Federated Learning Over-the-Air by Retransmissions

Henrik Hellström, *Graduate Student Member, IEEE*, Viktoria Fodor, *Member, IEEE*, and Carlo Fischione, *Senior Member, IEEE*

***Abstract*— Motivated by the increasing computational capabilities of wireless devices, as well as unprecedented levels of user- and device-generated data, new distributed machine learning (ML) methods have emerged. In the wireless community, Federated Learning (FL) is of particular interest due to its communication efficiency and its ability to deal with the problem of non-IID data. FL training can be accelerated by a wireless communication method called Over-the-Air Computation (AirComp) which harnesses the interference of simultaneous uplink transmissions to efficiently aggregate model updates. However, since AirComp utilizes analog communication, it introduces inevitable estimation errors. In this paper, we study the impact of such estimation errors on the convergence of FL and propose retransmissions as a method to improve FL accuracy over resource-constrained wireless networks. First, we derive the optimal AirComp power control scheme with retransmissions over static channels. Then, we investigate the performance of Over-the-Air FL with retransmissions and find two upper bounds on the FL loss function. Numerical results demonstrate that the power control scheme offers significant reductions in mean squared error. Additionally, we provide simulation results on MNIST classification with a deep neural network that reveals significant improvements in classification accuracy for low-SNR scenarios.**

***Index Terms*— Federated learning, over-the-air computation, retransmissions.**

I. Introduction

THE data collection rate in wireless devices is growing at an exceptional speed due to the increasing adoption of smartphones, tablets, and Internet of Things (IoT) devices [1], [2]. These devices are expected to provide a broad range of Artificial Intelligence (AI) services in Sixth Generation (6G) networks, such as predictive healthcare [3], search-and-rescue drones [4], and environmental monitoring [5]. As a consequence, new distributed machine learning methods, such as Federated Learning (FL), have become essential to enable privacy-preserving and communication-efficient model training [6]. A recent survey on open problems of FL argues that communication is often a primary bottleneck for FL because wireless links operate at low rates that can be both expensive and unreliable [7]. Communication-efficient FL is investigated thoroughly in [8], where various compression techniques such as quantization, random rotation, and sub-sampling are evaluated. In [9], [10], and [11] it is established that new wireless methods can greatly improve the communication efficiency of edge AI.

A novel approach to wireless communication, called Over-the-air computation (AirComp), has recently been adapted to support Machine Learning (ML) services [12], [13]. AirComp is an analog communication scheme that orders its users to communicate simultaneously over the same frequency band, thereby promoting interference. This interference is leveraged to compute a function of the transmitted messages by utilizing the superposition property of the wireless channel [14]. By appropriately precoding the transmitted signals and post-coding the received signal, all nomographic functions can be calculated over the air [15]. In FL, the central server is interested in collecting the arithmetic mean of model updates from the participating devices. Since the arithmetic mean is a nomographic function, AirComp is a suitable communication solution [11].

Compared to conventional point-to-point digital communications, AirComp is attractive from a communication-efficiency standpoint, with throughput gains approximately proportional to the number of users [12]. The reason for this drastic improvement is that the entire wireless spectrum can be utilized concurrently by all devices, rather than dividing it and allocating smaller resource blocks to each device. Additionally, AirComp obfuscates the participating users since the central server directly receives the arithmetic mean rather than the individual model updates, thereby enhancing privacy [16].

Currently, AirComp is reliant upon specialized hardware and fine synchronization that might be difficult to achieve in practice [17]. Additionally, AirComp is unable to guarantee perfect reconstruction of the transmitted messages at the receiver. Shannon's "fundamental theorem for a discrete channel with noise" establishes that for any degree of noise contamination, it is possible to communicate discrete data with an arbitrarily small frequency of errors [18]. However, to achieve a non-zero communication rate, redundant information must be transmitted in the form of a code. Since the information transmitted in AirComp is not discrete, existing codes do not appear to be directly applicable. Instead, AirComp settles for

Manuscript received 19 November 2021; revised 30 June 2022 and 3 February 2023; accepted 4 April 2023. Date of publication 26 April 2023; date of current version 12 December 2023. This work was supported in part by the Swedish Foundation for Strategic Research through the Software Artificial Intelligence for Communications (SAICOM) Project and in part by the Swedish Research Council. The associate editor coordinating the review of this article and approving it for publication was C. Brinton. *(Corresponding author: Henrik Hellström.)*

The authors are with the Network and Systems Engineering (NSE) and Digital Futures, KTH Royal Institute of Technology, 11428 Stockholm, Sweden (e-mail: hhells@kth.se; vjfodor@kth.se; carlofi@kth.se).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2023.3268742.

Digital Object Identifier 10.1109/TWC.2023.3268742

estimating the desired function as closely as possible, while retaining some non-zero estimation error [19]. In [20], it is proven that these errors harm the convergence properties of FL, both the rate of convergence and post-convergence loss.

In the current AirComp literature, the main way of reducing the estimation error is to optimize the transmission powers. In [19] and [21], the authors propose a closed-form power control scheme that minimizes the mean squared error (MSE) between the received signal and the desired function of the sources' messages under a peak transmission power constraint. For the case of multiple antennas, no closed-form power control scheme has been found, but [13] develops a strong heuristic by using a difference-of-convex-functions representation of the problem. In [22] and [23], the multi-antenna problem is coupled with wireless power transfer to improve the battery life of participating IoT devices. To further improve the power control, [24] proposes a gradient-statistics aware scheme that learns statistical properties of the model updates to improve the AirComp estimation error. In [25], the temporal structure of gradient sparsity is leveraged to develop a Bayesian prior that improves the estimation. Another common approach is to incorporate intelligent reflective surfaces with AirComp to reach substantially lower estimation errors [26], [27], [28].

As a general pattern, none of these works offer avenues to trade off communication resources for improved estimation. In digital communications, such communication-estimation trade-offs are the main way to reduce errors. For instance, it is standard to adaptively control the modulation order and coding rate to compensate for poor channels [29]. Unfortunately, none of these approaches are directly compatible with AirComp since the communication is analog. In this paper, we take a first step towards enabling this communication-estimation tradeoff for Over-the-Air federated learning with a system we call AirReComp. The contributions of this paper are summarized as follows.

- A power control scheme for AirReComp is proposed. The proposed scheme is proven to be globally optimal in terms of MSE between the estimated and desired function, given assumptions on the first and second moments of the local model updates.
- Upper bounds on the FL loss function are derived for single-epoch Lipschitz-smooth functions, both for the strongly convex and convex case.
- To further support the feasibility of AirReComp under non-convex functions, we provide numerical results with Deep Neural Networks (DNNs). These results suggest that AirReComp can beat state-of-the-art Over-the-Air FL in terms of classification accuracy.

The remainder of the paper is organized as follows. Section II introduces the system model. Section III presents and solves the power control problem to minimize the MSE between the desired and received sum. Section IV provides worst-case analyses on the performance of AirReComp in terms of two upper bounds on the FL loss function. In Section V, the proposed AirReComp scheme and the convergence bound are numerically evaluated for non-convex and convex loss functions. Finally, section VI concludes the paper and discusses future work.

Notation: z is a scalar, $\mathbf{z}$ is a vector, and $\mathbf{Z}$ is a matrix. Element i of vector $\mathbf{z}$ is expressed as $z^{(i)}$. To denote element-wise operations of vectors, we overload the scalar equivalent, e.g. $\mathbf{x}/\mathbf{y}$ is the element-wise division of $\mathbf{x}$ and $\mathbf{y}$. $\bar{z}$ denotes the complex conjugate. $\hat{z}$ denotes an estimate of z.

II. System Model

In this section, we describe the system model and the AirReComp algorithm. We consider a distributed ML system consisting of K single-antenna user devices each carrying a distinct dataset $\mathcal{D}_k$ and a single-antenna parameter server (PS) which can be reached by all devices in a single hop. The objective of the system is to solve the following optimization problem

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, F(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}}\, \frac{1}{K}\sum_{k=1}^{K} F_k(\mathbf{w}), \tag{1}$$

using the datasets at the user devices. The vector $\mathbf{w} \in \mathbb{R}^d$ is the $d \times 1$ parameter vector that defines the ML model, $F(\mathbf{w})$ is denoted the global loss function, and $F_k(\mathbf{w})$ is a local loss function.

The uplink wireless channel is modeled as a block-fading multiple access channel (MAC) with additive noise [30]. If the K users simultaneously transmit a vector $\mathbf{x}_k \in \mathbb{R}^d$ over the MAC, the PS receives

$$\mathbf{y} = \sum_{k=1}^{K} h_k \mathbf{x}_k + \mathbf{z}, \tag{2}$$

where $h_k \in \mathbb{C}$ denotes the channel coefficient from device k to the PS and $\mathbf{z} \in \mathbb{C}^d$ denotes additive white Gaussian noise (AWGN) with variance σ_z^2. Additionally, we consider retransmissions over this channel, where we assume that the coherence time of the wireless channel is long enough to accommodate M uplink transmissions. If the PS aggregates the result of these transmissions, we get

$$\mathbf{y} = M\sum_{k=1}^{K} h_k \mathbf{x}_k + \sum_{m=1}^{M} \mathbf{z}_m, \tag{3}$$

where the desired signal strength is increased by a factor M but the aggregate of the noise terms $\mathbf{z}_m$ is diminished due to the random sampling. These kinds of static fading channels exist in several practical wireless applications, such as industrial communications. As a conservative example, consider an IEEE 802.11 factory wireless sensor network with coherence times of around 100ms [31]. Such a network provides at least $L = 10$ parallel communication channels [32] and has a symbol period of less than $T = 10\mu s$ [33]. Considering a small neural network with $d = 10,000$ parameters, it takes $MdT/L = 10M$ ms to perform M uplink transmissions, which accommodates $M = 10$ transmissions within the coherence time. For other scenarios with fast-fading channels, we refer to our recent work [34].

For simplicity, we assume error-free broadcast transmission in the downlink, which is an acceptable approximation for most practical scenarios since the PS generally has much greater communication capability than the user devices [35].

A. Federated Learning Algorithm

FL is an iterative algorithm to solve (1), where each iteration is denoted a communication round and consists of downlink broadcast, model training at the user devices, and uplink aggregation.

Communication round n starts when the PS broadcasts the global model $\mathbf{w}_n$ to all user devices in the downlink. Upon receiving the model, user device k solves the local problem

$$\mathbf{w}_{n,k}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, F_k(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{\mathbf{u}_i \in \mathcal{D}_k} l(\mathbf{w}, \mathbf{u}_i) \tag{4}$$

where $\mathbf{u}_i$ denotes one training sample and $l(\mathbf{w}, \mathbf{u}_i)$ is the sample-wise loss function. Generally, (4) can not be solved exactly. Instead, each device runs E epochs of gradient descent to approximately solve (4) as follows

$$\mathbf{w}_{n,k}(i) \leftarrow \mathbf{w}_{n,k}(i-1) - \beta \nabla F_k(\mathbf{w}_{n,k}(i-1)), \quad \forall i = 1, \ldots, E, \tag{5}$$

where the first communication round is based on the global model, i.e. $\mathbf{w}_{n,k}(0) = \mathbf{w}_n$, and β is the step size. After executing E epochs, device k calculates a local model update as $\Delta\mathbf{w}_{n,k} = \mathbf{w}_n - \mathbf{w}_{n,k}(E)$. After all local model updates have been computed, they are transmitted in the uplink to the PS.

At the PS, the local model updates are aggregated to form a global model update. In this paper, we consider the original FedAvg update [6], written as

$$\Delta\mathbf{w}_n = \frac{1}{K}\sum_{k=1}^{K} \Delta\mathbf{w}_{n,k}. \tag{6}$$

Finally, the PS concludes the communication round by generating the next iteration of the model parameters

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \Delta\mathbf{w}_n. \tag{7}$$

The algorithm repeats for N communication rounds until $\mathbf{w}_N$ is generated as the final model.

B. Over-the-Air Computation Protocol

In the uplink aggregation step of FL, see (6), the PS reconstructs the sum of K model updates. In this section, we describe how this is achieved by AirComp. To start, the model updates are embedded into the transmit signals $\mathbf{x}_{n,k}$ as

$$\mathbf{x}_{n,k} = \Delta\mathbf{w}_{n,k} \frac{\bar{h}_{n,k}}{|h_{n,k}|}\sqrt{p_{n,k}}, \tag{8}$$

where $h_{n,k}$ is the channel coefficient of device k for communication round n. All devices transmit $\mathbf{x}_{n,k}$ simultaneously over the MAC (2), which yields the following received value at the PS

$$\mathbf{y}_n = \sum_{k=1}^{K} |h_{n,k}|\sqrt{p_{n,k}}\Delta\mathbf{w}_{n,k} + \mathbf{z}_n. \tag{9}$$

Ideally, the transmission powers would be chosen as $p_{n,k} = 1/|h_{n,k}|^2$, which would completely compensate for the fading effect. However, with a natural constraint on the maximum transmission power, $p_{n,k} = 1/|h_{n,k}|^2$ might be impossible to achieve. Because of this limitation and due to the additive noise, the PS can never perfectly reconstruct $\Delta\mathbf{w}_n$. Instead, it estimates $\Delta\mathbf{w}_n$ by dividing the received signal by a post-transmission scalar $\sqrt{\eta_n}$ and the number of devices K

$$\bar{\mathbf{y}}_n = \frac{\mathbf{y}_n}{\sqrt{\eta_n}K} = \sum_{k=1}^{K} \frac{|h_{n,k}|\Delta\mathbf{w}_{n,k}\sqrt{p_{n,k}}}{\sqrt{\eta_n}K} + \frac{\mathbf{z}_n}{\sqrt{\eta_n}K}. \tag{10}$$

In practice, the division of $\sqrt{\eta_n}K$ takes place in the baseband of the PS, i.e., an operation in the digital hardware of the receiver. Coupled with the transmission powers, $\sqrt{\eta_n}$ has an important role. We see that the ideal choice of the transmission powers is now $\sqrt{p_{n,k}} = \sqrt{\eta_n}/|h_{n,k}|$. As such, the selection of a small η_n will reduce the amount of energy required to invert a channel and thereby reduce the fading error. However, lowering η_n will also increase the relative power of the noise. Therefore, the post-transmission scalar $\sqrt{\eta_n}$ will play the role of a tradeoff parameter between the fading error and the noise-induced error [19].

In this work, we propose AirReComp, which considers retransmissions in the uplink aggregation step. Specifically, the devices transmit the same values in the uplink M times such that the signal part of (10) combines constructively, while the additive noise is different for each transmission. After receiving M values, the PS forms its estimate by calculating their arithmetic mean

$$\begin{aligned}\bar{\mathbf{y}}_n &= \frac{\sum_{m=1}^{M} \mathbf{y}_{n,m}}{M\sqrt{\eta_n}K} \\ &= \sum_{k=1}^{K} \frac{|h_{n,k}|\Delta\mathbf{w}_{n,k}\sqrt{p_{n,k}}}{\sqrt{\eta_n}K} + \sum_{m=1}^{M} \frac{\mathbf{z}_{n,m}}{M\sqrt{\eta_n}K}. \end{aligned}\tag{11}$$

Next, the PS takes the real part of $\bar{\mathbf{y}}_n$ to reduce the power of the noise

$$\Delta\hat{\mathbf{w}}_n = \operatorname{Re}(\bar{\mathbf{y}}_n) = \sum_{k=1}^{K} \frac{|h_{n,k}|\Delta\mathbf{w}_{n,k}\sqrt{p_{n,k}}}{\sqrt{\eta_n}K} + \sum_{m=1}^{M} \frac{\operatorname{Re}(\mathbf{z}_{n,m})}{M\sqrt{\eta_n}K}. \tag{12}$$

With appropriate choices of $p_{n,k}$ and η_n (elaborated upon in Section III), the estimate described in (12) can be a close estimate of $\Delta\mathbf{w}_n$. However, note that due to the analog modulation protocol, the norm of the model update $\|\Delta\hat{\mathbf{w}}_n\|$ depends on the transmission powers $p_{n,k}$. To ensure that the transmission protocol does not affect the length of the global update step, the PS updates the model as

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \frac{\Delta\hat{\mathbf{w}}_n}{c_1/K}, \tag{13}$$

where

$$c_1 := \frac{\sum_{k=1}^{K} \sqrt{p_{n,k}}|h_{n,k}|}{\sqrt{\eta_n}}. \tag{14}$$

This particular choice of normalization is motivated by that $\mathbb{E}[\Delta\hat{\mathbf{w}}_n/(c_1/K)] = \beta\Delta\mathbf{w}_n$, given that the model updates are independent and identically distributed (IID). See Appendix A, (43) for details.

The whole AirReComp process is summarized in Algorithm 1.

Algorithm 1 AirReComp

```
Parameter Server:
    initialize $\mathbf{w}_0$
for each round $n = 0, 1, \ldots, N-1$ do
    Parameter Server:
        broadcast $\mathbf{w}_n$ to devices
    Device $k$:
        $\mathbf{w}_{n,k}(E) \leftarrow$ Equation (5)
        $\Delta\mathbf{w}_{n,k} \leftarrow \mathbf{w}_n - \mathbf{w}_{n,k}(E)$
        $\mathbf{x}_{n,k} \leftarrow$ Equation (8)
    for each $m = 1, 2, \ldots, M$ do
        all devices simultaneously:
            transmit $\mathbf{x}_{n,k}$ to server
    end for
    Parameter Server:
        $\bar{\mathbf{y}}_n \leftarrow$ Equation (11)
        $\Delta\hat{\mathbf{w}}_n \leftarrow$ Equation (12)
        $\mathbf{w}_{n+1} \leftarrow$ Equation (13)
end for
```

Remark 1: The expected transmission power of $\mathbf{x}_{n,k}$ at device k is

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_{n,k}\|^2] &= \mathbb{E}[\|\Delta\mathbf{w}_{n,k}\frac{\bar{h}_{n,k}}{|h_{n,k}|}\sqrt{p_{n,k}}\|^2] \\ &= \mathbb{E}[\|\Delta\mathbf{w}_{n,k}\|^2]p_{n,k}, \end{aligned} \tag{15}$$

where $\mathbf{x}_{n,k}$ is defined in (8). Since the expected power is not $p_{n,k}$, a maximum transmission power constraint of $\overline{P}$, leads to $p_{n,k} \leq \overline{P}/(\max_k \mathbb{E}[\|\Delta\mathbf{w}_{n,k}\|^2])$. Throughout this paper, we refer to this maximum value as $P_{max,n} := \overline{P}/(\max_k \mathbb{E}[\|\Delta\mathbf{w}_{n,k}\|^2])$.

Remark 2: While retransmissions improve the ability to accurately estimate the global model update, the total training time is increased significantly. If the slowest device consumes approximately T_c seconds to solve (5) and T_u seconds for uplink communication, the total time spent in one communication round will be

$$T = (T_c + MT_u)N, \tag{16}$$

which is roughly proportional to M for communication-constrained systems. However, in the convergence bounds and numerical results we demonstrate that in low-SNR scenarios, this additional cost can be necessary to achieve sufficient performance.

Note that T_c corresponds to the slowest device due to the straggler problem of FL, which could potentially be improved through the use of coded computing [36] or by introducing a relay [37], which is not considered in this work. In practice, the time for transmission could be estimated using information about the wireless protocol [38] and the computational time could be estimated using standard formulas relating to the computational capacity of the devices [39].

III. Power Control

In this section, we consider a power control problem to minimize the mean-squared estimation error defined as

$$\mathbb{E}[(\Delta\mathbf{w}_n - \Delta\hat{\mathbf{w}}_n)^2], \tag{17}$$

where the expectation is taken over the AWGN, and $\Delta\mathbf{w}_n$ and $\Delta\hat{\mathbf{w}}_n$ are defined in (6) and (12) respectively. For mathematical tractability, as done in the related literature, we assume that these model updates are IID, zero mean, and have unit variance [19], [21]. To perform the minimization, we seek the optimal choice of the transmission powers $\sqrt{p_{n,k}}$ and the post-transmission scalar $\sqrt{\eta_n}$. Since we consider static fading coefficients, the power control problem only has to be solved once per communication round (the same solution is re-used for M transmissions). To model the limited transmission power of the devices, we consider the following constraint

$$p_{n,k} \leq P_{\max,n} \quad \forall k, \tag{18}$$

where $P_{\max,n}$ is defined in Remark 1. The minimization of (17) is formulated as

$$\begin{aligned} &\min_{\mathbf{p},\eta} \\ &\mathbb{E}\left[\left(\sum_{k=1}^{K}\frac{|h_k|\Delta\mathbf{w}_k\sqrt{p_k}}{\sqrt{\eta}K} + \sum_{m=1}^{M}\frac{\mathrm{Re}(\mathbf{z}_m)}{M\sqrt{\eta}K} - \sum_{k=1}^{K}\frac{\Delta\mathbf{w}}{K}\right)^2\right] \\ &\text{s.t. } p_k \leq P_{\max}, \quad \forall k \in \{1 < k < K\}, \end{aligned} \tag{19}$$

where the subscript n has been ommitted for brevity. Note that the number of transmissions M is given as an input parameter and is selected before the power control problem is solved.

Proposition 1: Problem (19) has a unique solution. The optimal post-transmission scalar is given by the solution to the K subproblems

$$\eta_n^* = \min_k \tilde{\eta}_{n,k}, \tag{20}$$

where

$$\tilde{\eta}_{n,k} = \left(\frac{\sum_{j=1}^{k}|h_{n,j}|^2 P_{\max,n} + \sigma_z^2/M}{\sum_{j=1}^{k}|h_{n,j}|\sqrt{P_{\max,n}}}\right)^2. \tag{21}$$

The optimal transmission powers are

$$p_{n,k}^* = \min\left(P_{\max,n}, \frac{\eta_n^*}{|h_{n,k}|^2}\right). \tag{22}$$

The proof of Proposition 1 follows the proof in [19] and is omitted from this paper.

Remark 3: From (21), we see that the post-transmission scalar η_n^* assumes a lower value when more retransmissions are used. As we increase the number of retransmissions, the signal-to-noise ratio (SNR) increases and consequently, the noise-induced error reduces. Therefore, the fading error becomes dominant and the optimal post-transmission scalar η_n^* is lowered to improve it.

Corollary 1: The optimal transmission powers p_k, given in Proposition 1, are decreasing in M.

Proof: From (21) it is clear that all $\eta_{n,k}$'s are strictly decreasing in M. The post-transmission scalar η_n is selected according to (20), which in turn is selected as the smallest *of K different $\eta_{n,k}$. The transmission powers $p_{n,k}$ is selected according to (22), from which it is clear that $p_{n,k}$ is decreasing in η_n, and therefore decreasing in M.* □

IV. Convergence Analysis

In this section, we analyze the learning performance of Algorithm 1. For the analysis, we assume that there is only one epoch of local training in each communication round ($E = 1$). Additionally, we assume that the channels remain static for the entire duration of the training process. As a result, we drop the n index in the channel coefficients h_k, the transmission powers p_k, and the post-transmission scalar η. The effect of dynamic channels is evaluated numerically in Section V. The performance is measured as the gap between the FL loss gap at communication round n, defined as

$$\mathbb{E}[F(\mathbf{w}_n)] - F(\mathbf{w}^*). \tag{23}$$

We derive two upper bounds on this loss gap, one for strongly-convex functions and one for convex functions. For both bounds, we use the following well-known lemma [20], [40].

Lemma 1: Let $F(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function with L-Lipschitz gradient. Then, the following inequality holds:

$$F(\mathbf{y}) - F(\mathbf{x}) - \nabla F(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \leq \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2. \tag{24}$$

Additionally, we make an assumption regarding the similarity of the local model updates $\Delta\mathbf{w}_{n,k}$ and the global model update $\Delta\mathbf{w}_n$ [24], [30].

Assumption 1: The local model updates $\Delta\mathbf{w}_{n,k}$ are assumed to be independent and unbiased estimates of the global model update $\Delta\mathbf{w}_n$.

$$\mathbb{E}[\Delta\mathbf{w}_{n,k}] = \Delta\mathbf{w}_n, \quad \forall k \in \{1,2,\ldots,K\}. \tag{25}$$

The local gradients and the global gradient are in general different. The difference has coordinate bounded variance [30]:

$$\mathbb{E}\left[\left(\nabla F_k^{(i)}(\mathbf{w}_n(0)) - \frac{1}{K}\sum_{k=1}^{K}\nabla F_k^{(i)}(\mathbf{w}_n(0))\right)^2\right] \leq (\sigma^{(i)})^2, \tag{26}$$

and as a consequence, the model update difference can be bounded as

$$\mathbb{E}[(\Delta w_{n,k}^{(i)} - \Delta w_n^{(i)})^2] \leq \beta^2(\sigma^{(i)})^2, \tag{27}$$

where $\Delta w_{n,k}^{(i)}$ is the i-th element of $\Delta\mathbf{w}_{n,k}$, and $(\sigma^{(i)})^2$ are the element-wise upper bounds. We will also use $\boldsymbol{\sigma} \in \mathbb{R}^d$ to denote the vector of variance bounds.

A. Strongly-Convex Loss

In this subsection, we assume that the FL loss is μ-strongly convex. For such a loss, we use the following lemma [20], [40]:

Lemma 2: Let $F(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$ be a μ-strongly convex function with L-Lipschitz gradient. Then, the following inequality holds:

$$\begin{aligned}&(\nabla F(\mathbf{x}) - \nabla F(\mathbf{y}))^T(\mathbf{x}-\mathbf{y})\\ &\geq \frac{\mu L}{\mu+L}\|\mathbf{x}-\mathbf{y}\|^2 + \frac{1}{\mu+L}\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|^2.\end{aligned} \tag{28}$$

Before we are ready to state the upper bound, we must also assert that the local step size β has been selected to be sufficiently small for convergence.

Assumption 2: Let the local step size β be

$$\beta < \min\left(\frac{\mu+L}{2\mu L}, \frac{2}{K(\mu+L)}\frac{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}{\sum_{k=1}^{K}p_k|h_k|^2}\right), \tag{29}$$

where p_k is determined according to Proposition 1.

Note that even though the power normalization step (13) makes the expected gradient norm independent of the power control parameters, the squared norm is still dependent. Therefore, the step length β must be selected with respect to the power, as seen in (29). We are now ready to give the first upper bound on the FL loss function (23). Given Assumption 1 and 2, the update described in (12) and (13) converges according to Proposition 2.

Proposition 2: Let

$$c_2 := 1 - 2\beta\frac{\mu L}{\mu+L}, \tag{30}$$

and

$$c_3 := \beta^2\|\boldsymbol{\sigma}\|^2 K\sum_{k=1}^{K}p_k|h_k|^2 + \frac{d\sigma_z^2}{M}. \tag{31}$$

Then the FL loss is upper bounded by

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_n)] - F(\mathbf{w}^*)\\ &\leq \frac{L}{2}c_2^n\mathbb{E}[r_0^2] + \frac{Lc_3}{2\left(\sum_{k=1}^{k}\sqrt{p_k}|h_k|\right)^2(1-c_2)},\end{aligned} \tag{32}$$

where $r_0 = \|\mathbf{w}_0 - \mathbf{w}^*\|$ is the distance between the initial weight vector and the optimal one, $\boldsymbol{\sigma}$ is a vector of the coordinate bounded variances from (27), and d is the number of model parameters.

Proof: The proof is provided in Appendix A. □

We refer to the first term on the RHS of (32) as the *diminishing term* because it approaches zero if $n \to \infty$. Along the same line, we refer to the other term as the *post-convergence term* because it remains non-zero even if $n \to \infty$. From (32), we know that the convergence rate of the diminishing term is $\mathcal{O}(c_2^n)$, typically called *linear convergence*. Implications of Proposition 2 are given in Section IV-C.

B. Convex Loss

In this subsection, we relax the assumption on strong convexity and develop a bound for Lipschitz smooth and convex loss functions. For this bound, we need a different guarantee on the fixed step size than for the strongly convex case.

Assumption 3: The fixed step size β is selected to satisfy:

$$0 < \beta < \frac{1}{LK}\frac{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}{\sum_{k=1}^{K}p_k|h_k|^2}. \tag{33}$$

Proposition 3: Consider Assumption 1 and 3. Then the FL loss is upper bounded by

$$\mathbb{E}[F(\mathbf{w}_n)] - F(\mathbf{w}^*) \leq \frac{1}{2n\beta}\mathbb{E}[r_0^2] + \frac{2+L}{2\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}c_3, \tag{34}$$

where c_3 is defined in (31).

Proof: The proof is provided in Appendix B. □

From (34), it is clear that the convergence rate of the diminishing term is $\mathcal{O}(1/n)$, typically called *sub-linear convergence*.

C. Discussion on Proposition 2 and 3

Since the propositions are upper bounds, we are discussing the worst-case properties of the FL loss using AirReComp. We are specifically interested in the impact of the number of retransmissions.

1) Diminishing Term: In both bounds, the diminishing term is unaffected by M. Similar results can be seen in the optimization literature. For instance, an illuminating parallel can be drawn to the convergence bound for mini-batch gradient descent (GD) [41]

$$f(x_n) \leq f(x^*) + \frac{\|x_0 - x^*\|^2}{2\beta n} + \frac{\beta\,\mathrm{var}(\mathbf{v})}{2}, \tag{35}$$

where $\mathrm{var}(\mathbf{v})$ is the variance introduced by the random selection of samples. Similar to our system, where the variance of the gradient is reduced by adding retransmissions, $\mathrm{var}(\mathbf{v})$ in mini-batch GD is reduced by increasing the batch size. As seen in (35), the effect of this variance reduction is only reflected in the post-convergence term, just as how M only shows up in the post-convergence term of (32) and (34).

2) Final Error: Since both bounds have post-convergence terms, the algorithm does not converge to a local optimum. Instead, the algorithm converges to a region of optimality, where the expected remaining loss gap is given by the post-convergence terms. There are two reasons why AirReComp does not converge exactly. Firstly, the channel noise (characterized by σ_z) causes unavoidable errors which prevent exact convergence. Secondly, the difference between local and global model updates (characterized by $\boldsymbol{\sigma}$) causes a global model update that differs from what is achieved in centralized gradient descent. This result aligns with what was found in [30].

We investigate the post-convergence terms of the two bounds closer. Since we are interested in the impact of retransmissions, we focus on the terms that are affected by M. To start, consider the post-convergence terms of (32) and (34), which can be expressed as

$$\frac{C}{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}\left(\beta^2\|\boldsymbol{\sigma}\|^2 K\sum_{k=1}^{K}p_k|h_k|^2 + \frac{d\sigma_z^2}{M}\right), \tag{36}$$

where $C := L/(2(1-c_2))$ and $C := (2+L)/2$ for (32) and (34), respectively. It is worth noting that the first term, caused by the gradient difference $\boldsymbol{\sigma}$ cannot be completely eliminated, even with perfect communication. In fact, by the Cauchy-Schwarz inequality, we can lower-bound the first term to

$$C\beta^2\|\boldsymbol{\sigma}\|^2\frac{K\sum_{k=1}^{K}p_k|h_k|^2}{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2} \geq C\beta^2\|\boldsymbol{\sigma}\|^2, \tag{37}$$

which is completely unaffected by the communication scheme. As for the noise-induced term, we see an improvement of order $\mathcal{O}(1/M)$. However, from Corollary 1, we know that this order is slightly diminished by the fact that the transmission powers p_k are decreasing in M. Since the relationship between p_k and M cannot be stated in closed form, we analyze the diminishing terms further in the numerical results.

V. Numerical Results

The performance of the proposed AirReComp system is now evaluated in terms of the model update estimation error, federated learning loss, and classification accuracy. Specifically, we have four goals with this section:

- To demonstrate the need for retransmission-aware power control, by comparing our proposed solution with the state-of-the-art single transmission schemes proposed in [19], [20], and [21] and a perfect communication baseline;
- To demonstrate that introducing retransmissions is also beneficial for non-convex loss functions. Note that our analytical results assumed convex loss functions;
- To demonstrate that the proposed method is viable both when the channels change every communication round (as assumed in Section II) and when the channels remain static for the entire training process (as assumed in Section IV);
- To demonstrate the rate that the post-convergence terms of the bounds we developed in Section IV are decreasing in M.

A. Power Control

In this subsection, we wish to evaluate the impact of our proposed power control scheme on the estimation error of the global update Δw_k. Specifically, we compare the MSE of the Δw_k for different choices of M, and compare the AirReComp power control scheme to the baseline solutions of [19] and [21] wh ere the power control algorithm is unaware of the number of retransmissions. For this, we consider the transmission of randomly generated scalars instead of running a complete FL simulation setup. For this simulation, we consider $K = 20$ users and varying noise powers σ_z^2. To simulate the network, we generate channel coefficients according to unit Rayleigh fading $h_k \sim \mathcal{N}(0, 1/2) + j\mathcal{N}(0, 1/2)$ and additive noise components as $z \sim \mathcal{N}(0, \sigma_z^2)$. The transmitted scalars Δw_k are generated according to the unit normal distribution, which matches the assumption in Section III. The maximum transmission power is selected as $P_{\max} = \overline{P} = 1$, according to Remark 1. The PS estimate of the arithmetic mean $\Delta\hat{w}$ is generated according to (13), where the transmission powers p_k and the post-transmission scalar $\sqrt{\eta}$ are selected according

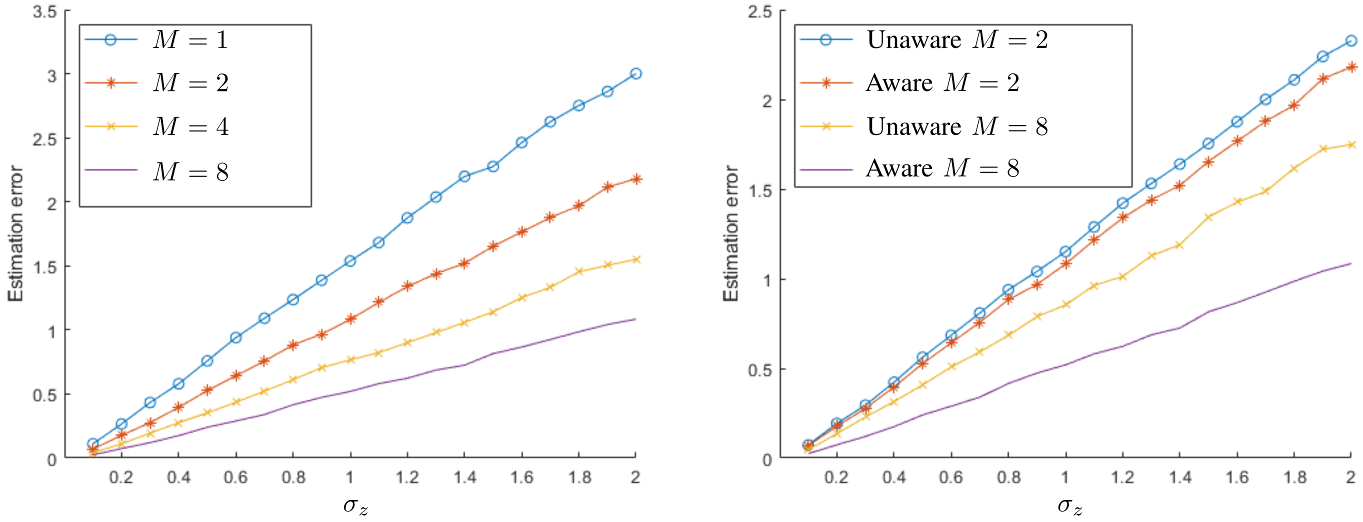


Fig. 1. Estimation error evaluation of AirReComp. We consider $K=20$ devices and evaluate the squared estimation error. Left: The estimation error of a single transmission ($M=1$) is compared to using retransmissions ($M>1$). Note that even though SNR scales linearly with the number of transmissions, the estimation error is not reduced as drastically. Right: The estimation error of optimal retransmission-aware power control is compared to a retransmission-unaware baseline. The results demonstrate the importance of designing the power control scheme with retransmissions in mind.

to Proposition 1. Upon calculating the estimate, it is compared to the true arithmetic mean of Δw_k according to

$$\text{MSE} = (\frac{1}{K}\sum_{k=1}^{K}\Delta w_k - \Delta\hat{w})^2. \tag{38}$$

This process is repeated 20,000 times for each value of σ_z^2. The resulting MSEs are averaged to form the plot in Fig. 1. In the left plot of the figure, the estimation errors using different numbers of transmissions M are illustrated. The plot demonstrates that the mean squared estimation error is approximately linear with the variance of the additive noise, regardless of M. In the right plot of Fig. 1, we compare the AirReComp power control scheme (Aware) to the retransmission-unaware scheme proposed in [19] and [21]. Their proposal is the optimal power control scheme for single retransmissions (Unaware). The numerical results demonstrate that AirReComp has a significantly lower estimation error when $M>1$. The gap between AirReComp and the baseline is also increasing with M, which demonstrates the importance of designing the power control scheme with retransmissions in mind. From the left plot, it is clear that the reduction in estimation error is worse than proportional to M. Instead, the system using $M=8$ achieves approximately three times lower estimation error that the baseline of $M=1$. Compared to using a forward error-correcting code, this result is significantly worse. However, since such codes are not compatible with analog communication, retransmissions are a good first step towards enabling a communication-estimation trade-off.

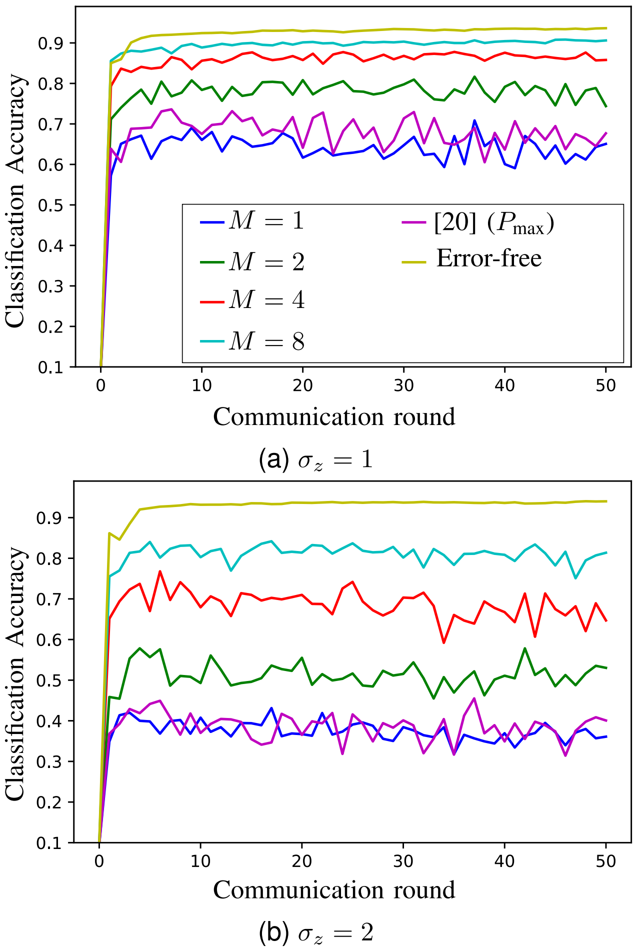
B. Federated Learning Convergence

In this subsection, we have two goals: to verify that the post-convergence classification accuracy is increasing in M for non-convex loss functions and to demonstrate the level of improvement compared to other baselines. For the FL simulation, the network setup is identical to Section V-A, except that σ_z^2 is fixed for each simulation and $K=10$. The ML task is multi-label classification on the MNIST dataset [42] with $|\mathcal{D}_k|=6000$ training samples per user device. The classifier is a DNN which consists of an input layer of 784 nodes, a hidden layer with 10 neurons, and an output layer of 10 neurons. The network is trained with a static learning rate of $\beta=0.1$, ReLU activation, sparse categorical cross-entropy loss, L2 regularization with $\epsilon=10^{-5}$, and without dropout. We run 2 epochs ($E=2$) per communication round, for $N=50$ rounds. The whole training process is repeated 10 times for each considered value of M, these results are then averaged to get the plots in Figs 2a and 2b.

AirReComp is compared to two baseline solutions. The first baseline (max power) is based on the scheme proposed in [20]. This scheme is a maximum power transmission scheme that does not require any channel information for the devices and thus allows for simple implementation. The second baseline (error-free) considers the case where there is no noise or fading and therefore that the server retrieves perfect copies of the model updates. Comparing AirReComp with the error-free baseline quantifies the performance gap caused by estimation errors of the model update aggregation.

In Figs 2a and 2b, the results of two simulations with $\sigma_z=1$ and $\sigma_z=2$ are presented. We wish to highlight that these simulations correspond to low SNR scenarios, because even though p_k has a maximum value of $P_{\max,n}=1$, the actual transmission power is much lower, as mentioned in Remark 1. In our simulations, we measured the average update norm to be $\mathbb{E}[\|\mathbf{w}_{n,k}\|^2]=329$. Because this is lower than the number of parameters in the model ($d=7960$), the average signal strength is less than 1. As a result, the average SNR was -5.3dB and -11.3dB for $\sigma_z=1$, and $\sigma_z=2$, respectively.

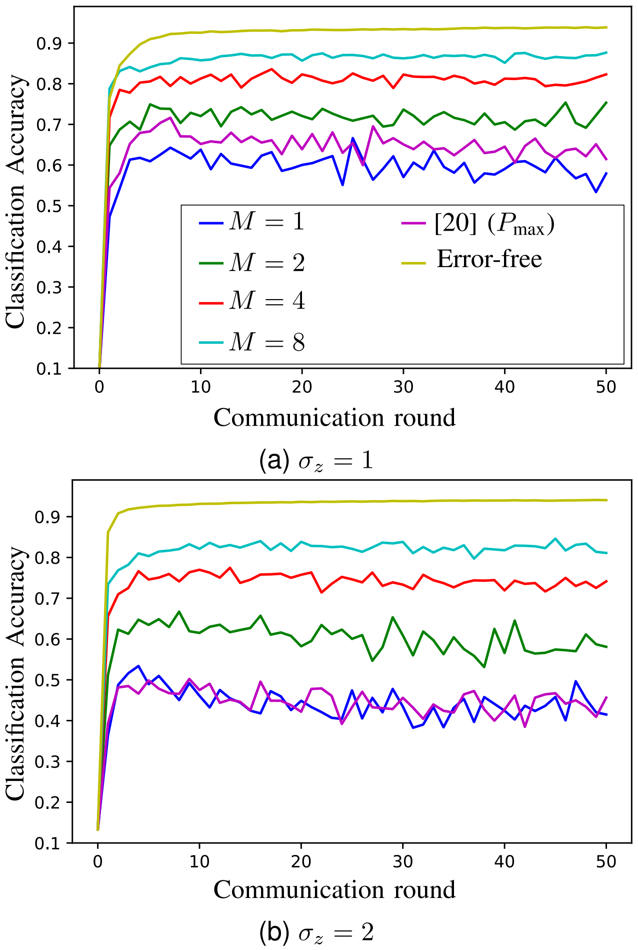
These results clearly demonstrate that the classification accuracy is improved as additional retransmissions are introduced, at least in low-SNR scenarios. While the convergence analysis in Section IV only holds for convex loss functions,



(a) $\sigma_z = 1$

(b) $\sigma_z = 2$

Fig. 2. Federated Learning performance with AirReComp. We consider $K = 20$ devices and train fully-connected DNNs over a multiple access channel with fading. We consider AirReComp with $M = 1, 2, 4, 8$ and two baselines. The first baseline corresponds to the max-power system and the second corresponds to the error-free system. Both plots correspond to low-SNR systems with varying levels of noise.



(a) $\sigma_z = 1$

(b) $\sigma_z = 2$

Fig. 3. Federated Learning performance with AirReComp. This simulation uses the same setup as in Figure 2 except that channels remain static throughout the training process. While the performance is not identical, the results demonstrate that the overall trend matches that of dynamic channels which change between communication rounds.

these results show that the method can also offer benefits for more complicated non-convex models, such as DNNs. Specifically, for $\sigma_z = 1$, the system with $M = 1$ achieved an average classification accuracy of 58%, while the best system with $M = 8$ achieved an average classification accuracy of 88%. The poor result of 58% is largely due to the low SNR of the network. This can also be seen by comparing Figs 2a and 2b, where the latter has a significantly wider gap between $M = 1$ and $M = 8$.

If we compare our proposal to the baselines, it is clear that the max-power basline performs closely to $M = 1$. This is unexpected since it does not perform any power control and therefore should be experiencing a worse MSE. However, it could potentially be explained by the assumption that $\mathbb{E}[(\Delta w_{n,k}^{(i)})^2] = 1$ in our power control scheme, which does not hold in practice. Alternatively, if the MSE improvement between the max-power baseline and $M = 1$ is minor, it might not have a noticeable effect on the classification accuracy when training with IID data, as suggested in [24].

While comparing to the error-free baseline, we notice that our proposal with $M = 8$ transmissions achieves 6% worse classification accuracy than perfect communication. This highlights the issue of estimation errors in FL performance and suggests that further improvements are necessary for low-SNR scenarios. One could always increase M, but at some point the increased communication cost causes digital communications to be a better alternative.

Finally, we provide a simulation for the case of static channels. The simulation setup is identical to that of Fig. 2 except that the same channel coefficients are used for all N communication rounds. As illustrated in Fig. 3, the classification accuracies are slightly worse for all systems (except for the error-free system) but the overall trend matches that of Fig. 2. A possible explanation for the performance decline is that, with static channels, any device that experiences a poor channel coefficient will consistently contribute less to the global update. Therefore, the knowledge contained in its dataset will be underrepresented, leading to model drift between its local model and the global model. Whereas in the dynamic case, where new channels are experienced for each communication round, the model drift would be corrected whenever a better channel is sampled.

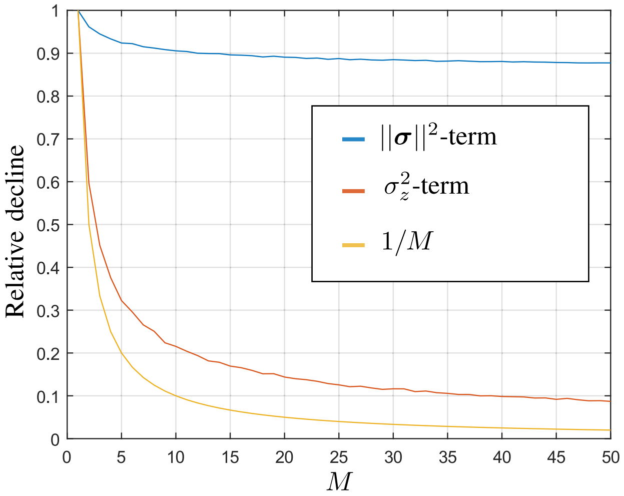


Fig. 4. The relative decline of the post-convergence terms in (32) and (34). The $\|\boldsymbol{\sigma}\|^2$-term refers to the term caused by the difference between the local and the global model updates, this term cannot be completely eliminated even with perfect communication, in contrast to the σ_z^2-term caused by the additive noise.

C. Convergence Bounds

In Section IV, we developed two bounds on the FL loss function and illustrated that the post-convergence terms are expected to decrease in M. However, due to the lack of a closed-form expression for the relationship between the transmission powers p_k and the number of uplink transmissions M, we were unable to provide the rate of decline with respect to M. Instead, we demonstrate the decline of these terms in this section. Specifically, there are two post-convergence terms for each bound, as expressed in (36).

In practice, the three learning-related variables $\|\boldsymbol{\sigma}\|^2$, L, and μ of the bounds are difficult to estimate. Therefore, rather than attempting to evaluate the absolute magnitude of the post-convergence terms, we are looking at the relative decline with respect to M. We define the relative decline as the quotient of the post-convergence term for M transmissions and the same term evaluated for one transmission, given by

$$\text{relative decline}(\|\boldsymbol{\sigma}\|^2) = \frac{\sum_{k=1}^{K} p_k(M)|h_k|^2}{\sum_{k=1}^{K} p_k(M=1)|h_k|^2} \cdot \frac{\left(\sum_{k=1}^{K} \sqrt{p_k(M=1)}|h_k|\right)^2}{\left(\sum_{k=1}^{K} \sqrt{p_k(M)}|h_k|\right)^2} \quad (39)$$

and

$$\text{relative decline}(\sigma_z^2) = \frac{\left(\sum_{k=1}^{K} \sqrt{p_k(M=1)}|h_k|\right)^2}{M\left(\sum_{k=1}^{K} \sqrt{p_k(M)}|h_k|\right)^2}, \quad (40)$$

where $p_k(M)$ is the transmission power of device k evaluated according to Proposition 1. For the simulation, we use the same network setup with $K = 20$ devices, Rayleigh fading, and $\sigma_z = 1$. We simulated the terms for 1,000 random realizations of the channels and averaged to get the results in Fig. 4.

As displayed in Fig. 4, the error caused by the difference between local and global gradients is hardly affected by introducing additional retransmissions. This is to be expected since the only improvement comes from the slight decrease of transmission powers that follow from an increased M, as highlighted in Corollary 1. The noise-induced error is however significantly improved, almost at the order of $\mathcal{O}(1/M)$, but with a gap due to the decreased transmission powers, as discussed in Section IV-C.

VI. Conclusion

In this paper, we propose retransmissions for Over-the-Air FL, in a system we call AirReComp. Arguably, this is the first work to enable a trade-off between communication resources and convergence speed for Over-the-Air FL. To improve the estimation error of AirReComp, we find a closed-form solution for optimal power control in the uplink. This power control solution shows that the number of retransmissions must be known by the transmitters to realize the MMSE estimator. We also prove two upper bounds on the FL loss for the AirReComp system, both for strongly-convex and convex loss functions. These bounds show that the post-convergence error of FL is strictly decreasing in the number of retransmissions, while the convergence rate is unaffected even though the estimation error of the updates is decreased. This contradicts to the findings of earlier works on AirComp [20], [30]. The reason is that those works do not normalize the update step, and thus the transmission scheme directly impacts the learning rate. We numerically verify the improved post-convergence performance for non-convex loss functions by training DNNs with AirReComp. The simulations also demonstrate that AirReComp can significantly outperform single uplink transmissions as well as full power baselines.

There is interesting open work on the reduction of estimation errors for Over-the-Air FL, including:

- **Gradient Statistics for Power Control** In a recent work [24], the authors proposed that online estimation of gradient statistics can significantly improve the power control of over-the-air FL. They found the optimal power control algorithm given that these statistics are known, but only for one-shot transmission. By combining this result with AirReComp, one could avoid the assumption that $\mathbb{E}[(\Delta w_{n,k}^{(i)})^2] = 1$ and find the optimal power control scheme for more realistic assumptions.
- **The consideration of fast-fading channels and diversity gains for AirReComp.** In this work, we consider a network with static channels, which restricts the improvements of retransmissions to reducing the power of the noise. In a fast-fading scenario, the problem changes substantially, especially the power control problem described in Section III. We have taken a first step in this direction in [34], where we show that by exploiting the ergodicity of the fast-fading channel, one can probabilistically guarantee unbiased over-the-air computation under peak transmission power constraints.
- **The consideration of non-IID data distributions** In this work, we only consider IID data distributions both in

the analytical and numerical results. It is likely that the importance of improved estimation is more pronounced when the datasets are non-IID, because the suppression of any individual update should cause greater harm to the convergence. There has been some work on over-the-air FL with non-IID data [43] but, as far as we are aware, no work that considers the gradient estimation.

- **Other methods of controlling the estimation error for Over-the-Air FL.** For instance, one could consider the possibility of a distributed channel code. With analog communication, this appears to be inapplicable, but with recent ideas of one-bit digital Over-the-Air Computation, there might be possibilities to explore in this direction [44], [45]. Additionally, one could consider combining the retransmission scheme with device selection for further improvements, as suggested in [46].
- **Tradeoff between transmission power and retransmissions.** Instead of focusing on adding retransmissions to improve the estimation error, one could consider changing the transmission power. Similar to the tradeoff explored in this work, there is a tradeoff between transmission power and convergence rate. Especially for low-powered IoT-devices, it would be interesting to analyze how much the transmission power could be reduced without significantly harming the FL performance.

Appendix A
Proof of Proposition 2

We start the proof by expressing the distance between the optimal global model $\mathbf{w}^*$ and the current global model $\mathbf{w}_n$ at communication round n as

$$r_n^2 := \|\mathbf{w}_n - \mathbf{w}^*\|^2. \tag{41}$$

This distance can be related to the FL loss function via Lemma 1 and Lemma 2. The plan for the proof is to utilize this relationship to form the upper bound. But before we get to that stage, we need to introduce the impact of AirReComp on the model update. To do so, we use (13) with (41) to express

$$\begin{aligned} r_{n+1}^2 &:= \|\mathbf{w}_n - \mathbf{w}^* - \frac{\Delta\hat{\mathbf{w}}_n}{c_1/K}\|^2 \\ &= r_n^2 - 2\frac{(\Delta\hat{\mathbf{w}}_n)^T}{c_1/K}(\mathbf{w}_n - \mathbf{w}^*) + \|\frac{\Delta\hat{\mathbf{w}}_n}{c_1/K}\|^2, \end{aligned} \tag{42}$$

where $\Delta\hat{\mathbf{w}}_n$ is the model update from (12) and c_1 is defined in (14). Next, we take the expectation of (42) with respect to $\Delta\mathbf{w}_{n,k}$ and $\mathbf{z}_m$. To do that, we first need to determine $\mathbb{E}\left[\Delta\hat{\mathbf{w}}_n\right]$ and $\mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right]$. Beginning with $\mathbb{E}\left[\Delta\hat{\mathbf{w}}_n\right]$, we use (12) to get

$$\begin{aligned} \mathbb{E}\left[\Delta\hat{\mathbf{w}}_n\right] &= \mathbb{E}[\Delta\mathbf{w}_n]\sum_{k=1}^{K}\frac{|h_k|\sqrt{p_k}}{\sqrt{\eta}K} \\ &= \frac{c_1}{K}\mathbb{E}[\Delta\mathbf{w}_n] = \beta\frac{c_1}{K}\mathbb{E}[\nabla F(\mathbf{w}_n)]. \end{aligned} \tag{43}$$

which has been simplified using Assumption 1 and the final equality holds since we assume there is only one epoch ($E = 1$) and therefore that the model update is the gradient of the global loss function. Next, we find $\mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right]$, once again using (12)

$$\begin{aligned} \mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right] &= \frac{1}{K^2\eta}\mathbb{E}\left[\|\sum_{k=1}^{K}\sqrt{p_k}|h_k|\Delta\mathbf{w}_{n,k}\|^2\right] \\ &\quad + \frac{1}{M^2K^2\eta}\mathbb{E}\left[\|\sum_{m=1}^{M}\mathrm{Re}(\mathbf{z}_m)\|^2\right]. \end{aligned} \tag{44}$$

The first term of (44) can be upper-bounded by the Cauchy-Schwartz inequality as follows

$$\|\sum_{k=1}^{K}\sqrt{p_k}|h_k|\Delta\mathbf{w}_{n,k}\|^2 \leq K\sum_{i=1}^{d}\left(\sum_{k=1}^{K}p_k|h_k|^2(\Delta w_{n,k}^{(i)})^2\right). \tag{45}$$

Then, we apply our assumption on the local model updates from (27) to get

$$\begin{aligned} \mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right] &\leq \frac{\beta^2}{K\eta}\sum_{k=1}^{K}p_k|h_k|^2(\|\boldsymbol{\sigma}\|^2 + \mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right]) \\ &\quad + \frac{d\sigma_z^2}{MK^2\eta}. \end{aligned} \tag{46}$$

With $\mathbb{E}\left[\Delta\hat{\mathbf{w}}_n\right]$ and $\mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right]$ evaluated in (43) and (46), we go back to the model distance. Taking the expectation on both sides of (42) yields

$$\begin{aligned} \mathbb{E}[r_{n+1}^2] &\leq \mathbb{E}[r_n^2] - 2\beta\mathbb{E}[\nabla F(\mathbf{w}_n)^T(\mathbf{w}_n - \mathbf{w}^*)] \\ &\quad + \frac{\beta^2K\sum_{k=1}^{K}p_k|h_k|^2}{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}(\|\boldsymbol{\sigma}\|^2 + \mathbb{E}[\|\nabla F(\mathbf{w}_n)\|^2]) \\ &\quad + \frac{d\sigma_z^2}{M\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}. \end{aligned} \tag{47}$$

Now we are ready to introduce the FL loss by utilizing strong convexity and Lipschitz smoothness. We do this by rewriting Lemma 2 to

$$\begin{aligned} \mathbb{E}\left[\nabla F(\mathbf{w}_n)^T(\mathbf{w}_n - \mathbf{w}^*)\right] &\geq \frac{\mu L}{\mu + L}\mathbb{E}[r_n^2] \\ &\quad + \frac{1}{\mu + L}\mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right], \end{aligned} \tag{48}$$

where we have utilized $\nabla F(\mathbf{w}^*) = 0$ for the final term on the RHS. Combining (47) and (48) yields

$$\begin{aligned} \mathbb{E}[r_{n+1}^2] &\leq \mathbb{E}[r_n^2] \\ &\quad - 2\beta\left(\frac{\mu L}{\mu + L}\mathbb{E}[r_n^2] + \frac{1}{\mu + L}\mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right]\right) \\ &\quad + \frac{\beta^2K\sum_{k=1}^{K}p_k|h_k|^2}{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}(\|\boldsymbol{\sigma}\|^2 + \mathbb{E}[\|\nabla F(\mathbf{w}_n)\|^2]) \\ &\quad + \frac{d\sigma_z^2}{M\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}. \end{aligned} \tag{49}$$

Since this expression is getting long, we use three constants c_2, c_3, and c_4 to simplify it. The model distance is then

$$\mathbb{E}[r_{n+1}^2] \le c_2\mathbb{E}[r_n^2] + \frac{1}{\left(\sum_{k=1}^{k}\sqrt{p_k}|h_k|\right)^2}c_3 + \beta c_4\mathbb{E}[\|\nabla F(\mathbf{w}_n)\|^2], \tag{50}$$

where c_2 and c_3 were defined in (30) and (31) respectively. Because of our choice of learning rate in Assumption 2, we have the following inequality for c_4

$$c_4 := \beta K\frac{\sum_{k=1}^{K} p_k|h_k|^2}{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2} - \frac{2}{\mu+L} < 0. \tag{51}$$

Since c_4 is less than zero, we can rewrite our bound in (50) as

$$\mathbb{E}[r_{n+1}^2] \le c_2\mathbb{E}[r_n^2] + \frac{1}{\left(\sum_{k=1}^{k}\sqrt{p_k}|h_k|\right)^2}c_3. \tag{52}$$

At this point, the bound is almost complete. The only thing that remains is to find an inequality comparing $\mathbb{E}[r_n^2]$ and $\mathbb{E}[r_0^2]$ instead of comparing two adjacent communication rounds. As such, we reduce the communication round counter by one and replace $\mathbb{E}[r_n^2]$ in (52) to get

$$\mathbb{E}[r_{n+1}^2] \le c_2^2\mathbb{E}[r_{n-1}^2] + (c_2+1)\frac{1}{\left(\sum_{k=1}^{k}\sqrt{p_k}|h_k|\right)^2}c_3. \tag{53}$$

By induction we have

$$\mathbb{E}[r_n^2] \le c_2^n\mathbb{E}[r_0^2] + \frac{1}{\left(\sum_{k=1}^{k}\sqrt{p_k}|h_k|\right)^2}c_3\sum_{i=0}^{n-1}c_2^i. \tag{54}$$

Then we apply $\sum_{i=0}^{n-1} c_2^i < \sum_{i=0}^{\infty} c_2^i = 1/(1-c_2)$ to achieve

$$\mathbb{E}[r_n^2] \le c_2^n\mathbb{E}[r_0^2] + \frac{c_3}{\left(\sum_{k=1}^{k}\sqrt{p_k}|h_k|\right)^2(1-c_2)}. \tag{55}$$

Finally, we utilize convexity and Lipschitz smoothness from (24) to relate the LHS of (55) to the FL loss, which yields

$$\mathbb{E}\left[F(\mathbf{w}_n)\right] - F(\mathbf{w}^*) \le \frac{L}{2}c_2^n\mathbb{E}[r_0^2] + \frac{Lc_3}{2\left(\sum_{k=1}^{k}\sqrt{p_k}|h_k|\right)^2(1-c_2)}, \tag{56}$$

which is the bound from Proposition 2. □

Appendix B
Proof of Proposition 3

Just as in the first proof, we utilize the properties of convexity and Lipschitz smoothness to relate the distance between the optimal global model $\mathbf{w}^*$ and the current global model $\mathbf{w}_n$ to the FL loss function. In contrast to the first proof, we use these properties immediately. Specifically, we start with Lemma 1 and take the expectation on both sides to get

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \le{}& \mathbb{E}\left[F(\mathbf{w}_n)\right] \\ &+ \mathbb{E}\left[\nabla F(\mathbf{w}_n)^T(\mathbf{w}_{n+1}-\mathbf{w}_n)\right] \\ &+ \frac{L}{2}\mathbb{E}\left[\|\mathbf{w}_{n+1}-\mathbf{w}_n\|^2\right].\end{aligned} \tag{57}$$

Then, we add the global update via (13) to get

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \le{}& \mathbb{E}\left[F(\mathbf{w}_n)\right] \\ &- \frac{1}{c_1/K}\mathbb{E}\left[\nabla F(\mathbf{w}_n)^T\Delta\hat{\mathbf{w}}_n\right] \\ &+ \frac{L}{2(c_1/K)^2}\mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right].\end{aligned} \tag{58}$$

Next, (43) gives us

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \le{}& \mathbb{E}\left[F(\mathbf{w}_n)\right] \\ &- \beta\mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right] \\ &+ \frac{L}{2(c_1/K)^2}\mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right].\end{aligned} \tag{59}$$

We insert the bound for $\mathbb{E}\left[\|\Delta\hat{\mathbf{w}}_n\|^2\right]$ from (46) to get

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \\ &\le \mathbb{E}\left[F(\mathbf{w}_n)\right] - \beta\mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right] \\ &\quad + \frac{L}{2}\frac{\beta^2 K\sum_{k=1}^{K}p_k|h_k|^2}{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}(\|\boldsymbol{\sigma}\|^2 + \mathbb{E}[\|\nabla F(\mathbf{w}_n)\|^2]) \\ &\quad + \frac{L}{2}\frac{d\sigma_z^2}{M\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}.\end{aligned} \tag{60}$$

Next, we recognize c_3 from (31) and substitute it into the bound

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \\ &\le \mathbb{E}\left[F(\mathbf{w}_n)\right] \\ &\quad - \beta\left(1 - \frac{KL\beta}{2}\frac{\sum_{k=1}^{K}p_k|h_k|^2}{\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}\right)\mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right] \\ &\quad + \frac{L}{2\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}c_3.\end{aligned} \tag{61}$$

This expression can be simplified by using Assumption 3 to bound the second term on the RHS

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \le{}& \mathbb{E}\left[F(\mathbf{w}_n)\right] \\ &- \frac{\beta}{2}\mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right] \\ &+ \frac{L}{2\left(\sum_{k=1}^{K}\sqrt{p_k}|h_k|\right)^2}c_3.\end{aligned} \tag{62}$$

Next, we are going to upper bound the first term on the RHS of (62). We use the following standard property of convexity (see equation 2.1.2 from [40]):

$$\mathbb{E}[F(\mathbf{w}_n)] \le F(\mathbf{w}^*) + \mathbb{E}[\nabla F(\mathbf{w}_n)^T(\mathbf{w}_n - \mathbf{w}^*)]. \tag{63}$$

Plug this into (62)

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \le F(\mathbf{w}^*) &+ \mathbb{E}[\nabla F(\mathbf{w}_n)^T(\mathbf{w}_n-\mathbf{w}^*)]\\ &- \frac{\beta}{2}\mathbb{E}\left[\|\nabla F(\mathbf{w}_n)\|^2\right]\\ &+ \frac{L}{2\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3.\end{aligned}\tag{64}$$

Some tedious algebraic manipulation transforms (64) to

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{w}_{n+1})\right]\\ &\le F(\mathbf{w}^*)\\ &\quad+\frac{1}{2\beta}\mathbb{E}\left[\|\mathbf{w}_n-\mathbf{w}^*\|^2-\|(\mathbf{w}_n-\mathbf{w}^*)-\beta\nabla F(\mathbf{w}_n)\|^2\right]\\ &\quad+\frac{L}{2\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3.\end{aligned}\tag{65}$$

Let $r_n^2 := \|\mathbf{w}_n-\mathbf{w}^*\|^2$ (same as (41)) and

$$\tilde{r}_{n+1}^2 := \|\mathbf{w}_n-\mathbf{w}^*-\beta\nabla F(\mathbf{w}_n)\|^2.\tag{66}$$

Then (65) becomes

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{n+1})\right] \le F(\mathbf{w}^*) &+ \frac{1}{2\beta}\mathbb{E}\left[r_n^2-\tilde{r}_{n+1}^2\right]\\ &+\frac{L}{2\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3.\end{aligned}\tag{67}$$

Now, just like in the previous proof, we want to form a bound with respect to r_0^2. However, instead of using induction, we use a telescoping sum. To set it up, we start by taking a sum of (67) over n communication rounds to get

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{w}_i)\right]-nF(\mathbf{w}^*) \le \frac{1}{2\beta}&\sum_{i=1}^n \mathbb{E}\left[r_{i-1}^2-\tilde{r}_i^2\right]\\ &+\frac{nL}{2\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3.\end{aligned}\tag{68}$$

The sum $\sum_{i=1}^n \mathbb{E}\left[r_{i-1}^2-\tilde{r}_i^2\right]$ can be rewritten as

$$\sum_{i=1}^n \mathbb{E}\left[r_{i-1}^2-\tilde{r}_i^2\right] = \mathbb{E}[r_0^2]-\mathbb{E}[\tilde{r}_n^2]+\sum_{i=0}^{n-2}\mathbb{E}[r_{i+1}^2-\tilde{r}_{i+1}^2]\tag{69}$$

and the middle terms $\sum_{i=0}^{n-2}\mathbb{E}[r_{i+1}^2-\tilde{r}_{i+1}^2]$ will be upper bounded to a constant. We develop this bound next. To start, we plug in the definition of r_{i+1}^2 and $\tilde{r}_{i+1}^2$ into $\mathbb{E}[r_{i+1}^2-\tilde{r}_{i+1}^2]$:

$$\begin{aligned}&\mathbb{E}\left[r_{i+1}^2-\tilde{r}_{i+1}^2\right]\\ &\quad=\mathbb{E}\Big[\|\mathbf{w}_{i+1}-\mathbf{w}^*\|^2-\|\mathbf{w}_i-\mathbf{w}^*-\beta\nabla F(\mathbf{w}_i)\|^2\Big].\end{aligned}\tag{70}$$

Applying (13) and doing some algebra yields

$$\begin{aligned}&\mathbb{E}\left[r_{i+1}^2-\tilde{r}_{i+1}^2\right]\\ &\quad=\frac{1}{c_1/K}\mathbb{E}\left[\frac{1}{c_1/K}\|\Delta\hat{\mathbf{w}}_i\|^2-2(\mathbf{w}_i-\mathbf{w}^*)^T\Delta\hat{\mathbf{w}}_i\right]\\ &\qquad+\beta\mathbb{E}\Big[2(\mathbf{w}_i-\mathbf{w}^*)^T\nabla F(\mathbf{w}_i)-\beta\|\nabla F(\mathbf{w}_i)\|^2\Big].\end{aligned}\tag{71}$$

Then we apply (43) to get

$$\mathbb{E}\left[r_{i+1}^2-\tilde{r}_{i+1}^2\right]=\mathbb{E}\left[\frac{\|\Delta\hat{\mathbf{w}}_i\|^2}{(c_1/K)^2}-\beta^2\|\nabla F(\mathbf{w}_i)\|^2\right].\tag{72}$$

Next, we insert $\|\Delta\hat{\mathbf{w}}_i\|^2$ from (44), apply Assumption 1, and do some algebra which yields

$$\begin{aligned}&\mathbb{E}\left[r_{i+1}^2-\tilde{r}_{i+1}^2\right]\\ &\quad=\mathbb{E}\Bigg[\beta^2\|\boldsymbol{\sigma}\|^2+\|\Delta\mathbf{w}_n\|^2\\ &\qquad+\frac{d\sigma_z^2}{M\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}-\beta^2\|\nabla F(\mathbf{w}_i)\|^2\Bigg].\end{aligned}\tag{73}$$

Since we assume $E = 1$, we have $\mathbb{E}[\|\Delta\mathbf{w}_n\|^2] = \beta^2\mathbb{E}[\|\nabla F(\mathbf{w}_n)\|^2]$, which yields

$$\begin{aligned}&\mathbb{E}\left[r_{i+1}^2-\tilde{r}_{i+1}^2\right]\\ &\quad=\frac{1}{\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}\left(\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2\beta^2\|\boldsymbol{\sigma}\|^2+\frac{d\sigma_z^2}{M}\right).\end{aligned}\tag{74}$$

Finally, we apply the Cauchy-Schwarz inequality to get

$$\mathbb{E}\left[r_{i+1}^2-\tilde{r}_{i+1}^2\right]\le\frac{c_3}{\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}\tag{75}$$

That concludes the upper bound on $\mathbb{E}\left[r_{i+1}^2-\tilde{r}_{i+1}^2\right]$ so we plug it back into (68) to get

$$\begin{aligned}&\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{w}_i)\right]-nF(\mathbf{w}^*)\\ &\quad\le\frac{nL}{2\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3\\ &\qquad+\frac{1}{2\beta}\mathbb{E}\left[r_0^2-\tilde{r}_n^2+\frac{(n-1)}{\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3\right].\end{aligned}\tag{76}$$

Since $\tilde{r}_n^2$ is positive, we can add one to the RHS of (76) without breaking the inequality. Similarly, we can add $c_3/\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2$ to the RHS, which together yields

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{w}_i)\right]-nF(\mathbf{w}^*) \le \frac{1}{2\beta}&\mathbb{E}\left[r_0^2\right]\\ &+\frac{2n+nL}{2\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3.\end{aligned}\tag{77}$$

Finally, we note that $\mathbb{E}\left[F(\mathbf{w}_n)\right]\le\mathbb{E}\left[F(\mathbf{w}_i)\right]$ for all $i\le n$. Therefore

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_n)\right]-F(\mathbf{w}^*) &\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{w}_i)\right]-F(\mathbf{w}^*)\\ &\le\frac{1}{2n\beta}\mathbb{E}\left[r_0^2\right]+\frac{2+L}{2\left(\sum_{k=1}^K \sqrt{p_k}|h_k|\right)^2}c_3,\end{aligned}\tag{78}$$

which is the bound from Proposition 3. □

References

[1] *Ericsson IoT Connections Outlook: Broadband IoT Set to Overtake 2G and 3G*. Accessed: Aug. 24, 2021. [Online]. Available: https://www.ericsson.com/en/mobility-report/dataforecasts/iot-connections-outlook

[2] *Pew Research Center: U.S. Technology Device Ownership 2015*. Accessed: Aug. 24, 2021. [Online]. Available: https://www.pewresearch.org/internet/2015/10/29/technology-device-ownership-2015/

[3] J. Jagannathan and U. Udaykumar, "Predictive modeling for improving healthcare using IoT: Role of predictive models in healthcare using IoT," in *Incorporating the Internet of Things in Healthcare Applications and Wearable Devices*. Hershey, PA, USA: IGI Global, 2020, pp. 243–254.

[4] B. Brik, A. Ksentini, and M. Bouaziz, "Federated learning for UAVs-enabled wireless networks: Use cases, challenges, and open problems," *IEEE Access*, vol. 8, pp. 53841–53849, 2020.

[5] L. Knoll, L. Breuer, and M. Bach, "Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning," *Sci. Total Environ.*, vol. 668, pp. 1317–1327, Jun. 2019.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[7] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, 2019.

[8] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2017, *arXiv:1610.05492*.

[9] H. Hellström et al., "Wireless for machine learning: A survey," *Found. Trends Signal Process.*, vol. 15, no. 4, pp. 290–399, 2022.

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

[11] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

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

[13] 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, Mar. 2020.

[14] M. Goldenbaum, H. Boche, and S. Stańczak, "Harnessing interference for analog function computation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 4893–4906, Oct. 2013.

[15] M. Goldenbaum, H. Boche, and S. Stańczak, "Analyzing the space of functions analog-computable via wireless multiple-access channels," in *Proc. IEEE 8th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Nov. 2011, pp. 779–783.

[16] B. Hasırcıoğlu and D. Gündüz, "Private wireless federated learning with anonymous over-the-air computation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5195–5199.

[17] M. Goldenbaum and S. Stanczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3863–3877, Sep. 2013.

[18] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jan. 1948.

[19] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7498–7513, Nov. 2020.

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

[21] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimization, analysis and scaling laws," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5488–5502, Aug. 2020.

[22] X. Li, G. Zhu, Y. Gong, and K. Huang, "Wirelessly powered data aggregation for IoT via over-the-air function computation: Beamforming and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3437–3452, Jul. 2019.

[23] J. M. Barros da Silva, K. Ntougias, I. Krikidis, G. Fodor, and C. Fischione, "Simultaneous wireless information and power transfer for federated learning," in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 296–300.

[24] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commu.*, vol. 20, no. 8, pp. 5115–5128, Aug. 2021.

[25] D. Fan, X. Yuan, and Y.-J. A. Zhang, "Temporal-structure-assisted gradient aggregation for over-the-air federated edge learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3757–3771, Oct. 2021.

[26] T. Jiang and Y. Shi, "Over-the-air computation via intelligent reflecting surfaces," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[27] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7595–7609, Jun. 2021.

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

[29] A. J. Goldsmith and S.-G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 595–602, May 1998.

[30] 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, Nov. 2021.

[31] D. Sexton, M. Mahony, M. Lapinski, and J. Werb, "Radio channel quality in industrial wireless sensor networks," in *Proc. Sensors Ind. Conf.*, Feb. 2005, pp. 88–94.

[32] S. Triverdi. *Wi-Fi 6 OFDMA: Resource Unit (RU) Allocations and Mappings*. Accessed: Jan. 10, 2022. [Online]. Available: https://www.blogs.cisco.com/networking/wi-fi-6-ofdma-resource-unit-ru-allocations-and-mappings

[33] N. Ravindranath, I. Singh, A. Prasad, and V. Rao, "Performance evaluation of IEEE 802.11 ac and 802.11 n using NS3," *Indian J. Sci. Technol.*, vol. 9, no. 26, pp. 1–8, Jul. 2016.

[34] H. Hellström, V. Fodor, and C. Fischione, "Unbiased over-the-air computation via retransmissions," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2022, pp. 782–787.

[35] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2019, pp. 1387–1395.

[36] S. Prakash et al., "Coded computing for low-latency federated learning over wireless edge networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 233–250, Jan. 2021.

[37] Z. Lin, H. Liu, and Y.-J.-A. Zhang, "Relay-assisted over-the-air federated learning," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2021, pp. 1–7.

[38] 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. 2019.

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

[40] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87. New York, NY, USA: Springer, 2003. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4419-8853-9#about-this-book

[41] R. Meka, "CS289ML: Notes on convergence of gradient descent," Dept. CS, Univ. California, Los Angeles, Los Angeles, CA, USA, Lect. Notes CS289ML, 2017. [Online]. Available: https://raghumeka.github.io/CS289ML/index.html

[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

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

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

[45] A. Şahin and R. Yang, "Over-the-air computation over balanced numerals," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 347–352.

[46] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4434–4449, Dec. 2021.

Henrik Hellström (Graduate Student Member, IEEE) received the bachelor's degree in electrical engineering, the master's degree in information and network engineering, and the Licentiate degree in electrical engineering from the KTH Royal Institute of Technology, in 2016, 2018, and 2022, respectively, where he is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science.

Previously, he was a temporary Contractor with ABB Corporate Research, assisting in the development of a novel PHY-layer protocol for ultra-low latency industrial communications called WirelessHP. His research interests include physical layer wireless, distributed machine learning, software-defined radio, and industrial communications. His current research focus is wireless for machine learning, i.e., wireless network protocols that are tailored to support distributed machine learning. Particularly, he focuses on over-the-air computation to combine the gradients of deep neural networks by utilizing the superposition property of electromagnetic waves. He serves as the Secretary for the IEEE Emerging Technology Initiative on Machine Learning for Communications (ETI MLC).

Viktoria Fodor (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer engineering from the Budapest University of Technology and Economics, Budapest, Hungary, in 1992 and 1999, respectively, and the Habilitation degree (docent) from the KTH Royal Institute of Technology, Sweden, in 2011. She is a Professor of communication networks with the KTH Royal Institute of Technology. In 1998, she was a Senior Researcher with Hungarian Telecommunication Company. Since 1999, she has been with KTH Royal Institute of Technology. She has published more than 100 scientific publications. Her current research interests include performance evaluation of networks and distributed systems, stochastic modeling, and protocol design, with focus on edge computing and machine learning over networks. She is an Associate Editor of IEEE TRANSACTIONS OF NETWORK AND SERVICE MANAGEMENT.

Carlo Fischione (Senior Member, IEEE) received the Laurea degree (summa cum laude) in electronic engineering and the Ph.D. degree in electrical and information engineering from the University of L'Aquila, Italy, in April 2001 and May 2005, respectively. He is a Full Professor of electrical engineering and computer science with the Division of Network and Systems Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. He is also the Director of the KTH-Ericsson Data Science Micro Degree Program directed to Ericsson globally. He is an Honorary Professor with the Department of Mathematics, Information Engineering, and Computer Science, University of L'Aquila. He has held research positions with the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015, as a Visiting Professor; Harvard University, Cambridge, MA, in 2015, as an Associate; and the University of California at Berkeley, CA, USA, from 2004 to 2005, as a Visiting Scholar, and from 2007 to 2008, as a Research Associate. He has coauthored over 250 publications, including two books, book chapters, international journals and conferences, and international patents. His research interests include applied optimization, wireless, sensor networks, the Internet of Things, and machine learning.

He is an Ordinary Member of the Italian Academy of History Deputazione Abruzzese di Storia Patria (DASP). He received a number of awards, such as the IEEE Communication Society S. O. Rice Award for the Best IEEE TRANSACTIONS ON COMMUNICATIONS Paper of 2018, the Best Paper Award of IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS in 2007, and the Starting Grant of the Swedish Research Council in 2007. He is the Chair of the IEEE Machine Learning for Communications Emerging Technologies Initiative and the Founding General Chair of the IEEE Communication Society flagship conference IEEE International Conference on Machine Learning for Communications and Networking (IEEE ICMLCN). He is an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS (machine learning for communications area) and IEEE TRANSACTIONS ON MACHINE LEARNING FOR COMMUNICATION AND NETWORKING. He was served as an Associate Editor for IFAC *Automatica*, from 2014 to 2019. He a Distinguished Lecturer of the IEEE Communication Society. He is a Co-Founder and a Scientific Advisor of ELK Audio.