This is the peer reviewd version of the followng article:

Hierarchical Over-the-Air Federated Edge Learning / Aygun, O., Kazemi, M., Gunduz, D., Duman, T.M.. - 2022-:(2022), pp. 3376-3381. (2022 IEEE International Conference on Communications, ICC 2022 COEX, kor 2022) [10.1109/ICC45855.2022.9839230].

Institute of Electrical and Electronics Engineers Inc.
Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

24/10/2026 05:36

(Article begins on next page)

Hierarchical Over-the-Air Federated Edge Learning

Ozan Aygün[1], Mohammad Kazemi[1], Deniz Gündüz[2] and Tolga M. Duman[1]
[1]*Dept. of Electrical and Electronics Engineering, Bilkent University*, Ankara, Turkey
[2]*Dept. of Electrical and Electronic Engineering, Imperial College London*, London, UK
{ozan, kazemi, duman}@ee.bilkent.edu.tr, d.gunduz@imperial.ac.uk

arXiv:2112.11167v1 [cs.LG] 21 Dec 2021

***Abstract*—Federated learning (FL) over wireless communication channels, specifically, over-the-air (OTA) model aggregation framework is considered. In OTA wireless setups, the adverse channel effects can be alleviated by increasing the number of receive antennas at the parameter server (PS), which performs model aggregation. However, the performance of OTA FL is limited by the presence of mobile users (MUs) located far away from the PS. In this paper, to mitigate this limitation, we propose hierarchical over-the-air federated learning (HOTAFL), which utilizes intermediary servers (IS) to form clusters near MUs. We provide a convergence analysis for the proposed setup, and demonstrate through theoretical and experimental results that local aggregation in each cluster before global aggregation leads to a better performance and faster convergence than OTA FL.**

***Index Terms*—machine learning, over-the-air communication, clustering, hierarchical federated learning.**

I. Introduction

Extensive amounts of collected data from various sources such as mobile phones and Internet-of-things (IoT) sensors have enabled the accelerating rise of machine learning (ML) algorithms, aiming to assemble all the data in a cloud server to obtain representative datasets for model training. This, however, brings out growing concerns regarding the privacy, cost, and latency of traditional ML algorithms. Firstly, data owners have become more sensitive about sharing their data; secondly, the increasing quality of data results in higher communication costs; and finally, solutions that work in real-time are faced with latency issues [1]. To overcome these problems, a decentralized approach called *federated learning* (FL) has been introduced, where the transmission of data is not required since models are trained locally instead of using a centralized server for training [2].

In FL, several data owners called mobile users (MUs) are selected based on some criteria such as their computing capability, data quality, available power, and location [3]. Each MU in the federation trains a local model using its own data and computing power in every iteration. After each MU completes its local stochastic gradient descent (SGD) computation, only the weight updates are sent to a parameter server (PS) that performs model aggregation and sends back the updated global model to MUs for the next iteration.

Despite its superiority over traditional ML, adverse channel effects in wireless setups and increased communication costs have arisen some concerns about the feasibility of conventional FL in practical scenarios. To address the communication cost concerns, over-the-air (OTA) aggregation [4] has become a popular method in wireless schemes thanks to its efficient strategy that allocates all the users to the same bandwidth, thereby handling the transmission and aggregation of the gradient updates simultaneously (over the air). For this framework, one approach to deal with the channel effects (particularly when there is no transmit side channel state information) is to increase the number of receive antennas at the PS [5]. Nevertheless, the disparity among the channel gains is still a critical factor when some MUs are far away from the PS.

Recent developments on FL include device selection algorithms [6], efficient communication schemes [4], [7]–[11], heterogeneity of data [12], and power and latency analysis [13], [14]. Although *Federated Averaging* [2] is the most common way to perform global aggregation in error-free setups, OTA communication has been preferred for wireless FL [5], [12], [15]. Furthermore, hierarchical federated learning (HFL) has been gaining increasing attention, where the objective is to utilize intermediate servers (IS) to form clusters to reduce communication costs. There exist studies on HFL on latency and power analysis [16], [17], resource allocation [18], [19], and performance analysis for non-independent and identically distributed (i.i.d.) data [20]. However, there is no work on HFL with OTA taking into account practical wireless channel models, which motivates this work.

In order to make distant MUs more resilient to the channel effects, we propose *hierarchical over-the-air federated learning* (HOTAFL), where MUs communicate with their corresponding ISs through wireless links. In this setup, each MU shares its local training result with its corresponding IS through OTA (cluster) aggregation. After several local iterations with the MUs in their clusters, the ISs send the results to the PS to complete the global aggregation, which constitutes one global iteration. We examine the performance of HOTAFL and compare the results with those of the conventional FL and error-free HFL both through analytical results and numerical experiments. The results show that the proposed framework outperforms conventional OTA FL and leads to a better model accuracy and faster convergence.

The paper is organized as follows: in Sections II and III, we introduce the HOTAFL framework as well as its communication model. In Section IV, we provide a convergence analysis of HOTAFL under convexity assumptions on the loss functions. We present our numerical results in Section V, and

Ozan Aygün's research in this study is supported by Turkcell A.S. within the framework of 5G and Beyond Joint Graduate Support Programme coordinated by Information and Communication Technologies Authority.

D. Gunduz acknowledges support from UK EPSRC through CHIST-ERA project CONNECT (CHISTERA-18-SDCDN-001, EPSRC-EP/T023600/1).

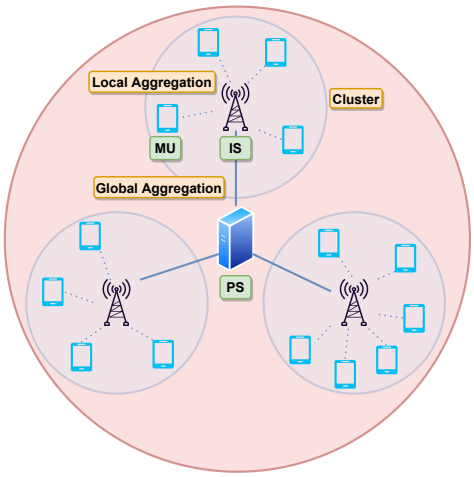

Fig. 1: HOTAFL system model.

conclude the paper in Section VI.

II. System Model

The objective of HOTAFL is to minimize a loss function $F(\boldsymbol{\theta})$ with respect to the model weight vector $\boldsymbol{\theta} \in \mathbb{R}^{2N}$, where $2N$ is the model dimension. Our system consists of C clusters each containing an IS and M MUs as depicted in Fig. 1. The dataset of the m-th MU in the c-th cluster is denoted as $\mathcal{B}_{m,c}$, and we define $B \triangleq \sum_{c=1}^{C}\sum_{m=1}^{M}|\mathcal{B}_{m,c}|$. We have

$$F(\boldsymbol{\theta}) = \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{|\mathcal{B}_{m,c}|}{B}F_{m,c}(\boldsymbol{\theta}), \tag{1}$$

where $F_{m,c}(\boldsymbol{\theta}) \triangleq \frac{1}{|\mathcal{B}_{m,c}|}\sum_{u\in\mathcal{B}_{m,c}} f(\boldsymbol{\theta}, u)$, with $f(\boldsymbol{\theta}, u)$ denoting the corresponding loss of u-th data sample.

We consider a hierarchical and iterative approach to minimize (1) consisting of global, local, and user iterations. In every cluster iteration, the MUs carry out τ user iterations on their own, then send their model updates to their corresponding ISs for local iteration. I local iterations are performed at the IS in every cluster before all the local models are forwarded to the PS for global aggregation. At the j-th user iteration of the i-th local iteration, the weight update is performed employing stochastic gradient descent (SGD) for the m-th user in the c-th cluster as follows

$$\boldsymbol{\theta}_{m,c}^{i,j+1}(t) = \boldsymbol{\theta}_{m,c}^{i,j}(t) - \eta_{m,c}^{i,j}(t)\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t)), \tag{2}$$

where $\eta_{m,c}^{i,j}(t)$ is the learning rate, $\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t))$ denotes the stochastic gradient estimate for the weight vector $\boldsymbol{\theta}_{m,c}^{i,j}(t)$ and a randomly sampled batch of data samples $\xi_{m,c}^{i,j}(t)$ from the dataset of the m-th user in the c-th cluster at the t-th global, i-th local and j-th user iteration. Initially, $\boldsymbol{\theta}_{m,c}^{1,1}(t) = \boldsymbol{\theta}_{IS,c}^{i}(t), \forall i \in [I]$ where $[I] \triangleq \{1, 2, \ldots, I\}$, and $\boldsymbol{\theta}_{IS,c}^{1}(t) = \boldsymbol{\theta}_{PS}(t)$, where $\boldsymbol{\theta}_{PS}(t)$ is the global model at the PS at the t-th global iteration and $\boldsymbol{\theta}_{IS,c}^{i}(t)$ denotes the local model of IS in the c-th cluster at the i-th local iteration. The purpose of employing ISs is to accumulate the local model differences within each cluster more frequently in smaller areas before obtaining the global model $\boldsymbol{\theta}_{PS}(t)$ for the next iteration. Also, note that $\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t))$ is an unbiased estimator of $\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t))$, i.e., $\mathbb{E}_{\xi}\left[\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t))\right] = \nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t))$, where the expectation is over the randomness due to the SGD.

III. Hierarchical Over-the-Air FL (HOTAFL)

A. Ideal Communication

We refer to the case where all the communication among all the units is error-free as the ideal communication scenario. In this case, after performing SGD, each MU calculates its model difference to be sent to its corresponding IS as

$$\Delta\boldsymbol{\theta}_{m,c}^{i}(t) = \boldsymbol{\theta}_{m,c}^{i,\tau+1}(t) - \boldsymbol{\theta}_{IS,c}^{i}(t). \tag{3}$$

Then, the local aggregation at the c-th cluster is performed as

$$\Delta\boldsymbol{\theta}_{IS,c}^{i}(t) = \frac{1}{M}\sum_{m=1}^{M}\Delta\boldsymbol{\theta}_{m,c}^{i}(t), \tag{4}$$

$$\boldsymbol{\theta}_{IS,c}^{i+1}(t) = \boldsymbol{\theta}_{IS,c}^{i}(t) + \Delta\boldsymbol{\theta}_{IS,c}^{i}(t). \tag{5}$$

After completing I local iterations in each cluster, ISs send their model updates to the PS, which can be written as

$$\Delta\boldsymbol{\theta}_{PS,c}(t) = \boldsymbol{\theta}_{IS,c}^{I+1}(t) - \boldsymbol{\theta}_{PS}(t). \tag{6}$$

The global update rule is $\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{C}\sum_{c=1}^{C}\Delta\boldsymbol{\theta}_{PS,c}(t)$. Using recursion, we can conclude that

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{MC}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{m=1}^{M}\Delta\boldsymbol{\theta}_{m,c}^{i}(t). \tag{7}$$

After the global aggregation, the model at the PS is updated as $\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t)$.

B. OTA Communication

We now consider the scheme referred as OTA communications, for which the links between the users and the ISs are wireless with OTA aggregation, however, the links between ISs and the PS is assumed to be error-free. Since a common wireless medium is used in local aggregations, noisy versions of the model updates $\Delta\boldsymbol{\theta}_{IS,c}(t)$ are received at the ISs. In our setup, the ISs are equipped with K antennas, and we assume perfect channel state information (CSI) at the receivers. For the k-th antenna, the received signal at the c-th IS can be written as[1]

$$\boldsymbol{y}_{IS,c,k}^{i}(t) = \sum_{m=1}^{M}\boldsymbol{h}_{m,c,k}^{i}(t)\circ\boldsymbol{x}_{m,c,k}^{i}(t) + \boldsymbol{z}_{IS,c,k}^{i}(t), \tag{8}$$

where $\circ$ denotes the element-wise product, $\boldsymbol{x}_{m,c,k}^{i}(t) \in \mathbb{C}^N$, $\boldsymbol{z}_{IS,c,k}^{i}(t) \in \mathbb{C}^N$ with independent and identically distributed (i.i.d.) entries $z_{IS,c,k}^{i,n}(t) \sim \mathcal{CN}(0, \sigma_z^2)$. The channel coefficients are modelled as $\boldsymbol{h}_{m,c,k}^{i}(t) = \sqrt{\beta_{m,c}}\,\boldsymbol{g}_{m,c,k}^{i}(t)$, where $\boldsymbol{g}_{m,c,k}(t) \in \mathbb{C}^N$ with entries $g_{m,c,k}^{i,n}(t) \sim \mathcal{CN}(0, \sigma_h^2)$ (i.e., Rayleigh fading), $\beta_{m,c}$ is the large-scale fading coefficient modeled as $\beta_{m,c} = (d_{m,c})^{-p}$, where p represents the path loss exponent, and $d_{m,c}$ denotes the distance between the m-th user in the c-th cluster and the IS in that cluster.

[1]Note that the setup here can be efficiently implemented in practice using orthogonal frequency-division multiplexing (OFDM).

1) Local Aggregation: In OTA communication, in order to increase the spectral efficiency, the model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}_{m,c}^{i,cx} \in \mathbb{C}^N$ with the following real and imaginary parts

$$\Delta\boldsymbol{\theta}_{m,c}^{i,re}(t) \triangleq \left[\Delta\theta_{m,c}^{i,1}(t), \Delta\theta_{m,c}^{i,2}(t), \ldots, \Delta\theta_{m,c}^{i,N}(t)\right]^T, \tag{9a}$$

$$\Delta\boldsymbol{\theta}_{m,c}^{i,im}(t) \triangleq \left[\Delta\theta_{m,c}^{i,N+1}(t), \Delta\theta_{m,c}^{i,N+2}(t), \ldots, \Delta\theta_{m,c}^{i,2N}(t)\right]^T. \tag{9b}$$

Under the assumption that there is no inter-cluster interference, the received signal for the k-th antenna in the c-th cluster at the i-th local iteration can be represented as

$$\boldsymbol{y}_{IS,c,k}^{i}(t) = P_t \sum_{m=1}^{M} \boldsymbol{h}_{m,c,k}^{i}(t) \circ \Delta\boldsymbol{\theta}_{m,c}^{i,cx}(t) + \boldsymbol{z}_{IS,c,k}^{i}(t), \tag{10}$$

where P_t is the transmit power constant at the t-th global iteration. Knowing the CSI perfectly, the c-th IS combines the received signals as $\boldsymbol{y}_{IS,c}^{i}(t) = \frac{1}{K}\sum_{k=1}^{K}\left(\sum_{m=1}^{M}\boldsymbol{h}_{m,c,k}^{i}(t)\right)^* \circ \boldsymbol{y}_{IS,c,k}^{i}(t)$. For the n-th symbol, it can be written as

$$\begin{aligned} y_{IS,c}^{i,n}(t) = & \underbrace{P_t \sum_{m=1}^{M}\Big(\frac{1}{K}\sum_{k=1}^{K}|h_{m,c,k}^{i,n}(t)|^2\Big)\Delta\theta_{m,c}^{i,n,cx}(t)}_{y_{IS,c}^{i,n,sig}(t) \text{ (signal term)}} \\ & + \underbrace{\frac{P_t}{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\ m'\neq m}}^{M}\sum_{k=1}^{K}(h_{m,c,k}^{i,n}(t))^* h_{m',c,k}^{i,n}(t)\Delta\theta_{m',c}^{i,n,cx}(t)}_{y_{IS,c}^{i,n,itf}(t) \text{ (interference term)}} \\ & + \underbrace{\frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K}(h_{m,c,k}^{i,n}(t))^* z_{c,k}^{i,n}(t)}_{y_{IS,c}^{i,n,no}(t) \text{ (noise term)}}. \end{aligned} \tag{11}$$

Aggregated model differences can be recovered by

$$\Delta\hat{\theta}_{IS,c}^{i,n}(t) = \frac{1}{P_t M \sigma_h^2 \bar{\beta}_c}\,\mathrm{Re}\{y_{IS,c}^{i,n}(t)\}, \tag{12a}$$

$$\Delta\hat{\theta}_{IS,c}^{i,n+N}(t) = \frac{1}{P_t M \sigma_h^2 \bar{\beta}_c}\,\mathrm{Im}\{y_{IS,c}^{i,n}(t)\}, \tag{12b}$$

where $\bar{\beta}_c = \sum_{m=1}^{M}\beta_{m,c}$. After estimating the model difference values, the cluster model update is written as

$$\boldsymbol{\theta}_{IS,c}^{i+1}(t) = \boldsymbol{\theta}_{IS,c}^{i}(t) + \Delta\hat{\boldsymbol{\theta}}_{IS,c}^{i}(t), \tag{13}$$

where $\Delta\hat{\boldsymbol{\theta}}_{IS,c}^{i}(t) = \left[\Delta\hat{\theta}_{IS,c}^{i,1}(t)\ \Delta\hat{\theta}_{IS,c}^{i,2}(t)\ \cdots\ \Delta\hat{\theta}_{IS,c}^{i,2N}(t)\right]^T$.

2) Global Aggregation: This part is similar to the case of ideal communication. The only difference is that the aggregated signals are estimates of the actual model differences. Letting $\boldsymbol{x}_{PS,c}(t)$ be the transmitted signal from the c-th IS, its n-th symbol can be written as

$$x_{PS,c}^{n}(t) = \Delta\theta_{PS,c}^{n}(t) + j\Delta\theta_{PS,c}^{n+N}(t). \tag{14}$$

Then, using (6), (11), (14) and recursion, the received signal for $1 \leq n \leq N$ (similarly for $N+1 \leq n \leq 2N$) can be written as

$$y_{PS}^{n}(t) = \sum_{c=1}^{C} x_{PS,c}^{n}(t) \tag{15}$$

$$\begin{aligned} = & \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y_{IS,c}^{i,n,sig}(t)\right\}}{P_t M \sigma_h^2}}_{y_{PS,1}^{n}(t)} + \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y_{IS,c}^{i,n,itf}(t)\right\}}{P_t M \sigma_h^2}}_{y_{PS,2}^{n}(t)} \\ & + \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y_{IS,c}^{i,n,no}(t)\right\}}{P_t M \sigma_h^2}}_{y_{PS,3}^{n}(t)}. \end{aligned} \tag{16}$$

The received signal at the PS is then recovered as $\Delta\hat{\theta}_{PS}^{n}(t) = \frac{1}{C}\,\mathrm{Re}\{y_{PS}^{n}(t)\}$, $\Delta\hat{\theta}_{PS}^{n+N}(t) = \frac{1}{C}\,\mathrm{Im}\{y_{PS}^{n}(t)\}$. Finally, the global aggregation is performed as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t), \tag{17}$$

where $\Delta\hat{\boldsymbol{\theta}}_{PS}(t) = \left[\Delta\hat{\theta}_{PS}^{1}(t)\ \Delta\hat{\theta}_{PS}^{2}(t)\ \cdots\ \Delta\hat{\theta}_{PS}^{2N}(t)\right]^T$.

IV. Convergence Analysis

In this section, we present a convergence analysis of the proposed HOTAFL algorithm. Define the optimal solution as $\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, the minimum values of the total and the local loss functions as $F^* = F(\boldsymbol{\theta}^*)$ and $F_{m,c}^*$, respectively, and the bias in the dataset as $\Gamma \triangleq F^* - \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{B_{m,c}}{B}F_{m,c}^* \geq 0$. In addition, assume that the learning rate of the overall system does not change in local iterations, i.e., $\eta_{m,c}^{i,j}(t) = \eta(t)$. Therefore, we can write the global update rule as

$$\boldsymbol{\theta}_{m,c}^{i,j+1}(t) = \boldsymbol{\theta}_{m,c}^{i,j}(t) - \eta(t)\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t)), \tag{18}$$

which can also be written as

$$\boldsymbol{\theta}_{m,c}^{i,j+1}(t) - \boldsymbol{\theta}_{m,c}^{i,1}(t) = -\eta(t)\sum_{l=1}^{j}\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,l}, \boldsymbol{\xi}_{m,c}^{i,l}(t)). \tag{19}$$

Assumption 1. *All the loss functions are L-smooth and μ-strongly convex; i.e., $\forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{2N}$, $\forall m \in [M], \forall c \in [C]$,*

$$F_{m,c}(\boldsymbol{v}) - F_{m,c}(\boldsymbol{w}) \leq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_{m,c}(\boldsymbol{w})\rangle + \frac{L}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2, \tag{20}$$

$$F_{m,c}(\boldsymbol{v}) - F_{m,c}(\boldsymbol{w}) \geq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_{m,c}(\boldsymbol{w})\rangle + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2. \tag{21}$$

Assumption 2. *The expected value of the squared l_2 norm of the stochastic gradients are bounded; i.e., $\forall j \in [\tau], i \in [I]$,*

$$\mathbb{E}_{\xi}\Big[\left\|\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t))\right\|_2^2\Big] \leq G^2, \tag{22}$$

which translates to $\forall n \in [2N]$, $\mathbb{E}_{\xi}\big[\nabla F_{m,c}(\theta_{m,c}^{i,j,n}, \xi_{m,c}^{i,j,n}(t))\big] \leq G$.

Theorem 1. *In HOTAFL, for $0 \leq \eta(t) \leq min\{1, \frac{1}{\mu\tau I}\}$, the global loss function can be upper bounded as*

$$\begin{aligned} & \mathbb{E}\big[\,\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2\,\big] \\ & \leq \Big(\prod_{a=1}^{t-1}X(a)\Big)\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{b=1}^{t-1}Y(b)\prod_{a=b+1}^{t-1}X(a), \end{aligned} \tag{23}$$

where $X(a) = (1 - \mu\eta(a)I\,(\tau - \eta(a)(\tau - 1)))$ and

$$Y(a)=\frac{\eta^2(a)\tau^2G^2I}{M^2C^2}\sum_{m_1=1}^{M}\sum_{c_1=1}^{C}\Big(\frac{\beta_{m_1,c_1}^2}{K\bar{\beta}_{c_1}^2}+\Big(\sum_{m_2=1}^{M}\sum_{c_2=1}^{C}A_1I\Big)\Big)$$
$$+\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\frac{\eta^2(a)\tau^2G^2I\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}$$
$$+\frac{\sigma_z^2IN}{P_a^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}$$
$$+(1+\mu(1-\eta(a))\,\eta^2(a)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}$$
$$+\eta^2(a)I(\tau^2+\tau-1)G^2+2\eta(a)I(\tau-1)\Gamma, \quad (24)$$

with $A_1=1-\frac{\beta_{m_1,c_1}}{\bar{\beta}_{c_1}}-\frac{\beta_{m_2,c_2}}{\bar{\beta}_{c_2}}+\frac{\beta_{m_1,c_1}\beta_{m_2,c_2}}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}}$.

Proof: Let us define auxiliary variable $\boldsymbol{v}(t+1)=\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)$. Then, we have

$$\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{\theta}^*\|_2^2=\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)+\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2$$
$$=\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\|_2^2+\|\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2$$
$$+2\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle. \quad (25)$$

Next, we provide upper bounds on the three terms of (25).

Lemma 1. *We have*

$$\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\big\|_2^2\Big]$$
$$\leq\frac{\eta^2(t)\tau^2G^2I}{M^2C^2}\sum_{m_1=1}^{M}\sum_{c_1=1}^{C}\Big(\frac{\beta_{m_1,c_1}^2}{K\bar{\beta}_{c_1}^2}+\Big(\sum_{m_2=1}^{M}\sum_{c_2=1}^{C}A_1I\Big)\Big)$$
$$+\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\frac{\eta^2(t)\tau^2G^2I\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}$$
$$+\frac{\sigma_z^2IN}{P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \quad (26)$$

Proof: See Appendix A. ■

Lemma 2. *We have*

$$\mathbb{E}\Big[\big\|v(t+1)-\boldsymbol{\theta}^*\big\|_2^2\Big]\leq(1-\mu\eta(t)I(\tau-\eta(t)(\tau-1)))\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big]$$
$$+(1+\mu(1-\eta(t))\,\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}$$
$$+\eta^2(t)I(\tau^2+\tau-1)G^2+2\eta(t)I(\tau-1)\Gamma. \quad (27)$$

Proof: The proof is similar to that of Lemma 2 in [5]. ■

Lemma 3. $\mathbb{E}\left[\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle\right]=0.$

Proof: We have $\mathbb{E}[\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle]=\mathbb{E}\Big[\langle\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t),\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\rangle\Big]$. Then, knowing that channel realizations are independent of the user and cluster updates at the same global iteration t, we have $\mathbb{E}\Big[\langle\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t),\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\rangle\Big]=0.$ ■

Recursively iterating through the results of Lemmas 1, 2, and 3 concludes the theorem. ■

Corollary 1. *Assuming L-smoothness, after T global iterations, the loss function can be upper bounded as*

$$\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T))-F^*\right]\leq\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T)-\boldsymbol{\theta}^*\|_2^2\right]$$
$$\leq\frac{L}{2}\Big(\prod_{n=1}^{T-1}X(n)\Big)\|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2+\frac{L}{2}\sum_{p=1}^{T-1}Y(p)\prod_{n=p+1}^{T-1}X(n). \quad (28)$$

Remark. Since the third term in $Y(a)$ is independent of $\eta(a)$, even for $\lim_{t\to\infty}\eta(t)=0$, we have $\lim_{t\to\infty}\mathbb{E}[F(\boldsymbol{\theta}_{PS}(t))]-F^*\neq 0$. This term is also inversely proportional to M, C, and K.

V. Simulation Results

We consider a hierarchical system with one PS and $C=4$ non-overlapping clusters, each containing one IS with $K=5MC$ receive antennas and $M=5$ MUs. Users are randomly placed in the clusters in such a way that their distance to the PS is between 0.5 and 3, while having a distance between 0.5 and 1 with their corresponding IS. We also define $\alpha=\frac{\sum_{m=1}^{M}\sum_{c=1}^{C}d_{m,c}}{\sum_{d=1}^{D}d_d}$ as a measure of relative closeness of the MUs to their corresponding IS compared to the PS, where $d_{m,c}$ is the distance between m-th user in c-th cluster to the c-th IS and d_d is the distance between the d-th user and the PS. α is set to 0.4 in the simulations. We use MNIST [21] and CIFAR-10 [22] datasets with Adam optimizer [23], and considered both i.i.d. and non-i.i.d. data distributions. In the i.i.d. case, data samples are randomly distributed among MUs, while in the non-i.i.d. case, the training data is divided into $5MC$ groups each consisting of data with only one label. Then, 5 groups are assigned to each user randomly. For CIFAR10, we use the neural network given in [5] with $2N=307498$ whereas for MNIST, we employ a one-layer neural network with 784 input and 10 output neurons with $2N=7850$.

Three scenarios are considered: baseline with error-free transmissions, FL with OTA aggregation over a wireless medium, and HOTAFL. We set the total number of global iterations T to 200, the mini-batch size to $|\boldsymbol{\xi}_{m,c}^i(t)|=500$, $\sigma_h^2=1$ and the path loss exponent p to 4. The noise variance is $\sigma_z^2=10$ for the MNIST, $\sigma_z^2=1$ for the CIFAR-10 training. Also, the power multiplier is set to $P_t=1+10^{-2}t$ for HOTAFL, $P_t=1.5+10^{-2}t$ for conventional FL, $t\in[T]$.

Accuracy plots are presented in Figs. 2-4, where $\bar{P}$ is the average transmit power. The results show that bringing the servers closer to the users enhances learning accuracy significantly. One reason for the improved performance is that the cluster structure enables the MUs share their model differences with a local server closer than the PS, reducing the adverse effects of the large-scale wireless channel effects. Another reason is that MUs receive updated models even without communicating with the PS due to local aggregations. We also observe that although more initial power is given to FL, the user updates do not reflect on the global model as much as HOTAFL does due to the effects of the wireless channel. More local iterations enables faster convergence but uses more transmit power due to the accumulating nature of IS. Even though the noise variance is high when compared to

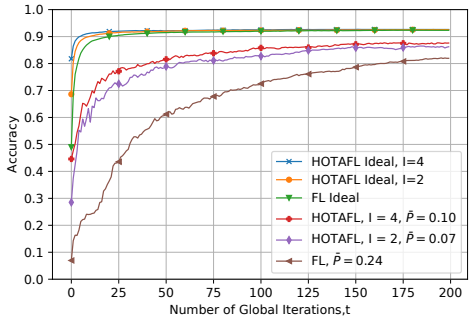

Fig. 2: Test accuracy for i.i.d. MNIST data with $\tau = 1$.

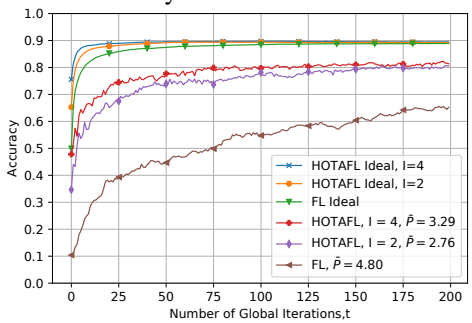

Fig. 3: Test accuracy for non-i.i.d. MNIST data with $\tau = 3$.

the transmit power, deploying $5MC = 100$ receive antennas almost mitigates the noise and the interference terms [5]. Increasing τ compensates the accuracy under more complex data structure. In Fig. 5, we compare the convergence rates of conventional FL and HOTAFL using the upper bound in (28), with $2N = 7850, L = 10, \mu = 1, G^2 = 1, \Gamma = 1, \eta(t) = 5{\cdot}10^{-2} - 2\cdot 10^{-5}t, P_t = 1 + 10^{-2}t, \beta = 3, \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$. It can be seen that the convergence rate of HOTAFL is very close to that of the ideal case, and it becomes almost the same when the number of local iterations is increased.

VI. Conclusions

In this work, we have proposed HOTAFL where ISs are employed to create clusters to bring the server-side closer to the areas where MUs are more densely located. Our framework includes OTA cluster aggregations, where the MUs send their model updates to the ISs through a wireless channel with path loss and fading. We have examined the performance and convergence rate of HOTAFL through theoretical limits as well as model training where MNIST and CIFAR-10 datasets are used with both i.i.d. and non-i.i.d. data distributions. The results show that employing a cluster-based hierarchical model outperforms the conventional FL.

Appendix A

For the n-th symbol, we have $\Delta\hat{\theta}^n_{PS}(t) = \sum_{l=1}^{3} \Delta\hat{\theta}^n_{PS,l}(t)$, so using the independence of channel coefficients, we have

$$\mathbb{E}\big[\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)\|_2^2\big] = \mathbb{E}\big[\big\|\Delta\hat{\boldsymbol{\theta}}_{PS}(t) - \Delta\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big]$$
$$= \sum_{n=1}^{2N}\big(\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t) - \Delta\theta^n_{PS}(t)\big)^2\big] + \sum_{l=2}^{3}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,l}(t)\big)^2\big]. \quad (29)$$

In the following lemmas, we will bound each of these terms.

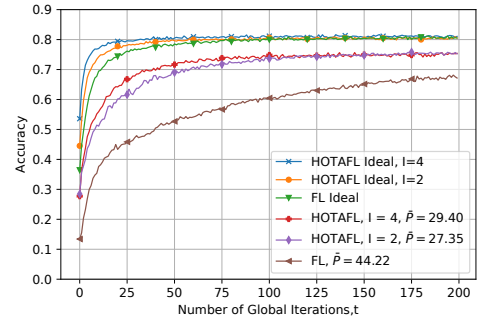

Fig. 4: Test accuracy for i.i.d. CIFAR-10 data with $\tau = 5$.

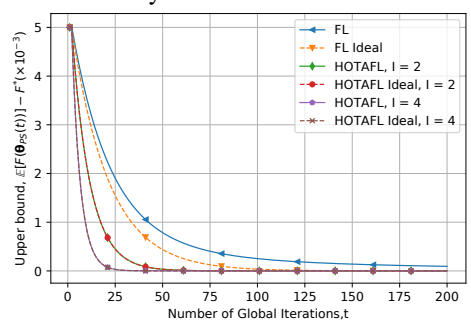

Fig. 5: Convergence rate for i.i.d. MNIST data with $\tau = 1$.

Lemma 4. *We have*

$$\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t) - \Delta\theta^n_{PS}(t)\big)^2\big]$$
$$= \frac{1}{M^2C^2}\sum_{m_1=1}^{M}\sum_{c_1=1}^{C}\sum_{i_1=1}^{I}\Big(\frac{\beta^2_{m_1,c_1}}{K\bar{\beta}^2_{c_1}}\mathbb{E}\big[\big\|\Delta\boldsymbol{\theta}^{i_1}_{m_1,c_1}(t)\big\|_2^2\big]$$
$$+\Big(\sum_{m_2=1}^{M}\sum_{c_2=1}^{C}\sum_{i_2=1}^{I}\sum_{n=1}^{2N}A_1\mathbb{E}\big[\Delta\theta^{i_1,n}_{m_1,c_1}(t)\Delta\theta^{i_2,n}_{m_2,c_2}(t)\big]\Big)\Big), \quad (30)$$

where $A_1 = 1 - \frac{\beta_{m_1,c_1}}{\bar{\beta}_{c_1}} - \frac{\beta_{m_2,c_2}}{\bar{\beta}_{c_2}} + \frac{\beta_{m_1,c_1}\beta_{m_2,c_2}}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}}$

Proof: Using (7) and (12), we have

$$\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t) - \Delta\theta^n_{PS}(t)\big)^2\big]$$
$$= \mathbb{E}\Big[\frac{1}{M^2C^2}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{i_1=1}^{I}\sum_{i_2=1}^{I}\Delta\theta^{i_1,n}_{m_1,c_1}(t)$$
$$\times\Delta\theta^{i_2,n}_{m_2,c_2}(t)\Big(1 - \frac{1}{K\sigma_h^2\bar{\beta}_{c_1}}\sum_{k_1=1}^{K}|h^{i_1,n}_{m_1,c_1,k_1}(t)|^2$$
$$-\frac{1}{K\sigma_h^2\bar{\beta}_{c_2}}\sum_{k_2=1}^{K}|h^{i_2,n}_{m_2,c_2,k_2}(t)|^2$$
$$+\frac{1}{K^2\sigma_h^4\bar{\beta}^2_{c_1}}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}|h^{i_1,n}_{m_1,c_1,k_1}(t)|^2|h^{i_2,n}_{m_2,c_2,k_2}(t)|^2\Big)\Big]. \quad (31)$$

Summing over all the symbols and using the independence of channel coefficients result in (30). ■

Lemma 5. *We have*

$$\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] = \sum_{m=1}^{M}\sum_{\substack{m'=1\\ m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}\mathbb{E}\big[\,\big\|\Delta\boldsymbol{\theta}^i_{m',c}(t)\big\|_2^2\,\big]. \quad (32)$$

Proof: For $1 \leq n \leq N$, using the independence of channel coefficients, we have

$$\begin{aligned}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] &= \mathbb{E}\Big[\Big(\sum_{m=1}^{M}\sum_{\substack{m'=1\\ m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{1}{MCK\sigma_h^2\bar{\beta}_c}\\ &\quad\times\sum_{k=1}^{K}\mathrm{Re}\big\{\big(h^{i,n}_{m,c,k}(t)\big)^* h^{i,n}_{m',c,k}(t)\Delta\theta^{i,n}_{m',c}(t)\big\}\Big)^2\Big]\\ &= \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{\substack{m'=1\\ m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\beta_{m,c}\beta_{m',c}}{2M^2C^2K\bar{\beta}_c^2}\\ &\quad\times\Big(\big(\Delta\theta^{i,n}_{m',c}(t)\big)^2+\big(\Delta\theta^{i,n+N}_{m',c}(t)\big)^2+\Delta\theta^{i,n}_{m,c}(t)\Delta\theta^{i,n}_{m',c}(t)\\ &\qquad -\Delta\theta^{i,n+N}_{m,c}(t)\Delta\theta^{i,n+N}_{m',c}(t)\Big)\Big] \quad (33)\end{aligned}$$

Obtaining the expressions for $N+1 \leq n \leq 2N$ in a similar manner and combining the two, results in (32). ■

Lemma 6.

$$\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] = \frac{\sigma_z^2 IN}{P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \quad (34)$$

Proof: Using the independence of channel coefficients, for $1 \leq n \leq N$, we have

$$\begin{aligned}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] &= \mathbb{E}\Big[\Big(\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{1}{P_tMCK\sigma_h^2\bar{\beta}_c}\\ &\quad\times\mathrm{Re}\big\{\big(h^{i,n}_{m,c,k}(t)\big)^* z^{i,n}_{c,k}(t)\big\}\Big)^2\Big]\\ &= \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{1}{P_t^2M^2C^2K^2\sigma_h^4\bar{\beta}_c^2}\\ &\quad\times\big(\mathrm{Re}\big\{\big(h^{i,n}_{m,c,k}(t)\big)^* z^{i,n}_{c,k}(t)\big\}\big)^2\Big]\\ &= \frac{\sigma_z^2 I}{2P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \quad (35)\end{aligned}$$

The same result holds for $N+1 \leq n \leq 2N$. Combining the two results concludes the lemma. ■

Combining the results in Lemmas 4, 5, and 6 and applying Assumption 2 with (19) completes the proof of Lemma 1.

References

[1] 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.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

[3] D. Gunduz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to earn at the edge," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 14–19, 2020.

[4] M. Mohammadi Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[5] M. M. Amiri, T. M. Duman, D. Gunduz, S. R. Kulkarni, and H. V. P. Poor, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5129–5143, 2021.

[6] M. M. Amiri, S. R. Kulkarni, and H. V. Poor, "Federated learning with downlink device selection," *arXiv preprint arXiv:2107.03510*, 2021.

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

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

[9] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, 2021.

[10] B. Tegin and T. M. Duman, "Blind federated learning at the wireless edge with low-resolution adc and dac," *IEEE Trans. on Wireless Commun.*, 2021.

[11] ——, "Federated learning over time-varying channels," in *IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, Dec. 2021.

[12] T. Sery, N. Shlezinger, K. Cohen, and Y. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Trans. Signal Process.*, vol. 69, pp. 3796–3811, 2021.

[13] C. T. Dinh, N. H. Tran, M. N. Nguyen, C. S. Hong, W. Bao, A. Y. Zomaya, and V. Gramoli, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Transactions on Networking*, vol. 29, no. 1, pp. 398–409, 2020.

[14] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, 2021.

[15] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. Vincent Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, pp. 1–16, 2021.

[16] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8866–8870.

[17] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

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

[19] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Local averaging helps: Hierarchical federated learning and convergence analysis," *arXiv preprint arXiv:2010.12998*, 2020.

[20] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[21] Y. LeCun, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1998.

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

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