Federated Learning in Heterogeneous Networks with Unreliable Communication

Paul Zheng, *Student Member, IEEE*, Yao Zhu, *Member, IEEE*, Yulin Hu, *Senior Member, IEEE*, Zhengming Zhang, *Member, IEEE*, and Anke Schmeink, *Senior Member, IEEE*

***Abstract*—In federated learning (FL), local workers learn a global model collaboratively using their local data by communicating trained models to a central server for privacy concerns. Due to its local nature, FL is typically subject to various heterogeneities, including system and statistical heterogeneity. To address these concerns, Federated Proximal (FedProx) has been considered a promising FL paradigm to provide more stable learning convergence in the presence of computation stragglers and statistical heterogeneity. However, in wireless networks with unreliable communication channels, the errors of packet transmissions should be considered, introducing additional heterogeneity. For the first time, we rigorously prove the convergence of FedProx in the presence of transmission packet errors in heterogeneous networks. In addition, we propose a joint client selection and resource allocation strategy that maximizes the number of effective participating users for convergence acceleration. The method is combined with a random weight mechanism to reduce the statistical bias caused by the client selection strategy. An efficient low-complexity algorithm for solving the optimization problem is developed. The proposed method achieves faster convergence and requires fewer communication rounds to attain accuracy than existing state-of-the-art client selection methods.**

***Index Terms*—Distributed learning, federated learning, wireless networks, packet error rate, convergence analysis**

I. Introduction

Data-driven machine learning approaches offer great potential for solving highly complex problems. The presence of a large amount of data is the cornerstone of machine learning. However, the vast amount of data generated by the growing number and variety of Internet-of-Things (IoT) devices makes machine learning computations intensive and training time prohibitively long. Decentralized learning schemes have been investigated to allocate the training task to several computation centers [2]. The excessive data quantity is also a cause for concern in terms of communication bandwidth and privacy. In this regard, a promising learning framework called federated learning (FL) is proposed in [3], [4]. In FL, the training process is carried out locally at each device while the trained models are aggregated in the central server. In particular, at each communication round, the central server sends the global learning model to a group of selected users. Then, users train the model locally with their own local data and upload the trained models to the server. At the central server, received models are averaged.

Due to its highly distributed nature, one of the major challenges in FL is to address the heterogeneity [5], [6]. Since users' preferences may differ, its data could also contain a statistical bias. In addition, the quantity of data acquired by each user is most likely unbalanced, which results in the system heterogeneity since the number of iterations in a single epoch differs if all users' local batch sizes are set as identical [7]. Moreover, each device's computing capability and availability vary. Stragglers, i.e., users who are extremely slow to train or update through wireless links, contribute to high latency since the server must wait for each user to update before proceeding to the next round. On the other hand, if the server moves to the next rounds once a specific time limit is met without waiting for the stragglers, their data information may never be used during the training phase. Considering these different kinds of heterogeneity, the machine learning target could be skewed with the conventional FL approach.

To tackle this issue, numerous solutions have been proposed. Momentum-based strategies have been presented in [8]–[10]. By maintaining double momentum buffers, [8] improves the training performance in a cross-silo FL context with non-i.i.d. data distributions. A generic algorithmic framework is proposed in [9] to reduce the client drift and adapt methods like momentum or ADAM to FL settings. Authors in [10] developed a momentum-based technique without introducing additional computation and communication load. Another line of research has been to use distillation-based methods as in [11], [12]. Authors in [11] resolve the heterogeneity problem by addressing the forgetting problem (of the local model forgetting the global model) approach. Knowledge distillation method usually requires a proxy dataset, which may not be realistic in FL settings; the authors in [12] offer a data-free method to address heterogeneous FL. In this work, we focus on strategies based on regularization [13]–[15] which, in general, add a regularization term to the local model update. Model-contrastive FL is proposed in [13]. By adding a model-contrastive loss for the similarity between model representations on the local loss, it provides accurate empirical results by correcting local model training. Both [14], [15] propose using a dynamic regularizer to the local model update based on the global model. Federated Proximal (FedProx) [15],

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the project "Open6GHub" (grant number: 16KISK012). Part of this work (Sections III, A.B. and IV. A.) has been presented at IEEE Globecom workshops in 2021 [1].

P. Zheng, Y. Zhu, Y. Hu are with Wuhan University, 430072 Wuhan, China and RWTH Aachen University, D-52074 Aachen, Germany, (email: $zheng|zhu|hu$@inda.rwth-aachen.de). Y. Hu is the corresponding author. Z. Zhang is with Southeast University, China (email: $zmzhang$@seu.edu.cn). A. Schmeink is with RWTH Aachen University, D-52074 Aachen, Germany. (email: $schmeink$@inda.rwth-aachen.de).

inspired by proximity operator [16], uses a regularizer with a fixed weight to control the distance between local loss and global loss. FedDyn [14] adds a regularizer such that the stationary point of local loss is also a stationary point of the global loss, hence enhancing the performance under non-i.i.d. setting.

Due to the existence of numerous possible frameworks, we have selected FedProx [15] in this work as the base method. This choice is motivated by its more general theoretical properties and ability to ensure stability in the presence of heterogeneities. In fact, it generalizes Federated Averaging (FedAvg) [3] and improves the convergence stability in case of non-i.i.d. dataset and partial computations for stragglers. In addition, FedProx offers a theoretical convergence analysis that accounts for these heterogeneities without requiring the loss functions to be strongly convex, which is generally essential in FedAvg's convergence analysis [17]. In fact, models in deep learning are non-convex in general. In addition, we note that the other mentioned solutions are either complementary (for momentum-based and distillation-based methods) or can be adapted (for FedDyn) to the solution in this work.

Moreover, applying FL in wireless communication networks (referred to as federated edge learning in the literature) introduces additional concerns and heterogeneity. For instance, local devices may lack sufficient energy for transmissions or training. Furthermore, poor channel conditions or computation power may result in high training latency. With consideration of those issues, how to optimally allocate bandwidth to those users is also a complex problem. As a result, resource allocation techniques for FL in wireless networks are required [18] and have been extensively studied under a variety of constraints and objectives. Authors in [19] propose a joint learning and communication resource allocation strategy to minimize the total energy consumption under a latency constraint. By incorporating a weighted proximal term to account for the heterogeneity, the work in [20] presents an efficient method to minimize the energy consumption or the completion time of FL training.

One line of research to counter the limited wireless bandwidths is by exploiting the superposition property of multi-access channels for the FL uplink model update, referred to as over-the-air (OTA) federated edge learning (Air-FEEL) [21]–[23]. FL model updates using analog OTA computation have first been proposed in [21]. Due to the large size of model parameters to be transmitted, one-bit quantization has been investigated in this scheme [22]. In addition, authors in [23] proposed an optimal power control to mitigate the analytical convergence bias in Air-FEEL. Despite the high communication efficiency, the AirComp technique introduces errors in the model updates and, in general requires massive client participation in local training. This high client participation can result in elevated network-wide energy consumption for the FL task.

On the other hand, in order to have exact model updates, applying FL in traditional wireless multi-access techniques is also investigated. Due to the limited bandwidth, only a small subset of total users are selected to engage in each FL communication round. Due to the high cost of updating the FL model in terms of wireless resources, the FL training process is urged to be accelerated. As a result, besides appropriate resource allocation strategies that take into account realistic limitations, client selection is another crucial factor in order to accelerate the training process. As selecting users only based on wireless resources often leads to biased learning results when data are non-identically distributed (non-i.i.d.), the user selection needs to be carefully investigated in each communication round. For instance, the work in [24] proposes a joint client selection and bandwidth allocation scheme to ensure long-term convergence performance and long-term energy constraints. The authors in [25] propose a probabilistic client selection method considering the importance of learning from both the user data via gradient divergence and the client's channel state information. Based upon it, in [26], by first deriving an analytical tight bound on the remaining communication round, a probabilistic client selection method having a more theoretical guarantee is proposed. An optimal joint client selection and resource allocation policy is provided in [27] under a total latency budget. A recent work [28] investigates FL in a hierarchical edge learning paradigm with a helper to support the server and client and suggests a joint helper selection and resource allocation. A user scheduling policy based on the measure of the age of update of local model updates and channel characteristics, as provided in [29], assists in improving the FL convergence rate. Maximizing clients per round is another strategy used to speed up FL convergence as applied in [30]–[33]. Theoretically, it has been shown that the convergence rate has a linear speedup with regard to the number of participating users not only in i.i.d. case, but also with non-i.i.d. data: [34] assuming the loss function is strongly convex, [35] with non-convex loss function but using two-sided learning rates. However, in these works including our earlier work [1], selecting more clients in a resource-limited setting always results in choosing clients with better channel conditions since few resources were required to achieve a successful transmission. The FL algorithm converges indeed faster but to a global model biased toward good channel clients, which is undesirable especially under high data heterogeneity correlated with the clients' spatial distribution.

In addition, the aforementioned works overlook the unreliability of wireless links. The transmission itself can be erroneous. The simple retransmission strategy leads to lengthy training times. However, simply dropping erroneous packets without further consideration causes convergence bias towards good channel users' data distribution and introduces another degree of heterogeneity. Therefore, it is important to investigate the effect of dropping erroneous packets on learning performance. A joint learning and communication scheme considering packet error is proposed in [36] by formulating a user selection optimization problem based on the convergence analysis of FL. However, the proposed user selection scheme is only feasible for balanced and i.i.d. dataset. The authors in [37] analyze three FL client selection methods in wireless networks under different signal-to-noise regimes. It, however, requires full involvement of all users' computation throughout each training round to exploit instantaneous channel state information for decision. The article [38] proposes an unbiased aggregation method based on averaging the model with correction weights. A client selection strategy is also proposed. It prioritizes users with weak channels to be selected more for fairness concerns. However, selecting weak channel users in a wireless network may be counter-intuitive and inefficient.

As described previously, choosing suitable users for training to accelerate convergence without introducing convergence bias while preserving energy efficiency, which prevents full computation participation, is a challenging issue. In this work, we present a combined energy-aware client selection and resource allocation policy for accelerating training while avoiding statistical bias due to the previous client selection method by adding a correction term that grows during the training process, it should be emphasized that the methodology applies not only to wireless systems that include packet dropping but can also serve as a general convergence correction framework to all other convergence acceleration client selection strategies based on solving an optimization problem that may cause FL convergence bias. Moreover, existing client selection strategies based on a combinatorial optimization problem [30]–[33] including our earlier work [1] are NP-hard and are not scalable with the dimension of the number of clients in an FL cross-device setting. In this work, we have proposed a very low complexity solution to the complex combinatorial client selection problem. Our main contributions are as follows:

- *For the first time*, we provide a rigorous convergence proof of the FedProx learning scheme in the presence of wireless transmission errors and aggregation weight correction for packet error. The novel convergence lemma ensures theoretical convergence when partial computation stragglers, non-i.i.d. data distribution, and packet error under fading channels are all included.
- We formulate an optimization problem addressing joint client selection and resource allocation to improve the convergence rate of the FL algorithm. The client selection takes into account the *effective participating clients* that includes the importance (to training) and average channel conditions, and certain random weights that change dynamically during the training process by using specifically designed function. This ensures rapid and stable early convergence while preventing biased convergence.
- FL cross-device scenario is proposed for large-scale applications involving hundreds to thousands of participating users and requiring numerous communication rounds to achieve convergence. As a result, the client selection method should be low complexity. We propose an efficient method for solving the proposed optimization problem based on the partial Lagrangian relaxation approach. A closed-form expression of the Lagrange dual function is provided, resulting in low complexity of the method.
- We validate the efficiency and accuracy of our suggested strategy for solving the optimization problem through simulation. We confirm our client selection is both accurate and fast learning convergent by testing on two datasets with different non-i.i.d. distributions.

The rest of the paper is organized as follows: Section II describes the system model and reviews traditional FedProx. In Section III, we analyze the convergence of the learning problem and formulate the convergence acceleration client selection problem. Section IV presents the optimal solution to the formulated problem and an efficient low-complexity suboptimal solution. We present simulations results in Section V and conclude the paper in Section VI.

II. System Description

A. System Model

Consider a wireless network of one base station (BS) and a set of N devices performing FL tasks. The distance between the BS and the device $k \in \{1, \dots, N\}$ is denoted by d_k. We assume $K \in \mathbb{N}^*$ resources blocks (RBs) are available for transmission during the training.

In FL, a learning task is performed cooperatively by the BS and user equipment (UEs). At each training round, the BS transmits the global model to K selected UEs via downlink transmission, which is assumed to be reliable due to BS's sufficient communication resources. Then, UEs perform the training with their own collected data. After certain local training epochs, selected UEs send their local models via uplink transmission to BS for aggregation to update the global model in a time-slot manner under a time budget of T for each user. The transmission is subject to Rayleigh fading and free space path loss (FSPL), where the channel gain is denoted by $h_k = \frac{o_k}{\text{FSPL}(d_k^2)}$ with o_k Rayleigh fading coefficient.

Due to limited power and limited resources of local devices, the uplink transmissions are not always reliable. Inspired by [36], we use a cyclic redundancy check (CRC) mechanism to check the data error in received local FL models at the BS. The packet is dropped once errors are detected during decoding. Retransmissions are omitted to maintain latency and communication efficiency, as the size of the model to learn may be substantial.

We considered the average packet error rate for user k over quasi-static Rayleigh fading and adopted the error rate expression according to [39] as the following:

$$q_k = 1 - \exp\left(-\frac{m}{\overline{\text{SNR}}_k(P_k)}\right), \tag{1}$$

with $m > 0$ a waterfall threshold, $\overline{\text{SNR}}_k(P_k) = \mathbb{E}_{h_k}[\frac{P_k h_k}{\mathcal{B}^U N_0}] = \frac{P_k \bar{h}_k}{\mathcal{B}^U N_0}$ the average signal-to-noise ratio (SNR); $\bar{h}_k$ is the average channel condition of k UE $\mathbb{E}_{h_k}[h_k]$; $\mathcal{B}^U$ is the uplink bandwidth; N_0 is the power spectral density of additive white noise; P_k is the transmit power of user k. This packet error model is tight in this scenario because of the large size of the local model to transmit. It is assumed that the whole packet is transmitted as a single packet[1]. We denote a multivariate random variable $\boldsymbol{h} = (h_1, ..., h_N)$ with h_k representing channel k's gain. We assume h_k between different UEs are all independent of each other and we denote q'_k the

[1] The assumption does not impact the validity of this work's designs because the packet error rate model is independent of packet size and a system design without this assumption would indicate that the overall success packet error rate will be the product of all packet successful transmissions. The system behavior stays similar and does not impact the overall idea of this work. More accurate design may be developed in our future work.

instantaneous packet error rate and $q_k = \mathbb{E}_{h_k}[q_k']$ the average packet error rate.

The computational power can be calculated as in [19], [40]. UE k spends the amount of computation energy:

$$E_k = \kappa \omega_k^2 C_k I_k J_k, \tag{2}$$

where κ is the effective switched capacitance that depends on the chip architecture, ω_k the computation capacity of user k, C_k (cycle/sample) the number of CPU cycles required for computing one sample data at user k, I_k the number of local iterations (number of local epochs) at user k, and J_k the local dataset size.

We assume that all users have the same computational power (ω_k, C_k are equal) and have the same amount of training epochs (I_k are equal). The system's heterogeneity exists as a result of each user's uneven quantity of data collected. Then, the total amount of energy spent on computation during the training round t can be expressed as follows:

$$E_{\text{train}}^{(t)} = \sum_k E_k = \theta \sum_{k \in S_t} J_k, \tag{3}$$

with $\theta = \kappa \omega_k^2 C_k I_k$ as ω_k, C_k, and I_k are equal among clients due to the previous assumption. The subset $S_t \subset \{1, ..., N\}$ corresponds to users selected for training at communication round t.

B. Traditional FedProx Federated Learning

We denote that each UE k collects and possesses a local dataset of input data $X_k = \{x_{k1}, ..., x_{kJ_k}\}$ and label of $Y_k = \{y_{k1}, ..., y_{kJ_k}\}$, where J_k defined as in (2). Then, the local loss function is defined as $F_k(w) = \frac{1}{J_k}\sum_{l=1}^{J_k} \ell(x_{kl}, y_{kl}; w)$, where $\ell(x_{kl}, y_{kl}; w)$

is the loss of the prediction on sample pair (x_{kl}, y_{kl}) with the model w. With the local data size ratio of user k, i.e., $p_k = \frac{J_k}{\sum_k J_k}$, the objective of the training of FL algorithm is to minimize the global loss function:

$$\min_w \quad f(w) = \sum_{k=1}^{N} p_k F_k(w) = \mathbb{E}_k[F_k(w)], \tag{4}$$

FL with FedProx constitutes the following steps:

1) Client selection and model broadcasting In the traditional FedProx algorithm, the BS selects a subset S_t of length K at random according to the selection probabilities p_k and sends the current global model $w^{(t)}$ of communication round t to local users.

2) Local training In FedAvg, $w_k^{(t+1)}$ is computed by several steps of stochastic gradient descent (SGD) directly over F_k. However, when strong statistical heterogeneity (non-i.i.d.) is present, i.e. F_k is different from f, the local model may converge to F_k and diverge from the global model f. FedProx is more stable and has better convergence with non-i.i.d. data because the proximal term includes a regularization term that encourages the local model to remain close to the current model.

In FedProx, UEs receive the global model, then approximately compute the proximal value of F_k (e.g., by certain steps of SGD):

$$w_k^{(t+1)} \approx \text{prox}_{\frac{1}{\mu}F_k}(w^{(t)}), \tag{5}$$

where $\text{prox}_{\frac{1}{\mu}F_k}(w^{(t)}) = \arg\min_w \left[z_k(w; w^{(t)})\right]$, with $z_k(w; w^{(t)})$ the regularized local loss function $z_k(w; w^{(t)}) = F_k(w) + \frac{\mu}{2}||w - w^{(t)}||^2$ with $\mu \geq 0$.

3) Model updates and server aggregation Then, each selected user sends its updated model to the server for aggregation:

$$w^{(t+1)} = \frac{1}{K}\sum_{k \in S_t} w_k^{(t+1)}, \tag{6}$$

where $w^{(t+1)}$ will be the global model to be sent to selected clients at the communication round $t + 1$. In the above traditional FedProx, the aggregation method is based on the assumption of error-free communications, whereas packet errors are ignored.

III. Convergence Analysis and Problem Formulation

In this section, we first show a necessary modification to the aggregation expression to account for the packet error. Subsequently, a convergence analysis of FedProx with this aggregation expression is provided. Then, we formulate an optimization problem aimed at maximizing the effective participating users under K wireless resource blocks to accelerate the convergence and propose a dynamic parameter to mitigate late convergence bias.

A. Fair Aggregation under Packet Errors

To start with, we in this subsection investigate a fair aggregation considering packet errors. In particular, we consider a practical assumption that the uplink transmissions are not arbitrarily reliable. We define $Z_k^{(t)} \in \{0, 1\}$, where $k \in \{1, \ldots, N\}$, a binomial distributed random variable of successful uplink transmission for user k at the communication round t, i.e., $Z_k^{(t)} \sim B(1 - q_k'^{(t)})$ and $Z_k^{(t)} = 1$ indicates a successful uplink transmission from device k, while $Z_k^{(t)} = 0$ when error occurs. We define $\boldsymbol{Z}^{(t)} = (Z_1^{(t)}, \ldots, Z_N^{(t)})$, $\boldsymbol{h}^{(t)} = (h_1^{(t)}, \ldots, h_N^{(t)})$ the instantaneous channel gain at t, and $\boldsymbol{q} = (q_1, \ldots, q_N)$ (as it was defined as an expected value independent of t). Inspired by [38], we propose a modified aggregation expression for FedProx considering packet error as:

$$w^{(t+1)} = \frac{1}{K}\sum_{k \in S_t}\left(\frac{Z_k^{(t)}}{1 - q_k} w_k^{(t+1)} + (1 - \frac{Z_k^{(t)}}{1 - q_k})w^{(t)}\right), \tag{7}$$

which is equivalent to

$$w^{(t+1)} = w^{(t)} + \frac{1}{K}\sum_{k \in S_t}\frac{Z_k^{(t)}}{1 - q_k}(w_k^{(t+1)} - w^{(t)}). \tag{8}$$

We can show that the expected value to packet error and channel fading of the update remains the same expression as (6):

$$\mathbb{E}_{\boldsymbol{Z}^{(t)}, \boldsymbol{h}^{(t)}}[w^{(t+1)}] = \frac{1}{K}\sum_{k \in S_t} w_k^{(t+1)}, \tag{9}$$

with $\mathbb{E}_{\boldsymbol{Z}^{(t)},\boldsymbol{h}^{(t)}}[Z_k] = \mathbb{E}_{\boldsymbol{h}^{(t)}}[1 - q_k'^{(t)}(h_k^{(t)})] = 1 - q_k$.

The previously mentioned FL scheme applied in unreliable network will be denoted as *unbiased* throughout the rest of the paper according to [17], [38]. Any other client selection than selecting client k with p_k probability will be considered as biased client selection.

B. FedProx Convergence under Unreliable Uplink Transmissions

Next, we investigate the convergence under non-i.i.d. data. Assumptions and definitions are similar to those in [15]. We first define a metric for local dissimilarity.

Definition 1. *(B-local dissimilarity) The local functions $(F_k)_{k\in\{1,\ldots,N\}}$ are B-locally dissimilar at w if $\mathbb{E}[\|\nabla F_k(w)\|^2] \leq \|\nabla f(w)\|^2 B^2$. We define a possible candidate $B(w) = \sqrt{\frac{\mathbb{E}[\|\nabla F_k(w)\|^2]}{\|\nabla f(w)\|^2}}$ for $\|\nabla f(w)\| \neq 0$ where ∇ is the gradient operator.*

The non-i.i.d. data distribution would manifest in the difference of local loss function F_k and its gradient. We state the assumption of dissimilarity based on the previous definition. The following assumption is first proposed in [15] for FedProx convergence analysis and shown in [15, Corollary 10] to be equivalent to the commonly-used bounded variance assumption, e.g., in [17].

Assumption 1. *(Bounded dissimilarity) For every $\epsilon > 0$, there exists $B_\epsilon > 0$ such that the local dissimilarity of $(F_k)_{k\in\{1,\ldots,N\}}$ $B(w)$ verifies that $B(w) \leq B_\epsilon$, $\forall w \in S_\epsilon^c = \{w|\ ||\nabla f(w)||^2 > \epsilon\}$.*

We further define the loss function's regularity and convexity-related assumption.

Assumption 2. *(Loss functions regularity) The loss function F_k are assumed non necessarily convex, L-Lipschitz smooth, i.e. for any x and y, $\|\nabla F_k(x) - \nabla F_k(y)\| \leq L\|x - y\|$, and in addition, there exists $L_- > 0$, such that $\nabla^2 F_k \succeq -L_- I$, with $\bar{\mu} = \mu - L_- > 0$.*

The L-Lipschitz regularity can be verified by all neural networks according to [41] and the bounded non-convexity assumption may hold by all loss functions if the weight space considered is constrained within a compact set throughout the training, and the lower bound assumption follows by the continuity of $\nabla^2 F_k$ in a compact set.

For allowing partial computation of stragglers, in the following convergence analysis, we denote $\gamma \in [0, 1]$ the local inexactness of local solution w^* defined by $||\nabla z_k(w^*; w_0)|| \leq \gamma||\nabla F_k(w_0)||$. Then, based on the above assumptions and definitions, we provide the convergence analysis of FedProx considering packet error rate in the following lemma.

Lemma 1. *(Non-convex FedProx convergence considering packet error rate) Suppose that $w^{(t)}$ is not a stationary solution. Let Assumption 1 ($B(w^{(t)}) \leq B$) and Assumption 2 hold. If μ, K, and γ defined previously in FedProx algorithm are chosen such that $\rho > 0$ defined as,*

$$\rho = \left(\frac{1}{\mu} - \frac{\gamma B}{\mu} - \frac{B(1+\gamma)\sqrt{2C(\boldsymbol{q})}}{\bar{\mu}\sqrt{K}} - \frac{L(1+\gamma)^2 B^2}{2\bar{\mu}^2}\right.$$
$$\left. - \frac{LB^2(1+\gamma)^2}{\bar{\mu}^2 K}\left(2\sqrt{2KC(\boldsymbol{q})} + 2C(\boldsymbol{q})\right)\right) > 0, \quad (10)$$

with $C(\boldsymbol{q})$ the parameter in addition to the traditional FedProx convergence theorem defined as

$$C(\boldsymbol{q}) = 1 + \frac{q_{\max}}{1 - q_{\max}}\left(1 + \frac{K}{2}\right), \quad (11)$$

where $q_{\max} = \max_k q_k$ is the maximum average (over fading) packet error rate of each user. At communication round t, we have the following expected decrease in the global objective:

$$\mathbb{E}_{S_t,\boldsymbol{h}^{(t)},\boldsymbol{Z}^{(t)}}[f(w^{(t+1)})] \leq f(w^{(t)}) - \rho||\nabla f(w^{(t)})||^2. \quad (12)$$

where S_t is the set of K devices randomly selected with probability p_k at communication round t.

The proof of this lemma is shown in the Appendix. The following remark explains the relation between this lemma result and the original FedProx convergence result.

Remark 1. *Lemma 1 shows that FedProx converges considering packet errors, i.e. the expected loss with regard to user selection, packet error, and fading reduces with each round if $\rho > 0$. In case of error-free transmission ($q_{\max} = 0$), i.e., $C(\boldsymbol{q}) = 1$, the original convergence condition of FedProx [15] coincides with our lemma. The function $C(\boldsymbol{q})$ increases monotonically to infinity in $[0, 1]$. To prevent the high instability that may be caused by $q_{\max}$ close to 1, i.e., $C(\boldsymbol{q}) = +\infty$, a constraint in the user selection is added to discard all users that have an average error probability higher than 0.9 to ensure stability of* (8).

The convergence rate of the FL algorithm can be inherited from FedProx [15] and is shown in the following remark.

Remark 2. ***Convergence rate**: According to [15], given $\epsilon > 0$, assume $B \geq B_\epsilon$ and all previous assumptions hold for each iteration t. Then, after $\mathscr{T} = O(\frac{f(w^{(0)}) - f^*}{\rho\epsilon})$ communication rounds,*

$$\frac{1}{\mathscr{T}}\sum_{t=1}^{\mathscr{T}} \mathbb{E}_{S_t,\boldsymbol{Z}^{(t)},\boldsymbol{h}^{(t)}}[||\nabla f(w^{(t)})||^2] \leq \epsilon. \quad (13)$$

Then we have,

$$\frac{1}{\mathscr{T}}\sum_{t=1}^{\mathscr{T}} \mathbb{E}_{S_t,\boldsymbol{Z}^{(t)},\boldsymbol{h}^{(t)}}[||\nabla f(w^{(t)})||^2] = O\left(\frac{f(w^{(t)}) - f^*}{\rho\mathscr{T}}\right). \quad (14)$$

It should be pointed out that the value of ρ also depends on the packet error rate.

The following remark analyzes the impact of wireless communication parameters on the FL convergence rate.

Remark 3. *In terms of wireless communication, the parameter ρ is only dependent on $\frac{C(\boldsymbol{q})}{K}$. All other parameters are loss* *function-related or learning-related. The term $\frac{C(\boldsymbol{q})}{K}$ can be written and derived as $\frac{C(\boldsymbol{q})}{K} = \frac{2+K}{2K(1-q_{\max})} - \frac{1}{2}$. Therefore, the term $\frac{C(\boldsymbol{q})}{K}$ is decreasing with regard to both K and $(1-q_{\max})$.*

We aim at improving the convergence rate, which directly depends on ρ. However, it seems intractable due to the complexity of the expression of ρ. We observe that ρ is highly dependent on both factors K and $1-q_{\max}$ according to Remark 3. As K may be constrained by the available RB for transmission, we will attempt to maximize the convergence rate by optimizing a joint quantity under this wireless constraint in the following section.

C. Problem Formulation

Increasing the convergence rate has the advantage of reducing the number of communication rounds required to converge, which results in significant time and energy savings for computation and communication. As mentioned in the introduction and Remark 3 where the monotony of convergence rate with regard to K and $1 - q_{\max}$ is specified, the number of successfully transmitted participating clients per round can be maximized in order to increase the convergence rate at the cost of FL convergence bias. In this work, we are considering K fixed resource blocks in the network and we will instead attempt to maximize the number of effective participating users, as explained in the following.

When packet errors occur, the actual number of users contributing to training is the number of successfully transmitted users instead of K, called *effective* participating users written as:

$$\sum_{k=1}^{N} a_k^{(t)} \eta_k^{(t)} (1 - q_k(P_k^{(t)})), \tag{15}$$

with $a_k^{(t)} \in \{0, 1\}$ represents the user selection, $\eta_k^{(t)} > 0$ the importance attached to user k at the communication round t which will be specified later, and $P_k^{(t)}$ the power allocation at t. The term *effective* stands for both channel quality and local data importance captured by $\eta_k^{(t)}$, the joint consideration of both terms exists already in the literature, e.g. [25]. Assuming that $\eta_k^{(t)}$ captures perfectly the importance of each local client's data to the global model training, in this case, finding users with largest $\eta_k^{(t)}(1-q_k(P_k^{(t)}))$ results in finding more valuable clients to the training which are susceptible to be successfully transmitted. Thus, our objective is to maximize the sum of this quantity for improving the convergence rate under a given K resource block.

The weights $\eta_k^{(t)} > 0$ should reflect dynamically the importance of each user k model during the training process. If $\eta_k^{(t)}$ were chosen constant throughout the training, for example $\eta_k^{(t)} = p_k$, the same users will be chosen in each round because only the average channel condition is taken into account for client selection, as the training duration is typically too long for the channel to remain static. Therefore, we follow the similar approaches proposed in [25], [28], where the importance weights $\eta_k^{(t)}$ are introduced to take into account each user's model importance.

However, even if we assume that $\eta_k^{(t)}$ completely captures each user's model importance, this user selection is still biased toward good channel users, regardless of the design of $\eta_k^{(t)}$. In cases of a strong correlation between channel condition and data distribution, prioritizing good channel user updates leads to inaccurate or unstable convergence. Introducing some random explorations stabilizes the convergence. Therefore, it is necessary to incorporate random weights representing non-biased user selection and increase their relative weights during training.

As a result, we introduce the integer $\xi_t \in \{0, 1, ..., N\}$ to represent the progression of the training process. Based on this value, the method balances the convergence acceleration user selection with $\eta_k^{(t)}(1 - q_k(P_k^{(t)}))$ and the non-biased random user selection. It is defined as the number of users which have been selected and have successfully transmitted their updates at least one time until the communication round t. We define the set $\bar{S}_t$ of the set of selected users who have successfully transmitted their updates. We further denote the set of all users who successfully participated in the training up to the end of communication round t by $\mathcal{T}_t = \bigcup_{r=0}^{t} \bar{S}_r$. The parameter ξ_t can be defined as the cardinality of the set $\mathcal{T}_t$:

$$\xi_t = \text{card}(\mathcal{T}_t). \tag{16}$$

Next, we introduce the weights deduced from ξ_t for balancing the convergence acceleration scheme and unbiased convergence. We define the function value $\varphi(\xi_t)$ for the convergence acceleration term to maximize the effective participating users, and $\psi(\xi_t)$ for fully random unbiased user selection strategy only based on $\{p_k\}_k$. Note that the system benefits from random user selection for avoiding convergence bias, only once the majority of the user's data has been incorporated into our model and the convergence acceleration has occurred. As a result, the intended function shape is flat for small values of ξ_t in order for the convergence-accelerating client selection to be dominant, and steep for large values of ξ_t close to N (the maximum of ξ_t. Therefore, for satisfying such function shape, we construct the following functions:

$$\forall \xi_t = 1, \ldots, N, \; \varphi(\xi_t) = \frac{1 - e^{-\frac{N-\xi_t}{N}M}}{1 - e^{-M}}, \; \psi(\xi_t) = 1 - \varphi(\xi_t), \tag{17}$$

where $M > 0$ is an arbitrary shape factor. The bigger M is, the slower φ will decrease at the beginning and the steeper at the end, as shown in the illustration of some cases in Fig. 1. Therefore, the convergence acceleration term dominates at the start of the training but diminishes when most users' updates have been taken into account to ensure accurate convergence. A similar decaying and increasing function approach has been proposed in [42]. The relative novelty of this work lies mainly in the use of ξ_t, which is more adapted to packet error scenarios, the two terms that are weighted, and the function shapes.

For the simplicity of notation, we omit the index t for the decision variable $\boldsymbol{a}$ and $\boldsymbol{P}$. As a result, the proposed optimization to jointly accelerate convergence and avoid bias

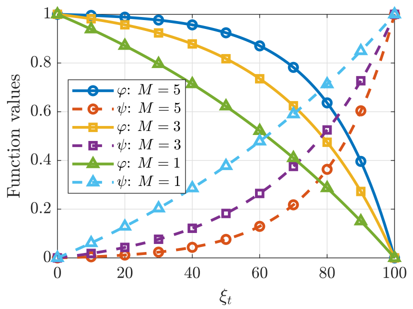


Fig. 1. Illustration of φ and $\psi = 1 - \varphi$ to ξ_t with different values of shape factor M for a total number of users of $N = 100$.

for each communication round t is formulated as follows:

$$\max_{\boldsymbol{a},\boldsymbol{P}} \quad \sum_{k=1}^{N} a_k \left(\eta_k^{(t)} (1 - q_k(P_k)) \varphi(\xi_t) + \Lambda_{k,t} \psi(\xi_t) \right), \tag{18a}$$

$$\text{s.t.} \quad \boldsymbol{a} \in \{0,1\}^N, \tag{18b}$$

$$\sum_{k=1}^{N} a_k = K, \tag{18c}$$

$$\rho(q_{\max}(\boldsymbol{a} \odot \boldsymbol{P})) > 0, \tag{18d}$$

$$\forall k \in \{1, \dots, N\},\ a_k P_{k,\min} \leq a_k P_k \leq P_{\max}, \tag{18e}$$

$$\sum_k a_k (P_k + \frac{\theta J_k}{T}) \leq \frac{E_{\max}}{T}, \tag{18f}$$

where the parameter $\Lambda_{k,t}$ is taken as a random weight as follows:

$$\Lambda_{k,t} = \text{rand}_{k,t}()^{\frac{1}{p_k}}, \tag{19}$$

where $\text{rand}_{k,t}()$ is the sample for UE-k at the communication round t from a uniform random distribution over $[0, 1)$. It has been shown in [43] that selecting K greatest value of $\Lambda_{k,t}$ at each communication round is equivalent to random weighted selection without replacement of K users with the weights $\{p_k\}_{k=1}^N$. Note that the client selection method eventually converges, i.e., when $\xi_t \approx N$, and will not stall at some small and particular sets of clients, which could damage the convergence performance under heterogeneity, because when ξ_t is small, the design of $\eta_k^{(t)}$ favors selecting clients that have not successfully updated their models. When ξ_t is large, the random weights dominated by $\psi(\xi_t)$ will have the chance to pick clients that have not updated yet and increase ξ_t. The weighted random selection fully takes place after $\xi_t \approx N$ ensuring a non-biased FL convergence.

The constraint (18b) imposes that $a_k = 1$ if user k is selected for participation and $a_k = 0$ otherwise. The constraint (18c) limits that only K users are selected in each round for computation and communication due to limited wireless communication resources. The convergence condition of FedProx is ensured by (18d). The operator $\odot$ denotes element-wise multiplication for vectors. The constraint indicates that the selected clients' power must ensure a low packet error rate in order for the FL algorithm to converge. Constraint (18e) shows the power of transmission is limited by the maximum capacity of the device and that a lower threshold of $\overline{\text{SNR}}_k$ for sufficient reliable transmission is set at: $\frac{\overline{\text{SNR}}_k(P_{k,\min})}{m} = \frac{P_{k,\min}\bar{h}_k}{m\mathcal{B}^U N_0} = \frac{1}{2}$. The lower power constraint also serves to preserve stability while aggregation (8). The client selection relies only on average channel information $\bar{h}_k$ since uplink transmission occurs after users receive and train their models and so the instantaneous channel information can not be used. A total network-wide energy budget $E_{\max}$ for training (computation and communication) is imposed in (18f) as the network should be energy-aware and needs to support other missions at the meantime.

D. Choice of $\eta_k^{(t)}$

The choice of the parameter $\eta_k^{(t)}$ is critical for ensuring convergence acceleration by capturing the importance of each client's data to the training, especially when the data are non-i.i.d.. For example, if $\eta_k^{(t)}$ were chosen uniformly for all users, the optimal solution to the problem (18) is always to select the K best average channel users. If $\eta_k^{(t)}$ were kept constant throughout the training process, the same K users are selected at each round. Then, the parameter ξ_t describing the training process remains unchanged according to (16). The user selection is stalled at this point and the final convergence is strongly biased toward these clients.

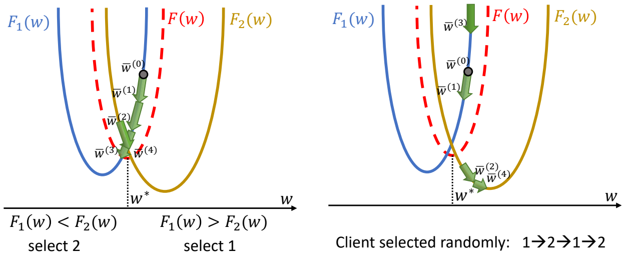
To address this issue, factor $\eta_k^{(t)}$ needs to represent the importance of the local data to the current global model. Several candidates exist in the literature such as the Age-of-Updates (AoU) [29] or the local loss values $F_k(w^{(t)})$ with $w^{(t)}$ the current global model as in [44], [45]. We will use local loss values as η_k throughout the rest of this work. Simulation results in section V will show that choosing AoU or local loss gives comparable results.

In fact, simply selecting clients with higher loss can result in faster convergence [44] and a simple illustration of how this works can is provided in Fig. 2. As it is a special case of FL, we adopt a different notation than formerly introduced, with $\bar{w}^{(i)}$ the model at the communication round i. Random client selection can lead to updates that deteriorate the training by deviating from the optimal model w^* as shown by $\bar{w}^{(2)}$ and $\bar{w}^{(4)}$ in Fig. 2b, whereas selecting higher loss clients always yields updates close to the optimal as in Fig. 2a. Local loss values are thus an excellent candidate for representing the importance of the local model to the global model.

The losses of the current global model on local datasets are inaccessible to the BS unless additional communication rounds and forward propagation computation are added. In order to save communication/computation rounds for energy and latency concerns, we utilize the accumulated averaged loss over local iterations during training $F_k(w_k^{(t+1)})$ as an approximation to $F_k(w^{(t+1)})$, as briefly proposed in [44]. Therefore, we have

$$\eta_k^{(t+1)} = \frac{1}{E} \sum_{l=1}^{E} \frac{1}{|\xi_k^{(l)}|} \sum_{\xi \in \xi_k^{(l)}} f(w_{k,l}^{(t+1)}; \xi) \tag{20}$$

where E the number of epochs, $w_{k,l}^{(t+1)}$ is the model of the training round $t+1$ for user k at the l-th epoch, we have $w_k^{(t+1)} = w_{k,E}^{(t+1)}$ and $\xi_k^{(l)}$ the mini-batch at epoch l. These values are recorded during the training and can be sent together



(a) Selecting higher loss clients (b) Random client selection

Fig. 2. An simple example illustrating the advantage of selecting higher loss clients: w^* the optimal model; one client selected at each communication round with $\bar{w}^{(i)}$ the global model at communication round t.

with the model parameter updates. The additional cost is negligible considering the size of model parameters.

E. Method Summary

The whole algorithm is shown in Algorithm 1. In addition to applying FedProx in a packet loss scheme, a client selection considering jointly training and transmission packet error is proposed. The cost is negligible as it only requires each user to record the average loss during training and transmit it to the BS for client selection, together with the updated models and values, to update the value $\eta_k^{(t)}$. The BS also records how many times each user has successfully transmitted for updating ξ_t. The model aggregation formula is applied to a packet error scenario with (8).

Algorithm 1: Proposed whole framework

```
Initialize: $K$, $T$, $\mu$, $\gamma$, $w^{(0)}$, all wireless factors, $\boldsymbol{S}^{(0)}$, $\boldsymbol{\eta}^{(0)} = 0$
for $t = 0, 1, \dots$ do
    BS selects a subset $S_t$ of length $K$ solving the problem (18) depending on $\{\eta_k^{(t)}\}_{1 \le k \le N}$, $\xi_t$ and other average wireless factors.
    The BS sends $w^{(t)}$ to all selected devices.
    for selected device $k \in S_t$ in parallel do
        Compute $w_k^{(t+1)}$ that is a $\gamma_k^{(t)}$-inexact minimizer of $\arg\min_w z_k(w, w^{(t)})$ (6) and store the loss value at each epochs.
        Compute the average loss value $\eta_k^{(t+1)}$.
        Send $w_k^{(t+1)}$ and $\eta_k^{(t+1)}$ back to the BS.
    end
    BS records successfully transmitted users to update the set of successfully updated clients set $\mathcal{T}_t$ in order to calculate the value $\xi_t$ by (16).
    BS aggregates the received $w_k^{(t+1)}$ according to (8) and obtains $w^{(t+1)}$.
end
```

IV. Joint User Selection and Power Allocation for FL Under Packet Error

The problem formulated in (18) is a non-linear mix-integer optimization problem. The solution is typically NP-hard. In this section first the naif exact approach solution is explained before presenting the low-complexity sub-optimal solution.

A. Naif Approach

Since the variable $\boldsymbol{a} \in \{0, 1\}^N$ is binary, the optimization problem (18) is a mixed integer problem, which can be solved by decomposing the original problem via going through all possible combinations of $\boldsymbol{a}$ while solving those sub-problems with regard to $\boldsymbol{P}$.

In particular, for any user selection $\tilde{\boldsymbol{a}}$, where $\tilde{\boldsymbol{a}} \in \{0, 1\}^N$ and $\sum_{k=1}^N \tilde{a}_k = K$, we denote the set $\tilde{S}_t = \{k | \tilde{a}_k = 1\}$. We have the corresponding sub-problem with variable of $\boldsymbol{P}$:

$$\begin{aligned} \max_{\boldsymbol{P}} \quad & \sum_{k \in \tilde{S}_t} g_k(P_k) \\ \text{s.t.} \quad & \text{(18d)} - \text{(18f)} \end{aligned} \tag{21}$$

Denoting $g_k(P_k) = \eta_k^{(t)} \exp\left(-\frac{\beta}{P_k \bar{h}_k}\right)\varphi(\xi_t) + \Lambda_{k,t}\psi(\xi_t)$ with $\beta = m\mathcal{B}^U N_0$ (the index t of g_k is omitted for the sake of clarity of notation). The following lemma shows that the sub-problem is a convex optimization problem.

Lemma 2. *The objective of* (21) *is concave in the feasible region, and therefore problem* (21) *is a convex optimization problem.*

Proof. It can be noticed that the objective is concave only if for any $k = \{1, ..., N\}$, g_k is concave. The second order derivative of g_k is:

$$g_k''(x) = \eta_k^{(t)} \varphi(\xi_t) \frac{\beta}{\bar{h}_k x^3} e^{-\frac{\beta}{\bar{h}_k x}} \left(\frac{\beta}{\bar{h}_k x} - 2 \right). \tag{22}$$

For g_k to be concave, a sufficient and necessary condition is if $\frac{\bar{h}_k x}{\beta} \geq \frac{1}{2}$, i.e. $\frac{\text{SNR}_k(P_k)}{m} \geq \frac{1}{2}$ which is satisfied in constraint (18e) by definition of $P_{k,\min}$. The constraint (18d) can be easily transformed into a convex constraint form by exploiting the monotonicity of ρ with regard to P_k. The constraint is equivalent to having a lower bound of P_k for all selected client k, therefore a convex constraint and can thus be merged into the constraints (18e). □

According to Lemma 2, the optimal solutions can be efficiently found by the standard convex optimization algorithm with a complexity of $\mathcal{O}(K^2)$ [46], [47]. To screen all possibilities of $\boldsymbol{a}$, we have $\binom{N}{K}$ possibilities which have asymptotically $\mathcal{O}(N^K/K!)$ operations, resulting in the overall complexity of $\mathcal{O}(N^K K^2/K!)$.

B. An Efficient Approach via Lagrangian Relaxation

Although the proposed naif approach can provide globally optimal solutions, its complexity is high. Especially, FL scenarios typically involve a huge number of users and require numerous communication rounds. As a result, the efficiency of user selection and power allocation computation has a substantial effect on the global convergence time.

Therefore, we propose an efficient approach based on Lagrange relaxation (LR) [48] to solve Problem (18), where some constraints are included within the objective with the help of

Lagrange multipliers. In particular, we formulate a Lagrangian relaxed problem $\mathcal{P}(\lambda)$ by incorporating (18f) into the objective function with $\lambda \geq 0$ as follows:

$$\begin{aligned} \max_{\boldsymbol{a},\boldsymbol{P}} \quad & \sum_{k=1}^{N} a_k \Big(\eta_k^{(t)} (1 - q_k(P_k)) \varphi(\xi_t) + \Lambda_{k,t} \psi(\xi_t) - \\ & \lambda \big(P_k + \frac{\theta J_k}{T} \big) \Big) + \lambda \frac{E_{\max}}{T}, \\ \text{s.t.} \quad & \text{(18b) – (18e).} \end{aligned}$$

We denote the objective function $\sum_k^N a_k g_k^{(\lambda)}(P_k)$ with $g_k^{(0)} = g_k$ as previously defined and we denote, for UE k, $g_k^{(\lambda)*}$ and P_k^* respectively the optimal value and optimal point of the following one-variable optimization problem:

$$\begin{aligned} \max_{P_k} \quad & \eta_k (1 - q_k(P_k)) \varphi(\xi_t) + \Lambda_{k,t} \psi(\xi_t) - \lambda \big(P_k + \frac{\theta J_k}{T} \big), \\ \text{s.t.} \quad & P_{k,\min} \leq P_k \leq P_{\max}. \end{aligned} \tag{23}$$

The solution of this problem $g_k^{(\lambda)*}$ can be attained, because the objective is continuous and the feasible region is compact. As the constraints of (23) affect only one user at a time, the following lemma indicates that the solution of (23) can easily be derived from the optimum of (23).

Lemma 3. *The optimal solution to the problem* $\mathcal{P}(\lambda)$ *corresponds to*

$$a_k = \begin{cases} 1 & \textit{if } k \in \textit{arg } K\textit{-max } g_k^{(\lambda)*} \\ 0 & \textit{otherwise} \end{cases}$$

and the optimal power is $\{P_k^*\}_{k=1}^N$.

Proof. As the only joint constraint is relaxed into the objective, we can write the problem as:

$$\begin{aligned} \max_{\boldsymbol{a}} \quad & \sum_{k=1}^{N} a_k \max_{P_{k,\min} \leq P_k \leq P_{\max}} g_k^{(\lambda)}(P_k) \\ \text{s.t.} \quad & \boldsymbol{a} \in \{0,1\}^N, \\ & \sum_{k=1}^{N} a_k = K. \end{aligned}$$

The problem is reduced to choose the K highest values of $\max_{P_{k,\min} \leq P_k \leq P_{\max}} g_k^{(\lambda)}(P_k)$. □

It remains to solve the convex optimization (23) (objective concave in the feasible region according to Lemma 2) for each user $k = 1, ..., N$. The following lemma describes the optimal solution to the problem.

Lemma 4. *The optimal solution to the problem* (23) *is:*

$$P_k^* = \begin{cases} P_{k,\min} & \textit{if } g_k^{(\lambda)\prime}(P_{k,\min}) \leq 0 \\ P_{\max} & \textit{if } g_k^{(\lambda)\prime}(P_{\max}) \geq 0 \\ -\frac{b_k}{2\mathcal{W}(-\sqrt{\lambda b_k / c_k}/2)} & \textit{otherwise.} \end{cases} \tag{24}$$

where $b_k = \beta / \bar{h}_k$ *and* $c_k = \eta_k \varphi(\xi_t)$ *and the function* $\mathcal{W}$ *denotes the Lambert W function.*

Proof. Let $\mu_1, \mu_2 \in \mathbb{R}$ be the dual variable relating to the constraint $P_{k,\min} - P_k \leq 0$ and $P_k - P_{\max} \leq 0$. The KKT conditions can be written as:

$$\begin{cases} P_{k,\min} \leq P_k \leq P_{max} \\ \mu_1, \mu_2 \geq 0 \\ \mu_1 (P_k - P_{k,\min}) = 0, \quad \mu_2 (P_k - P_{\max}) = 0 \\ g_k'(P_k) = \lambda - \mu_1 + \mu_2. \end{cases} \tag{25}$$

We remind from the proof in the Lemma 2 that g_k' is non-increasing in the feasible region $[P_{k,\min}, P_{\max}]$. From the KKT conditions:

- If $P_k = P_{k,\min}$, then $\mu_2 = 0$ and $\mu_1 > 0$. We have $g_k'(P_{k,\min}) < \lambda$.
- If $P_k = P_{\max}$, then $\mu_1 = 0$ and $\mu_2 > 0$. We have $g_k'(P_{\max}) > \lambda$.
- If $P_k \in (P_{k,\min}, P_{\max})$, then by complementary slackness, $\mu_1 = \mu_2 = 0$. We have $g_k'(P_k^*) = \lambda$. We remind that λ is not a variable in this optimization problem.

For any $x \in (P_{k,\min}, P_{\max})$, $g_k'(x) = c_k \frac{b_k}{x^2} e^{-\frac{b_k}{x}}$ with $b_k > 0$ and $c_k > 0$ defined previously. The equation can be written as follows:

$$\begin{aligned} \frac{b_k^2}{4x^2} e^{-\frac{b_k}{x}} = \frac{\lambda b_k}{4 c_k} \quad (> 0) \iff & -\frac{b_k}{2x} e^{-\frac{b_k}{2x}} = -\frac{1}{2} \sqrt{\frac{\lambda b_k}{c_k}} \\ \iff & \mathcal{W}\Big(-\frac{1}{2} \sqrt{\frac{\lambda b_k}{c_k}}\Big) = -\frac{b_k}{2x}, \end{aligned} \tag{26}$$

The last equivalence is due to the definition of the Lambert W function. The closed-form solution of P_k^* in (24) follows. The lemma is proved because $g_k^{(\lambda)\prime} = g_k^{(0)\prime} - \lambda$. □

Algorithm 2: Solve (18) by Lagrangian relaxation

```
Initialize: $\varepsilon > 0$, itermax$\in \mathbb{N}^*$, $\lambda_{\min,\text{feasible}} = +\infty$, and
  $\lambda_{\max,\text{unfeasible}} = 0$.
while iter $\leq$ itermax and $\lambda_{\min,feasible} - \lambda_{\max,unfeasible} > \varepsilon$ do
    $\lambda = \frac{\lambda_{\max,\text{unfeasible}} + \lambda_{\min,\text{feasible}}}{2}$.
    Solve the problem $\mathcal{P}(\lambda)$ via Lemma 3 and 4 and obtain
      solutions $(\boldsymbol{a}_\lambda^*, \boldsymbol{P}_\lambda^*)$.
    if $(\boldsymbol{a}_\lambda^*, \boldsymbol{P}_\lambda^*)$ is feasible, i.e. constraint (18f) satisfied then
        $\lambda_{\min,\text{feasible}} = \lambda$
    else
        $\lambda_{\max,\text{unfeasible}} = \lambda$
    end
    iter ← iter + 1.
end
return $(\boldsymbol{a}_\lambda^*, \boldsymbol{P}_\lambda^*)$ solution of $\mathcal{P}(\lambda_{\min,\text{feasible}})$.
```

We can obtain a lower bound of the optimal solution of the initial mixed-integer problem (18) by finding the optimal λ^* minimizing the optimum of $\mathcal{P}(\lambda)$. Strong duality cannot be guaranteed as our initial problem is a mixed-integer problem. In addition, the solution of $\mathcal{P}(\lambda^*)$ does not ensure feasibility in the primal problem. We propose to obtain a sub-optimal feasible solution obtained by Algorithm 2. The algorithm is based on bisection search [49] to find the optimal primal-feasible dual solution λ_{feas}^*. The approach is illustrated in

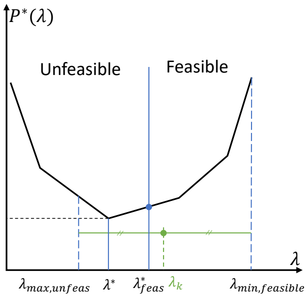


Fig. 3. Explanation of Algorithm 2 for a bisection algorithm to find the best primal-feasible dual problem solution. $\mathcal{P}^*(\lambda)$ is the optimal value of $\mathcal{P}(\lambda)$ in (23).

TABLE I
PARAMETER VALUES USED IN SIMULATIONS

Parameter	Value	Parameter	Value
N_0	-150 dBm/Hz	$\mathcal{B}^U$	1 MHz
P_{max}	10 mW	m	0.023 dB
σ(fading)	1	$E_{\max}$	30 mJ
T_{slot}	1.3 s	frequency	2.4 GHz
κ	10^{-28}	ω_k	2 GHz
C_k	2000	I_k	20
M	3		

Fig. 3. In fact, we initialize with the primal-feasible dual variable bounds $\lambda_{\text{min,feasible}}$ as infinity for primal feasible solution and $\lambda_{\text{max,unfeasible}}$ as zero for ignoring the hard constraint. Note that $\lambda_{\text{max,unfeasible}} < \lambda_{\text{min,feasible}}$ because the relaxed primal constraint becomes tighter when λ becomes greater. The problem $\mathcal{P}(\lambda)$ is solved via Lemma 3 and 4. As a closed-form solution exists in (24), the highest complexity operation is to sort N values when finding the K greatest values of $g_k^{(\lambda)}(P^*_{k,\lambda})$ as in Lemma 3, so has the complexity of $\mathcal{O}(N\log(N))$, since the accuracy of the bisection and other operations are of a constant term with regard to N and K. The bisection search depends on the accuracy of the solution to achieve and is of constant order regarding N and K. Therefore, our solution has an overall complexity of $\mathcal{O}(N\log(N))$ and is well-suited for large-scale FL networks.

V. SIMULATIONS

A. Settings

We assume there are N user devices uniformly distributed over a circular network area of radius $R = 1000\,\text{m}$ (by default) serving by a BS. The $K = 10$ (by default) uplink resource blocks are allocated for model update transmissions. We assume a free space path loss model and Rayleigh fading channels. Other wireless parameters are given in Table I.

FL is performed on two datasets:

a) MNIST dataset [50]: consists of 60000 28 × 28 grayscale images of handwritten digits between 0 to 9;

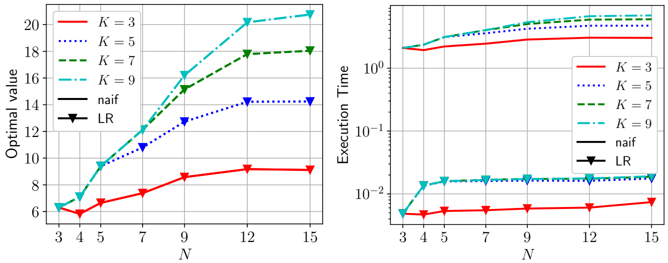
b) CIFAR-10 dataset [51]: consists of 60000 32x32 colour images in 10 classes;

The MNIST and CIFAR-10 datasets are split into an unbalanced quantity of user's data following a power law as done in [15], [17]. Two non-i.i.d. distributions are considered: each device contains samples of two classes of data; data follow Dirichlet distribution with $\alpha = 0.5$, parameter quantifying the similarity of data between UEs as in [52].

For the MNIST dataset, we train a three-layer neural network with 512 hidden units at each hidden layer. We use rectified linear unit as activation functions for the hidden layers. A 20 % dropout layer is employed after each hidden layer to avoid over-fitting. CIFAR-10 dataset is trained on vgg11 [53]. Details of parameters are specified in Table II. Local SGD solver is used for solving the local optimization problem in FedProx. We use batch sizes of 64. We use the learning rate at 0.1 for MNIST and 0.01 for CIFAR-10 and $\mu = 1$ after a grid search. The number of local epochs is set to be $E = 20$.

TABLE II
TRAINING MODELS AND PARAMETERS USED FOR EACH DATASET.

	N	K	Data Distribution	Model
MNIST	500	10	2 digits and Dirichlet	three layer NN
CIFAR-10	500	10	2 digits and Dirichlet	vgg11 [53]



(a) Optimal value attained

(b) Execution time needed.

Fig. 4. Comparison of exact (naif) method and the proposed method (LR) against total number of users N. Lines with markers are obtained by the proposed method, and lines without markers by exact method.

First, we compare the complexity improvement of the algorithm we proposed based on Lagrangian relaxation while maintaining high optimality. Secondly, we show that the client selection strategy proposed increases and stabilizes the convergence speed during training.

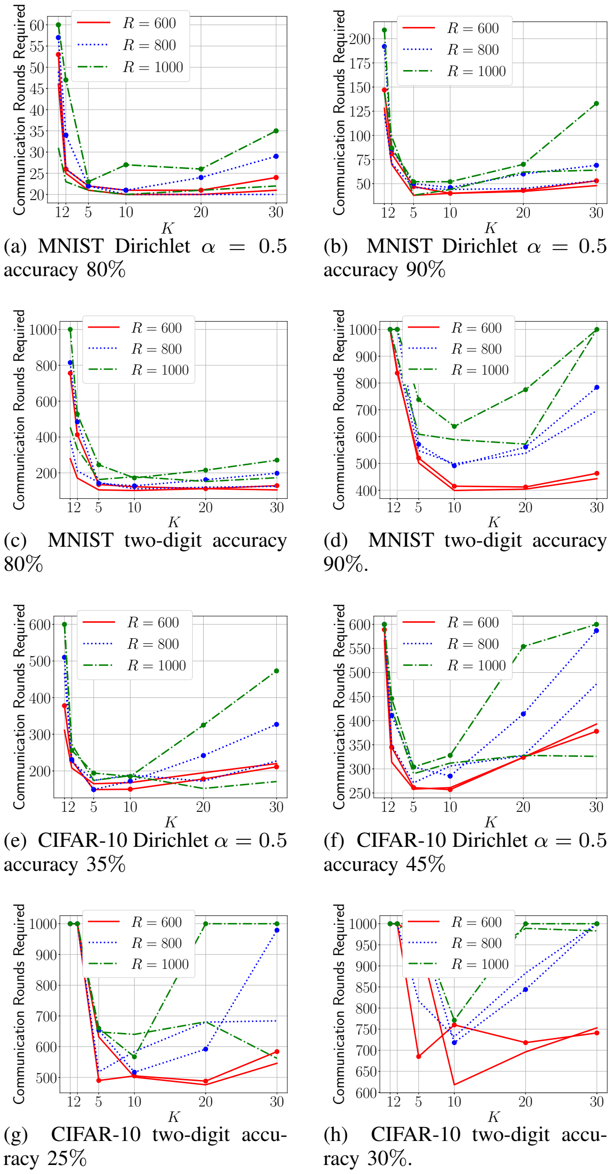
B. Lagrangian Relaxation Optimization Results

We evaluate the performance of Algorithm 2 to solve the optimization problem (18). Only in this subsection, values of $\eta_k^{(t)}$ were simulated as a uniform distribution in $[1, 3]$. The obtained optimal value with different values of N and K are shown in Fig. 4a. We assume that when $N < K$, only N RBs are used. The proposed LR solution is compared to the exact (naif) method presented in IV.A.. We observe that our algorithm (LR) method achieves almost the same optimality as the exact method for all combinations of (K, N).

Execution times are shown in Fig. 4b for the same combinations of (K, N) in the previous comparison. Our algorithm spends on average 100 less time than the exact method by achieving identical performance. This confirms the low complexity advantage of our approach.

C. Optimization Learning Performance Gain

The advantage of the aggregation considering packet error rate as in (8) is already shown in [38]. This aggregation is employed for all following experiments without further specification. We compare the improvement of the convergence rate using our client selection method with four other methods: *best loss*: the K largest approximated loss' users are selected [44]; *uniform sampling*: K users are randomly scheduled at each



(a) MNIST Dirichlet $\alpha = 0.5$ accuracy 80%

(b) MNIST Dirichlet $\alpha = 0.5$ accuracy 90%

(c) MNIST two-digit accuracy 80%

(d) MNIST two-digit accuracy 90%.

(e) CIFAR-10 Dirichlet $\alpha = 0.5$ accuracy 35%

(f) CIFAR-10 Dirichlet $\alpha = 0.5$ accuracy 45%

(g) CIFAR-10 two-digit accuracy 25%

(h) CIFAR-10 two-digit accuracy 30%.

Fig. 5. Number of communication rounds of FL using *weighted random* and our method as client selection method to attain specific test accuracy levels with different numbers of active users K with three levels of cell sizes. A total of 1000 communication rounds are evaluated except 600 for the CIFAR-10 Dirichlet case. Line with marker denotes *weighted random*; line without marker is the proposed method.

round with equal probabilities; *weighted random sampling*: K users are randomly selected with probabilities p_k (data size ratio of user k) with the aggregation weight correction [38]; *best channel*: the K best average channel users are selected; *weighted random w/o aggregation correction*: weighted random sampling but without the aggregation weight correction as in original FL papers [3], [15]. For all the above methods, if the user selection method already considers data size, the aggregation formula is exactly (8) by default. Otherwise, the data size weights p_k are considered in the aggregation.

To confirm the convergence acceleration approach, Fig. 5 and Table III, Table IV show the number of communication rounds needed to attain certain accuracy levels of test accuracy with regard to the number of resource blocks K for MNIST and CIFAR-10 datasets. A maximum of 1000 communication rounds are evaluated for all cases except 600 for the CIFAR-10 Dirichlet distribution case. Results are also compared with regard to different cell radii. The average channel quality is better when the cell radius size is smaller, as we assume N users are uniformly distributed in a cell size of R. “X" in the tables denotes the case where both methods did not reach the accuracy level. A moving average of window size of 11 for Dirichlet distribution and 41 for two-digits case is applied for smoothing. The proposed method is compared with classic user selection method *weighted random*. For MNIST dataset, the proposed method requires almost consistently fewer communication rounds to achieve certain levels of accuracy, except in a few rare cases with minor differences. For some extreme

TABLE III

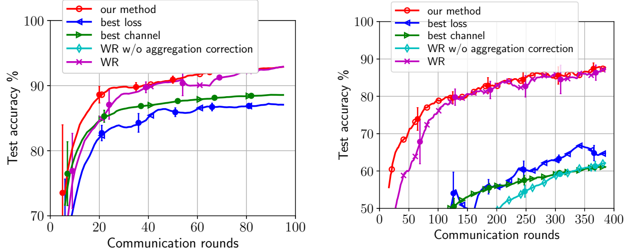
DIFFERENCE OF NUMBER OF COMMUNICATION ROUNDS OF *weighted random* AND OUR METHOD TO ATTAIN SPECIFIC TEST ACCURACY LEVELS $\{80\,\%, 90\,\%\}$ WITH DIFFERENT NUMBERS OF ACTIVE USERS K WITH THREE LEVELS OF CELL SIZES $\{600\,\mathrm{m}, 800\,\mathrm{m}, 1000\,\mathrm{m}\}$. A TOTAL OF 1000 COMMUNICATION ROUNDS ARE EVALUATED. THE POSITIVE NUMBER MEANS THAT OUR METHOD ACHIEVES THE ACCURACY FASTER. THE TEST IS DONE ON MNIST DATASET.

K			1	2	5	10	20	30
Cell radius (m)	non-i.i.d. type	accuracy level						
$R = 600$	Dirichlet	80%	7	1	1	1	1	3
		90%	19	11	9	0	1	5
	2 digits	80%	477	242	30	21	0	24
		90%	X	-10	18	16	8	20
$R = 800$	Dirichlet	80%	8	8	0	1	4	9
		90%	70	15	4	2	15	16
	2 digits	80%	436	273	-5	17	42	73
		90%	0	108	23	-8	23	87
$R = 1000$	Dirichlet	80%	29	24	2	7	5	13
		90%	65	-12	14	8	8	69
	2 digits	80%	546	194	82	-7	63	98
		90%	X	103	129	49	203	X

TABLE IV

DIFFERENCE OF NUMBER OF COMMUNICATION ROUNDS OF FL USING *weighted random* AND OUR METHOD AS CLIENT SELECTION METHOD TO ATTAIN SPECIFIC TEST ACCURACY LEVELS WITH VARIOUS NUMBERS OF ACTIVE USERS K WITH THREE LEVELS OF CELL SIZES. A TOTAL OF 1000 COMMUNICATION ROUNDS ARE EVALUATED EXCEPT 600 FOR THE DIRICHLET CASE. THE TEST IS DONE ON CIFAR-10 DATASET.

K			1	2	5	10	20	30
Cell radius (m)	non-i.i.d. type	accuracy level						
$R = 600$	Dirichlet	35%	45	19	-17	1	-17	-8
		45%	28	27	6	11	1	-5
	2 digits	25%	X	X	-125	25	-5	11
		30%	X	X	-224	105	4	-53
$R = 800$	Dirichlet	35%	37	23	5	-5	57	147
		45%	X	16	4	9	70	127
	2 digits	25%	X	X	36	-65	-9	285
		30%	X	X	190	-19	-3	X
$R = 1000$	Dirichlet	35%	X	9	8	15	218	329
		45%	X	-37	23	36	240	312
	2 digits	25%	X	X	37	-53	317	553
		30%	X	X	63	-107	X	17



(a) 2 digits, window size 31

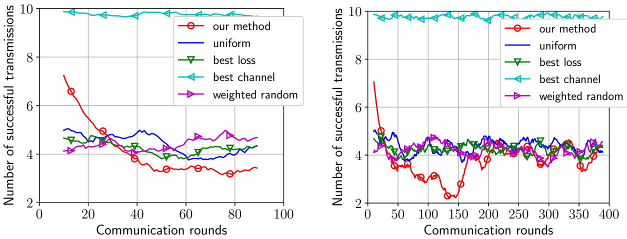
(b) Dirichlet distribution $\alpha = 0.5$

Fig. 6. Test accuracy evolution for MNIST dataset.

cases as $K = 1$ and accuracy of $80\,\%$, the proposed method provides an advantage of 400-500 communication rounds. We observe that the number of communication rounds needed is not monotone to K. This is due to the fact that an overall energy budget exists. More transmission resources here imply a higher user participation. Low energy left for transmission after local training leads to less reliable transmission and then slower convergence. According to the figure and the table, we observe that in general, the proposed method's improvement is more significant for extreme values of K, i.e., $K = 1, 2$ or $K = 20, 30$. The case $K = 30$ for $R = 1000\,\mathrm{m}$ is an exception because both user selections didn't attain 90% of accuracy within 1000 communication rounds. We further notice that bigger cell sizes, so less good average channel conditions of users, resulting in a greater improvement of the proposed methods in terms of communication rounds. As for CIFAR-10 dataset, in general the same observations and conclusions hold. However, we observe there are more cases where the proposed method is outperformed by the baseline, in general with minor differences, except for the two-digits cases when $R = 600\,m$ with significant performance loss. We conjecture the reason is that the higher complexity of the dataset and the model may cause more random behaviors due to the sensibility to the very low number of local dataset sizes and also may make the convergence trend different than MNIST. Despite this, it is worth noting that the improvement of the convergence in some other cases, e.g., $R = 1000\,m$ and $K = 20, 30$ is far more significant, with at least 200 communication rounds of gain.

In the following, we show further details of the simulations and explain why only the results of the proposed method and weighted random were compared previously. Fig. 6 shows the training accuracy evolution of MNIST dataset with two different kinds of non-i.i.d.. All five user selection methods are compared. We notice first that for both non-i.i.d. cases, only the proposed method and the weighted random selection can achieve the highest accuracy level, the difference is more significant when the degree of non-i.i.d. is more important, i.e. for 2 digits data distribution for each user. That illustrates the reason why only these two methods are compared previously. Furthermore, the proposed method converges faster than any other method. For instance, for Dirichlet distribution results, weighted random method is slower than the best channel method to achieve 80% of accuracy, but still slower than the proposed method.

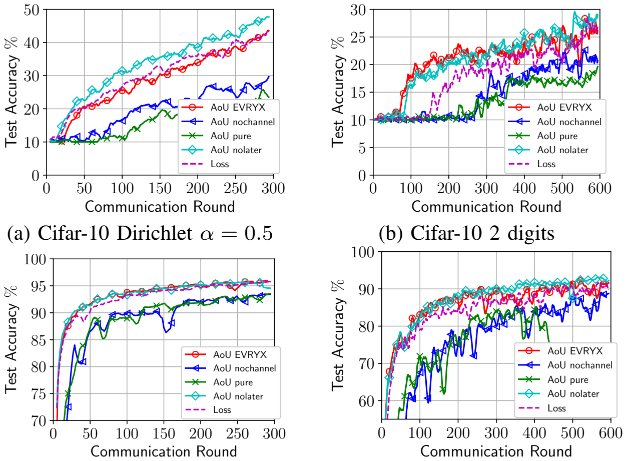
In Fig. 7, a smooth curve of the number of failed transmissions at each communication round is shown. The proposed method starts by taking users with good channel conditions, resulting in lower failing transmission rates. It explains the faster convergence rate of the proposed method compared to others because other methods have, on average, 6 against 4 successful packet transmission during the first 20 communication rounds. The proposed method then seeks to obtain models of less good channel users to improve training accuracy. Comparing the learning performance in Fig. 6, it continues to improve at the same pace as other methods even though the average successful transmission is lower, showing the need to update the appropriate important models. Once the model stabilizes, i.e. $\varphi(\xi_t) = 0$, the proposed method will coincide with weighted random by its own construction to ensure convergence stability.



(a) For Dirichlet distribution $\alpha = 0.5$

(b) For two-digit non-i.i.d.

Fig. 7. Number of successful transmissions occurred at each communication round comparing different client selection methods. Curve smoothed with window size of 21.



(a) Cifar-10 Dirichlet $\alpha = 0.5$

(b) Cifar-10 2 digits

(c) MNIST Dirichlet $\alpha = 0.5$

(d) MNIST 2 digits

Fig. 8. Comparison of different integration of AoU approaches with the proposed η_k as approximated local loss and framework.

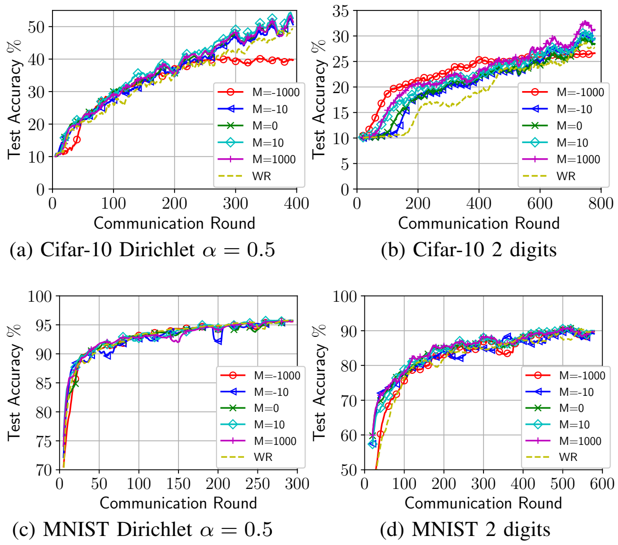
D. Choice and Effect of η_k and M

It is mentioned in III.D. that the choice of the importance metric $\eta_k^{(t)}$ in (20) is only one potential candidate. Choosing it constant and equal is equivalent to *best channel* compared previously and its poor results confirm the importance of choice for it. We provide some results using AoU [29] as η_k in Fig. 8. In the spirit of the ablation study, the following cases are compared:

1) *AoU EVRYX*: $\eta_k =$ AoU, everything else stays the same.
2) *AoU no channel*: We solve (18) considering constant and equal channel term $1 - q_k(P_k)$.
3) *AoU pure*: directly selecting the largest AoU value users (then do a greedy power allocation for this user selection as in [29]).
4) *AoU nolater*: assuming $\varphi = 1$ and $\psi = 0$.
5) *Loss*: the proposed method with η_k as in (20).

We observe that *AoU nochannel* and *AoU pure* performs significantly less well than the other methods, i.e., showing the importance of considering channel condition in the proposed *effective participating clients* user selection (15) jointly with the model importance as in (20).

We observe that *AoU nolater* performs the best in all cases except CIFAR-10 2 digits. *AoU EVRYX* and *Loss* have close



(a) Cifar-10 Dirichlet $\alpha = 0.5$ (b) Cifar-10 2 digits

(c) MNIST Dirichlet $\alpha = 0.5$ (d) MNIST 2 digits

Fig. 9. M effects. Negative value of M denotes the symmetry function of original Ψ with regard to identity with paramter M.

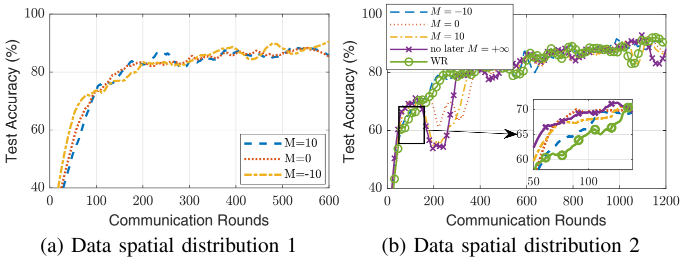
performance. We confirm the effect of *effective participating clients* in user selection and conclude that AoU may be a better metric for model importance as the local loss (20). Other model importance metrics may be considered in this proposed framework.

It can be observed that the performance is better without the impact of the random weights $\Lambda_{k,t}$. It does make sense in the tested scenarios since there is no correlation between local data distribution and UE's channel conditions and the proposed method can explore a large set of clients that have relatively good channels. This can be confirmed in Fig. 9. We consider M negative the function symmetric of φ to $x \mapsto 1-x$ of φ. We note that $\lim_{M\to+\infty} \varphi = 1$, corresponding to the case without the impact of $\Lambda_{k,t}$. We observed that the difference in performance is small and in general the more M is large, the better the performance.

To see the effect of M and $\Lambda_{k,t}$, we evaluate settings with a strong correlation between data distribution and channel condition. We consider two data spatial distributions (DSD) in the non-i.i.d. case with 2 digits data per UE. For DSD 1, the clients with the data label $\{0,1\}$ and $\{1,2\}$ are set close to the BS with a distance of $100\,m$ and all other clients are at $1000\,m$. The DSD 2 is similar to DSD 1 except users with labels $\{8,9\},\{9,0\},\{7,8\}$ are located at $1300\,m$. The results have been shown in Fig. 10. For DSD 2, a large value of M leads to higher convergence speed at the beginning as observed previously, however, a significant accuracy drop occurs just afterward with high values of M. For DSD 1, also smaller values of M is preferable. In the DSD2 case where strong data and space correlation exists, *weighted random* performs well. This confirms the importance of the choice of M: when the data distribution and the UEs' spatial distribution is weak, M is encouraged to be large, and vice versa, especially when some UEs are under weak channel condition.

VI. Conclusion

We developed a low-complexity convergence acceleration FL framework in heterogeneous and unreliable communication networks. In fact, FL must take into account network heterogeneity in terms of statistics and systems, limited communication resources, and the unreliability of wireless communications in order to be applied in practice. This work proposed the use of FedProx to handle heterogeneity. We provided a convergence analysis of FedProx under transmission packet error conditions and proposed a client selection strategy combined with resource allocation to accelerate training convergence by maximizing the *effective participating users* as FL communication rounds reduction is a primary issue. Contrary to most existing optimization problem-based client selection strategies for convergence acceleration, the proposed method simultaneously avoids biased convergence. The approach makes use of a parameter to take into account the learning process progression in order to dynamically adjust the weights between various variables in our client selection strategy, thereby avoiding convergence bias that occurs due to the strong correlation between data distribution and UEs' spatial distribution. In order to guarantee the scalability of the approach and ensure a low wall-clock convergence time, a highly computation-efficient Lagrangian relaxation-based method has been proposed to obtain a sub-optimal solution to the described joint client selection and power allocation problem. According to the simulation results, the proposed efficient sub-optimal solution approaches the optimal solution and has extremely low complexity. The results demonstrate that the proposed client selection technique accelerates the early convergence under no data-spatial correlation case, resulting in fewer communication rounds to achieve learning accuracy levels, while ensuring a stable convergence when the data-spatial correlation is present. In a more realistic wireless communication environment, the packet to send may also be divided into numerous packets, and more complicated retransmission mechanisms may be used. Adapting the proposed analysis and method with more advanced concerns will be addressed in our future work. Reproducible codes are available at https://github.com/paulzhengfr/FedproxPER.



(a) Data spatial distribution 1 (b) Data spatial distribution 2

Fig. 10. M effect under strong correlation between local data distribution and UE's distance to the BS. Data spatial distribution (DSD) 1: users with labels $\{0,1\},\{1,2\}$ located at $100\,m$ and others at $1000\,m$; DSD 2: DSD 1 but with users with labels $\{8,9\},\{9,0\},\{7,8\}$ located at $1300\,m$.

Appendix
Proof of Lemma 1

The packet error that occurred during aggregation only intervenes at the inequality [15, eq. (15)] of FedProx proof which was to show that:

$$\mathbb{E}_{S_t}[||w^{(t+1)} - \bar{w}^{(t+1)}||^2] \leq \frac{1}{K}\mathbb{E}_k[||w_k^{(t+1)} - \bar{w}^{(t+1)}||^2] \leq \frac{2B^2}{K}\frac{(1+\gamma)^2}{\bar{\mu}^2}||\nabla f(w^{(t)})||^2, \tag{27}$$

with $\bar{w}^{(t+1)} = \mathbb{E}_k[w_k^{(t+1)}]$. We need to bound the LHS of the previous inequality by adding the consideration of the packet error and the channel condition, because the actual packet error depends on the instantaneous channel condition. The quantity to bound can be expressed as $\mathbb{E}_{S_t,\boldsymbol{h}^{(t)},\boldsymbol{Z}^{(t)}}[||w^{(t+1)} - \bar{w}^{(t+1)}||^2]$. For notation simplicity, we will omit the index t on $\boldsymbol{h}$ and $\boldsymbol{Z}$.

We replace the term $w^{(t+1)}$ by its aggregation expression when packet error is present (8):

$$\mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}[||w^{(t+1)} - \bar{w}^{(t+1)}||^2] = \mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}\Big[\Big|\Big|\frac{1}{K}\sum_{k\in S_t}\Big(\frac{Z_k}{1-q_k}(w_k^{(t+1)} - \bar{w}^{(t+1)}) + (1 - \frac{Z_k}{1-q_k})(w^{(t)} - \bar{w}^{(t+1)})\Big)\Big|\Big|^2\Big],$$

where Z_k is the random variable of successful packet transmission.

We first give the relation between the expected value over randomly selected set S_t and the expected value over all client k:

$$\mathbb{E}_{S_t}\Big[\sum_{k\in S_t} X_k\Big] = K\,\mathbb{E}_k[X_k]. \tag{28}$$

Each client k is sampled with the probability p_k. The above equality can be proven by denoting $S_t = \{i_1, \ldots, i_K\}$, then for any $\{x_k\}_k$ as in [17],

$$\mathbb{E}_{S_t}\sum_{k\in S_t} x_k = \mathbb{E}_{S_t}\sum_{k=1}^{K} x_{i_k} = K\,\mathbb{E}_{S_t}\, x_{i_1} = K\,\mathbb{E}_k[x_k].$$

The square norm of the sum of two terms is separated as any cross scalar product is zero as

$$\mathbb{E}_{S_t}\left[\sum_{k\in S_t} w_k^{(t+1)} - \bar{w}^{(t+1)}\right] = K(\mathbb{E}_k[w_k^{(t+1)}] - \bar{w}^{(t+1)}) = 0, \tag{29}$$

by definition of $\bar{w}^{(t+1)}$. We denote $\mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}[||w^{(t+1)} - \bar{w}^{(t+1)}||^2] = A + D$ with

$$A = \mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}\Big[\Big|\Big|\frac{1}{K}\sum_{k\in S_t}\frac{Z_k}{1-q_k}(w_k^{(t+1)} - \bar{w}^{(t+1)})\Big|\Big|^2\Big], \tag{30}$$

$$D = \mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}\Big[\Big|\Big|\frac{1}{K}\sum_{k\in S_t}\left(1 - \frac{Z_k}{1-q_k}\right)(w^{(t)} - \bar{w}^{(t+1)})\Big|\Big|^2\Big]. \tag{31}$$

We have $\mathbb{E}_{h_k,Z_k}(Z_k) = 1 - q_k$. By developing the square term in A and exploiting the independence of cross terms, we observe that all cross terms are zero from (29), term A is derived as:

$$A = \frac{1}{K^2}\,\mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}\Big[\sum_{k\in S_t}\frac{Z_k^2}{(1-q_k)^2}||w_k^{(t+1)} - \bar{w}^{(t+1)}||^2\Big]. \tag{32}$$

We have $Z_k = Z_k^2 = \{0,1\}$ and by (28):

$$A = \frac{1}{K}\,\mathbb{E}_{k,\boldsymbol{h},\boldsymbol{Z}}\Big[\frac{Z_k}{(1-q_k)^2}||w_k^{(t+1)} - \bar{w}^{(t+1)}||^2\Big]. \tag{33}$$

Only Z_k is dependent on h and Z,

$$\begin{aligned} A &= \frac{1}{K}\,\mathbb{E}_k\Big[\frac{\mathbb{E}_{\boldsymbol{h},\boldsymbol{Z}}[Z_k]}{(1-q_k)^2}||w_k^{(t+1)} - \bar{w}^{(t+1)}||^2\Big] \\ &= \frac{1}{K}\,\mathbb{E}_k[\frac{1}{1-q_k}||w_k^{(t+1)} - \bar{w}^{(t+1)}||^2] \\ &\leq \frac{1}{K(1-q_{\max})}\,\mathbb{E}_k[||w_k^{(t+1)} - \bar{w}^{(t+1)}||^2] \\ &\leq \frac{2B^2}{K(1-q_{\max})}\frac{(1+\gamma)^2}{\bar{\mu}^2}||\nabla f(w^{(t)})||^2. \end{aligned} \tag{34}$$

The last inequality comes from proven results in the proof of [15]. In the term D, the difference of the weight term is independent of any random variable, then we have,

$$D = \frac{1}{K^2}||w^{(t)} - \bar{w}^{(t+1)}||^2\,\mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}\Big[\Big|\sum_{k\in S_t}(1 - \frac{Z_k}{1-q_k})\Big|^2\Big]. \tag{35}$$

By Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} D &\leq \frac{1}{K^2}||w^{(t)} - \bar{w}^{(t+1)}||^2 K\,\mathbb{E}_{S_t,\boldsymbol{h},\boldsymbol{Z}}\Big[\sum_{k\in S_t}\Big|1 - \frac{Z_k}{1-q_k}\Big|^2\Big] \\ &\leq ||w^{(t)} - \bar{w}^{(t+1)}||^2\,\mathbb{E}_{k,\boldsymbol{h},\boldsymbol{Z}}\Big[(1 - \frac{Z_k}{1-q_k})^2\Big], \quad \text{as in (28)} \\ &\leq ||w^{(t)} - \bar{w}^{(t+1)}||^2\Big(1 - 2\,\mathbb{E}_k[\frac{\mathbb{E}_{\boldsymbol{h},\boldsymbol{Z}}[Z_k]}{1-q_k}] + \mathbb{E}_k[\frac{\mathbb{E}_{\boldsymbol{h},\boldsymbol{Z}}[Z_k^2]}{(1-q_k)^2}]\Big) \\ &\leq ||w^{(t)} - \bar{w}^{(t+1)}||^2\Big(\mathbb{E}_k\Big[\frac{1}{1-q_k}\Big] - 1\Big) \\ &\leq ||w^{(t)} - \bar{w}^{(t+1)}||^2\frac{q_{\max}}{1-q_{\max}} \\ &\leq \frac{B^2(1+\gamma)^2}{\bar{\mu}^2}||\nabla f(w^{(t)})||^2\frac{q_{\max}}{1-q_{\max}}. \end{aligned} \tag{36}$$

Finally,

$$\mathbb{E}_{S_t,h,Z}[||w^{(t+1)} - \bar{w}^{(t+1)}||^2] \leq \frac{B^2(1+\gamma)^2}{\bar{\mu}^2}||\nabla f(w^{(t)})||^2\frac{2}{K}\underbrace{\Big(1 + \frac{q_{\max}}{1-q_{\max}}(1 + \frac{K}{2})\Big)}_{C(\boldsymbol{q})} \tag{37}$$

holds, and the lemma is proved.

References

[1] P. Zheng, Y. Zhu, Z. Zhang, Y. Hu, and A. Schmeink, "Federated learning in heterogeneous networks with unreliable communication," in *IEEE Globecom Workshops (GC Wkshps)*, Madrid, Spain, 2021, pp. 1–6.

[2] D. Peteiro-Barral and B. Guijarro-Berdiñas, "A survey of methods for distributed machine learning," *Progress in Artificial Intelligence*, vol. 2, no. 1, pp. 1–11, 2013.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, Fort Lauderdale, Florida, USA, 2017.

[4] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NeurIPS Workshop*, Barcelona SPAIN, 2016.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[6] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[7] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. Vincent Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *NeurIPS*, 2020.

[8] A. Xu and H. Huang, "Coordinating momenta for cross-silo federated learning," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8735–8743.

[9] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh, "Breaking the centralized barrier for cross-device federated learning," in *NeurIPS*, vol. 34, 2021, pp. 28 663–28 676.

[10] E. Ozfatura, K. Ozfatura, and D. Gündüz, "FedADC: Accelerated federated learning with drift control," in *2021 IEEE Int. Symp. on Information Theory (ISIT)*, 2021, pp. 467–472.

[11] G. Lee, M. Jeong, Y. Shin, S. Bae, and S.-Y. Yun, "Preservation of the global knowledge by not-true distillation in federated learning," in *NeurIPS*, 2022.

[12] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *ICML*, 2021.

[13] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proc. of the IEEE/CVF Conf. on CVPR*, 2021.

[14] D. A. E. Acar, Y. Zhao, R. Matas, M. Mattina, P. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *ICLR*, 2021.

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

[16] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[17] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-iid data," in *ICLR*, 2020.

[18] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, 2021.

[19] 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.*, vol. 20, no. 3, pp. 1935–1949, 2021.

[20] V.-D. Nguyen, S. K. Sharma, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Efficient federated learning algorithm for resource allocation in wireless IoT networks," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3394–3409, Mar. 2021.

[21] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2020.

[22] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, 2021.

[23] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 342–358, 2022.

[24] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, 2021.

[25] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.

[26] M. Zhang, G. Zhu, S. Wang, J. Jiang, Q. Liao, C. Zhong, and S. Cui, "Communication-efficient federated edge learning via optimal probabilistic device scheduling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8536–8551, 2022.

[27] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 453–467, 2021.

[28] W. Wen, Z. Chen, H. H. Yang, W. Xia, and T. Q. S. Quek, "Joint scheduling and resource allocation for hierarchical federated edge learning," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2022.

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

[30] Y. Sun, S. Zhou, and D. Gündüz, "Energy-aware analog aggregation for federated learning with redundant data," in *IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–7.

[31] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–7.

[32] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, 2020.

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

[34] Z. Qu, K. Lin, J. Kalagnanam, Z. Li, J. Zhou, and Z. Zhou, "Federated learning's blessing: FedAvg has linear speedup," in *ICLR Workshop*, 2021.

[35] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," in *ICLR*, 2021.

[36] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2021.

[37] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, pp. 317–333, 2020.

[38] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5136–5151, 2021.

[39] Y. Xi, A. Burr, J. Wei, and D. Grace, "A general upper bound to evaluate packet error rate over quasi-static fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 5, pp. 1373–1377, 2011.

[40] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, 2016.

[41] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree, "Regularisation of neural networks by enforcing Lipschitz continuity," *Machine Learning*, vol. 110, no. 2, pp. 393–416, 2021.

[42] J. Leng, Z. Lin, M. Ding, P. Wang, D. Smith, and B. Vucetic, "Client scheduling in wireless federated learning based on channel and learning qualities," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 732–735, 2022.

[43] L. Hübschle-Schneider and P. Sanders, "Parallel weighted random sampling," in *27th Annu. European Sympos. on Algorithms (ESA)*, Aarhus, Denmark, 2019.

[44] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," in *AISTATS*, Mar. 2022.

[45] R. Balakrishnan, M. Akdeniz, S. Dhakal, and N. Himayat, "Resource management and fairness for federated learning over wireless edge networks," in *IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, 2020, pp. 1–5.

[46] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[47] S. Arora and B. Barak, *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[48] M. L. Fisher, "The Lagrangian relaxation method for solving integer programming problems," *Management science*, vol. 27, no. 1, pp. 1–18, 1981.

[49] R. Hamming, *Numerical methods for scientists and engineers*. Courier Corporation, 2012.

[50] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[51] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[52] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," in *NeurIPS Workshop*, 2019.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR San Diego, CA, USA, 2015*, 2015.