Wireless Quantized Federated Learning: A Joint Computation and Communication Design

Pavlos S. Bouzinis, *Graduate Student Member, IEEE*, Panagiotis D. Diamantoulakis, *Senior Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—Recently, federated learning (FL) has sparked widespread attention as a promising decentralized machine learning approach which provides privacy and low delay. However, communication bottleneck still constitutes an issue, that needs to be resolved for an efficient deployment of FL over wireless networks. In this paper, we aim to minimize the total convergence time of FL, by quantizing the local model parameters prior to uplink transmission. More specifically, the convergence analysis of the FL algorithm with stochastic quantization is firstly presented, which reveals the impact of the quantization error on the convergence rate. Following that, we jointly optimize the computing and communication resources as well as the number of quantization bits, in order to guarantee minimized convergence time, subject to energy and quantization error requirements. The impact of the quantization error on the convergence time is evaluated and the trade-off among model accuracy and timely execution is revealed. Moreover, the proposed method is shown to result in faster convergence compared with baseline schemes. Finally, useful insights for the selection of the quantization error tolerance are provided.**

***Index Terms*—wireless federated learning, quantization, convergence time minimization**

I. Introduction

Future wireless networks are envisioned to support ubiquitous artificial intelligent services [1]. Conventional machine learning techniques are usually conducted in a centralized manner, where the data are uploaded and processed at a single entity, e.g., a central server [2]. However, the growing computing capabilities of devices have paved the way for realizing distributed learning frameworks. Among the decentralized approaches, federated learning (FL) has shown great promise in preserving data privacy and providing low delay [3], [4]. In FL, users collaboratively build a shared learning model, without exposing their raw data to the server or any other residual participant. The server redistributes the global model back to the users, while the whole procedure is repeated until the convergence of the global model. In this manner, FL is inherently privacy-preserving and reduces the communication load. However, the wireless environment imposes some distinctive challenges, owing to the limited wireless resources, unreliable links, etc., which degrade the performance of FL and need to be efficiently addressed [5]–[8].

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 957406. P. S. Bouzinis, P. Diamantoulakis, and G. K. Karagiannidis are with Wireless Communication and Information Processing Group (WCIP), Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece. G. K. Karagiannidis is also with Cyber Security Systems and Applied AI Research Center, Lebanese American University (LAU), Lebanon. e-mails: {mpouzinis, padiaman, geokarag}@auth.gr

A. Related Works and Motivation

In the context of wireless networks, several works examined and optimized the performance of FL in various aspects, such as model accuracy, timely execution and energy efficiency. More specifically, in [9], a joint learning and communication framework was proposed, in order to minimize the training loss in the presence of packet errors, while in [10] the objective was to minimize the total energy consumption by optimizing both the computation and communication resources. Moreover in [11], authors focus on minimizing the convergence time of the FL process by efficiently scheduling the participating devices, while in [12], the convergence time was minimized by considering both the impact of training and communication. In [13], the FL global loss function was minimized under total convergence time constraints, by jointly allocating the bandwidth and scheduling the users, while in [14], the impact of various scheduling policies on the FL convergence rate was examined. Finally, [15] proposed a coded federated learning scheme, which exploits coding techniques to introduce redundant computations to the FL server.

Although previous works focused on achieving a timely FL execution, communication bottleneck may be still incurred by the limited bandwidth and the large size of the local training parameters. To alleviate this burden, gradient compression techniques, such as quantization, have been proposed to further improve the communication efficiency in FL [16]–[18]. Through this technique, a quantized version of the local model is being transmitted to the server, aiming to achieve faster communication without deteriorating the FL model accuracy. In this direction, authors in [19], examined quantization schemes for the uplink and downlink communication in FL and rigorously proved the respective convergence rate upper bounds. In [20], a lossy FL scheme is introduced where both global and local updates are quantized before being transmitted, while the convergence behavior is analyzed. Furthermore, in [21], a universal vector quantization scheme for FL was proposed accompanied by an analysis for the respective error distortion, while authors showed that the error vanishes as the number of users grows. Moreover, in [22], the FL convergence bound was minimized, in a multiple access channel scenario of FL, while an efficient utilization of the quantization levels was proposed. In [23], the quantization error was minimized subject to uplink transmission delay and outage constraints per

round, in a wireless environment. In [24], authors proposed a one-bit quantization scheme for over-the-air FL, while they also examined the convergence rate under fading channels. In [25], a heterogeneous quantization scheme was proposed towards minimizing the convergence upper bound as a function of the heterogeneous quantization errors of all clients. In [26], a decentralized quantization algorithm based on the alternating direction method of multipliers was proposed, which resulted in decreased communication cost in comparison with its non-quantized counterpart. Finally, in [27] and [28], the use of weight quantization methods in FL over wireless networks is examined, towards minimizing either the energy consumption or the training time.

At this point, we clarify that none of the aforementioned works [19]–[27], which applied quantization methods, focused on minimizing the total convergence time of the FL, which is a metric of paramount importance for meeting the low latency requirements of the next generation of wireless networks. Moreover, the convergence time is one of the most critical metric for evaluating the communication efficiency of quantization methods, since their primary goal is to reduce communication bottlenecks. Although, authors in [19]–[21] and [24], [26] examined the behavior of quantization schemes in FL, they did not propose techniques for optimizing the performance of quantization in a wireless environment and addressing its underlying constraints in the quantization process, e.g., allocating communication and computation resources towards achieving timely FL execution. Regarding the works that performed optimization methods, [22], [23] and [25] focused on minimizing the quantization error -and not the convergence time- without jointly considering the computation, the communication resources, as well as the quantization policy. Specifically, the analysis in [22] assumed a simplistic model for the transmission rate constraints, without including the wireless factors and the computation delay. Similarly, [25] minimizes the convergence upper bound due to quantization without considering any additional constraints, neither wireless nor computation ones. Moreover, although [23] focuses on a wireless FL scenario, the local computation time and subsequently the adjustment of users' local CPU frequency were neglected, which affect both the convergence time and the energy requirements. Also, in [27], the energy consumption in wireless FL is minimized, while authors solely consider the optimal selection of the quantization precision and do not focus on optimizing computational or radio resources. To this end, in [28], the authors aim to minimize the training/convergence time in FL over wireless networks, without neither considering a variable computational power nor optimizing the transmit power. Hence, the trade-off between computational and communication power is neglected, and thus, its effects on the convergence time is absent. Moreover, the number of quantization bits is selected to be fixed throughout the whole training phase.

By taking the above into consideration, prior works seem to neglect the joint design of computational and communication resources when applying weight quantization in wireless FL. Therefore, the efficient utilization of the available computation, radio resources and quantization strategy in FL over wireless networks, is an issue that has not been yet resolved. This joint optimization of the available resources in a holistic approach, could further accelerate the training time of FL. Furthermore, the trade-off among model accuracy and fast convergence, owing to the local updates' quantization, is not well-examined in the previous works, where the convergence rate is mainly investigated with regards to the global communication rounds and not the total evolution of time. Specifically, a large number of quantization bits may increase precision, in the expense of slow communication with the server per FL round, since the training parameters' size also increases and leads to higher transmission delay. On the contrary, smaller quantization level may lead to slightly decreased model performance, though, with lower delay per round and potentially faster convergence. However, low precision updates may be communication efficient in terms of latency per round, but require increased number of communication rounds until convergence. Therefore, it is not evident that the utilization of few quantization bits can always lead to faster convergence, with respect to the evolution of time. The aforementioned fact is not clearly presented in the literature and needs further investigation.

B. Contributions

Driven by the aforementioned considerations, we study the quantization of the local model parameters of each user and aim to minimize the total convergence time of FL, under energy consumption and quantization error constraints. The latter aims to retain the model accuracy in desirable levels. Moreover, the considered minimization is conducted with respect to the unit of time, and not purely the evolution of global FL rounds, since the wireless factors, the available resources and the quantization policy affect the duration of each round, which is critical for the evaluation of the convergence time. The main contributions of this paper are summarized as:

- We present a rigorous convergence analysis of the FL process by considering stochastic quantization of the local parameters. The impact of the quantization error on the convergence bound is revealed, while useful insights for the optimality gap are provided. Furthermore, throughout the derivation of the resulted bound, we do not enforce the quantization bits to be a function of convergence related parameters, which is usually assumed for guaranteeing convergence [19], [26]. Note that by setting the quantization bits in advance, may not lead to minimized convergence time, since the number of quantization bits also influence the uplink transmission duration of the training parameters, owing to the limitations of the wireless medium. Therefore, we allow the proper adjustment of the quantization bits when dealing with wireless systems' constraints, and thus, targeting to the timely completion of the FL task.
- We minimize the convergence time of FL, i.e, the total duration of all global rounds, subject to energy consumption and quantization error constraints. The latter have resulted from the convergence analysis and aim to guarantee sufficiently high precision and subsequently

high model accuracy. To this end, we jointly optimize the computation and communication resources, as well as the number of quantization bits of each user. After some mathematical manipulations, the resulted convex problem is solved with the *Lagrange dual decomposition* and closed-form solutions are derived, in terms of the Lagrange multipliers (LMs).

- Through simulations, the performance of local parameters' quantization is evaluated. Specifically, we investigate the impact of the quantization error tolerance on the convergence time and model accuracy, while the trade-off among model accuracy and fast convergence is exhibited. In addition to this, it is shown that low precision quantization cannot always achieve fast convergence. This result is related with the increased number of global rounds that low precision quantization requires, which may prevail over the low transmission delay per round. Moreover, numerical results validate the effectiveness of the proposed optimization towards minimizing the convergence time, in comparison with baseline schemes, revealing the significance of jointly adjusting the radio and computation resources, as well as the quantization bits. Finally, driven by the theoretical convergence analysis, we study the effects of decaying the quantization error tolerance along with the evolution of the training, instead of keeping it constant. The simulation results corroborate the effectiveness of this approach, which demonstrates increased convergence rate without model accuracy degradation. In essence, this observation coincides with the concept of "later-is-better" [29], which implies that reserving FL-related resources in the early stages of the training process and spending them in the final stages, may be beneficial for the performance.

II. System model

A. FL model

We consider a wireless FL system, consisting of a set $\mathcal{K} = \{1, 2, ..., K\}$ of K users and a base station (BS) co-located with a server, while hereinafter we use the terms BS and server interchangeably. To tackle with the straggler effect, we assume that a only a subset $\mathcal{N} \subseteq \mathcal{K}$ of users is participating in the FL process, with cardinality $|\mathcal{N}| = N \leq K$. The user selection policy will be described later on this work. Each user $n \in \mathcal{N}$, poses a local dataset $\mathcal{D}_n^{\mathrm{L}} = \{\boldsymbol{x}_{n,k}, y_{n,i}\}_{k=1}^{D_n^{\mathrm{L}}}$, where $D_n^{\mathrm{L}} = |\mathcal{D}_n^{\mathrm{L}}|$, $\boldsymbol{x}_{n,k}$ is the k-th input data vector of user n, while $y_{n,k}$ is the corresponding output. The whole dataset is denoted as $\mathcal{D} = \underset{n\in\mathcal{N}}{\cup} \mathcal{D}_n^{\mathrm{L}}$, while the size of all training data among the participating users is $D = \sum_{n=1}^{N} D_n^{\mathrm{L}}$. The loss function of user n, is defined as [8]

$$F_n(\boldsymbol{w}) \triangleq \frac{1}{D_n^{\mathrm{L}}} \sum_{k\in\mathcal{D}_n^{\mathrm{L}}} \phi(\boldsymbol{w}, \boldsymbol{x}_{n,k}, y_{n,k}), \quad \forall n \in \mathcal{N}, \tag{1}$$

where $\phi(\boldsymbol{w}, \boldsymbol{x}_{n,k}, y_{n,k})$ captures the error of the d-dimensional model parameter $\boldsymbol{w} \in \mathbb{R}^d$ for the input-output pair $\{\boldsymbol{x}_{n,k}, y_{n,k}\}$. The goal of the training process is to find the global parameter $\boldsymbol{w}$, which minimizes the loss function on the whole dataset, i.e., $F(\boldsymbol{w}) = \sum_{n=1}^{N} p_n F_n(\boldsymbol{w})$, where $p_n = \frac{D_n^{\mathrm{L}}}{D}$. Hereinafter, for ease of presentation we consider that $p_n = \frac{1}{N}$, $\forall n \in \mathcal{N}$.

We assume that the whole FL process consists of T global rounds, denoted as $t \in \mathcal{T} = \{0, ..., T-1\}$. The subset of users scheduled for participation in the t-th round is denoted as $\mathcal{N}(t)$, with $|\mathcal{N}(t)| = N$, $\forall t$. During the t -th global round, each user receives the global parameter $\boldsymbol{w}(t)$ from the server, and performs τ steps of the stochastic gradient descent (SGD) method. The i-th step of SGD, $\forall n \in \mathcal{N}(t)$, is given as

$$\boldsymbol{w}_n^i(t) = \boldsymbol{w}_n^{i-1}(t) - \eta(t) \nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t)), \ i = 1, ..., \tau, \tag{2}$$

where $\boldsymbol{w}_n^0(t) \triangleq \boldsymbol{w}(t)$. Moreover, $\eta(t)$ represents the learning rate, while $\boldsymbol{\xi}_n^{i-1}(t) \subseteq \mathcal{D}_n^{\mathrm{L}}$ is a mini-batch, which is sampled uniformly at random from the local dataset $\mathcal{D}_n^{\mathrm{L}}$ of user n. Therefore, it holds $\mathbb{E}[\nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))] = \nabla F_n(\boldsymbol{w}_n^{i-1}(t))$, where the expectation is taken with respect to the randomness of the stochastic gradient function. Furthermore, we assume that at the first global round, the server initializes $\boldsymbol{w}(0)$. After terminating the local training, user n transmits the weight differential to the server, given as

$$\begin{aligned} \Delta \boldsymbol{w}_n(t) &= \boldsymbol{w}_n^{\tau}(t) - \boldsymbol{w}_n^0(t) = \boldsymbol{w}_n^{\tau}(t) - \boldsymbol{w}(t) \\ &= -\eta(t) \sum_{i=1}^{\tau} \nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t)). \end{aligned} \tag{3}$$

The selection of transmitting the weight differential instead of simply transmitting the latest local weight $\boldsymbol{w}_n^{\tau}(t)$, is related with the quantization scheme that will be used and discussed later on this work. Following that, the global model at the server's side, in round t, is updated as follows

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \sum_{n\in\mathcal{N}(t)} \frac{1}{N} \Delta \boldsymbol{w}_n(t). \tag{4}$$

At last, the global model is broadcast to the devices, while the whole process is repeated for T rounds, until the convergence of the global model.

B. Quantization model

As mentioned previously, at time step t, each user $n \in \mathcal{N}(t)$ calculates its local model $\Delta \boldsymbol{w}_n(t) = (\Delta w_{n,1}(t), ..., \Delta w_{n,d}(t))^{\top}$. In order to prevent a wasteful overuse of resources, we assume that users send to the server a quantized version of $\Delta \boldsymbol{w}_n(t)$, which is denoted as $Q(\Delta \boldsymbol{w}_n(t))$, where $Q(\cdot)$ denotes the quantization function. Therefore, the global model update at the server, is actually given as

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \sum_{n\in\mathcal{N}(t)} \frac{1}{N} Q(\Delta \boldsymbol{w}_n(t)). \tag{5}$$

We also highlight that in [19] it was shown that by transmitting the weight differential $\Delta \boldsymbol{w}_n(t)$, a faster convergence is achieved, while it is also adopted in our work. Following that, for each element $j \in \{1, 2, ..., d\}$ of $\Delta \boldsymbol{w}_n(t)$, it holds $|\Delta w_{n,j}(t)| \in [\Delta w_n^{\min}(t), \Delta w_n^{\max}(t)]$, where $\Delta w_n^{\min}(t) \triangleq$

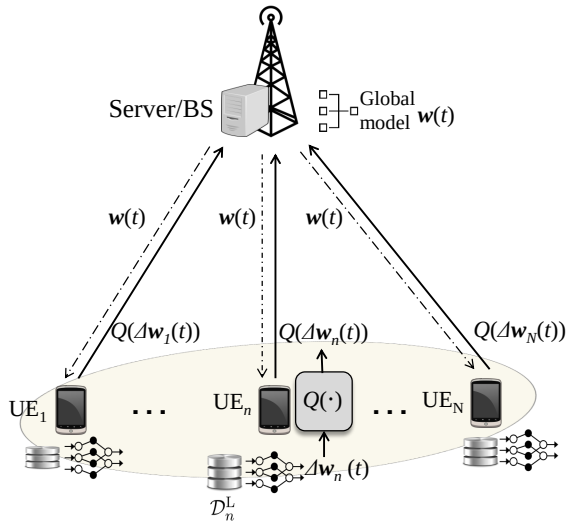


Fig. 1. Federated learning with local model quantization.

$\min\{|\Delta \boldsymbol{w}_n(t)|\}$ and $\Delta w_n^{\max}(t) \triangleq \max\{|\Delta \boldsymbol{w}_n(t)|\}$. Moreover, we assume that $\Delta w_{n,j}(t)$ is quantized according to the stochastic quantization method [19]. That is, with $B_n(t)$ quantization bits, user n can divide the interval $[\Delta w_n^{\min}(t), \Delta w_n^{\max}(t)]$ into the following ς intervals: $I_1 = [s_0, s_1]$, $I_2 = [s_1, s_2], ..., I_\varsigma = [s_{\varsigma-1}, s_\varsigma]$, where $\varsigma = 2^{B_n(t)} - 1$ and

$$s_k = \Delta w_n^{\min}(t) + k \frac{\Delta w_n^{\max}(t) - \Delta w_n^{\min}(t)}{2^{B_n(t)} - 1}, \tag{6}$$

where $k = 0, 1, ..., 2^{B_n(t)} - 1$. Therefore, if the parameter $\Delta w_{n,j}(t)$ falls into the interval I_k, it will be quantized as

$$Q(\Delta w_{n,j}(t)) = \begin{cases} s_{k-1} \cdot \mathrm{sign}(\Delta w_{n,j}(t)), & \text{w.p. } \frac{s_k - |\Delta w_{n,j}(t)|}{s_k - s_{k-1}} \\ s_k \cdot \mathrm{sign}(\Delta w_{n,j}(t)), & \text{w.p. } \frac{|\Delta w_{n,j}(t)| - s_{k-1}}{s_k - s_{k-1}} \end{cases} \tag{7}$$

where w.p. stands for "with probability". Furthermore, the overall d-dimensional quantized local model is denoted as

$$Q(\Delta \boldsymbol{w}_n(t)) = (Q(\Delta w_{n,1}(t)), ..., Q(\Delta w_{n,d}(t)))^\top, \quad \forall n, t. \tag{8}$$

while its size is given by $S_n(t) = d(B_n(t)+1)+m$ (bits), since each element of the quantized model vector is represented by $B_n(t)$ bits plus one bit for the sign specification. Also, m bits are needed to specify the values of $\Delta w_n^{\max}(t)$ and $\Delta w_n^{\min}(t)$.

C. Computation and Communication model

During the t-th global round, the server is broadcasting the global model $\boldsymbol{w}(t)$ to all users. We consider that the downlink transmission latency is negligible, since the transmit power of the BS is much larger than that of the devices, while also the same message is broadcast to all users. In addition to this, we assume that the downlink is error-free. In the continue, we slightly abuse the notation by dropping t, while the following expressions could refer to any arbitrary round. The time for local computations by device n, for τ SGD steps, in order to generate the local model, is

$$l_n^{\mathrm{c}} = \tau \frac{c_n D_n}{f_n}, \quad \forall n \in \mathcal{N}, \tag{9}$$

where f_n is the CPU cycle frequency of user n, D_n is the size of the mini-batch (in bits), while c_n denotes the number of CPU cycles for user n to perform one sample of data during the local model training. The energy consumption for the local computations, is given as [30]

$$E_n^{\mathrm{c}} = \tau \zeta c_n D_n f_n^2, \quad \forall n \in \mathcal{N}, \tag{10}$$

where ζ is a constant parameter related with the hardware architecture of device n. Finally, we assume that the time duration dedicated for generating $Q(\Delta \boldsymbol{w}_n)$ through the quantization process, is negligible.

Following the local training, each device uploads the quantized training parameters, $Q(\Delta \boldsymbol{w}_n)$, to the BS. Similarly to [31], we assume that the considered transmission is carried out via time-division multiple access (TDMA), while this choice is not restrictive, since other schemes such as frequency-division multiple access (FDMA) can also be applied. To successfully upload the training parameters within l_n^{up} uplink time duration, the n-th user should satisfy the condition

$$l_n^{\mathrm{up}} W \log_2 \left(1 + \frac{g_n E_n}{l_n^{\mathrm{up}} W N_0}\right) \geq S_n, \quad \forall n \in \mathcal{N}, \tag{11}$$

where W is the available bandwidth, E_n is the transmit energy, S_n is the size of the quantized training parameters and N_0 is the power of the additive white Gaussian noise (AWGN). Moreover, $g_n = |h_n|^2 d_n^{-\beta}$ denotes the channel gain, where the complex random variable $h_n \sim \mathcal{CN}(0, 1)$ is the small scale fading, d_n is the distance between user n and the BS and β is the path loss exponent. Moreover, we assume that the channel gain is quasi-static and stays unchanged during a single global round, while we also consider perfect CSI both in BS's and users' side. Following that, we assume that the parameter transmission phase begins after the termination of the local computations phase by each user. Therefore, for the local computation duration l^{c}, it holds that

$$l^{\mathrm{c}} \geq l_n^{\mathrm{c}} = \tau \frac{c_n D_n}{f_n}, \quad \forall n \in \mathcal{N}. \tag{12}$$

Since all users should terminate the computation and uplink transmission phases, so as the server can receive each local model and subsequently update the global model, the total duration of a FL global round is given as

$$l^{\mathrm{r}} = l^{\mathrm{c}} + \sum_{n \in \mathcal{N}} l_n^{\mathrm{up}}, \tag{13}$$

which is the sum of the computation latency and the transmission latency among all users and is depicted in Fig. 2. At this point, it is clarified that the participating users are selected with respect to their channel gains [32], in order to exclude users who suffer from bad channel conditions and subsequently avoid large transmission delays. By adopting this approach, the server selects only the N strongest among the total K users for participation, i.e., $\mathcal{N} = \{n \in \mathcal{K} \,|\, g_n \geq g_{[N]}\}$,

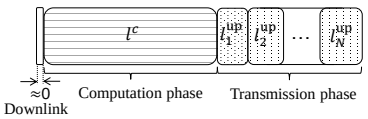


Fig. 2. Duration of a global FL round.

where $g_{[N]}$ denotes the N-th largest channel gain among all users.

III. Convergence Analysis

In this section we investigate the convergence behavior of FL with stochastic quantization and stochastic gradient function. Firstly, with regards to the stochastic quantization scheme, we proceed to the formulation of the following lemma:

Lemma 1: $Q(\Delta \boldsymbol{w}_n(t))$ is an unbiased estimator of $\Delta \boldsymbol{w}_n(t)$, i.e.,

$$\mathbb{E}\left[Q(\Delta \boldsymbol{w}_n(t))\right] = \Delta \boldsymbol{w}_n(t), \tag{14}$$

while it also holds that

$$\mathbb{E}\left[\left\|Q(\Delta \boldsymbol{w}_n(t)) - \Delta \boldsymbol{w}_n(t)\right\|_2^2\right] \leq \frac{\delta_n^2(t)}{\left(2^{B_n(t)}-1\right)^2} \triangleq J_n^2(t), \tag{15}$$

where $\delta_n(t) \triangleq \sqrt{\frac{d}{4}\left(\Delta w_n^{\max}(t) - \Delta w_n^{\min}(t)\right)^2}$.

Proof: The proof can be found in [19], [23]. ∎

Next, we make the following common assumptions for the functions $F_1, F_2, ..., F_N$, in order to facilitate the convergence analysis [33].

Assumption 1: $F_n, \forall n \in \mathcal{N}$, are all L-smooth, i.e., $\forall \boldsymbol{w}', \boldsymbol{w} \in \mathbb{R}^d$: $F_n(\boldsymbol{w}') \leq F_n(\boldsymbol{w}) + \langle \boldsymbol{w}' - \boldsymbol{w}, \nabla F_n(\boldsymbol{w})\rangle + \frac{L}{2}\left\|\boldsymbol{w}' - \boldsymbol{w}\right\|_2^2$.

Assumption 2: F_n, $\forall n \in \mathcal{N}$, are all μ-strongly convex, i.e., $\forall \boldsymbol{w}', \boldsymbol{w} \in \mathbb{R}^d$: $F_n(\boldsymbol{w}') \geq F_n(\boldsymbol{w}) + \langle \boldsymbol{w}' - \boldsymbol{w}, \nabla F_n(\boldsymbol{w})\rangle + \frac{\mu}{2}\left\|\boldsymbol{w}' - \boldsymbol{w}\right\|_2^2$.

Assumption 3: The expected squared norm of stochastic gradients is uniformly bounded $\forall n, t, i$, i.e., $\mathbb{E}\left[\left\|\nabla F_n(\boldsymbol{w}_n^i(t), \boldsymbol{\xi}_n^i(t))\right\|_2^2\right] \leq G^2$.

Assumption 4: The variance of stochastic gradients in each user is bounded $\forall n, t, i$, i.e., $\mathbb{E}\left[\left\|\nabla F_n(\boldsymbol{w}_n^i(t), \boldsymbol{\xi}_n^i(t)) - \nabla F_n(\boldsymbol{w}_n^i(t))\right\|_2^2\right] \leq \sigma_n^2$.

Moreover, we define $\Gamma \triangleq F(\boldsymbol{w}^*) - \sum_{n=1}^N p_n F_n^*$, where F_n^* denotes the minimum value of $F_n(\cdot)$, while Γ quantifies the degree of NON-IID, among users' datasets. Taking these into account, we introduce the following theorem:

Theorem 1: Let Assumptions 1 to 4 hold. By selecting a diminishing learning rate $\eta(t) = \frac{2}{\mu(\gamma+t)}$ and $\gamma > \max\{2, \frac{2}{\mu}, \frac{L}{\mu}\}$, the upper bound of $\mathbb{E}\left[F(\boldsymbol{w}(T)) - F(\boldsymbol{w}^*)\right]$ is given by

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{w}(T))\right] - F(\boldsymbol{w}^*) \leq \\ \frac{L}{2}\frac{1}{\gamma+T}\left(\frac{4U}{\mu^2} + \gamma\mathbb{E}\left[\left\|\boldsymbol{w}(0) - \boldsymbol{w}^*\right\|_2^2\right]\right) \\ + \underbrace{\frac{L}{2}\sum_{j=0}^{T-1}\left[\sum_{n\in\mathcal{N}(j)}\frac{1}{N}\frac{\delta_n^2(j)}{\left(2^{B_n(j)}-1\right)^2}\prod_{i=j+1}^{T-1}\left(1 - \frac{2}{\gamma+i}\right)\right]}_{\text{Impact of the quantization error on the convergence}}\end{aligned} \tag{16}$$

where

$$\begin{aligned}U = \tau^2\sum_{n=1}^K\frac{\sigma_n^2}{K} + 2L\tau^2\Gamma + (\mu+2)\frac{\tau(\tau-1)(2\tau-1)}{6}G^2 \\ + \tau G^2 + \frac{K-N}{N(K-1)}\tau^2G^2,\end{aligned} \tag{17}$$

while the expectation is taken with respect to the stochastic gradient function, the stochastic quantization scheme and the randomness in user selection.

Proof: See Appendix A. ∎

It is clarified that the assumption of equal dataset sizes among users can be dropped by applying some modifications in the proof, as in [33]. Specifically, by scaling the local functions as $\tilde{F}_n(\boldsymbol{w}) = q_n K F_n(\boldsymbol{w})$, where $q_n = \frac{D_n^{\mathrm{L}}}{\sum_{n\in\mathcal{K}} D_n^{\mathrm{L}}}$, $\forall n \in \mathcal{K}$, the global function can be equivalently written as $F(\boldsymbol{w}) = \frac{1}{K}\sum_{n=1}^K \tilde{F}_n(\boldsymbol{w})$. Additionally, $\boldsymbol{w}(t+1)$ should be rewritten as $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \sum_{n\in\mathcal{N}(t)} p_n Q(\Delta \boldsymbol{w}_n(t))$. With these transformations and some algebraic manipulations, i.e., σ_n and J_n are replaced by $\tilde{\sigma}_n \triangleq \sqrt{Kq_n}\sigma_n$ and $\tilde{J}_n \triangleq \sqrt{Np_n}J_n$, Theorem 1 will still hold [33]. Hence, the presented results can be generalized for the case of unbalanced dataset sizes among users.

As one can observe, for large T, the first term of the upper bound tends to zero with rate $\mathcal{O}(\frac{1}{T})$. However, the second term which is related with the quantization error, creates a gap between $\mathbb{E}\left[F(\boldsymbol{w}(T))\right]$ and $F(\boldsymbol{w}^*)$. Inspired by [20], we present the following interesting comment. For small j, the term $\prod_{i=j+1}^{T-1}\left(1 - \frac{2}{\gamma+i}\right)$ tends to zero, since $1 - \frac{2}{\gamma+i} < 1$, $\forall i$. Therefore, the effect of the quantization error in the early stages of the training process vanishes over time. Hence, it is discernible that during the early training rounds, the quantization error would not contribute to increasing the optimality gap. Nevertheless, in order to further mitigate the impact of the quantization error, an increased number of quantization bits B_n may be selected. However, such choice may result to increased latency during the local parameter transmission phase, while the considered trade-off is studied later. Finally, when $B_n(j) \to \infty$, $\forall n, j$, the optimality gap is zero and the convergence bound of Theorem 1 coincides with that of a lossless FL model. When it also holds $N = K$, i.e., the full-user participation scenario, the convergence bound is reduced to that of the vanilla FedAvg algorithm [33].

At this point, it should be highlighted that in the convergence analysis of [19], where a stochastic quantization scheme was also considered, authors concluded that the quantization error does not create an optimality gap, i.e.,

$\mathbb{E}\left[F(\boldsymbol{w}(T)) - F(\boldsymbol{w}^*)\right]$ tends asymptotically to zero. However, in their analysis, they selected specific values for the quantization bits $B_n(t)$, given as a function either of the learning rate $\eta(t)$ or residual convergence-related parameters. Also, in [26], authors introduced a relationship among the quantization bits of two consecutive communication rounds towards guaranteeing convergence. In opposition to these, in our analysis we do not restrict $B_n(t)$ to take certain values. In this manner, the values of $B_n(t)$ are not being enforced by any parameter or the previous rounds' bits selection $B_n(t-1)$. This fact is of significant importance, since the constraints imposed by the wireless environment could affect the selection of $B_n(t)$, i.e., it cannot be always feasible or communication efficient to pre-assign specific values to $B_n(t)$, resulted from the convergence analysis. The constraints on $B_n(t)$ would be obvious through the optimization problem that is formulated in the subsequent section.

IV. Convergence Time Minimization

A. Problem Formulation

Our objective goal is to minimize the total convergence time of the FL task, i.e., the overall latency across all FL rounds, under energy and quantization error constraints, with the latter aiming to retain the optimality gap of the upper bound of $\mathbb{E}\left[F(\boldsymbol{w}(T))\right] - F(\boldsymbol{w}^*)$, at small levels. Note that the upper bound is affected by the quantization error through the term $\sum_n \frac{1}{N}\frac{\delta_n^2(t)}{\left(2^{B_n(t)}-1\right)^2}$, as concluded in Theorem 1. Therefore, it is obvious that by increasing the number of quantization bits $B_n(t)$, the quantization error decreases. However, this strategy increases the size $S_n(t)$ of the local model parameters and thus, may result in increased transmission latency. Taking this into account, it is important to balance the considered trade-off, among model accuracy and fast convergence. Hence, we formulate the following optimization problem

$$
\begin{aligned}
\min_{l^{\mathrm{c}},\boldsymbol{E},\boldsymbol{B},\boldsymbol{f},\boldsymbol{l}^{\mathrm{up}}} \quad & \sum_{t=0}^{T-1} l^{\mathrm{r}}(t) \\
\textbf{\textit{s.t.}} \quad & \mathrm{C}_1: l_n^{\mathrm{up}}(t) W \log_2\left(1+\frac{g_n(t)E_n(t)}{l_n^{\mathrm{up}}(t)WN_0}\right) \\
& \qquad \geq d(B_n(t)+1)+m, \quad \forall n \in \mathcal{N}(t), \forall t, \\
& \mathrm{C}_2: \tau\zeta c_n D_n f_n^2(t) + E_n(t) \leq E_n^{\max}(t), \ \forall n \in \mathcal{N}(t), \\
& \mathrm{C}_3: \sum_{n\in\mathcal{N}(t)} \frac{1}{N}\frac{\delta_n^2(t)}{\left(2^{B_n(t)}-1\right)^2} \leq \epsilon(t), \quad \forall t, \\
& \mathrm{C}_4: l^{\mathrm{c}}(t) \geq \tau\frac{c_n D_n}{f_n(t)}, \quad \forall n \in \mathcal{N}(t), \quad \forall t, \\
& \mathrm{C}_5: 0 \leq f_n(t) \leq f_n^{\max}(t), \quad B_n(t) \in \mathbb{Z}_+, \quad \forall t,
\end{aligned}
\tag{18}
$$

where C_1 is related with the successful transmission of the local training parameters, C_2 indicates that the dedicated energy both for computation and transmission purposes, cannot exceed the maximum available energy of the n-th user at the t-th round, i.e., $E_n^{\max}(t)$. Moreover, C_3 implies that the quantization error should not exceed a required tolerance, $\epsilon(t)$, at the respective round. We also clarify that it is reasonable to constrain the quantization error per global FL round t, since there is no coupling of the error among different global rounds, as observed in (16). Finally, C_4 stems from (12), while $f_n^{\max}$ denotes the maximum CPU clock speed of user n. Also, note that the quantization bits $B_n(t)$ are positive integers. Moreover, we highlight that the selection of increased number of bits $B_n(t)$, leads to better model precision, which is reflected in C_3 through the selection of the error tolerance $\epsilon(t)$. However, it is observed from C_1 that such policy also increases the transmission delay and subsequently the total convergence time. To this end, it is evident from C_1 that the number of quantization bits $\boldsymbol{B}(t)$ should be carefully adjusted and not pre-determined [19], due to its coupling with the uplink transmission time intervals $\boldsymbol{l}^{\mathrm{up}}(t)$, which directly affect the total convergence time. Hence, by pre-assigning a dedicated number of quantization bits only to meet theoretical convergence guarantees, may highly increase the uplink transmission interval and subsequently slow down the convergence with respect to time unit. Finally, it is clarified that although the effects of the channel gains are not directly present in the optimality gap of Theorem 1, it is evident that through C_1, the channel gains will affect the optimal value of $\boldsymbol{B}(t)$. Therefore, there is an underlying dependency among the channel conditions and the value of the optimality gap, which can be attributed to the number of quantization bits, $\boldsymbol{B}(t)$.

B. Proposed Solution

It should be highlighted that problem (18) is intractable in the current form, since at the t-th round the channel gains $g(t'), \forall t' > t$ are unknown. However, this is not restricting, since we can address this issue by solving the problem round-by-round, in an online fashion. In addition to this, we relax $B_n \in \mathbb{Z}_+$ to $B_n \geq 1, \forall n \in \mathcal{N}$. Thus, the problem in (18) should be solved in each global round, while hereinafter the t notation is dropped for simplicity. Next, by observing that C_4 is equivalent to: $f_n \geq \tau\frac{c_n D_n}{l^{\mathrm{c}}}, \forall n \in \mathcal{N}$, we introduce the following proposition:

Proposition 1: The optimal $f_n, \forall n \in \mathcal{N}$, satisfy

$$
f_n^* = \tau\frac{c_n D_n}{l^{\mathrm{c}*}}, \quad \forall n \in \mathcal{N}, \tag{19}
$$

with $l^{\mathrm{c}*} \geq a_1 \triangleq \max_{n\in\mathcal{N}}\left\{\frac{\tau c_n D_n}{f_n^{\max}}\right\}$.

Proof: First, by manipulating C_4, it is straightforward to show that $l^{\mathrm{c}} \geq a_1$. Following that, let consider a known $\bar{l}^{\mathrm{c}}$. By observing C_2, it is obvious that the selection of larger f_n decreases the value of E_n. Moreover, it easy to verify that $l_n^{\mathrm{up}} W \log_2\left(1+\frac{g_n E_n}{l_n^{\mathrm{up}} W N_0}\right)$ in C_1, is an increasing function w.r.t. both E_n and l_n^{up}. Therefore, the selection of smaller E_n will lead to increased l_n^{up}, while the objective is to minimize l_n^{up}. From the aforementioned, f_n should be selected as small as possible, given the local computation duration, which from C_4 concludes to (19). ■

Proposition 1 implies that the CPU clock speed $f_n, \forall n$, should be selected in such a way that all users terminate the com-

TABLE I
LIST OF NOTATIONS

Parameter	Description	Parameter	Description
$\mathcal{K}$	Set of users	$\mathcal{N}(t)$	Set of participating users at round t
$\mathcal{D}_n^{\mathrm{L}}$	Local dataset of user n	$F_n(\cdot)$	Loss function of user n
$F(\cdot)$	Global loss function	$\boldsymbol{w}(t)$	Global weight at round t
T	Number of global rounds	τ	Number of local iterations
$\boldsymbol{w}_n^i$	Weight of user n at the i-th local iteration	$\Delta\boldsymbol{w}_n(t)$	Local model update of user n at t-th round
$\boldsymbol{\xi}_n^i$	Minibatch of user n at the i-th local iteration	$\eta(t)$	Learning rate at the t-th round
$Q(\cdot)$	Quantization function	$B_n(t)$	Number of quantization bits of user n at round t
f_n	CPU frequency of user n	l_n^{up}	Uplink time-slot duration of user n
ϵ	quantization error tolerance	E_n	Transmit energy of user n
l^{c}	Local computation time duration	$\boldsymbol{w}^*$	$\mathrm{argmin}_{\boldsymbol{w}} F(\boldsymbol{w})$

putation phase concurrently. By exploiting Proposition 1, the problem in (18) can be re-written as

$$\begin{aligned}
\min_{l^{\mathrm{c}},\boldsymbol{E},\boldsymbol{B},\boldsymbol{l}^{\mathrm{up}}} \quad & l^{\mathrm{c}} + \sum_{n\in\mathcal{N}} l_n^{\mathrm{up}} \\
\textbf{\textit{s.t.}} \quad & \mathrm{C}_1: l_n^{\mathrm{up}} W \log_2\left(1+\frac{g_n E_n}{l_n^{\mathrm{up}} W N_0}\right) \\
& \quad \geq d(B_n+1)+m, \quad \forall n\in\mathcal{N}, \\
& \mathrm{C}_2: \frac{\zeta\tau^3 c_n^3 D_n^3}{{l^{\mathrm{c}}}^2} + E_n \leq E_n^{\max}, \quad \forall n\in\mathcal{N}, \\
& \mathrm{C}_3: \sum_{n\in\mathcal{N}} \frac{1}{N}\frac{\delta_n^2}{\left(2^{B_n}-1\right)^2} \leq \epsilon, \\
& \mathrm{C}_4: l^{\mathrm{c}} \geq a_1, \quad E_n, l_n^{\mathrm{up}} \geq 0, \forall n\in\mathcal{N}, \\
& \mathrm{C}_5: B_n \geq 1, \forall n\in\mathcal{N}.
\end{aligned} \tag{20}$$

It can be easily shown that the problem in (20) is jointly convex with respect to all the considered variables, while the proof is omitted due to space limitations. The problem in (20) will be solved via the *Lagrange dual decomposition*. Firstly, the Lagrangian function can be written as

$$\begin{aligned}
\mathcal{L}(l^{\mathrm{c}},\boldsymbol{l}^{\mathrm{up}},\boldsymbol{E},\boldsymbol{B},\boldsymbol{\lambda}) &= l^{\mathrm{c}} + \sum_{n\in\mathcal{N}} l_n^{\mathrm{up}} \\
&+ \sum_{n\in\mathcal{N}} \lambda_{1,n}\left(d(B_n+1)+m - l_n^{\mathrm{up}} W\log_2\left(1+\tfrac{g_n E_n}{l_n^{\mathrm{up}} W N_0}\right)\right) \\
&+ \sum_{n\in\mathcal{N}} \lambda_{2,n}\left(\frac{\zeta\tau^3 c_n^3 D_n^3}{{l^{\mathrm{c}}}^2} + E_n - E_n^{\max}\right) \\
&+ \lambda_3\left(\sum_{n\in\mathcal{N}} \frac{1}{N}\frac{\delta_n^2}{\left(2^{B_n}-1\right)^2} - \epsilon\right) + \lambda_4(a_1 - l^{\mathrm{c}}) \\
&+ \sum_{n\in\mathcal{N}} \lambda_{5,n}(1-B_n),
\end{aligned} \tag{21}$$

where $\boldsymbol{\lambda} = (\lambda_{1,1},...,\lambda_{2,1},...,\lambda_{5,N}) \geq 0$ ('$\geq$' denotes the component-wise inequality) is the LM vector and $\lambda_{1,n},\lambda_{2,n},\lambda_3,\lambda_4,\lambda_{5,n}, \forall n \in \mathcal{N}$, are associated with the constraints C_i, $i=1,...,5$, respectively. Following that, the dual function is given as

$$\mathcal{G}(\boldsymbol{\lambda}) = \min_{l^{\mathrm{c}},\boldsymbol{E},\boldsymbol{B},\boldsymbol{l}^{\mathrm{up}}} \mathcal{L}(l^{\mathrm{c}},\boldsymbol{l}^{\mathrm{up}},\boldsymbol{E},\boldsymbol{B},\boldsymbol{\lambda}), \tag{22}$$

while the corresponding dual problem can be written as

$$\max_{\boldsymbol{\lambda}} \ \mathcal{G}(\boldsymbol{\lambda}). \tag{23}$$

Since the primal problem is convex and the Slater's conditions are satisfied, strong duality holds, i.e., solving the dual in (23) is equivalent to solving the primal problem in (20), [34]. According to the Karush-Kuhn-Tucker (KKT) conditions, the optimal solution to the problem should satisfy

$$\nabla\mathcal{L}(l^{\mathrm{c}*},\boldsymbol{l}^{\mathrm{up}*},\boldsymbol{E}^*,\boldsymbol{B}^*,\boldsymbol{\lambda}^*) = 0. \tag{24}$$

Thus, by taking $\frac{\partial\mathcal{L}}{\partial l^{\mathrm{c}}} = 0$ and $\frac{\partial\mathcal{L}}{\partial E_n} = 0$, $\forall n\in\mathcal{N}$, leads to

$$l^{\mathrm{c}*} = \sqrt[3]{\frac{2\zeta\tau^3\sum_n \lambda_{2,n}^* c_n^3 D_n^3}{1-\lambda_4^*}}, \tag{25}$$

and

$$E_n^* = l_n^{\mathrm{up}*} W\left(\frac{\lambda_{1,n}^*}{\lambda_{2,n}^*\ln(2)} - \frac{N_0}{g_n}\right), \forall n\in\mathcal{N}. \tag{26}$$

From (26), we observe that $\lambda_{2,n}^* \neq 0, \forall n \in \mathcal{N}$. Taking this into account, according to the complementary slackness conditions which require

$$\lambda_{2,n}^*\left(\frac{\zeta\tau^3 c_n^3 D_n^3}{{l^{\mathrm{c}*}}^2} + E_n^* - E_n^{\max}\right) = 0, \quad \forall n\in\mathcal{N}, \tag{27}$$

the constraint C_2 should be satisfied with equality [34], leading to

$$E_n^* = E^{\max} - \frac{\zeta\tau^3 c_n^3 D_n^3}{{l^{\mathrm{c}*}}^2}, \quad \forall n\in\mathcal{N}. \tag{28}$$

This is reasonable, since it indicates that users should utilize their whole available energy, towards minimizing the objective function. Following that, $\frac{\partial\mathcal{L}}{\partial l_n^{\mathrm{up}}} = 0, \forall n\in\mathcal{N}$, it holds that

$$\frac{W\lambda_{1,n}^*}{\ln(2)}\frac{-(1+\frac{b}{l_n^{\mathrm{up}*}}) + 1 + (1+\frac{b}{l_n^{\mathrm{up}*}})\ln\left(1+\frac{b}{l_n^{\mathrm{up}*}}\right)}{1+\frac{b}{l_n^{\mathrm{up}*}}} = 1, \tag{29}$$

where $b = \frac{E_n^* g_n}{W N_0}$. The manipulation of (29), results to

$$l_n^{\mathrm{up}*} = -\frac{g_n E_n^*}{W N_0(1+\mathcal{W}_0^{-1}(\psi_n))}, \quad \forall n\in\mathcal{N}, \tag{30}$$

where $\mathcal{W}_0$ is the principal branch of the Lambert W function [35] and ψ_n is given by

$$\psi_n = -\frac{2^{-\frac{1}{W\lambda_{1,n}^*}}}{e}, \quad \forall n\in\mathcal{N}. \tag{31}$$

Furthermore, it is easy to verify from (29) that $\lambda_{1,n}^* \neq 0$, $\forall n \in \mathcal{N}$, since the case $\lambda_{1,n}^* = 0$, leads to a contradiction. This

result indicates that users should spend their resources so as to transmit exactly S_n bits to the server, which is observed from the right-hand-side of C_1. Thus, C_1 is satisfied with equality, yielding

$$B_n^* = \frac{l_n^{\mathrm{up}*} W}{d} \log_2 \left(1 + \frac{g_n E_n^*}{l_n^{\mathrm{up}*} W N_0}\right) - \frac{m}{d} - 1, \quad \forall n \in \mathcal{N}. \tag{32}$$

Finally, by taking $\frac{\partial \mathcal{L}}{\partial B_n} = 0$, $\forall n \in \mathcal{N}$, yields

$$\left(2^{B_n^*} - 1\right)^3 = \frac{1}{N} \frac{2\lambda_3^* \ln(2) \delta_n^2}{d\lambda_{1,n}^* - \lambda_{5,n}^*} 2^{B_n^*}. \tag{33}$$

From (33), it is obvious that $\lambda_3^* \neq 0$, since the case $\lambda_3^* = 0$ leads to $B_n^* = 0$, which is infeasible. Therefore, also constraint C_3 is satisfied with equality.

According to the previous analysis, the optimal variables $l^{\mathrm{c}*}, \boldsymbol{E}^*, \boldsymbol{l}^{\mathrm{up}*}, \boldsymbol{B}^*$, have been given in closed forms in terms of the LMs, by the equations (25), (28), (30), (32), respectively. Note that given the LMs, each optimization variable can be directly calculated, with the aforementioned order of appearance, by using the respective equations. Subsequently, the LMs can be updated iteratively via the subgradient method [36] towards solving the dual problem, while the primal variables can be calculated through the LMs. The algorithm of the above procedure, which outputs the solution of problem (23) is presented below. The number of iterations of the sub-gradient method is of the order $\mathcal{O}(\frac{1}{\tilde{\epsilon}^2})$, where $\tilde{\epsilon}$ is the required tolerance [36].

Algorithm 1 Solution of (20)

1: **Initialize** $k \leftarrow 0$, $l^{\mathrm{c}(0)}, \boldsymbol{E}^{(0)}, \boldsymbol{l}^{\mathrm{up}(0)}, \boldsymbol{B}^{(0)}$
2: **Repeat**
3: **Update** the LM vector $\boldsymbol{\lambda}^{(k+1)}$ (refer to Appendix C).
4: **Calculate** $l^{\mathrm{c}(k+1)}, \boldsymbol{E}^{(k+1)}, \boldsymbol{l}^{\mathrm{up}(k+1)}, \boldsymbol{B}^{(k+1)}$ using the
5: equations (25), (28), (30), (32)
6: $k \leftarrow k + 1$
7: **Until** convergence
8: **Output** $l^{\mathrm{c}*}, \boldsymbol{E}^*, \boldsymbol{l}^{\mathrm{up}*}, \boldsymbol{B}^*$ and $\boldsymbol{f}^*$ from (22)

Following this analysis, recall that $B_n, \forall n \in \mathcal{N}$, should finally take integer values. Therefore, after obtaining the solutions to the problem, B_n^* should be rounded to the smallest integer which is greater or equal to B_n^*. Thus we set $\tilde{B}_n^* = \lceil B_n^* \rceil$, in order to guarantee that C_3 is still satisfied, i.e., the quantization error tolerance constraint is not violated. However, since with the selection of B_n^* it was previously shown that C_1 is satisfied with equality, by plugging $\tilde{B}_n^*$ into the problem, C_1 will be now violated, due to the fact that $\tilde{B}_n^* > B_n^*$. Therefore, to address this issue, the problem in (20) has to be resolved for a fixed value of $\tilde{B}_n^*$, i.e.,

$$\min_{l^{\mathrm{c}}, \boldsymbol{E}, \boldsymbol{l}^{\mathrm{up}}; \tilde{\boldsymbol{B}}^*} \quad l^{\mathrm{c}} + \sum_{n=1}^{N} l_n^{\mathrm{up}}, \qquad \textbf{\textit{s.t.}} \quad \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_4, \tag{34}$$

and finally the optimal variables can be obtained. It should be clarified that (34) can be solved similarly to (20) for a fixed $\tilde{\boldsymbol{B}}^*$. Hence, (34) can be solved with a slight modification of Algorithm 1, where $\boldsymbol{B}$ is no longer an optimization variable and can be treated as a constant, assigned with the value $\tilde{\boldsymbol{B}}^*$. To this end, the CPU frequency $\boldsymbol{f}^*$ is given by (19), which concludes the overall solution to the problem.

TABLE II
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
$f_n^{\max}$	1.5 GHz	D_n	1 Mbit
W	0.3 MHz	N_0	-174 dBm/Hz
ζ	10^{-27}	c_n	$\sim \mathcal{U}(10, 40)$
N	10 users	d_n	$\sim \mathcal{U}(0, 1000\mathrm{m})$
$E_n^{\max}$	0.3 Joule	m	64 bits
d	23820	β	3.75

V. NUMERICAL EXPERIMENTS AND PERFORMANCE EVALUATION

For the simulation results, we assume that the users are uniformly distributed in a circle with radius 1000m, while the server/BS is located at the center of the circle. Also, $N = 10$ users are selected for participation in each global round. The rest of the simulation parameters are presented in Table II, and retain their respective values unless specified otherwise.

We select the FL task to be the image classification on the widely-known MNIST dataset [37]. We assume that each user carries 200 data samples and trains a fully-connected feed-forward neural network with a single hidden layer, consisting of 30 nodes, while the softmax is the activation function of the output layer. Thus, the total model parameters are $d = 23860$ (i.e., $784 \times 30 + 30 \times 10 = 23820$ weights and $30 + 10 = 40$ biases). The mini-batch size has been set as $|\boldsymbol{\xi}_n^i(t)| = 50$, $\forall n, i, t$. Moreover, the Adam optimizer is utilized for the local training [38]. Following that, we consider two different cases of training data distributions. Firstly, for the case of IID data distribution among users, the training data is shuffled and randomly assigned to each user. Secondly, for the non-IID scenario, the training data is sorted by labels and each user is equipped only with 5 labels. For both cases, the datasets among users are non-overlapping. All results have been conducted on the MNIST dataset, unless specified otherwise. Finally, although the convergence analysis assumed convex problems, it will be evident in the following subsections that useful insights can still be derived and theoretical observations be validated via numerical results.

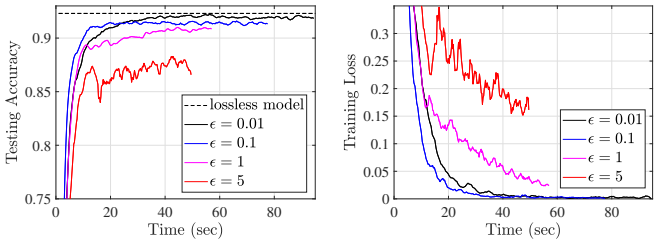
A. Effects of the quantization error tolerance on the convergence time and model accuracy

In Fig. 3(a) and Fig. 3(b), the testing accuracy and training loss, respectively, are evaluated for various values of the quantization error tolerance, ϵ. Note that the proposed optimization method has been utilized in order to extract all figures. In the considered simulations, we set $\tau = 2$ local iterations for each user, while we set the total number of global communication rounds equal to $T = 225$. It should be highlighted though, that the x axis illustrates the time in seconds, and not purely the evolution of the global rounds. We made this choice, since for different tolerance values, the duration of a global round also varies. In this manner, the comparison on the convergence time between various values of ϵ can be fairly conducted. Also, we clarify that total the training time in seconds, given the T global rounds, differs among different choices of ϵ. From Fig. 3(a), it is evident that for smaller ϵ, higher testing accuracy is achieved, which

also approximates the performance of the lossless model, i.e., the FL model without applying quantization. Moreover, high values of ϵ may lead to decreased model accuracy, e.g., when $\epsilon = 5$. Another interesting observation comes as follows. The case $\epsilon = 0.1$, demonstrates the highest convergence rate in the early stages of the training process, which is also greater than the case $\epsilon = 0.01$. This outcome is related with the duration of each global communication round. Smaller values of ϵ, translate to the selection of more quantization bits, which in turn result in increased transmission latency during each communication round. More specifically, in Fig. 4, the average delay per global round and the average number of quantization bits per user per global round are presented. It can been seen that the case $\epsilon = 0.01$ presents the highest average delay per round, among the rest choices of ϵ. Therefore, this fact can slow down the convergence speed in the early stages of the training, although finally the highest accuracy is achieved, owing this to the selection of more quantization bits which guaranty high precision. Based on the above, it is evident that a model accuracy/fast convergence trade-off is present, among the cases $\epsilon = 0.01$ and $\epsilon = 0.1$.

At this point, it is significant to highlight the following observation. The cases $\epsilon = 1$ and $\epsilon = 5$ do not really contribute neither in model accuracy nor in fast convergence, since they are totally outperformed by the cases $\epsilon = 0.1$ and $\epsilon = 0.01$, in both aspects. Even when targeting smaller accuracy values, the higher-ϵ cases fail to converge faster than the cases $\epsilon = 0.1$ and $\epsilon = 0.01$. An interpretation of this result is the following: The large number of communication rounds until convergence, which occur from the low precision quantization, prevails over the low-latency per round. Therefore, by selecting a relatively loose quantization error tolerance, which is equivalent to utilizing a few quantization bits, may not result in any gains or offer any benefits. This contradicts the fact which implies that by using a small number of quantization bits, communication efficiency is always achieved. Accordingly, in Fig. 5, the testing accuracy and training loss are evaluated for the NON-IID scenario. For this example, we set $\tau = 3$ and $T = 150$. Similar behavior with the IID case is observed, while the model accuracy is degraded.

To further showcase the impact of the quantization error tolerance on the performance, we also ran experiments on the CIFAR-10 dataset [39]. We assume that each user is equipped with 2500 training samples (IID) and trains a convolutional neural network (CNN) of the following structure: A 3×3 convolutional layer with 32 channels, a 5×5 convolutional layer with 64 channels, both followed by 2×2 max pooling and relu activation, a fully connected layer with 64 units and relu activation, and a final output layer with softmax. The second convolutional layer is also followed by a dropout layer w.p. 0.2. The number of the training parameters is $d = 315,018$. The mini-batch size is selected as $|\boldsymbol{\xi}_n^i(t)| = 128, \forall n, i, t.$. We also set $\tau = 2$ and $T = 40$. Following that, in Fig. 6, various values of ϵ are illustrated. The results are very similar to the MNIST dataset experiments. Specifically, it is evident that the curve which results in faster convergence, is the $\epsilon = 20$, presenting a slight accuracy decrease compared to the case $\epsilon = 1$. Hence, the trade-off among timely convergence and model accuracy is again highlighted. Finally, we clarify that the selection of the ϵ values is different to that of the MNIST dataset. This can be attributed to the different range of the training parameters for different dataset and neural network architecture, i.e., the CNN in the case of CIFAR-10. Thus, the range of $\delta_n(t)$ in (15), which is proportional to the model size d, is affected. As a consequence, the range of ϵ is also being affected through constraint C_3. It is noted that the values $1, 20, 50, 100$ of ϵ, correspond to $6.38, 4.07, 3.52, 3.13$ average quantization bits per user per round, respectively.



(a) Testing accuracy versus time. (b) Training loss versus time.

Fig. 3. IID scenario.

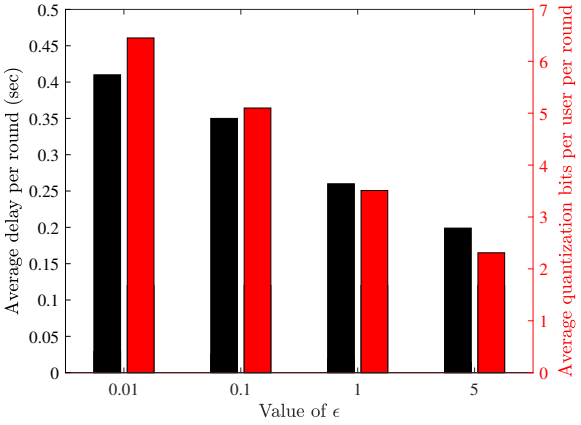
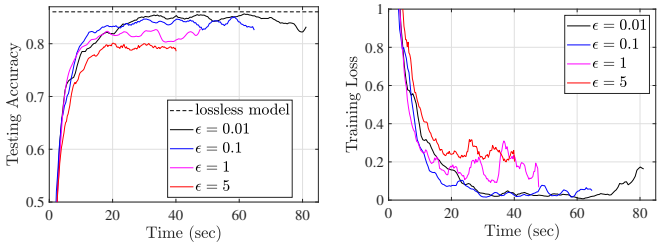


Fig. 4. Average delay per round and average quantization bits per user per round, for the IID scenario.



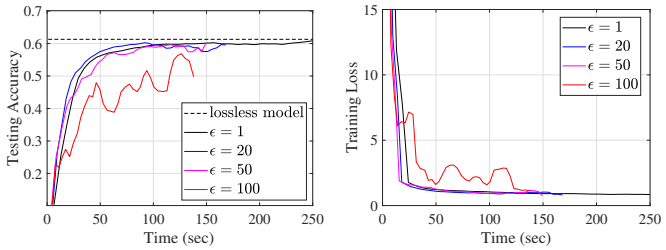
(a) Testing accuracy versus time. (b) Training loss versus time.

Fig. 5. NON-IID scenario.

B. Comparison with baseline schemes

In Fig. 7, we compare the performance of the proposed optimization scheme with some baseline schemes. Firstly, we

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2023.3258485



(a) Testing accuracy versus time. (b) Training loss versus time.

Fig. 6. Experiments on CIFAR-10.

consider the *enhanced FedAvg (e-FedAvg)* baseline scheme. In *e-FedAvg*, the proposed optimization method is adopted, but the number of quantization bits is pre-assigned as $\bar{B}_n = 16, \forall n \in \mathcal{N}$ in each global round. Hence, we select an adequate number of quantization bits to model the lossless behavior of the standard FedAvg algorithm, since the implementation of a pure lossless model would require the usage of infinite number of quantization bits, and thus, infinite transmission time which is impractical. The term *enhanced* justifies the optimization of the available resources, since the vanilla FedAvg does not perform any optimization. Moreover, the *equal slots allocation* scheme also exploits the proposed optimizations' method, however it assigns equal duration uplink time slots to all users. Finally, the *equal energy allocation* scheme assigns equal energy values for both the computation and transmission phase, while the rest of the optimization is conducted according to the proposed method. For the considered simulation, we set the quantization error tolerance $\epsilon = 0.01$, consider an IID scenario and set $T = 225$ global rounds. By observing Fig. 7, it can be seen that all schemes demonstrate almost identical testing accuracy, which is related with the selected quantization error tolerance. However, it is clearly seen that the proposed scheme dominates all baseline schemes, in terms of convergence time, which is the objective goal of the proposed optimization. Thus, Fig. 7 highlights the significance of the proposed scheme, which jointly takes into account the communication and computation resources, as well as the quantization bits allocation. Also, the case of *e-FedAvg* leads to very high latency until convergence, without offering further accuracy gain. Therefore, it is evident that when aiming towards fast convergence of the FL process, the number of quantization bits ought to be wisely selected.

C. The dynamical adjustment of quantization error tolerance

In the continue, we examine the effects of dynamically adjusting the quantization error tolerance ϵ throughout the training process. Recall that in Theorem 1, it was evident that in the early training stages, the quantization error has not large impact in the optimality gap. Driven by this fact, we now focus on decaying ϵ along with the evolution of the global rounds. Specifically, we consider that $\epsilon(i+1) = r \cdot \epsilon(i)$, where $0 < r < 1$ is a constant, $i = 1, ..., 223$, while we initialize $\epsilon(0) = 0.1$. By setting $r = 10^{-1/224}$, it is easy to verify that $\epsilon(224) = 0.01$, i.e., ϵ is equal to 0.1 in the first round and equal to 0.01 in the final round, given that $T = 225$. Following that, in Fig. 8, we compare the performance of the considered technique, i.e., *decaying* ϵ, with the standard cases of a constant ϵ throughout the training. It can be observed that by dynamically decreasing ϵ, the convergence rate is significantly increased. More specifically, in the very early stages of the training, higher ϵ values contribute to fast communication with the server, while along with the reduction of ϵ, the precision is gradually increased. This policy results in fast convergence and notable performance, in comparison with the rest fixed ϵ cases and especially with the stringent case, where $\epsilon = 0.01$. Specifically, although the testing accuracy is almost identical between the *decaying* ϵ policy and the case $\epsilon = 0.01$, the former converges after 22 seconds of training, while the latter after about 40 seconds. Therefore, the effectiveness of decreasing the quantization error tolerance along with the evolution of the training, which enforces the gradual increase of the number of quantization bits throughout the training, is corroborated.

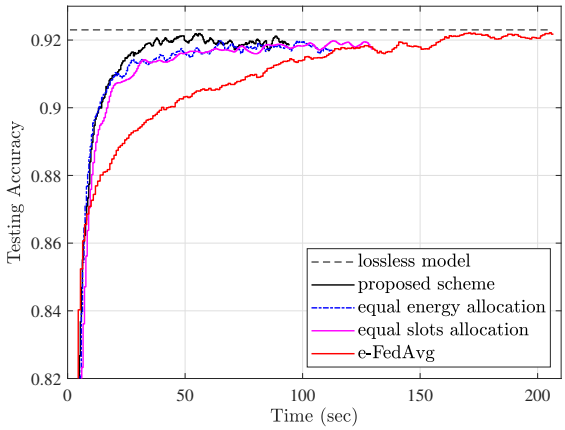


Fig. 7. Comparison of the proposed scheme with baseline schemes.

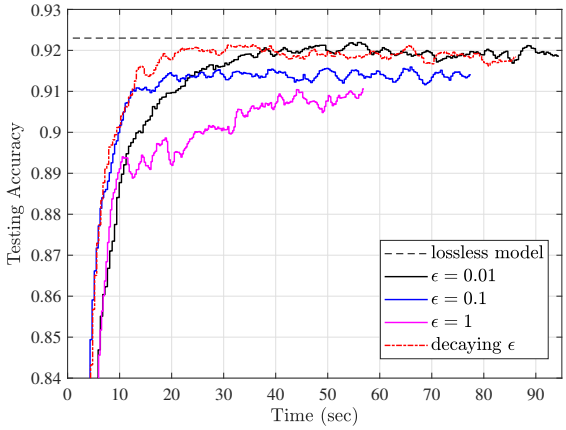


Fig. 8. The impact of dynamically adjusting ϵ, on the convergence rate and testing accuracy.

VI. Conclusions

In this paper, we studied and optimized the performance of FL over wireless networks by considering the quantization of

the local model parameters. More specifically, we have jointly optimized the communication and computation resources, as well as the quantization bits allocation, focusing on minimizing the total convergence time of FL subject to energy constraints and quantization error tolerance. The optimization problem was coupled with the convergence analysis, aiming to control the impact of the quantization error and subsequently balance the trade-off between model accuracy and fast convergence. Simulations are conducted, where the considered trade-off is examined and the effectiveness of the proposed method in accelerating the convergence speed, is verified. Also, the results indicate that the selection of the quantization error tolerance is critical for achieving enhanced performance in FL, while efficient techniques are presented which result in increased convergence rate.

APPENDIX A
PROOF OF THEOREM 1

For proving Theorem 1, we adopt the methodologies of [33] and [32]. Firstly, we define the auxiliary variable $\boldsymbol{v}(t)$ as

$$\boldsymbol{v}(t+1) = \boldsymbol{w}(t) + \sum_{n\in\mathcal{N}(t)} \frac{1}{N}\Delta\boldsymbol{w}_n(t), \tag{35}$$

which represents a lossless model's update during the $(t+1)$-th round by considering the set of selected users. Following that, we also define $z(t)$ as

$$\boldsymbol{z}(t+1) = \boldsymbol{w}(t) + \sum_{n=1}^{K} \frac{1}{K}\Delta\boldsymbol{w}_k(t), \tag{36}$$

which denotes the lossless model's update when all networks users are participating. Recall that

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \sum_{n\in\mathcal{N}(t)} \frac{1}{N}Q(\Delta\boldsymbol{w}_n(t)). \tag{37}$$

Following that, we have

$$\begin{aligned}
&\|\boldsymbol{w}(t+1) - \boldsymbol{w}^*\|_2^2 \\
&\quad= \|\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1) + \boldsymbol{v}(t+1) - \boldsymbol{w}^*\|_2^2 \\
&\quad= \|\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1)\|_2^2 + \|\boldsymbol{v}(t+1) - \boldsymbol{w}^*\|_2^2 \\
&\qquad+ 2\langle \boldsymbol{w}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{w}^*\rangle.
\end{aligned} \tag{38}$$

In the continue, the average of the right-hand-side terms in (38) are bounded and presented in Lemmas 2-4 respectively.

Lemma 2: We have

$$\mathbb{E}\left[\|\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1)\|_2^2\right] \le \sum_{n\in\mathcal{N}(t)} \frac{1}{N} J_n^2(t).$$

Proof:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1)\|_2^2\right] \\
&\quad= \mathbb{E}\left[\left\|\sum_{n\in\mathcal{N}(t)} \frac{1}{N}(Q(\Delta\boldsymbol{w}_n(t)) - \Delta\boldsymbol{w}_n(t))\right\|_2^2\right] \\
&\quad\overset{(a)}{\le} \sum_{n\in\mathcal{N}(t)} \frac{1}{N}\mathbb{E}\left[\|(Q(\Delta\boldsymbol{w}_n(t)) - \Delta\boldsymbol{w}_n(t))\|_2^2\right] \\
&\quad\overset{(b)}{\le} \sum_{n\in\mathcal{N}(t)} \frac{1}{N} J_n^2(t),
\end{aligned} \tag{39}$$

where (a) follows from the convexity of $\|\cdot\|_2^2$ and the fact that $\sum_{n=1}^{N} \frac{1}{N} = 1$, while (b) follows from Lemma 1. ■

Lemma 3: We have

$$\mathbb{E}\left[\|\boldsymbol{v}(t+1) - \boldsymbol{w}^*\|_2^2\right] \le -\mu\eta(t)\mathbb{E}\left[\|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2\right] + \eta^2(t)U \tag{40}$$

where

$$\begin{aligned}
U = \tau^2 \sum_{n=1}^{K} \frac{\sigma_n^2}{K} + 2L\tau^2\Gamma + (\mu+2)\frac{\tau(\tau-1)(2\tau-1)}{6}G^2 \\
+ \tau G^2 + \frac{K-N}{N(K-1)}\tau^2 G^2.
\end{aligned} \tag{41}$$

Proof: See Appendix B. ■

Lemma 4: We have

$$\mathbb{E}\left[2\langle \boldsymbol{w}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{w}^*\rangle\right] = 0. \tag{42}$$

Proof: Since it holds

$$\mathbb{E}\left[Q(\Delta\boldsymbol{w}_n(t))\right] = \Delta\boldsymbol{w}_n(t), \tag{43}$$

and

$$\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1) = \sum_{n\in\mathcal{N}(t)} \frac{1}{N}(Q(\Delta\boldsymbol{w}_n(t)) - \Delta\boldsymbol{w}_n(t)), \tag{44}$$

it is straightforward to conclude to (42). ■

Following that, by combining the results in Lemmas 2-4, (38) leads to

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{w}(t+1) - \boldsymbol{w}^*\|_2^2\right] \le (1-\eta(t)\mu)\mathbb{E}\left[\|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2\right] \\
+ \eta^2(t)U + \sum_{n\in\mathcal{N}(t)} \frac{1}{N} J_n^2(t).
\end{aligned} \tag{45}$$

Let $\Delta_t = \mathbb{E}\left[\|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2\right]$. (45) can be re-written as

$$\Delta_{t+1} \le (1-\eta(t)\mu)\Delta_t + \eta^2(t)U + \sum_{n\in\mathcal{N}(t)} \frac{1}{N} J_n^2(t). \tag{46}$$

Next, we will show that $\Delta_t \le \frac{\nu}{\gamma+t} + \Psi(t)$ where

$$\Psi(x) \triangleq \sum_{j=0}^{x-1} \sum_{n\in\mathcal{N}(j)} \frac{1}{N} J_n^2(j) \prod_{i=j+1}^{x-1} (1-\eta(i)\mu), \quad x \ge 1, \tag{47}$$

by selecting a diminishing learning rate $\eta(t) = \frac{\beta}{\gamma+t}$, with $\beta \ge \frac{1}{\mu}$, $\gamma \ge \beta$ such that $\eta(0) \le 1$, $\gamma \ge \beta\mu$ such that $\eta(0) \le \frac{1}{\mu}$ and $\nu \le \max\left\{\frac{\beta^2 U}{\beta\mu - 1}, \gamma\Delta_0\right\}$. Also, since we required $\eta(t) \le \frac{1}{L\tau}$, it should also hold $\gamma \ge \frac{L}{\mu}$. Similarly to [33], via induction we have

$$\begin{aligned}
\Delta_{t+1} &\le (1-\eta(t)\mu)\Delta_t + \eta^2(t)U + \sum_{n\in\mathcal{N}(t)} \frac{1}{N} J_n^2(t) \\
&\le \left(1 - \frac{\beta\mu}{\gamma+t}\right)\left(\frac{\nu}{\gamma+t} + \Psi(t)\right) + \frac{\beta^2 U^2}{(\gamma+t)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2}\nu + \left(\frac{\beta^2 U^2}{(\gamma+t)^2} - \frac{\beta\mu - 1}{(t+\gamma)^2}\nu\right) \\
&\quad+ \Psi(t+1) \\
&\le \frac{\nu}{t+\gamma+1} + \Psi(t+1).
\end{aligned} \tag{48}$$

Following that, we have

$$\begin{aligned}\nu &\leq \max\left\{\frac{\beta^2 U}{\beta\mu-1}, \gamma\Delta_0\right\} \leq \frac{\beta^2 U}{\beta\mu-1} + \gamma\Delta_0 \\ &\overset{\beta=\frac{2}{\mu}}{\leq} \frac{4U}{\mu^2} + \gamma\mathbb{E}\left[\|\boldsymbol{w}(0)-\boldsymbol{w}^*\|_2^2\right].\end{aligned} \tag{49}$$

By substituting (49) in (48) and by using the fact that $F(\cdot)$ is L-smooth, which gives

$$\begin{aligned}F(\boldsymbol{w}(T)) - F(\boldsymbol{w}^*) &\leq \langle \boldsymbol{w}(T)-\boldsymbol{w}^*, \nabla F(\boldsymbol{w}^*)\rangle \\ &\quad + \frac{L}{2}\|\boldsymbol{w}(T)-\boldsymbol{w}^*\|_2^2 \\ &\leq \frac{L}{2}\|\boldsymbol{w}(T)-\boldsymbol{w}^*\|_2^2,\end{aligned} \tag{50}$$

since $\nabla F(\boldsymbol{w}^*)=0$, the proof of Theorem 1 is completed.

Appendix B
Proof of Lemma 3

Firstly, we have

$$\begin{aligned}&\|\boldsymbol{v}(t+1)-\boldsymbol{w}^*\|_2^2 \\ &\quad= \|\boldsymbol{v}(t+1)-\boldsymbol{z}(t+1)+\boldsymbol{z}(t+1)-\boldsymbol{w}^*\|_2^2 \\ &\quad= \|\boldsymbol{z}(t+1)-\boldsymbol{w}^*\|_2^2 + \|\boldsymbol{v}(t+1)-\boldsymbol{z}(t+1)\|_2^2 \\ &\qquad+ 2\langle \boldsymbol{v}(t+1)-\boldsymbol{z}(t+1), \boldsymbol{z}(t+1)-\boldsymbol{w}^*\rangle.\end{aligned} \tag{51}$$

The first term in the right-hand-side (RHS) of (51) can be expanded as

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{z}(t+1)-\boldsymbol{w}^*\|_2^2\right] \\ &\quad= \mathbb{E}\left[\|\boldsymbol{w}(t)-\boldsymbol{w}^*\|_2^2\right] + \underbrace{\mathbb{E}\left[\left\|\sum_{n=1}^K \frac{1}{K}\Delta\boldsymbol{w}_n(t)\right\|_2^2\right]}_{A_1} \\ &\qquad+ \underbrace{2\mathbb{E}\left[\left\langle \boldsymbol{w}(t)-\boldsymbol{w}^*, \sum_{n=1}^K \frac{1}{K}\Delta\boldsymbol{w}_n(t)\right\rangle\right]}_{A_2}.\end{aligned} \tag{52}$$

For A_1, we have that

$$\begin{aligned}A_1 &= \mathbb{E}\left[\left\|\sum_{n=1}^K \frac{1}{K}\left(-\eta(t)\sum_{i=1}^{\tau}\nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))\right)\right\|_2^2\right] \\ &\overset{(a)}{\leq} \eta^2(t)\tau\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[\left\|\nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))\right\|_2^2\right] \\ &= \eta^2(t)\tau\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\Big[\|\nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t)) \\ &\quad- \nabla F_n(\boldsymbol{w}_n^{i-1}(t)) + \nabla F_n(\boldsymbol{w}_n^{i-1}(t))\|_2^2\Big] \\ &\overset{(b)}{\leq} \eta^2(t)\tau\left(\tau\sum_{n=1}^K\frac{\sigma_n^2}{K} + \sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[\left\|\nabla F_n(\boldsymbol{w}_n^{i-1}(t))\right\|_2^2\right]\right) \\ &\overset{(c)}{\leq} \eta^2(t)\tau^2\sum_{n=1}^K\frac{1}{K}\sigma_n^2 \\ &\quad+ 2L\eta^2(t)\tau\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F_n^*\right],\end{aligned} \tag{53}$$

where (a) follows from the convexity of $\|\cdot\|_2^2$, (b) from Assumption 4 and $\mathbb{E}[\nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))] = \nabla F_n(\boldsymbol{w}_n^{i-1}(t))$, while (c) from the L-smoothness of F_n, which implies that [34]:

$$\left\|\nabla F_n(\boldsymbol{w}_n^{i-1}(t))\right\|_2^2 \leq 2L(F_n(\boldsymbol{w}_n^{i-1}(t)) - F_n^*). \tag{54}$$

In the following we bound the last term in (52), A_2, as:

$$\begin{aligned}A_2 &= 2\sum_{n=1}^K\frac{1}{K}\mathbb{E}\left[\langle \boldsymbol{w}(t)-\boldsymbol{w}^*, \Delta\boldsymbol{w}_n(t)\rangle\right] \\ &= 2\eta(t)\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[\langle \boldsymbol{w}^*-\boldsymbol{w}(t), \nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))\rangle\right] \\ &= 2\eta(t)\times \\ &\quad\Bigg[\underbrace{\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[\langle \boldsymbol{w}_n^{i-1}(t)-\boldsymbol{w}(t), \nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))\rangle\right]}_{B_1} \\ &\quad+ \underbrace{\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[\langle \boldsymbol{w}^*-\boldsymbol{w}_n^{i-1}(t), \nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))\rangle\right]}_{B_2}\Bigg].\end{aligned} \tag{55}$$

Next, for B_1 we have

$$\begin{aligned}B_1 &\overset{(a)}{\leq} \eta(t)\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\Big[\frac{1}{\eta(t)}\left\|\boldsymbol{w}_n^{i-1}(t)-\boldsymbol{w}(t)\right\|_2^2 \\ &\quad+ \eta(t)\left\|\nabla F_n(\boldsymbol{w}_n^{i-1}(t), \boldsymbol{\xi}_n^{i-1}(t))\right\|_2^2\Big] \\ &\overset{(b)}{\leq} \sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t)-\boldsymbol{w}(t)\right\|_2^2\right] + \eta^2(t)\tau G^2,\end{aligned} \tag{56}$$

where (a) follows from the Cauchy-Schwarz inequality in combination with the inequality

$$2\left(\frac{x}{\sqrt{\eta(t)}}\right)(\sqrt{\eta(t)}y) \leq \frac{x^2}{\eta(t)} + \eta(t)y^2, \tag{57}$$

while (b) follows from Assumption 3. Next, we bound B_2 from (55) as:

$$\begin{aligned}B_2 &\leq 2\eta(t)\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\left[\langle \boldsymbol{w}^*-\boldsymbol{w}_n^{i-1}(t), \nabla F_n(\boldsymbol{w}_n^{i-1}(t))\rangle\right] \\ &\overset{(a)}{\leq} 2\eta(t)\sum_{n=1}^K\frac{1}{K}\sum_{i=1}^{\tau}\mathbb{E}\Big[F_n(\boldsymbol{w}^*) - F_n(\boldsymbol{w}_n^{i-1}(t)) \\ &\quad- \frac{\mu}{2}\left\|\boldsymbol{w}_n^{i-1}(t)-\boldsymbol{w}^*\right\|_2^2\Big],\end{aligned} \tag{58}$$

where (a) follows from the μ-strong convexity of F_n, $\forall n \in \mathcal{N}$. By combining (53) and (55) we conclude to

$$\begin{aligned}
&A_1 + A_2 \\
&\leq \eta^2(t)\tau^2 \sum_{n=1}^{K} \frac{\sigma_n^2}{K} + \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right] \\
&\quad + \eta^2(t)\tau G^2 - 2\eta(t) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[\frac{\mu}{2}\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}^*\right\|_2^2\right] \\
&\quad + 2L\eta^2(t)\tau \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F_n^*\right] \\
&\quad - 2\eta(t) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F_n(\boldsymbol{w}^*)\right].
\end{aligned} \tag{59}$$

Next, we bound the last two terms in (59), which we denote as C. That gives

$$\begin{aligned}
C &= 2L\eta^2(t)\tau \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F_n^*\right] \\
&\quad - 2\eta(t) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F_n(\boldsymbol{w}^*)\right] \\
&= -2\eta(t)(1 - L\eta(t)\tau) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F_n^*\right] \\
&\quad + 2\eta(t) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}^*) - F_n^*\right].
\end{aligned} \tag{60}$$

We can now write C as

$$\begin{aligned}
C &= -2\eta(t)(1 - L\eta(t)\tau) \\
&\quad \times \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F(\boldsymbol{w}^*)\right] \\
&\quad + (2\eta(t) - 2\eta(t)(1 - L\eta(t)\tau)) \\
&\quad \times \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F(\boldsymbol{w}^*) - F_n^*\right] \\
&\leq -2\eta(t)(1 - L\eta(t)\tau) \\
&\quad \times \underbrace{\sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F(\boldsymbol{w}^*)\right]}_{D} \\
&\quad + 2L\eta^2(t)\tau^2\Gamma.
\end{aligned} \tag{61}$$

To bound D from (61), we write

$$\begin{aligned}
D &\overset{(a)}{=} \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}_n^{i-1}(t)) - F(\boldsymbol{w}(t))\right] \\
&\quad + \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[F_n(\boldsymbol{w}(t)) - F(\boldsymbol{w}^*)\right] \\
&\geq \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[\langle \nabla F_n(\boldsymbol{w}(t)), \boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\rangle\right] \\
&\quad + \tau\mathbb{E}\left[(F(\boldsymbol{w}(t)) - F(\boldsymbol{w}^*))\right] \\
&\overset{(b)}{\geq} -\frac{1}{2} \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\Big[\left\|\eta(t)\nabla F_n(\boldsymbol{w}(t))\right\|_2^2 \\
&\quad + \frac{1}{\eta(t)}\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\Big] + \tau\mathbb{E}\left[F(\boldsymbol{w}(t)) - F(\boldsymbol{w}^*)\right] \\
&\overset{(c)}{\geq} -\sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\Big[\eta(t)L(F_n(\boldsymbol{w}(t)) - F(\boldsymbol{w}^*)) \\
&\quad + \frac{1}{2\eta(t)}\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\Big] + \tau\mathbb{E}\left[F(\boldsymbol{w}(t)) - F(\boldsymbol{w}^*)\right],
\end{aligned} \tag{62}$$

where in (a) we have used that $\sum_{n=1}^{K} \frac{1}{K} F_n(\boldsymbol{w}(t)) = F(\boldsymbol{w}(t))$, (b) follows from Cauchy-Schwarz inequality and (c) from (54). Next, by plugging D in C we get

$$\begin{aligned}
C &\leq 2\eta(t)(1 - L\eta(t)\tau) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\Big[\eta(t)L(F_n(\boldsymbol{w}(t)) \\
&\quad - F(\boldsymbol{w}^*)) + \frac{1}{2\eta(t)}\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\Big] + 2L\eta^2(t)\tau^2\Gamma \\
&\quad - 2\eta(t)(1 - L\eta(t)\tau)\mathbb{E}\left[\tau(F(\boldsymbol{w}(t)) - F(\boldsymbol{w}^*))\right] \\
&\overset{(a)}{\leq} 2L\eta^2(t)\tau^2\Gamma + \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right],
\end{aligned} \tag{63}$$

where (a) holds for $\eta(t) \leq \frac{1}{\tau L}$, since $F(\boldsymbol{w}(t)) - F(\boldsymbol{w}^*) \geq 0$, $\forall t$.

By plugging (53) and (55) in (59), yields

$$\begin{aligned}
A_1 + A_2 &\leq \eta^2(t)\tau^2 \sum_{n=1}^{K} \tfrac{\sigma_n^2}{K} + \eta^2(t)\tau G^2 + 2L\eta^2(t)\tau^2\Gamma \\
&\quad + 2\sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right] \\
&\quad - \mu\eta(t) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=1}^{\tau} \mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}^*\right\|_2^2\right] \\
&\overset{(a)}{\leq} \eta^2(t)\tau^2 \sum_{n=1}^{K} \sigma_n^2 + \eta^2(t)\tau G^2 + 2L\eta^2(t)\tau^2\Gamma \\
&\quad + 2\sum_{n=1}^{K} \frac{1}{K} \sum_{i=2}^{\tau} \mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right] \\
&\quad - \mu\eta(t) \sum_{n=1}^{K} \frac{1}{K} \sum_{i=2}^{\tau} \mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}^*\right\|_2^2\right] \\
&\quad - \mu\eta(t)\mathbb{E}\left[\left\|\boldsymbol{w}(t) - \boldsymbol{w}^*\right\|_2^2\right],
\end{aligned} \tag{64}$$

and now we have

$$\begin{aligned}
&A_1 + A_2 \\
&\overset{(b)}{\leq} \eta^2(t)\tau^2 \sum_{n=1}^{K} \frac{\sigma_n^2}{K} + \eta^2(t)\tau G^2 + 2L\eta^2(t)\tau^2\Gamma \\
&\quad + 2\sum_{n=1}^{K}\frac{1}{K}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right] \\
&\quad + \mu\eta(t)(1-\eta(t))\Bigg(- (\tau-1)\mathbb{E}\left[\|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2\right] \\
&\quad + \frac{1}{\eta(t)}\sum_{n=1}^{K}\frac{1}{K}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right]\Bigg) \\
&\quad - \mu\eta(t)\mathbb{E}\left[\|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2\right],
\end{aligned} \tag{65}$$

where in (a) we used that $\boldsymbol{w}_n^0(t) \triangleq \boldsymbol{w}(t)$, while (b) follows from the fact that:

$$\begin{aligned}
&-\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}^*\right\|_2^2 \\
&\quad = -\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2 - \|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2 \\
&\qquad - 2\langle \boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t), \boldsymbol{w}(t) - \boldsymbol{w}^*\rangle \\
&\quad \overset{(c)}{\leq} -\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2 - \|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2 \\
&\qquad + \frac{1}{\eta(t)}\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2 + \eta(t)\left\|\boldsymbol{w}(t) - \boldsymbol{w}^*\right\|_2^2 \\
&\quad = \left(\frac{1}{\eta(t)} - 1\right)\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2 \\
&\qquad - (1-\eta(t))\left\|\boldsymbol{w}(t) - \boldsymbol{w}^*\right\|_2^2,
\end{aligned} \tag{66}$$

where (c) follows from Cauchy-Schwarz inequality, combined with the inequality in (57). By further expanding (65), we get

$$\begin{aligned}
&A_1 + A_2 \\
&\leq \eta^2(t)\tau^2 \sum_{n=1}^{K} \frac{\sigma_n^2}{K} + \eta^2(t)\tau G^2 + 2L\eta^2(t)\tau^2\Gamma \\
&\quad - \mu\eta(t)(\tau - \eta(t)(\tau-1))\mathbb{E}\left[\|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2\right] \\
&\quad + (2 + \mu - \mu\eta(t))\sum_{n=1}^{K}\frac{1}{K}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right], \\
&\overset{(a)}{\leq} \eta^2(t)\tau^2 \sum_{n=1}^{K} \frac{\sigma_n^2}{K} + \eta^2(t)\tau G^2 + 2L\eta^2(t)\tau^2\Gamma \\
&\quad - \mu\eta(t)\mathbb{E}\left[\|\boldsymbol{w}(t) - \boldsymbol{w}^*\|_2^2\right] \\
&\quad + (2 + \mu)\sum_{n=1}^{K}\frac{1}{K}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right],
\end{aligned} \tag{67}$$

where in (a) we have used that $0 < \eta(t) \leq 1$, which also implies that $\tau - \eta(t)(\tau-1) \geq 1$. Finally, the last term in (67), can be bounded as follows:

$$\begin{aligned}
&\sum_{n=1}^{K}\frac{1}{K}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{w}_n^{i-1}(t) - \boldsymbol{w}(t)\right\|_2^2\right] \\
&\quad = \eta^2(t)\sum_{n=1}^{K}\frac{1}{K}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\sum_{j=1}^{i}\nabla F_n(\boldsymbol{w}_n^{j-1}(t), \boldsymbol{\xi}_n^{j-1}(t))\right\|_2^2\right] \\
&\quad \overset{(a)}{\leq} \eta^2(t)\sum_{i=2}^{\tau} i^2 G^2 = \eta^2(t)\frac{\tau(\tau-1)(2\tau-1)}{6}G^2,
\end{aligned} \tag{68}$$

where (a) follows from the convexity of $\|\cdot\|_2^2$ and Assumption 3.

According to [33, Lemma 5], for the second term in the RHS of (51) it holds

$$\mathbb{E}\left[\|\boldsymbol{v}(t+1) - \boldsymbol{z}(t+1)\|_2^2\right] \leq \frac{K-N}{N(K-1)}\eta^2(t)\tau^2 G^2. \tag{69}$$

Finally the third term in the RHS of (51) vanishes, since $\mathbb{E}\left[\boldsymbol{v}(t)\right] = \boldsymbol{z}(t)$, where the expectation is taken with respect to the randomness in user selection [33]. By substituting (68) in (67) and also using (69), the proof of Lemma 3 is completed.

Appendix C
LMs Update

The LMs can be updated as follows:

$$\begin{aligned}
\lambda_{1,n}^{(k+1)} = \Bigg[\lambda_{1,n}^{(k)} + a^{(k)}\Bigg(&d(B_n^{(k)}+1) + m \\
&- l_n^{\mathrm{up}(k)} W \log_2\left(1 + \tfrac{g_n E_n^{(k)}}{l_n^{\mathrm{up}(k)} W N_0}\right)\Bigg)\Bigg]^+,
\end{aligned} \tag{70}$$

$$\lambda_{2,n}^{(k+1)} = \left[\lambda_{2,n}^{(k)} + a^{(k)}\left(\frac{\zeta\tau^3 c_n^3 D_n^3}{l^{\mathrm{c}(k)^2}} + E_n^{(k)} - E_n^{\max}\right)\right]^+, \tag{71}$$

$$\lambda_3^{(k+1)} = \left[\lambda_3^{(k)} + a^{(k)}\left(\sum_{n\in\mathcal{N}}\frac{1}{N}\frac{\delta_n^2}{\left(2^{B_n^{(k)}} - 1\right)^2} - \epsilon\right)\right]^+, \tag{72}$$

$$\lambda_4^{(k+1)} = \left[\lambda_4^{(k)} + a^{(k)}\left(a_1 - l^{\mathrm{c}(k)}\right)\right]^+, \tag{73}$$

$$\lambda_{5,n}^{(k+1)} = \left[\lambda_{5,n}^{(k)} + a^{(k)}\left(1 - B_n^{(k)}\right)\right]^+, \tag{74}$$

where k denotes the iteration index, $[\cdot]^+ = \min(\cdot, 0)$ and $a^{(k)}$ is a positive diminishing step size, a selection which guarantees the convergence of the subgradient method [36].

References

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[2] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, 2019.

[3] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[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. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless federated learning (WFL) for 6G networks - Part I: Research challenges and future trends," *IEEE Commun. Lett.*, pp. 1–1, 2021.

[7] ——, "Wireless federated learning (WFL) for 6G networks - Part II: The compute-then-transmit NOMA paradigm," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 8–12, 2022.

[8] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 2020.

[9] 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, 2020.

[10] 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, 2020.

[11] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, 2020.

[12] S. Wan, J. Lu, P. Fan, Y. Shao, C. Peng, and K. B. Letaief, "Convergence analysis and system design for federated learning over wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3622–3639, 2021.

[13] 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, 2020.

[14] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, 2019.

[15] J. S. Ng, W. Y. B. Lim, Z. Xiong, X. Cao, D. Niyato, C. Leung, and D. I. Kim, "A hierarchical incentive design toward motivating participation in coded federated learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 359–375, 2021.

[16] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[17] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. 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.

[18] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *arXiv preprint arXiv:1812.07210*, 2018.

[19] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE J. Sel. Areas Commun.*, 2020.

[20] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," *arXiv preprint arXiv:2006.10672*, 2020.

[21] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Uveqfed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, 2020.

[22] W.-T. Chang and R. Tandon, "Communication efficient federated learning over multiple access channels," *arXiv preprint arXiv:2001.08737*, 2020.

[23] Y. Wang, Y. Xu, Q. Shi, and T.-H. Chang, "Quantized federated learning under transmission delay and outage constraints," *IEEE J. on Sel. Areas in Commun.*, vol. 40, no. 1, pp. 323–341, 2022.

[24] 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, 2020.

[25] S. Chen, C. Shen, L. Zhang, and Y. Tang, "Dynamic aggregation for heterogeneous quantization in federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6804–6819, 2021.

[26] A. Elgabli, J. Park, A. S. Bedi, C. B. Issaid, M. Bennis, and V. Aggarwal, "Q-gadmm: Quantized group admm for communication efficient decentralized machine learning," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 164–181, 2020.

[27] M. Kim, W. Saad, M. Mozaffari, and M. Debbah, "On the tradeoff between energy, precision, and accuracy in federated quantized neural networks," in *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 2194–2199.

[28] P. Liu, J. Jiang, G. Zhu, L. Cheng, W. Jiang, W. Luo, Y. Du, and Z. Wang, "Training time minimization for federated edge learning with optimized gradient quantization and bandwidth allocation," *Frontiers of Information Technology & Electronic Engineering*, vol. 23, no. 8, pp. 1247–1263, 2022.

[29] C. Shen, J. Xu, S. Zheng, and X. Chen, "Resource rationing for wireless federated learning: Concept, benefits, and challenges," *IEEE Commun. Mag.*, vol. 59, no. 5, pp. 82–87, 2021.

[30] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, 2016.

[31] N. H. Tran, W. Bao, A. Zomaya, M. N. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1387–1395.

[32] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3643–3658, 2021.

[33] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[34] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[35] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.

[36] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *lecture notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, pp. 2004–2005, 2003.

[37] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

Pavlos S. Bouzinis received the Diploma Degree (5 years) in Electrical and Computer Engineering from the Aristotle University of Thessaloniki (AUTH), Greece, in 2019, where he is currently pursuing his PhD with the Department of Electrical and Computer Engineering. Also, he is a member of the Wireless Communications & Information Processing (WCIP) Group. His current research interests include resource allocation in wireless networks, optimization theory, mobile edge computing, and federated learning.

Panagiotis D. Diamantoulakis (Senior Member, IEEE) received the Diploma (five years) and PhD from the Department of Electrical and Computer Engineering (ECE), Aristotle University of Thessaloniki (AUTH), Greece, in 2012 and 2017, respectively. Since 2017, he works as a Post-doctoral Fellow in Wireless Communications & Information Processing (WCIP) Group at AUTH and since 2021, he is a visiting Assistant Professor in the Key Lab of Information Coding and Transmission at Southwest Jiaotong University (SWJTU), China. His research interests include optimization theory and applications in wireless networks and smart grids, game theory, and optical wireless communications. He serves as an Editor for IEEE Wireless Communications Letters, IEEE Open Journal of the Communications Society, Physical Communications (Elsevier), and Frontiers in Communications and Networks.

George K. Karagiannidis (M'96-SM'03-F'14) is currently Professor in the Electrical & Computer Engineering Dept. of Aristotle University of Thessaloniki, Greece and Head of Wireless Communications & Information Processing (WCIP) Group. He is also Faculty Fellow in the Cyber Security Systems and Applied AI Research Center, Lebanese American University. His research interests are in the areas of Wireless Communications Systems and Networks, Signal processing, Optical Wireless Communications, Wireless Power Transfer and Applications and Communications & Signal Processing for Biomedical Engineering.

Dr. Karagiannidis was in the past Editor in several IEEE journals and from 2012 to 2015 he was the Editor-in Chief of IEEE Communications Letters. From September 2018 to June 2022 he served as Associate Editor-in Chief of IEEE Open Journal of Communications Society. Currently, he is in the Steering Committee of IEEE Transactions on Cognitive Communications & Networks.

Recently, he received three prestigious awards: The 2021 IEEE ComSoc RCC Technical Recognition Award, the 2018 IEEE ComSoc SPCE Technical Recognition Award and the 2022 Humboldt Research Award from Alexander von Humboldt Foundation.

Dr. Karagiannidis is one of the highly-cited authors across all areas of Electrical Engineering, recognized from Clarivate Analytics as Web-of-Science Highly-Cited Researcher in the eight consecutive years 2015-2022.