Fast Convergence Algorithm for Analog Federated Learning

Shuhao Xia† , Jingyang Zhu†, Yuhan Yang†, Yong Zhou†, Yuanming Shi† and Wei Chen*
† School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
* Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, 100084, China
Email: xiashh@shanghaitech.edu.cn, shiym@shanghaitech.edu.cn

ICC 2021 - IEEE International Conference on Communications | 978-1-7281-7122-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICC42927.2021.9500875

***Abstract*—In this paper, we consider federated learning (FL) over a noisy fading multiple access channel (MAC), where an edge server aggregates the local models transmitted by multiple end devices through over-the-air computation (AirComp). To realize efficient analog federated learning over wireless channels, we propose an AirComp-based FedSplit algorithm, where a threshold-based device selection scheme is adopted to achieve reliable local model uploading. In particular, we analyze the performance of the proposed algorithm and prove that the proposed algorithm linearly converges to the optimal solutions under the assumption that the objective function is strongly convex and smooth. We also characterize the robustness of proposed algorithm to the ill-conditioned problems, thereby achieving fast convergence rates and reducing communication rounds. A finite error bound is further provided to reveal the relationship between the convergence behavior and the channel fading and noise. Our algorithm is theoretically and experimentally verified to be much more robust to the ill-conditioned problems with faster convergence compared with other benchmark FL algorithms.**

I. Introduction

As an emerging decentralized machine learning solution, federated learning (FL) has recently attracted considerable attention from both academia and industry. In FL, multiple devices with their local datasets collaboratively train a global model, where only local model updates instead of raw data are transmitted to the parameter server, thereby significantly reducing the bandwidth requirement and providing additional privacy protection [1]–[4]. Most of the existing studies on FL focused on the reduction of the volume of model exchange without explicitly taking into account the impact of wireless channels. However, FL has a wide range of applications in wireless networks, e.g., Internet of Things (IoT) [5], autonomous driving [6]. Therefore, it is essential to investigate the impact of the physical characteristics (e.g., channel distortion, noise) of the wireless medium on the convergence rate and the optimality of FL algorithms.

Many digital communication based approaches have recently been proposed to facilitate FL in wireless networks [7]–[11], where each edge device is assigned an orthogonal channel to upload its local model. In particular, the authors in [7] formulated a joint resource allocation and user selection problem for FL, where a closed-form expression for the expected convergence rate of the FL was derived to establish an explicit relationship between the packet error rates and the FL performance. To reduce the total learning-and-communication latency, the authors in [8] partitioned the learning task into multiple sub-tasks, which are allocated to different edge devices for parallel training. In addition, the authors in [9], [10] further enhanced the communication efficiency of wireless federated learning systems by proposing efficient resource management mechanisms. As the orthogonal channels are required to enable concurrent local model uploading to avoid interference, the aforementioned studies may not be communication-efficient, especially when the number of edge devices is large.

To support communication-efficient design, over-the-air computation (AirComp), as a promising analog multiple access scheme, is capable of achieving ultra-fast model aggregation for FL by allowing concurrent transmission from edge devices over the same frequency channel and exploiting the waveform superposition property of MAC [12]–[17]. In particular, the authors in [12] proposed a joint device selection and beamforming design to accelerate the convergence of analog federated learning. In [13], a broadband analog aggregation scheme over MAC was proposed to reduce the communication latency. However, these studies did not analyze the convergence performance of FL algorithms. On the other hand, the authors in [14] proposed a distributed stochastic gradient descent (SGD) algorithm, in which each device transmits a sparse gradient estimate over MAC. In [15], the authors developed the gradient based multiple access (GBMA) algorithm, which is proved to achieve the same convergence rate as the centralized gradient descent (GD) algorithm in large-scale networks. Although the convergence analysis was provided, the aforementioned studies suffer from a high iteration complexity and high communication overhead under the ill-conditioned setting, which is well-known to be a performance-limiting factor. Very recently, the authors in [18] developed the FedSplit algorithm based on the operator splitting procedure to achieve fast convergence even in the ill-conditioned setting. However, both the convergence rate and the optimality of the FedSplit algorithm in wireless networks has not been studied, which motivates this work.

In this paper, we consider a wireless FL problem over a noisy fading MAC. Due to the distortion and noise caused by MAC, the performance of FL algorithms over wireless channels are significantly degraded, especially under the ill-conditioned setting. To address these issues, we propose an AirComp-based FedSplit algorithm, in which the edge server aims to recover the aggregation of local models computed by the end devices via AirComp at each communication round. We exploit a threshold-based device selection scheme to achieve reliable communication. For strongly convex and smooth local loss functions, we prove that the proposed algorithm can linearly converge to optimal points. Furthermore, we establish an error bound in term of the expected loss of the objective function to reveal the impact of channel fading and noise over convergence behavior. Finally, our theoretical results are well verified through numerical experiments under various parameter settings.

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant U20A20159.

978-1-7281-7122-7/21/$31.00 ©2021 IEEE

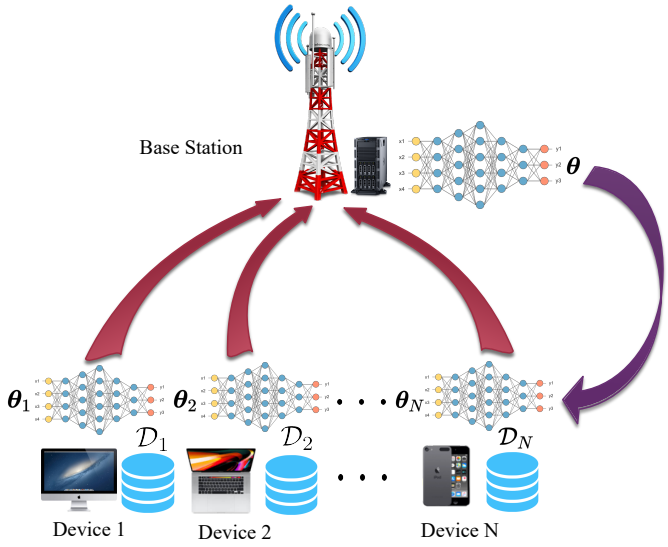

Fig. 1. Illustration of the wireless federated edge learning system consisting of N edge devices and one edge server.

Notations: All vectors are considered to be column vectors. We use boldface lowercase (uppercase) letters to represent vectors (matrices). We denote the identity matrix by $\mathbf{I}$, the set of real values by $\mathbb{R}$, the cardinality of set A by $|A|$ and ℓ_2-norm of vector $\boldsymbol{x}$ by $\|\boldsymbol{x}\|$. In addition, the function f is defined to be ℓ-strongly convex, if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \frac{\ell}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$$

for all $\boldsymbol{x}, \boldsymbol{y}$. Similarly, the function f is defined to be L-smooth, if

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$$

for all $\boldsymbol{x}, \boldsymbol{y}$.

II. System Model and Problem Formulation

A. Federated Optimization

We consider a federated edge learning system consisting of N single-antenna edge devices indexed by set $\mathcal{N} = \{1, 2, \ldots, N\}$ and a computing enabled edge server equipped with a single antenna, as illustrated in Fig. 1. Each device n is associated with its own local dataset $\mathcal{D}_n$, and all edge devices collaboratively learn a shared global model by communicating with the edge server.

In federated learning systems, the goal is to learn a shared global model by minimizing the sum of the devices' local loss function. Therefore, the problem can be formulated as the following consensus federated optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\theta}, \{\boldsymbol{\theta}_n\}_{n=1}^N}{\text{minimize}} \quad & F(\boldsymbol{\theta}) \triangleq \textstyle\sum_{n=1}^N f_n(\boldsymbol{\theta}) \\ \text{subject to} \quad & \boldsymbol{\theta}_n = \boldsymbol{\theta}, \forall n \in \mathcal{N}, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the global model with dimension d. For each device n, $\boldsymbol{\theta}_n \in \mathbb{R}^d$ is the local model and f_n is the local loss function defined by the learning task and the local dataset $\mathcal{D}_n$.

B. FedSplit

To solve problem (1), we adopt the FedSplit algorithm proposed in [18], which is based on the operator splitting procedures. As depicted in Fig. 1, at each t-th communication round, the edge server broadcasts the current global model $\boldsymbol{\theta}^t$ to all edge devices via the downlink channel through the digital communication. Hence, the power constraint of the edge server is not as strict as the edge devices, and the downlink communication is assumed to be error-free [14], [15], [19]. Consequently, each device receives the current shared global model $\boldsymbol{\theta}^t$ without distortion. Based on the received current global model $\boldsymbol{\theta}^t$ and the local dataset $\mathcal{D}_n$, device n updates the local model $\boldsymbol{\theta}_n^{t+1}$ with two steps as follows.

1) Local prox step:
$$\boldsymbol{\theta}_n^{t+1/2} \triangleq \text{prox}_{s,n}(2\boldsymbol{\theta}^t - \boldsymbol{\theta}_n^t), \tag{2}$$
2) Local centering step:
$$\boldsymbol{\theta}_n^{t+1} \triangleq \boldsymbol{\theta}_n^t + 2(\boldsymbol{\theta}_n^{t+1/2} - \boldsymbol{\theta}^t), \tag{3}$$

where the proximal operator $\text{prox}_{s,n}$ is defined by

$$\text{prox}_{s,n}(\boldsymbol{z}) \triangleq \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\min}\ f_n(\boldsymbol{x}) + \frac{1}{2s}\|\boldsymbol{z}-\boldsymbol{x}\|_2^2, \tag{4}$$

for some step size $s > 0$.

After updating local models, the edge devices transmit a function of the local model over a wireless fading multiple access channel (MAC) to the edge server. By exploiting over-the-air computation (AirComp), the edge server aggregates all devices local models in one channel use [12], which significantly reduces communication latency. Based on the received signals, the edge server is able to obtain an estimate $\hat{\boldsymbol{\theta}}^{t+1}$ of the average global model

$$\boldsymbol{\theta}^{t+1} \triangleq \frac{1}{N}\sum_{n=1}^N \boldsymbol{\theta}_n^{t+1}. \tag{5}$$

The whole procedure will continue until meeting a convergence condition.

Consider the case when each local loss function f_n is both ℓ_n-strongly convex and L_n-smooth. We define the condition number of the problem $\kappa = \frac{L^*}{\ell_*}$ where $\ell_* = \min_{n\in N} \ell_n$ is the smallest strong convexity constant and $L^* = \max_{n\in N} L_n$ is the largest Lipschitz constant. Then we have the following results, which has been proved in [18, Section 5].

Theorem 1. *Assuming that $\{\boldsymbol{\theta}_n^*\}$ are fixed points for the **FedSplit** algorithm. Then for any initialization $\boldsymbol{\theta}^1 \in \mathbb{R}^d$ and step size $s = 1/\sqrt{\ell_* L^*}$,*

1) *the algorithm has an optimal solution $\boldsymbol{\theta}^* = \frac{1}{N}\sum_{n=1}^N \boldsymbol{\theta}_n^*$ to the problem (1);*
2) *the iterates (5) satisfy*
$$\left\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*\right\| \leq \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^t \sqrt{\delta_0},$$
where $\delta_0 = \frac{1}{N}\sum_{n=1}^N \left\|\boldsymbol{\theta}_n^0 - \boldsymbol{\theta}_n^*\right\|^2$;
3) *the iteration complexity is*
$$T(\epsilon, \kappa) = O\left(\sqrt{\kappa}\log(1/\epsilon)\right),$$
to achieve an ϵ-accurate solution, i.e., $\|\boldsymbol{\theta}^T - \boldsymbol{\theta}^*\| \leq \epsilon$.

Remark 1: To address problem (1), a number of different methods have been proposed, e.g., FedAvg [20] and FedProx [21]. However, these methods guarantee convergence to fixed points, but not necessarily optimal solutions, even in strongly convex settings [18], [22]. Furthermore, the iteration complexities of these algorithms will increase significantly when the problem becomes ill-conditioning. In constrast, the FedSplit algorithm enjoys global optimality and linear convergence rate when the local loss functions are strongly convex. In addition, the FedSplit algorithm is robust to the condition number of the problem. Nevertheless, the authors in [18] consider both uplink and downlink channels are error-free, which is indeed unpractical due to limited communication resources in wireless federated learning. Hence, in this paper, we study a more

practical implementation of the FedSplit algorithm over a noisy fading MAC.

C. Communication Protocol

In this paper, all edge devices communicate with the edge server over the shared wireless MAC channel via AirComp. In this case, coding is unnecessary for an AirComp system to achieve the optimal tradeoff between computation rate and accuracy [23]. Hence, this paper adopts an uncoded nonorthogonal multiple access (NOMA) protocol. Under this setup, we assume a block flat-fading channel, where the channel coefficient remains same during one communication block. Each block is assumed to contain d time slots, so that a d-dimensional local model is allowed to transmit within one block. Due to limited memory and computational capacity, the models on edge devices are usually tiny, and may consist of only thousands of parameters [24]. Since typical coherence blocks also have the same order of magnitude [25], it is possible to transmit a model vector in one transmission block. For large model dimensions, we can transmit the models during multiple consecutive coherence blocks, which will slightly affect the analysis. Hence, this paper mainly focuses on the former case.

In this paper, the number of blocks is assumed to be equal to the number of iterations, so that all devices upload their local models in the t-th iteration corresponding to the t-th block. Then, the received signal at the edge server is given by

$$\boldsymbol{y}^t = \sum_{n=1}^{N} h_n^t \boldsymbol{x}_n^t + \boldsymbol{w}^t, \tag{6}$$

where $h_n^t \in \mathbb{C}$ is the channel coefficient for device n in the t-th block; $\boldsymbol{w}_n^t \in \mathbb{C}^d$ denotes the additive noise i.i.d according to $\mathcal{CN}(0, \sigma_w^2 \mathbf{I})$; and transmitted signal $\boldsymbol{x}_n^t$ encodes the information about the local model $\boldsymbol{\theta}_n^t$. In addition, the transmit power constraint of each device is given by

$$\mathbb{E}\left[\|\boldsymbol{x}_n^t\|_2^2\right] \leq P_0, \quad \forall n \in \mathcal{N}, \tag{7}$$

with P_0 as the maximum transmit power.

Based on the received signal $\boldsymbol{y}^t$, the edge server needs to recover the average global model (5) via AirComp. However, due to the distortion and noise caused by wireless channels, the edge server can only use the perturbed information about local models received from the devices to update the global model. In addition, these factors will greatly affect the convergence of the FedSplit algorithm. Hence, this paper aims to develop a reliable transceiver strategy based on the FedSplit algorithm in wireless communication systems. The strategy includes precoding local models at edge devices and recovering the average global model at edge server. In the following, we will propose the **AirComp based FedSplit Algorithm**, and then provide convergence analysis in section III.

D. AirComp Based FedSplit Algorithm

In this paper, we assume that perfect CSI are available on all devices and the edge sever, which can be achieved by pilot based methods. For implementation of AirComp, each device is required to perform magnitude alignment to reduce the received signal to the desired average global model (5). By exploiting the knowledge of CSI, each device is able to implement channel inversion by multiplying the local model by its inverse channel coefficient. Specifically, in the t-th iteration, device n encodes its local model $\boldsymbol{\theta}_n^t$ into the transmitted signal $\boldsymbol{x}_n^t$ via

$$\boldsymbol{x}_n^t \triangleq \sqrt{\alpha^t} \frac{(h_n^t)^{\mathrm{H}}}{|h_n^t|^2} \boldsymbol{\theta}_n^t, \tag{8}$$

where $\sqrt{\alpha^t}$ is a uniform scaling factor in the t-th iteration. The uniform scaling factor $\sqrt{\alpha^t}$ satisfies the power constraint (7), likely,

$$\|\boldsymbol{x}_n^t\|_2^2 = \left\| \sqrt{\alpha^t} \frac{(h_n^t)^{\mathrm{H}}}{|h_n^t|^2} \boldsymbol{\theta}_n^t \right\|_2^2 \leq P_0, \forall n, \tag{9}$$

which implies $\sqrt{\alpha^t} \triangleq \min_{n \in \mathcal{N}} \frac{|h_n^t| \sqrt{P_0}}{\|\boldsymbol{\theta}_n^t\|_2}$. However, it is obvious to note that weak channels (i.e., $|h_t^n| \approx 0$) results in the small scaling factor $\sqrt{\alpha^t}$. Consequently, the received signals will be weakened and the interference caused by channel noise will significantly increase. This suggests that uniform channel inversion may not be always desirable and the optimal power-control policy for AirComp should be adapted to multiuser CSI. Therefore, we propose a binary scheme of device selection based on multiuser CSI. In particular, a threshold γ is set for device selection, and edge devices observing fading coefficients of a smaller magnitude than γ do not transmit in the corresponding communication round. Under this scheme, the transmitted signals (8) become

$$\boldsymbol{x}_n^t = \sqrt{\alpha^t} \beta_n^t \frac{(h_n^t)^{\mathrm{H}}}{|h_n^t|^2} \boldsymbol{\theta}_n^t, \tag{10}$$

where the indicator of device selection β_n^t is defined by

$$\beta_n^t = \begin{cases} 0, & |h_n^t| < \gamma; \\ 1, & |h_n^t| \geq \gamma, \end{cases} \tag{11}$$

with some threshold $\gamma \geq 0$. Hence, the scaling factor becomes $\sqrt{\alpha^t} = \min_{n \in \mathcal{B}} \frac{|h_n^t| \sqrt{P_0}}{\|\boldsymbol{\theta}_n^t\|_2}$.

To simplify, we denote $\mathcal{B}^t \subseteq \mathcal{N}$ as the set of participating devices indices subject to $\beta_n^t = 1$. Since the edge server is assumed to know all CSI, it also knows $\mathcal{B}^t$ by (11). Hence, the average global model (5) can be recovered by the edge server as follows,

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{t+1} &\triangleq \frac{1}{\sqrt{\alpha^t}|\mathcal{B}^t|} \boldsymbol{y}^t \\ &= \frac{1}{\sqrt{\alpha^t}|\mathcal{B}^t|} \left(\sum_{n=1}^{N} h_n^t \sqrt{\alpha^t} \beta_n^t \boldsymbol{\theta}_n^t + \boldsymbol{w}^t \right) \\ &= \frac{1}{|\mathcal{B}^t|} \sum_{n \in \mathcal{B}^t} \boldsymbol{\theta}_n^t + \frac{\boldsymbol{w}^t}{\sqrt{\alpha^t}|\mathcal{B}^t|} \\ &= \frac{1}{|\mathcal{B}^t|} \sum_{n \in \mathcal{B}^t} \boldsymbol{\theta}_n^t + \tilde{\boldsymbol{w}}^t \\ &= \bar{\boldsymbol{\theta}}^t + \tilde{\boldsymbol{w}}^t, \end{aligned} \tag{12}$$

where $\bar{\boldsymbol{\theta}}^t \triangleq \frac{1}{|\mathcal{B}^t|} \sum_{n \in \mathcal{B}^t} \boldsymbol{\theta}_n^t$ and $\tilde{\boldsymbol{w}}^t$ is the equivalent additive noise according to $\mathcal{CN}(0, \frac{\sigma_w^2}{\alpha^t |\mathcal{B}^t|^2} \mathbf{I})$. The resulting algorithm with T communication rounds is summarized in Algorithm 1.

III. Convergence Analysis

In this section, we will provide the convergence analysis of the AirComp based FedSplit algorithm and prove that it can converge to the global optimal solution under strongly convex and smooth local loss functions.

The main strategy of our proof is to introduce two sequences $\{\boldsymbol{\theta}^t\}$ and $\{\hat{\boldsymbol{\theta}}^t\}$ generated by (5) and (12), respectively. While the sequence $\{\hat{\boldsymbol{\theta}}^t\}$ is perturbed by the channel gain and noise, we still can establish a single step recursive bound for the error $\mathbb{E}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\|^2$. Then by exploiting the results of Theorem 1, we further characterize the convergence of $\{\boldsymbol{\theta}^t\}$. Before presenting our main results, we first have the following assumptions that our analysis is based on.

Algorithm 1: AirComp based FedSplit Algorithm

Input : Initial $\boldsymbol{\theta}^0$, threshold γ, step size s, max number of rounds T

Initialization for each device: $\boldsymbol{\theta}_n^0 = \boldsymbol{\theta}^0, n \in \mathcal{N}$ with the initial $\boldsymbol{\theta}^0$;

for $t = 0, 1, \ldots, T$ **do**

  All devices receive the current estimate $\hat{\boldsymbol{\theta}}^t$;

  for each $n \in \mathcal{N}$ **do in parallel**

    Updating $\boldsymbol{\theta}_n^{t+1}$ via (2) and (3) ;

    Checking the channel state h_n^t and determining β_n^t via (11) ;

    if $\beta_n^t = 1$ **then**

      Transmitting $\boldsymbol{x}_n^t$ encoded via (10) over the MAC (6);

    end

  end

  The edge server receives $\boldsymbol{y}^t$, recovers $\boldsymbol{\theta}^{t+1}$ via (12), and then broadcasts $\hat{\boldsymbol{\theta}}^{t+1}$ back to all the devices via an error-free channel ;

end

A. Preliminaries

Assumption 1. *The local loss function f_n is both ℓ_n-strongly convex and L_n-smooth for any $n \in \mathcal{N}$.*

Assumption 2. *The local model is bounded by a universal constant $G > 0$, likely, $\|\boldsymbol{\theta}_n^t\|^2 \leq G^2, \forall t, n$.*

Assumption 3. *At each communication iteration, the set of participating devices B_t satisfies $|B_t| = B \leq N$ and is uniformly distributed over all the subsets of $\mathcal{N}$.*

Assumptions 1 and 2 are commonly used in analyzing FL algorithm for many learning-based tasks, i.e., linear regression and logistic regression. Assumption 3 can be implemented by the following distributed mechanism. At each communication round t, we choose the top B devices among the participating device set $\mathcal{B}^t$ in term of their CSI, i.e., $|h_n^t|$, to transmit their signals. If the event of $|\mathcal{B}_t| < B$ happens, the devices need wait to the next communication round. Actually, the probability of the event is very small especially when N is large and γ is small. This mechanism guarantees $|\mathcal{B}_t| = B$ at each iteration. Notice that Assumption 3 is only used for convergence analysis. In fact, when AirComp based FedSplit is implemented without such a mechanism, i.e., $|\mathcal{B}_t|$ is random, it will achieve similar convergence characteristics. The authors in [26] have made similar assumptions.

B. Main Results

Based on the above assumptions, we present the convergence of the sequence $\{\hat{\boldsymbol{\theta}}^t\}$ as follows.

Theorem 2. *Consider the system model specified in Section II. Let $\boldsymbol{\theta}^*$ denote the solution of the optimization problem* (1)*. When Assumptions (1-3) holds and setting the step size $s = \frac{1}{\sqrt{\ell_* L^*}}$, then it holds that*

$$\mathbb{E}[F(\hat{\boldsymbol{\theta}}^t)] - F(\boldsymbol{\theta}^*) \leq \frac{\delta_0}{2L}\rho^t + \frac{G^2}{2B^2L}\left(B + \frac{d\sigma_w^2}{\gamma^2 P_0}\right), \quad (13)$$

where $\delta_0 = \frac{1}{N}\sum_{n=1}^N \left\|\boldsymbol{\theta}_n^0 - \boldsymbol{\theta}_n^*\right\|^2$, $\rho = \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^2$ and $L = \sum_{n=1}^N L_n$.

Proof. See Section A. □

Remark 2: Theorem 2 shows that it is able to achieve the convergence rate of the FedSplit algorithm, i.e., linear convergence. Note that the iteration complexity remains $T(\epsilon, \kappa) = O\left(\sqrt{\kappa}\log(1/\epsilon)\right)$ as claimed in Theorem 1, whereas the GD based algorithms for the wireless FL problem proposed in [14], [15] are linearly dependent of the condition number, i.e., $T(\epsilon, \kappa) = O\left(\kappa \log(1/\epsilon)\right)$. Hence, we can conclude that our algorithm is more robust to the ill-conditioned problems. What's more, Theorem 2 establishes a finite bound of the estimation error for strongly convex and smooth local loss functions over fading MAC. The error bound is characterized by two terms, the *initial distance* due to the error in the initial estimate and the *additive noise* caused by the channel noise. In the case of error-free channels, our algorithm can be reduced to the FedSplit algorithm, and thus achieves the same performance. In addition, the design of the threshold γ will greatly affect the additive noise term, which will be discussed in our future work.

IV. Numerical Experiments

In this section we numerically evaluate the performance of AirComp based FedSplit Algorithm by presenting a typical example, i.e., linear regression. Local dataset of each device $\mathcal{D}_n$ is randomly generated by linear model

$$Y_n = X_n\boldsymbol{\theta}_0 + v_n, \forall n \in \mathcal{N}, \quad (14)$$

where $Y_n \in \mathbb{R}^{m_n}$ is a output vector with m_n elements related to the design matrix X_n, $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ is generated by sampling from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I}_d)$ and the noise vectors are independently generated according to $\mathcal{N}(0, \sigma^2\mathbf{I}_{m_n})$. The details of generating X_n will be discussed in Sections IV-A and IV-B. We use a linear least square loss function for each device n, given by

$$f_n(\boldsymbol{\theta}) = \frac{1}{2}\left\|Y_n - X_n\boldsymbol{\theta}\right\|^2, \quad (15)$$

which is strongly convex and differentiable.

We evaluate the algorithms in term of the expected loss of the objective values by running an algorithm, i.e., $\mathbb{E}[F(\hat{\boldsymbol{\theta}}^t)] - F(\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is the solution to federated optimization problem (1). In addition, we compare the AirComp based FedSplit algorithm with the Gradient Based Multiple Access (GBMA) algorithm proposed in [15], which is also developed to solve the federated learning problem over MAC. All the experiments will be performed p times, and we take the average of the results. In the following, we consider two different settings for problem conditioning.

A. Well-conditioned Setting

In this case, we generate random matrices $X_n \in \mathbb{R}^{m_n \times d}$ with $(X_n)_{uv} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$, for all $n \in \mathcal{N}$, $u \in [m_n]$ and $v \in [d]$. The simulation parameters are set as

$$\begin{aligned} &p = 20, \quad N = 100, \quad m_n = 200, \\ &d = 6, \quad \sigma^2 = 0.25, \quad \sigma_w^2 = 1, \quad \gamma = 0.5, \end{aligned} \quad (16)$$

thus satisfying that all X_n are full rank due to $m_n \gg d$ for all device n, which is called as well-conditioned setting.

Except GBMA, we also compare AirComp based FedSplit Algorithm with following algorithms: (i) original FedSplit algorithm; (ii) FedSGD algorithm, $e = 1$, which is the original version of GBMA without channel distortion, where e is the number of local gradient steps. After defining these parameters, we run simulations according to (14) and (15) with

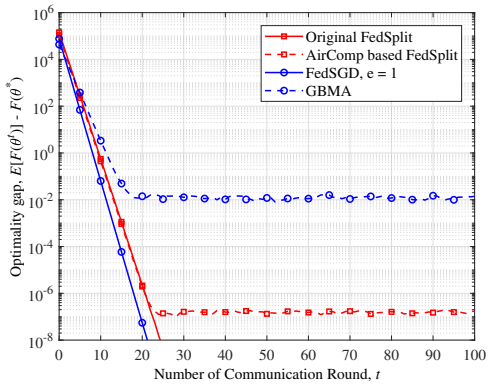

Fig. 2. Simulation results for linear regression under well-conditioned setting, plotting log optimality gap versus number of communication round t.

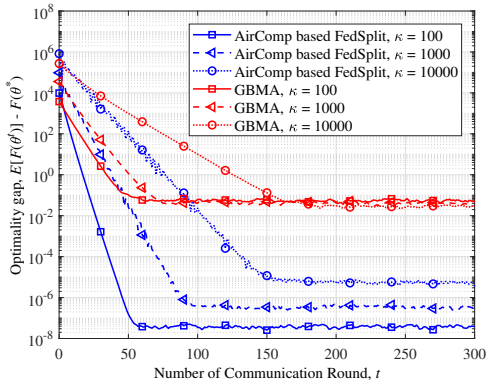

Fig. 3. Simulation results for linear regression under ill-conditioned setting, plotting log optimality gap versus number of communication round t.

Rayleigh channel gain and additive noise defined before. The simulation results are shown in Fig. 2.

As illustrated in Fig. 2, although both GBMA and AirComp based FedSplit achieve a linear convergence rate, GBMA has a larger error gap, i.e., 10^{-2} while AirComp based FedSplit can converge to a more accurate solution, i.e., 10^{-4}. Besides, AirComp based FedSplit and original FedSplit have the same convergence rate at the beginning. However, due to fading channels and additive noise, there is still a boundary between the solution obtained by the former and the latter, which corresponds to our theoretical analysis.

B. Ill-conditioned Setting

To verify the effectiveness of our algorithm under ill-conditioned setting, we consider the linear regression problem with different ill-conditioned setting in term of the condition number κ. The detail of generating the design matrix X_n under different condition numbers can be found in [18, Section 4.3]. Similarly, we use i.i.d Rayleigh channel gains and additive Gaussian noise for each device to simulate real situation. The other parameters are set as follows,

$$\begin{aligned} &p = 20, \quad N = 100, \quad m_n = 200, \\ &d = 6, \quad \sigma^2 = 1, \quad \sigma_w^2 = 1, \quad \gamma = 0.5. \end{aligned} \tag{17}$$

By running experiments for $\kappa \in \mathcal{K} = \{10^2, 10^3, 10^4\}$, we are able to plot the log gaps of different κ from the two algorithms versus communication round t. Fig. 3 shows that all the algorithms can achieve linear convergence rate. Besides, when the condition number κ increases, the convergence will slow down. In particular, the convergence rate of AirComp based FedSplit is less sensitive than GBMA in ill-condition cases, which means AirComp based FedSplit is more robust to the condition number than GBMA. In other words, for ill-conditioned problems and wireless environment, AirComp based FedSplit is faster and able to achieve higher accuracy compared with GBMA.

V. Conclusion

In this paper, we studied a wireless FL problem over a noisy fading MAC. To tackle the performance degradation in ill-conditioned settings, we proposed the AirComp based FedSplit algorithm, where the edge server recovered the noisy aggregation of local models transmitted by the end devices via AirComp. We provided the convergence analysis for the proposed algorithm that linearly converges to the optimal solutions for strongly convex and smooth loss functions. The robustness of the proposed algorithm to ill-conditioned problems with fast convergence was verified by theoretical results and numerical experiments.

Appendix A
Proof of Theorem 1

According to Assumption 1, the objective function F is also L-smooth with the Lipschitz constant $L = \sum_{n=1}^{N} L_n$. By the smoothness of the objective function F, we have that

$$\mathbb{E}[F(\hat{\boldsymbol{\theta}}^t)] - F(\boldsymbol{\theta}^*) \leq \frac{L}{2}\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2\right]. \tag{18}$$

By introducing the auxiliary sequence $\{\boldsymbol{\theta}^t\}$, we can rearrange the error term $\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2\right]$ as follows

$$\begin{aligned} &\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2\right] = \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t + \boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2\right] \\ &= \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\|^2\right] + 2\mathbb{E}\left[(\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t)^\top(\boldsymbol{\theta}^t - \boldsymbol{\theta}^*)\right] \\ &+ \mathbb{E}\left[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2\right]. \end{aligned}$$

According to (12) and Assumption 3, it is easy to verify that $\hat{\boldsymbol{\theta}}^t$ is an unbiased estimator of $\boldsymbol{\theta}^t$ as follows,

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\theta}}^t] &= \mathbb{E}\left[\frac{1}{B}\sum_{n\in\mathcal{B}^t}\boldsymbol{\theta}_n^t + \tilde{\boldsymbol{w}}\right] = \mathbb{E}\left[\frac{1}{B}\sum_{n\in\mathcal{B}^t}\boldsymbol{\theta}_n^t\right] + \mathbb{E}[\tilde{\boldsymbol{w}}] \\ &\overset{\text{(i)}}{=} \frac{1}{B}\mathbb{E}\left[\sum_{n\in\mathcal{B}^t}\boldsymbol{\theta}_n^t\right] \overset{\text{(ii)}}{=} \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\theta}_n^t = \boldsymbol{\theta}^t, \end{aligned}$$

where (i) comes from that $\tilde{\boldsymbol{w}}^t$ is zero-mean and (ii) can be easily derived from [27, Lemma 4]. Hence,

$$\mathbb{E}\left[(\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t)^\top(\boldsymbol{\theta}^t - \boldsymbol{\theta}^*)\right] = 0,$$

which further implies

$$\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2\right] = \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\|^2\right] + \mathbb{E}\left[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2\right].$$

The error term $\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^*\|^2\right]$ can be divided into two terms $\mathcal{A}_1 = \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\|^2\right]$ and $\mathcal{A}_2 = \mathbb{E}\left[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2\right]$. In the following, we establish the upper bounds for these two terms respectively in oder to bound the error term.

Recall that $\hat{\boldsymbol{\theta}}^t = \bar{\boldsymbol{\theta}}^t + \tilde{\boldsymbol{w}}^t$ in (12), we obtain that

$$\mathcal{A}_1 = \mathbb{E}\left[\left\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\right\|^2\right] + 2\mathbb{E}\left[\left(\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\right)^\top\tilde{\boldsymbol{w}}^t\right] + \mathbb{E}\left[\|\tilde{\boldsymbol{w}}^t\|^2\right].$$

Since the equivalent noise $\tilde{\boldsymbol{w}}^t$ is independent of the models and zero-mean, the term $\mathbb{E}\left[\left(\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\right)^\top \tilde{\boldsymbol{w}}^t\right]$ is zero. In addition, the noise also satisfies

$$\mathbb{E}\left[\|\tilde{\boldsymbol{w}}^t\|^2\right] = \frac{d\sigma_w^2}{\alpha^t B^2} \leq \frac{d\sigma_w^2 G^2}{\gamma^2 B^2 P_0}, \tag{19}$$

where the last inequality comes from that the definition of α_t and Assumption 2.

To further bound the term $\mathbb{E}\left[\left\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\right\|^2\right]$, we refer to the proof of [27, Lemma 5] and Assumption 2. Then,

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{n\in\mathcal{B}^t}\boldsymbol{\theta}_n^t - \boldsymbol{\theta}^t\right\|^2\right] \\
&= \frac{1}{B^2}\mathbb{E}\left[\left\|\sum_{i=1}^{B}(\boldsymbol{\theta}_{n_i}^t - \boldsymbol{\theta}^t)\right\|^2\right] = \frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}\left[\left\|\boldsymbol{\theta}_{n_i}^t - \boldsymbol{\theta}^t\right\|^2\right] \\
&= \frac{1}{B}\sum_{n=1}^{N}\frac{1}{N}\left\|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}^t\right\|^2 \\
&= \frac{1}{B}\sum_{n=1}^{N}\frac{1}{N}\left(\|\boldsymbol{\theta}_n^t\|^2 - 2(\boldsymbol{\theta}_n^t)^\top\boldsymbol{\theta}^t + \|\boldsymbol{\theta}^t\|^2\right) \\
&= \frac{1}{BN}\sum_{n=1}^{N}\|\boldsymbol{\theta}_n^t\|^2 - \frac{1}{B}\|\boldsymbol{\theta}^t\|^2 \leq \frac{1}{BN}\sum_{n=1}^{N}\|\boldsymbol{\theta}_n^t\|^2 \leq \frac{1}{B}G^2.
\end{aligned} \tag{20}$$

Substitute (19) and (20) into $\mathcal{A}_1$, we arrive at

$$\mathcal{A}_1 \leq \frac{1}{B}G^2 + \frac{d\sigma_w^2 G^2}{\gamma^2 B^2 P_0} = \frac{G^2}{B^2}\left(B + \frac{d\sigma_w^2}{\gamma^2 P_0}\right). \tag{21}$$

As for the term $\mathcal{A}_2$, we exploit the result of [18, Theorem 3] as follows

$$\mathcal{A}_2 \leq \rho^t \frac{1}{N}\sum_{n=1}^{N}\left\|\boldsymbol{\theta}_n^0 - \boldsymbol{\theta}_n^*\right\|^2,$$

where $\rho = \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^2$. Combining the bounds of $\mathcal{A}_1$ and $\mathcal{A}_2$ with (18) yields the stated claim.

REFERENCES

[1] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[2] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, 2019.

[3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6g: Ai empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[4] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Commun. Surveys Tuts*, vol. 22, no. 4, pp. 2167–2191, 2020.

[5] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, p. 19, Jan. 2019.

[6] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y. C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 2020.

[7] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2020.

[8] D. Wen, M. Bennis, and K. Huang, "Joint parameter-and-bandwidth allocation for improving the efficiency of partitioned edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8272–8286, 2020.

[9] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, 2020.

[10] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.

[11] M. Chen, H. Vincent Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2020.

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

[13] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2020.

[14] M. Mohammadi 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.

[15] 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.

[16] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, "Federated machine learning for intelligent iot via reconfigurable intelligent surface," *IEEE Netw.*, vol. 34, no. 5, pp. 16–22, 2020.

[17] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Lataief, "Federated Learning via Intelligent Reflecting Surface," *arXiv e-prints*, p. arXiv:2011.05051, Nov. 2020.

[18] R. Pathak and M. J. Wainwright, "FedSplit: An algorithmic framework for fast federated optimization," in *Proc. 32th Neural Inf. Process. Sys. (NeurIPS)*, May 2020.

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

[20] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 54, Apr. 2017, p. 1273–1282.

[21] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," in *Proc. 3rd Conf. Mach. Learn. Syst. (MLSys)*, Mar. 2020.

[22] G. Malinovsky, D. Kovalev, E. Gasanov, L. Condat, and P. Richtárik, "From Local SGD to Local Fixed-Point Methods for Federated Learning," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, Apr. 2020.

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

[24] S. Ravi, "Efficient on-device models using neural projections," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 5370–5379.

[25] S. Wang, K. Guan, D. He, G. Li, X. Lin, B. Ai, and Z. Zhong, "Doppler shift and coherence time of 5g vehicular channels at 3.5 ghz," in *2018 IEEE International Symposium on Antennas and Propagation USNC/URSI National Radio Science Meeting*, 2018, pp. 2005–2006.

[26] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-Air Federated Learning from Heterogeneous Data," *arXiv e-prints*, p. arXiv:2009.12787, Sep. 2020.

[27] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *Proc. 8th Int. Conf. Learn. Repre. (ICLR)*, Apr 2020.