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

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9839230

***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 severely 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 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 devices such as mobile phones and Internet-of-things (IoT) sensors have enabled the accelerating rise of machine learning (ML) algorithms. Traditionally, ML algorithms require all the data to be collected at a cloud server for model training, which raises concerns regarding privacy, cost, and latency. Firstly, data owners may be sensitive about sharing their personal data; secondly, the increasing quality and volume of collected 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 models are trained locally instead of using a centralized server for training [2].

In FL, several data owners, called mobile users (MUs), are selected at each iteration based on some criteria such as their computing capability, available power, and location [3]. The parameter server (PS) sends the current global model to the selected MUs. Each of these MUs trains a local model by carrying out multiple stochastic gradient descent (SGD) iterations using its own data and computing power. Then, each MU sends only the weight updates to the PS, which performs model aggregation to update the global model. These steps are repeated until a convergence criterion is met.

Despite its superiority over traditional ML, adverse channel effects in wireless setups and increased communication costs pose challenges for 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 thanks to its efficient strategy that allocates all the MUs 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 concern, e.g., when some MUs are far away from the PS, as this would introduce bias across the updates.

Recent developments in FL include device selection algorithms [6], efficient communication schemes [4], [7]–[11], addressing of 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 is 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 the 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 study.

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 specific communication model and the HOTAFL framework, respectively. In Section IV, we provide a convergence analysis of HOTAFL under certain convexity

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).

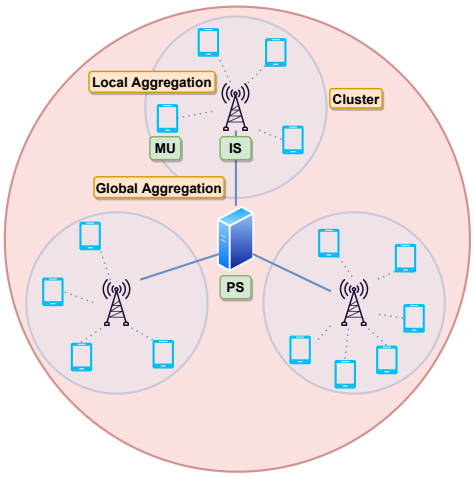

Fig. 1: HOTAFL system model.

assumptions on the loss function. We present our numerical results in Section V, and 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, and 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 MU 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 MU 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 groups 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 estimate 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 SGD.

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

A. Ideal Communication

We refer to the case in which 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

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

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

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

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

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{MC} \sum_{c=1}^{C} \sum_{i=1}^{I} \sum_{m=1}^{M} \Delta\boldsymbol{\theta}_{m,c}^{i}(t). \tag{6}$$

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 MUs 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 and no CSI at the MUs. For the k-th antenna, the received signal at the c-th IS can be written as[1]

$$\boldsymbol{y}_{IS,c,k}^{i}(t) = \sum_{m=1}^{M} \boldsymbol{h}_{m,c,k}^{i}(t) \circ \boldsymbol{x}_{m,c,k}^{i}(t) + \boldsymbol{z}_{IS,c,k}^{i}(t), \tag{7}$$

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 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 MU 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 communications, in order to increase the spectral efficiency, the model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}^{i,cx}_{m,c}(t) \in \mathbb{C}^N$ with the following real and imaginary parts

$$\Delta\boldsymbol{\theta}^{i,re}_{m,c}(t) \triangleq \left[\Delta\theta^{i,1}_{m,c}(t), \Delta\theta^{i,2}_{m,c}(t), \ldots, \Delta\theta^{i,N}_{m,c}(t)\right]^T, \quad (8a)$$

$$\Delta\boldsymbol{\theta}^{i,im}_{m,c}(t) \triangleq \left[\Delta\theta^{i,N+1}_{m,c}(t), \Delta\theta^{i,N+2}_{m,c}(t), \ldots, \Delta\theta^{i,2N}_{m,c}(t)\right]^T. \quad (8b)$$

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

$$\boldsymbol{y}^i_{IS,c,k}(t) = P_t \sum_{m=1}^{M} \boldsymbol{h}^i_{m,c,k}(t) \circ \Delta\boldsymbol{\theta}^{i,cx}_{m,c}(t) + \boldsymbol{z}^i_{IS,c,k}(t), \quad (9)$$

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

$$y^{i,n}_{IS,c}(t) = \underbrace{P_t \sum_{m=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}|h^{i,n}_{m,c,k}(t)|^2\right)\Delta\theta^{i,n,cx}_{m,c}(t)}_{y^{i,n,sig}_{IS,c}(t) \text{ (signal term)}}$$
$$+ \underbrace{\frac{P_t}{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{k=1}^{K}(h^{i,n}_{m,c,k}(t))^* h^{i,n}_{m',c,k}(t)\Delta\theta^{i,n,cx}_{m',c}(t)}_{y^{i,n,itf}_{IS,c}(t) \text{ (interference term)}}$$
$$+ \underbrace{\frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K}(h^{i,n}_{m,c,k}(t))^* z^{i,n}_{c,k}(t)}_{y^{i,n,no}_{IS,c}(t) \text{ (noise term)}}. \quad (10)$$

Aggregated model differences can be recovered by

$$\Delta\hat{\theta}^{i,n}_{IS,c}(t) = \frac{1}{P_t M\sigma_h^2\bar{\beta}_c}\,\mathrm{Re}\{y^{i,n}_{IS,c}(t)\}, \quad (11a)$$

$$\Delta\hat{\theta}^{i,n+N}_{IS,c}(t) = \frac{1}{P_t M\sigma_h^2\bar{\beta}_c}\,\mathrm{Im}\{y^{i,n}_{IS,c}(t)\}, \quad (11b)$$

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

$$\boldsymbol{\theta}^{i+1}_{IS,c}(t) = \boldsymbol{\theta}^i_{IS,c}(t) + \Delta\hat{\boldsymbol{\theta}}^i_{IS,c}(t), \quad (12)$$

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

2) Global Aggregation: This part is similar to ideal communication. The only difference is that the aggregated signals are the 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

$$x^n_{PS,c}(t) = \Delta\theta^n_{PS,c}(t) + j\Delta\theta^{n+N}_{PS,c}(t). \quad (13)$$

Then, using (5), (10), (13) and recursion, the received signal for $1 \le n \le N$ (similarly for $N+1 \le n \le 2N$) can be written as

$$y^n_{PS}(t) = \sum_{c=1}^{C} x^n_{PS,c}(t) \quad (14)$$
$$= \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y^{i,n,sig}_{IS,c}(t)\right\}}{P_t M\sigma_h^2}}_{y^n_{PS,1}(t)} + \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y^{i,n,itf}_{IS,c}(t)\right\}}{P_t M\sigma_h^2}}_{y^n_{PS,2}(t)}$$
$$+ \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y^{i,n,no}_{IS,c}(t)\right\}}{P_t M\sigma_h^2}}_{y^n_{PS,3}(t)}. \quad (15)$$

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

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t), \quad (16)$$

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

IV. Convergence Analysis of HOTAFL

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 $\Gamma \triangleq F^* - \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{B_{m,c}}{B}F^*_{m,c} \ge 0$. In addition, assume that the learning rate of the overall system does not change in local iterations, i.e., $\eta^{i,j}_{m,c}(t) = \eta(t)$. Therefore, we can write the global update rule as

$$\boldsymbol{\theta}^{i,j+1}_{m,c}(t) = \boldsymbol{\theta}^{i,j}_{m,c}(t) - \eta(t)\nabla F_{m,c}(\boldsymbol{\theta}^{i,j}_{m,c}(t), \boldsymbol{\xi}^{i,j}_{m,c}(t)), \quad (17)$$

which can also be written as

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

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

$$F_{m,c}(\boldsymbol{v}) - F_{m,c}(\boldsymbol{w}) \le \langle \boldsymbol{v}-\boldsymbol{w}, \nabla F_{m,c}(\boldsymbol{w})\rangle + \frac{L}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2, \quad (19)$$

$$F_{m,c}(\boldsymbol{v}) - F_{m,c}(\boldsymbol{w}) \ge \langle \boldsymbol{v}-\boldsymbol{w}, \nabla F_{m,c}(\boldsymbol{w})\rangle + \frac{\mu}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2. \quad (20)$$

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

$$\mathbb{E}_\xi\left[\left\|\nabla F_{m,c}(\boldsymbol{\theta}^{i,j}_{m,c}(t), \boldsymbol{\xi}^{i,j}_{m,c}(t))\right\|_2^2\right] \le G^2, \quad (21)$$

which translates to $\forall n \in [2N]$, $\mathbb{E}_\xi\left[\nabla F_{m,c}(\theta^{i,j,n}_{m,c}, \xi^{i,j,n}_{m,c}(t))\right] \le G$.

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

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2\right]$$
$$\le \left(\prod_{a=1}^{t-1}X(a)\right)\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{b=1}^{t-1}Y(b)\prod_{a=b+1}^{t-1}X(a), \quad (22)$$

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

$$Y(a)=\frac{\eta^2(a)\tau^2G^2I}{M^2C^2}\sum_{m_1=1}^{M}\sum_{c_1=1}^{C}\Big(\frac{\beta_{m_1,c_1}^2}{K\bar{\beta}_{c_1}^2}+\Big(\sum_{m_2=1}^{M}\sum_{c_2=1}^{C}A_1I\Big)\Big)$$
$$+\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\frac{\eta^2(a)\tau^2G^2I\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}$$
$$+\frac{\sigma_z^2IN}{P_a^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}$$
$$+(1+\mu(1-\eta(a))\,\eta^2(a)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}$$
$$+\eta^2(a)I(\tau^2+\tau-1)G^2+2\eta(a)I(\tau-1)\Gamma, \quad (23)$$

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 an auxiliary variable $\boldsymbol{v}(t+1)\triangleq\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)$. Then, we have

$$\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{\theta}^*\|_2^2=\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)+\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2$$
$$=\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\|_2^2+\|\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2$$
$$+2\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle. \quad (24)$$

Next, we provide upper bounds on the three terms of (24).

Lemma 1. *We have*

$$\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\big\|_2^2\Big]$$
$$\leq\frac{\eta^2(t)\tau^2G^2I}{M^2C^2}\sum_{m_1=1}^{M}\sum_{c_1=1}^{C}\Big(\frac{\beta_{m_1,c_1}^2}{K\bar{\beta}_{c_1}^2}+\Big(\sum_{m_2=1}^{M}\sum_{c_2=1}^{C}A_1I\Big)\Big)$$
$$+\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\frac{\eta^2(t)\tau^2G^2I\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}$$
$$+\frac{\sigma_z^2IN}{P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \quad (25)$$

Proof: See Appendix A. ∎

Lemma 2. *We have*

$$\mathbb{E}\Big[\big\|v(t+1)-\boldsymbol{\theta}^*\big\|_2^2\Big]\leq(1-\mu\eta(t)I(\tau-\eta(t)(\tau-1)))\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big]$$
$$+(1+\mu(1-\eta(t))\,\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}$$
$$+\eta^2(t)I(\tau^2+\tau-1)G^2+2\eta(t)I(\tau-1)\Gamma. \quad (26)$$

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*

$$\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T))-F^*\right]\leq\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T)-\boldsymbol{\theta}^*\|_2^2\right]$$
$$\leq\frac{L}{2}\Big(\prod_{n=1}^{T-1}X(n)\Big)\|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2+\frac{L}{2}\sum_{p=1}^{T-1}Y(p)\prod_{n=p+1}^{T-1}X(n). \quad (27)$$

Remark. Since the third term in $Y(a)$ is independent of $\eta(a)$, even for $\lim_{t\to\infty}\eta(t)=0$, we have $\lim_{t\to\infty}\mathbb{E}[F(\boldsymbol{\theta}_{PS}(t))]-F^*\neq0$. $Y(a)$ is also proportional to I, meaning that more cluster aggregations do not always provide faster convergence. However, since the MUs face lower path losses in HOTAFL than in the conventional FL, it can reach a higher accuracy. Moreover, increasing the number of clusters C leads to a faster convergence, however, at the cost of employing more ISs.

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. MUs are randomly placed in the clusters in such a way that their distance to the PS is between 0.5 and 3, and between 0.5 and 1 to their corresponding IS.

We use the MNIST [21] and CIFAR-10 [22] datasets with Adam optimizer [23], and consider 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 samples with the same label. Then, 5 groups are assigned to each MU randomly. For CIFAR-10, we use the neural network given in [5] with $2N=307498$ whereas for MNIST, we employ a one-layer network with $2N=7850$.

Three scenarios are considered: baseline with error-free transmissions, FL with OTA, i.e., ISs are not employed, and all the MUs aggregate parameters at the PS, and HOTAFL. We set the total number of global iterations T to 200, the mini-batch size to $|\boldsymbol{\xi}_{m,c}^i(t)|=500$, the path loss exponent p to 4, $\sigma_h^2=1$, $\sigma_z^2=10$ for the MNIST, and $\sigma_z^2=1$ 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 with the selected geometry, bringing the servers closer to the MUs enhances the learning accuracy significantly. One reason is that the cluster structure enables the MUs share their model differences with a local IS closer than the PS, reducing the adverse effects of the large-scale wireless channel. 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 received signals are distorted more compared to those of HOTAFL due to the more severe wireless channel effects. More local iterations enable faster convergence at the cost of increased transmit power. Increasing τ compensates the accuracy under a more complex model. In Fig. 5, we compare the convergence rates

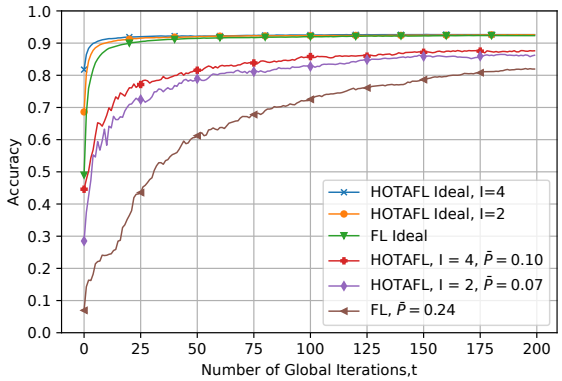

Fig. 2: Test accuracy for i.i.d. MNIST data with $\tau = 1$.

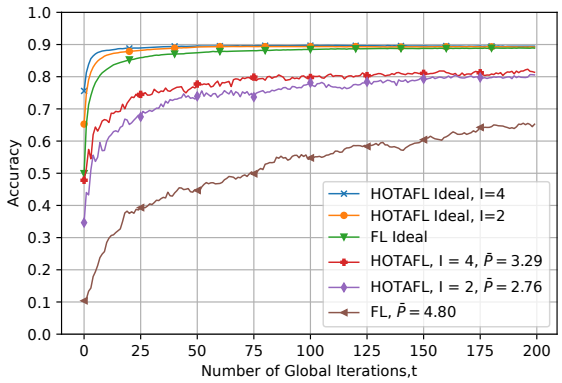

Fig. 3: Test accuracy for non-i.i.d. MNIST data with $\tau = 3$.

of conventional FL and HOTAFL using the upper bound in (27), with $2N = 7850, L = 10, \mu = 1, G^2 = 1, \Gamma = 1, \eta(t) = 5\cdot10^{-2} - 2\cdot10^{-5}t, P_t = 1+10^{-2}t, \beta_{m,c} = 1, \forall m \in [M], \forall c \in [C], \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$. The convergence rate of HOTAFL and the ideal case are very close, and they become almost the same when the number of local iterations is increased.

VI. Conclusions

We have proposed HOTAFL, which enables geographically localized model aggregation by employing ISs located in the areas where the MUs are more densely located. Our framework includes OTA cluster aggregations, which allows the MUs to simultaneously transmit and aggregate their model updates at the ISs over a wireless channel with path-loss and fading. We have analyzed the convergence rate of HOTAFL, and examined its performance with different datasets and data distributions. The results show that HOTAFL outperforms the conventional FL significantly. As a future work, one can consider inter-cluster interference as well as fading channels between ISs and PS.

Appendix A

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

$$\mathbb{E}\left[||\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)||_2^2\right] = \mathbb{E}\left[\left\|\Delta\hat{\boldsymbol{\theta}}_{PS}(t) - \Delta\boldsymbol{\theta}_{PS}(t)\right\|_2^2\right]$$
$$= \sum_{n=1}^{2N} \left(\mathbb{E}\left[\left(\Delta\hat{\theta}^n_{PS,1}(t) - \Delta\theta^n_{PS}(t)\right)^2\right] + \sum_{l=2}^{3} \mathbb{E}\left[\left(\Delta\hat{\theta}^n_{PS,l}(t)\right)^2\right]. \quad (28)$$

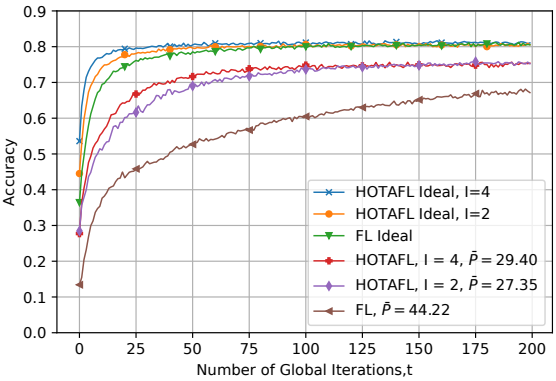

Fig. 4: Test accuracy for i.i.d. CIFAR-10 data with $\tau = 5$.

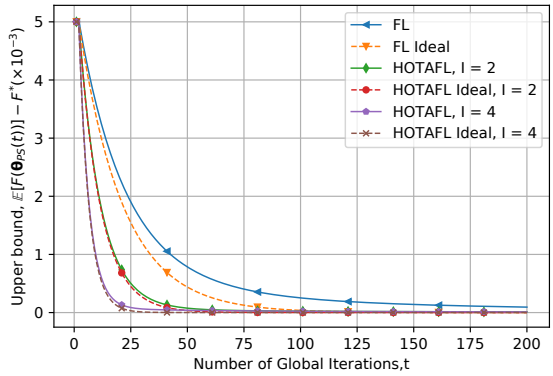

Fig. 5: Convergence rate for i.i.d. MNIST data with $\tau = 1$.

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

Lemma 4. $\sum_{n=1}^{2N} \mathbb{E}\left[\left(\Delta\hat{\theta}^n_{PS,1}(t) - \Delta\theta^n_{PS}(t)\right)^2\right]$

$$= \frac{1}{M^2C^2} \sum_{m_1=1}^{M} \sum_{c_1=1}^{C} \sum_{i_1=1}^{I} \Bigg(\frac{\beta^2_{m_1,c_1}}{K\bar{\beta}^2_{c_1}} \mathbb{E}\left[\left\|\Delta\boldsymbol{\theta}^{i_1}_{m_1,c_1}(t)\right\|_2^2\right]$$
$$+ \Bigg(\sum_{m_2=1}^{M} \sum_{c_2=1}^{C} \sum_{i_2=1}^{I} \sum_{n=1}^{2N} A_1 \mathbb{E}\left[\Delta\theta^{i_1,n}_{m_1,c_1}(t) \Delta\theta^{i_2,n}_{m_2,c_2}(t)\right] \Bigg) \Bigg), \quad (29)$$

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 (6) and (11), we have

$$\mathbb{E}\left[\left(\Delta\hat{\theta}^n_{PS,1}(t) - \Delta\theta^n_{PS}(t)\right)^2\right]$$
$$= \mathbb{E}\Bigg[\frac{1}{M^2C^2} \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{c_1=1}^{C} \sum_{c_2=1}^{C} \sum_{i_1=1}^{I} \sum_{i_2=1}^{I} \Delta\theta^{i_1,n}_{m_1,c_1}(t)$$
$$\times \Delta\theta^{i_2,n}_{m_2,c_2}(t) \Bigg(1 - \frac{1}{K\sigma_h^2\bar{\beta}_{c_1}} \sum_{k_1=1}^{K} |h^{i_1,n}_{m_1,c_1,k_1}(t)|^2$$
$$- \frac{1}{K\sigma_h^2\bar{\beta}_{c_2}} \sum_{k_2=1}^{K} |h^{i_2,n}_{m_2,c_2,k_2}(t)|^2$$
$$+ \frac{1}{K^2\sigma_h^4\bar{\beta}^2_{c_1}} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} |h^{i_1,n}_{m_1,c_1,k_1}(t)|^2 |h^{i_2,n}_{m_2,c_2,k_2}(t)|^2 \Bigg) \Bigg]. \quad (30)$$

Summing over all the symbols and using the independence of channel coefficients result in (29). ∎

Lemma 5. $\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big]$

$$= \sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}\mathbb{E}\big[\,\big\|\Delta\boldsymbol{\theta}^i_{m',c}(t)\big\|_2^2\big]. \quad (31)$$

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

$$\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] = \mathbb{E}\Big[\Big(\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{1}{MCK\sigma_h^2\bar{\beta}_c}$$
$$\times\sum_{k=1}^{K}\mathrm{Re}\left\{\big(h^{i,n}_{m,c,k}(t)\big)^*h^{i,n}_{m',c,k}(t)\Delta\theta^{i,n}_{m',c}(t)\right\}\Big)^2\Big]$$
$$= \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\beta_{m,c}\beta_{m',c}}{2M^2C^2K\bar{\beta}_c^2}$$
$$\times\Big(\big(\Delta\theta^{i,n}_{m',c}(t)\big)^2+\big(\Delta\theta^{i,n+N}_{m',c}(t)\big)^2+\Delta\theta^{i,n}_{m,c}(t)\Delta\theta^{i,n}_{m',c}(t)$$
$$-\Delta\theta^{i,n+N}_{m,c}(t)\Delta\theta^{i,n+N}_{m',c}(t)\Big)\Big] \quad (32)$$

Obtaining the expressions for $N+1 \leq n \leq 2N$ in a similar manner and combining the two, results in (31). ∎

Lemma 6.

$$\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] = \frac{\sigma_z^2IN}{P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \quad (33)$$

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

$$\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] = \mathbb{E}\Big[\Big(\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{1}{P_tMCK\sigma_h^2\bar{\beta}_c}$$
$$\times\mathrm{Re}\left\{\big(h^{i,n}_{m,c,k}(t)\big)^*z^{i,n}_{c,k}(t)\right\}\Big)^2\Big]$$
$$= \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{1}{P_t^2M^2C^2K^2\sigma_h^4\bar{\beta}_c^2}$$
$$\times\Big(\mathrm{Re}\left\{\big(h^{i,n}_{m,c,k}(t)\big)^*z^{i,n}_{c,k}(t)\right\}\Big)^2\Big]$$
$$= \frac{\sigma_z^2I}{2P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \quad (34)$$

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

Combining the results in Lemmas 4, 5, and 6 and applying Assumption 2 with (18) 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.