Power Minimization in Federated Learning with Over-the-air Aggregation and Receiver Beamforming

Faeze Moradi Kalarde
faeze.moradi@mail.utoronto.ca
University of Toronto
Toronto, ON, Canada

Ben Liang
liang@ece.utoronto.ca
University of Toronto
Toronto, ON, Canada

Min Dong
min.dong@ontariotechu.ca
Ontario Tech University
Oshawa, ON, Canada

Yahia A. Eldemerdash Ahmed
yahia.ahmed@ericsson.com
Ericsson Canada
Ottawa, ON, Canada

Ho Ting Cheng
ho.ting.cheng@ericsson.com
Ericsson Canada
Ottawa, ON, Canada

ABSTRACT

Combining over-the-air uplink transmission and multi-antenna beamforming can improve the efficiency of federated learning (FL). However, to mitigate the significant aggregation error due to communication noise and signal distortion, pre-processing of device signals and post-processing at the server are required. In this paper, we study the optimization of receiver beamforming and device transmit weights in over-the-air FL, to minimize the total transmit power in each communication round while guaranteeing the convergence of FL. We establish sufficient convergence conditions based on the analysis of gradient descent with error and formulate a power minimization problem. An alternating optimization approach is then employed to decompose the problem into tractable subproblems, and efficient solutions are developed for these subproblems. Our proposed method is evaluated through simulation on standard image classification tasks, demonstrating its effectiveness in achieving substantial reductions in transmit power compared with existing alternatives.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Distributed computing methodologies**.

KEYWORDS

Federated Learning, Over-the-air Computation, Power Consumption, Multi-antenna Beamforming

ACM Reference Format:
Faeze Moradi Kalarde, Ben Liang, Min Dong, Yahia A. Eldemerdash Ahmed, and Ho Ting Cheng. 2023. Power Minimization in Federated Learning with Over-the-air Aggregation and Receiver Beamforming. In *Proceedings of the Int'l ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems (MSWiM '23), October 30-November 3, 2023, Montreal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3616388.3617534

MSWiM '23, October 30-November 3, 2023, Montreal, QC, Canada

ACM ISBN 979-8-4007-0366-9/23/10…$15.00
https://doi.org/10.1145/3616388.3617534

1 INTRODUCTION

Federated learning (FL) leverages the computational capabilities of edge devices without necessitating the transmission of their private training datasets. It operates as an iterative algorithm where, during each iteration (i.e., communication round), each device computes the gradient of its local loss function and sends it to a central server for aggregation. However, the signal transmission between the edge devices and the server can impose significant stress on communication resources, especially in scenarios involving a substantial number of devices and limited available bandwidth. In such scenarios, conventional orthogonal multiple access methods may not suffice to enable the transmission of updates from the devices. To mitigate the burden of communication overhead, a method known as analog (i.e., over-the-air) aggregation has emerged. This approach involves the simultaneous analog transmission of individual models by edge devices over a shared wireless uplink channel, enabling natural model summation through superposition. This over-the-air computation technique has gained increasing attention due to its efficient utilization of bandwidth and reduced communication latency in contrast to conventional transmission techniques over orthogonal channels [2, 3, 8, 24, 30].

Over-the-air computation is susceptible to significant aggregation error, which propagates over the FL iterations. The aggregation error arises from two primary sources: noise and channel distortion. Hence, pre-processing of the devices' signal prior to transmission, as well as post-processing of the received signal at the server, is necessary to mitigate the inconsistencies between the actual received signal at the server and the desired signal. Furthermore, transmit power minimization plays a critical role in facilitating efficient system operation, as it reduces device energy usage and also minimizes the interference to outside receivers. Careful design of transmit weights for individual devices and receiver beamforming is necessary to minimize the transmit power while also guaranteeing convergence of FL to an optimal model.

A considerable number of works in FL focus on minimizing or bounding power consumption by considering both communication and computation factors [1, 4, 9, 12, 18, 20, 29, 31, 34]. However, these works only concern FL with digital communication. Their techniques do not apply to our problem. Several recent studies have

addressed power efficiency in FL with over-the-air aggregation. Both [36] and [25] employed truncated channel inversion to devise transmit weights for devices. The device selection process in both studies relies on a threshold for channel strength to guarantee adherence to the average transmit-power constraint for each individual device. The authors of [30] aimed to minimize the cumulative training loss while considering individual long-term transmit-power constraints at the mobile devices, by designing the transmit weights based on channel inversion. The authors of [10] aimed to minimize the global loss optimality gap by jointly designing the devices transmit weights and server power while considering constraints on either the individual or total uplink transmit power of devices during each communication round. The authors of [27] introduced an online energy-aware dynamic worker scheduling policy. This policy was designed to optimize the average number of workers scheduled, taking into consideration a long-term energy constraint. The energy-aware dynamic device scheduling algorithm in [28] aims to optimize training performance within energy constraints, accounting for both communication and computation energy factors. However, all these works have focused on scenarios where the server operates with a single antenna, thereby overlooking the design of receiver beamforming techniques.

When the server is equipped with multiple antennas, it was shown in [13, 35] that beamforming techniques can be employed to reduce the impact of noise and channel distortion in over-the-air computation. These studies optimized the receiver beamforming with the objective of reducing the mean squared error (MSE) while ensuring that the average transmit power for each device is bounded. It was shown in [33] that the method in [13] can be applied to improve FL performance. Both [32] and [14] studied the joint optimization of receiver beamforming and device selection to maximize the number of selected devices while ensuring both a bounded MSE and bounded average transmit power. In [19], for a reconfigurable intelligent surface (RIS)-assisted system, receiver beamforming, device selection, and RIS phase shift were optimized to increase the convergence rate. However, none of the mentioned works in this category aimed to minimize power consumption.

In contrast to the existing works, the objective of this study is to optimize both the multi-antenna receiver beamforming and the transmit weights of devices in FL with over-the-air aggregation. This naturally includes device selection since assigning zero transmit weight to a device is equivalent to deselecting the device. We aim to minimize the average total transmit power of devices while ensuring the convergence of FL to an optimal point. Specifically, our contribution can be summarized as follows:

- We consider distributed gradient descent in the presence of error and derive a set of sufficient conditions for convergence of FL to the optimal model. These conditions account for the impact of over-the-air aggregation, receiver beamforming, and device transmit weights on the discrepancy between the signal used by the server for model updates and the global loss gradient. Our analysis substantially differs from existing works on FL over noisy communication with beamforming design, since it ensures the convergence of FL to the optimum, whereas existing works utilize upper bounds on the optimality gap that generally do not diminish to zero.
- We formulate an optimization problem to minimize the total device transmit power, with the derived convergence conditions serving as constraints. An alternating optimization approach is used to solve the resulting bi-convex problem. Most importantly, we show how to transform each subproblem into a convex quadratic programming form that can be solved efficiently.
- We experiment with FL for image classification over a simulated wireless network. We show that for a wide range of parameter settings, the proposed method achieves significantly reduced transmit power compared with state-of-the-art benchmarks.

The rest of this paper is organized as follows. In Section 2, we present the system model and problem formulation. Section 3 describes the transmit and receiver beamforming design. Simulation results and conclusion are provided in Sections 4 and 5 respectively.

2 SYSTEM MODEL AND PROBLEM FORMULATION

2.1 FL System

We consider a wireless network comprising a central server and M edge devices. Each device, denoted by index m, contains a local training dataset of size K_m represented by $\mathcal{D}_m = \{(\mathbf{x}_{m,k}, y_{m,k}) : 1 \leq k \leq K_m\}$, where $\mathbf{x}_{m,k}$ is the k-th data feature vector and $y_{m,k}$ is its corresponding label. The aim of the edge devices is to cooperatively train a global model on the server, capable of predicting the true labels of data feature vectors for all devices while ensuring the privacy of their local datasets. We define the empirical local training loss function for device m as follows:

$$F_m(\mathbf{w}; \mathcal{D}_m) \triangleq \frac{1}{K_m} \sum_{k=1}^{K_m} l(\mathbf{w}; \mathbf{x}_{m,k}, y_{m,k}), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^D$ is the global model parameter vector and $l(\cdot)$ is the sample-wise training loss associated with each data sample. Then the global training loss function is

$$F(\mathbf{w}) = \frac{1}{K} \sum_{m=1}^{M} K_m F_m(\mathbf{w}; \mathcal{D}_m), \tag{2}$$

where $K = \sum_m K_m$ is the total number of training samples over all devices. In this study, we adopt the conventional Federated Stochastic Gradient Descent (FedSGD) technique for iterative model training in FL, where the server updates the model parameters using an aggregation of the gradients derived from the local loss functions of all devices [21]. The main objective of this study is to determine the optimal global model $\mathbf{w}^*$ that minimizes the global training loss function $F(\mathbf{w})$. We refer to each cycle of the algorithm as a communication round. In the t-th communication round, the following operations are executed:

(1) **Downlink phase:** The server broadcasts the model parameter vector $\mathbf{w}_t$ to all devices.
(2) **Gradient computation:** Each device m computes the gradient of its local loss function, given by $\mathbf{g}_{m,t} \triangleq \nabla F_m(\mathbf{w}_t; \mathcal{D}_m)$, where $\nabla F_m(\mathbf{w}_t; \mathcal{D}_m)$ is the gradient of $F_m(\cdot)$ at $\mathbf{w}_t$.
(3) **Uplink phase:** The devices transmit their local gradients to the server via the uplink wireless channels.

(4) **Model updating:** The server computes a weighted aggregation of the local gradients to update the global model. In an ideal scenario where the local gradients can be received accurately at the server, $\mathbf{r}_t \triangleq \sum_{m=1}^{M} K_m \mathbf{g}_{m,t}$ is utilized to update $\mathbf{w}_t$. However, in practical settings, only an approximation $\hat{\mathbf{r}}_t$ is feasible at the server due to the effects of wireless channels and noise. Therefore, the server updates the global model as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma_t \frac{\mathcal{R}(\hat{\mathbf{r}}_t)}{\sum_{m=1}^{M} K_m}, \tag{3}$$

where γ_t is the learning rate in round t and $\mathcal{R}(\cdot)$ returns the real part of a complex variable.

2.2 FL with Over-the-Air Analog Aggregation

We assume that each device is equipped with a single antenna, while the server is equipped with N antennas. The wireless uplink channel between device m and the server during communication round t is represented by the complex-valued vector $\mathbf{h}_{m,t} \in \mathbb{C}^N$. We assume that the server has perfect knowledge of $\mathbf{h}_{m,t}$ at the beginning of each round t.

We utilize over-the-air computation for efficient aggregation of the local gradients at the server. This is achieved through analog aggregation over multiple access channels, as proposed in [36]. In each communication round t, the devices send their local gradients to the server simultaneously using the same frequency resource. The D entries of the local gradient of each device are transmitted over D time slots. Specifically, in time slot d, each selected device m sends the d-th entry of its local gradient, denoted by $g_{m,t}[d]$, which is normalized by $v_{m,t} = \frac{\|\mathbf{g}_{m,t}\|}{\sqrt{D}}$ and then adjusted by the transmit weight $a_{m,t} \in \mathbb{C}$. The transmitted signal by device m in round t is denoted by $\mathbf{z}_{m,t}$ and its d-th entry is defined as

$$z_{m,t}[d] = a_{m,t} \frac{g_{m,t}[d]}{v_{m,t}}, \tag{4}$$

which implies the average transmit power is $\mathbb{E}[|z_{m,t}[d]|^2] = |a_{m,t}|^2$. In this work, we assume the average transmit power of each device is bounded by P_0, i.e., $|a_{m,t}|^2 \leq P_0$, $\forall m, \forall t$. The corresponding received signal at the server in round t and in time slot d is denoted by $\mathbf{y}_{d,t}$ and is given by

$$\mathbf{y}_{d,t} = \sum_{m=1}^{M} \mathbf{h}_{m,t} z_{m,t}[d] + \mathbf{n}_{d,t}, \tag{5}$$

where $\mathbf{n}_{d,t} \in \mathbb{C}^N$ is the Circularly Symmetric Complex Gaussian (CSCG) noise vector, i.e., $\mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ and is independently and identically distributed (i.i.d.) over t and d. Each device m also sends $v_{m,t}$ to the server in each communication round. As $v_{m,t}$ is only a scalar, we assume it is sent over a separate digital channel and is received by the server perfectly.

The server applies receiver beamforming to process the received signal. Let $\mathbf{f}_t \in \mathbb{C}^N$ denote the receiver beamforming vector at round t. The post-processed received signal in the t-th communication round and in the d-th time slot is given by

$$\hat{r}_t[d] = \mathbf{f}_t^H \mathbf{y}_{d,t} = \sum_{m=1}^{M} \mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t} \frac{g_{m,t}[d]}{v_{m,t}} + \mathbf{f}_t^H \mathbf{n}_{d,t}. \tag{6}$$

The server uses $\hat{\mathbf{r}}_t \triangleq [\hat{r}_t[1], ..., \hat{r}_t[D]]^T$ to update $\mathbf{w}_t$ based on (3).

2.3 Problem Formulation

We aim to minimize the total uplink transmit power over all devices during each communication round while guaranteeing the convergence of the model to the optimal model. Specifically, we design the transmit scalars, $a_{m,t}$ for all devices and receiver beamforming, $\mathbf{f}_t$ in each communication round t, such that convergence of $\mathbf{w}_t$ to an optimal model is guaranteed. This optimization problem in round t can be formulated as follows:

$$\min_{\mathbf{f}_t, \{a_{m,t}\}} \quad \sum_{m=1}^{M} |a_{m,t}|^2 \tag{7a}$$

$$\text{s.t.} \quad \text{Convergence of FL to optimum,} \tag{7b}$$

$$|a_{m,t}|^2 \leq P_0, \ \forall m. \tag{7c}$$

Note that even though we do not consider device selection explicitly in our problem formulation, problem (7) does capture device selection since if a device's transmit weight is set to zero, it transmits nothing to the server.

3 BEAMFORMING DESIGN FOR POWER MINIMIZATION

In problem (7), constraint (7b) is not an explicit function of optimization variables $\{\mathbf{f}_t\}$ and $\{a_{m,t}\}$, so solving this problem is challenging. We must first analyze the convergence of the gradient descent method in the presence of errors and beamforming.

3.1 Training Convergence Analysis

We rewrite the global model update at the server in (3) as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma_t \, \mathbf{s}_t, \tag{8}$$

where $\mathbf{s}_t \triangleq \frac{\mathcal{R}(\hat{\mathbf{r}}_t)}{K}$. We may equivalently rewrite $\mathbf{s}_t$ as follows:

$$\mathbf{s}_t = \nabla F(\mathbf{w}_t) + \mathbf{e}_t, \tag{9}$$

where $\nabla F(\mathbf{w}_t)$ is the gradient of the global loss function at $\mathbf{w}_t$ which is the desired information, and $\mathbf{e}_t$ is the error vector. Based on (9), the expression for $\mathbf{e}_t$ can be written as

$$\begin{aligned}
\mathbf{e}_t &= \mathbf{s}_t - \nabla F(\mathbf{w}_t) \\
&= \frac{\mathcal{R}(\hat{\mathbf{r}}_t)}{K} - \frac{\sum_{m=1}^{M} K_m \mathbf{g}_{m,t}}{K} \\
&= \frac{1}{K} \sum_{m} \left(\frac{\mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]}{v_{m,t}} - K_m \right) \mathbf{g}_{m,t} + \frac{1}{K} \mathcal{R} \begin{pmatrix} \mathbf{f}_t^H \mathbf{n}_{1,t} \\ \vdots \\ \mathbf{f}_t^H \mathbf{n}_{D,t} \end{pmatrix},
\end{aligned} \tag{10}$$

where the second equality comes from the fact that based on (2), $\nabla F(\mathbf{w}_t) = \sum_{m=1}^{M} \frac{K_m}{K} \mathbf{g}_{m,t}$, and the third equality follows the definition of $\hat{\mathbf{r}}_t$ in (6).

We consider the following assumptions on the global loss function, which are common in the stochastic optimization literature [6, 23].

A1: $F(\mathbf{w})$ is differentiable and its minimizer is denoted by $\mathbf{w}^*$.

A2: $F(\mathbf{w})$ is μ-strongly convex, i.e., $\exists \mu > 0, \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^D$:

$$F(\mathbf{w}) \geq F(\mathbf{w}') + (\mathbf{w} - \mathbf{w}')^T \nabla F(\mathbf{w}') + \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}'\|_2^2. \tag{11}$$

A3: $\nabla F(\mathbf{w})$ is L-Lipschitz continuous, i.e., $\exists L > 0, \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^D$:

$$\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|. \tag{12}$$

Theorem 1. *Suppose **A1**, **A2**, and **A3** are satisfied and the following conditions hold:*

C1: $\|\mathbb{E}[\ \mathbf{e}_t|\mathbf{w}_t\ ]\| \leq \alpha\|\nabla F(\mathbf{w}_t)\|$, *for some* $0 \leq \alpha < 1, \forall t$.
C2: $\mathbb{E}[\ \|\mathbf{e}_t\|^2|\mathbf{w}_t\ ] \leq \delta\|\nabla F(\mathbf{w}_t)\|^2 + \beta$, *for some* $\delta, \beta \geq 0, \forall t$.
C3: $\gamma_t \geq 0,\ \Sigma_{t=0}^{\infty}\gamma_t = \infty,\ \lim_{t\to\infty}\gamma_t \to 0$.

Then, for any initial point $\mathbf{w}_0$, *the expected optimality gap converges to zero, i.e.,* $\lim_{t\to\infty}\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \to 0$.

Proof. See Appendix A. □

*Remark 1: One way to satisfy **C3** is by choosing a constant learning rate during the early communication rounds, i.e., ($\gamma_t = \gamma, 0 \leq t \leq T$) and setting γ_t as a harmonic series for the rest of the rounds, i.e., ($\gamma_t = \frac{\gamma}{t}, t \geq T$).*

3.2 Problem Reformulation

Based on Theorem 1, we replace constraint (7b) by **C1** and **C2** and rewrite problem (7) as follows:

$$\min_{\mathbf{f}_t, \{a_{m,t}\}} \sum_{m=1}^{M} |a_{m,t}|^2 \tag{13a}$$
$$\text{s.t.} \quad \|\mathbb{E}[\ \mathbf{e}_t|\mathbf{w}_t\ ]\| \leq \alpha\|\nabla F(\mathbf{w}_t)\|, \tag{13b}$$
$$\mathbb{E}[\ \|\mathbf{e}_t\|^2|\mathbf{w}_t\ ] \leq \delta\|\nabla F(\mathbf{w}_t)\|^2 + \beta, \tag{13c}$$
$$|a_{m,t}|^2 \leq P_0,\ \forall m. \tag{13d}$$

In communication round t, the server aims to solve problem (13) centrally to determine the receiver beamforming and transmit weights of all devices. However, the LHS of (13b) and (13c) are functions of $\{\mathbf{g}_{m,t}\}$, which are unknown to the server. To overcome this challenge, we replace the LHS of (13b) and (13c) by their upper bounds provided by the following lemmas:

Lemma 1. *Given* $\mathbf{w}_t$, *the norm of expected error is bounded by*

$$\|\mathbb{E}[\ \mathbf{e}_t|\mathbf{w}_t\ ]\| \leq \frac{\sqrt{D}}{K}\sum_m |K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]|. \tag{14}$$

Proof. See Appendix B. □

Lemma 2. *Given* $\mathbf{w}_t$, *the expected error norm squared is bounded by*

$$\mathbb{E}[\|\mathbf{e}_t\|^2|\mathbf{w}_t] \leq \frac{D}{K^2}\Big(\sum_m |K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]|\Big)^2 + \frac{D\sigma_n^2\|\mathbf{f}_t\|^2}{2K^2} \tag{15}$$

Proof. See Appendix C. □

Now, $\|\nabla F(\mathbf{w}_t)\|$, on the RHS of (13b) and (13c), is still unknown to the server. The server can approximate it by

$$\|\nabla F(\mathbf{w}_t)\| = \|\sum_{m=1}^{M}\frac{K_m}{K}\mathbf{g}_{m,t}\| \overset{(a)}{\leq} \sum_{m=1}^{M}\frac{K_m}{K}\|\mathbf{g}_{m,t}\| \overset{(b)}{=} \sum_{m=1}^{M}\frac{K_m}{K}\sqrt{D}v_{m,t}\,, \tag{16}$$

where (a) is obtained by the Triangle Inequality and (b) follows the definition of $v_{m,t}$. Note that if the local gradients $\{\mathbf{g}_{m,t}\}_{m=1}^M$ have the same direction, the inequality (a) can be replaced by equality and therefore, (16) is exact. In situations where the local gradients are similar (such as i.i.d. data distribution over devices), (16) is a close approximation for $\|\nabla F(\mathbf{w}_t)\|$. We denote this approximation by V_t and thus, the server solves the following optimization problem in each round:

$$\min_{\mathbf{f}_t, \{a_{m,t}\}} \quad \sum_{m=1}^{M} |a_{m,t}|^2 \tag{17a}$$
$$\text{s.t.} \quad \frac{\sqrt{D}}{K}\sum_m |K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]| \leq \alpha V_t, \tag{17b}$$
$$\frac{D}{K^2}\Big(\sum_m |K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]|\Big)^2 + \frac{D\sigma_n^2\|\mathbf{f}_t\|^2}{2K^2} \leq \delta V_t^2 + \beta, \tag{17c}$$
$$|a_{m,t}|^2 \leq P_0,\ \forall m. \tag{17d}$$

Remark 2: Note that, even though we have used an approximation in problem (17), *for any V_t that satisfies $\frac{V_t}{\|\nabla F(\mathbf{w}_t)\|} < \frac{1}{\alpha}, \forall t$, given* (17b) *and* (17c), *there exists a set of values $\alpha' \geq \alpha$, $\beta' = \beta$, and $\delta' \geq \delta$ that satisfy **C1** and **C2**, so FL convergence to the optimum is ensured by any feasibility solution to problem* (17).

3.3 Proposed Alternating Optimization Approach

Despite having a convex objective function, problem (17) is non-convex due to the multiplication of optimization variables, $\mathbf{f}_t$ and $\{a_{m,t}\}$ in its constraints. However, given $\mathbf{f}_t$, constraints (17b) and (17c) are convex in $\{a_{m,t}\}$, and given $\{a_{m,t}\}$, they are convex in $\mathbf{f}_t$. Therefore, problem (17) is bi-convex and an alternating optimization approach can be used to find a partial optimum solution [7].

3.3.1 Optimizing transmit weights $\{a_{m,t}\}$. Since the LHS of (17b) and the first term of LHS of (17c) are identical, when $\mathbf{f}_t$ is given, (17b) and (17c) can be combined into a single constraint as follows:

$$\min_{\{a_{m,t}\}} \quad \sum_{m=1}^{M} |a_{m,t}|^2 \tag{18a}$$
$$\text{s.t.} \quad \frac{\sqrt{D}}{K}\sum_m |K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]| \leq \min\Big(\alpha V_t, \sqrt{\delta V_t^2 + \beta - \frac{D\sigma_n^2\|\mathbf{f}_t\|^2}{2K^2}}\Big) \tag{18b}$$
$$|a_{m,t}|^2 \leq P_0,\ \forall m. \tag{18c}$$

Lemma 3. *Denote the optimal solution of problem* (18a) *by $\{a_{m,t}^*\}$. Its phase, defined in $[0, 2\pi)$, satisfies*

$$\angle\, a_{m,t}^* = -\angle\, \mathbf{f}_t^H \mathbf{h}_{m,t}, \forall m. \tag{19}$$

Proof. See Appendix D. □

Let $b_{m,,t} \triangleq |a_{m,t}|$. We now need to solve (18a) only with respect to $\{b_{m,t}\}$:

$$\min_{\{b_{m,t}\}} \quad \sum_{m=1}^{M} b_{m,t}^2 \tag{20a}$$

$$\text{s.t.} \quad \frac{\sqrt{D}}{K} \sum_{m} |K_m v_{m,t} - |\mathbf{f}_t^H \mathbf{h}_{m,t}|\, b_{m,t}| \leq \min(\alpha V_t, \sqrt{\delta V_t^2 + \beta - \frac{D\sigma_n^2 \|\mathbf{f}_t\|^2}{2K^2}}) \tag{20b}$$

$$b_{m,t} \leq \sqrt{P_0}, \ \forall m. \tag{20c}$$

We further introduce M auxiliary variables $\mathbf{q} = [q_1, q_2, ..., q_M]^T$ to transform (20b) into $2M + 1$ equivalent linear constraints. The new problem is

$$\min_{\{b_{m,t}\},\mathbf{q}} \quad \sum_{m=1}^{M} b_{m,t}^2 \tag{21a}$$

$$\text{s.t.} \quad K_m v_{m,t} - |\mathbf{f}_t^H \mathbf{h}_{m,t}|\, b_{m,t} \leq q_m, \forall m, \tag{21b}$$

$$|\mathbf{f}_t^H \mathbf{h}_{m,t}|\, b_{m,t} - K_m v_{m,t} \leq q_m, \forall m, \tag{21c}$$

$$\frac{\sqrt{D}}{K} \sum_{m=1}^{M} q_m \leq \min(\alpha V_t, \sqrt{\delta V_t^2 + \beta - \frac{D\sigma_n^2 \|\mathbf{f}_t\|^2}{2K^2}}) \tag{21d}$$

$$b_{m,t} \leq \sqrt{P_0}, \ \forall m. \tag{21e}$$

This is a convex problem since the objective function is quadratic and the constraints are linear. It can be efficiently solved using quadratic programming techniques with standard solvers such as CVXPY.

3.3.2 Optimizing receiver beamforming $\mathbf{f}_t$. As the objective function in problem (17) is independent of $\mathbf{f}_t$, we minimize the LHS of constraint (17c), which represents the expected deviation from the true gradient. Thus, given $\{a_{m,t}\}$, we have

$$\min_{\mathbf{f}_t} \quad \frac{D}{K^2} (\sum_{m} |K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]|)^2 + \frac{D\sigma_n^2 \|\mathbf{f}_t\|^2}{2K^2} \tag{22a}$$

$$\text{s.t.} \quad \frac{\sqrt{D}}{K} \sum_{m} |K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]| \leq \alpha V_t. \tag{22b}$$

By again introducing M auxiliary variables $\mathbf{q} = [q_1, q_2, ..., q_M]^T$ to problem (22), in a form slightly different from problem (21) above, we have

$$\min_{\mathbf{f}_t,\mathbf{q}} \quad \frac{D}{K^2} (\|\mathbf{q}\|^2 + \frac{\sigma_n^2}{2} \|\mathbf{f}_t\|^2) \tag{23a}$$

$$\text{s.t.} \quad K_m v_{m,t} - \mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}] \leq q_m, \forall m, \tag{23b}$$

$$\mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}] - K_m v_{m,t} \leq q_m, \forall m, \tag{23c}$$

$$\frac{\sqrt{D}}{K} \sum_{m}^{M} q_m \leq \alpha V_t. \tag{23d}$$

This again is a convex problem with quadratic objective function and linear constraints, allowing an efficient numerical solution.

3.3.3 Finding an initial feasible point. To initiate the alternating optimization approach for problem (17) using our solutions in Sections 3.3.1 and 3.3.2, it is important to identify an appropriate initial feasible point. By assuming a given receiver beamforming vector $\mathbf{f}_t$, we can ensure that the left-hand side (LHS) of constraint (17b) and the first term of the LHS of constraint (17c) are set to zero by setting

$$a_{m,t} = \frac{K_m v_{m,t}}{\mathbf{f}_t^H \mathbf{h}_{m,t}}, \forall m. \tag{24}$$

This ensures the satisfaction of constraint (17b). In the subsequent step, our objective is to find some $\mathbf{f}_t$, that minimizes the LHS of constraint (17c) while simultaneously meeting constraint (17d). Therefore, we aim to solve the following optimization problem

$$\min_{\mathbf{f}_t} \quad \frac{D\sigma_n^2 \|\mathbf{f}_t\|^2}{2K^2} \tag{25a}$$

$$\text{s.t.} \quad K_m^2 v_{m,t}^2 \leq P_0 |\mathbf{f}_t^H \mathbf{h}_{m,t}|^2, \forall m. \tag{25b}$$

We observe that problem (25) is equivalent to the problem of quality-of-service single-group downlink multicast beamforming [5, 26]. In that problem, the base station (BS) aims to transmit a common message to all devices, with the objective of optimizing the multicast beamformer to minimize the transmit power while satisfying the SNR target for each device. Specifically, $\mathbf{f}_t$ represents the transmit beamformer, and $\frac{K_m^2 v_{m,t}^2}{P_0}$ denotes the SNR target for device m. Although the multicast beamforming problem is typically NP-hard, it can be addressed using the Successive Convex Approximation (SCA) method [5], which offers a convergence guarantee to a stationary point.

Upon solving (25), two scenarios arise. In the first case, if the optimal value of the objective function obtained through the SCA technique is less than $\delta V_t^2 + \beta$, the approach successfully identifies a feasible point. Conversely, if the optimal value exceeds $\delta V_t^2 + \beta$, the method fails to find a feasible point. This outcome may be attributed to either the inefficiency of the method or the infeasibility of (17). In that case, we arbitrarily select an initial point.

3.4 Optimality and Complexity Analysis

In the proposed alternating optimization approach to solve the biconvex problem (17), since the solutions in Sections 3.3.1 and 3.3.2 are optimal, we decrease the objective (17a) in each iteration. Furthermore, the constraints (17b), (17c), and (17d) are always satisfied. As a result, the proposed method is guaranteed to converge to a partial optimum solution [7].

In Section 3.3.1, a quadratic problem with $2M$ variables and $3M + 1$ constraints is solved. Thus, the computational complexity is $O(M^3)$. In Section 3.3.2, a quadratic problem with $M + N$ variables and $2M + 1$ constraints is solved. Therefore, the complexity is $O(\min(M + N, 2M + 1)^3) = O(M^3)$. In Section 3.3.3, the computational complexity of solving the feasibility problem by the SCA method is $O(\min(N, M)^3)$. Therefore, the overall computational complexity of each iteration is $O(M^3)$.

4 SIMULATION RESULTS

In this section, we evaluate the efficacy of our proposed method for image classification over MNIST [17] and CIFAR-10 [15] datasets,

with logistic regression and convolutional neural networks (CNN), respectively. Even though CNNs are non-convex, we will show that the proposed method remains effective in such an application.

We consider $M = 10$ devices and $N = 16$ antennas. The distance of device m from the parameter server d_m is sampled from a uniform distribution between 10 and 100 meters and the path loss follows the COST Hata model [16], i.e., PL[dB] = 139.1+35.22log(d_m[km]). The channel vector for device m is constant during the training and sampled from a Complex Normal distribution, i.e., $\mathbf{h}_{m,t} = \mathbf{h}_m \sim \mathcal{CN}(\mathbf{0}, \frac{1}{PL}\mathbf{I}_{N\times N})$. We assume $P_0 = 27$dBm and the power of noise σ_n^2, which may also account for external interference, changes in a range from -80dBm to -68dBm. For comparison, we consider two approaches as benchmarks:

(1) **Minimum Mean Squared Error (MMSE):** In each communication round the transmit weights and the receiver beamforming are optimized to minimize the MSE between the received signal at the server and the desired signal. Specifically $\{a_{m,t}\}$ are set based on [14, 19, 32] by zero forcing as (24). Thus, the MMSE problem reduces to the same problem as (25), and the same SCA method can be used to design $\mathbf{f}_t$.
(2) **Greedy Spatial Device Selection (GSDS)[22]:** Joint design of receiver beamforming and device selection is achieved by minimizing an upper bound given in [19] of the optimality gap. Since our work does not consider using RIS, we utilize a simplified method for this minimization problem, which we refer to as GSDS. In each iteration of the GSDS, the device with the highest channel alignment with the previously selected devices is added to the set, and the receiver beamforming is designed for the updated set using the SCA method discussed in [5]. Finally, the set of selected devices and its corresponding beamforming vector that results in the minimum value of the objective function is chosen.

4.1 Logistic Regression on MNIST Dataset

In the MNIST dataset, each data sample is a labeled grey-scaled handwritten digit image of size 28×28 pixels, i.e., $\mathbf{x}_k \in \mathbb{R}^{784}$, with a label $y_k \in \{0, 1, ..., 9\}$ to indicate its class. There are 60000 training and 10000 test samples. We consider training a multinomial logistic regression classifier with cross-entropy loss. The model parameters for each class consist of 784 weights and a bias term. We use a regularization term for the global loss function as $\frac{\mu}{2}\|\mathbf{w}\|^2$ where μ is the regularization constant and set to $\mu = 10^{-4}$.

An equal number of data samples from different classes are uniformly randomly distributed among the devices so that $K_m = 5420$. The learning rates for the MMSE method, GSDS method, and our proposed method have been tuned for optimal performance and set to 0.25, 0.25, and 0.95, respectively. All methods employ the full local batch for gradient computation. In our proposed method, the parameter values are set to $\delta = 6$, $\beta = 0$, and $\alpha = 0.5$ through hyperparameter tuning.

Figs. 1(a) and 1(b) respectively show the average power consumption per round and the number of rounds to reach a target test accuracy of 85%, under various noise power levels. Both figures exhibit averaged results over 20 channel and noise realizations, accompanied by 95% confidence intervals. They both demonstrate the significant advantages of the proposed method over MMSE and GSDS, as it can reduce the power consumption by a factor of more than 10^3 and simultaneously reduce the number of required rounds for convergence. Moreover, the proposed method is adaptive to changes in noise power. As the noise power increases, the proposed method adjusts the power consumption accordingly. In contrast, the benchmark methods use the full available power without considering the level of noise power.

Fig. 1(c) illustrates the total energy consumption over all communication rounds to achieve the target accuracy. We assume a symbol duration of 1μ second for the transmission of each gradient entry. This figure demonstrates that the proposed method provides significant savings over MMSE and GSDS, due to its reduced power requirement and faster convergence.

4.2 CNN on CIFAR-10 Dataset

In the CIFAR-10 dataset, each data sample consists of a colored image of size $3 \times 32 \times 32$ pixels, i.e., $\mathbf{x}_k \in \mathbb{R}^3 \times \mathbb{R}^{32} \times \mathbb{R}^{32}$ and a label $y_k \in \{0, 1, ..., 9\}$ which indicate the class of the image. There are 50000 training and 10000 test samples. As this dataset is more complex, we adopt a more advanced CNN model, namely the Residual Network (ResNet) with 14 layers (ResNet-14) [11], and utilize the cross-entropy loss function. During training, we use the data augmentation approach in [11], which includes padding the image with 4 pixels on each side and randomly sampling a 32×32 crop from either the padded image or its horizontal flip.

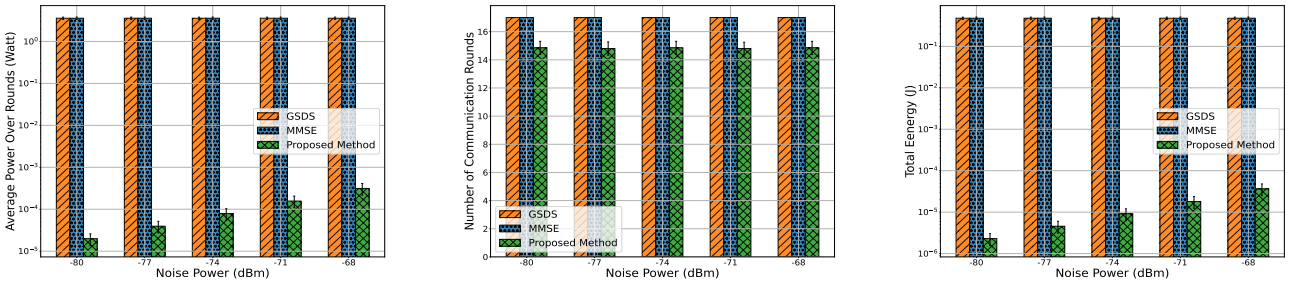
Similarly to the previous scenario, we evenly distribute each class of training samples among the devices, so that $K_m = 5000$. During each communication round, the devices calculate their local gradient using a batch of data of size 10 from their respective local dataset. The learning rate is set to 0.01, with SGD momentum 0.9 and weight decay 10^{-4}. For our proposed method, the parameter values are set to $\delta = 0.5$, $\beta = 0$, and $\alpha = 10^{-4}$ after conducting a grid search for hyperparameter tuning. The batch size and the learning rate of the MMSE and GSDS methods are set to 10 and 0.01 respectively after hyperparameter tunning.

Fig. 2(a) and Fig. 2(b) respectively show the average power consumption of devices per round and the number of communication rounds required to achieve 70% test accuracy under various noise power levels. The 95% confidence intervals are calculated based on 20 different channel and noise realizations. Even for non-convex CNN, the proposed method achieves a substantial reduction in average power consumption while incurring only a small increase in the number of communication rounds.

Fig. 2(c) depicts the total energy consumption over all communication rounds to achieve the target test accuracy, considering a symbol duration time of 1μ second for the transmission of each gradient entry. As this figure implies, despite the small increase in communication rounds, the proposed method is far superior in energy conservation, reducing the energy usage by 20 to 300 times.

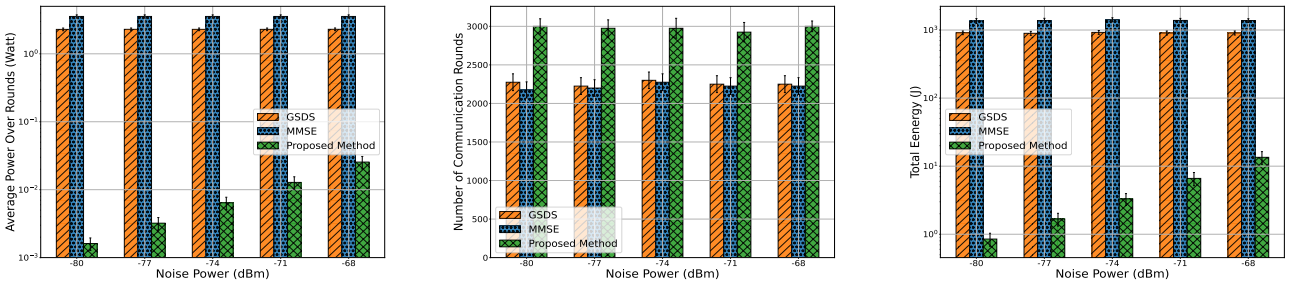
5 CONCLUSION

In this paper, our objective is to optimize the receiver beamforming and device transmit weights in an FL system with over-the-air aggregation, focusing on minimizing the average transmit power consumption while ensuring convergence of FL to an optimal point. We establish new sufficient conditions for FL convergence with analog



(a) (b) (c)

Figure 1: (a) Average transmit power over rounds, (b) number of communication rounds, (c) total energy over rounds for MNIST



(a) (b) (c)

Figure 2: (a) Average transmit power over rounds, (b) number of communication rounds, (c) total energy over rounds for CIFAR-10

transmission and beamforming, which we then use as constraints in formulating a power minimization problem. This problem is bi-convex and solved using an alternating optimization approach, where we transform each sub-problem into a convex quadratic programming form. Through simulation on wireless FL with various datasets and learning models, we demonstrate the effectiveness of our proposed method in achieving lower power consumption in comparison with the MMSE and GSDS benchmarks.

ACKNOWLEDGMENTS

This work was funded in part by Ericsson Canada and by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Abdullatif Albaseer, Mohamed Abdallah, Ala Al-Fuqaha, and Aiman Erbad. 2022. Fine-Grained Data Selection for Improved Energy Efficiency of Federated Edge Learning. *IEEE Transactions on Network Science and Engineering* 9, 5 (2022), 3258–3271.

[2] Mohammad Mohammadi Amiri and Deniz Gündüz. 2020. Federated learning over wireless fading channels. *IEEE Trans. Wireless Commun.* 19, 5 (2020), 3546–3557.

[3] Mohammad Mohammadi Amiri and Deniz Gündüz. 2020. Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air. *IEEE Trans. Signal Process.* 68 (2020), 2155–2169.

[4] Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H Vincent Poor, and Shuguang Cui. 2020. A joint learning and communications framework for federated learning over wireless networks. *IEEE Trans. Wireless Commun.* 20, 1 (2020), 269–283.

[5] Min Dong and Qiqi Wang. 2020. Multi-group multicast beamforming: Optimal structure and efficient algorithms. *IEEE Trans. Signal Process.* 68 (2020), 3738–3753.

[6] Michael P Friedlander and Mark Schmidt. 2012. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing* 34, 3 (2012), A1380–A1405.

[7] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research* 66, 3 (2007), 373–407.

[8] Huayan Guo, An Liu, and Vincent KN Lau. 2020. Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis. *IEEE Internet of Things Journal* 8, 1 (2020), 197–210.

[9] Kun Guo, Zihan Chen, Howard H. Yang, and Tony Q. S. Quek. 2022. Dynamic Scheduling for Heterogeneous Federated Learning in Private 5G Edge Networks. *IEEE Journal of Selected Topics in Signal Processing* 16, 1 (2022), 26–40.

[10] Wei Guo, Ran Li, Chuan Huang, Xiaoqi Qin, Kaiming Shen, and Wei Zhang. 2022. Joint Device Selection and Power Control for Wireless Federated Learning. *IEEE J. Sel. Areas Commun.* 40, 8 (Aug. 2022), 2395–2410.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[12] Yun Ji, Zhoubin Kou, Xiaoxiong Zhong, Hangfan Li, Fan Yang, and Sheng Zhang. 2022. Client Selection and Bandwidth Allocation for Federated Learning: An Online Optimization Perspective. In *Proc. IEEE Global Commun. Conf. (GLOBECOM).*

[13] Tao Jiang and Yuanming Shi. 2019. Over-the-air computation via intelligent reflecting surfaces. In *Proc. IEEE Global Commun. Conf. (GLOBECOM).*

[14] Minsik Kim, A. Lee Swindlehurst, and Daeyoung Park. 2023. Beamforming Vector Design and Device Selection in Over-the-Air Federated Learning. *IEEE Trans. Wireless Commun.* 54, 6 (2023), 2239–2251.

[15] Alex Krizhevsky and Geoffrey E. Hinton. 2009. *Learning multiple layers of features from tiny images.* Master's thesis. Univ. of Toronto (UofT), Toronto, Canada.

[16] Premchandra Kumar, Bhushan Patil, and Suraj Ram. 2015. Selection of radio propagation model for long-term evolution (LTE) network. *International Journal*

of Engineering Research and General Science 3, 1 (2015), 373–379.
[17] Yann LeCun, Corinna Cortes, and Christopher J. Burges. 1998. The MNIST Database of Handwritten Digits. *[Online]. Available: http://yann.lecun.com/exdb/mnist/* (1998).
[18] Yangchen Li, Ying Cui, and Vincent Lau. 2021. Optimization-Based GenQSGD for Federated Edge Learning. In *Proc. IEEE Global Commun. Conf. (GLOBECOM).*
[19] Hang Liu, Xiaojun Yuan, and Ying-Jun Angela Zhang. 2021. Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach. *IEEE Trans. Wireless Commun.* 20, 11 (2021), 7595–7609.
[20] Bing Luo, Xiang Li, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. 2021. Cost-Effective Federated Learning in Mobile Edge Networks. *IEEE Journal on Selected Areas in Communications* 39, 12 (2021), 3606–3621.
[21] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS).*
[22] Faeze Moradi Kalarde, Min Dong, Ben Liang, Ho Ting Cheng, and Yahia Ahmed. 2023. Joint beamforming and device selection in federated learning with over-the-air aggregation. *arXiv preprint arXiv:2302.14336* (2023).
[23] Boris T. Polyak. 1987. *Introduction to optimization.* Optimization Software.
[24] Tomer Sery and Kobi Cohen. 2020. On analog gradient descent learning over multiple access fading channels. *IEEE Trans. Signal Process.* 68 (2020), 2897–2911.
[25] Tomer Sery, Nir Shlezinger, Kobi Cohen, and Yonina C Eldar. 2021. Over-the-air federated learning from heterogeneous data. *IEEE Trans. Signal Process.* 69 (2021), 3796–3811.
[26] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo. 2006. Transmit beamforming for physical-layer multicasting. *IEEE Trans. Signal Process.* 54, 6 (Jun. 2006), 2239–2251.
[27] Yuxuan Sun, Sheng Zhou, and Deniz Gündüz. 2020. Energy-Aware Analog Aggregation for Federated Learning with Redundant Data. In *Proc. IEEE Int. Conf. on Communications (ICC).*
[28] Yuxuan Sun, Sheng Zhou, Zhisheng Niu, and Deniz Gündüz. 2022. Dynamic Scheduling for Over-the-Air Federated Edge Learning With Energy Constraints. *IEEE Journal on Selected Areas in Communications* 40, 1 (2022), 227–242. https://doi.org/10.1109/JSAC.2021.3126078
[29] Shuo Wan, Jiaxun Lu, Pingyi Fan, Yunfeng Shao, Chenghui Peng, and Khaled B. Letaief. 2021. Convergence Analysis and System Design for Federated Learning Over Wireless Networks. *IEEE Journal on Selected Areas in Communications* 39, 12 (2021), 3622–3639.
[30] Juncheng Wang, Min Dong, Ben Liang, Gary Boudreau, and Hatem Abou-zeid. 2022. Online Model Updating with Analog Aggregation in Wireless Edge Learning. In *Proc. IEEE Conf. on Computer Communications (INFOCOM).*
[31] Jie Xu and Heqiang Wang. 2021. Client Selection and Bandwidth Allocation in Wireless Federated Learning Networks: A Long-Term Perspective. *IEEE Transactions on Wireless Communications* 20, 2 (2021), 1188–1200.
[32] Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. 2020. Federated learning via over-the-air computation. *IEEE Trans. Wireless Commun.* 19, 3 (2020), 2022–2035.
[33] Kai Yang, Yuanming Shi, Yong Zhou, Zhanpeng Yang, Liqun Fu, and Wei Chen. 2020. Federated machine learning for intelligent IoT via reconfigurable intelligent surface. *IEEE Network* 34, 5 (2020), 16–22.
[34] Jingjing Zheng, Kai Li, Eduardo Tovar, and Mohsen Guizani. 2021. Federated learning for energy-balanced client selection in mobile edge computing. In *Proc. IEEE Int. Wireless Commun. Mobile Comput. Conf. (IWCMC).*
[35] Guangxu Zhu, Li Chen, and Kaibin Huang. 2018. MIMO over-the-air computation: Beamforming optimization on the Grassmann manifold. In *Proc. IEEE Global Commun. Conf. (GLOBECOM).*
[36] Guangxu Zhu, Yong Wang, and Kaibin Huang. 2019. Broadband analog aggregation for low-latency federated edge learning. *IEEE Trans. Wireless Commun.* 19, 1 (2019), 491–506.

A PROOF OF THEOREM 1

We will utilize Theorem 3 on page 53 of [23], which is reproduced below for clarity.

Theorem 2 ([23]). *Assume the following conditions hold:*

(1) *The distribution of $\mathbf{s}_t$ depends only on $\mathbf{w}_t$ and t, and given the sequence $\{\mathbf{w}_t\}$, the sequence $\{\mathbf{s}_t\}$ is independent.*
(2) *There is a scalar Lyapunov function $V(\cdot) \geq 0$ that is differentiable and $\nabla V(\cdot)$ is L-Lipschitz continuous.*
(3) *Process $\mathbf{s}_t$ is pseudogradient in relation to $V(\mathbf{w}_t)$, i.e., $\langle \nabla V(\mathbf{w}_t), \mathbb{E}[(\mathbf{s}_t|\mathbf{w}_t)]\rangle \geq bV(\mathbf{w}_t)$, where $b > 0$ is a constant scalar and $\langle \cdot, \cdot \rangle$ represents the inner product.*
(4) *The following growth condition on $\mathbf{s}_t$ is satisfied:*

$$\mathbb{E}[\|\mathbf{s}_t\|^2|\mathbf{w}_t] \leq \sigma^2 + \tau\langle \nabla V(\mathbf{w}_t), \mathbb{E}[(\mathbf{s}_t|\mathbf{w}_t)]\rangle, \tag{26}$$

where $\tau \geq 0$ and σ^2 are two constants.
(5) *The initial point satisfies $\mathbb{E}[V(\mathbf{w}_0)] < \infty$.*
(6) *The learning rate is such that:*

$$\gamma_t \geq 0, \sum_{t=0}^{\infty} \gamma_t = \infty, \limsup_{t\to\infty} \gamma_t < \frac{2}{L\tau}, \forall t. \tag{27}$$

Let either $\sigma^2 = 0$ or $\lim_{t\to\infty} \gamma_t \to 0$. Then the gradient descent method in (8) results in $\lim_{t\to\infty} \mathbb{E}[V(\mathbf{w}_t)] \to 0$.

Consider the Lyapunov function $V(\mathbf{w}) = F(\mathbf{w}) - F(\mathbf{w}^*)$. We will show that under assumptions **A1**-**A3** and conditions **C1**-**C3**, the conditions stated in the above theorem are satisfied, which directly implies our theorem statement. From the definition of $\mathbf{s}_t$, we have

$$\mathbf{s}_t = \frac{\mathcal{R}(\hat{\mathbf{r}}_t)}{K} = \frac{1}{K}\sum_m \left(\frac{\mathcal{R}[\mathbf{f}_t^H \mathbf{h}_{m,t} a_{m,t}]}{v_{m,t}}\right)\mathbf{g}_{m,t} + \frac{1}{K}\mathcal{R}\begin{pmatrix} \mathbf{f}_t^H \mathbf{n}_{1,t} \\ \vdots \\ \mathbf{f}_t^H \mathbf{n}_{D,t} \end{pmatrix}. \tag{28}$$

Based on (28), given $\mathbf{w}_t$, $\mathbf{s}_t$ is only a function of noise in round t, and since $\{\mathbf{n}_{d,t}\}$ is independent of the noise in other rounds and also independent of $\mathbf{w}_t$, the values of $\mathbf{s}_t$ in different rounds are independent of each other, given $\mathbf{w}_t$, and hence the first condition is satisfied in Theorem 2.

Since $V(\mathbf{w})$ is a difference of two L-Lipschitz continuous functions, it is also L-Lipschitz continuous, and thus the second condition in Theorem 2 is satisfied.

In order to fulfill the third condition of Theorem 2, we compute the following:

$$\begin{aligned}
\langle \nabla V(\mathbf{w}_t), \mathbb{E}[\mathbf{s}_t|\mathbf{w}_t]\rangle &\overset{(a)}{=} \langle\, \nabla F(\mathbf{w}_t), \mathbb{E}[\nabla F(\mathbf{w}_t) + \mathbf{e}_t|\mathbf{w}_t]\,\rangle \\
&\overset{(b)}{=} \|\nabla F(\mathbf{w}_t)\|^2 + \nabla F(\mathbf{w}_t)^T \mathbb{E}[\mathbf{e}_t|\mathbf{w}_t] \\
&\overset{(c)}{\geq} \|\nabla F(\mathbf{w}_t)\|^2 - \|\nabla F(\mathbf{w}_t)\|\,\|\mathbb{E}[\mathbf{e}_t|\mathbf{w}_t]\| \\
&\overset{(d)}{\geq} (1-\alpha)\|\nabla F(\mathbf{w}_t)\|^2 \\
&\overset{(e)}{\geq} 2\mu(1-\alpha)V(\mathbf{w}_t),
\end{aligned} \tag{29}$$

where (a) follows the definition of $\mathbf{s}_t$ in (9) and the fact that $\nabla V(\mathbf{w}) = \nabla F(\mathbf{w})$. Inequality (c) follows the Cauchy–Schwartz Inequality, and (d) follows our assumption in **C1**. Inequality (e) follows the fact that for strongly convex functions $\|\nabla F(\mathbf{w}_t)\|^2 \geq 2\mu(F(\mathbf{w}_t) - F(\mathbf{w}^*)) = 2\mu V(\mathbf{w}_t)$. So, the third condition of Theorem 2 holds with $b = 2\mu(1-\alpha)$. To satisfy the fourth condition of Theorem 2, we compute the following:

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{s}_t\|^2|\mathbf{w}_t] \\
&\quad\overset{(a)}{=} \mathbb{E}[\|\nabla F(\mathbf{w}_t) + \mathbf{e}_t\|^2] \\
&\quad\overset{(b)}{=} \|\nabla F(\mathbf{w}_t)\|^2 + 2\nabla F(\mathbf{w}_t)^T \mathbb{E}[\mathbf{e}_t|\mathbf{w}_t] + \mathbb{E}[\|\mathbf{e}_t\|^2|\mathbf{w}_t] \\
&\quad\overset{(c)}{\leq} \|\nabla F(\mathbf{w}_t)\|^2 + 2\nabla F(\mathbf{w}_t)^T \mathbb{E}[\mathbf{e}_t|\mathbf{w}_t] + \delta\|\nabla F(\mathbf{w}_t)\|^2 + \beta \\
&\quad\overset{(d)}{\leq} (1+\delta)\|\nabla F(\mathbf{w}_t)\|^2 + 2\|\nabla F(\mathbf{w}_t)\|\,\|\mathbb{E}[\mathbf{e}_t|\mathbf{w}_t]\| + \beta
\end{aligned}$$

$$\overset{(e)}{\leq} (1+\delta+2\alpha)\|\nabla F(\mathbf{w}_t)\|^2+\beta$$
$$\overset{(f)}{\leq} (1+\delta+2\alpha)\frac{\langle \nabla V(\mathbf{w}_t), \mathbb{E}[\mathbf{s}_t|\mathbf{w}_t]\rangle}{(1-\alpha)}+\beta, \tag{30}$$

where (a) follows the definition of $\mathbf{s}_t$ in (9), (c) follows our assumption in **C2**, (d) follows the Cauchy–Schwartz Inequality, (e) follows our assumption in **C1**, and finally (f) follows inequality (d) in (29). Therefore, the fourth condition is satisfied with $\sigma^2=\beta$ and $\tau=\frac{1+\delta+2\alpha}{1-\alpha}$.

The last two conditions of Theorem 2 are satisfied since the loss function value at the starting point is bounded, and **C3** guarantees the bounds on the learning rate outlined in the theorem.

B PROOF OF LEMMA 1

According to (10),

$$\mathbb{E}[\mathbf{e}_t|\mathbf{w}_t]=\mathbb{E}[\frac{1}{K}\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}+\frac{1}{K}\mathcal{R}\begin{pmatrix}\mathbf{f}_t^H\mathbf{n}_{1,t}\\ \vdots\\ \mathbf{f}_t^H\mathbf{n}_{D,t}\end{pmatrix}]$$
$$=\frac{1}{K}\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}, \tag{31}$$

where the reason for the second equality is that $\mathbf{n}_{d,t}$ has a zero mean, which implies the expectation of the second term is zero. Therefore,

$$\|\mathbb{E}[\mathbf{e}_t|\mathbf{w}_t]\|\overset{(a)}{\leq}\frac{1}{K}\sum_m|\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m|\|\mathbf{g}_{m,t}\|$$
$$\overset{(b)}{=}\frac{1}{K}\sum_m|\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m|\sqrt{D}v_{m,t}$$
$$=\frac{\sqrt{D}}{K}\sum_m|\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]-K_mv_{m,t}|, \tag{32}$$

where (a) follows the Triangle Inequality and (b) follows the definition of $v_{m,t}$.

C PROOF OF LEMMA 2

Let's define $\tilde{\mathbf{n}}_t\triangleq\begin{pmatrix}\mathcal{R}[\mathbf{f}_t^H\mathbf{n}_{1,t}]\\ \vdots\\ \mathcal{R}[\mathbf{f}_t^H\mathbf{n}_{D,t}]\end{pmatrix}$. Since $\mathbf{n}_{d,t}$ has zero mean, $\mathbb{E}[\tilde{\mathbf{n}}_t]=\mathbf{0}$. The d-th entry of $\tilde{\mathbf{n}}_t$ is denoted by $\tilde{\mathbf{n}}_t[d]$. Since $\mathbf{n}_{d,t}\sim\mathcal{CN}(\mathbf{0},\sigma_n^2\mathbf{I})$, the following relations hold for the statistics of its real and imaginary parts:

$$\mathbb{E}[\mathcal{R}[\mathbf{n}_{d,t}]\mathcal{R}[\mathbf{n}_{d,t}]^T]=\mathbb{E}[\mathcal{I}[\mathbf{n}_{d,t}]\mathcal{I}[\mathbf{n}_{d,t}]^T]=\frac{\sigma_n^2}{2}I, \tag{33}$$
$$\mathbb{E}[\mathcal{R}[\mathbf{n}_{d,t}]\mathcal{I}[\mathbf{n}_{d,t}]^T]=\mathbb{E}[\mathcal{I}[\mathbf{n}_{d,t}]\mathcal{R}[\mathbf{n}_{d,t}]^T]=\mathbf{0}. \tag{34}$$

Therefore, the variance of $\tilde{\mathbf{n}}_t[d]$ can be computed as

$$\mathbb{E}[\tilde{\mathbf{n}}_t[d]^2]=\mathbb{E}[\mathcal{R}[\mathbf{f}_t^H\mathbf{n}_{d,t}]^2]=\frac{\sigma_n^2}{2}\|\mathbf{f}_t\|^2, \tag{35}$$

and subsequently,

$$\mathbb{E}[\|\tilde{\mathbf{n}}_t\|^2]=\frac{D\sigma_n^2\|\mathbf{f}_t\|^2}{2}. \tag{36}$$

Based on the definition of $\mathbf{e}_t$ in (10):

$$\mathbb{E}[\|\mathbf{e}_t\|^2|\mathbf{w}_t]$$
$$\overset{(a)}{=}\mathbb{E}\Big[\|\frac{1}{K}\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}+\frac{\tilde{\mathbf{n}}_t}{K}\|^2|\mathbf{w}_t\Big]$$
$$\overset{(b)}{=}\mathbb{E}[\|\frac{1}{K}\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}\|^2|\mathbf{w}_t]$$
$$+\mathbb{E}[\Big(\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}\Big)^H\frac{\tilde{\mathbf{n}}_t}{K^2}|\mathbf{w}_t]$$
$$+\mathbb{E}[\frac{\tilde{\mathbf{n}}_t^H}{K^2}\Big(\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}\Big)|\mathbf{w}_t]$$
$$+\mathbb{E}[\frac{\|\tilde{\mathbf{n}}_t\|^2}{K^2}|\mathbf{w}_t]$$
$$\overset{(c)}{=}\|\frac{1}{K}\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}\|^2$$
$$+\Big(\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}\Big)^H\mathbb{E}[\frac{\tilde{\mathbf{n}}_t}{K^2}]$$
$$+\mathbb{E}[\frac{\tilde{\mathbf{n}}_t^H}{K^2}]\Big(\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}\Big)+\mathbb{E}[\frac{\|\tilde{\mathbf{n}}_t\|^2}{K^2}]$$
$$\overset{(d)}{=}\|\frac{1}{K}\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}\|^2+\frac{D\sigma_n^2\|\mathbf{f}_t\|^2}{2K^2}$$
$$\overset{(e)}{\leq}\Big(\frac{1}{K}\sum_m|\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m|\,\|\mathbf{g}_{m,t}\|\Big)^2+\frac{D\sigma_n^2\|\mathbf{f}_t\|^2}{2K^2}$$
$$\overset{(f)}{=}\frac{D}{K^2}\Big(\sum_m|\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]-K_mv_{m,t}|\Big)^2+\frac{D\sigma_n^2\|\mathbf{f}_t\|^2}{2K^2}, \tag{37}$$

where (a) follows the definition of $\mathbf{e}_t$; (b) expands the error norm squared to four terms; (c) follows the fact that given $\mathbf{w}_t$, $\frac{1}{K}\sum_m(\frac{\mathcal{R}[\mathbf{f}_t^H\mathbf{h}_{m,t}a_{m,t}]}{v_{m,t}}-K_m)\mathbf{g}_{m,t}$ is deterministic, and $\tilde{\mathbf{n}}_t$ is independent of $\mathbf{w}_t$; (d) follows the fact that $\mathbb{E}[\tilde{\mathbf{n}}_t]=\mathbf{0}$ and (36); (e) follows the Triangle Inequality; and finally (f) follows the definition of $v_{m,t}$.

D PROOF OF LEMMA 3

The proof is by contradiction. Let's denote $\phi_{m,t}\triangleq\angle a_{m,t}^*$ and $\psi_{m,t}=\angle\mathbf{f}_t^H\mathbf{h}_{m,t}$. Suppose $\phi_{m,t}\neq-\psi_{m,t}$. Let's define another set of transmit weights $\tilde{a}_{m,t}\triangleq|a_{m,t}^*|\cos(\phi_{m,t}+\psi_{m,t})e^{-j\psi_{m,t}},\forall m$. We have

$$\mathcal{R}[\tilde{a}_{m,t}\mathbf{f}_t^H\mathbf{h}_{m,t}]=|a_{m,t}^*|\cos(\phi_{m,t}+\psi_{m,t})|\mathbf{f}_t^H\mathbf{h}_{m,t}| \tag{38}$$
$$=\mathcal{R}[a_{m,t}^*\mathbf{f}_t^H\mathbf{h}_{m,t}],\forall m. \tag{39}$$

Therefore, $\{\tilde{a}_{m,t}\}$ can satisfy constraint (18b). Moreover, $|\tilde{a}_{m,t}|=|a_{m,t}^*|\cos(\phi_{m,t}+\psi_{m,t})<|a_{m,t}^*|,\forall m$ and hence $\{\tilde{a}_{m,t}\}$ satisfies (18c) with a lower value for the objective function. This contradicts the optimality of $\{a_{m,t}^*\}$.