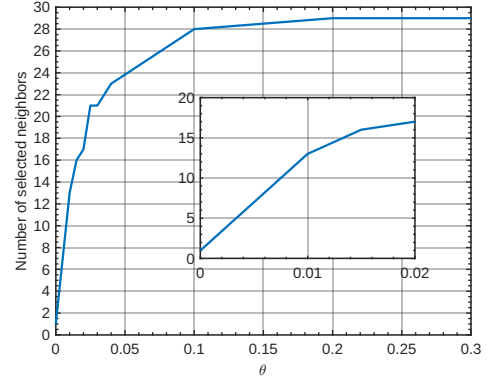


Fig. 1. Effect of regularization on neighbor selection rate.

without regularization, the problem is reduced to selecting a single neighbor, enough to avoid an indeterminate form of the objective function, while obtaining a high knowledge gain-to-energy ratio. This hinders the local model's ability to generalize. In Fig. 1, we plot the number of selected neighbors as a function of θ for a specific D-FL node with 30 neighbors, generated using random values for knowledge gain and energy consumption. Results demonstrate the importance of the regularization term to avoid selecting a single neighbor. Constraint (13a) guarantees that at least one neighbor is selected to avoid an indeterminate form while asserting that at most all neighbors are selected. For the sake of simplicity, we define by $\beta_k = \sigma(w_k)$ the probability that neighbor k is selected as a peer. The objective is to learn $(w_k)_{k \in \mathcal{K}_i}$ for each node i and, according to its $(\beta_k)_{k \in \mathcal{K}_i}$, decides under a certainty threshold the neighbors that will be peered. Since the objective function of (13) is not concave, it cannot be solved directly. Nevertheless, (13) is differentiable, thus rendering its resolution with gradient-based algorithms feasible.

V. SIMULATION EXPERIMENTS AND RESULTS

A. Simulation Setup

OCD-FL is implemented using Torch inside a Python environment. We adopt a sparse topology, where $N = 20$ nodes are randomly placed on a 2-dimensional bounded rectangular surface. For each node i , P_i^t and B_i are uniformly distributed in the intervals $[10, 21]$ dBm and $[5, 20]$ MHz, respectively, $\forall i \in \mathcal{N}$. Antenna gains are $G_i^t = G_i^r = 0$ dBi, $\forall i \in \mathcal{N}$. The signal frequency $f = 1$ GHz, $c = 3 \cdot 10^8$ m/sec, $d_i^{\max} = 2$ km, $\forall i \in \mathcal{N}$, the size of data is $S = 87$ Kbits (MNIST) and $S = 23$ Mbits (CIFAR-10), $n = 2$ (suburban), and $\mu = 2$.

Our experiments are performed on two different datasets, MNIST and CIFAR-10 [23], [24]. Although both datasets consist of 60,000 training samples and 10,000 test samples, MNIST has square images ($28 \times 28 \times 1$ pixels) of handwritten digits (from 0 to 9), while CIFAR-10 contains colored square images ($32 \times 32 \times 3$ pixels) of 10 different object classes. Training dataset is split into equally-sized $N = 20$ subsets following Dirichlet distribution with $\alpha = 1$ (non-IID scenario, i.e., number of samples in classes are significantly unbalanced), and $\alpha = 100$ (IID scenario, i.e., number of samples

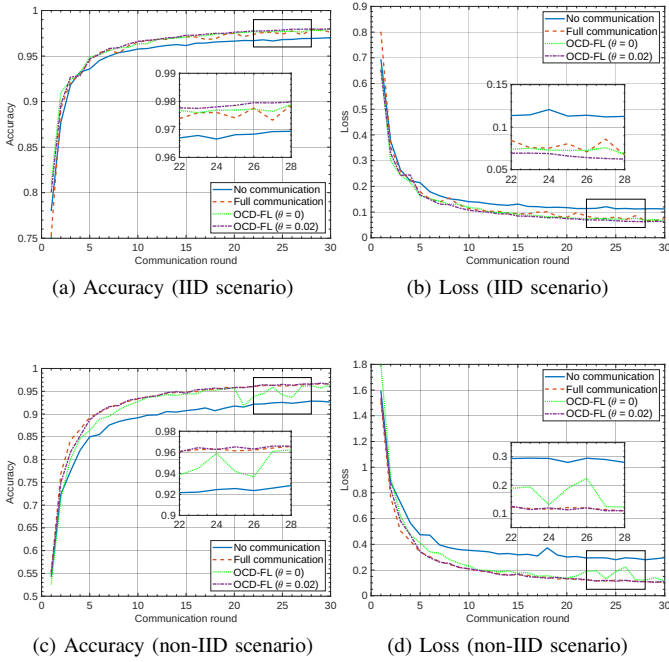


Fig. 2. Avg. accuracy and loss (MNIST, different schemes).

per class are approximately equal). In any case, the testing subset is IID to avoid non-biased evaluations.³

Different models are implemented to fit adequately each dataset. For MNIST, we used a model with 2.5×5 convolutional neural network (CNN) hidden layers and ReLU activation functions, for a total number of parameters of 21,840. For CIFAR-10, the model is more complex with six 3×3 CNN layers, for a total number of parameters of 5,852,234. During local updates, an NVIDIA Tesla T4 processing unit, along with a CUDA environment, was used to speed up computations.

B. Simulation Results

OCD-FL is evaluated against baseline schemes “No communication” where each client trains exclusively locally, and “Full communication” where any node communicates with all its neighbors. Note that we considered “OCD-FL ($\theta = 0$)” since it is analogous to the lower case of (13) where $\sum_{k \in \mathcal{K}_i} \sigma(w_k) = 1$, while “Full communication” reflects the upper case of (13) where $\sum_{k \in \mathcal{K}_i} \sigma(w_k) = K_i$.

In Fig. 2, we illustrate the FL performances, in terms of accuracy and loss, for the proposed method when applied to the MNIST dataset, and compared to the benchmarks, under IID and non-IID scenarios. For the IID scenario, the proposed OCD-FL method ($\theta = 0$ and $\theta = 0.02$) outperforms all benchmarks. Indeed, by setting θ to 0.02, we introduce a regularization term that promotes collaboration with a wider range of neighbors. Such results highlight the importance of controlled collaboration between nodes to achieve consensus on efficient models. In the meanwhile, the “No communication” scheme provides the worst results. This is expected since each client relies only on its knowledge for training. In

³With a non-IID testing dataset, overfitting is a prominent risk that yields biased evaluations of local models, thus hindering the performance of D-FL.

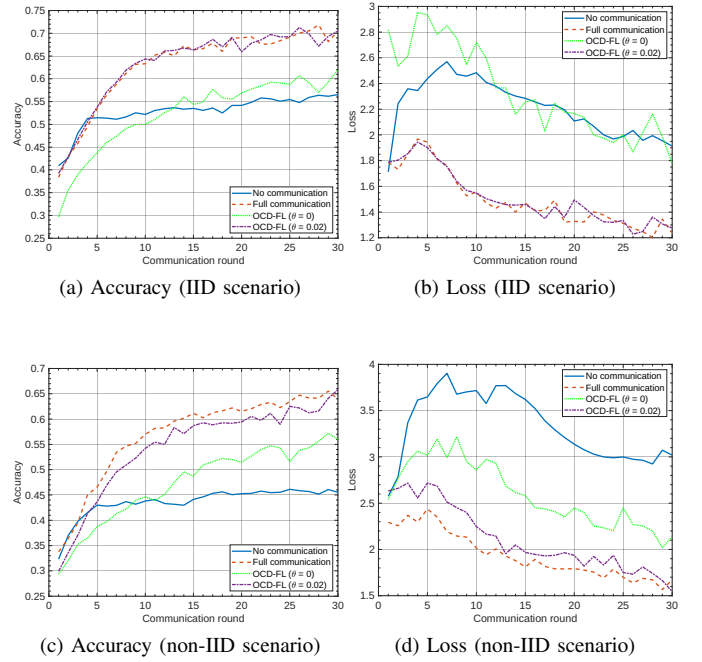


Fig. 3. Avg. accuracy and loss (CIFAR-10, different schemes).

the non-IID scenario, the performance gap between the “No collaboration” scheme and the other ones increases. This is due to the higher complexity of training in the non-IID setting. Moreover, the proposed OCD-FL method ($\theta = 0.02$) still outperforms all other methods, while the performance of OCD-FL ($\theta = 0$) degrades below that of “Full communication”. Indeed, since $\theta = 0$, regularization is eliminated. Hence, our scheme limits its peer selection for each node to a small number, which may not be sufficient to train efficiently in the non-IID scenario. Indeed, our scheme struggles to achieve a consensus on an efficient model, and the instability of the associated learning curve highlights the network’s inability to generalize. Note that the initial increasing trend in the loss curve is due to gradient instability during the first few rounds. The increase is less significant under “Full Communication” and “OCD-FL ($\theta = 0.02$)” since efficient model aggregation contributes to faster gradient stability. This increase is not observed on the MNIST simpler dataset, where it only takes the gradients on a small number of training rounds to stabilize.

Fig. 3 presents the same results as in Fig. 2, but for the CIFAR-10 dataset. As it can be seen, for the IID scenario, “OCD-FL ($\theta = 0.02$)” is capable of providing similar performances, in terms of accuracy and loss, to “Full communication”, while the gap with “OCD-FL ($\theta = 0$)” and “No communication” is very significant. For instance, after 20 rounds, the gap in accuracy is approximately 10%. In the non-IID scenario, “Full communication” presents the best performances, while “OCD-FL ($\theta = 0.02$)” falls slightly behind, by about 2% in terms of accuracy. “OCD-FL ($\theta = 0$)”, although not the best scheme, is still significantly outperforming “No communication”. Note that, even though our scheme is not the best in CIFAR-10 with non-IID, an optimal θ might be determined, which would provide very close performances to

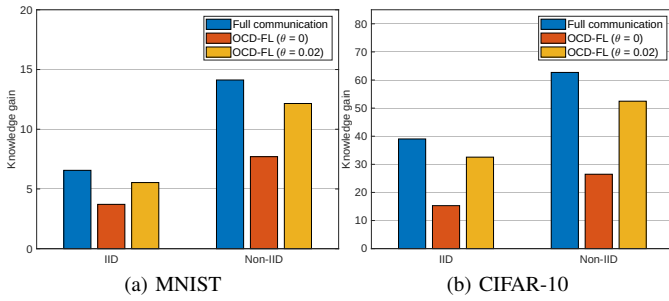


Fig. 4. Knowledge gain (different schemes and scenarios).

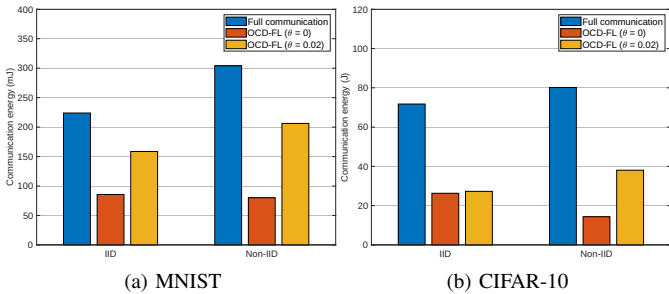


Fig. 5. Consumed communication energy (different schemes and scenarios).

the “Full communication” scheme.

Fig. 4 presents the average knowledge gain under different scenarios. Although OCD-FL ($\theta = 0$) fails to promote knowledge sharing between clients, it manages to compete with “Full Communication” when applied with regularization $\theta = 0.02$, in IID and non-IID settings. With non-IIDness, the resulting knowledge gain is significantly high. This is due to the model performance disparity between clients.

Similarly, in Fig. 5, we depict the communication energy consumed by each system with OCD-FL or “Full communication” in IID and non-IID scenarios, and for MNIST and CIFAR-10 datasets. “Full communication” consumed the highest amounts of energy in any setting, since it relies on communications between all N clients. In contrast, our OCD-FL scheme consumes less energy between 30% and 80% than “Full communication”. This is mainly due to the accurate selection of peers for model sharing.

VI. CONCLUSION

In this paper, we proposed a novel distributed FL scheme, called OCD-FL. The latter systematically selects neighbors for peer-to-peer FL collaboration. Our solution incorporates a trade-off between knowledge gain and energy efficiency. To do so, the developed peer selection strategy was assimilated into a regularized multi-objective optimization problem aiming to maximize knowledge gain while consuming minimum energy. The OCD-FL method was evaluated in terms of FL accuracy, loss, and energy consumption, and compared against baselines and under several scenarios. OCD-FL proved its capability to achieve consensus on an efficient FL model while significantly reducing communication energy consumption between 30% and 80%, compared to the best benchmark. Although we conducted a comprehensive evaluation of OCD-FL, several

aspects of the network’s layout present interesting research opportunities, such as the impact of a time-varying topology on model convergence. Moreover, despite the adoption of federated averaging in our work, this research serves as a proof of concept and lays the groundwork for future exploration of other distributed FL systems, where different FL aggregation techniques might be experimented.

REFERENCES

- [1] D. Xu *et al.*, “Edge intelligence: Empowering intelligence to the edge of network,” *Proc. IEEE*, vol. 109, no. 11, pp. 1778–1837, Nov. 2021.
- [2] Q. Li *et al.*, “A survey on federated learning systems: Vision, hype and reality for data privacy and protection,” *IEEE Trans. Knowl. Data Engineer.*, vol. 35, no. 4, pp. 3347–3366, 2023.
- [3] L. Li *et al.*, “A survey on federated learning,” in *Proc. IEEE Int. Conf. Control & Autom. (ICCA)*, 2020, pp. 791–796.
- [4] Z. Yang *et al.*, “Energy efficient federated learning over wireless communication networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, 2021.
- [5] X. Zhang *et al.*, “Energy efficient federated learning over cooperative relay-assisted wireless networks,” in *Proc. IEEE Glob. Commun. Conf.*, 2022, pp. 179–184.
- [6] H. Wang *et al.*, “Optimizing federated learning on non-IID data with reinforcement learning,” in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 1698–1707.
- [7] J. Han *et al.*, “Heterogeneity-aware adaptive federated learning scheduling,” in *Proc. IEEE Int. Conf. Big Data*, 2022, pp. 911–920.
- [8] E. T. M. Belrán *et al.*, “Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges,” *IEEE Commun. Surv. Tuts.*, pp. 1–1, 2023.
- [9] J. Zheng *et al.*, “Federated learning for energy-balanced client selection in mobile edge computing,” in *Proc. Int. Wireless Commun. Mob. Comput. (IWCMC)*, 2021, pp. 1942–1947.
- [10] Z. Li *et al.*, “Towards effective clustered federated learning: A peer-to-peer framework with adaptive neighbor matching,” *IEEE Trans. Big Data*, pp. 1–16, 2022.
- [11] W. Liu *et al.*, “Decentralized federated learning: Balancing communication and computing costs,” *IEEE Trans. Sig. Info. Process. Netw.*, vol. 8, pp. 131–143, 2022.
- [12] J. Du *et al.*, “Gradient and channel aware dynamic scheduling for over-the-air computation in federated edge learning systems,” *IEEE J. Sel. Ar. Commun.*, vol. 41, no. 4, pp. 1035–1050, Apr. 2023.
- [13] S. Zhang *et al.*, “Blockchain and federated deep reinforcement learning based secure cloud-edge-end collaboration in power IoT,” *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 84–91, 2022.
- [14] Y. Xiao *et al.*, “Fully decentralized federated learning-based on-board mission for UAV swarm system,” *IEEE Commun. Lett.*, vol. 25, no. 10, pp. 3296–3300, 2021.
- [15] D. B. Johnson and D. A. Maltz, *Dynamic Source Routing in Ad Hoc Wireless Networks*. Boston, MA: Springer US, 1996, pp. 153–181.
- [16] H. Friis, “A note on a simple transmission formula,” *Proc. of IRE*, vol. 34, no. 5, pp. 254–256, May 1946.
- [17] A. I. Pérez-Neira and M. R. Campalans, “Chapter 2 - different views of spectral efficiency,” in *Cross-Layer Resource Allocation in Wireless Communications*, A. I. Pérez-Neira and M. R. Campalans, Eds. Oxford: Academic Press, 2009, pp. 13–33.
- [18] S. Ruder, “An overview of gradient descent optimization algorithms,” *CoRR*, vol. abs/1609.04747, 2016. [Online]. Available: <http://arxiv.org/abs/1609.04747>
- [19] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.
- [20] A. F. M. S. Shah *et al.*, “Survey and performance evaluation of multiple access schemes for next-generation wireless communication systems,” *IEEE Access*, vol. 9, pp. 113428–113442, Aug. 2021.
- [21] T. Li *et al.*, “Federated learning: Challenges, methods, and future directions,” *IEEE Sig. Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.
- [22] C. Cortes *et al.*, “L2 regularization for learning kernels,” 2012. [Online]. Available: <https://arxiv.org/abs/1205.2653>
- [23] L. Deng, “The MNIST database of handwritten digit images for machine learning research,” *IEEE Sig. Process. Mag.*, vol. 29, no. 6, pp. 141–142, 2012.
- [24] A. Krizhevsky, “Learning multiple layers of features from tiny images,” *University of Toronto*, 05 2012.