Federated Learning over Noisy Channels

Xizixiang Wei and Cong Shen
Department of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22904, USA
{xw8cw,cong}@virginia.edu

***Abstract*—Does Federated Learning (FL) work when *both* uplink and downlink communications have errors? How much communication noise can FL handle and what is its impact to the learning performance? This work is devoted to answering these practically important questions by explicitly incorporating both uplink and downlink noisy channels in the FL pipeline. We present a rigorous convergence analysis of FL over simultaneous uplink and downlink noisy communication channels, and characterize the sufficient conditions for FL to maintain the same convergence rate scaling as the ideal case of no communication error. The analysis reveals that, in order to maintain the $\mathcal{O}(1/T)$ convergence rate of FEDAVG with perfect communications, the uplink and downlink signal-to-noise-ratio (SNR) should be controlled such that they scale as $\mathcal{O}(t^2)$ where t is the index of communication rounds. This key result leads to a transmit power control policy for analog aggregation, whose performance is shown to be superior over the standard method via extensive numerical experiments using real-world FL tasks.**

***Index Terms*—Federated Learning; Convergence Analysis; Noisy Communications.**

I. INTRODUCTION

Federated Learning (FL) is an emerging distributed machine learning (ML) paradigm that has many attractive properties. In particular, FL caters to the growing trend that massive amount of the real-world data is generated exogenously at the edge devices. For better privacy protection, it is desirable to keep the data locally at the device and enable distributed model training, which motivated the development of FL [1]. The power of FL has been realized in commercial devices (e.g., Pixel 2 uses FL to train ML models to personalize user experience) and ML tasks (e.g., Gboard uses FL for keyboard prediction) [2].

It is well known that communication is one of the primary bottlenecks for FL [1], [3]. However, existing research has largely focused on either reducing the number of communication rounds [4], or decreasing the size of the payload for transmission [5]. This is because in most FL literature that deal with communication efficiency, it is often assumed that a perfect communication "tunnel" has been established (e.g., using existing Wi-Fi or cellular architecture [2], [3]), and the task of improving communication efficiency largely resides on the ML design that trades off computation and communication. There are also recent studies that focus on the communication system design, particularly for wireless FL [5]–[9], but the focus has been on resource allocation, device selection, or cellular system design.

While the early studies provide a glimpse of the potential of optimizing communication for learning, the important issue of *noisy communications* for both uplink (clients send local models to the parameter server) and downlink (server sends global model to clients) have not been well investigated. In particular, it is often taken for granted that standard signal processing and communication techniques can be directly applied to FL. We show in this paper that this can be highly suboptimal because they are mostly designed for independent and identically distributed (IID) sources over time, while the communicated model updates (both uplink and downlink) in FL represent a long-term process consisting of many progressive learning rounds that collectively determine the final learning outcome. Channel noise and bit/packet error rates cannot be directly translated to the ultimate model accuracy and convergence rate. It is thus of utmost importance to rethink the wireless system design that caters to the unique characteristics of FL.

The goal of this paper is to answer the following fundamental question: how much communication noise can FL handle, and what is its impact to the learning performance? Towards this end, we first describe a complete FL system where *both* model upload and download take place over noisy channels, which is novel as all prior works either study uplink or downlink noisy communications, but not both. We then present the first major contribution of this work – a novel convergence analysis of the standard FEDAVG scheme under non-IID datasets, partial clients participation, and noisy downlink and uplink channels. More importantly, the analysis reveals that, in order to maintain the same $\mathcal{O}(1/T)$ convergence rate of FEDAVG with perfect communications, the uplink and downlink signal-to-noise-ratio (SNR) should be controlled such that they scale as $\mathcal{O}(t^2)$ where t is the index of communication rounds. This key result leads to the second major contribution of this work – a transmit power control method for analog aggregation that achieves the same model accuracy and convergence rate scaling of FEDAVG without noisy communications. The power control policy satisfies the sufficient SNR conditions of the convergence analysis, and its effectiveness is fully corroborated in the numerical experiments using standard real-world datasets.

The remainder of this paper is organized as follows. The system model that captures the noisy channels of FL in both uplink and downlink is described in Section II. Theoretical analysis is presented in Section III, which leads to a transmit power control policy that is described in Section IV. Experimental results are given in Section V, followed by the conclusions in Section VI.

II. System Model

We first introduce the standard FL problem formulation, and then describe a complete FL pipeline where both model upload and download take place over noisy channels.

A. FL Problem Formulation

The general federated learning problem setting follows the standard model in the original paper [1]. In particular, we consider a FL system with one central parameter server (e.g., base station) and a set of at most N clients (e.g., mobile devices). Client n stores a local dataset $\mathcal{D}_n = \{\mathbf{z}_i\}_{i=1}^{D_n}$ with its size denoted by D_n. Datasets across devices are assumed to be non-IID and disjoint. The maximum data size when all devices participate in FL is $D = \sum_{n=1}^{N} D_n$. The loss function $f(\mathbf{w}, \mathbf{z})$ measures how well a ML model with parameter $\mathbf{w} \in \mathbb{R}^d$ fits a particular data sample $\mathbf{z}$. Without loss of generality, we assume that $\mathbf{w}$ has zero-mean and unit-variance elements[1], i.e., $\mathbb{E}||w_i||_2^2 = 1 \ \ \forall i = 1 \cdots d$. For the n-th device, its local loss function $F_n(\cdot)$ is defined by

$$F_n(\mathbf{w}) \triangleq \frac{1}{D_n} \sum_{\mathbf{z} \in \mathcal{D}_n} f(\mathbf{w}, \mathbf{z}).$$

The goal of wireless FL is for the base station to learn a *global* machine learning (ML) model based on the distributed *local* datasets at the N clients, by coordinating and aggregating the training processes at individual clients without accessing the raw data. Specifically, the global optimization objective over all N clients is given by

$$F(\mathbf{w}) \triangleq \sum_{n=1}^{N} \frac{D_n}{D} F_n(\mathbf{w}) = \frac{1}{D} \sum_{n=1}^{N} \sum_{\mathbf{z} \in \mathcal{D}_n} f(\mathbf{w}, \mathbf{z}). \tag{1}$$

The global loss function measures how well the model fits the entire corpus of data on average. The learning objective is to find the best model parameter $\mathbf{w}^*$ that minimizes the global loss function: $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$. Let F^* and F_k^* be the minimum value of F and F_k, respectively. Then, $\Gamma = F^* - \frac{1}{N} \sum_{k=1}^{N} F_k^*$ quantifies the degree of non-IID as in [11].

B. FL over Noisy Uplink and Downlink Channels

We consider a generic FL framework where *partial* client participation and *non-IID* local datasets, two critical features that separate FL from conventional distributed ML, are explicitly captured. However, both the upload and download transmissions take place over noisy communication channels. The overall system diagram is depicted in Fig. 1. In particular, the FL-over-noisy-channel pipeline works by iteratively executing the following steps at the t-th learning round, $\forall t = 1, \cdots, T$.

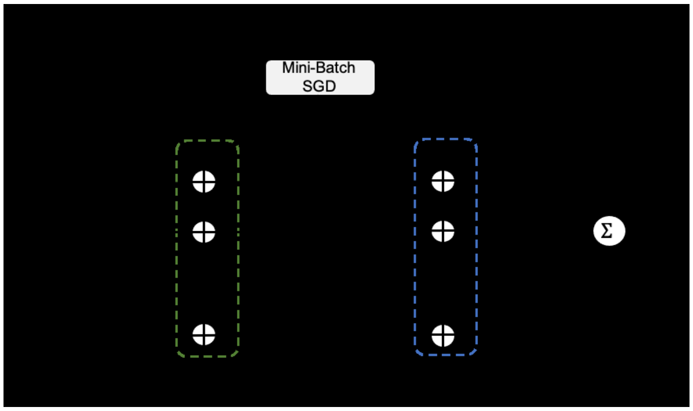

Fig. 1. End-to-end FL system diagram in the tth communication round. The impact of noisy channels in both uplink and downlink is captured.

1) **Downlink communication for global model download.** The centralized server broadcasts the current global ML model, which is described by the latest weight vector $\mathbf{w}_{t-1}$, to a set of randomly selected clients denoted as $\mathcal{S}_t$ with $|\mathcal{S}_t| = K$. Because of the imperfection introduced in communications, client k receives a noisy version of $\mathbf{w}_{t-1}$:

$$\hat{\mathbf{w}}_{t-1}^k = \mathbf{w}_{t-1} + \mathbf{e}_t^k, \tag{2}$$

where $\mathbf{e}_t^k = [e_{t,1}^k, \cdots, e_{t,i}^k, \cdots, e_{t,d}^k]^T \in \mathbb{R}^d$ is the d-dimensional downlink *effective noise* vector at client k and time t. We assume that $\mathbf{e}_t^k$ is a zero-mean random vector consisting of IID elements with bounded variance: $\mathbb{E}||e_{t,i}^k||_2^2 = \zeta_{t,k}^2, \forall t, k, i$; $\mathbb{E}||\mathbf{e}_t^k||_2^2 = d\zeta_{t,k}^2, \forall t, k$. Note that the effective noise term does not necessarily correspond to only the channel noise – it also captures post-processing errors in a communication transceiver, such as estimation, decoding and demodulation, frequency offset and phase noise, etc. We define the receive *local* (post-processing) SNR for the k-th client at the t-th communication round as

$$\mathsf{SNR}_{t,k}^{\mathsf{L}} = \frac{\mathbb{E}||\mathbf{w}_{t-1}||_2^2}{\mathbb{E}||\mathbf{e}_t^k||_2^2} = \frac{1}{\zeta_{t,k}^2}. \tag{3}$$

2) **Local computation.** Each client uses its local data to train a local ML model improved upon the received global ML model. In this work, we assume that *mini-batch stochastic gradient descent (SGD)* is used in training. Note that this is the most common training method in modern ML tasks, e.g., deep neural networks. Specifically, mini-batch SGD operates by updating the weight $\mathbf{w}_{t-1}^k$ iteratively (for E steps in each learning round) at device k as follows.

$$\begin{aligned} \text{Initialization:} \quad & \mathbf{w}_{t,0}^k = \hat{\mathbf{w}}_{t-1}^k, \\ \text{Iteration:} \quad & \mathbf{w}_{t,\tau}^k = \mathbf{w}_{t,\tau-1}^k - \eta_t \nabla F_k(\mathbf{w}_{t,\tau-1}^k, \xi_\tau^k), \\ & \forall \tau = 1, \cdots, E, \\ \text{Output:} \quad & \mathbf{w}_t^k = \mathbf{w}_{t,E}^k, \end{aligned}$$

where ξ_τ^k is a batch of data points that are sampled independently and uniformly at random from the local dataset of client k in the τ-th iteration of mini-batch SGD.

3) **Uplink communication for local model upload.** The K participating clients upload their latest local models to the server. More specifically, client k transmits a vector $\mathbf{x}_t^k$ to the server at the t-th round. We again consider the practical

[1] The parameter normalization and de-normalization method can be found in the appendix in [10].

case where the upload communication is erroneous, and the server receives a noisy version of the individual weight vectors from each client due to various imperfections in the uplink communications (e.g. channel noise, transmitter and receiver distortion, processing error). The received vector can be written as

$$\hat{\mathbf{x}}_t^k = \mathbf{x}_t^k + \mathbf{n}_t^k, \tag{4}$$

where $\mathbf{n}_t^k \in \mathbb{R}^d$ is the d-dimensional effective uplink noise vector for decoding client k's model at time t. We assume that $\mathbf{n}_t^k$ is a zero-mean random vector consisting of IID elements with bounded variance: $\mathbb{E}||n_{t,i}^k||_2^2 = \sigma_{t,k}^2, \forall t, k, i$; $\mathbb{E}||\mathbf{n}_t^k||_2^2 = d\sigma_{t,k}^2, \forall t, k$. For mathematical simplicity, we again assume that each element of the transmitted signal $\mathbf{x}_t^k$ has zero-mean and unit-variance elements, i.e., $\mathbb{E}||x_{t,i}^k||_2^2 = 1$, $\forall t, k, i$. The receive SNR at the *server* for decoding k-th client's signal $\mathbf{x}_t^k$ can be written as

$$\mathsf{SNR}_{t,k}^{\mathsf{S}} = \frac{\mathbb{E}||\mathbf{x}_t^k||_2^2}{\mathbb{E}||\mathbf{n}_t^k||_2^2} = \frac{1}{\sigma_{t,k}^2}. \tag{5}$$

Here we consider the direct model transmission scheme[2] in the uplink. The K participating clients upload the latest local models $\mathbf{w}_t^k$ themselves to the server, i.e., $\mathbf{x}_t^k = \mathbf{w}_t^k$.

4) **Global aggregation.** The server aggregates the received local models to generate a new global ML model following the standard FEDAVG [1]: $\mathbf{w}_t = \sum_{k \in \mathcal{S}_t} \frac{D_k}{\sum_{i \in \mathcal{S}_t} D_i} \hat{\mathbf{x}}_t^k$. The server then moves on to the $(t+1)$-th round. For ease of exposition and to simply the analysis, we assume in the remainder of the paper that the local dataset sizes at all devices are the same: $D_i = D_j$, $\forall i, j \in [N]$, and focus on the general case of randomly selected K out of N clients participating in the server aggregation with non-IID datasets. The aggregation can be simplified as

$$\mathbf{w}_t = \frac{1}{K} \sum_{k \in \mathcal{S}_t} \hat{\mathbf{x}}_t^k = \frac{1}{K} \sum_{k \in \mathcal{S}_t} \left(\mathbf{w}_t^k + \mathbf{n}_t^k \right). \tag{6}$$

the SNR for the *global* model (after aggregation) can be written as

$$\mathsf{SNR}_t^G = \frac{\mathbb{E}|| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k ||^2}{\mathbb{E}|| \sum_{k \in \mathcal{S}_t} \mathbf{n}_t^k ||^2} = \frac{\mathbb{E}|| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k ||^2}{d\sigma_t^2}, \tag{7}$$

where $\sigma_t^2 \triangleq \sum_{k \in \mathcal{S}_t} \sigma_{t,k}^2$.

III. Convergence Analysis of FL over Noisy Channels

In this section, we analyze the convergence of FEDAVG in the presence of both uplink and downlink communication noise when direct model transmission is adopted for local model upload. We first make the following standard assumptions that are widely used in the convergence analysis of FEDAVG; see [11], [13], [14].

Assumption 1. *1) L-smooth:* $\|\nabla F_k(\mathbf{w}) - \nabla F_k(\mathbf{v})\| \leq L \|\mathbf{w} - \mathbf{v}\|, \forall\ \mathbf{v}, \mathbf{w}$.

*2) μ-**strongly convex:*** $(\nabla F_k(\mathbf{w}) - \nabla F_k(\mathbf{v}))^T (\mathbf{w} - \mathbf{v}) \geq \mu \|\mathbf{w} - \mathbf{v}\|^2, \forall\ \mathbf{v}, \mathbf{w}$.

*3) **Bounded variance for mini-batch SGD:** The variance of stochastic gradients at any client $k = 1, \cdots, N$ satisfies:*

$$\mathbb{E} \|\nabla F_k(\mathbf{w}, \xi) - \nabla F_k(\mathbf{w})\|^2 \leq \delta_k^2,$$

for any mini-batch data ξ at client k.

*4) **Uniformly bounded gradient:** $\mathbb{E} \|\nabla F_k(\mathbf{w}, \xi)\|^2 \leq H^2$ $\forall k = 1, \cdots, N$ and any mini-batch data ξ at client k.*

For simplicity, we consider *homogeneous* noise power levels at the uplink and downlink, i.e., we assume

$$\sigma_{t,k}^2 = \bar{\sigma}_t^2, \quad \text{and} \quad \zeta_{t,k}^2 = \bar{\zeta}_t^2, \quad \forall t \in [T], k \in [N]. \tag{8}$$

Theorem 1. *Define $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, E\}$. Choose learning rate $\eta_t = \frac{2}{\mu(\gamma + t)}$ and adopt a SNR control policy that scales the effective uplink and downlink noise power over t such that:*

$$\bar{\sigma}_t^2 \leq \frac{4K}{\mu^2(\gamma + t - 1)^2} \sim \mathcal{O}\left(\frac{1}{t^2}\right) \tag{9}$$

$$\bar{\zeta}_t^2 \leq \frac{4N}{\mu^2(\gamma + t)(\gamma + t - 2)} \sim \mathcal{O}\left(\frac{1}{t^2}\right). \tag{10}$$

where $\bar{\sigma}_t^2$ and $\bar{\zeta}_t^2$ represent the individual client effective noise in the uplink and downlink, respectively, which are defined in Eqn. (8)*. Then, under Assumption 1, the convergence of* FEDAVG *with non-IID datasets and partial clients participation satisfies*

$$\mathbb{E} \|\mathbf{w}_T - \mathbf{w}^*\|^2 \leq \frac{8\kappa + E}{\gamma + T} \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{4D}{\mu^2(\gamma + T)}, \tag{11}$$

where $D = \sum_{k=1}^N \frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2 + 2d$.

Due to the space limitation, we omit the proof of Theorem 1. The complete proof can be found in [12]. Theorem 1 guarantees that even under *simultaneous* uplink and downlink noisy communications, the same $\mathcal{O}(1/T)$ convergence rate of FEDAVG with perfect communications can be achieved if we control the effective noise power of both uplink and downlink to scale at rate $\mathcal{O}(1/t^2)$ and choose the learning rate at $\mathcal{O}(1/t)$ over t. We note that the choice of η_t to scale as $\mathcal{O}(1/t)$ is well-known in distributed and federated learning [11], [13], [14], which essentially controls the "noise" that is inherent to the stochastic process in SGD to gradually shrink as the FL process converges. The fundamental idea that leads to Theorem 1 is to *control the "effective channel noise" to not dominate the "effective SGD noise"*, i.e., to always have the effective channel noise floor to be below that of the SGD noise. This idea is both critical and utilized in a non-trivial fashion in the proof of Theorem 1.

We further note that although the requirement of Theorem 1 is presented in terms of the effective noise power, what ultimately matters is the signal-to-noise ratio defined in

[2] We extend it to a more practical model differential transmission scheme, where the difference of latest model and the previous global model is transmitted, in the online version [12].

Section II-B. There exist signal processing and communication techniques that can satisfy the requirement by either increasing the signal power (e.g., transmit power control) or reducing the post-processing noise power (e.g., diversity combining). In the following section, we shown one such design example that controls the effective noise by controlling the transmit power.

IV. Transmit Power Control for Analog Aggregation in FL

An immediate engineering question following the theoretical analysis is how we can realize the SNR control policy in Theorem 1. One natural approach is *transmit power control*, which can alter the receive SNR while satisfying certain constraints. In this work, we propose a power control policy for the analog aggregation framework in [10] as an example to demonstrate the communication system design for FL tasks in presence of communication noise.

Consider a communication system where several narrowband orthogonal channels (e.g. sub-carriers in OFDM, time slots in TDMA, or eigenchannels in MIMO) are shared by K random selected clients in an uplink model update phase of a communication round in FedAvg. Each element in the transmitted model $\mathbf{w} \in \mathbb{R}^d$ is allocated and transmitted in a narrowband channel and aggregated automatically over the air [10]. Denote the received signal of each element $i = 1, \cdots, d$ in the t-th communication round as

$$y_{t,i} = \frac{1}{K} \sum_{k=1}^{K} r_{t,k}^{-\alpha/2} h_{t,k,i} p_{t,k,i} w_{t,k,i} + n_{t,i} \;\; \forall k \in \mathcal{S}_t, \quad (12)$$

where $r_{t,k}^{\alpha/2}$ and $h_{t,k,i} \in \mathcal{CN} \sim (0,1)$ are the large-scale and small-scale fading factors of the channel, respectively, $n_{t,i} \in \mathcal{CN} \sim (0,1)$ is the IID additive Gaussian white noise, and $p_{t,k,i}$ denotes the transmit power based on the power control policy. We assume perfect channel state information at the transmitters (CSIT). Due to the aggregation requirement of federated learning, we adopt the channel inversion rule:

$$p_{t,k,i} = \frac{\sqrt{\rho_t^{\text{UL}}}}{r_{t,k}^{-\alpha/2} h_{t,k,i}}, \quad (13)$$

where ρ_t is a scalar. Hence, the received SNR of the element $w_{t,i}$ can be written as

$$\mathsf{SNR}_t^G = \mathbb{E} \left\| \sum_{i=1}^{d} \frac{\frac{\sqrt{\rho_t^{\text{UL}}}}{K} \sum_{k \in \mathcal{S}_t} w_{t,k,i}}{n_{t,i}} \right\|^2 = \frac{\rho_t^{\text{UL}} \mathbb{E} \| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k \|^2}{dK^2}. \quad (14)$$

According the Theorem 1, we need to have

$$\rho_t^{\text{UL}} = \frac{K}{\sigma_t^2} \geq \frac{\mu^2 (\gamma + t - 1)^2}{4K} \sim \mathcal{O}(t^2) \quad (15)$$

in the uplink to ensure the convergence of FedAvg.

In the downlink case, when the server broadcasts the global model to K randomly selected clients, the receive signal of the i-th element for the k-th user in the t-th communication round is

$$y_{t,n,i} = r_{t,n}^{-\alpha/2} h_{t,n,i} \sqrt{\rho_t^{\text{DL}}} w_{t,i} + e_{t,n,i} \;\; \forall n = 1 \cdots K, \quad (16)$$

where $e_{t,k,i} \in \mathcal{CN} \sim (0,1)$ is the IID additive Gaussian white noise, ρ_t is the transmitted power at the server and $p_{t,k,i} = \frac{1}{r_{t,k}^{-\alpha/2} h_{t,k,i}}$ is the channel inversion factor applying at the clients. The downlink SNR for the k-th user is

$$\mathsf{SNR}_{t,n,i}^{L} = r_{t,n}^{-\alpha} \rho_t^{\text{DL}}. \quad (17)$$

Instead of keeping ρ_t^{DL} as a constant, we derive the following policy based on Theorem 1 to guarantee the convergence of FedAvg:

$$\rho_t^{\text{DL}} \geq \frac{r_{t,k}^{\alpha} \mu^2 (\gamma + t)(\gamma + t - 2)}{4N} \sim \mathcal{O}(t^2). \quad (18)$$

By applying this power control policy, federated learning tasks are able to achieve better performances under the same total energy budget. This will also be numerically validated in the experiment section.

Remark 1. *We note that transmit power control is not the only approach to have an increased effective SNR of model parameters in FL. Methods such as increasing quantization bit of parameters and applying diversity combining in wireless communication systems may also be adapted to implement this general SNR control policy.*

V. Simulation Results

A. Experiment Setup

We consider a communication system with narrowband uplink and downlink parallel channels for FL tasks. For simplicity, we assume that every channel has the same noise level. During each communication round of FL, each parameter of the ML model is transmitted in one of the narrowband channels. Suppose that the total communication rounds is T and both the uplink and downlink total power budget is $P = \sum_{t=1}^{T} P_t$, where P_t is the transmission power of the t-th round. We consider the following three schemes.

1) **Noise free (ideal).** This is the ideal case where there is no noise in either uplink or downlink channels, i.e., the accurate model parameters are perfectly received at both server and clients. This servers as the best-case performance.
2) **Equal power allocation.** In each communication round, the uplink and downlink transmission power is the same, i.e., $P_t = P/T, \;\; \forall t$. This represents the current state of the art in [10].
3) $\mathcal{O}(t^2)$**-increased power allocation.** Transmission power is increased at the rate of $\mathcal{O}(t^2)$ with the communication rounds, i.e., the received SNR of the model parameters is increased and the effective noise of the signal is decreased with the progress of FL. With the total budget P, it is easy to see that $P_t = 6Pt^2/(T(T+1)(2T+1)), \;\; \forall t$.

We use the standard image classification FL tasks to evaluate the performance of the above schemes. The following three widely utilized datasets are adopted.

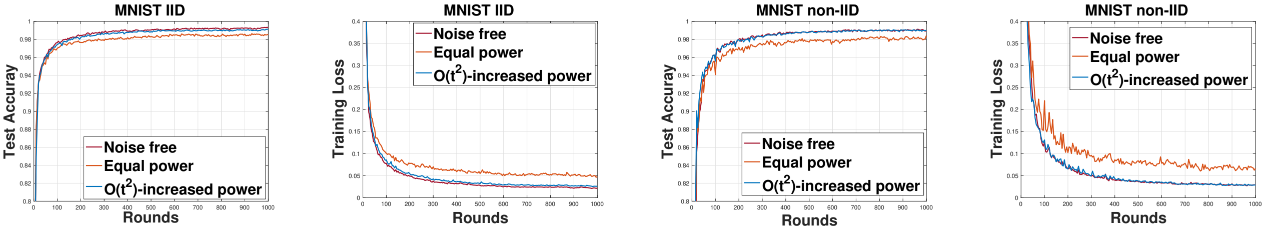

Fig. 2. Comparing test accuracy and training loss for noise-free (ideal), equal power allocation scheme and $\mathcal{O}(t^2)$-increased power allocation in IID and Non-IID data partitions on MNIST dataset.

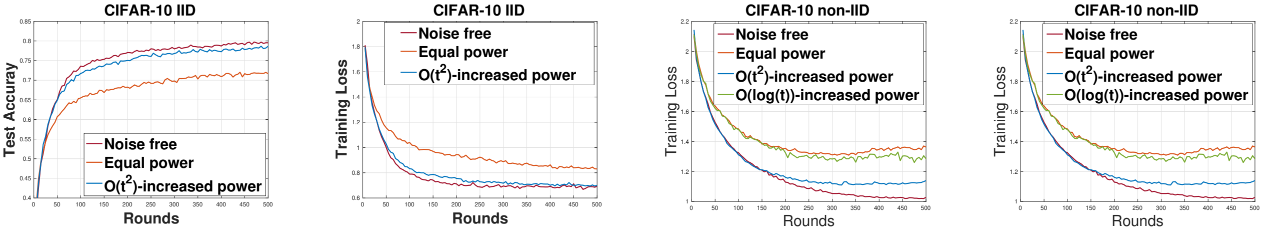

Fig. 3. Comparing test accuracy and training loss for noise-free (ideal), equal power allocation scheme and $\mathcal{O}(t^2)$-increased power allocation in IID and Non-IID data partitions on CIFAR-10 dataset.

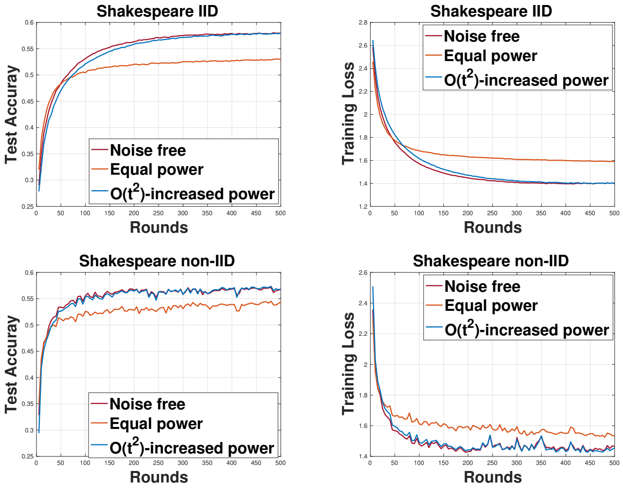

Fig. 4. Comparing test accuracy and training loss for noise-free (ideal), equal power allocation scheme and $\mathcal{O}(t^2)$-increased power allocation in IID and Non-IID data partitions on Shakespeare dataset.

1) **MNIST.** The training sets are evenly partitioned over $N = 2000$ clients each containing 30 examples and we set $K = 20$ per round (1% of total users). For the **IID** case, the data is shuffled and randomly assigned to each client while for the **non-IID** case, the data is sorted by labels, divided into 4000 partitions, and each client is then randomly assigned 2 partitions with 1 or 2 labels. The CNN model has two 5×5 convolution layers, a fully connected layer with 512 units and ReLU activation, and a final output layer with softmax. The first convolution layer has 32 channels while the second one has 64 channels, and both are followed by 2×2 max pooling. The following parameters are used for training: local batch size $BS = 5$, the number of local epochs $E = 1$, and learning rate $\eta = 0.065$.
2) **CIFAR-10.** We set $N = 100$ and $K = 10$ for i.i.d while $N = K = 10$ for non-IID We train a CNN model with two 5×5 convolution layers (both with 64 channels), two fully connected layers (384 and 192 units respectively) with ReLU activation and a final output layer with softmax. The two convolution layers are both followed by 2×2 max pooling and a local response norm layer. The training parameters are: (a) **IID**: $BS = 50$, $E = 5$, learning rate initially sets to $\eta = 0.15$ and decays every 10 rounds with rate 0.99; (b) **non-IID**: $BS = 100$, $E = 1$, $\eta = 0.1$ and decay every round with rate 0.992.
3) **Shakespeare.** This dataset is built from *The Complete Works of William Shakespeare* and each speaking role is viewed as a client. Hence, the dataset is naturally unbalanced and non-IID since the number of lines and speaking habits of each role vary significantly. There are totally 1129 roles in the dataset [15]. We randomly pick 300 of them and build a dataset with 794659 training examples and 198807 test examples. We also construct an IID dataset by shuffling the data and redistribute evenly to 300 roles and set $K = 10$. The ML task is the next-character prediction, and we use a classifier with an 8D embedding layer, two LSTM layers (each with 256 hidden units) and a softmax output layer with 86 nodes. The training parameters are: $BS = 20$, $E = 1$, learning rate initially sets to $\eta = 0.8$ and decays every 10 rounds with rate 0.99.

B. Experiment Results

The final model accuracies (after FL is complete) of the three schemes on MNIST, CIFAR-10 and Shakespeare datasets in both IID and non-IID configurations are summarized in Table I. As shown in the Fig. 2, we see that the proposed $\mathcal{O}(t^2)$-increased power allocation scheme achieves higher test accuracy and lower train loss than the equal power allocation scheme under the same energy budget on MNIST. In particular, $\mathcal{O}(t^2)$-increased power allocation scheme achieves 0.6%

TABLE I
PERFORMANCE SUMMARY OF THREE SCHEMES ON MNIST, CIFAR-10 AND SHAKESPEARE DATASETS UNDER IID AND NON-IID SETTINGS.

Dataset	Scheme	Accuracy	Percentage*	Accuracy	Percentage*
		IID		non-IID	
MNIST	Noise free	99.3%	100%	99.1%	100%
	Increased power	99.1%	99.8%	99.0%	99.9%
	Equal power	98.5%	99.2%	98.4%	99.3%
CIFAR-10	Noise free	79.5%	100%	63.3%	100%
	Increased power	78.9%	99.2%	59.6%	94.2%
	Equal power	71.7 %	90.2%	49.8%	78.7%
Shakespeare	Noise free	57.8%	100%	56.8%	100%
	Increased power	57.8%	100%	56.4%	99.3%
	Equal power	52.9 %	91.5%	54.4%	95.8%

*: The percentage columns represent the FL accuracy against the ideal case of noise-free accuracy.

higher test accuracy than that of equal power allocation scheme both in IID and non-IID data partitions. It might appear that the gain is not significant, but the reason is mostly due to that MNIST classification is a rather simple task. In fact, this gain is more notable under the more challenging CIFAR-10 and Shakespeare datasets as shown in Fig. 3 and Fig. 4. Comparing with the equal power allocation scheme, which achieves 90.2% and 78.7% of the ideal (noise free) test accuracy in IID and non-IID data partitions under CIFAR-10 dataset, the proposed method achieves 99.2% (IID) and 94.2% (non-IID) of the ideal (noise free) test accuracy respectively after $T = 500$ communication rounds, which is significant. Specifically, the training loss (test accuracy) of equal power allocation scheme increases (decreases) during the later 350th to 500th round in the non-IID case, implying that a non-increasing SNR may occur deterioration in the convergence of FL for difficult ML tasks. Similarly, under Shakespeare dataset, the equal power allocation scheme achieves 91.5% (IID) and 95.8% (non-IID) of the ideal (noise free) test accuracy, while the proposed method improves 8.5% and 3.5%, respectively. All of the three tasks have significant accuracy improvement due to the $\mathcal{O}(t^2)$ power control. We conclude that by allocating the transmission power at the rate of $\mathcal{O}(t^2)$ with the communication round t, we can achieve lower training loss and better test accuracy of the FL tasks than equal power allocation scheme under the same energy budget.

VI. CONCLUSION

We have investigated FL over noisy channels, where a FEDAVG pipeline with both uplink and downlink communication noise was studied. By theoretically analyzing its convergence under noisy communications in both directions, we have proved that the same $\mathcal{O}(1/T)$ convergence rate scaling of FEDAVG under perfect communications can be achieved if the uplink and downlink SNRs are controlled as $\mathcal{O}(t^2)$ over noisy channels. Inspired by this critical result, we proposed a transmit power control policy for the analog aggregation FL system. Extensive experimental results have corroborated the theoretical analysis and demonstrated the performance superiority of the fine-tuned power allocation scheme over baseline methods under the same total power budget.

ACKNOWLEDGEMENT

The work was partially supported by the US National Science Foundation (NSF) under Grant ECCS-2033671.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *AISTATS*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[2] K. Bonawitz *et al.*, “Towards federated learning at scale: System design,” in *The 2nd SysML Conference*, 2019, pp. 1–15.

[3] 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.

[4] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, “FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization,” in *AISTATS*, 2020.

[5] S. Zheng, C. Shen, and X. Chen, “Design and analysis of uplink and downlink communications for federated learning,” *IEEE J. Select. Areas Commun.*, vol. 39, no. 7, pp. 2150–2167, July 2021.

[6] J. Xu and H. Wang, “Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, 2021.

[7] T. Nishio and R. Yonetani, “Client selection for federated learning with heterogeneous resources in mobile edge,” in *IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–7.

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

[9] H. H. Yang, Z. Liu, T. Q. 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] 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.

[11] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of FedAvg on non-IID data,” in *International Conference on Learning Representations*, 2020.

[12] X. Wei and C. Shen, “Federated learning over noisy channels: Convergence analysis and design examples,” *CoRR*, vol. abs/2101.02198, 2021. [Online]. Available: https://arxiv.org/abs/2101.02198

[13] S. U. Stich, “Local SGD converges fast and communicates little,” in *ICLR*, 2018.

[14] C. Shen and S. Chen, “Federated learning with heterogeneous quantization,” in *ACM Symposium on Edge Computing – Workshop on Edge Computing and Communications (EdgeComm)*, November 2020.

[15] S. Caldas *et al.*, “LEAF: A benchmark for federated settings,” *arXiv preprint arXiv:1812.01097*, 2018.