Performance Analysis for Channel-Weighted Federated Learning in OMA Wireless Networks

Na Yan, Kezhi Wang, Cunhua Pan and Kok Keong Chai

***Abstract*—To alleviate the negative impact of noise on wireless federated learning (FL), we propose a channel-weighted aggregation scheme of FL (CWA-FL), in which the parameter server (PS) makes aggregation of the gradients according to the channel conditions of devices. In the proposed scheme, the gradients are transmitted to the PS in an uncoded way through an orthogonal multiple access (OMA) channel, which can avoid the synchronization issue among devices faced by over-the-air FL. The convergence analysis of CWA-FL is conducted and the theoretical results show that the scheme can converge with the rate of $\mathcal{O}\left(\frac{1}{T}\right)$. Simulation results show that the proposed scheme performs better than the equal-weighted aggregation scheme of FL (EWA-FL) and is more robust to noise.**

***Intex Terms* - Federated learning, aggregation of gradients, orthogonal multiple access, convergence analysis.**

I. Introduction

Federated learning (FL) [1] has been proposed as a distributed machine learning technique, where edge devices collaboratively train a model using only locally available data with the help of a parameter server (PS). In wireless FL, each device trains model or computes gradient locally and then sends the updated model or the gradient to the PS through wireless channel for centralized aggregation.

To improve the performance of wireless FL, the authors in [2] investigated the convergence of FL over a noisy downlink and the case of noise in both uplink and downlink was studied in [3]. The analysis in [2, 3] demonstrated that the noise in wireless communications makes the gradients received at the PS less accurate, therefore, has a negative effect on learning performance. To alleviate the impact of noise, the authors in [4] proposed a robust FL method by formulating the training problem as a parallel optimization. Other works [5, 6] directly used the noisy gradients to perform global updates instead of using strategies to remove the effect of noise. All the above researches make aggregation by simply averaging or using dataset size to weight each gradient or model, which is the most basic aggregation way but has little help on mitigating the negative impact of noise on gradients.

This work of Na Yan was supported by China Scholarship Council. *(Corresponding author: Kezhi Wang and Cunhua Pan.)*

Na Yan and Kok Keong Chai are with School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: n.yan, michael.chai@qmul.ac.uk).

Kezhi Wang is with Department of Computer and Information Sciences, Northumbria University, NE2 1XE, Newcastle upon Tyne, U.K. (e-mail: kezhi.wang@northumbria.ac.uk).

Cunhua Pan is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. (email: cpan@seu.edu.cn).

Most of wireless FL studies [7–9] considered that the gradients are transmitted with digital transmission. However, for the low-cost, low-power devices, e.g., wireless IoT sensors, analog transmission may be more suitable due to the lack of analog-to-digital (A/D) function of these devices. Additionally, the communication efficiency could be greatly improved with analog transmission by avoiding quantization and channel encoding/decoding, which is very attractive for low-latency applications at the edge of wireless networks [10]. On the other hand, some works [5, 10] have tried to use analog transmission to achieve over-the-air aggregation of gradients by exploiting the superposition property of a wireless multiple access channel (MAC). However, over-the-air computation (Aircomp) requires a stringent synchronization among devices, which is quite challenging in realistic scenarios. Furthermore, the effective signal-to-noise ratio (SNR) of the system will be limited by the device with the worst channel condition [11], which might not be suitable for the power-limited edge networks. By contrast, with orthogonal multiple access (OMA) channel, there are no such limitations.

Against the above background, in this paper, we consider the uncoded gradients are transmitted from devices to the PS through an OMA channel, which can relieve the synchronization requirement faced by over-the-air FL. Additionlly, we propose a channel-weighted aggregation scheme of FL (CWA-FL) where PS makes gradient aggregation according to the channel conditions of devices to alleviate the bad effect of the noise on learning performance. We theoretically prove that the proposed scheme can converge with the rate of $\mathcal{O}\left(\frac{1}{T}\right)$. Simulation results show that the proposed scheme performs better than the equal-weighted aggregation scheme of FL (EWA-FL) and the performance superiority is more significant in the case with larger power of noise and fewer devices.

II. System Model

We consider a wireless FL system as shown in Fig. 1, where K edge devices, denoted by $\mathcal{K} = \{1, 2, ..., K\}$ are connected to a PS for centralized aggregation through an OMA channel. Suppose that each device holds the dataset $\mathcal{D}_k$ of size D_k and we assume that $D_1 = \cdots D_k = \cdots D_K$ for simplicity. The goal of the learning is to minimize the global loss function as shown in (1):

$$\min_{\boldsymbol{w}} \left\{ F\left(\boldsymbol{w}\right) \triangleq \frac{1}{K} \sum\nolimits_{k=1}^{K} F_k\left(\boldsymbol{w}\right) \right\}, \tag{1}$$

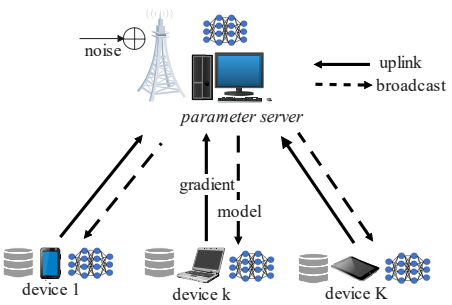


Fig. 1: The OMA wireless FL system.

where $\boldsymbol{w} \in \mathbb{R}^d$ denotes the model parameter to be optimized. The local objective function $F_k(\boldsymbol{w})$ is defined as follows,

$$F_k(\boldsymbol{w}) \triangleq \frac{1}{D_k} \sum\nolimits_{(\mathbf{u},v)\in\mathcal{D}_k} f(\boldsymbol{w};(\mathbf{u},v)), \tag{2}$$

where $f(\cdot,\cdot)$ denotes the loss function corresponding to specific learning model and $(\mathbf{u}, v)$ is one of the data samples.

III. Channel-weighted aggregation scheme of FL and convergence analysis

We propose the CWA-FL scheme to reduce the negative impact of noise on FL in this section.

A. CWA-FL: channel-weighted aggregation scheme of FL

In this paper, we assume that the channel state information (CSI) of the devices can be estimated on the PS. Then, the detailed process of CWA-FL can be given as follows.

(a) At the beginning of each round t, PS broadcasts the latest global parameter $\tilde{\boldsymbol{w}}^t$ to the devices.

(b) Each device firstly sets $\boldsymbol{w}_k^t = \tilde{\boldsymbol{w}}^t$, and then selects a batch of data samples to compute its gradient as follows,

$$\nabla F_k\left(\boldsymbol{w}_k^t, \xi_k^t\right) = \frac{1}{b_k} \sum\nolimits_{(\mathbf{u},v)\in\xi_k^t} \nabla f\left(\boldsymbol{w}_k^t;(\mathbf{u},v)\right), \tag{3}$$

where ξ_k^t is the batch of samples and b_k is the size of ξ_k^t.

(c) Then, devices send the gradients to the PS and the input signal of device k is given by,

$$\boldsymbol{x}_k^t = p_k^t \nabla F_k\left(\boldsymbol{w}_k^t, \xi_k^t\right), \tag{4}$$

where p_k^t is the power scaling factor and is required to satisfy $p_k^t = \frac{\sqrt{P_k}}{\left\|\nabla F_k\left(\boldsymbol{w}_k^t,\xi_k^t\right)\right\|_2}$ for the transmit power P_k constraint. The received signal at the PS from device k is given by

$$\boldsymbol{y}_k^t = h_k^t \boldsymbol{x}_k^t + \boldsymbol{n}^t = h_k^t p_k^t \nabla F_k\left(\boldsymbol{w}_k^t, \xi_k^t\right) + \boldsymbol{n}^t, \tag{5}$$

where $h_k^t \in \mathbb{R}^+$ is the real channel gain coefficient we assumed for simplicity [12], and $\boldsymbol{n}^t \in \mathbb{R}^d$ is the received noise, following the distribution of $\mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$.

(d) Upon receiving all the gradients, PS makes gradient aggregation by,

$$\widehat{\nabla F}\left(\tilde{\boldsymbol{w}}^t\right) = \frac{1}{C^t} \sum_{k=1}^{K} \boldsymbol{y}_k^t = \sum_{k=1}^{K} \frac{c_k^t}{C^t} \left(\nabla F_k\left(\boldsymbol{w}_k^t, \xi_k^t\right) + \frac{\boldsymbol{n}^t}{c_k^t}\right), \tag{6}$$

where $c_k^t = h_k^t p_k^t$ denotes the channel conditions of device k in round t and $C^t = \sum_{k=1}^K c_k^t$. Different from EWA-FL where the PS simply averages the recovered gradients by $\widehat{\nabla F}'\left(\tilde{\boldsymbol{w}}^t\right) = \frac{1}{K}\sum_{k=1}^K \left(\nabla F_k\left(\boldsymbol{w}_k^t, \xi_k^t\right) + \frac{\boldsymbol{n}^t}{c_k^t}\right)$, the CWA-FL assigns smaller weight to the gradient from the device with poor channel quality, to alleviate the distortion of the aggregated gradient.

(e) Finally, PS performs global model update based on the aggregated gradient as follows,

$$\tilde{\boldsymbol{w}}^{t+1} = \tilde{\boldsymbol{w}}^t - \eta^t \widehat{\nabla F}\left(\tilde{\boldsymbol{w}}^t\right), \tag{7}$$

where η^t is the learning rate in round t.

We assume that the PS is a more capable node with sufficient energy resources. Therefore, the broadcast of the global parameter is error-free [6, 10] .

B. Convergence analysis of CWA-FL

We first show the notations and assumptions applied in the following analysis.

1) Notations: Motivated by [13], we define two virtual sequences to denote the aggregated full gradient and stochastic gradient respectively as follows,

$$\bar{\boldsymbol{g}}^t = \sum\nolimits_{k=1}^K \frac{c_k^t}{C^t} \nabla F_k\left(\boldsymbol{w}_k^t\right), \boldsymbol{g}^t = \sum\nolimits_{k=1}^K \frac{c_k^t}{C^t} \nabla F_k\left(\boldsymbol{w}_k^t, \xi_k^t\right). \tag{8}$$

2) Assumptions: For analysis, we provide the following assumptions on loss functions, defined in (2).

Assumption 1. *For each k, $F_k(\cdot)$ is L-smooth, i.e., for all $\boldsymbol{\iota}'$ and $\boldsymbol{\iota}$, one has,*

$$F_k(\boldsymbol{\iota}') - F_k(\boldsymbol{\iota}) \leqslant (\boldsymbol{\iota}' - \boldsymbol{\iota})^{\mathrm{T}} \nabla F_k(\boldsymbol{\iota}) + \frac{L}{2}\|\boldsymbol{\iota}' - \boldsymbol{\iota}\|_2^2. \tag{9}$$

Assumption 2. *For each k, $F_k(\cdot)$ is μ-strongly convex, i.e., for all $\boldsymbol{\iota}'$ and $\boldsymbol{\iota}$, one has,*

$$F_k(\boldsymbol{\iota}') - F_k(\boldsymbol{\iota}) \geqslant (\boldsymbol{\iota}' - \boldsymbol{\iota})^{\mathrm{T}} \nabla F_k(\boldsymbol{\iota}) + \frac{\mu}{2}\|\boldsymbol{\iota}' - \boldsymbol{\iota}\|_2^2. \tag{10}$$

Assumption 3. *Assume that the stochastic gradient is an unbiased estimate of the full gradient,*

$$\mathbb{E}\left[\nabla F_k(\boldsymbol{w}, \xi_k)\right] = \nabla F_k(\boldsymbol{w}). \tag{11}$$

The variance of the local gradient for each k satisfies,

$$\mathbb{E}\left[\|\nabla F_k(\boldsymbol{w}, \xi_k) - \nabla F_k(\boldsymbol{w})\|_2^2\right] \leqslant \delta_k^2, \tag{12}$$

where ξ_k denotes the data chosen from $\mathcal{D}_k$.

3) Convergence analysis: We give the following lemmas based on the above definitions and assumptions.

Lemma 1. *Assume that Assumption 1 holds and $\boldsymbol{w}^* = [w_1^*, \cdots, w_d^*]$, $\boldsymbol{w}_k^* = \left[w_{k,1}^*, \cdots, w_{k,d}^*\right]$ are the globally optimal model and the locally optimal model of device k. Then, for each device k, the upper bound of the gap between $F_k(\boldsymbol{w}^*)$ and $F_k(\boldsymbol{w}_k^*)$ is given by,*

$$F_k(\boldsymbol{w}^*) - F_k(\boldsymbol{w}_k^*) \leqslant \tau, \tag{13}$$

where $\tau = \max_k \left\{\frac{Ld}{2}\left(\max_i \left\{\left|w_i^* - w_{k,i}^*\right|\right\}\right)^2\right\}$.

Proof. According to to Assumption 1, one has $F_k(\boldsymbol{w}^*) - F_k(\boldsymbol{w}_k^*) \overset{(a)}{\leqslant} \frac{L}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_k^*\|_2^2$ where (a) comes from the fact that $\nabla F_k(\boldsymbol{w}_k^*) = 0$. Then, applying $\|\cdot\|_2 \leqslant \sqrt{d}\|\cdot\|_\infty$, one completes the proof. □

Lemma 2. *Assume that Assumption 3 holds, then, the variance of the aggregated gradient is bounded by,*

$$\mathbb{E}\left(\|\boldsymbol{g}^t - \bar{\boldsymbol{g}}^t\|_2^2\right) \leqslant \sum_{k=1}^K \frac{c_k^t}{C^t}\delta_k^2. \tag{14}$$

Proof. See Appendix A. □

Lemma 3. *Assume that Assumption 1 to Assumption 3 hold. A constant κ and η^t satisfy $\frac{1}{\kappa} \leqslant \eta^t \leqslant \frac{1}{L}$. One has,*

$$\mathbb{E}\left[\left\|\tilde{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^*\right\|_2^2\right] \leqslant (1-\mu\eta^t)\left\|\tilde{\boldsymbol{w}}^t - \boldsymbol{w}^*\right\|_2^2 + (\eta^t)^2 X^t, \tag{15}$$

where $X^t = d\left(\frac{K\sigma}{C^t}\right)^2 + \sum_{k=1}^K \frac{c_k^t}{C^t}\delta_k^2 + 2\kappa\tau$.

Proof. See Appendix B. □

We define F^* as the training loss of the optimal model and one can obtain the optimality gap based on the above lemmas.

Theorem 1. *Assume that Assumption 1 to Assumption 3 hold and there is a constant κ satisfies $\frac{1}{\kappa} \leqslant \eta^t = \frac{2}{\mu t + 2L}$. When the training process terminates after T rounds and $\tilde{\boldsymbol{w}}^T$ is returned as the final solution, the bound of the optimality gap can be given by,*

$$\mathbb{E}\left[F\left(\tilde{\boldsymbol{w}}^T\right)\right] - F^* \leqslant \frac{L}{\mu T + 2L}\left(\frac{2\chi}{\mu} + L\left\|\tilde{\boldsymbol{w}}^0 - \boldsymbol{w}^*\right\|_2^2\right), \tag{16}$$

where $\chi = \max_t\{X^t\}$ *and* $X^t = d\left(\frac{K\sigma}{C^t}\right)^2 + \sum_{k=1}^K \frac{c_k^t}{C^t}\delta_k^2 + 2\kappa\tau$.

Proof. See Appendix C. □

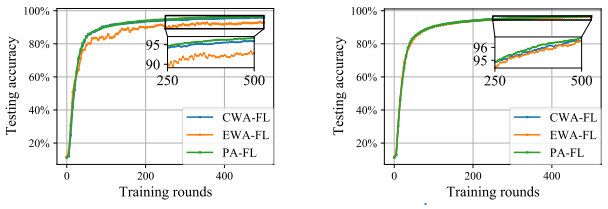
From (16), one can find that the optimality gap decreases with T, and will go to zero when T approaches infinity, which means that the proposed CWA-FL can converge with the rate of $\mathcal{O}\left(\frac{1}{T}\right)$.

IV. Simulation Results

In this section, we evaluate the performance of CWA-FL by comparing it with the perfect aggregation scheme of FL (PA-FL) without noise distortion and the EWA-FL.

We assume that the wireless channels from edge devices to the PS follow Rayleigh fading in different communication rounds. The transmit power budgets at each device are assumed to be the same and are set as $P_k = 30\text{dBm}$ [11]. We evaluate the proposed scheme through training Convolutional Neural Network (CNN) [14] on the popular MNIST [15] dataset. The batch size is set as $b_k = 64$ and we choose 5 batches every round for computing the gradients. We set the initial learning rate as $\eta^0 = 0.1$ and it decreases at a rate of 0.99 every round.

Fig. 2 plots the learning performance of different aggregation schemes under different variances of noise. One can find that the CWA-FL performs better than the EWA-FL in all cases. The performance of CWA-FL is quite close to the PA-FL even in the case with $K = 10$ while the EWA-FL obtains a poor learning accuracy. Therefore, the proposed scheme will consume fewer resources because it requires less device participation. Particularly, in the case of a larger variance of noise as shown in Fig. 2 (c)-(d), the performance superiority of CWA-FL is more significant, which means that the proposed scheme is more robust to noise. This is because the CWA-FL can effectively reduce the distortion of the aggregated gradient by assigning different weights to the noisy gradients in the aggregation process.



(a) $K = 10, \sigma^2 = 0.02$ (b) $K = 100, \sigma^2 = 0.02$

(c) $K = 10, \sigma^2 = 0.06$ (d) $K = 100, \sigma^2 = 0.06$

Fig. 2: Performance of different aggregation schemes under different values of noise variance.

V. Conclusion

In this paper, we considered that the gradients are transmitted to the PS via OMA channel to avoid the synchronization issue of Aircomp. We proposed a CWA-FL scheme to alleviate the distortion of the aggregated gradient by assigning smaller weight to the gradient of the device with poor channel quality. We then proved that the proposed scheme can converge with the rate of $\mathcal{O}\left(\frac{1}{T}\right)$. Simulation results have shown that the proposed scheme performs better than EWA-FL and is more robust to the situation that the number of devices is small and the power of noise is large.

Appendix A
Proof of Lemma 2

Accroding to (8), the variance of the aggregated gradient can be bounded by,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{g}^t - \bar{\boldsymbol{g}}^t\right\|_2^2\right] \\
&\overset{(a)}{\leqslant} \sum_{k=1}^K \frac{c_k^t}{C^t}\mathbb{E}\left[\left\|\nabla F_k\left(\tilde{\boldsymbol{w}}^t, \xi_k^t\right) - \nabla F_k\left(\tilde{\boldsymbol{w}}^t\right)\right\|_2^2\right] \\
&\overset{(b)}{\leqslant} \sum_{k=1}^K \frac{c_k^t}{C^t}\delta_k^2,
\end{aligned} \tag{17}$$

where (a) comes from Jensen's Inequality and (b) is from Assumption 3. □

APPENDIX B
PROOF OF LEMMA 3

The expression of $\left\|\tilde{\boldsymbol{w}}^{t+1}-\boldsymbol{w}^*\right\|_2^2$ is shown as follows,

$$\begin{aligned}
&\left\|\tilde{\boldsymbol{w}}^{t+1}-\boldsymbol{w}^*\right\|_2^2=\left\|\tilde{\boldsymbol{w}}^t-\eta^t\boldsymbol{g}^t-\tfrac{\eta^tK\boldsymbol{n}^t}{C^t}-\boldsymbol{w}^*\right\|_2^2\\
&=\underbrace{\left\|\tilde{\boldsymbol{w}}^t-\eta^t\bar{\boldsymbol{g}}^t+\eta^t\bar{\boldsymbol{g}}^t-\eta^t\boldsymbol{g}^t-\tfrac{\eta^tK\boldsymbol{n}^t}{C^t}-\boldsymbol{w}^*\right\|_2^2}_{A}+\underbrace{\left(\eta^t\right)^2\left\|\boldsymbol{g}^t-\bar{\boldsymbol{g}}^t\right\|_2^2}_{B}\\
&\underbrace{-2\eta^t\left\langle\tilde{\boldsymbol{w}}^t-\eta^t\bar{\boldsymbol{g}}^t-\frac{\eta^tK\boldsymbol{n}^t}{C^t}-\boldsymbol{w}^*,\boldsymbol{g}^t-\bar{\boldsymbol{g}}^t\right\rangle}_{C}.
\end{aligned}\tag{18}$$

It naturally follows from Assumption 3 that $\mathbb{E}[C]=0$. Then, the expression of $\mathbb{E}[A]$ can be given by

$$\begin{aligned}
&\mathbb{E}[A]=\mathbb{E}[\underbrace{\left\|\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*-\eta^t\bar{\boldsymbol{g}}^t\right\|_2^2}_{A_1}]+\mathbb{E}[\underbrace{\left\|\frac{\eta^tK\boldsymbol{n}^t}{C^t}\right\|_2^2}_{A_2}]\\
&\underbrace{-2\left\langle\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*-\eta^t\bar{\boldsymbol{g}}^t,\mathbb{E}\left[\frac{\eta^tK\boldsymbol{n}^t}{C^t}\right]\right\rangle}_{A_3},
\end{aligned}\tag{19}$$

where $A_3=0$ because $\mathbb{E}[\boldsymbol{n}^t]=0$. The expression of $\mathbb{E}[A_1]$ can be given by $\mathbb{E}[A_1]=\|\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*\|_2^2+\underbrace{\left(\eta^t\right)^2\left\|\bar{\boldsymbol{g}}^t\right\|_2^2}_{A_{1-1}}\underbrace{-2\eta^t\left\langle\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*,\bar{\boldsymbol{g}}^t\right\rangle}_{A_{1-2}}$. The bound of $\mathbb{E}[A_{1-1}]$ can be given by $\mathbb{E}[A_{1-1}]=\left(\eta^t\right)^2\left\|\sum_{k=1}^K\frac{c_k^t}{C^t}\nabla F_k\left(\boldsymbol{w}_k^t\right)\right\|_2^2$
$\overset{(a)}{\leqslant}\left(\eta^t\right)^2\sum_{k=1}^K\frac{c_k^t}{C^t}\left\|\nabla F_k\left(\boldsymbol{w}_k^t\right)\right\|_2^2$
$\overset{(b)}{\leqslant}\underbrace{2L\left(\eta^t\right)^2\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}_k^*\right)\right)}_{A_{1-1-1}}$, where (a) comes from Jensen's Inequality and (b) comes from Assumption 1 and the property of smooth function [16] as $\left\|\nabla F_k\left(\boldsymbol{w}_k^t\right)\right\|_2^2\leqslant 2L\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}_k^*\right)\right)$. The bound of $\mathbb{E}[A_{1-2}]$ can be given by

$$\begin{aligned}
&\mathbb{E}[A_{1-2}]=-2\eta^t\left\langle\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*,\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\nabla F_k\left(\boldsymbol{w}_k^t\right)\right\rangle\\
&=-2\eta^t\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left\langle\boldsymbol{w}_k^t-\boldsymbol{w}^*,\nabla F_k\left(\boldsymbol{w}_k^t\right)\right\rangle\\
&\overset{(a)}{\leqslant}-2\eta^t\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}^*\right)+\frac{\mu}{2}\left\|\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*\right\|_2^2\right)\\
&=\underbrace{-2\eta^t\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}^*\right)\right)}_{A_{1-2-1}}-\mu\eta^t\left\|\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*\right\|_2^2,
\end{aligned}\tag{20}$$

where (a) comes from Assumtion 2. Then, the bound of $\mathbb{E}[A_{1-1-1}]+\mathbb{E}[A_{1-2-1}]$ can be given by

$$\begin{aligned}
&\mathbb{E}[A_{1-1-1}]+\mathbb{E}[A_{1-2-1}]\\
&=-2\eta^t\left(1-L\eta^t\right)\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}^*\right)\right)\\
&+2L\left(\eta^t\right)^2\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}^*\right)-F_k\left(\boldsymbol{w}_k^*\right)\right)\overset{(a)}{\leqslant}2L\tau\left(\eta^t\right)^2\\
&\underbrace{-2\eta^t\left(1-L\eta^t\right)\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}^*\right)\right)}_{A_4},
\end{aligned}\tag{21}$$

where (a) follows from Assumption 3. The bound of $\mathbb{E}[A_4]$ can be given by,

$$\begin{aligned}
&\mathbb{E}[A_4]=-2\eta^t\left(1-L\eta^t\right)\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}^*\right)\right)\\
&=-2\eta^t\left(1-L\eta^t\right)\left[\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}_k^*\right)\right)\right.\\
&\left.+\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}_k^*\right)-F_k\left(\boldsymbol{w}^*\right)\right)\right]\\
&\overset{(a)}{\leqslant}2\eta^t\left(1-L\eta^t\right)\textstyle\sum_{k=1}^K\frac{c_k^t}{C^t}\left(F_k\left(\boldsymbol{w}^*\right)-F_k\left(\boldsymbol{w}_k^*\right)\right)\\
&\overset{(b)}{\leqslant}2\eta^t\left(1-L\eta^t\right)\tau,
\end{aligned}\tag{22}$$

where (a) comes from the fact that $F_k\left(\boldsymbol{w}_k^t\right)-F_k\left(\boldsymbol{w}_k^*\right)\geqslant 0$ and $\eta^t\leqslant\frac{1}{L}$. (b) is from Lemma 1.

$\mathbb{E}[A_2]$ can finally be given by $\mathbb{E}[A_2]=\left\|\frac{\eta^tK\boldsymbol{n}^t}{C^t}\right\|_2^2=d\left(\frac{\eta^tK\sigma}{C^t}\right)^2$. By plugging $\mathbb{E}(A_1)$, $\mathbb{E}(A_2)$, $\mathbb{E}(A_4)$ into $\mathbb{E}(A)$ and applying $\frac{1}{\kappa}\leqslant\eta^t$, we have $\mathbb{E}[A]\leqslant\left(1-\mu\eta^t\right)\left\|\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*\right\|_2^2+2\left(\eta^t\right)^2\kappa\tau+d\left(\frac{\eta^tK\sigma}{C^t}\right)^2$. From Lemma 1, the bound of $\mathbb{E}[B]$ can be given as $\mathbb{E}[B]\leqslant\left(\eta^t\right)^2\sum_{k=1}^K\frac{c_k^t}{C^t}\delta_k^2$.

By using the expression of $\mathbb{E}\left[\left\|\tilde{\boldsymbol{w}}^{t+1}-\boldsymbol{w}^*\right\|_2^2\right]$ and plugging $\mathbb{E}[A]$, $\mathbb{E}[B]$, $\mathbb{E}[C]$ into it, we obtain (15). Therefore, the proof is completed. □

APPENDIX C
PROOF OF THEOREM 1

Similar to [13], we define $\Delta^t=\mathbb{E}\left[\left\|\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*\right\|_2^2\right]$. It thus follows that $\Delta^{t+1}\leqslant\left(1-\mu\eta^t\right)\Delta^t+\left(\eta^t\right)^2X^t$. Let $\chi=\max_t\{X^t\}$ and $\eta^t=\frac{\alpha}{t+\beta}$ for some $\alpha\geqslant\frac{1}{\mu}$ and $\beta>1$ so that $\eta^0\leqslant\min\left\{\frac{1}{\mu},\frac{1}{L}\right\}=\frac{1}{L}$. We will prove $\Delta^t\leqslant\frac{\lambda}{t+\beta}$ where $\lambda=\max\left\{\frac{\alpha^2\chi}{\alpha\mu-1},\beta\Delta^0\right\}$ by induction as follows.

Firstly, the inequality naturally holds for $t=0$ according to the definition of λ.

Then, assume that the inequality holds for some $t>0$, it follows that,

$$\begin{aligned}
&\Delta^{t+1}\leqslant\left(1-\mu\eta^t\right)\Delta^t+\left(\eta^t\right)^2X^t\leqslant\left(1-\tfrac{\alpha\mu}{t+\beta}\right)\tfrac{\lambda}{t+\beta}\\
&+\tfrac{\alpha^2X^t}{(t+\beta)^2}=\tfrac{t+\beta-1}{(t+\beta)^2}\lambda+\underbrace{\left[\frac{\alpha^2X^t}{(t+\beta)^2}-\frac{\alpha\mu-1}{(t+\beta)^2}\lambda\right]}_{\leqslant 0}\\
&\leqslant\tfrac{t+\beta-1}{(t+\beta)^2-1}\lambda\leqslant\tfrac{\lambda}{(t+1)+\beta}.
\end{aligned}\tag{23}$$

Specifically, if we choose $\alpha=\frac{2}{\mu}$ and $\beta=2\frac{L}{\mu}$, then $\eta^t=\frac{2}{\mu t+2L}$. Then, we have

$$\begin{aligned}
&\lambda=\max\left\{\tfrac{\alpha^2\chi}{\alpha\mu-1},\beta\Delta^0\right\}\\
&\leqslant\tfrac{\alpha^2\chi}{\alpha\mu-1}+\beta\Delta^0=\tfrac{4\chi}{\mu^2}+\tfrac{2L}{\mu}\left\|\tilde{\boldsymbol{w}}^0-\boldsymbol{w}^*\right\|_2^2.
\end{aligned}\tag{24}$$

Finally, we complete the proof as $\mathbb{E}[F(\tilde{\boldsymbol{w}}^t)]-F^*\overset{(a)}{\leqslant}\frac{L}{2}\left\|\tilde{\boldsymbol{w}}^t-\boldsymbol{w}^*\right\|_2^2=\frac{L}{2}\frac{\lambda}{t+\beta}\overset{(b)}{\leqslant}\frac{L}{\mu t+2L}\left(\frac{2\chi}{\mu}+L\left\|\tilde{\boldsymbol{w}}^0-\boldsymbol{w}^*\right\|_2^2\right)$, where (a) comes from the L-smoothness of $F(\cdot)$ and the fact that $\nabla F\left(\boldsymbol{w}^*\right)=0$, and (b) comes from (24). □

REFERENCES

[1] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6g communications: Challenges, methods, and future directions," *China Communications*, vol. 17, no. 9, pp. 105–118, 2020.

[2] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. Vincent Poor, "Convergence of federated learning over a noisy downlink," *Early access in IEEE Transactions on Wireless Communications*, 2021.

[3] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *arXiv preprint arXiv:2101.02198*, 2021.

[4] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3452–3464, 2020.

[5] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 170–185, 2020.

[6] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2897–2911, 2020.

[7] W. Y. B. Lim, J. S. Ng, Z. Xiong, D. Niyato, C. Miao, and D. I. Kim, "Dynamic edge association and resource allocation in self-organizing hierarchical federated learning networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3640–3653, 2021.

[8] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2020.

[9] 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 Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 359–375, 2021.

[10] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[11] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2604–2609.

[12] B. Hasırcıoğlu and D. Gündüz, "Private wireless federated learning with anonymous over-the-air computation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5195–5199.

[13] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2019.

[14] Y. Luo, J. Xu, W. Xu, and K. Wang, "Sliding differential evolution scheduling for federated learning in bandwidth-limited networks," *IEEE Communications Letters*, vol. 25, no. 2, pp. 503–507, 2020.

[15] M. P. Uddin, Y. Xiang, J. Yearwood, and L. Gao, "Robust federated averaging via outlier pruning," *IEEE Signal Processing Letters*, early access, 2021.

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