Optimizing Federated Averaging over Fading Channels

Yujia Mu* Cong Shen* Yonina C. Eldar†
* Department of Electrical and Computer Engineering, University of Virginia, USA
† Department of Electrical Engineering, Weizmann Institute of Science, Israel

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834609

***Abstract*—Deep fading represents the typical error event when communicating over wireless channels. We show that deep fading is particularly detrimental for federated learning (FL) over wireless communications. In particular, the celebrated FEDAVG and several of its variants break down for FL tasks when deep fading exists in the communication phase. The main contribution of this paper is an optimal global model aggregation method at the parameter server, which allocates different weights to different clients based on not only their learning characteristics but also the instantaneous channel state information at the receiver (CSIR). This is accomplished by first deriving an upper bound on the parallel stochastic gradient descent (SGD) convergence over fading channels, and then solving an optimization problem for the server aggregation weights that minimizes this upper bound. The derived optimal aggregation solution is closed-form, and achieves the well-known $\mathcal{O}(1/t)$ convergence rate for strongly-convex loss functions under arbitrary fading and decaying learning rates. We validate our approach using several real-world FL tasks.**

I. INTRODUCTION

Despite recent progress on optimizing communications for distributed machine learning (ML), the understanding of how wireless impacts its convergence (especially for federated learning (FL)) is still limited. For example, it is well-known that *channel fading* is a unique characteristic that differentiates wireless communication from others (e.g., wired or deep space communication), and we now have a very good understanding of the random fading effect on wireless communications [1], [2]. However, such understanding only partially applies to wireless FL, because its objective is not limited to achieving low error probability with high data rate [3]. In particular, the goal of communication in FL is to have an accurate *final* ML model *with fast convergence*, which relies on the progressive communication rounds that *collectively* determine the learning performance. It is unclear how, in the presence of individual and highly dynamic fading channel conditions across different clients and across different learning rounds, the server should aggregate the heterogeneously erroneous local model estimates to have an accurate global model. In fact, as we report in the experiment results using real-world FL tasks, the *de facto* fading-unaware federated averaging (FEDAVG) [4], including several of its variants that exclusively utilize ML model training enhancements, break down when deep fading exists in the upload communication phases. This highlights the need to explicitly incorporate fading in the global model aggregation.

FL over noisy or fading channels has been studied in e.g., [5]–[8], which largely focus on unifying the local model upload and the global model aggregation with over-the-air (OTA) computation, which leverages the superposition nature of the wireless medium to directly "compute" the desired model aggregation. These designs however all require transmitter-side channel state information (CSIT), while this work only assumes receiver-side channel state information (CSIR). A direct consequence of no CSIT is that we cannot preclude clients whose channels are in deep fade to participate in FL, which is commonly used in OTA FL [6], [7]. Neither can we pre-cancel the fading effect at the client as in [7] which, combined with OTA computation, removes the fading effect and only needs to address the channel noise.

Another line of research related to this paper is the design of global model aggregation methods at the parameter server [9]. A related work [10] studies server aggregation for fading channels, where the aggregation is based only on channel statistics. Our work, on the other hand, focuses on model aggregation to account for the heterogeneous and random instantaneous fading channels, with a goal of directly minimizing the convergence rate upper bound.

In this work, we first analyze the impact of wireless channels on the convergence of distributed SGD. We then optimize FEDAVG to accommodate heterogeneous and deep-faded wireless channels in the model upload phases. Towards this end, we focus on an uplink fading communication model and a distributed SGD system, and consider the channel fading effect on uploading the model updates. To derive an optimal federated averaging method that incorporates channel fading, we analyze the convergence of *parallel SGD* over fading channels with an arbitrary feasible set of aggregation weights. We then minimize the convergence bound by adjusting the FEDAVG weight selection. The optimization problem is convex, with a simple closed-form solution. We show that with an additional (weaker) non-uniformly bounded variance assumption, the optimal weight assignment can be similarly obtained when minimizing the drift term, and the resulting upper bound highlights how fading affects the convergence under optimal server aggregation. We validate these claims by performing real-world FL tasks, including both the relatively easy MNIST handwritten digit classification and the more difficult CIFAR-10 image classification tasks. The experimental

The work was partially supported by the US National Science Foundation (NSF) under ECCS-2033671 and ECCS-2029978.

results indicate that fading-aware model aggregation not only significantly outperforms vanilla FEDAVG and its variants, but in fact approaches the performance with neither fading nor noise, i.e., FEDAVG under perfect communications [4], [11].

The remainder of this paper is organized as follows. Section II introduces the system model, including the problem formulation of distributed SGD, the standard FL pipeline, and communication over fading channels. The novel design of optimal server aggregation in the presence of fading channels is presented in Section III. Experimental results are reported in Section IV. Finally, Section V concludes the paper.

II. SYSTEM MODEL

A. Distributed SGD

We study the standard empirical risk minimization (ERM) problem in ML:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{m}\sum_{\mathbf{z}\in\mathcal{D}} l(\mathbf{x};\mathbf{z}), \tag{1}$$

where $\mathbf{x}\in\mathbb{R}^d$ is the machine learning model variable that one would like to optimize, $l(\mathbf{x};\mathbf{z})$ is the loss function evaluated at model $\mathbf{x}$ and data sample $\mathbf{z}=(\mathbf{z}_{\text{in}}, z_{\text{out}})$, which describes an input-output relationship of $\mathbf{z}_{\text{in}}$ and its label z_{out}, and $F:\mathbb{R}^d\to\mathbb{R}$ is a differentiable loss function averaged over the total dataset $\mathcal{D}$ with size m. We assume that there is a latent distribution ν that guides the generation of the global dataset $\mathcal{D}$, i.e., every data sample $z\in\mathcal{D}$ is drawn independently and identically distributed (IID) from ν. We denote $\mathbf{x}^* \triangleq \arg\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$, $f^* \triangleq F(\mathbf{x}^*)$.

Distributed and decentralized ML, including FL, aims at solving the ERM problem (1) by using a set of clients that run local computations in parallel, hence achieving a wall-clock speedup compared with the centralized training paradigm. We consider a distributed ML system with one central parameter server (e.g., at the base station) and a set of n clients (e.g., mobile devices). Mathematically, (1) can be equivalently formulated as

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i\in[n]} \frac{m_i}{m} F_i(\mathbf{x}), \tag{2}$$

where $F_i(\mathbf{x})$ is the local loss function at client i, defined as the average loss over its local dataset $\mathcal{D}_i$ with size m_i. We make the standard assumption that local datasets are disjoint, i.e., $\mathcal{D}=\cup_{i\in[n]}\mathcal{D}_i$ and $m=\sum_{i\in[n]} m_i$. This work focuses on the *full clients participation* setting, where all n clients participate in every round of distributed SGD. To ease the exposition and simplify the analysis, we also assume that all clients have the same local dataset size, i.e., $m_i = m_j, \forall i,j\in[n]$.

B. FEDAVG Pipeline

The considered distributed ML pipeline follows the standard framework of FEDAVG [4], with an explicit consideration of fading channels in the upload communication phase. In particular, the FEDAVG pipeline works by iteratively executing the following steps at the t-th learning round, for an $t\in[T]\triangleq\{1,2,\cdots,T\}$.

(1) Download the global model. The server broadcasts the current global model $\mathbf{x}_t$ to all clients. In a wireless FL setting, the base station typically has significantly more transmit power and communication resources than the clients, which are mostly battery-powered mobile devices with limited capabilities. As a result, it is often assumed that the download communication phase is error-free [6], [8], [12]. We follow the same assumption of *perfect* download communication, which results in all clients having the exact information of $\mathbf{x}_t$.

(2) Local model update at clients. Client i updates the received global model $\mathbf{x}_t$ based on its local dataset $\mathcal{D}_i$. In this work, we assume that SGD is used in the model training. Specifically, SGD at client i operates by updating the weight iteratively (for E steps in each learning round) as follows:

$$\begin{aligned}\text{Initialization: } & \mathbf{x}^i_{t,0} = \mathbf{x}_t,\\ \text{Iteration: } & \mathbf{x}^i_{t,\tau} = \mathbf{x}^i_{t,\tau-1} - \eta_t \nabla f_i(\mathbf{x}^i_{t,\tau-1}), \forall \tau = 1,\cdots,E,\\ \text{Output: } & \mathbf{x}^i_{t+1} = \mathbf{x}^i_{t,E},\end{aligned}$$

where we define

$$f_i(\mathbf{x}) \triangleq l(\mathbf{x};\xi_i), \qquad f(\mathbf{x}) \triangleq l(\mathbf{x};\xi). \tag{3}$$

Here ξ_i and ξ are data points sampled independently and uniformly at random (u.a.r.) from the local dataset of client i and global dataset, respectively.

(3) Upload local models. After the local model update, client i needs to transmit $\mathbf{s}_{i,t} \triangleq \mathbf{x}^i_{t+1} - \mathbf{x}_t$ to the parameter server so that it can recover $\mathbf{x}^i_{t+1}$. With the fading channel model and the transmitter/receiver processing described in Section II-C, the parameter server is able to recover

$$\mathbf{y}_{i,t} = \mathbf{s}_{i,t} + \mathbf{z}_{i,t} \tag{4}$$

for client i's model update, where $\mathbf{z}_{i,t}$ is the post-processing effective noise vector whose characteristics are given in Section II-C.

(4) Global aggregation. The server aggregates the received local models in (4) to generate a new global ML model $\mathbf{x}_{t+1}$. We follow the spirit of FEDAVG [4] but leave open the aggregation weights design:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \sum_{i\in[n]} p_{i,t}\mathbf{y}_{i,t}$$

with $\sum_{i\in[n]} p_{i,t} = 1$, for any $t\in[T]$.

C. Communication over Fading Channels

We now elaborate on the upload communication phase in the previous section, where each client i aims at sending vector $\mathbf{s}_{i,t}$ to the server over wireless fading channels. At the transmitter (client i), a total transmit power of P is enforced, and the transmitted symbol vector can be written as

$$\tilde{\mathbf{s}}_{i,t} = \sqrt{\frac{P}{\mathbb{E}\left\|\mathbf{s}_{i,t}\right\|^2}}\mathbf{s}_{i,t}. \tag{5}$$

It can be easily verified that the power constraint P is satisfied. Note that (5) requires the second-order statistics of $\mathbf{s}_{i,t}$. This can be obtained with a method similar to [7, Sec. III-A].

We consider that the transmission of $\tilde{\mathbf{s}}_{i,t}$ experiences a random fading channel fluctuation $h_{i,t}$ for each of its d elements. This holds when the underlying channel follows a *block fading* model [1], [2], where the channel remains constant for a duration of at least d symbol periods (i.e., the coherence time is larger than d), and then changes independently to another value following its distribution (e.g., Gaussian).

We then have the received signal

$$\tilde{\mathbf{y}}_{i,t} = h_{i,t}\tilde{\mathbf{s}}_{i,t} + \tilde{\mathbf{z}}_{i,t} = h_{i,t}\sqrt{\frac{P}{\mathbb{E}\left\|\mathbf{s}_{i,t}\right\|^2}}\mathbf{s}_{i,t} + \tilde{\mathbf{z}}_{i,t}, \quad (6)$$

where $\tilde{\mathbf{z}}_{i,t}$ denotes an additive white Gaussian noise (AWGN) vector with d independent Gaussian elements of mean zero and variance N_0. The receive SNR for client i is

$$\mathsf{SNR}_{i,t} = \frac{P g_{i,t}}{N_0} \quad (7)$$

with the channel power $g_{i,t}$ defined as $g_{i,t} \triangleq \left\|h_{i,t}\right\|^2$. We assume a *coherent receiver* with perfect channel state information at the receiver (CSIR). Hence, the receiver computes an unbiased estimate of $\mathbf{s}_{i,t}$ as

$$\mathbf{y}_{i,t} = \frac{1}{h_{i,t}}\sqrt{\frac{\mathbb{E}\left\|\mathbf{s}_{i,t}\right\|^2}{P}}\tilde{\mathbf{y}}_{i,t} = \mathbf{s}_{i,t} + \mathbf{z}_{i,t}, \quad (8)$$

where the post-processing noise vector $\mathbf{z}_{i,t}$ remains Gaussian distributed but with mean zero and variance $\mathbb{E}\left\|\mathbf{s}_{i,t}\right\|^2/\mathsf{SNR}_{i,t}$.

Because the channel coefficient $h_{i,t}$ is random, the receive SNR is also a random variable. Deep fading refers to the event when $g_{i,t} \ll 1$, which results in a very low instantaneous SNR. In addition, the random channel coefficients $\{h_{i,t}\}_{i\in[n]}$ across different clients at the same time t also naturally lead to heterogeneous receive SNRs for estimating $\{\mathbf{s}_{i,t}\}_{i\in[n]}$. Such SNR heterogeneity and deep fading effects are particularly detrimental to the performance of standard model aggregation, as will be demonstrated in Section IV.

III. Optimal Aggregation over Fading Channels

Our goal is to design the aggregation weights $\{p_{i,t} : i \in [n], t \in [T]\}$ at the server that incorporate the SNR heterogeneity and deep fading effect. To accomplish this goal, we first study the convergence behavior with an arbitrarily feasible choice of $\{p_{i,t}\}$. We then minimize the derived convergence upper bound by selecting the optimal weights. The decoupling of $\{p_{i,t} : i \in [n]\}$ at each round t naturally arises because the existing convergence analyses for distributed SGD, including this paper, are built on recursively reducing the gap to the optimal model. Optimizing such a recursive bound for each individual round t naturally reveals an optimal solution of $\{p_{i,t} : i \in [n]\}$ for every t.

Due to space limitation, we only report the optimal design for *parallel SGD*, which performs model aggregation after every SGD step, i.e., $E = 1$. The extension to *local SGD* [11], where $E > 1$, is left to the journal version. We focus on the L-smooth and μ-strongly convex loss function $l(\mathbf{x};\xi)$, as formally stated in Definitions 1 and 2 [13].

Definition 1 *$l(\mathbf{x};\xi)$ is L-smooth: $\|\nabla l(\mathbf{x},\xi) - \nabla l(\mathbf{y},\xi)\| \leq L\|\mathbf{x}-\mathbf{y}\|$ for $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ and $\xi \in \mathcal{D}$.*

Definition 2 *$l(\mathbf{x};\xi)$ is μ-strongly convex: $\langle \nabla l(\mathbf{x};\xi) - \nabla l(\mathbf{y};\xi), \mathbf{x}-\mathbf{y}\rangle \geq \mu\|\mathbf{x}-\mathbf{y}\|^2$ for $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ and $\xi \in \mathcal{D}$.*

Our first important result is presented in Proposition 1.

Proposition 1 *Consider L-smooth and μ-strongly convex loss functions f and f_i, with IID local datasets at all clients. We select the stepsize η_t to satisfy $\eta_t < 1/(2\mu), \forall t \in [T]$. If we assume unbiased SGD at all clients, i.e., $\mathbb{E}\left[\nabla f_i(\mathbf{x})\right] = \nabla F(\mathbf{x})$, for $i \in [n]$, the following inequality holds for parallel SGD:*

$$\mathbb{E}\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|^2 \leq (1-2\mu\eta_t)\,\mathbb{E}\left\|\mathbf{x}_t - \mathbf{x}^*\right\|^2 + \eta_t^2 \sum_{i\in[n]} p_{i,t}^2 \left(1 + \frac{1}{\mathsf{SNR}_{i,t}}\right)\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t)\right\|^2, \forall t \in [T] \quad (9)$$

where $\forall t \in [T]$, $\{p_{i,t} : i \in [n]\}$ in an arbitrary set of feasible aggregation weights in FedAvg *that satisfy*

$$\sum_{i\in[n]} p_{i,t} = 1 \quad \textit{and} \quad 0 \leq p_{i,t} \leq 1. \quad (10)$$

We note that Proposition 1 does not have any of the conventional assumptions of either uniformly bounded variance or uniformly bounded second moment, or both, for SGD [11], [14]. This is because Proposition 1 only characterizes the per-round gap reduction, and still has a second moment term that we choose not to bound (for now). The rationale, however, is that regardless of the assumptions we take to bound this term (and hence derive a final convergence rate upper bound), the optimization of the aggregation weight $\{p_{i,t} : i \in [n]\}$ at each round t is not affected for parallel SGD. In fact, Proposition 2 is established based on this generality.

Proof Sketch. We introduce an auxiliary error-free global model $\bar{\mathbf{x}}_{t+1} = \sum_{i\in[n]} p_{i,t}\mathbf{x}_{t+1}^i$. We have that

$$\begin{aligned}\mathbb{E}\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|^2 &= \mathbb{E}\left\|\bar{\mathbf{x}}_{t+1} - \mathbf{x}^*\right\|^2 + \mathbb{E}\left\|\mathbf{x}_{t+1} - \bar{\mathbf{x}}_{t+1}\right\|^2 \\ &\quad + 2\mathbb{E}\left\langle \sum_{i\in[n]} p_{i,t}\mathbf{z}_{i,t}, \bar{\mathbf{x}}_{t+1} - \mathbf{x}^* \right\rangle \\ &= \mathbb{E}\left\|\bar{\mathbf{x}}_{t+1} - \mathbf{x}^*\right\|^2 + \mathbb{E}\left\|\mathbf{x}_{t+1} - \bar{\mathbf{x}}_{t+1}\right\|^2, \end{aligned} \quad (11)$$

where the last equality holds since the noise vector is zero mean and the channel noise randomness is independent of the SGD randomness. We then separately bound the two terms in (11). The first part can be bounded in a standard way while the second term produces the desired SNR terms.

Proposition 2 *If $\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t)\right\|^2$ can be uniformly bounded independently of client $i \in [n]$, then the optimal aggregation weights $\{p_{i,t} : i \in [n]\}$ that minimize the per-round convergence upper bound in Proposition 1 are given as*

$$p_{i,t}^* = \frac{\mathsf{SNR}_{i,t}/\left(1+\mathsf{SNR}_{i,t}\right)}{\sum_{j\in[n]} \mathsf{SNR}_{j,t}/\left(1+\mathsf{SNR}_{j,t}\right)}. \quad (12)$$

We note that the condition of Proposition 2 only requires that $\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_t\right)\right\|^2$ has an upper bound that holds for any $i \in [n]$; it does not have to be instance-dependent or constant. In fact, the commonly adopted assumptions of uniformly bounded variance or uniformly bounded second moment both satisfy this requirement. On the other hand, removing this constraint does not fundamentally change the result in Proposition 2, and we will see a generalized result in Theorem 1.

Proof Sketch. Because the fading channel and noise only affect the drift term (the second term on the RHS of (9)), and the assumption that $\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_t\right)\right\|^2$ can be uniformly bounded independently of client $i \in [n]$ leads to the drift term being independent of i, minimizing the convergence upper bound in Proposition 1 is equivalent to solving the following optimization problem at round t:

$$\begin{aligned} \underset{\{p_{i,t}:i\in[n]\}}{\text{minimize}} \quad & \sum_{i\in[n]} p_{i,t}^2\left(1+\frac{1}{\mathsf{SNR}_{i,t}}\right) \\ \text{subject to} \quad & \sum_{i\in[n]} p_{i,t}=1;\ 0\le p_{i,t}\le 1, \forall i\in[n]. \end{aligned} \tag{13}$$

It is straightforward to verify that this is a convex optimization problem. Standard derivation applying the Karush–Kuhn–Tucker (KKT) condition [13] directly leads to (12).

Propositions 1 and 2 establish the per-round model convergence behavior and the corresponding optimal aggregation weight design. On the one hand, we are often interested in having a final convergence rate upper bound that characterizes the relationship between convergence and number of clients n, number of rounds T, and channel conditions $\{\mathsf{SNR}_{i,t}\}$. On the other hand, as mentioned in Proposition 2, we can have a more refined control of the convergence if we make client-dependent assumptions about the second-order behavior of clients SGD and improve the corresponding convergence derivation. The following theorem accomplishes these two goals.

Theorem 1 *Assume L-smooth and μ-strongly convex loss functions f_i with unbiased SGD on IID local datasets at clients. We further assume that the loss functions have client-dependent bounded variances: $\mathbb{E}\left\|\nabla f_i\left(\mathbf{x}_t\right)-\nabla F\left(\mathbf{x}_t\right)\right\|^2 \le \sigma_i^2, \forall i \in [n]$. Denote $R_{i,t} = \frac{\mathsf{SNR}_{i,t}}{\sigma_i^2(1+\mathsf{SNR}_{i,t})}, R_t = \sum_{i\in[n]} R_{i,t}, B_t = \min_{i\in[n]}\left\{\sigma_i^2\left(1+\mathsf{SNR}_{i,t}\right)\right\}$. The optimal aggregation weights $\{p_{i,t} : i \in [n]\}$ that minimize the drift term in the convergence rate upper bound are given as*

$$p_{i,t}^* = \frac{R_{i,t}}{R_t} = \frac{R_{i,t}}{\sum_{j\in[n]} R_{j,t}}. \tag{14}$$

Furthermore, if the stepsize is selected as $\eta_t = \frac{2}{\mu(t+\alpha)}$ where $\alpha = 2\max\left\{\frac{L^2}{\mu^2}\left(1+\frac{1}{B_t R_t}\right), 1\right\} - 1$, then the convergence of parallel SGD with server model aggregation weights given in (14) *satisfies*

$$\mathbb{E}\left[F\left(\mathbf{x}_t\right)\right] - f^* \le \frac{L}{2\left(t+\alpha\right)}\left(\frac{4}{\mu^2 R_t} + \left(1+\alpha\right)\left\|\mathbf{x}_1-\mathbf{x}^*\right\|^2\right) \tag{15}$$

for any $t \ge 1$.

We remark that Theorem 1 uses an assumption of client-dependent bounded variances, which is weaker than the uniformly bounded variances that is often required in the analysis. Additionally, this is different than the uniformly bounded second moment assumption in Proposition 2, which is why the optimal weights in (14) are slightly different than (12). In practice, σ_i may not be readily available to FL, and thus we use (12) in the experiments.

Proof Sketch. We start with the same step (11) as in the proof of Proposition 1. Now, since each client i has a different variance bound for the SGD, we can show that:

$$\mathbb{E}\left\|\sum_{i\in[n]} p_{i,t}\nabla f_i\left(\mathbf{x}_t\right)\right\|^2 \le \sum_{i\in[n]} p_{i,t}^2\sigma_i^2 + L^2\mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|^2. \tag{16}$$

Similar to the derivation of (16), we have

$$\begin{aligned} \mathbb{E}\left\|\mathbf{x}_{t+1}-\bar{\mathbf{x}}_{t+1}\right\|^2 \le \eta_t^2 \sum_{i\in[n]} \frac{p_{i,t}^2\sigma_i^2}{\mathsf{SNR}_{i,t}} \\ + L^2\eta_t^2 \sum_{i\in[n]} \frac{p_{i,t}^2}{\mathsf{SNR}_{i,t}}\mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|^2. \end{aligned} \tag{17}$$

Defining $\Delta_t = \mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|^2$ and combining (11) and (17), we have

$$\begin{aligned} \Delta_{t+1} \le \left(1-2\mu\eta_t + L^2\eta_t^2\left(1+\sum_{i\in[n]}\frac{p_{i,t}^2}{\mathsf{SNR}_{i,t}}\right)\right)\Delta_t \\ + \eta_t^2\sum_{i\in[n]} p_{i,t}^2\sigma_i^2\left(1+\frac{1}{\mathsf{SNR}_{i,t}}\right). \end{aligned} \tag{18}$$

It is clear that the drift term in the per-round convergence bound has a similar structure as the objective function of (13), and we can apply the same steps to derive the optimal weights and prove the theorem. The notable difference is the additional terms in the coefficient of Δ_t, which can be bounded using the properties of η_t.

IV. Experimental Results

A. Setup

To understand the impact of fading on distributed ML and evaluate the performance of the proposed aggregation method, we carry out a standard FL experiment of the CIFAR-10 classification task [15]. We report experimental results for both *IID* and *non-IID* datasets, as well as *full* and *partial* clients participation. Note that these results extend beyond the setting for the theoretical analysis. The experimental setup follows [4] and complete details remain the same as [16]. When fading channel is considered, we simulate a standard Rayleigh fading channel with average channel power of 1. The transmit power is normalized to be 1, thus leading to the default average signal-to-noise ratio (SNR) to be 0 dB.

We consider the following schemes in the experiments. (1) **Perfect_comm**: the ideal case with perfect communication. (2) **AWGN**: no fading but channel has AWGN, leading to equal

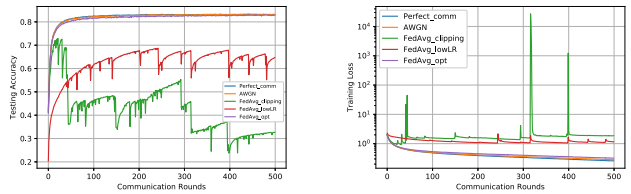

Fig. 1. IID local datasets and full clients participation.

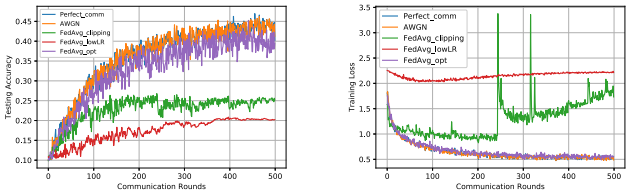

Fig. 2. Non-IID local datasets and partial clients participation.

aggregation weight. (3) **FedAvg_clipping**: vanilla FEDAVG with equal aggregation weight. However, we have observed in all experiments that directly using vanilla FEDAVG fails to converge for all tasks when fading is present. We are thus forced to implement ML enhancements. Since there are gradient explosions in FEDAVG, the popular *gradient clipping* method [17] is used. (4) **FedAvg_lowLR**: in order to improve the FedAvg_clipping performance, we combine the low learning rate method [18] with gradient clipping. (5) **FedAvg_opt**: the method we proposed in this paper.

All of the reported results are obtained by averaging over five independent runs based on different data distribution, client selection and random seed of the noise. We also report the final test accuracy, which is averaged over the last ten rounds, as the performance of the final global model.

B. Results

We can see from Fig. 1 that both variants of the original FEDAVG have poor performance (testing accuracy and training loss), and even fail to converge in some case. The issues mainly come from the "deep drops" that largely reset the training. These deep drops are caused by the random fading effect, not the channel noise – the supporting evidence is that the AWGN case does not have this performance degradation, and all schemes experience the same realization of channel noises (when applicable). We also remind the reader that this is already after averaging over five independent runs, which "masks" the severity of each valley (since it is unlikely different runs have such drops at the same round) but increases the number of rounds this may happen.

More importantly, we note that the proposed FedAvg_opt is very effective in handling the channel fading – without using any of the ML tricks it is able to achieve comparable performance as Perfect_comm. The same observations also hold for non-IID local datasets and partial clients participation, as shown in Fig. 2.

V. CONCLUSION

We have shown that the channel fading effect is detrimental to the performance of the popular *federated averaging* that is widely adopted in distributed machine learning. Deep fading, as opposed to channel noise, is the dominating effect in disrupting the convergence of distributed ML. We analyzed the convergence of *parallel SGD* under channel fading and arbitrary server aggregation. We then minimized the resulting convergence bound by adjusting the aggregation weights, which led to a closed-form solution. This solution further led to the main theoretical result that even under random channel fading across clients and communication rounds, federated averaging with optimized weights can still maintain the same $\mathcal{O}(1/t)$ convergence for strongly-convex loss functions – the same as when there is neither fading nor noise.

REFERENCES

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
[2] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.
[3] C. Shen, J. Xu, S. Zheng, and X. Chen, "Resource rationing for wireless federated learning: Concept, benefits, and challenges," *IEEE Commun. Mag.*, vol. 59, no. 5, pp. 82–87, May 2021.
[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.
[5] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.
[6] 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.
[7] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Trans. Signal Processing*, vol. 69, pp. 3796–3811, 2021.
[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] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proceedings of the 36th International Conference on Machine Learning*, June 2019, pp. 4615–4625.
[10] S. Lee, C. Park, S.-N. Hong, Y. C. Eldar, and N. Lee, "Bayesian federated learning over wireless networks," *arXiv preprint arXiv:2012.15486*, 2020.
[11] S. U. Stich, "Local SGD converges fast and communicates little," in *International Conference on Learning Representations*, 2019.
[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] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
[14] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," in *ICML Workshop on Coding Theory for Machine Learning*, 2019.
[15] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., April 2009.
[16] 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.
[17] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*. PMLR, 2013, pp. 1310–1318.
[18] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 437–478.