Weighted Over-the-Air Federated Learning

Seyed Mohammad Azimi-Abarghouyi*, Leandros Tassiulas†, and Carlo Fischione*

*KTH Royal Institute of Technology, Stockholm, Sweden, {seyaa, carlofi}@kth.se
†Yale University, New Haven, CT USA, leandros.tassiulas@yale.edu

***Abstract*—This paper introduces a new federated learning scheme that leverages over-the-air computation. The novel feature of this scheme is the proposal to employ adaptive weights during aggregation, as opposed to predefined weights in existing over-the-air schemes. This can mitigate the impact of wireless channel conditions on learning performance, without needing channel state information at transmitter side (CSIT). We derive convergence bound for the proposed scheme, supplemented with design insights. Accordingly, we propose an aggregation selection problem and develop an efficient algorithm to solve it, yielding optimized weights for the aggregation. Finally, through numerical experiments, we validate the effectiveness of the proposed scheme. Even with the challenges posed by channel conditions and device heterogeneity, the proposed scheme significantly surpasses other over-the-air schemes, including the one with CSIT.**

***Index Terms*—Federated learning, wireless multiple access channel, over-the-air computation, analog communications.**

I. Introduction

As wireless edge devices grow in prevalence, training a global model from their diverse data is essential. However, transferring data to a central server is impractical due to latency, power, bandwidth, and privacy concerns. Federated learning (FL) addresses this by enabling on-device machine learning without data transfer [2]. In FL, model training occurs locally on each device, with iterative model updates and aggregation at a parameter server until convergence. Conducted over resource-constrained and unreliable wireless networks, FL faces significant communication challenges as devices and the server share the same wireless medium. Traditional methods, using digital communications and orthogonal multiple access, separate communication from computation, requiring each device to transmit individually. While this avoids interference, it results in high communication latency and significant resource demands [3].

Over-the-air computation [4] is a promising scheme based on analog communications that leverages the superposition property of wireless channels, allowing simultaneous multiple access transmissions from edge devices within a single resource block. Based on this scheme, an approach known as over-the-air FL is proposed to perform aggregation under the interference, by merging communication and computation. Over-the-air FL can function with remarkably fewer resources and lower latency compared to FL using orthogonal transmissions [3]. However, the aggregation process in over-the-air FL is prone to estimation errors due to the severe effects of wireless channels.

Most prior research on over-the-air FL has assumed that perfect channel state information at the transmitter side (CSIT) is required for all devices. By employing a power allocation strategy for channel compensation, this information helps adjust transmission powers and phases to correct the misalignment between wireless channels and predefined aggregation weights, thereby minimizing estimation errors [5]–[15]. However, in addition to the need for extra on-device hardware for precise channel adjustments, this approach imposes a significant burden on channel estimation training and feedback mechanisms for each device. This results in increased latency before each transmission and a notable reduction in both spectral and energy efficiency. Additionally, poor channel conditions may either prevent a device from contributing to the learning process or require it to use substantial transmission power. Thus, this approach presents certain challenges when applied to a large number of devices, particularly those with limited power and hardware capabilities. The common strategy to power allocation is truncated power allocation, where each transmitter only needs to know its own channel [5]–[11]. Another strategy involves joint device selection and power allocation schemes [12]–[15]. These studies require global knowledge of all channels at every device before each transmission to centralize the allocation process. At its core, the device selection aims to include the maximum number of devices in each communication round, as the aggregation selection objective, ensuring that an estimation cost term remains beneath a tolerable threshold.

Many current wireless systems transmit blindly using a constant power. Indeed, apart from eliminating the need for CSIT, there are several advantages to transmitting without explicitly compensating for the channel. Primarily, it facilitates maintaining the average transmission energy of the signal regardless of the channel conditions. Also, it prevents the enlargement of the dynamic range of the transmitted signal without any adaptation for channel compensations, thereby significantly reducing the complexity of hardware implementation requirements. Lastly, efforts to correct the channel at

This paper is a conference version of [1], focusing on a special case that accounts for a uniform batch size and homogeneous computational capabilities across devices.

the transmitter could be compromised by channel estimation errors, leading to values at the receiver that are multiplied by unpredictable gains [16]. Taking these into account, the idea of a blind over-the-air FL approach has become increasingly prominent. The studies in [16], [17] delve into blind over-the-air FL without any compensation for the detrimental effects of wireless channels. In [18]–[20], blind over-the-air schemes are presented that leverage multiple antennas and rely on channel state information at the receiver side (CSIR), resulting in significantly higher performance. Nonetheless, these schemes necessitate large enough multiple receiver antennas.

In this work, we propose a novel over-the-air FL scheme named weighted over-the-air FL (`WAFeL`) to counteract the negative impact of wireless channels on the learning convergence performance by leveraging adaptive aggregation weights. Importantly, `WAFeL` operates as a blind scheme, eliminating the need for CSIT. Without channel compensation at the transmitter end, this approach has been shown to effectively mitigate aggregation estimation error caused by varying wireless channel conditions across devices, even when using a single-antenna server. This sets it apart from other blind schemes that either forgo compensation altogether or need an extensive number of antennas for the same. In summary, our proposal offers the following major contributions.

- We propose a generalized approach to aggregation, `WAFeL`, which differs from the conventional method of using predefined weights, such as equal or proportional to local dataset sizes. The seminal paper on FL [2] originally presented this aggregation method assuming ideal transmission circumstances and perfect aggregation estimation. However, subsequent research on over-the-air FL has continued to use the same aggregation method, regardless of interference and transmission imperfections [5]–[19]. In contrast, our work employs adaptive aggregation weights to mitigate aggregation estimation error and its effects, while a desirable learning performance is ensured.
- In order to accomplish the proposed weighted aggregation, we propose a new receiver architecture at the server side that incorporates both the real and imaginary parts of the signal, along with an equalization vector. This is then optimized with the objective of reducing the mean squared error (MSE) of the estimation to a minimum.
- Based on a basic set of broadly accepted principles, we analyze the convergence rate of `WAFeL` for any given aggregation weights. Our analysis reveals that equal aggregation weights are optimal when there is no estimation error. We then formulate an aggregation selection problem with a unique objective and constraint derived from the error term in the convergence analysis. This formulation integrates both communication and learning factors—specifically the MSE and mismatch terms—to identify optimal aggregation weights that adapt to system conditions in each communication round. This supports our integrated approach to communication and learning. Additionally, we propose an efficient algorithm for solving the selection problem, with a complexity that scales as K^3, where K represents the number of devices.
- Our experimental findings show that the aggregation weights designed by `WAFeL` exhibit significantly improved learning accuracy when compared to other over-the-air schemes, specifically about 15% over the CSIT-equipped scheme [5] and 30% over the non-CSIT scheme [16]. Notably, `WAFeL` achieves this performance without requiring CSIT, which makes it highly promising. Moreover, the learning performance closely approximates the ideal scenario of error-free orthogonal transmission.

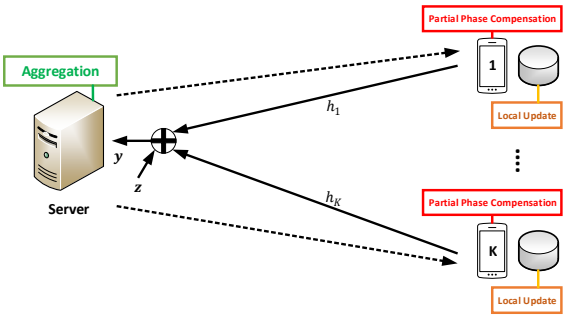

Fig. 1: Over-the-air FL system. The phase compensation is imperfect with accuracy range of $[0, \frac{\pi}{2})$.

Notation: All vectors are in column form. For vector $\mathbf{a}$ (boldface), a_i is its i-th component, and $\|\mathbf{a}\| = \sqrt{\mathbf{a}^\top \mathbf{a}}$ its Euclidean norm. $\mathbf{I}_n$ is the $n \times n$ identity matrix. $\mathbf{1}$ is the all one vector. $\mathbf{a} \odot \mathbf{b}$ is the Hadamard product of vectors $\mathbf{a}$ and $\mathbf{b}$.

II. System Model

In this section, we present the wireless setup for FL and outline its core algorithmic principles.

A. Setup

There are K devices and a single server as the basic setup for FL systems, as shown in Fig. 1. All nodes are equipped with a single antenna, reflecting the widespread use and broad applicability of single-antenna units in wireless systems. This underscores their importance as the fundamental scenario for FL. The downlink channels from the server to the devices are considered error-free, consistent with [5]–[19]. The uplink channel from each device k to the server at communication round t is modeled by $h_{k,t} = |h_{k,t}|e^{\angle h_{k,t}} \in \mathbb{C}$, where $|h_{k,t}|$ is the channel gain and $\angle h_{k,t}$ is the channel phase. The same transmission power constraint P is considered for all the devices. However, asymmetric power constraints can be incorporated by scaling the channel coefficients appropriately. Let the entire channel vector $\mathbf{h}_t = [h_{1,t}, \ldots, h_{K,t}]^\top$. The server is the only node that knows $\mathbf{h}_t$ as the CSIR *after the transmission*, while each device k knows only a partial estimation of its channel phase $\angle h_{k,t}$ with an accuracy range of $[0, \frac{\pi}{2})$. The purpose of such a *quadrant phase estimation* with a wide range of inaccuracy, accounting for a quarter of the entire possible range, is to enable each device to adjust its

phase so that all channels are observed as positive at the server. This ensures that the channels do not alter the sign of the transmitted data. Such partial estimation is also considered in studies like [16], [17]. Therefore, perfect fine synchronization is not needed, thanks to the acceptable range of uncertainty in channel phase. It is worth noting that many real-world wireless systems have CSIR and phase estimation. Contrastingly, the majority of over-the-air FL schemes [5]–[15] necessitate CSIT for all devices *before every transmission*. This includes precise values of $|h_{k,t}|$ and $\angle h_{k,t}$ for each device k and mandates perfect fine synchronization to fully counteract the channels.

Under our central assumption, we either lack knowledge of the channel gain or the information we do have contains substantial inaccuracies, thus complicating its effective use. Instead, our methodology gravitates towards the partial channel phase, providing a more reliable and consistent foundation for our proposed scheme in Section III. This approach achieves consistent average signal energy across varying channel conditions, minimizes the dynamic range of the signal, and simplifies the complexity of device hardware.

B. Learning Algorithm

Device $k \in \{1, \ldots, K\}$ possesses its own local (private) dataset $\mathcal{D}_k$. The learning model is parametrized by the parameter vector $\mathbf{w} = [w_1, \ldots, w_s]^\top \in \mathbb{R}^{s\times 1}$, where s is the model size. Then, the local loss function of the model vector $\mathbf{w}$ on $\mathcal{D}_k$ is

$$F_k(\mathbf{w}) = \frac{1}{D_k} \sum_{\xi_i \in \mathcal{D}_k} \ell(\mathbf{w}, \xi_i), \tag{1}$$

where $D_k = |\mathcal{D}_k|$ is the size of the dataset and the function $\ell(\mathbf{w}, \xi_i)$ represents the sample-wise loss, measuring the prediction error of $\mathbf{w}$ on the sample ξ_i. Following this, the global loss function applied to all distributed datasets, denoted as $\cup_{k=1}^K \mathcal{D}_k$, is

$$F(\mathbf{w}) = \frac{1}{\sum_{k=1}^K D_k} \sum_{k=1}^K D_k F_k(\mathbf{w}). \tag{2}$$

The goal of the training procedure is to discover an optimal parameter vector $\mathbf{w}$ that minimizes $F(\mathbf{w})$, expressed as

$$\mathbf{w}^* = \min_{\mathbf{w}} F(\mathbf{w}). \tag{3}$$

A widely used FL algorithm for solving (3) is `FedAvg` [2], which is outlined below.

Consider a particular round $t \in \{0, \ldots, T-1\}$, where T denotes the number of rounds. In this round, each device k first updates its own learning model via τ local training epochs, each based on a randomly sampled mini-batch $\boldsymbol{\xi}_k^i$ with size B drawn from $\mathcal{D}_k$, as

$$\mathbf{w}_{k,t,i+1} = \mathbf{w}_{k,t,i} - \mu_t \nabla F_k(\mathbf{w}_{k,t,i}, \boldsymbol{\xi}_k^i), \forall i \in \{0, \ldots, \tau-1\}, \tag{4}$$

where μ_t is the learning rate at the round t. Then, each device k uploads the local model $\mathbf{w}_{k,t} = \mathbf{w}_{k,t,\tau}$ to the server for aggregation. As the ideal aggregation, the global gradient can be obtained as an average of model parameters from all the devices with equal weighting[0], as

$$\mathbf{w}_{\mathrm{G},t+1} = \frac{1}{K} \sum_{k=1}^K \mathbf{w}_{k,t}. \tag{5}$$

Next, the server broadcasts the obtained global model $\mathbf{w}_{\mathrm{G},t+1}$ to the devices, based on which each device k updates its initial state for the next round as $\mathbf{w}_{k,t+1,0} = \mathbf{w}_{\mathrm{G},t+1}$.

III. `WAFeL`: Weighted Over-the-Air Scheme

The `WAFeL` constructs a form of the aggregation based on the additive nature of wireless multiple-access channels. This scheme instead of the ideal aggregation in (5) generally has the goal to recover a weighted aggregation as

$$\mathbf{w}_{\mathrm{G},t+1} = \sum_{k=1}^K \alpha_{k,t} \mathbf{w}_{k,t}, \tag{6}$$

where $\alpha_{k,t} \geq 0$ is the weight of device k in the aggregation at round t, such that $\sum_{k=1}^K \alpha_{k,t} = 1$. Let $\boldsymbol{\alpha}_t = [\alpha_{1,t}, \ldots, \alpha_{K,t}]^\top$ be the weight vector.

Unlike the common over-the-air FL schemes in [5]–[15], which rely on power allocation with enforced average and maximum power constraints and inherently involve device selection, `WAFeL` allows every device to contribute to the learning process without such limitations. Instead, each device is assigned an aggregation weight that is specifically designed based on its individual learning and communication conditions. The `WAFeL` comprises two principal components: the transmission scheme deployed at the devices and the aggregation scheme implemented at the server, as follows.[1]

A. Transmission Scheme

The model parameters at each device are normalized before transmission to have zero mean and unit variance. Normalizing the parameters offers two benefits. First, when the parameters have zero-mean entries, the estimate obtained in the sequel is unbiased. Second, when the entries have unit variance, the power of the received signal and the estimation error do not depend on the specific values of the model parameters.

The local model parameter vector at a device k is normalized as $\bar{\mathbf{w}}_k = (\mathbf{w}_k - \mu_k \mathbf{1})/\sigma_k$, where μ_k and σ_k denote the mean and standard deviation of the s entries of the model vector given by

$$\mu_k = \frac{1}{s} \sum_{i=1}^s w_{k,i}, \ \sigma_k^2 = \frac{1}{s} \sum_{i=1}^s (w_{k,i} - \mu_k)^2. \tag{7}$$

Each device k shares the two scalars (μ_k, σ_k) to the server in an error-free manner. This information is however negligible compared to the model parameters.

Subsequently, after the partial phase correction, each device transmits its normalized model parameter vector by scaling it with $\sqrt{P}$, resulting in a transmission power of P.

[0]This results from using a consistent batch size B across all devices. It is also worth noting that most over-the-air schemes [5]–[19] are based on equal aggregation weights.

[1]In this section, we ignore the iteration index for simplicity of presentation.

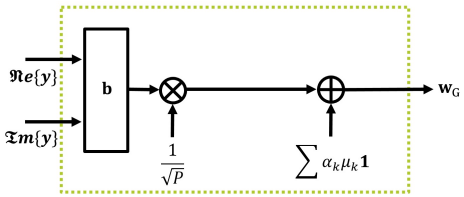

Fig. 2: Receiver architecture at the server.

B. Aggregation Scheme

After simultaneous transmission of all the devices within a single resource block, the baseband received signal $\mathbf{y} \in \mathbb{C}^{s\times 1}$ at the server is

$$\mathbf{y} = \sum_{k=1}^{K} \sqrt{P} h_k \bar{\mathbf{w}}_k + \mathbf{z}, \tag{8}$$

where $\mathbf{z} \in \mathbb{C}^{s\times 1}$ is the additive white Gaussian noise (AWGN), where each entry has variance σ_z^2. Then, (8) has the following real-valued representation

$$\mathbf{Y} = \sqrt{P}\mathbf{H}\bar{\mathbf{W}} + \mathbf{Z}, \tag{9}$$

where $\bar{\mathbf{W}} = [\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_K]^\top$ and

$$\mathbf{Y} = \begin{bmatrix} \mathfrak{Re}\{\mathbf{y}^\top\} \\ \mathfrak{Im}\{\mathbf{y}^\top\} \end{bmatrix}, \ \mathbf{H} = \begin{bmatrix} \mathfrak{Re}\{\mathbf{h}^\top\} \\ \mathfrak{Im}\{\mathbf{h}^\top\} \end{bmatrix}, \ \mathbf{Z} = \begin{bmatrix} \mathfrak{Re}\{\mathbf{z}^\top\} \\ \mathfrak{Im}\{\mathbf{z}^\top\} \end{bmatrix}.$$

Directly from (9), the server estimates the aggregation (6) as

$$\hat{\mathbf{w}}_G^\top = \frac{1}{\sqrt{P}}\mathbf{b}^\top\mathbf{Y} + \sum_{k=1}^{K} \alpha_k \mu_k \mathbf{1}^\top, \tag{10}$$

where $\mathbf{b} \in \mathbb{R}^{2\times 1}$ is employed as an equalization vector. The rationalization is that (6) is recovered from $\frac{1}{\sqrt{P}}\mathbf{b}^\top\mathbf{Y}$. However, due to its zero mean resulting from the transmit normalization, the appropriate mean $\sum_{k=1}^{K} \alpha_k \mu_k \mathbf{1}^\top$ is added to ensure accurate recovery, leading to an unbiased estimation. Fig. 2 provides a schematic representation of the receiver architecture designed for this aggregation process.

We can rewrite $\hat{\mathbf{w}}_G^\top$ as

$$\sum_{k=1}^{K} \alpha_k \mathbf{w}_k^\top + \frac{1}{\sqrt{P}}\mathbf{b}^\top\mathbf{Y} - \sum_{k=1}^{K}\left(\alpha_k \mathbf{w}_k^\top - \alpha_k \mu_k \mathbf{1}^\top\right) = \sum_{k=1}^{K} \alpha_k \mathbf{w}_k^\top + \mathbf{b}^\top\mathbf{H}\bar{\mathbf{W}} + \frac{1}{\sqrt{P}}\mathbf{b}^\top\mathbf{Z} - \sum_{k=1}^{K} \alpha_k \sigma_k \bar{\mathbf{w}}_k^\top. \tag{11}$$

Thus, the estimation error in recovering the weighted aggregation (6) is measured in terms of the MSE as

$$\begin{aligned} &\mathrm{MSE}(\boldsymbol{\alpha}) = \\ &\mathbb{E}\left\{\left\|\mathbf{b}^\top\mathbf{H}\bar{\mathbf{W}} + \frac{1}{\sqrt{P}}\mathbf{b}^\top\mathbf{Z} - (\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top\bar{\mathbf{W}}\right\|^2\right\} \\ &= \left(\|\mathbf{b}^\top\mathbf{H} - (\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top\|^2 + \frac{1}{\mathrm{SNR}}\|\mathbf{b}\|^2\right)s, \end{aligned} \tag{12}$$

where the independence of $\bar{\mathbf{w}}_k$ and $\bar{\mathbf{w}}_{k'}$, $\forall k \neq k'$, is assumed, similar to other over-the-air FL works [9]–[15], [18], [19]. In (12), $\mathrm{SNR} = P/\sigma_z^2$ is the signal-to-noise ratio. The vector $\mathbf{b}$ that minimizes the estimation MSE (12) is presented in the next theorem.

Theorem 1: The optimal equalization vector for a given $\boldsymbol{\alpha}$ is

$$\mathbf{b}_{\text{opt}}^\top = (\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top\mathbf{H}^\top\left(\frac{1}{\mathrm{SNR}}\mathbf{I}_2 + \mathbf{H}\mathbf{H}^\top\right)^{-1}. \tag{13}$$

Proof: Expanding

$$\begin{aligned} &\|\mathbf{b}^\top\mathbf{H} - (\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top\|^2 + \frac{1}{\mathrm{SNR}}\|\mathbf{b}\|^2 = \left(\mathbf{b}^\top\mathbf{H} - (\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top\right) \\ &\times\left(\mathbf{H}^\top\mathbf{b} - \boldsymbol{\alpha}\odot\boldsymbol{\sigma}\right) + \frac{1}{\mathrm{SNR}}\|\mathbf{b}\|^2 = \mathbf{b}^\top\mathbf{H}\mathbf{H}^\top\mathbf{b} \\ &- 2\mathbf{b}^\top\mathbf{H}(\boldsymbol{\alpha}\odot\boldsymbol{\sigma}) + (\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top(\boldsymbol{\alpha}\odot\boldsymbol{\sigma}) + \frac{1}{\mathrm{SNR}}\mathbf{b}^\top\mathbf{b}, \end{aligned} \tag{14}$$

and taking derivative from the result with respect to $\mathbf{b}$, we obtain $2\mathbf{H}\mathbf{H}^\top\mathbf{b} - 2\mathbf{H}(\boldsymbol{\alpha}\odot\boldsymbol{\sigma}) + \frac{2}{\mathrm{SNR}}\mathbf{b}$, which is equal to zero at (13). ∎

Replacing $\mathbf{b}_{\text{opt}}$ in (12), we obtain

$$\begin{aligned} &\mathrm{MSE}(\boldsymbol{\alpha}) = s(\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top\times \\ &\left(\mathbf{I}_K - \mathbf{H}^\top\left(\frac{1}{\mathrm{SNR}}\mathbf{I}_2 + \mathbf{H}\mathbf{H}^\top\right)^{-1}\mathbf{H}\right)(\boldsymbol{\alpha}\odot\boldsymbol{\sigma}), \end{aligned} \tag{15}$$

which, using the matrix inversion lemma [21], can be written as

$$\begin{aligned} &\mathrm{MSE}(\boldsymbol{\alpha}) = s(\boldsymbol{\alpha}\odot\boldsymbol{\sigma})^\top\left(\mathbf{I}_K + \mathrm{SNR}\mathbf{H}^\top\mathbf{H}\right)^{-1}\times \\ &(\boldsymbol{\alpha}\odot\boldsymbol{\sigma}) = s\boldsymbol{\alpha}^\top\mathrm{diag}(\boldsymbol{\sigma})\left(\mathbf{I}_K + \mathrm{SNR}\mathbf{H}^\top\mathbf{H}\right)^{-1}\times \\ &\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{\alpha}. \end{aligned} \tag{16}$$

IV. Convergence Analysis

The analysis of `WAFeL` in terms of the convergence rate is presented in the following theorem. Notably, our analysis is based on the minimal set of assumptions commonly found in the literature as

Assumption 1 (Lipschitz-Continuous Gradient): The gradient of the loss function $F(\mathbf{w})$, as represented in (2), is characterized by Lipschitz continuity with a non-negative constant $L > 0$. This implies that for any pair of model vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, we have

$$F(\mathbf{w}_2) \leq F(\mathbf{w}_1) + \nabla F(\mathbf{w}_1)^T(\mathbf{w}_2 - \mathbf{w}_1) + \frac{L}{2}\|\mathbf{w}_2 - \mathbf{w}_1\|^2, \tag{17}$$

$$\|\nabla F(\mathbf{w}_2) - \nabla F(\mathbf{w}_1)\| \leq L\|\mathbf{w}_2 - \mathbf{w}_1\|. \tag{18}$$

Assumption 2 (Gradient Variance Bound): The local stochastic gradient estimate for device k at $\mathbf{w}_k$, using a mini-batch $\boldsymbol{\xi}_k$ with size B, is an unbiased estimate of the ground-truth gradient $\nabla F(\mathbf{w}_k)$ with bounded variance

$$\mathbb{E}\left\{\|\nabla F_k(\mathbf{w}_k, \boldsymbol{\xi}_k) - \nabla F(\mathbf{w}_k)\|^2\right\} \leq \frac{\sigma_g^2}{B}. \tag{19}$$

Theorem 2: Let $1-\frac{L^2\eta^2}{2}\tau(\tau-1)-L\eta\tau\geq 0$ and $\boldsymbol{\alpha}_t$ as the weight vector for each round $t\in\{0,\ldots,T-1\}$, then the convergence rate of WAFeL under Assumptions 1 and 2 is

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\{\|\nabla F(\mathbf{w}_{\mathrm{G},t})\|^2\right\}\leq\frac{2\left(F(\mathbf{w}_{\mathrm{G},0})-F^*\right)}{\eta\tau T}+L^2\eta^2\frac{\tau-1}{2}\frac{\sigma_{\mathrm{g}}^2}{B}+\frac{L}{\eta\tau T}\sum_{t=0}^{T-1}\mathcal{I}_t(\boldsymbol{\alpha}_t),\tag{20}$$

where $F^*=F(\mathbf{w}^*)$ comes from the problem (3) as the optimal loss value, and

$$\mathcal{I}_t(\boldsymbol{\alpha}_t)=\eta^2\frac{\sigma_{\mathrm{g}}^2}{B}\tau\|\boldsymbol{\alpha}_t\|^2+\mathrm{MSE}_t(\boldsymbol{\alpha}_t),\tag{21}$$

where MSE_t is given in (16).

Proof: The proof is based on [1, Appendix A], where $\mathbf{b}_{\mathrm{s}}=\frac{1}{B}\mathbf{1}$, followed by simplifications. ∎

Remark 1: The error term $\mathcal{I}_t(\boldsymbol{\alpha}_t)$ in the convergence rate incorporates learning and communication factors such as the estimation error measured by MSE including wireless channels and SNR, the aggregation weights, the learning rate, the batch size, the number of local steps, and the gradient variance bound.

From Theorem 2, the convergence rate of FedAvg in Subsection II.B—under error-free transmission and aggregation (5)—serves as a specific case of WAFeL and is presented below.

Corollary 1: Let $1-\frac{L^2\eta^2}{2}\tau(\tau-1)-L\eta\tau\geq 0$ and $\boldsymbol{\alpha}_t=\frac{1}{K}\mathbf{1}$ for each round $t\in\{0,\ldots,T-1\}$. Under error-free transmission, the convergence rate is

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\{\|\nabla F(\mathbf{w}_{\mathrm{G},t})\|^2\right\}\leq\frac{2\left(F(\mathbf{w}_{\mathrm{G},0})-F^*\right)}{\eta\tau T}+L^2\eta^2\frac{\tau-1}{2}\frac{\sigma_{\mathrm{g}}^2}{B}+\frac{L}{\eta\tau T}\sum_{t=0}^{T-1}\mathcal{I},\tag{22}$$

where

$$\mathcal{I}=\eta^2\frac{\sigma_{\mathrm{g}}^2}{B}\tau\frac{1}{K}.\tag{23}$$

Remark 2: The effect of errors in WAFeL is an increase in the error term $\mathcal{I}$ by $\eta^2\frac{\sigma_{\mathrm{g}}^2}{B}\tau\left(\|\boldsymbol{\alpha}_t\|^2-\frac{1}{K}\right)+\mathrm{MSE}_t(\boldsymbol{\alpha}_t)$ for each round t. This increase is due not only to the MSE term resulting from estimation error, but also from the mismatch between the weight vector $\boldsymbol{\alpha}_t$ and $\frac{1}{K}\mathbf{1}$, the latter being used for the ideal aggregation in (5).

Remark 3: The mismatch term $\|\boldsymbol{\alpha}_t\|^2-\frac{1}{K}$ is related to the learning facet of WAFeL and the MSE term $\mathrm{MSE}_t(\boldsymbol{\alpha}_t)$ is due to the communication facet of WAFeL.

Remark 4: The mismatch term $\|\boldsymbol{\alpha}_t\|^2-\frac{1}{K}$ is scaled by $\eta^2\frac{\sigma_{\mathrm{g}}^2}{B}\tau$. This indicates that the effect of the mismatch is reinforced by an increase in the local training parameters η and τ.

Remark 5: When there is no estimation error, the minimum of the error term $\mathcal{I}_t(\boldsymbol{\alpha}_t)$ happens at $\boldsymbol{\alpha}_t=\frac{1}{K}\mathbf{1}$, as the mismatch term become zero. That is why conventionally the ideal aggregation in the the primary FedAvg in [2] and the literature aligns with (5). It is demonstrated here that in the presence of estimation error, the optimal aggregation that minimizes the term $\mathcal{I}_t(\boldsymbol{\alpha}_t)$ deviates from the ideal aggregation. This is further discussed in Section V.

V. Aggregation Selection

To select the weight vectors $\boldsymbol{\alpha}_t,\forall t\in\{0,\ldots,T-1\}$, the goal is to minimize the error term $\mathcal{I}_t(\boldsymbol{\alpha}_t)$ in the convergence rate described in Theorem 2, as

$$\boldsymbol{\alpha}_t=\arg\min_{\boldsymbol{\alpha}\geq\mathbf{0}\setminus\{\mathbf{0}\},\ \mathbf{1}^\top\boldsymbol{\alpha}=1}\mathcal{I}_t(\boldsymbol{\alpha}).\tag{24}$$

This approach is inspired by *Remark 5*. However, given that the precise determination of the gradient variance bound σ_{g}^2 and the Lipschitz constant L in (21) necessitates an understanding of gradient data statistics—which is unavailable in many applications—the optimal evaluation of (24) becomes impractical.[2] Therefore, in pursuit of our universality goal, we introduce an alternative aggregation selection problem based on the key components of $\mathcal{I}_t(\boldsymbol{\alpha}_t)$, specifically the MSE and mismatch terms, which does not depend on any prior knowledge. The proposed formulation treats the mismatch term as the primary objective to be minimized, while the MSE of the estimation error is constrained by a predefined threshold th. Drawing inspiration from *Remark 3*, this approach favors the learning aspect over the communication aspect of WAFeL, as learning is the intended goal of FL. The resulting problem is defined as follows.

$$\boldsymbol{\alpha}_t=\arg\min_{\boldsymbol{\alpha}\geq\mathbf{0}\setminus\{\mathbf{0}\}}\|\boldsymbol{\alpha}\|^2,\tag{25}$$

subject to

$$\boldsymbol{\alpha}^\top\mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K+\mathrm{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha}\leq\mathrm{th},$$
$$\mathbf{1}^\top\boldsymbol{\alpha}=1.$$

The problem (25) is convex. Using the dual Lagrangian method [22], it can be transformed to

$$\max_{\lambda>0}\min_{\alpha\setminus\{\mathbf{0}\}}\Big\{\mathcal{L}(\lambda,\boldsymbol{\alpha})=\boldsymbol{\alpha}^\top\boldsymbol{\alpha}+\lambda\boldsymbol{\alpha}^\top\mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K+\mathrm{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha}=\boldsymbol{\alpha}^\top\times\left(\mathbf{I}_K+\lambda\mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K+\mathrm{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t)\right)\boldsymbol{\alpha}\Big\},\tag{26}$$

subject to $\mathbf{1}^\top\boldsymbol{\alpha}=1$. In (26), the following problem

$$g(\lambda)=\min_{\alpha\setminus\{\mathbf{0}\}}\boldsymbol{\alpha}^\top\times\left(\mathbf{I}_K+\lambda\mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K+\mathrm{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t)\right)\boldsymbol{\alpha},\tag{27}$$

subject to $\mathbf{1}^\top\boldsymbol{\alpha}=1$, has a closed-form solution, which is obtained in the next theorem. Before that, let us define

$$\mathbf{G}_t(\lambda)=\mathbf{I}_K+\lambda\mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K+\mathrm{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t).\tag{28}$$

[2]Most FL studies likewise do not assume such knowledge [5]–[19].

Algorithm 1 Aggregation Selection

Initialize $\lambda^{(0)}$ and $\boldsymbol{\alpha}^{(0)} = \frac{1}{K}\mathbf{1}$

Iterate

Update $\boldsymbol{\alpha}^{(j)}$ as in (29) with $\lambda^{(j-1)}$.

Update $\lambda^{(j)}$ as $\lambda^{(j)} = \lambda^{(j-1)} + t\left(\boldsymbol{\alpha}^{(j)\top}\text{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K + \text{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\text{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha}^{(j)} - \text{th}\right)$.

Until $\left|\lambda\left(\boldsymbol{\alpha}^\top\text{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K + \text{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\text{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha} - \text{th}\right)\right| \leq \epsilon$.

Theorem 3: The optimized weight vector, as the solution to (27), is

$$\boldsymbol{\alpha} = \frac{\mathbf{G}_t(\lambda)^{-1}\mathbf{1}}{\mathbf{1}^\top\mathbf{G}_t(\lambda)^{-1}\mathbf{1}}, \tag{29}$$

Proof: We use the Karush-Kuhn-Tucker (KKT) conditions [22] to solve (27). The corresponding Lagrangian function is as follows

$$\mathcal{L}_c(\xi, \boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top\mathbf{G}_t(\lambda)\boldsymbol{\alpha} + \xi\boldsymbol{\alpha}^\top\mathbf{1}, \tag{30}$$

which, after taking derivative, leads to

$$2\mathbf{G}_t(\lambda)\boldsymbol{\alpha} + \xi\mathbf{1} = \mathbf{0}, \tag{31}$$

and then

$$\boldsymbol{\alpha} = -\frac{\xi}{2}\mathbf{G}_t^{-1}(\lambda)\mathbf{1}. \tag{32}$$

To meet the condition $\mathbf{1}^\top\boldsymbol{\alpha} = 1$, ξ is obtained as

$$\xi = -\frac{2}{\mathbf{1}^\top\mathbf{G}_t^{-1}(\lambda)\mathbf{1}} < 0, \tag{33}$$

which, when replaced in (32), completes the proof. Since $\mathbf{G}_t(\lambda)$ is a positive definite matrix with all non-negative entries due to the phase compensation, its inverse $\mathbf{G}_t^{-1}(\lambda)$ is also positive definite with all non-negative entries. Consequently, all the elements of $\mathbf{G}_t^{-1}(\lambda)\mathbf{1}$ are non-negative, leading to the conclusion that all the elements of the optimal $\boldsymbol{\alpha}$ are non-negative, as defined in (6). ■

Now, the remaining work is to maximize $g(\lambda)$ with respect to nonnegative λ. We use the subgradient method [22] to find a solution. The subgradient direction of the function $g_1(\lambda)$ is determined as $\boldsymbol{\alpha}^\top\text{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K + \text{SNR}\mathbf{H}_t^\top\mathbf{H}_t\right)^{-1}\text{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha} - \text{th}$, where $\boldsymbol{\alpha}$ is given in (29) with the fixed λ. Subsequently, the solution for problem (25) can be searched in an iterative fashion summarized in Algorithm 1.

In Algorithm 1, calculating the inverse of a $K \times K$ matrix, as required by (29), has a computational complexity of $\mathcal{O}(K^3)$. Therefore, for a maximum number of iterations $n_{\max}$, the total complexity is $\mathcal{O}(n_{\max}K^3)$.

VI. Experimental Results

The learning task is the classification on the standard MNIST dataset. The classifier model is implemented using a convolutional neural network (CNN), which consists of two 3×3 convolution layers with ReLU activation (the first with 32 channels, the second with 64), each followed by a 2×2 max pooling; a fully connected layer with 128 units and ReLU activation; and a final softmax output layer. We take into account both i.i.d. and non-i.i.d. distribution of dataset samples across the devices. In the non-i.i.d. scenario, each device holds samples from only two classes, and the quantity of samples varies among devices. The performance is measured as the learning accuracy with reference to the test dataset over global iteration count t. Each performance result is evaluated as the average of 20 realization samples to account for Gaussian channel distribution, i.e., the channel gain with the Rayleigh fading $\sim$ exp(1) and the channel phase (after the phase compensation) with the uniform distribution $\sim \mathcal{U}(0, \frac{\pi}{2})$.

TABLE I: Parameter Values

K	SNR	τ	μ
30	10	3	0.01

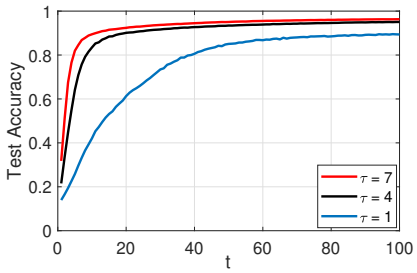

Fig. 3: i.i.d. case.

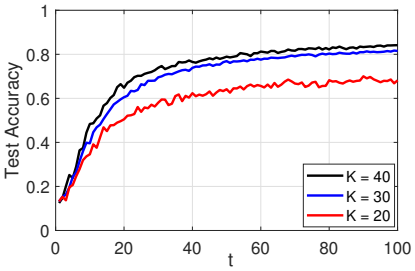

Fig. 4: non-i.i.d. case.

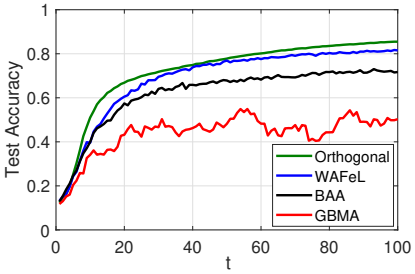

Fig. 5: non-i.i.d. case.

Fig. 3 shows the accuracy for different local iterations τ in the i.i.d. scenario. As observed, increasing τ or t improves the learning performance. It further exhibits a marked improvement when integrating multiple local iterations in comparison to using just a single local iteration.

In Fig. 4, the accuracy is shown for different numbers of devices K in the non-i.i.d. scenario. The performance improves as K increases because more devices participates

in the learning process. However, this improvement comes at the cost of increased MSE, resulting in a tradeoff. Specifically, for higher values of K, this performance improvement tends to diminish.

In Fig. 5, we compare WAFeL with well-known benchmark schemes from the literature in the non-i.i.d. scenario. Among the benchmarks using analog over-the-air computation similar to WAFeL are the BAA scheme, which uses truncated power allocation with accurate knowledge of CSIT [5], and the GBMA scheme, which compensates for phase only at the transmitters [16]. For both BAA and GBMA, perfect fine synchronization is considered. Additionally, the ideal error-free performance is considered using FL with digital orthogonal transmissions, requiring at least $K = 30$ times more resource blocks. For a consistent comparison, FedAvg with the same $\tau = 3$ is implemented across all the schemes. While the BAA scheme possesses a distinct complexity level owing to its need for perfect CSIT and channel compensation requirements, the blind GBMA method lacks any optimization algorithm or processing, setting it apart from WAFeL which incorporates an algorithm of complexity $\mathcal{O}(K^3)$.

It is observed that WAFeL significantly outperforms both the BAA and GBMA schemes. For example, at $t = 100$, the improvement of WAFeL is around 15% and 30% compared to BAA and GBMA, respectively. Moreover, it closely approximates the performance achieved by the orthogonal case. Surprisingly, WAFeL, which incorporates partial phase compensation, outperforms schemes like BAA that have perfect gain and phase compensation. This performance improvement is attributed to the utilization of optimized adaptive weights in the aggregation process, which was not previously identified, coupled with the unique receiver structure. It is noteworthy that in contrast to BAA where each device needs to adhere to both average and maximum power constraints for power allocation and there is an inherent device selection, the weight allocation in WAFeL is not subjected to any limitations and all the devices contribute to the learning.

VII. Conclusions

In this paper, we introduced a unique over-the-air federated learning scheme incorporating a novel weighted aggregation approach. In the scheme, the adaptive choice of aggregation weights helps counteract the impacts of wireless channels on performance, eliminating the need for channel compensation at the transmitting end. We analyzed the convergence rate of the learning process for the scheme as a function of the aggregation weights, which also includes both communication and learning factors. To select the aggregation weights, we proposed an aggregation selection problem based on the analysis. We also presented an efficient algorithm to solve the problem. Despite channel conditions and device heterogeneity, experimental results demonstrated the high learning accuracy of the proposed scheme, surpassing existing solutions even the one with channel compensation at transmitters. Moreover, the proposed scheme can closely attain the level of performance that would be expected in the absence of any errors.

Acknowledgment

This project has received funding from the European Union's Horizon Europe programme under grant agreement No. 101137954. The authors would like to thank the rest of the BATTwin consortium for supporting this research.

References

[1] S. M. Azimi-Abarghouyi and L. Tassiulas, "Over-the-air federated learning via weighted aggregation," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18240-18253, Dec. 2024.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," *AISTATS*, pp. 1273-1282, 2017.

[3] H. Hellstrom, J. M. B. da Silva Jr, M. M. Amiri, M. Chen, V. Fodor, H. V. Poor, and C. Fischione, "Wireless for machine learning: A survey," *Foundations and Trends@ in Signal Processing*, vol. 15, no. 4, pp. 290-399, 2022.

[4] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498-3516, Oct. 2007.

[5] 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, Jan. 2020.

[6] M. Mohammadi Amiri and D. Gunduz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546-3557, May 2020.

[7] G. Zhu, Y. Du, D. Gunduz, 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, Mar. 2021.

[8] Z. Lin, X. Li, V. K. N. Lau, Y. Gong, and K. Huang, "Deploying federated learning in large-scale cellular networks: Spatial convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1542-1556, Mar. 2022.

[9] X. Cao, G. Zhu, J. Xu, and S. Cui, "Transmission power control for over-the-air federated averaging at network edge," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1571-1586, May. 2022.

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

[11] S. M. Azimi-Abarghouyi and V. Fodor, "Scalable hierarchical over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8480-8496, Aug. 2024.

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

[13] A. Bereyhi, A. Vagollari, S. Asaad, R. R. Muller, W. Gerstacker, and H. V. Poor, "Device scheduling in over-the-air federated learning via matching pursuit," *IEEE Trans. Signal Process.*, vol. 71, pp. 2188-2203, June 2023.

[14] Z. Wang, et al., "Federated learning via intelligent reflecting surface," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 808-822, Feb. 2022.

[15] M. Kim, A. L. Swindlehurst, and D. Park, "Beamforming vector design and device selection in over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7464-7477, Nov. 2023.

[16] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Sig. Process.*, vol. 68, pp. 2897-2911, 2020.

[17] R. Paul, Y. Friedman, and K. Cohen, "Accelerated gradient descent learning over multiple access fading channels," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 532-547, Feb. 2022.

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

[19] S. M. Azimi-Abarghouyi and L. R. Varshney, "Federated learning via lattice joint source-channel coding," *IEEE Int. Symp. Inf. Theory (ISIT)*, Athens, Greece, July 2024.

[20] S. M. Azimi-Abarghouyi and L. R. Varshney, "Compute-update federated learning: A lattice coding approach," *IEEE Trans. Signal Process.*, vol. 72, pp. 5213-5227, Nov. 2024.

[21] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Rev.*, vol. 23, no. 1, pp. 53-60, Jan. 1981.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.