

# FairEnergy: Contribution-Based Fairness meets Energy Efficiency in Federated Learning

Ouiame Marnissi, Hajar EL Hammouti, El Houcine Bergou

College of Computing, Mohammed VI Polytechnic University, Ben Guerir, Morocco.

{ouiame.marnissi, hajar.elhammouti, elhoucine.bergou}@um6p.ma

**Abstract**—Federated learning (FL) enables collaborative model training across distributed devices while preserving data privacy. However, balancing energy efficiency and fair participation while ensuring high model accuracy remains challenging in wireless edge systems due to heterogeneous resources, unequal client contributions, and limited communication capacity. To address these challenges, we propose FairEnergy, a fairness-aware energy minimization framework that integrates a contribution score capturing both the magnitude of updates and their compression ratio into the joint optimization of device selection, bandwidth allocation, and compression level. The resulting mixed-integer non-convex problem is solved by relaxing binary selection variables and applying Lagrangian decomposition to handle global bandwidth coupling, followed by per-device subproblem optimization. Experiments on non-IID data show that FairEnergy achieves higher accuracy while reducing energy consumption by up to 79% compared to baseline strategies.

**Index Terms**—Client selection, fairness, federated learning, energy efficiency, resource allocation, model compression.

## I. INTRODUCTION

Federated learning (FL) has emerged as a decentralized approach that allows a set of distributed devices to jointly train a shared machine learning model (ML) while keeping their data local [1]. Instead of uploading raw data, devices periodically exchange model updates with a coordinating server, thus preserving privacy and efficiently exploiting the computation and storage resources available at the network edge. Despite these advantages, practical deployment in wireless and edge environments faces two major challenges: *energy efficiency* and *fair participation*. Communication between devices and the central server typically dominates energy consumption, while the repeated exclusion of low-resource or low-quality clients leads to severe fairness degradation and biased models. Balancing these competing objectives remains an open problem. To reduce the communication cost in FL, various compression approaches have been introduced. They aim to decrease the size of transmitted updates by applying sparsification [2] or quantization[3], where only a subset or a lower-precision representation of model parameters is sent to the server. Such schemes substantially decrease communication overhead but may degrade model accuracy if not carefully designed or dynamically adapted [4].

In parallel, resource allocation strategies have been developed to optimally distribute communication resources such as bandwidth and power among devices [5], [6]. For instance, the work in [5] jointly optimize bandwidth and computation resources to minimize training costs, while authors in [6]

design an energy-efficient allocation scheme for heterogeneous edge devices.

Another effective way to reduce communication overhead is client selection, where only a subset of devices participates in each round. Selecting clients with more informative or energy-efficient updates can accelerate convergence and reduce redundant communication [7]–[9]. Early approaches such as FedCS [7] prioritized devices based on channel and computational conditions to meet latency constraints, while more recent studies have incorporated data-related or gradient-based criteria to improve convergence speed and model generalization [8], [10]. However, these strategies often overlook the long-term fairness of participation: devices with weaker channels or limited energy are repeatedly excluded, leading to model bias and slower convergence in heterogeneous networks.

Recent efforts have highlighted the importance of fairness in FL as clients differ in data distribution, resources, and connectivity. Naive selection can systematically exclude certain devices, leading to biased models and slower convergence. Approaches such as biased client selection analysis [9] and fairness-aware client scheduling [11] aim to balance participation by adjusting selection probabilities or reweighting local contributions. More advanced frameworks consider multiple fairness criteria simultaneously, including participation frequency, data quality, and resource heterogeneity [12], [13]. While these methods promote balanced participation, they often ignore communication efficiency, compression, and energy constraints, which are critical in wireless and edge environments. A unified framework that jointly optimizes energy efficiency, update quality, and long-term fairness remains largely unexplored.

In this paper, we propose a Fairness-Aware Energy Minimization Framework for FL, termed *FairEnergy*. The core idea is to promote energy-efficient and equitable participation among devices through a fairness-driven selection mechanism. Specifically, we define a score function based on the update magnitude and compression ratio. These scores guide the joint optimization of device selection, compression ratio, and bandwidth allocation to minimize total communication energy while ensuring fair long-term participation across heterogeneous devices.

Our main contributions are summarized as follows:

- We formulate a fairness-aware energy minimization problem that jointly optimizes client selection, compression, and bandwidth allocation. The formulation integrates two

important metrics: a *contribution score* combining update magnitude and compression, and a *long-term fairness metric* to ensure equitable participation among clients.

- We develop a computationally efficient solution via Lagrangian relaxation and dual decomposition. The method yields an intuitive threshold rule where a client is selected only if the benefit of its update outweighs the combined cost of energy and bandwidth. Optimal bandwidth allocation is determined via Golden Section Search (GSS), while Lagrange multipliers for bandwidth and fairness constraints are updated via subgradient ascent to ensure global feasibility.
- Through experiments on non-IID FMNIST data, we demonstrate that FairEnergy achieves higher model accuracy while significantly reducing total energy consumption compared to benchmark strategies. Specifically, FairEnergy requires roughly 71% less energy than Score-Max and 79% less than EcoRandom to reach the same target accuracy.

The remainder of this paper is organized as follows. In section II, we present the system model. Section III introduces the fairness-aware contribution metrics. In Section IV, we formulate the joint optimization problem and present a Lagrangian based solution in Section V. Section VI summarizes the per-round algorithm and analyzes its computational complexity. Simulation experiments are discussed in Section VII. Finally, conclusions are drawn in Section VIII.

## II. SYSTEM MODEL

### A. Learning Setup

We consider  $N$  connected clients that communicate with a central server over the wireless network. Each client trains the ML model on its private dataset  $\mathcal{D}_i$  of size  $|\mathcal{D}_i|$ . The global learning objective is to minimize the empirical risk:

$$\min_{\mathbf{w}} F(\mathbf{w}) = \sum_{i=1}^N \frac{|\mathcal{D}_i|}{|\mathcal{D}|} F_i(\mathbf{w}),$$

where  $F_i$  is the local loss function of device  $i$  and  $w$  is the global model. We let the binary variable  $x_i^r \in \{0, 1\}$  indicate whether client  $i$  is selected ( $x_i^r = 1$ ) or not in round  $r$ .

The training is performed in synchronous rounds  $r = 1, 2, \dots$ . At the beginning of round  $r$ , the server broadcasts the current global model  $\mathbf{w}^r$  to a subset of selected devices. Each selected device then computes a local model update  $\mathbf{u}_i^{r+1}$ , i.e., the gradient of its local loss function  $\nabla F_i(\mathbf{w}^{r+1})$  and sends it to the server. Finally, the server aggregates these local updates to obtain the new global model  $\mathbf{w}^{r+1}$ .

In this work, we focus on the communication aspect of FL, as it remains one of the main bottlenecks in distributed training [14]. In particular, the frequent exchange of model updates between clients and the server leads to significant communication overhead, especially in constrained networks. Consequently, we concentrate on minimizing the communication energy consumption, leaving the computation cost of local training outside the scope of the optimization [2]. In the

following, we present the communication model adopted to characterize uplink transmission and energy consumption.

### B. Communication Model

Each device  $i$  communicates with the server via a wireless uplink channel. Let  $B_i^r$  denote the bandwidth assigned to device  $i$  in round  $r$ , and  $\gamma_i^r \in [\gamma_{\min}, 1]$  denote its sparsity ratio, which represents the fraction of non-zero coefficients in the compressed model update.

The size of the transmitted payload is  $\gamma_i^r S + I$ , where  $S$  is the size of the full model update and  $I$  is the overhead for encoding the indices of the non-zero coefficients [15].

We assume an additive Gaussian noise channel with channel gain  $h_i^r$ , and noise spectral density  $N_0$ . The achievable uplink rate then follows the Shannon capacity formula:

$$R_i^r = B_i^r \log_2 \left( 1 + \frac{P_i h_i^r}{N_0 B_i^r} \right),$$

where  $P_i$  is the transmit power of device  $i$ .

The communication time of device  $i$  in round  $r$  is

$$T_i^r = \frac{\gamma_i^r S + I}{R_i^r},$$

and the corresponding communication energy is

$$E_i^r = P_i T_i^r.$$

A multi-objective trade-off is at the heart of efficient and fair FL. Allocating more bandwidth ( $B_i^r$ ) or transmitting less compressed updates (higher  $\gamma_i^r$ ) can enhance learning efficiency but escalates communication energy. Conversely, aggressive compression or selecting only clients with good channels saves immediate energy but risks discarding informative updates from other clients, thereby degrading model accuracy and more importantly violating long-term fairness among clients. Therefore, in subsequent sections, we aim to answer the following question: How can the server jointly optimize client selection, bandwidth allocation, and compression ratios to minimize total communication energy, while simultaneously ensuring high model accuracy and guaranteeing long-term fairness for all participating clients?

## III. FAIRNESS-AWARE CONTRIBUTION METRICS

To balance efficiency with fairness in resource-constrained FL, we introduce two key metrics: a *contribution score* that quantifies the utility of a compressed client update, and a *long-term fairness metric* that tracks participation history.

### A. Contribution Score

Client heterogeneity leads to unequal contributions to the global model. Prioritizing clients with more impactful updates can accelerate convergence [8], yet in resource-constrained environments, updates must often be compressed, which reduces communication cost but also limits the information available to the server.

To capture an update's importance after compression, we define the *contribution score* for client  $i$  in round  $r$  as:

$$s_i^r(\gamma_i^r) = \|\mathbf{u}_i^r\| \cdot \gamma_i^r,$$

where  $\|u_i^r\|$  is the  $L_2$  norm of the update  $u_i^r$ . The  $L_2$  norm of the update reflects the overall magnitude of the changes contributed by the client, while the compression factor scales this magnitude according to the information actually sent. This score provides a simple and practical measure of each client's effective contribution.

### B. Long-Term Fairness Metric

The proposed fairness metric is framed as long-term participation fairness, in that we track the effective participation rate of each client and enforce a minimum threshold on that rate. Prior works have argued for fairness in client selection by ensuring each client gets an equitable number of participation opportunities across rounds [12] and by incorporating clients' past selection history in the selection process [11].

To capture this long-term participation, we adopt an exponential moving average (EMA):

$$q_i^r = \rho q_i^{r-1} + (1 - \rho) x_i^r, \quad (1)$$

where  $\rho$  controls the memory of past rounds and the initialization of  $q_i^0$  is discussed in Section VI-A. This formulation allows us to maintain a smooth and responsive measure of participation: clients that have been selected less frequently in recent rounds have a lower  $q_i^r$  and are thus more likely to be prioritized in future selections. By tuning  $\rho$ , we can control how strongly recent participation affects the metric, providing a flexible way to ensure long-term fairness. This also enables the selection framework to dynamically boost under-selected clients and maintain equitable participation over time. Finally, we impose a minimum participation threshold:

$$q_i^r \geq \pi_{\min}, \quad \forall i \in \{1, \dots, N\},$$

ensuring that every client contributes to the learning process at least at a minimal rate  $\pi_{\min}$  over time.

By combining  $s_i^r$  and  $q_i^r$ , the framework simultaneously prioritizes high-quality updates and guarantees fair long-term participation across all devices. These metrics serve as the basis for jointly selecting devices, allocating resources, and adapting compression ratios.

In the next section, we formalize this design as a fairness-aware energy minimization problem, where the goal is to optimize communication energy while ensuring both update quality and equitable participation.

## IV. PROBLEM FORMULATION

In this section, we formulate the fairness-aware energy minimization problem. At each global round  $r$ , the server aims to select a subset of participating devices while jointly optimizing their communication parameters (compression ratio and bandwidth allocation) to minimize the total communication energy. Each device  $i$  contributes to the global model with a score  $s_i^r(\gamma_i^r)$  that reflects the relevance and quality of its local update. To prevent over-representation of a few high-scoring devices, we introduce a long-term fairness constraint ensuring that each device participates in the training process with at

least a minimum rate  $\pi_{\min}$ . This translates into the following problem:

$$\underset{\{x_i^r, \gamma_i^r, B_i^r\}_{i=1}^N}{\text{minimize}} \quad \sum_{i=1}^N \left( x_i^r E_i^r(\gamma_i^r, B_i^r) - \eta x_i^r s_i^r(\gamma_i^r) \right) \quad (2a)$$

$$\text{subject to} \quad \sum_{i=1}^N x_i^r B_i^r \leq B_{\text{tot}}, \quad (2b)$$

$$\gamma_{\min} \leq \gamma_i^r \leq 1, \quad \forall i, \quad (2c)$$

$$x_i^r \in \{0, 1\}, \quad \forall i, \quad (2d)$$

$$q_i^r \geq \pi_{\min}, \quad \forall i. \quad (2e)$$

Constraint (2b) ensures that the total allocated bandwidth does not exceed the available system budget. Constraint (2c) bounds the compression ratio within a feasible range. The binary variable  $x_i^r$  in (2d) indicates whether device  $i$  participate in round  $r$ . Finally, constraint (2e) ensures that each device maintains a minimum participation ratio over time, where  $q_i^r$  is the EMA fairness metric defined in equation (1)<sup>1</sup>.

Problem (2) is mixed-integer, non-convex optimization problem. The binary participation variables  $x_i^r$  create combinatorial complexity, while the coupling between selection, compression, and bandwidth makes joint optimization challenging.

To address these challenges, we adopt a Lagrangian relaxation approach [16] that introduces dual variables associated with bandwidth and fairness constraints. The problem is decomposed to simpler per-device subproblems. Each device can then optimize its local compression and bandwidth decisions independently, while the server iteratively adjusts dual variables to enforce global fairness and bandwidth feasibility.

In the following section, we detail the Lagrangian formulation and the iterative resolution process adopted in our setup.

## V. PROBLEM RESOLUTION WITH LAGRANGIAN RELAXATION

To efficiently solve the per-round optimization problem (2), we adopt a relaxation-decomposition strategy that enables distributed per-device updates while ensuring both bandwidth feasibility and long-term fairness. Concretely, we first relax the binary selection variables  $x_i^r \in \{0, 1\}$  to the continuous interval  $[0, 1]$ . Next, we form the partial Lagrangian by dualizing only the coupling constraints (bandwidth and fairness), which allows the problem to decompose across devices and enables efficient distributed optimization. Finally we recover feasible binary decisions through a repair step.

### A. Lagrangian Relaxation and Problem Decomposition

After relaxing  $x_i^r \in \{0, 1\}$  to  $x_i^r \in [0, 1]$ , we introduce dual variables  $\lambda^r \geq 0$  for the bandwidth constraint in (2b),

<sup>1</sup>Although the global loss convergence is not explicitly added as a constraint in the studied optimization, our design is supported by our prior theoretical work [8], which established that selection based on update magnitude accelerates convergence. Here, we extend this result to a more practical setting with compression and fairness criteria, and empirically show convergence in Section VII, leaving a full theoretical analysis for future work.

and  $\mu_i^r \geq 0$  for the fairness constraint in (2e). The partial Lagrangian is therefore

$$\begin{aligned} \mathcal{L}(\mathbf{x}^r, \boldsymbol{\gamma}^r, \mathbf{B}^r; \lambda^r, \boldsymbol{\mu}^r) = & \sum_{i=1}^N x_i^r (E_i^r(\gamma_i^r, B_i^r) - \eta s_i^r(\gamma_i^r)) \\ & + \lambda^r \left( \sum_{i=1}^N x_i^r B_i^r - B_{\text{tot}} \right) + \sum_{i=1}^N \mu_i^r (\pi_{\min} - q_i^{r-1}). \end{aligned} \quad (3)$$

By substituting the explicit form of the fairness metric in equation (1) into the Lagrangian and rearranging the terms, the global Lagrangian can be separated into a sum of independent, per-device Lagrangians:

$$\begin{aligned} \mathcal{L}_i(x_i^r, \gamma_i^r, B_i^r; \lambda^r, \boldsymbol{\mu}^r) = & x_i^r \left( E_i^r(\gamma_i^r, B_i^r) + \lambda^r B_i^r - \eta s_i^r(\gamma_i^r) - \mu_i^r (1 - \rho) \right) \\ & + \mu_i^r (\pi_{\min} - \rho q_i^{r-1}), \end{aligned} \quad (4)$$

where the term  $\mu_i^r (\pi_{\min} - \rho q_i^{r-1})$  is a constant with respect to the current round's decision variables.

Thus, for fixed duals  $(\lambda^r, \boldsymbol{\mu}^r)$ , the global problem decomposes into  $N$  independent local minimization problems.

$$\min_{x_i^r \in [0,1], \gamma_i^r \in [\gamma_{\min}, 1], B_i \geq 0} \mathcal{L}_i(x_i^r, \gamma_i^r, B_i^r).$$

### B. Per-Device Minimization

Because the function  $\mathcal{L}_i$  is *affine* in  $x_i^r$ , the minimizer over  $x_i^r \in [0, 1]$  lies at the boundaries of the feasible interval.

$$x_i^r = \begin{cases} 1, & E_i^r(\gamma_i^r, B_i^r) + \lambda^r B_i^r < \eta s_i^r(\gamma_i^r) + \mu_i^r (1 - \rho), \\ 0, & \text{otherwise.} \end{cases}$$

Hence, the relaxed solution yields a binary decision: the device participates ( $x_i = 1$ ) only if the benefit of its update outweighs the associated energy and bandwidth costs. We therefore consider the two possible cases:

*Case 1: Device not selected ( $x_i^r = 0$ )*

The Lagrangian reduces to a constant:

$$\mathcal{L}_i^{\text{not}} = \mu_i^r (\pi_{\min} - \rho q_i^{r-1}),$$

and no further optimization over  $\gamma_i^r$  or  $B_i^r$  is required.

*Case 2: Device selected ( $x_i^r = 1$ )*

We must solve the joint compression and bandwidth allocation subproblem:

$$\mathcal{L}_i^{\text{sel}} = \min_{\gamma_i^r, B_i^r} \phi_i^r(\gamma_i^r, B_i^r) - \mu_i^r (1 - \rho) + \mu_i^r (\pi_{\min} - \rho q_i^{r-1}),$$

where the core optimization is captured in the function:

$$\phi_i^r(\gamma_i^r, B_i^r) = E_i^r(\gamma_i^r, B_i^r) + \lambda^r B_i^r - \eta s_i^r(\gamma_i^r). \quad (5)$$

Only this term depends on the decision variables  $\gamma_i^r$  and  $B_i^r$ .

### C. Joint Optimization of $(\gamma_i^r, B_i^r)$ for Selected Clients

When a client is selected, the minimization of the Lagrangian over  $(\gamma_i^r, B_i^r)$  reduces to the following problem

$$\min_{\gamma_i^r, B_i^r} \phi_i^r(\gamma_i^r, B_i^r).$$

The function  $\phi_i^r$  is not convex in  $B_i^r$ , but exhibits a *unimodal* shape: for small  $B_i^r$ , transmission energy dominates;

as  $B_i^r$  increases, energy decreases sharply, then flattens as the achievable rate saturates; finally, the linear term  $\lambda^r B_i^r$  grows, increasing the total cost. This decreasing-flattening-increasing pattern produces a single minimum, making the problem well-suited to derivative-free one-dimensional search methods such as the Golden Section Search (GSS) [17], [18].

We therefore solve the joint minimization in three steps:

- 1) Discretize  $\gamma_i^r$  on a small finite grid  $\Gamma = \{\gamma^{(1)}, \dots, \gamma^{(G)}\}$ .
- 2) For each  $\gamma^{(g)} \in \Gamma$ , find  $B_i^{r,*}(\gamma^{(g)}) = \arg \min_{B_i^r} \phi_i^r(\gamma^{(g)}, B_i^r)$  using GSS.
- 3) Evaluate  $\phi_i^r(\gamma^{(g)}, B_i^{r,*}(\gamma^{(g)}))$  and choose  $(\gamma_i^{r,*}, B_i^{r,*}) = \arg \min_{\gamma^{(g)} \in \Gamma} \phi_i^r(\gamma^{(g)}, B_i^{r,*}(\gamma^{(g)}))$ .

## VI. ALGORITHM AND DUAL UPDATES

In the following, we provide a brief summary of the per-resolution of the FairEnergy optimization algorithm, followed by an analysis of its computational complexity.

---

### Algorithm 1 Per-Round FairEnergy Optimization

---

- 1: **Input:** previous fairness metrics  $q_i^{r-1}$ , parameters  $\eta, \rho, \Gamma$ , step sizes  $\alpha_\lambda, \alpha_\mu > 0$
  - 2: **Initialize:** dual variables  $\lambda^r, \boldsymbol{\mu}^r$
  - 3: **repeat**
  - 4:     **for** each device  $i = 1, \dots, N$  **do**
  - 5:         For each  $\gamma_i^r \in \Gamma$ , use GSS to find  $B_i^{r,*}(\gamma_i^r)$  that minimizes  $\phi_i^r$  defined in equation (5).
  - 6:         Select  $(\gamma_i^{r,*}, B_i^{r,*}) = \arg \min_{\gamma_i^r \in \Gamma} \phi_i^r(\gamma_i^r, B_i^{r,*}(\gamma_i^r))$ .
  - 7:         Determine selection:  
$$x_i^{r,*} = \begin{cases} 1, & E_i^r(\gamma_i^{r,*}, B_i^{r,*}) + \lambda^r B_i^{r,*} < \eta s_i^r(\gamma_i^{r,*}) + \mu_i^r (1 - \rho), \\ 0, & \text{otherwise.} \end{cases}$$
  - 8:         Update fairness metric:  $q_i^{r+1} = \rho q_i^r + (1 - \rho) x_i^{r,*}$
  - 9:         Update fairness dual:  
$$\mu_i^r \leftarrow [\mu_i^r + \alpha_\mu (\pi_{\min} - \rho q_i^{r-1} - (1 - \rho) x_i^{r,*})]^+$$
  - 10:      **end for**
  - 11:      Server updates bandwidth dual:  
$$\lambda^r \leftarrow [\lambda^r + \alpha_\lambda (\sum_i x_i^{r,*} B_i^{r,*} - B_{\text{tot}})]^+$$
  - 12: **until** convergence
- 

### A. Algorithm

Algorithm 1 summarizes the per-round resolution of the proposed FairEnergy optimization. For a given round  $r$ , each device  $i$  first solves a local optimization over the feasible compression ratios  $\gamma_i^r \in \Gamma$  to obtain the corresponding optimal bandwidth allocation  $B_i^{r,*}(\gamma_i^r)$  via GSS. The fairness metric  $q_i^r$  is subsequently updated using an EMA to capture the device's participation history. For feasibility, we initialize  $q_i^0$  sufficiently large so that in the early rounds, fairness is not restrictive and selection is mainly driven by minimizing energy and maximizing score. As rounds progress,  $q_i^r$  naturally evolves to enforce fairness. Finally, the dual variables  $\lambda$  and  $\boldsymbol{\mu}$  associated

respectively with the total bandwidth and fairness constraints are updated as shown in lines 10 and 12 of Algorithm 1. These updates follow the standard projected subgradient ascent rule commonly used in dual decomposition methods [16], [19], where the subgradients correspond to the constraint violations of the bandwidth and fairness limits. In practice, these dual variables are typically refined over several inner iterations to ensure convergence of the Lagrangian subproblem within each communication round. However, to keep notation concise and the algorithm readable, we present only one generic update per round, which implicitly captures this iterative process.

### B. Complexity Analysis

The computational complexity of FairEnergy is dominated by the per-device joint optimization of the compression ratio ( $\gamma_i^r$  and bandwidth  $B_i^r$ ). For each device,  $\gamma_i^r$  is discretized over a grid of size  $G$ , and for each grid point,  $B_i^r$  is optimized via the GSS method with  $T_{\text{GSS}}$  function evaluations, yielding a per-device complexity of  $\mathcal{O}(G T_{\text{GSS}})$ . Subsequent computations of the selection metric, relaxed selection, and local fairness updates are negligible in comparison. The global bandwidth dual update  $\lambda^r$  requires a sum over  $N$  devices, giving  $\mathcal{O}(N)$  complexity. Consequently, the total per-round complexity is  $\mathcal{O}(N G T_{\text{GSS}})$ .

## VII. SIMULATION RESULTS

In this section, we evaluate the performance of our proposed framework in terms of test accuracy and communication energy. Experiments are implemented in Keras with TensorFlow. We train a convolutional neural network with approximately 2 million parameters on a non-IID partitioned FMNIST, a dataset of fashion products from 10 categories. We consider a network with  $N = 50$  devices. The total available bandwidth is  $B_{\text{tot}} = 10$  MHz, and transmission power for each device is uniformly distributed in  $[0.1, 0.3]$  mW. To model heterogeneous data distributions, local datasets are sampled using a Dirichlet distribution  $Dir_K(\beta)$  with concentration parameter  $\beta = 0.3$  [20]. Compression ratios are constrained in  $[0.1, 1]$ , and long-term fairness is enforced with threshold  $\pi_{\min} = 0.2$ . The participation memory for fairness tracking is  $\rho = 0.6$  and the learning rate is of 0.01.

To ensure a fair comparison, the number of selected devices in the baseline schemes is fixed to the mean number of devices selected by FairEnergy across rounds.

We evaluate the performance of FairEnergy to the following baselines designed to isolate components of our approach:

- ScoreMax: selects devices with the highest contribution scores, ignoring energy and fairness. To isolate the effect of contribution score, updates are transmitted at full precision ( $\gamma_i = 1$ ), and the available bandwidth  $B_{\text{tot}}$  is equally divided among the selected devices. This approach isolates the effect of importance-driven selection and has been extensively investigated in [8], [21].
- EcoRandom: a communication-efficient random selection baseline that removes both fairness and contribution-awareness. In each round, devices are chosen randomly,

and all selected devices transmit using the minimum compression ratio and bandwidth observed in FairEnergy. This configuration drives EcoRandom toward the lowest possible communication energy, in line with the principle of minimizing transmission cost explored in several state-of-the-art energy-aware FL studies [4], [22].

These benchmarks allow us to assess the trade-offs between energy efficiency, model performance, and fairness under different device selection policies.

Figure 1 shows that FairEnergy and ScoreMax achieve the highest test accuracy. Both methods select devices based on their contribution importance, leading to the inclusion of clients providing the most informative updates. In contrast, EcoRandom has slow convergence over the rounds mainly because of the large compression levels applied to reduce transmission energy. Notably, FairEnergy also applies compression, yet it maintains accuracy comparable to ScoreMax throughout the rounds. This is because FairEnergy balances contribution with fairness in client participation, ensuring that both high-quality and underrepresented devices contribute to model training.

In Figure 2, we report the average energy consumption per global round. As expected, EcoRandom achieves the lowest per-round energy consumption due to its aggressive compression and favorable bandwidth allocation. FairEnergy follows closely, consuming slightly more energy. In contrast, ScoreMax operates with no compression and with uniform bandwidth across devices, which leads to consistently higher energy consumption over the rounds.

In Figure 3, we compare the total cumulative energy required to reach a target accuracy of 80%. The results clearly highlights the effectiveness of the proposed FairEnergy framework, which achieves the target accuracy while consuming substantially less energy than the baselines. Specifically, FairEnergy requires roughly 71% less energy than ScoreMax and 79% less than EcoRandom. While EcoRandom consumes less energy per round, its slower convergence leads to a higher overall energy cost to reach the desired accuracy. ScoreMax, which operates without compression and uniform bandwidth, also requires significantly more energy due to inefficient communication. These results demonstrate that combining contribution-aware device selection with fairness-driven participation effectively reduces the total energy needed to achieve a target accuracy.

Table I: Device participation statistics across strategies.

Strategy	Min	Max	Std
FairEnergy	401	413	2.85
ScoreMax	50	697	180.93
EcoRandom	369	423	15.79

Table I highlights how FairEnergy enforces fairness in device participation, maintaining a tight selection range across devices. EcoRandom, being a random selection baseline,

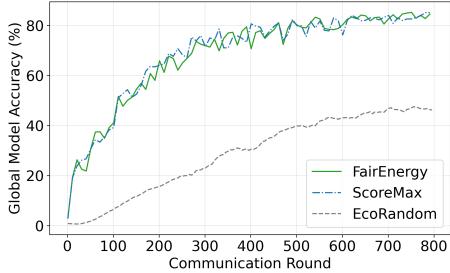


Figure 1: Test accuracy per round.

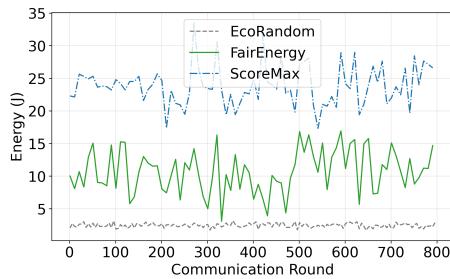


Figure 2: Energy consumption per round.

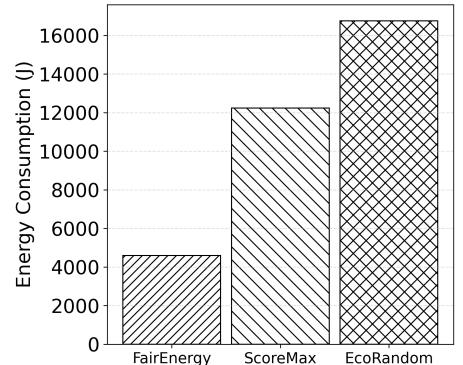


Figure 3: Total energy needed to achieve a target accuracy.

shows moderate variability, while ScoreMax prioritizes only the most informative clients, resulting in highly uneven participation.

### VIII. CONCLUSION

In this work, we proposed FairEnergy, a joint fairness and energy-aware framework for FL that optimizes device selection, bandwidth allocation, and compression ratios. By using a contribution score combining update magnitude and compression ratio, along with long-term fairness constraints, our approach ensures efficient communication, equitable participation, and faster model convergence. The proposed per-round resolution, based on Lagrangian relaxation and efficient per-device minimization, enables scalable implementation across large networks. Experiments demonstrate that FairEnergy achieves substantial energy savings while maintaining high model accuracy. Future work will extend this framework to dynamic channel environments and joint optimization of computation and communication resources.

### REFERENCES

- [1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [2] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang, “Gradient sparsification for communication-efficient distributed optimization,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [3] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani, “Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization,” in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2021–2031.
- [4] Ouiame Marnissi, Hajar El Hammouti, and El Houcine Bergou, “Adaptive sparsification and quantization for enhanced energy efficiency in federated learning,” *IEEE Open Journal of the Communications Society*, 2024.
- [5] Van-Dinh Nguyen, Shree Krishna Sharma, Thang X Vu, Symeon Chatzinotas, and Björn Ottersten, “Efficient federated learning algorithm for resource allocation in wireless iot networks,” *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3394–3409, 2020.
- [6] Deepali Kushwaha and Rajesh M Hegde, “Optimal device selection and resource allocation in federated learning,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [7] Takayuki Nishio and Ryo Yonetani, “Client selection for federated learning with heterogeneous resources in mobile edge,” in *ICC 2019-2019 IEEE international conference on communications (ICC)*. IEEE, 2019, pp. 1–7.
- [8] Ouiame Marnissi, Hajar El Hammouti, and El Houcine Bergou, “Client selection in federated learning based on gradients importance,” in *AIP Conference Proceedings*. AIP Publishing LLC, 2024, vol. 3034, p. 100005.
- [9] Yae Jee Cho, Jianyu Wang, and Gauri Joshi, “Towards understanding biased client selection in federated learning,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10351–10375.
- [10] Yae Jee Cho, Jianyu Wang, and Gauri Joshi, “Client selection in federated learning: Convergence analysis and power-of-choice selection strategies,” *arXiv preprint arXiv:2010.01243*, 2020.
- [11] Yuxin Shi, Zelei Liu, Zhan Shi, and Han Yu, “Fairness-aware client selection for federated learning,” in *2023 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2023, pp. 324–329.
- [12] Simin Javaherian, Sanjeev Panta, Shelby Williams, Md Sirajul Islam, and Li Chen, “Fedfair 3: Unlocking threefold fairness in federated learning,” in *ICC 2024-IEEE International Conference on Communications*. IEEE, 2024, pp. 3622–3627.
- [13] Ying-Chi Mao, Li-Juan Shen, Jun Wu, Ping Ping, and Jie Wu, “Federated dynamic client selection for fairness guarantee in heterogeneous edge computing,” *Journal of Computer Science and Technology*, vol. 39, no. 1, pp. 139–158, 2024.
- [14] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek, “Robust and communication-efficient federated learning from non-iid data,” *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.
- [15] Mohammad Mohammadi Amiri and Deniz Gündüz, “Federated learning over wireless fading channels,” *IEEE transactions on wireless communications*, vol. 19, no. 5, pp. 3546–3557, 2020.
- [16] Dimitri P Bertsekas, “Nonlinear programming,” *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.
- [17] Heng Wang, Aijun Liu, Xiaofei Pan, and Jianfei Yang, “Optimization of power allocation for multiusers in multi-spot-beam satellite communication systems,” *Mathematical Problems in engineering*, vol. 2014, no. 1, pp. 780823, 2014.
- [18] Jack Kiefer, “Sequential minimax search for a maximum,” *Proceedings of the American mathematical society*, vol. 4, no. 3, pp. 502–506, 1953.
- [19] Daniel Pérez Palomar and Mung Chiang, “A tutorial on decomposition methods for network utility maximization,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.
- [20] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He, “Federated learning on non-iid data silos: An experimental study,” in *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 2022, pp. 965–978.
- [21] Wenlin Chen, Samuel Horvath, and Peter Richtarik, “Optimal client sampling for federated learning,” *arXiv preprint arXiv:2010.13723*, 2020.
- [22] Minsu Kim, Walid Saad, Mohammad Mozaffari, and Merouane Debbah, “Green, quantized federated learning over wireless networks: An energy-efficient design,” *IEEE transactions on wireless communications*, vol. 23, no. 2, pp. 1386–1402, 2023.