Gradient Statistics Aware Power Control for Over-the-Air Federated Learning in Fading Channels

Naifu Zhang and Meixia Tao
Dept. of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
Emails: {arthaslery, mxtao}@sjtu.edu.cn

***Abstract*—To enable communication-efficient federated learning, fast model aggregation can be designed using over-the-air computation (AirComp). In order to implement a reliable and high-performance AirComp over fading channels, power control at edge devices is crucial. Existing works assume that the signal to be aggregated from each device is identically distributed, and normalized with zero mean and unit variance. This assumption, however, does not hold for gradient aggregation in machine learning because gradient statistics are unknown for each device. In this paper, we study the optimal power control problem for efficient over-the-air FL by taking gradient statistics into account. Our goal is to minimize the model aggregation error measured by mean square error (MSE) by jointly optimizing the transmit power of each device and the denoising factor at the edge server. We first derive the optimal solution in closed form where the gradient first-order and second-order statistics are known. We then propose a method to estimate the gradient statistics based on the historical aggregated gradients and then dynamically adjust the transmit power on devices over each training iteration. Experiment results show that our proposed power control is better than existing full power transmission and threshold-based power control in both model accuracy and convergence rate.**

I. Introduction

Federated learning (FL) [1]–[3] is a new edge learning framework that enables many edge devices to collaboratively train a machine learning model without exchanging datasets under the coordination of an edge server in wireless networks. Compared with traditional learning at a centralized data center, FL offers several distinct advantages, such as preserving privacy, reducing network congestion, and leveraging distributed on-device computation. In FL, each edge device downloads a shared model from the edge server, computes an update to the current model by learning from its own local dataset, then sends this update to the edge server. Therein, the updates are averaged to improve the shared model.

The communication cost is the main bottleneck in FL since a large number of participating edge devices send their updates to the edge server at each round of the model training. Existing methods to obtain communication-efficient FL can be mainly divided into three categories, model parameter compression [4], [5], gradient sparsification [6], [7], and infrequent local update [1], [8]. Recently, a fast model aggregation approach is proposed for FL by applying over-the-air computation (AirComp) principle [9], such as in [10]–[12]. Such AirComp-based FL, referred to as *over-the-air FL* in this work, can dramatically save the uplink communication bandwidth compared with existing approaches.

Due to the channel fading, device selection and power control are crucial to achieve a reliable and high-performance over-the-air FL. In [13], the authors jointly optimize the transmit power at edge devices and the receive scaling factor (known as denoising factor) at the edge server. It is shown that the optimal power control in static channels exhibits a threshold-based structure. Namely, each device applies channel-inversion power control if its quality indicator exceeds the optimal threshold, and applies full power transmission otherwise. For AirComp-based gradient aggregation in FL, the work [10] introduces a truncation-based approach for excluding the edge devices with deep fading channels to strike a good balance between learning performance and aggregation error. The work [11] proposes a joint device selection and receiver beamforming design method to find the maximum selected devices with MSE requirements to improve the learning performance. These works [10], [11], [13] assume that the signal (i.e., the gradient) to be aggregated from each device is IID, and normalized with zero mean and unit variance. By exploiting the sparsity pattern in gradient vectors, the work [12] projects the gradient estimate in each device into a low-dimensional vector and transmits only the important gradient entries while accumulating the error from previous iterations. Therein, a channel-inversion like power control scheme, similar to those in [10], [11], [13] is designed so that the gradient vectors sent from selected devices are aligned at the edge server.

Note that all the exiting works on power control for over-the-air FL have overlooked the following statistical characteristics of gradients. The gradient distribution over each iteration is independent but not necessarily identically distributed, and even in the same iteration, the distribution of each element of the gradient vector can be non-identical. A general observation is that the gradient distribution changes over iterations and is different in each feature dimension. In addition, if the gradient statistics are unknown for each device, normalizing the gradient to a distribution with zero mean and unit variance is unrealistic. As such, due to the neglect of gradient statistics, the existing power control methods for over-the-air FL would

This work is supported by the NSF of China under grant 61941106 and the National Key R&D Project of China under grant 2019YFB1802702.

978-1-7281-7440-2/20/$31.00 ©2020 IEEE

perform poorly in practice.

Motivated by the above issue, in this paper, we study the optimal power control problem for over-the-air FL over fading channels by taking gradient statistics into account. Our goal is to minimize the model aggregation error measured by MSE, and hence improve the convergence rate of FL, through the joint optimization of the transmit power of each device and the denoising factor at the edge server. The main contributions of this work are outlined below:

- *Optimal power control with known gradient statistics:* We first derive the MSE expression of gradient aggregation at each training round when the first- and second-order statistics of the gradient vectors are known. We then formulate a joint optimization problem of transmit power and denoising factor for MSE minimization subject to individual peak power constraints at edge devices. Be decomposing this non-convex problem into subproblems defined on different subregions, we obtain the optimal power control strategy in closed form. It is found that there is an optimal index threshold, below which the devices transmit with full power and above which the devices transmit at the power such that the equivalent weight of their gradients for aggregation are equalized.
- *Adaptive power control with unknown gradient statistics:* We propose an adaptive power control algorithm that estimates the gradient statistics based on the historical aggregated gradients and then dynamically adjusts power values in each iteration. The communication cost consumed by estimating the gradient statistics are negligible compared to the transmission of the entire gradient vector.

Experiment results show that the over-the-air FL with the proposed adaptive power control obtains a much faster convergence rate than that with existing power control methods (full power transmission and threshold-based power control).

II. System Model

A. Federated Learning Over Wireless Networks

We consider an FL setting over a wireless network where a shared AI model (e.g., a classifier) is trained collaboratively across K edge devices via the coordination of an edge server as shown in Fig. 1. Let $\mathcal{K} = \{1, ..., K\}$ denote the set of edge devices. Each device $k \in \mathcal{K}$ collects a fraction of labelled training data via interaction with its own users, constituting a local dataset, denote as $\mathcal{D}_k$. The loss function measuring the model error is defined as

$$L(\boldsymbol{w}) = \sum_{k \in \mathcal{K}} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} L_k(\boldsymbol{w}), \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^D$ denotes the D-dimensional model parameter to be learned, $L_k(\boldsymbol{w}) = \frac{1}{|\mathcal{D}_k|} \sum_{i \in \mathcal{D}_k} l_i(\boldsymbol{w})$ is the loss function of device k with $l_i(\boldsymbol{w})$ being the sample-wise loss function, and $\mathcal{D} = \bigcup_k \mathcal{D}_k$ is the global dataset. The minimization of $L(\boldsymbol{w})$ is typically carried out through stochastic gradient descent (SGD) algorithm, where device k's local dataset $\mathcal{D}_k$ is split into mini-batches of size B and at each iteration

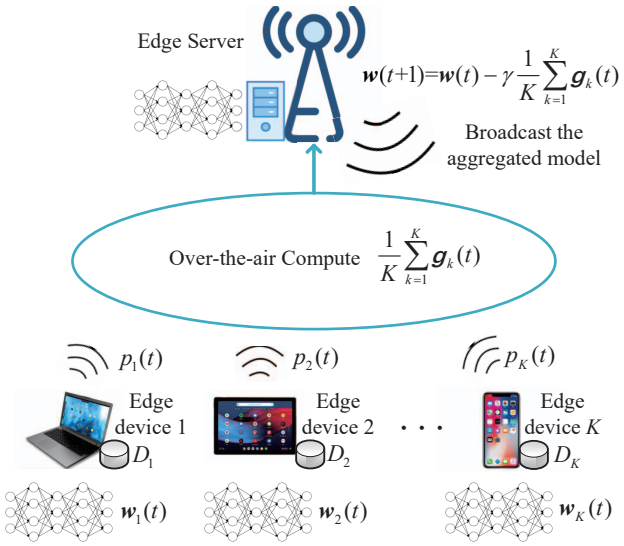

Fig. 1. Illustration of over-the-air federated learning.

$t = 1, 2, ...$, we draw one mini-batch $\mathcal{B}_k(t)$ randomly, and update the model parameter as

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \gamma \frac{1}{K} \sum_{k \in \mathcal{K}} \nabla L_{k,t}^{SGD}(\boldsymbol{w}(t)), \tag{2}$$

with γ being the learning rate and $L_{k,t}^{SGD}(\boldsymbol{w}) = \frac{1}{B} \sum_{i \in \mathcal{B}_k(t)} l_i(\boldsymbol{w})$. The mean of the gradient $\nabla L_{k,t}^{SGD}(\boldsymbol{w}(t))$ in SGD is equal to the gradient $\nabla L(\boldsymbol{w}(t))$ in GD while the variance depends on the mini-batch size and distribution of data (IID or non-IID).

B. Over-The-Air Computation for Gradient Aggregation

We consider block fading channels, where the wireless channels remain unchanged within the duration of each iteration in FL but may change independently from one iteration to another. We define the duration of one iteration as one time block, indexed by $t \in \mathbb{N}$. It is assumed that the channel coefficients over different time blocks are generated from a stationary and ergodic process. Let $\boldsymbol{g}_k(t) \triangleq \nabla L_{k,t}^{SGD}(\boldsymbol{w}(t)) \in \mathbb{R}^D$ denote the gradient vector computed on device k at time block t. The following are key assumptions on the distribution of each entry, $g_{k,d}(t)$, of $\boldsymbol{g}_k(t)$:

- The gradient elements $\{g_{k,d}(t)\}, \forall k \in \mathcal{K}$, are independent and identically distributed over devices k's. This is a default assumption since the distributions of the local datasets are unknown to the edge server and thus are treated equally.
- The gradient elements $\{g_{k,d}(t)\}, \forall t \in \mathbb{N}$, are independent but not identically distributed over iterations t's. The non-identical distribution is valid since the gradient values in general change dynamically at the beginning, then gradually approach to zero as the training goes on.
- The gradient elements $\{g_{k,d}(t)\}, \forall d \in \{1, 2, ..., D\}$, are independent but not identically distributed over gradient vector dimension d's. This assumption is valid as long as the features in a data sample are independent but not identically distributed.

The gradient of interest at the edge server at time block t is given by

$$\boldsymbol{g}(t) = \frac{1}{K}\sum_{k\in\mathcal{K}} \boldsymbol{g}_k(t). \tag{3}$$

At each time block t, all the devices transmit their gradient vectors $\boldsymbol{g}_k(t)$ concurrently in an analog manner, following the AirComp principle. Each transmission block takes a duration of D slots, one slot for one entry in the D-dimensional gradient vector. Each gradient vector $\boldsymbol{g}_k(t)$ is multiplied with a pre-process factor, denoted as $b_k(t)$. The received signal vector at the edge server is given by

$$\boldsymbol{y}(t) = \sum_{k\in\mathcal{K}} b_k(t)h_k(t)\boldsymbol{g}_k(t) + \boldsymbol{n}(t), \tag{4}$$

where $h_k(t)$ denotes the channel coefficient from device k to the edge server and $\boldsymbol{n}(t)$ denotes the additive white Gaussian noise (AWGN) vector with each element having zero mean and variance of σ_n^2. To compensate the channel phase offset and control the actual transmit power at each device, we let $b_k(t) = \frac{\sqrt{p_k(t)}e^{-j\theta_k(t)}}{B_k(t)}$, where $p_k(t) \geq 0$ denotes the transmit power at device $k \in \mathcal{K}$ at each time block t, $\theta_k(t)$ is the phase of $h_k(t)$, and $B_k(t) \triangleq \|\boldsymbol{g}_k(t)\| = \sqrt{\sum_{d=1}^{D} g_{k,d}^2(t)}$ denotes the norm of the gradient $g_k(t)$. Here, we have assumed that each device k can estimate perfectly the channel phase $\theta_k(t)$. By such design of $b_k(t)$, we can rewrite (4) as

$$\boldsymbol{y}(t) = \sum_{k\in\mathcal{K}} \frac{\sqrt{p_k(t)}|h_k(t)|}{B_k(t)}\boldsymbol{g}_k(t) + \boldsymbol{n}(t). \tag{5}$$

Each device $k \in \mathcal{K}$ has a peak power budget P_k, i.e.,

$$p_k(t) \leq P_k, \forall k \in \mathcal{K}, \forall t \in \mathbb{N}. \tag{6}$$

Upon receiving $\boldsymbol{y}(t)$, the edge server applies a denoising factor, denoted by $\eta(t)$, to recover the gradient of interest as

$$\hat{\boldsymbol{g}}(t) = \frac{\boldsymbol{y}(t)}{K\sqrt{\eta(t)}}, \tag{7}$$

where the factor $1/K$ is employed for the averaging purpose.

C. Performance Measure

We are interested in minimizing the distortion of the recovered gradient $\hat{\boldsymbol{g}}(t)$, with respect to (w.r.t.) the ground true gradient $\boldsymbol{g}(t)$. The distortion at a given iteration t is measured by the instantaneous MSE defined as

$$\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{g}}(t) - \boldsymbol{g}(t)\|^2] \\
=&\frac{1}{K^2}\mathbb{E}\left[\left\|\frac{\boldsymbol{y}(t)}{\sqrt{\eta(t)}} - \sum_{k\in\mathcal{K}}\boldsymbol{g}_k(t)\right\|^2\right] \\
=&\frac{1}{K^2}\left[\sum_{d=1}^{D}\sigma_d^2(t)\sum_{k\in\mathcal{K}}\left(\frac{\sqrt{p_k(t)}|h_k(t)|}{\sqrt{\eta(t)}B_k(t)} - 1\right)^2\right. \\
&\left. + \sum_{d=1}^{D} m_d^2(t)\left(\frac{1}{\sqrt{\eta(t)}}\sum_{k\in\mathcal{K}}\frac{\sqrt{p_k(t)}|h_k(t)|}{B_k(t)} - K\right)^2 + \frac{D\sigma_n^2}{\eta(t)}\right],
\end{aligned} \tag{8}$$

where the expectation is over the distribution of the transmitted gradients $\boldsymbol{g}_k(t)$ and the received noise $\boldsymbol{n}(t)$. Note that the gradient norm $B_k(t)$ of each device k can be transmitted to the edge server with negligible communication cost, thus $B_k(t)$ is considered as a known value. In (8), $m_d(t)$ and $\sigma_d^2(t)$ denote the mean (first-order statistics) and variance (second-order statistics) of $g_d(t)$, the d-th entry of gradient $\boldsymbol{g}(t)$ at iteration t, respectively.

D. Gradient Statistics

To facilitate the estimation of gradient statistics to be presented in Section IV, we introduce two alternative parameters. Let $\alpha(t)$ denote the mean squared norm (MSN) of $\boldsymbol{g}(t)$, i.e., $\mathbb{E}[\|\boldsymbol{g}(t)\|^2]$, which is given by

$$\alpha(t) = \sum_{d=1}^{D}\left(\sigma_d^2(t) + m_d^2(t)\right), \tag{9}$$

and let $\beta(t)$ denote the squared multivariate coefficient of variation (SMCV) of $\boldsymbol{g}(t)$, which is given by

$$\beta(t) = \frac{\sum_{d=1}^{D}\sigma_d^2(t)}{\sum_{d=1}^{D} m_d^2(t)}. \tag{10}$$

Then the MSE in (8) can be rewritten as:

$$\begin{aligned}
&\frac{1}{K^2}\left[\frac{\beta(t)\alpha(t)}{\beta(t)+1}\sum_{k\in\mathcal{K}}\left(\frac{\sqrt{p_k(t)}|h_k(t)|}{\sqrt{\eta(t)}B_k(t)} - 1\right)^2\right. \\
&\left. + \frac{\alpha(t)}{\beta(t)+1}\left(\frac{1}{\sqrt{\eta(t)}}\sum_{k\in\mathcal{K}}\frac{\sqrt{p_k(t)}|h_k(t)|}{B_k(t)} - K\right)^2 + \frac{D\sigma_n^2}{\eta(t)}\right].
\end{aligned} \tag{11}$$

III. Optimal Power Control with Known Gradient Statistics

In this section, we formulate and solve the optimal power control problem for minimizing MSE when the gradient statistics $\alpha(t)$ and $\beta(t)$ are known. Due to page limit, the proofs of all the lemmas and theorems in this section are omitted. They can be found in the extended version [14].

For convenience, we omit iteration index t in this section. For each device $k \in \mathcal{K}$, we define its *gain level* with power p and denoising factor η as

$$G_k(p,\eta) = \frac{\sqrt{p}|h_k|}{\sqrt{\eta}B_k}, \tag{12}$$

which indicates the weight of the gradient from device k in the global gradient aggregation (7). Furthermore, we define *capability* of device k as its gain level with peak power P_k and unit denoising factor $\eta = 1$, i.e., $C_k = \frac{\sqrt{P_k}|h_k|}{B_k}$. Then we rank each device according to its capability as:

$$C_1 \leq ... \leq C_k \leq ... \leq C_K. \tag{13}$$

The optimal power control problem is formulated as

$$\begin{aligned}
\text{P1}: \quad \min \quad & \frac{\beta\alpha}{\beta+1}\sum_{k\in\mathcal{K}}\left(G_k(p_k,\eta) - 1\right)^2 \\
& + \frac{\alpha}{\beta+1}\left(\sum_{k\in\mathcal{K}} G_k(p_k,\eta) - K\right)^2 + \frac{D\sigma_n^2}{\eta}
\end{aligned} \tag{14}$$

$$s.t. \qquad 0 \leq p_k \leq P_k, \quad \forall k \in \mathcal{K} \tag{15}$$

$$\eta \geq 0. \tag{16}$$

Different from the power control problem in [13], the objective function in our P1 contains not only the individual misalignment error ($\frac{\beta\alpha}{\beta+1}\sum_{k\in\mathcal{K}}(G_k(p_k,\eta)-1)^2$), but also the composite misalignment error ($\frac{\alpha}{\beta+1}(\sum_{k\in\mathcal{K}}G_k(p_k,\eta)-K)^2$) and the two errors cannot be minimized simultaneously. Problem P1 is non-convex in general. Even if the denoising factor η is given, problem P1 is still hard to solve due to the coupling of each power control p_k. In the following, we derive some properties of the optimal solution.

Lemma 1: Let η^* denote the optimal denoising factor for problem P1. It must satisfy $\eta^* \geq C_1^2$.

Lemma 1 reduces the range of denoising factor. Based on Lemma 1, we have the following lemma.

Lemma 2: The optimal power control policy satisfies $p_k^* = P_k, \forall k \in \{1,...,l\}, p_k^* < P_k, \forall k \in \{l+1,...,K\}$ for some $l \in \mathcal{K}$.

Lemma 2 shows that solving problem P1 is equivalent to minimizing the objective function in the following K subregions, denoted as $\{\mathcal{M}_l\}$, for $l = 1,...,K$, and comparing their corresponding optimal values to obtain the minimum one.

$$\mathcal{M}_l = \Bigg\{\boldsymbol{p}\in\mathbb{R}^K | p_k = P_k, \forall k \in \{1,...,l\}, \\ 0 \leq p_k < P_k, \forall k \in \{l+1,...,K\}\Bigg\}, \forall l \in \mathcal{K}. \tag{17}$$

To facilitate the derivation, we denote $\tilde{\mathcal{M}}_l$ as a relaxed version of $\mathcal{M}_l$ by removing the condition $p_k < P_k$ for $k \in \{l+1,...,K\}$, which is given by

$$\tilde{\mathcal{M}}_l = \Bigg\{\boldsymbol{p}\in\mathbb{R}^K | p_k = P_k, \forall k \in \{1,...,l\}, \\ p_k \geq 0, \forall k \in \{l+1,...,K\}\Bigg\}, \forall l \in \mathcal{K}. \tag{18}$$

Taking the derivative of the objective function (14) w.r.t. p_k equating it to zero for all $k \in \{l+1,...,K\}$, we obtain the optimal power control $\tilde{p}_k(l)$ in the l-th relaxed subregion at any given η as

$$\tilde{p}_k(l) = \left[\frac{\beta+K-\sum_{i=1}^{l}G_i(P_i,\eta)}{\beta+K-l}\right]^2 \cdot \frac{B_k^2\eta}{|h_k|^2}, k \in \{l+1,...,K\}. \tag{19}$$

Note that by such power control in (19) the gain level $G_k(\tilde{p}_k(l),\eta)$ of each device $k \in \{l+1,...,K\}$ is same. Substituting (19) back to (14), and letting the derivative of the objective function (14) w.r.t. η be zero, we can derive a closed-form optimal solution for denoising factor η in the l-th relaxed subregion, given by

$$\sqrt{\tilde{\eta}(l)} = \frac{\frac{\beta\alpha}{\beta+1}\sum_{k=1}^{l}C_k^2 + \frac{\beta\alpha}{(\beta+K-l)(\beta+1)}\left(\sum_{k=1}^{l}C_k\right)^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l)(\beta+1)}\sum_{k=1}^{l}C_k}. \tag{20}$$

Note that $\tilde{p}_k(l)$ may be not less than its power constraint P_k for some $k \in \{l+1,...,K\}$, and thus the corresponding $\tilde{\boldsymbol{p}}(l)$ does not lie in the subregion $\mathcal{M}_l$.

Lemma 3: For the problem defined in the l-th relaxed subregion $\tilde{\mathcal{M}}_l$, if $\exists k \in \{l+1,...,K\}, \tilde{p}_k(l) \geq P_k$, the optimal power $\boldsymbol{p}^*$ of Problem P1 must not be in $\mathcal{M}_l$.

Lemma 3 shows that only the power control $\tilde{\boldsymbol{p}}(l)$'s with $\forall k \in \{l+1,...,K\}, p_k(l) < P_k$ are legal candidates of Problem P1. Let $\mathcal{L}$ denote the index set of relaxed subregions with legal candidate power control. Note that $\mathcal{L}$ is non-empty because $\tilde{\boldsymbol{p}}(K)$ is always a legal power control candidate. Then, we only need to compare the legal candidate values to obtain the minimum MSE:

$$l^* = \arg\min_{l\in\mathcal{L}} V_l, \tag{21}$$

where V_l is the optimal value of (14) in subregion $\mathcal{M}_l$. The optimal solution to problem P1 is derived as follows.

Theorem 1: The optimal transmit power at each device that solves problem P1 is given by

$$p_k^* = \begin{cases} P_k, & \forall k \in \{1,...,l^*\} \\ \left[\frac{\beta+K-\sum_{i=1}^{l^*}G_i(P_i,\eta^*)}{\beta+K-l^*}\right]^2 \cdot \frac{B_k^2\eta^*}{|h_k|^2}, & \forall k \in \{l^*+1,...,K\}, \end{cases} \tag{22}$$

and the optimal denoising factor at the edge server is given by

$$\sqrt{\eta^*} = \frac{\frac{\beta\alpha}{\beta+1}\sum_{k=1}^{l^*}C_k^2 + \frac{\beta\alpha}{(\beta+K-l^*)(\beta+1)}\left(\sum_{k=1}^{l^*}C_k\right)^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l^*)(\beta+1)}\sum_{k=1}^{l^*}C_k}, \tag{23}$$

where l^* is given in (21).

Remark 1: Theorem 1 shows that devices $k \in \{1,...,l^*\}$ with capability not higher than device l^* transmit their gradients with full power, i.e., $p_k = P_k$, while devices $k \in \{l^*+1,...,K\}$ with capability higher than device l^* transmit gradients with the power so that they have the same gain level $G_k(p_k^*,\eta^*) = \frac{\beta+K-\sum_{i=1}^{l^*}G_i(P_i,\eta^*)}{\beta+K-l^*}$, somewhat analogous to channel inversion.

IV. Adaptive Power Control Design with Unknown Gradient Statistics

In this section, we consider practical scenarios where $\alpha(t)$ and $\beta(t)$ are unknown and may vary over time. We propose a method to estimate $\alpha(t)$ and $\beta(t)$ in each time block and then perform adaptive power control based on the estimated results.

A. Parameters Estimation

1) Estimation of $\alpha(t)$: Since we assume that the instantaneous gradient norm of each device, $B_k(t)$, is known at the edge server, then by definition (9), we can estimate the gradient MSN as

$$\hat{\alpha}(t) = \frac{1}{K}\sum_{k\in\mathcal{K}}B_k^2(t). \tag{24}$$

Algorithm 1 FL Process with Adaptive Power Control

1: Initialize $\boldsymbol{w}(0)$ in edge server, $\hat{\beta}(1)$;
2: **for** time block $t = 1, ..., T$ **do**
3: Edge server broadcasts the global model $\boldsymbol{w}(t)$ to all edge devices $k \in \mathcal{K}$;
4: **for** each device $k \in \mathcal{K}$ in parallel **do**
5: $\boldsymbol{g}_k(t) = \nabla L_{k,t}^{SGD}(\boldsymbol{w}(t))$;
6: $B_k(t) = \sqrt{\sum_d g_{k,d}^2(t)}$;
7: Upload $B_k(t)$ to edge server;
8: **end for**
9: Edge server estimates $\hat{\alpha}(t)$ based on (24);
10: Edge server obtains the optimal power control $\boldsymbol{p}^*(t)$ based on (22) and the optimal denoising factor $\eta^*(t)$ based on (23);
11: Edge server sends $p_k^*(t)$ to device k for all $k \in \mathcal{K}$;
12: **for** each device $k \in \mathcal{K}$ in parallel **do**
13: Transmit gradient $\boldsymbol{g}_k(t)$ with power $p_k^*(t)$ to edge server using AirComp;
14: **end for**
15: Edge server receives $\boldsymbol{y}(t)$ and recovers $\hat{\boldsymbol{g}}(t)$ based on (7);
16: Edge server estimates $\hat{\beta}(t+1)$ based on (25);
17: Edge server updates global model $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \gamma\hat{\boldsymbol{g}}(t)$;
18: **end for**
19: Edge server returns $\boldsymbol{w}(T+1)$;

2) Estimation of $\beta(t)$: By definition in (10), the gradient SMCV $\beta(t)$ depends on $m_d(t)$ and $\sigma_d(t)$. Knowing $m_d(t)$ of each dimension d is equivalent to recovering the gradient of interest in (7) at the edge server. Thus we cannot estimate $\beta(t)$ in advance before each device sending its gradient at time block t. As can be verified by simulation, $\beta(t)$ changes slowly over iteration t. Thus we propose to estimate $\beta(t)$ based on the aggregated gradient at time block $t-1$ as below:

$$\hat{\beta}(t) = \frac{\hat{\alpha}(t-1) - \sum_{d=1}^{D} \hat{g}_d^2(t-1)}{\sum_{d=1}^{D} \hat{g}_d^2(t-1)}, \tag{25}$$

where $\hat{\alpha}(t-1)$ estimates $\sum_d \sigma_d^2(t-1) + m_d^2(t-1)$ and $\sum_d \hat{g}_d^2(t-1)$ estimates $\sum_d m_d^2(t-1)$.

B. FL with adaptive power control

In this subsection, we propose the FL process with adaptive power control, which is presented in Algorithm 1. The algorithm has three steps. First, each device locally takes one step of SGD on the current model using its local dataset (line 5). After that each device calculates the norm of its local gradient and uploads it to the edge server with conventional digital transmission (line 6 and line 7). Second, the edge server estimates parameters $\alpha(t)$ and $\beta(t)$ (line 9 and line 16). Then the optimal power control and denoising factor are obtained based on (22) and (23), respectively (line 10). Third, the edge server informs the optimal power control to each device and each device transmits local gradient with the assigned power simultaneously using AirComp to the edge server in an analog manner (line 12-14).

V. Experiment Results

A. Experiment Setup

We conducted experiments on a simulated environment with 100 edge devices, where $K = 10$ devices are selected at random at each iteration to participate in the model training. The wireless channels from each device to the edge server follow IID Rayleigh fading, such that h_k's are modeled as IID complex Gaussian variables with zero mean and unit variance. For each device $k \in \mathcal{K}$, we define $\text{SNR}_k = \mathbb{E}\left[\frac{P_k|h_k|^2}{D\sigma_n^2}\right] = \frac{P_k}{D\sigma_n^2}$ as the average received SNR.

1) Baselines: We compare the proposed power control scheme with the following baseline approaches:

- *Error-free transmission:* the aggregated gradient is updated without any transmission error, which is equivalent to the centralized SGD algorithm.
- *Threshold-based power control in [13]:* this is the power control scheme given in [13], which assumed that signals are normalized. Note that it is actually the special case of our proposed power control scheme with $\beta \to \infty$ by considering the individual mis-alignment error only in Problem P1.
- *Full power transmission:* all devices transmit with full power P_k and the edge server applies the optimal denoising factor in (23), where $l^* = K$.

2) Datasets: We evaluate the training of convolutional neural networks on the MNIST dataset. It consists of 10 categories ranging from digit 0 to 9 and a total of 60000 labeled data samples (we use 50000 samples for training and 10000 samples for testing).

3) Data Distribution: We simulate two types of data partitions among the mobile devices, i.e., the IID setting and non-IID one. For the former, we randomly partition the training samples into 100 equal shards, each of which is assigned to one particular device. While for the latter, we first sort the data by digit label, divide it into 200 shards of 300 images, and randomly assign 2 shards to each device.

B. Experiment Results

The test accuracy over different iterations is shown in Fig. 2 and Fig. 3 with IID and non-IID datasets, respectively, where the average received SNR is set as 10dB and is equal for all devices. It is observed that the proposed adaptive power control scheme outperforms the threshold-based power control scheme and full power transmission scheme in both IID and non-IID dataset partitions. In particular, the threshold-based power control scheme has a much lower accuracy compared to our adaptive power control scheme in the IID partition. This is because in this case, the gradient SMCV is small and thus the MSE is dominated by the composite misalignment error. As a result, the threshold-based power control that considers the individual misalignment error only is much inferior. The full power transmission scheme has a much lower accuracy in

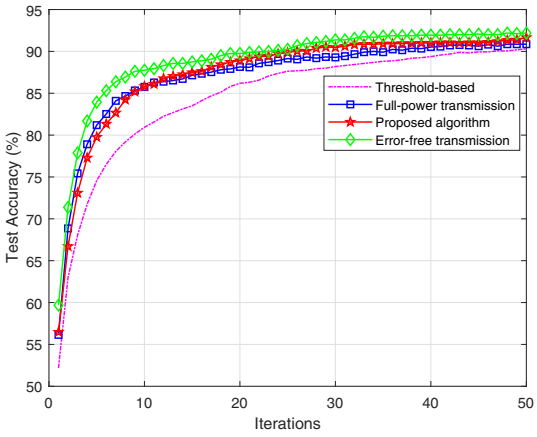

Fig. 2. Performance comparison on IID MNIST Dataset with SNR=10dB.

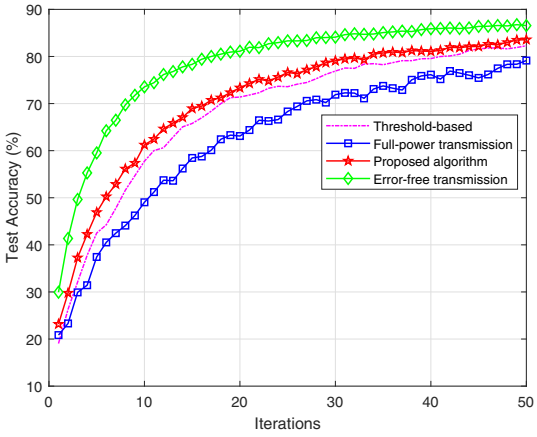

Fig. 3. Performance comparison on Non-IID MNIST Dataset with SNR=10dB.

the non-IID partition. This is because the gradient SMCV is large and therefore the full power transmission scheme fails to minimize the individual misalignment error that dominates the MSE in this case.

Fig. 4 shows the test accuracy with SNR=5dB for all devices and non-IID data partition. It is observed that proposed adaptive power control scheme outperforms both the threshold-based power control and the full power transmission schemes throughout the iterations. This indicates that exploiting the gradient statistics for power control is particularly beneficial in the low SNR region for boosting the learning performance.

VI. Conclusion

This work studied the power control optimization problem for the over-the-air federated learning over fading channels by taking the gradient distribution into account. The structure of the optimal power control for SGD learning mainly depends on the multivariate coefficient of variation of the gradient β as well the mean squared value of the gradient α. We propose an adaptive power control scheme based on the estimated parameter $\hat{\alpha}$ and $\hat{\beta}$. Experiment results confirm that

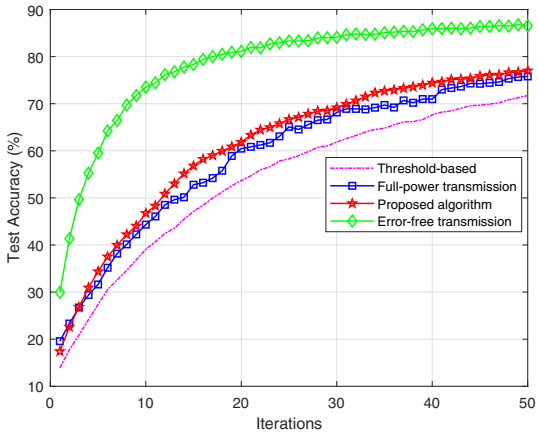

Fig. 4. Performance comparison on Non-IID MNIST Dataset with SNR=5dB.

the accuracy of the over-the-air FL with our proposed adaptive power control outperforms the existing works, especially at low SNR region and with non IID data distribution.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[2] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, p. 12, 2019.

[4] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via randomized quantization and encoding," *Advances in Neural Information Processing Systems 30*, vol. 3, pp. 1710–1721, 2018.

[5] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," *arXiv preprint arXiv:1704.05021*, 2017.

[7] Y. Tsuzuku, H. Imachi, and T. Akiba, "Variance-based gradient compression for efficient distributed deep learning," *arXiv preprint arXiv:1802.06058*, 2018.

[8] 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 Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[9] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Transactions on information theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

[10] G. Zhu, Y. Wang, and K. Huang, "Low-latency broadband analog aggregation for federated edge learning," *arXiv preprint arXiv:1812.11494*, 2018.

[11] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *arXiv preprint arXiv:1812.11750*, 2018.

[12] M. M. Amiri and D. Gunduz, "Federated learning over wireless fading channels," *arXiv preprint arXiv:1907.09769*, 2019.

[13] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimal power control for over-the-air computation in fading channels," *arXiv preprint arXiv:1906.06858*, 2019.

[14] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning in fading channels," *arXiv preprint arXiv:2003.02089*, 2020.