Cost-Effective Federated Learning: A Unified Approach to Device and Training Scheduling

Xiaobing Chen, Xiangwei Zhou, Hongchao Zhang, Mingxuan Sun, and Taibiao Zhao

***Abstract*—Federated learning enables decentralized model training across numerous devices without data centralization, leveraging model updates to enhance privacy and reduce communication overhead. Despite its advantages, federated learning systems must be optimized for cost efficiency, considering the limited computational capabilities and battery life of edge devices. Current research often focuses on minimizing either time or energy costs but rarely both, and does not jointly optimize the parameters of device and training scheduling in the presence of system and data heterogeneity. In our paper, we formulate a novel joint optimization problem for device and training scheduling that minimizes the total cost of federated learning while ensuring model convergence. We propose a new device scheduling scheme, Group Scheduling on Orthogonal Frequency-Division Multiple Access (GS-OFDMA), to improve time efficiency and develop an iterative algorithm to tackle the resulting mixed integer nonlinear programming problem. Our experimental results show that our approach significantly reduces the total cost by at least 35% across different real-world datasets and data distributions in comparison with random participant selection.**

I. Introduction

Traditional machine learning models, which centralize vast data volumes for training, face hurdles in scenarios where large-scale data aggregation is impractical or sensitive, especially due to limited bandwidth and strict data privacy laws. To address the limitations of centralized learning, federated learning presents a decentralized framework that enables model training across multiple edge devices without collecting the raw data. Interchanging only the model updates between the server and clients not only greatly reduces the communication overhead but also effectively avoids the private data leakage [1].

In a federated learning system, the cost efficiency during training is important in evaluating the utility of the system design [2]. Empirical methods like [3, 4] track client training metrics, such as loss and time cost, to guide participant selection. However, these heuristic approaches often lack theoretical guarantees for convergence.

Research in federated learning increasingly focuses on cost minimization, aiming to reduce time [5, 6] or energy costs [7, 8] through various control variables and constraints. These control variables fall into two categories: device scheduling [5, 6, 9] and training scheduling [10–12]. Device scheduling involves selecting participants for each round and organizing model transmissions, enhancing model convergence through data diversity and reducing costs by excluding inefficient participants. Training scheduling involves configuring the training process, such as the number of participants per round [11], the number of local iterations [10], and the number of communication rounds between clients and the server. Without proper training scheduling, issues such as client shift can compromise model convergence and increase resource usage.

However, current studies aimed at optimizing the cost efficiency of federated learning systems presents several notable limitations. First, a significant portion of these studies tend to focus singularly on either reducing time cost [5, 9, 13] or energy consumption [7, 8], but not both. Second, a comprehensive optimization that jointly considers both device scheduling and training scheduling is often absent [10, 11]. Last, in the problem formulation, system heterogeneity and data heterogeneity are not jointly considered [13, 14].

To address the limitations inherent in current research, we present a novel joint optimization problem on both device and training scheduling to minimize the training cost of federated learning with convergence guarantee. The main contributions of our paper are as follows:

1) We formulate a new cost-minimization problem to jointly optimize the parameters of device scheduling and training scheduling. To make the problem practical, both system and data heterogeneity are factored in.
2) To improve the time efficiency, we propose a novel device scheduling scheme: Group Scheduling on Orthogonal Frequency-Division Multiple Access (GS-OFDMA). By ordering and grouping participants by their time cost per round, GS-OFDMA allows more participants in one round to find the optimal number of participants.
3) We derive the analytical expression for the overall training cost of federated learning with probabilistic participant selection. We develop an iterative algorithm integrating coordinate descent and polyhedral active set algorithm (PASA) to solve the challenging mixed integer nonlinear programming (MINLP) problem.
4) Through experiments on real datasets, we evaluate the total cost and convergence speed of the model training configured based on the solution to our optimization. The proposed integrative algorithm shows near optimal performance on total cost and convergence speed.

II. System Model

A. Federated Learning with Client-Selection Probability

Assume that there are one server and N clients with index set $\mathcal{N} = \{1, 2, ..., N\}$ in a federated learning system. Client

This paper was supported in part by the National Science Foundation under Grants 1943486, 1946231, 2110722, 2246757, and 2332011.

X. Chen, X. Zhou, H. Zhang, M. Sun, and T. Zhao are with Louisiana State University, Baton Rouge, LA, 70803, USA (e-mail: xwzhou@lsu.edu).

i possesses a unique and private dataset $\mathcal{D}_i = \{\xi_j^i \mid j = 1, 2, ..., |\mathcal{D}_i|\}$ of size $|\mathcal{D}i|$, where ξ_j^i denotes the j-th data sample of client i. The composite dataset across all clients is represented as $\mathcal{D} = \bigcup_{i\in\mathcal{N}} \mathcal{D}_i$ and of size $|\mathcal{D}|$. The local loss function on dataset $\mathcal{D}_i$ is defined by

$$F_i(x) \coloneqq \frac{1}{|\mathcal{D}_i|} \sum_{\xi_j^i \in \mathcal{D}_i} F_i(x; \xi_j^i), \tag{1}$$

where x represents the model parameters.

The objective of a federated learning system is to find an optimal model parameter x to minimize a global loss function $f(x)$ on all the distributed datasets, defined by

$$\min_x f(x) \coloneqq \sum_{i=1}^{N} d_i F_i(x), \tag{2}$$

where $d_i = |\mathcal{D}_i|/|\mathcal{D}|$ denotes the ratio of data volume at client i and $\sum_{i=1}^{N} d_i = 1$.

FedAvg [1] addresses data privacy in federated learning by training locally and transmitting only model updates from clients to the server. To reduce communication overhead, FedAvg employs multi-epoch local training and selects a random subset of participants for each training round.

We extend this approach by introducing probabilistic participant selection in federated learning. This method considers T communication rounds between participants and the server to achieve model convergence. The stopping criterion for these rounds is defined as: $\mathbb{E}(f(x_T)) - f^* \leq \epsilon$, where f^* denotes the minimum global loss and $\epsilon > 0$.

In each round, a training process consists of model broadcasting, local training, model transmission, and aggregation. Initially, the server chooses M clients to participate in the t-th training round, forming participant set $\mathcal{M}^{(t)}$ based on the participant selection probability $\mathbf{p} = [p_1, p_2, ..., p_N]$, and sends out the global model parameters x_t to these participants. Each participant then synchronizes its local model with the global one, trains the model using stochastic gradient descent (SGD) algorithm, updates the model over I iterations, and transmits the updated model parameters back to the server. Finally, the server aggregates all received updates to renew the global model by

$$x_{t+1} = \sum_{i\in\mathcal{M}^{(t)}} \frac{d_i}{Mp_i} y_{t,I}^i. \tag{3}$$

B. Device Scheduling Scheme

Effective participation scheduling is crucial for time efficiency in federated learning, influenced by the wireless transmission model, which includes time-sharing [2, 11] and frequency-sharing protocols [8, 15]. In mobile network scenarios with variable participant resources, we introduce a novel scheduling approach based on Orthogonal Frequency-Division Multiple Access (OFDMA), termed Group Scheduling on OFDMA protocol (GS-OFDMA). GS-OFDMA's key innovation lies in allowing more participants within the same bandwidth per round, achieved through group submissions. This strategy aims to reduce stochastic variance and enhance model convergence speed.

Specifically, suppose that the total bandwidth in OFDMA is divided into S sub-channels, where each sub-channel is exclusively used by one participant. Unlike existing work [2, 8, 16] with OFDMA that only samples S participants per round, GS-OFDMA selects $M = KS$ participants according to participation probability $\mathbf{p}$, where K is an integer and $1 \leq K \leq \lceil \frac{N}{S} \rceil$. Let $\delta_{i,l}$ denote the one-iteration training time of client i in each round and $\delta_{i,a}^{(t)}$ be the model transmission time of client i in round t. GS-OFDMA is described by the following protocols.

1) The server estimates $\{\delta_{i,l} : i = 1, 2, ..., KS\}$ and sorts them in terms of local computing time such that $\delta_{1,l} \leq \delta_{2,l} \leq ... \leq \delta_{M,l}$.
2) Participants are divided into K groups according to $\{\delta_{i,l}\}$, i.e., $G_i = \{(i-1)S+1, (i-1)S+2, ..., iS\}, i = 1, 2, ..., K$.
3) Participants in one group submit their updates simultaneously. Model transmission proceeds from G_1 to G_K and the next group starts transmitting only if all the updates of the current group are received by the server.
4) Model transmission of this round ends after K groups finish uploading updates.

C. Cost Evaluation of Federated Learning

In terms of cost, time cost and energy cost are crucial for efficient federated learning system design. Similar to existing work in wireless federated learning [11], we assume that mobile devices have relatively stable computation capacity but face dynamic communication environments.

1) Time Cost: Let $[S, 2S, ..., KS]$ be the indexes of straggler in each group in terms of communication time. We assume that all the participants of a group can finish the local training before the previous group finishes the model transmission.

According to the proposed GS-OFDMA scheduling, $\delta^{(t)}$ can be calculated by the summation of the longest computation time in the first group and the longest communication time of each group.

$$\delta^{(t)} = \delta_{S,l} I + \sum_{k=1}^{K} \delta_{kS,a}^{(t)}. \tag{4}$$

As a result, the total time cost of T global rounds can be calculated as

$$\Delta = \sum_{t=1}^{T} (\delta_{S,l} I + \sum_{k=1}^{K} \delta_{kS,a}^{(t)}). \tag{5}$$

2) Energy Cost: Similar to the time cost, let $e_{i,l}$ be the one-iteration energy cost of client i in each round and $e_{i,a}^{(t)}$ be energy cost of model transmission in round t. The energy cost of client i in round t is given by

$$e_i^{(t)} = e_{i,l} I + e_{i,a}^{(t)}. \tag{6}$$

Since device scheduling schemes do not change the total energy cost, the cumulative energy cost of T rounds can be

expressed as

$$e = \sum_{t=1}^{T} \sum_{i \in \mathcal{M}^{(t)}} e_i^{(t)}. \tag{7}$$

D. Problem Formulation

Cost evaluation reveals that the efficiency of federated learning systems is influenced by several key factors: the participation selection probability $\mathbf{p}$, number of global rounds T, number of scheduled groups K in GS-OFDMA, and number of local iterations I. Here, $\mathbf{p}$ is the parameter for device scheduling and (K, I, T) are for training scheduling. We propose to integrate the time and energy costs of federated learning into one adjusted cost C_T, given by

$$C_T = \alpha\Delta + (1-\alpha)e, \tag{8}$$

where $0 \leq \alpha \leq 1$ represents the relative importance of each cost and provides flexibility to the cost metric.

The objective is to minimize the adjusted cost of federated learning training, while satisfying the model convergence requirement. Therefore, we formulate the joint device and training scheduling optimization problem as follows

$$\min_{\mathbf{p},T,K,I} \quad \mathbb{E}(C_T) \tag{P1}$$

$$\text{s.t.} \quad \mathbb{E}(f(x_T)) - f^* \leq \epsilon, \tag{P1a}$$

$$\sum_{i=1}^{N} p_i = 1, \tag{P1b}$$

$$1 \leq K \leq \lceil \frac{N}{S} \rceil, K \in \mathbb{Z}^+, \tag{P1c}$$

$$1 \leq T \leq T_{max}, 1 \leq I \leq I_{max}, T, I \in \mathbb{Z}^+, \tag{P1d}$$

where T_{max} and I_{max} denote the maximum numbers of global rounds and local iterations, respectively.

III. Joint Device and Training Scheduling Optimization

In this section, we transform problem P1 into a more tractable form and develop an iterative algorithm to find the optimal solutions.

A. Analytical Expression of $\mathbb{E}(C_T)$

The probability of client i participating in round t can be given by

$$P(i \in \mathcal{M}^{(t)}) = 1 - (1-p_i)^M \overset{(a)}{\approx} Mp_i, \tag{9}$$

where (a) approximates $(1-p_i)^M$ by its zero and first order terms.

Even though client i could be selected multiple times in one round, the energy cost will be counted only once (one-time local training and one-time model transmission). Therefore, we have

$$\begin{aligned}\mathbb{E}(e) &= \mathbb{E}\left(\sum_{t=1}^{T} \sum_{i \in \mathcal{M}^{(t)}} e_i^{(t)}\right) = \sum_{t=1}^{T}\sum_{i=1}^{N} P(i \in \mathcal{M}^{(t)}) e_i^{(t)} \\ &\approx \sum_{t=1}^{T}\sum_{i=1}^{N} Mp_i e_i^{(t)} = KTS\sum_{i=1}^{N} p_i(e_{i,l}I + \bar{e}_{i,a}),\end{aligned} \tag{10}$$

where the average energy cost of model transmission of client i is denoted by $\bar{e}_{i,a} = \frac{1}{T}\sum_{t=1}^{T} e_{i,a}^{(t)}$.

For $\mathbb{E}(\Delta)$, we have

$$\mathbb{E}(\Delta) = \mathbb{E}\left(\sum_{t=1}^{T}\left(\delta_{S,l}I + \sum_{k=1}^{K}\delta_{kS,a}^{(t)}\right)\right). \tag{11}$$

To find the expectation of the first term in (11) which denotes the longest local training time, we define q_i to be the probability of client i being selected in round t and with the longest training time of the first group, which is equivalent to that only clients $1, 2, ..., i$ are candidates in the first group. Therefore, we have

$$\begin{aligned} q_i &= \mathbb{P}(i \text{ has the longest training time of the first group}) \\ &= \sum_{m=1}^{S}\binom{S}{m} p_i^m \left(\sum_{j=1}^{i-1} p_j\right)^{S-m} \\ &= \left(\sum_{j=1}^{i} p_j\right)^S - \left(\sum_{j=1}^{i-1} p_j\right)^S, \end{aligned} \tag{12}$$

where the last equality uses the binomial theorem. Then the first term in (11) becomes

$$\mathbb{E}\left(\sum_{t=1}^{T}\delta_{S,l}I\right) = \sum_{t=1}^{T}\sum_{i=1}^{N} q_i\delta_{i,l}I. \tag{13}$$

The second term $\{\delta_{kS,a}^{(t)} : k = 1, ..., K\}$ in (11) denotes the largest model transmission time in each scheduling group and can vary in different rounds. Since the model transmission time of each group is independent, we approximate the transmission time of each group by the first group and have

$$\mathbb{E}\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{kS,a}^{(t)}\right) = \sum_{t=1}^{T} K\sum_{i=1}^{N} q_i\delta_{i,a}^{(t)}. \tag{14}$$

With (13) and (14), we have

$$\mathbb{E}(\Delta) = T\sum_{i=1}^{N} q_i\left(\delta_{i,l}I + K\bar{\delta}_{i,a}\right), \tag{15}$$

where $\bar{\delta}_{i,a} = \frac{1}{T}\sum_{t=1}^{T}\delta_{i,a}^{(t)}$ denotes the average time cost of model transmission of client i.

The above formulation is still very hard to optimize because q_i includes the polynomial term of p_i with the order of S. For analytical tractability, we approximate q_i with $q_i \approx p_i$. Note that there are two cases where p_i is equivalent to q_i in terms of $\mathbb{E}(\Delta)$. 1) When $S = 1$, we can easily show that $q_i = p_i$. 2) When $\delta_{i,l} = \delta_{j,l}$ and $\bar{\delta}_{i,a} = \bar{\delta}_{j,a}$ for $i \neq j$, we have

$$\sum_{i=1}^{N} q_i\left(\delta_{i,l}I + K\bar{\delta}_{i,a}\right) = \sum_{i=1}^{N} p_i\left(\delta_{i,l}I + K\bar{\delta}_{i,a}\right).$$

Therefore, we formally define the approximated $\tilde{\mathbb{E}}(\Delta)$ as

$$\tilde{\mathbb{E}}(\Delta) = T\sum_{i=1}^{N} p_i\left(\delta_{i,l}I + K\bar{\delta}_{i,a}\right). \tag{16}$$

With the analytical expressions of $\mathbb{E}(e)$ and $\tilde{\mathbb{E}}(\Delta)$, the

approximation of total cost $\tilde{\mathbb{E}}(C_T)$ is

$$\begin{aligned}\tilde{\mathbb{E}}(C_T) &= \alpha\tilde{\mathbb{E}}(\Delta) + (1-\alpha)\mathbb{E}(e) \\ &= T\sum_{i=1}^{N} p_i(Iw_{i,l}(K) + w_{i,a}(K)),\end{aligned} \tag{17}$$

where $w_{i,l}(K)$ and $w_{i,a}(K)$ are functions of K defined as follows:

$$\begin{aligned}w_{i,l}(K) &= \alpha\delta_{i,l} + (1-\alpha)KSe_{i,l}, \\ w_{i,a}(K) &= K\left(\alpha\bar{\delta}_{i,a} + (1-\alpha)S\bar{e}_{i,a}\right).\end{aligned} \tag{18}$$

Here, $w_{i,l}(K)$ and $w_{i,a}(K)$ can also be regarded as "pseudo costs" of local training and model transmission of client i, respectively.

B. Approximation Optimization Problem of P1

To approximate the ϵ-convergence constraint, we utilize the convergence result [16]:

$$\mathbb{E}(f(x_T)) - f^* \leq \frac{1}{T}\left[AI\left(\frac{1}{KS}\sum_{i=1}^{N}\frac{C_i}{p_i} + D\right) + \frac{B}{I}\right], \tag{19}$$

where A, B, $\{C_i\}$ and D are constants related to the local loss functions and data heterogeneity among clients. To reformulate problem P1, we use (19) to replace the convergence constraint and (17) as the objective function. We have

$$\min_{\mathbf{p},K,T,I} \quad T\sum_{i=1}^{N} p_i(Iw_{i,l}(K) + w_{i,a}(K)) \tag{P2}$$

$$\text{s.t.} \quad \frac{1}{T}\left[AI\left(\frac{1}{KS}\sum_{i=1}^{N}\frac{C_i}{p_i} + D\right) + \frac{B}{I}\right] \leq \epsilon, \tag{P2a}$$

$$\text{(P1b)}, \text{(P1c)}, \text{(P1d)}.$$

C. Solution to P2

With continuous ($\mathbf{p}$) and integer (K, T, I) variables, P2 is a Mixed Integer Nonlinear Programming (MINLP) problem, which is difficult to solve directly in general. We present an iterative algorithm to effectively solve problem P2.

We first apply linear programming relaxation to convert integer variables K, T, I to be continuous variables. Observing T in the objective of P2 and constraint (P2a), we have that the optimal T^* minimizing the objective of P2 also maximizes the LHS of constraint (P2a). T^* satisfies

$$T^* = \frac{1}{\epsilon}\left[AI\left(\frac{1}{KS}\sum_{i=1}^{N}\frac{C_i}{p_i} + D\right) + \frac{B}{I}\right]. \tag{20}$$

Then problem P2 is converted to

$$\begin{aligned}\min_{\mathbf{p},K,I} \quad & \left(\frac{AI}{KS}\sum_{i=1}^{N}\frac{C_i}{p_i} + \frac{AI^2D + B}{I}\right) \\ & \cdot\left(\sum_{i=1}^{N} p_i(Iw_{i,l}(K) + w_{i,a}(K))\right)\end{aligned} \tag{P3}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} p_i = 1, \tag{P3a}$$

$$1 \leq K \leq N/S, \tag{P3b}$$

$$1 \leq I \leq I_{max}. \tag{P3c}$$

Achieving the global minimizer of Problem P3 is challenging due to its non-convex nature. Notably, the variables $(K, I, \mathbf{p})$ are independent within the constraints, with $\mathbf{p}$ impacting only the coefficient terms of K and I. Therefore, we decompose Problem P3 into two sub-problems and effectively solve the sub-problems to obtain the suboptimal solution to Problem P3.

Given any $\mathbf{p}$, the first sub-problem with (K, I) is

$$\min_{K,I} \quad \left(\frac{a_0 I}{K} + \frac{a_1 I^2 + B}{I}\right)(a_2 KI + a_3 K + a_4 I) \tag{P4}$$

$$\text{(P3b)}, \text{(P3c)}, \tag{21}$$

where $a_0 = \frac{A}{S}\sum_{i=1}^{N}\frac{C_i}{p_i}, a_1 = AD, a_2 = (1-\alpha)S\sum_{i=1}^{N} p_i e_{i,l}, a_3 = \sum_{i=1}^{N} p_i(\alpha\bar{\delta}_{i,a} + (1-\alpha)S\bar{e}_{i,a})$ and $a_4 = \alpha\sum_{i=1}^{N} p_i\delta_{i,l}$ are positive constants.

We present the following theorem for Problem P4.

Theorem 1. *Given positive constants a_0, a_1, a_2, a_3, a_4, suppose a set $H = [1, N/S] \times [1, I_{max}]$ and a function $h : H \to \mathbb{R}$ is defined as*

$$h(K, I) = \left(\frac{a_0 I}{K} + \frac{a_1 I^2 + B}{I}\right)(a_2 KI + a_3 K + a_4 I).$$

Then H is a biconvex set and h is a biconvex function on H. Therefore, Problem P4 is a biconvex optimization problem.

Thanks to its biconvexity, we can use coordinate descent algorithm to find the local minimum by alternatively updating K and I while fixing one of them and solving the corresponding convex optimization problem.

Given (K, I) from Problem P4, we have the second sub-problem given by

$$\min_{\mathbf{p}} \quad \left(b_0\sum_{i=1}^{N}\frac{C_i}{p_i} + b_1\right)\left(\sum_{i=1}^{N} p_i(Iw_{i,l}(K) + w_{i,a}(K))\right) \tag{P5}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} p_i = 1, \tag{P5a}$$

where $b_0 = \frac{AI}{KS}$ and $b_1 = AID + \frac{B}{I}$ are positive constants.

The challenge posed by Problem P5 is characterized by nonlinear optimization under polyhedral constraints. The Polyhedral Active Set Algorithm (PASA) as detailed by [17] is adopted to solve it. With $(K^*, I^*, \mathbf{p}^*)$, we update T by (20). The process continues by iteratively solving Problems P4 and P5 until the sequence of objective values of Problem P3 converges.

IV. Experiments

A. Experimental Settings

Datasets and Predictive Model: We utilize the EMNIST_LETTERS and FASHION_MNIST datasets. EMNIST_LETTERS contains 26 lowercase English letter images, and FASHION_MNIST features ten different image classes. For our experiments, we split each dataset into private datasets

for clients and a test dataset for evaluating our method. We employ the LeNet-5 as our classification model.

Data Heterogeneity: To simulate real-world data distributions in federated learning, we employ three data partitioning approaches: one I.I.D. and two non-I.I.D. configurations. *I.I.D.*: Training data is equally and randomly distributed among clients, ensuring a balanced class representation in each client's dataset. *Class*: Each client receives data from C randomly selected classes, where $C = 13$ for EMNIST_LETTERS and $C = 5$ for FASHION_MNIST. This approach represents 50% of classes in each dataset. *Dirichlet*: Data is distributed following a Dirichlet distribution with a parameter of 0.1, resulting in clients having varied data volumes and class distributions.

System Parameters: We set the total count of clients as $N = 10$ for FASHION_MNIST and $N = 40$ for EMNIST_LETTERS. We use the default Adam optimizer settings from TensorFlow. For our communication system, the total bandwidth is fixed at 2 MHz, with $S = 2$ sub-channels for FASHION_MNIST and $S = 5$ for EMNIST_LETTERS, which leads to $K \in [1, 5]$ and $K \in [1, 8]$ respectively. The size of the transmitted model is 2 million bits. Energy costs for clients, measured in millijoules (mJ), follow a normal distribution: $e_{i,l} \sim \mathcal{N}(10, 2)$ and $\bar{e}_{i,a} \sim \mathcal{N}(20, 4)$. Time costs are also modeled heterogeneously. For FASHION_MNIST, we generate $\delta_{i,l} \sim \mathcal{N}(5, 1)$ and $\bar{\delta}_{i,a} \sim \mathcal{N}(260, 100)$. For EMNIST_LETTERS, we have $\delta_{i,l} \sim \mathcal{N}(10, 2)$ and $\bar{\delta}_{i,a} \sim \mathcal{N}(560, 200)$. The parameter α is set to 0.5 by default.

B. Experiment Results

1) Optimal Selection Probability: To evaluate our optimal solution of participant selection probability $\mathbf{p}*$, we fix (K, I) and solve Problem P3 to obtain $\mathbf{p}*$. Specifically, $(K, I) = (2, 120)$ for FASHION_MNIST and $(K, I) = (2, 40)$ for EMNIST_LETTERS. We adopt three commonly used baselines: *Uniform Selection* [1]. This scheme samples participants uniformly, i.e., $p_1 = p_2 = \cdots = p_N = 1/N$. *Norm Selection* [16]. Participants are chosen in accordance with p_i, where $p_i = \frac{d_i G_i}{\sum_{i=1}^{N} d_i G_i}$. *Ratio Selection* [18]. This technique selects participants with $p_i = d_i$.

In a broad range of data distributions and datasets, our scheduling method consistently outperforms the other methods in all metrics. As detailed in TABLE I, our approach achieves a minimum cost reduction of 35% across all the settings in comparison with the uniform method. Specifically, in terms of the total cost, our approach achieves from a 40.97% to 61.92% reduction on the FASHION_MNIST dataset and even greater efficiency on the EMNIST_LETTERS dataset, reducing costs by 57.65% to 78.01%.

Notably, the robustness of our device scheduling stands in contrast to the setting-dependent performance of the norm and ratio methods. In challenging scenarios characterized by high data heterogeneity (Dir) and a large number of clients (EMNIST_LETTERS), our method prevails, achieving nearly an 80% reduction in cost in comparison with the uniform method.

2) Convergence Performance: To evaluate the convergence performance of our solution (K^*, I^*), we demonstrate the training loss curves with respect to total cost C_T under *Dir* setting in Figures 1, where the optimal (K, I) by grid search and other (K, I) combinations are compared with those under the same participant selection probability $\mathbf{p}^*$.

As shown, the performance of our proposed solution (K^*, I^*) is notably distinct. Our solution exhibits a convergence trajectory that is on par with the optimal benchmark. Specially, in the stage characterized by a steep decline in loss, our solution (K^*, I^*) often reaches convergence quicker than the optimal grid search.

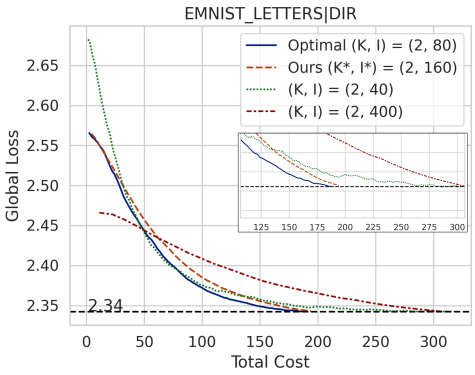
It is also noteworthy how various (K, I) combinations yield different convergence behaviors and overall total costs. Specifically, the number of participants, linear with K, compared with the number of local iterations I, appears to have more significant influence on cost efficiency.

V. Conclusion

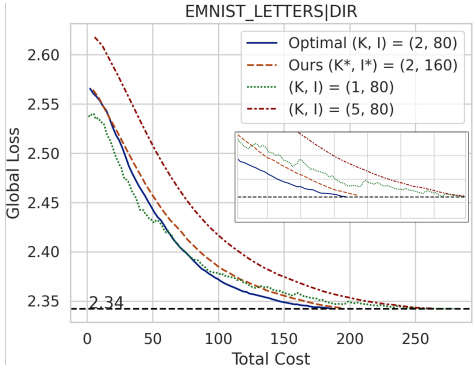
This paper presents a joint optimization problem for device and training scheduling in federated learning, addressing system and data heterogeneity. We have developed an

Data Dist.	Device Scheduling	FASHION_MNIST			EMNIST_LETTERS		
		Time Cost	Energy Cost	Total Cost	Time Cost	Energy Cost	Total Cost
IID	Uniform	541.31	1870.07	1205.69	746.88	2007.96	1377.42
	Norm	526.27	1816.45	1171.36	801.30	2160.90	1481.10
	Ratio	556.83	1933.35	1245.09	560.12	1503.83	1031.98
	GS-OFDMA	**297.24 (-45.09%)***	**1040.79 (-44.34%)**	**669.02 (-44.51%)**	**312.56 (-58.15%)**	**854.11 (-57.46%)**	**583.34 (-57.65%)**
Class	Uniform	458.33	1583.16	1020.75	961.03	2581.09	1771.06
	Norm	598.55	2048.02	1323.29	722.63	1936.35	1329.49
	Ratio	462.18	1606.59	1034.39	857.12	2296.36	1576.74
	GS-OFDMA	**296.71 (-35.26%)**	**908.33 (-42.63%)**	**602.52 (-40.97%)**	**345.83 (-64.01%)**	**938.78 (-63.63%)**	**642.31 (-63.73%)**
Dir	Uniform	1178.36	4100.03	2639.20	782.95	2097.54	1440.25
	Norm	523.56	1920.63	1222.10	462.23	1269.25	865.74
	Ratio	767.22	2735.80	1751.51	655.98	1783.85	1219.92
	GS-OFDMA	**419.65 (-64.39%)**	**1590.41 (-61.21%)**	**1005.03 (-61.92%)**	**168.13 (-78.53%)**	**465.27 (-77.82%)**	**316.70 (-78.01%)**

TABLE I: Cost Evaluation for Scheduling Schemes. **(-45.09%)*** denotes the cost reduced from the uniform method.



(a) Performance across different values of I.



(b) Performance across different values of K.

Fig. 1: Convergence performance on EMNIST_LETTERS.

iterative algorithm that combines coordinate descent with the polyhedral active set algorithm to solve the mixed integer nonlinear programming challenge. Empirical tests on real-world datasets demonstrate that our approach achieves superior cost efficiency in comparison with traditional federated learning algorithms across various datasets and data distributions.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[2] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, 2019, pp. 1387–1395.

[3] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Oort: Efficient federated learning via guided participant selection." in *Proc. USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2021, pp. 19–35.

[4] C. Li, X. Zeng, M. Zhang, and Z. Cao, "PyramidFL: A fine-grained client selection framework for efficient federated learning," in *Proc. Annu. Int. Conf. Mob. Comput. Netw. (MobiCom)*, 2022, p. 158–171.

[5] 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.

[6] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun. (TWC)*, vol. 20, no. 4, pp. 2457–2471, 2021.

[7] X. Mo and J. Xu, "Energy-efficient federated edge learning with joint communication and computation design," *J. Commn. Net*, vol. 6, no. 2, pp. 110–124, 2021.

[8] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun. (TWC)*, vol. 20, no. 3, pp. 1935–1949, 2021.

[9] L. WANG, W. WANG, and B. LI, "CMFL: Mitigating communication overhead for federated learning," in *Proc. IEEE Int. Conf. Distrib. Compt. Syst. (ICDCS)*, 2019, pp. 954–964.

[10] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun. (JSAC)*, vol. 37, no. 6, pp. 1205–1221, 2019.

[11] B. Luo, X. Li, S. Wang, J. Huang, and L. Tassiulas, "Cost-effective federated learning in mobile edge networks," *IEEE J. Sel. Areas Commun. (JSAC)*, vol. 39, no. 12, pp. 3606–3621, 2021.

[12] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, "HFEL: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning," *IEEE Trans. Wireless Commun. (TWC)*, vol. 19, no. 10, pp. 6535–6548, 2020.

[13] W. Chen, S. Horvath, and P. Richtarik, "Optimal client sampling for federated learning," *Trans. Machine Learning Research (TMLR)*, 2022.

[14] H. Ma, H. Guo, and V. K. N. Lau, "Communication-efficient federated multitask learning over wireless networks," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 609–624, 2023.

[15] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient radio resource allocation for federated edge learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[16] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," in *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, 2022, pp. 1739–1748.

[17] W. W. Hager and H. Zhang, "An active set algorithm for nonlinear optimization with polyhedral constraints," *Science China Mathematics*, vol. 59, no. 8, pp. 1525–1542, 2016.

[18] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Conf. Machine Learning and Syst. (MLSys)*, vol. 2, 2020, pp. 429–450.