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

Naifu Zhang and Meixia Tao

Abstract

Federated learning (FL) is a promising technique that enables many edge devices to train a machine learning model collaboratively in wireless networks. By exploiting the superposition nature of wireless waveforms, over-the-air computation (AirComp) can accelerate model aggregation and hence facilitate communication-efficient FL. Due to channel fading, power control is crucial in AirComp. Prior works assume that the signal to be aggregated from each device, i.e., local gradient, can be normalized with zero mean and unit variance. In FL, however, gradient statistics vary over both training iterations and feature dimensions, and are unknown in advance. This paper studies the power control problem for over-the-air FL by taking gradient statistics into account. The goal is to minimize the aggregation error by jointly optimizing the transmit power at each device and the denoising factor at the edge server. We obtain the optimal policy in closed form when gradient statistics are given. Notably, we show that the optimal transmit power at each device is continuous and monotonically decreases with the squared multivariate coefficient of variation (SMCV) of gradient vectors. We also propose an estimation method of gradient statistics with negligible communication cost. Experimental results demonstrate high learning performance by using the proposed scheme.

Index Terms

Federated learning, over-the-air computation, power control, fading channel.

I. Introduction

The proliferation of mobile devices such as smartphones, tablets, and wearable devices has revolutionized people's daily lives. Due to the growing computation and sensing capabilities of these devices, a wealth of data has been generated each day. This has promoted a wide range of artificial intelligence (AI) applications such as image recognition and natural language processing. Traditional machine learning procedure, including both training and inference, relies on cloud computing on a centralized data center with computing, storage, and full access to the entire data set. Wireless edge devices are thus required to transmit their collected data to a central parameter server, which can be very costly in terms of energy and bandwidth consumption, and unfavorable due to response delay and privacy concerns. It is thus increasingly desired to let edge devices engage in the learning process by keeping the collected data locally and performing training/inference either collaboratively or individually. This emerging technology is known as Edge Machine Learning [2] or Edge Intelligence [3].

Federated learning [4]–[6] 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

N. Zhang and M. Tao are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, P. R. China. (email: arthaslery@sjtu.edu.cn; mxtao@sjtu.edu.cn.). Part of this work is to be presented at IEEE ICC 2020 workshop [1]

a shared model from the edge server, computes an update to the current model by learning from its 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 [7], [8], gradient sparsification [9], [10], and infrequent local update [4], [11]. Nevertheless, the communication cost of FL is still proportional to the number of edge devices, and thus inefficient in large-scale environment. Recently, a fast model aggregation approach is proposed for FL by applying over-the-air computation (AirComp) principle [12], such as in [13]–[16]. This is accomplished by exploiting the waveform superposition nature of the wireless medium to compute the desired function of the distributed local gradients (i.e., the weighted average function) by concurrent transmission. Such AirComp-based FL, referred to as *over-the-air FL* in this work, can dramatically save the uplink communication bandwidth.

Due to the channel fading, device selection and power control are crucial to achieve a reliable and high-performance over-the-air FL. In [17], the authors jointly optimize the transmit power at edge devices and the receive scaling factor (known as denoising factor) at the edge server for minimizing the mean square error (MSE) of the aggregated signal. It is shown that the optimal transmit power 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. The work [17], however, is purely for AirComp-based signal aggregation (not in the context of learning), where the signal on each device is assumed to be independent and identically distributed (IID) with zero mean and unit variance. For AirComp-based gradient aggregation in FL, the work [14] 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 [13] proposes a joint device selection and receiver beamforming design method to find the maximum number of selected devices with MSE requirements to improve the learning performance. As in [17], it is assumed in both [13] and [14] that the signal (i.e., the local 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 [16] 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 [13], [14], [17] is designed so that the gradient vectors sent from the 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 training iterations is independent but not necessarily identical; and even in the same iteration, the distribution of each entry 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 distribution is unknown for each device, normalizing the gradient to a distribution with zero mean and unit variance is infeasible. As such, due to the neglect of the above characteristics of gradient distribution, the existing power control methods for over-the-air FL may not perform efficiently in practice.

Motivated by the above issue, in this paper, we study the optimal power control problem for over-the-air FL in fading channels by taking gradient statistics into account. Our goal is to minimize the MSE of the aggregated model, and hence improve the accuracy of FL, by jointly optimizing the transmit power at each device and the denoising factor at the edge server given the first-order and second-order statistics of gradients at each iteration. 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 iteration of the model training when the first-order and second-order statistics of the gradient vectors are known. We then formulate a joint optimization problem of transmit power at edge devices and denoising factor at the edge server for MSE minimization subject to individual peak power constraints at edge devices. By decomposing this non-convex problem into subproblems defined on different subregions, we obtain the optimal power control strategy in closed form. Unlike the existing threshold-based power control for normalized signal in [17], our optimal transmit power depends not only on the channel quality and noise power, but also heavily on the gradient statistics. In particular, we find that the relative dispersion of the gradient values, i.e., the squared multivariate coefficient of variation (SMCV) of the gradient vectors, plays a key role in the optimal transmit power control. Specifically, we prove that the optimal transmit power of each device is a continuous and monotonically decreasing function of the gradient SMCV.
- *Optimal power control in special cases:* In the special case where the gradient SMCV approaches infinity, which could happen when the model training converges and/or the dataset is highly non-IID, we show that there is an optimal threshold for the aggregation capabilities of the devices, below which the devices transmit with full power and above which the devices transmit at the power to equalize the weights of their gradients for aggregation to one. In the other special case where the gradient SMCV approaches zero, which could happen when the model training just begins, we show that the optimal power control is to let all the devices transmit with their peak powers.
- *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 the power values in each iteration based on the estimation results. The communication cost consumed by estimating the gradient statistics is negligible compared to the transmission of the entire gradient vector.

To evaluate the efficiency of the proposed power control scheme, we implement the FL in PyTorch for AI applications of three datasets: MNIST, CIFAR-10 and SVHN. Experimental results demonstrate that the over-the-air FL with the proposed adaptive power control obtains higher model accuracy than that with existing power control methods (full power transmission and threshold-based power control for normalized signal [17]). In particular, while the full power transmission performs poorly in high SNR region and non-IID data distribution and the threshold-based power control for normalized signal [17] performs poorly in low SNR region and IID data distribution, the proposed power control can perform very well over a wide range of scenarios by exploiting the gradient statistics.

The rest of this paper is organized as follows. The over-the-air FL system is modeled in Section II. Section III describes the optimal power control strategy with known gradient statistics. In Section IV, we introduce an adaptive power control scheme when the gradient statistics are unknown in advance. Section V provides experimental results. Finally we conclude the paper in Section VI.

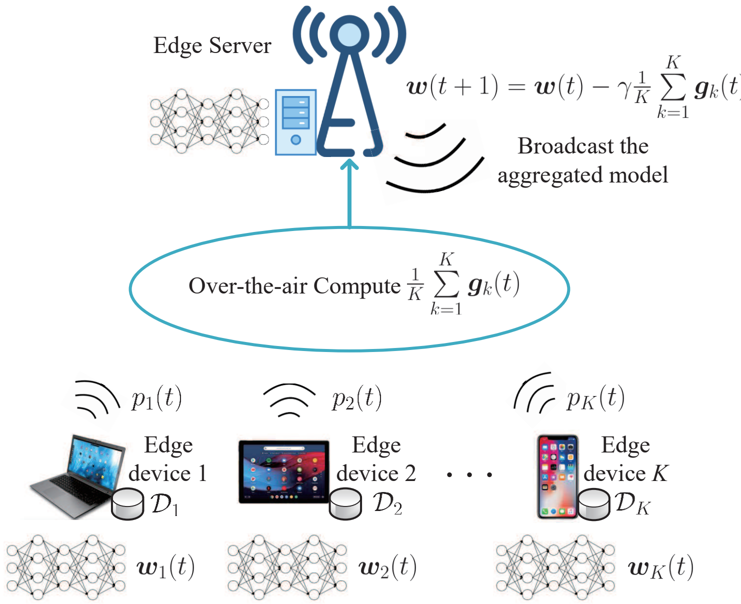


Fig. 1. Illustration of over-the-air federated learning.

II. System Model

A. Federated Learning Over Wireless Networks

We consider a wireless FL framework as illustrated in Fig. 1, where a shared AI model (e.g., a classifier) is trained collaboratively across K single-antenna edge devices via the coordination of a single-antenna edge server. 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, denoted 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 quantifying the prediction error of the model $\boldsymbol{w}$ on the local dataset collected at the k-th device, with $l_i(\boldsymbol{w})$ being the sample-wise loss function, and $\mathcal{D}=\bigcup_{k\in\mathcal{K}}\mathcal{D}_k$ is the union of all datasets. 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 $t=1,2,...$, 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 $L_{k,t}^{SGD}(\boldsymbol{w})=\frac{1}{B}\sum_{i\in\mathcal{B}_k(t)}l_i(\boldsymbol{w})$. Note that 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 gradient descent (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 $g_{k,d}(t), d \in \{1, ..., D\}$ 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 at different devices, whether identical or non-identical, are unknown to the edge server and thus the distributions of the local gradients trained from these local datasets are treated equally.
- The gradient elements $\{g_{k,d}(t)\}, \forall t \in \mathbb{N}$, are independent but non-identically distributed over iterations t's. In other words, the gradient distribution is non-stationary over time. The non-stationary distribution is valid since the gradient values in general change rapidly 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 non-identically distributed over gradient vector dimension d's. This assumption is valid as long as the features in a data sample are independent but non-identically distributed, which is typically the case.

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

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

To obtain (3), all the devices transmit their gradient vectors $\boldsymbol{g}_k(t)$ concurrently in an analog manner, following the AirComp principle as shown in Fig. 1. 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-processing 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 complex-valued channel coefficient from device k to the edge server and $\boldsymbol{n}(t)$ denotes the additive white Gaussian noise (AWGN) vector at the edge server 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 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 gradient norm of device k. Here, we have assumed that each device k can estimate perfectly its own channel phase $\theta_k(t)$. To design the optimal power control policy $p_k(t)$, for $k \in \mathcal{K}$, we also assume that the edge server knows the channel amplitude $|h_k|$ of all devices. 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.,

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

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

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

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

Observing (8) closely, we find that the MSE consists of three components, representing the individual misalignment error (the first term), the composite misalignment error (the second term), and the noise-induced error (the third term), respectively. The individual misalignment error is weighted by the gradient variance $\sum_{d=1}^{D} \sigma_d^2(t)$ while the composite misalignment error is weighted by the gradient mean $\sum_{d=1}^{D} m_d^2(t)$. In the special case when the gradient has zero mean, the MSE in (8) reduces to that in [17] where the composite misalignment error is absent. Our objective is to minimize MSE in (8), by jointly optimizing the transmit power $p_k(t)$ at all devices and the denoising factor $\eta(t)$ at the edge server, subject to the individual power budget in (6).

D. Gradient Statistics

In general, the individual misalignment error and the composite misalignment error in MSE of the gradient aggregation (8) cannot be minimized simultaneously due to the peak power budget on each device. It is difficult to capture the tradeoff between the two errors by directly using

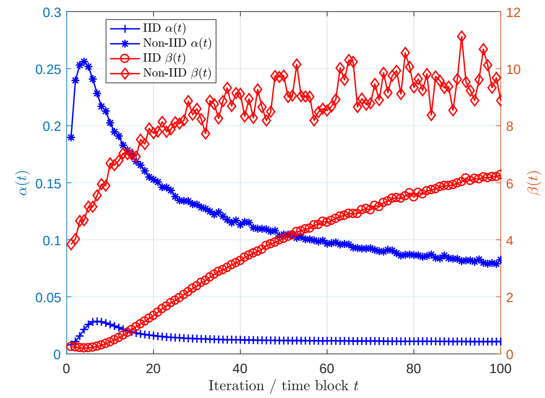


Fig. 2. Experimental results of MSN $\alpha(t)$ (left y-axis in linear scale) and SMCV $\beta(t)$ (right y-axis) over iterations for dataset MNIST, where the number of edge devices is 10 and the local mini-batch size is 20.

their respective weights, namely, the gradient variance and gradient mean. To tackle this issue, we introduce two alternative parameters of gradient statistics in this subsection.

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

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

While $\alpha(t)$ measures the average absolute gradient values at iteration t, $\beta(t)$ is a measure of relative dispersion of the gradient values at iteration t. Shortly later, we shall show that the MSE of the model aggregation is mainly dominated by $\beta(t)$, rather than $\alpha(t)$. Figs. 2-4 illustrate the experimental results of the alternative gradient statistics $\alpha(t)$ and $\beta(t)$ of three datasets, MNIST, CIFAR-10, and SVHN, respectively, where the gradients are updated ideally without any transmission error. Both IID and non-IID partitions are considered for the training dataset. Each value of $\alpha(t)$ and $\beta(t)$ is obtained by averaging over 300 model trainings. It is observed that as the training time increases, the gradient MSN $\alpha(t)$ decreases while the gradient SMCV $\beta(t)$ increases for all the three datasets. This result is in agreement with the intuition that the absolute gradient values in SGD-based learning gradually vanish when the training iteration goes on, but their relative variation remains significant over each iteration. It is also observed that the gradient SMCV $\beta(t)$ in non-IID partition is much larger than that in IID partition for all the three datasets. This indicates that the gradient distribution with non-IID dataset partition is more dispersive than that with IID dataset partition as expected.

By (9) and (10), the MSE in (8) can be rewritten as (omitting the constant coefficient $1/K^2$ for convenience)

$$\mathrm{MSE}(t) = \frac{\beta(t)\alpha(t)}{\beta(t)+1} \sum_{k\in\mathcal{K}} \left(\frac{\sqrt{p_k(t)}|h_k(t)|}{\sqrt{\eta(t)}B_k(t)} - 1 \right)^2$$

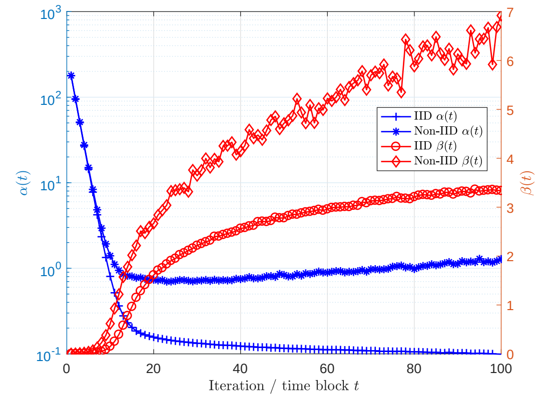


Fig. 3. Experimental results of MSN $\alpha(t)$ (left y-axis in log scale) and SMCV $\beta(t)$ (right y-axis) over iterations for dataset CIFAR-10, where the number of edge devices is 10 and the local mini-batch size is 200.

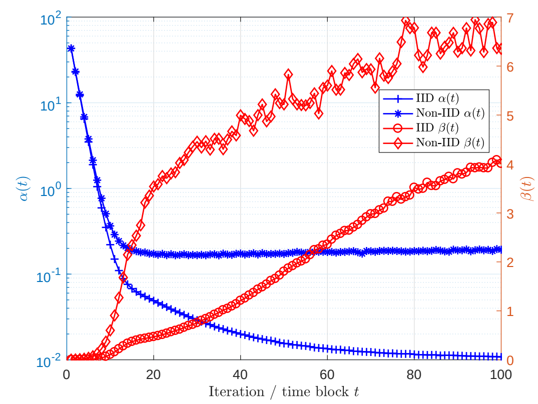


Fig. 4. Experimental results of MSN $\alpha(t)$ (left y-axis in log scale) and SMCV $\beta(t)$ (right y-axis) over iterations for dataset SVHN, where the number of edge devices is 10 and the local mini-batch size is 300.

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

It is seen from (11) that while the gradient MSN $\alpha(t)$ appears linearly in the weights of both individual and composite misalignment errors, the gradient SMCV $\beta(t)$ plays a more distinguishing role in the MSE expression. In particular, when $\beta(t)\to 0$, which could be the case when the model training just begins (as could be verified by Figs. 2-4), the individual signal misalignment error can be neglected. When $\beta(t)\to\infty$, which could be the case when the model training converges or during the middle of the training when the dataset is highly non-IID (as could be verified by Figs. 2-4), the composite signal misalignment error disappears.

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. For convenience, we omit iteration index t in this section. For each device $k\in\mathcal{K}$, we define its *aggregation 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)[1]. Furthermore, we define *aggregation capability* of device k as its aggregation level with peak power P_k and unit denoising factor $\eta = 1$, i.e., $C_k = G_k(P_k, 1) = \frac{\sqrt{P_k}|h_k|}{B_k}$. Without loss of generality, we assume that

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

A. Power Control Problem for General Case

In this subsection, we consider the optimal power control problem for MSE minimization in the general case with arbitrary β. The problem is formulated as

$$\begin{aligned}
\mathcal{P}_1: \quad \min \quad & \frac{\beta\alpha}{\beta+1}\sum_{k\in\mathcal{K}}\left(G_k(p_k,\eta)-1\right)^2+\frac{\alpha}{\beta+1}\left(\sum_{k\in\mathcal{K}}G_k(p_k,\eta)-K\right)^2+\frac{D\sigma_n^2}{\eta} && \text{(14a)}\\
s.t. \quad & 0\leq p_k\leq P_k, \quad \forall k\in\mathcal{K} && \text{(14b)}\\
& \eta\geq 0. && \text{(14c)}
\end{aligned}$$

Different from the power control problem in [17], the objective function in (14a) 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 ($\frac{\alpha}{\beta+1}(\sum_{k\in\mathcal{K}}G_k(p_k,\eta)-K)^2$). Problem $\mathcal{P}_1$ is non-convex in general. Even if the denoising factor η is given, problem $\mathcal{P}_1$ is still hard to solve due to the coupling of each transmit power p_k. In the following, we present some properties of the optimal solution.

Lemma 1: The optimal denoising factor η^* for problem $\mathcal{P}_1$ satisfies $\eta^* \geq C_1^2$.

Proof: Please refer to Appendix A. ■

Lemma 1 reduces the range of η^* and shows that the optimal transmit power of the 1-st device is $p_1^* = P_1$.

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

Proof: Please refer to Appendix B. ■

Based on Lemma 2, solving problem $\mathcal{P}_1$ can be equivalent to minimizing the objective function in the following K exclusive subregions of the global power region, denoted as $\{\mathcal{M}\}_{l\in\mathcal{K}}$ and comparing their corresponding optimal solutions to obtain the global optimal solution:

$$\mathcal{M}_l=\left\{[p_1,\cdots,p_k]\in\mathbb{R}^K|p_k=P_k,\forall k\in\{1,...,l\},0\leq p_k<P_k,\forall k\in\{l+1,...,K\}\right\}. \tag{15}$$

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

$$\tilde{\mathcal{M}}_l=\left\{[p_1,\cdots,p_k]\in\mathbb{R}^K|p_k=P_k,\forall k\in\{1,...,l\},p_k\geq 0,\forall k\in\{l+1,...,K\}\right\}. \tag{16}$$

[1]The weight should be 1 for all devices in the ideal case.

For the sub-problem defined in each relaxed subregion $\tilde{\mathcal{M}}_l$, for $l \in \mathcal{K}$, taking the derivative of the objective function (14a) w.r.t. p_k and equating it to zero for all $k \in \{l+1, ..., K\}$, we obtain the optimal transmit power $\tilde{p}^*_{l,k}$ at any given η as

$$\tilde{p}^*_{l,k} = \left[\frac{\beta + K - \sum_{i=1}^{l} G_i(P_i, \eta)}{\beta + K - l}\right]^2 \cdot \frac{B_k^2 \eta}{|h_k|^2}, k \in \{l+1, ..., K\}. \tag{17}$$

Note that by such power control in (17), the aggregation level $G_k(\tilde{p}^*_{l,k}, \eta)$ of each device $k \in \{l+1, ..., K\}$ is the same, given by

$$G_0(l) = \frac{\beta + K - \frac{1}{\sqrt{\eta}} \sum_{i=1}^{l} C_i}{\beta + K - l}. \tag{18}$$

Substituting (17) back to (14a) and letting its derivative w.r.t. η be zero, we derive the optimal denoising factor η in closed-form for problem $\mathcal{P}_1$ defined in the l-th relaxed subregion, i.e.,

$$\sqrt{\tilde{\eta}^*_l} = \frac{\frac{\beta\alpha}{\beta+1} \sum\limits_{i=1}^{l} C_i^2 + \frac{\beta\alpha}{(\beta+K-l)(\beta+1)} \left(\sum\limits_{i=1}^{l} C_i\right)^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l)(\beta+1)} \sum\limits_{i=1}^{l} C_i}. \tag{19}$$

Note that $\tilde{p}^*_{l,k}$ may be not less than its power constraint P_k for some $k \in \{l+1, ..., K\}$ and thus the corresponding $\tilde{\boldsymbol{p}}^*_l$ may not lie in the subregion $\mathcal{M}_l$. If this happens, the optimal transmit power of the sub-problem defined in subregion $\mathcal{M}_l$ is irrelevant and does not need to be considered. This is given in the following lemma.

Lemma 3: For the power control problem $\mathcal{P}_1$ defined in the l-th relaxed subregion $\tilde{\mathcal{M}}_l$, if $\exists k \in \{l+1, ..., K\}$ such that $\tilde{p}^*_{l,k} \geq P_k$, the global optimal power $\boldsymbol{p}^* \triangleq [p^*_1, ...p^*_K]$ of problem $\mathcal{P}_1$ must not be in $\mathcal{M}_l$.

Proof: Please refer to Appendix C. ■

Lemma 3 shows that only the transmit power vectors $\tilde{\boldsymbol{p}}^*_l$'s with elements satisfying $\tilde{p}^*_{l,k} < P_k, \forall k \in \{l+1, ..., K\}$ are legal candidates of problem $\mathcal{P}_1$. Let $\mathcal{L}$ denote the index set of the corresponding relaxed subregions. Note that $\mathcal{L}$ is non-empty because $\tilde{\boldsymbol{p}}^*_K$ is always a legal transmit power candidate. Then, we only need to compare the legal candidate values to obtain the minimum MSE

$$l^* = \arg\min_{l \in \mathcal{L}} V_l, \tag{20}$$

where V_l is the optimal value of (14a) in subregion $\mathcal{M}_l$ of $\mathcal{P}_1$ and can be easily obtained by substituting (17) and (19) back to (14a). The optimal solution to problem $\mathcal{P}_1$ is given as follows.

Theorem 1: The optimal transmit power at each device that solves problem $\mathcal{P}_1$ is given by

$$p^*_k = \begin{cases} P_k, & \forall k \in \{1, ..., l^*\} \\ \left[\frac{\beta+K-\sum_{i=1}^{l^*} G_i(P_i, \eta^*)}{\beta+K-l^*}\right]^2 \cdot \frac{B_k^2 \eta^*}{|h_k|^2}, & \forall k \in \{l^*+1, ..., K\}, \end{cases} \tag{21}$$

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

$$\sqrt{\eta^*} = \frac{\frac{\beta\alpha}{\beta+1} \sum\limits_{i=1}^{l^*} C_i^2 + \frac{\beta\alpha}{(\beta+K-l^*)(\beta+1)} \left(\sum\limits_{i=1}^{l^*} C_i\right)^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l^*)(\beta+1)} \sum\limits_{i=1}^{l^*} C_i}, \tag{22}$$

where l^* is given in (20).

Proof: Please refer to Appendix D. ■

Remark 1: Theorem 1 shows that these devices $k \in \{1, ..., l^*\}$ with aggregation capability not higher than that of device l^* should transmit their gradients with full power, i.e., $p_k = P_k$, while devices $k \in \{l^*+1, ..., K\}$ with aggregation capability higher than that of device l^* should transmit with the power so that they have the same aggregation level, given by

$$G_0^* = \frac{\beta + K - \sum_{i=1}^{l^*} G_i(P_i, \eta^*)}{\beta + K - l^*}, \tag{23}$$

somewhat analogous to channel inversion.

B. On The Optimal Transmit Power Function

In this subsection, we analyse the optimal transmit power $\boldsymbol{p}^*$ as a vector function of the gradient SMCV β and the noise variance σ_n^2, i.e., $\boldsymbol{p}^*(\beta, \sigma_n^2) = [p_1^*(\beta, \sigma_n^2), ..., p_K^*(\beta, \sigma_n^2)]$. Note that the vector function $\boldsymbol{p}^*(\beta, \sigma_n^2)$ might have abrupt changes at some (β, σ_n^2) due to the discrete nature of the optimal device threshold l^* in Theorem 1. In this work, however, we can show that $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is continuous[2] everywhere for all $(\beta \geq 0, \sigma_n^2 \geq 0)$. Define $\mathcal{X}_l \subseteq \mathbb{R}^2$ as the domain of the vector function $\boldsymbol{p}^*(\beta, \sigma_n^2)$ when $\boldsymbol{p}^*(\beta, \sigma_n^2)$ lies in the subregion $\mathcal{M}_l$, for $l \in \mathcal{K}$. That is, $\forall(\beta, \sigma_n^2) \in \mathcal{X}_l$, one can have $l^*(\beta, \sigma_n^2) = l$.

We first show that $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is continuous and monotonic within each domain $\mathcal{X}_l$, $l \in \mathcal{K}$. Let $\tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)$ denote the optimal power of the sub-problem defined in the l-th relaxed subregion $\tilde{\mathcal{M}}_l$, $\forall l \in \mathcal{K}$. Based on (17) and (19), it is obvious that $\tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)$ is continuous within each domain $\mathcal{X}_l$. Moreover, $\tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)$ increases monotonically with the noise variance σ_n^2 since enlarging σ_n^2 increases both $\tilde{\eta}_l^*$ and $\tilde{p}_{l,k}^*$, for $k \in \{l+1, ..., K\}$. Furthermore, $\tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)$ decreases monotonically with the gradient SMCV β since it can be easily shown that $\frac{\partial \tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)}{\partial \beta}$ is always negative. Thus $\tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)$ is monotonic within each domain $\mathcal{X}_l$, $l \in \mathcal{K}$. By definition, within each domain $\mathcal{X}_l$, $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is equal to $\tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)$, thus, the vector function $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is also continuous and monotonic within each domain $\mathcal{X}_l$, for $l \in \mathcal{K}$.

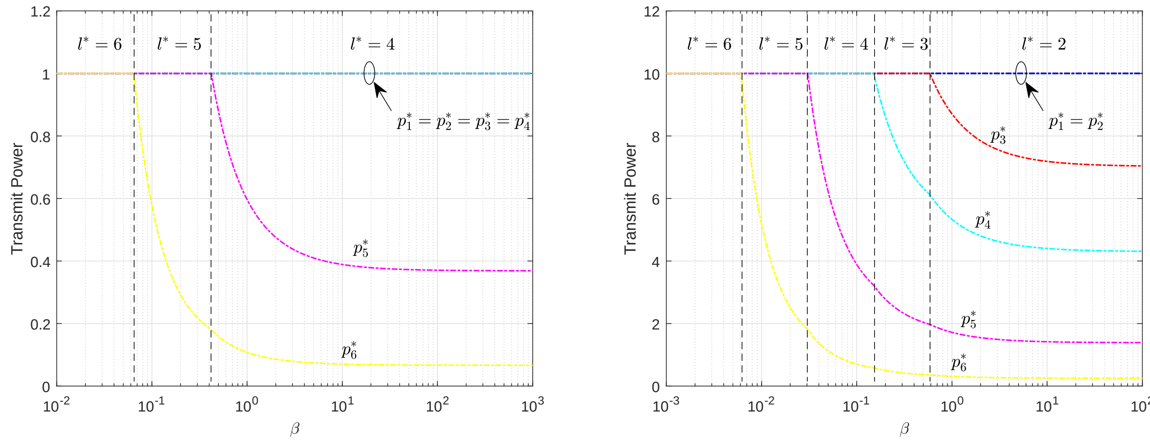
Then we find the boundary of each domain $\mathcal{X}_l$, for $l \in \mathcal{K}$ and the corresponding optimal transmit power $\boldsymbol{p}^*$. To this end, we need the following lemma on the lower and upper bounds of the optimal transmit power values.

Lemma 4: The optimal transmit power $\boldsymbol{p}^*$ lies in subregion $\mathcal{M}_l$ if and only if the optimal transmit power $\tilde{\boldsymbol{p}}_l^*$ in the l-th relaxed subregion $\tilde{\mathcal{M}}_l$ satisfies: $(\frac{C_l B_k}{|h_k|})^2 \leq \tilde{p}_{l,k}^* < (\frac{C_{l+1} B_k}{|h_k|})^2, \forall k \in \{l+1, ..., K\}$.

Proof: Please refer to Appendix E. ■

Lemma 4 shows that in each domain $\mathcal{X}_l, l \in \mathcal{K}$, the range of $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is left-closed and right-open intervals in $\left[(\frac{C_l B_k}{|h_k|})^2, (\frac{C_{l+1} B_k}{|h_k|})^2\right), \forall k \in \{l+1, ..., K\}$. Recall that $\tilde{\boldsymbol{p}}_l^*(\beta, \sigma_n^2)$ is continuous and monotonic. The optimal transmit power $\boldsymbol{p}^*(\beta, \sigma_n^2) = \boldsymbol{p}_l^*(\beta, \sigma_n^2)$ sits on the lower bound of the range when (β, σ_n^2) is at the lower boundary of domain $\mathcal{X}_l$, denoted as $\mathcal{L}_l \triangleq \left\{(\beta, \sigma_n^2) | \tilde{p}_{l,k}^*(\beta, \sigma_n^2) = (\frac{C_l B_k}{|h_k|})^2, \forall k \in \{l+1, ..., K\}\right\}$, and $\boldsymbol{p}^*(\beta, \sigma_n^2) = \boldsymbol{p}_l^*(\beta, \sigma_n^2)$ approaches the

[2] $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is a continuous vector valued function if and only if each element $p_k^*(\beta, \sigma_n^2)$, for $k \in \{1, ..., K\}$ is a continuous function.



(a) Average received SNR=0dB.

(b) Average received SNR=10dB.

Fig. 5. Illustration of the optimal transmit power $\boldsymbol{p}^*$ as a function of gradient SMCV β.

upper bound of the range when (β, σ_n^2) approaches the upper boundary of domain $\mathcal{X}_l$, denoted as $\mathcal{U}_l \triangleq \left\{(\beta, \sigma_n^2) | \tilde{p}_{l,k}^*(\beta, \sigma_n^2) = (\frac{C_{l+1}B_k}{|h_k|})^2, \forall k \in \{l+1, ..., K\}\right\}$.

Next, we consider the continuity of $\boldsymbol{p}^*(\beta, \sigma_n^2)$ at boundaries $\mathcal{L}_l$ and $\mathcal{U}_l$ for each $l \in \mathcal{K}$ in the following lemma.

Lemma 5: The optimal transmit power function $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is continuous at $\mathcal{U}_l = \mathcal{L}_{l+1}$, for $l \in \{1, ..., K-1\}$.

Proof: Please refer to Appendix F. ■

Finally, we can formally conclude the following property of the optimal transmit power function with respect to the gradient statistics and noise variance.

Property 1: The optimal transmit power $\boldsymbol{p}^*(\beta, \sigma_n^2)$ of problem $\mathcal{P}_1$ is a continuous and monotonic vector function in the entire domain of $(\beta \geq 0, \sigma_n^2 \geq 0)$. Moreover, it decreases with β and increases with σ_n^2.

Take a system with $K = 6$ devices for illustration. Fig. 5 shows the optimal transmit power of each device p_k^* as a function of gradient SMCV β, together with the corresponding optimal index l^*. The results are based on one channel realization of each device, $|h_k|$, taken independently from normalized Rayleigh distribution, given by $[0.50, 0.82, 0.85, 1.16, 2.09, 2.83]$. The peak power constraint of each device is set to be same, resulting in the same average received SNR, given by $\frac{P_k}{D\sigma_n^2} = $ 5dB and 10dB, respectively. The gradient norms of devices, B_k, are $[0.23, 0.31, 0.26, 0.28, 0.28, 0.16]$, noise variance $\sigma^2 = 1$ and $D = 1$. Fig. 5 clearly verifies that the optimal transmit power $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is a continuous and monotonically decreasing vector function of the gradient SMCV β. In particular, it is seen that when $\beta \to 0$, all the devices transmit with full power; when beta increases, the power of the devices with large aggregation capability decreases, then gradually approaches a constant, and the larger the aggregation capability is, the smaller the constant transmit power is. In addition, it is observed from Fig. 5 that when the peak power budget increases, more devices can transmit with less power than its peak value. Equivalently, the optimal transmit power decreases when the noise variance decreases.

C. Power Control Problem for Special Cases

In this subsection, we provide some discussions on the optimal power control policy in the two special cases where the gradient SMCV $\beta \to \infty$ and $\beta \to 0$, respectively.

1) $\beta \to \infty$: As discussed in Section II-D, this case may happen when the model training converges and/or the dataset is highly non-IID. In this case, the composite signal misalignment error in the MSE expression disappears. Thus problem $\mathcal{P}_1$ reduces to the power control problem in [17]. To be self-contained, we re-state the problem and the solution as follows

$$\mathcal{P}_2: \quad \min \quad \alpha \sum_{k\in\mathcal{K}} (G_k(p_k,\eta)-1)^2 + \frac{D\sigma_n^2}{\eta} \tag{24a}$$

$$s.t. \qquad 0 \le p_k \le P_k, \quad \forall k \in \mathcal{K} \tag{24b}$$

$$\eta \ge 0. \tag{24c}$$

Theorem 2 ([17]): The optimal transmit power that solves problem $\mathcal{P}_2$ has a threshold-based structure, i.e.,

$$p_k^* = \begin{cases} P_k, & \forall k \in \{1,...,l^*\} \\ \frac{B_k^2\eta^*}{|h_k|^2}, & \forall k \in \{l^*+1,...,K\}, \end{cases} \tag{25}$$

where the optimal denoising factor is given as

$$\eta^* = \left(\frac{\alpha\sum_{i=1}^{l^*} C_i^2 + D\sigma_n^2}{\alpha\sum_{i=1}^{l^*} C_i}\right)^2, \tag{26}$$

and l^* is given in (20). Furthermore, it holds that $C_{l^*}^2 \le \eta^* \le C_{l^*+1}^2$.

Proof: Please refer to [17, Theorem 1]. ∎

Note that the optimal denoising factor η^* for problem $\mathcal{P}_2$ can be interpreted as the threshold, since whether a device transmits with full power or not depends entirely on the comparison between its aggregation capability C_k and $\sqrt{\eta^*}$. Moreover, this threshold increases with σ_n^2. In the extreme case when the noise power is very large and so is the threshold, all the devices shall transmit with full power. Nevertheless, such threshold interpretation of η^* does not hold for problem $\mathcal{P}_1$ with general β.

2) $\beta \to 0$: As discussed in Section II-D, $\beta \to 0$ could happen when the model training just begins. In this case, the individual signal misalignment error disappears in the MSE expressions. The original problem $\mathcal{P}_1$ reduces to

$$\mathcal{P}_3: \quad \min \quad \alpha\left(\sum_{k\in\mathcal{K}} G_k(p_k,\eta) - K\right)^2 + \frac{D\sigma_n^2}{\eta} \tag{27a}$$

$$s.t. \qquad 0 \le p_k \le P_k, \quad \forall k \in \mathcal{K} \tag{27b}$$

$$\eta \ge 0. \tag{27c}$$

Theorem 3: The optimal transmit power that solves problem $\mathcal{P}_3$ is full power transmission, i.e.,

$$p_k^* = P_k, \quad \forall k \in \mathcal{K}, \tag{28}$$

and the optimal denoising factor is given by

$$\eta^* = \left(\frac{\alpha \big(\sum_{i \in \mathcal{K}} C_i \big)^2 + D\sigma_n^2}{\alpha K \sum_{i \in \mathcal{K}} C_i} \right)^2. \tag{29}$$

Proof: Please refer to Appendix G. ■

Remark 2: The optimal solution of problem $\mathcal{P}_3$ is the special case of the solution of problem $\mathcal{P}_1$ with $l^* = K$. Note that the direction of the gradient vector received from each device at the edge server is independent to the power of the transmitting device. Thus, increasing the power of all devices can reduce the noise-induced error when the composite signal misalignment error is fixed.

IV. Adaptive Power Control with Unknown Gradient Statistics

In this section, we consider the practical scenario where the gradient statistics $\alpha(t)$ and $\beta(t)$ are unknown. We propose a method to estimate $\alpha(t)$ and $\beta(t)$ in each time block and then devise an adaptive power control scheme based on the optimal solution of problem $\mathcal{P}_1$ by using the estimated $\alpha(t)$ and $\beta(t)$.

A. Parameters Estimation

In this subsection, we propose a method to estimate $\alpha(t)$ and $\beta(t)$ at each time block t directly based on their definitions in (9) and (10), respectively.

1) Estimation of $\alpha(t)$: Note that the instantaneous gradient norm of each device, $B_k(t)$, is assumed to be available at the edge server with negligible cost. By definition (9), we can estimate the gradient MSN as

$$\hat{\alpha}(t) = \frac{1}{K} \sum_{k \in \mathcal{K}} B_k^2(t). \tag{30}$$

2) Estimation of $\beta(t)$: By definition in (10), the gradient SMCV $\beta(t)$ depends on $m_d(t)$ and $\sigma_d(t)$. Before each device sending its gradient at time block t, we cannot estimate $\beta(t)$ in advance. However, from the experimental results of real datasets in Figs. 2-4, it can be observed that $\beta(t)$ changes slowly over iterations t. Thus we propose to estimate $\beta(t)$ based on the aggregated gradient at time block $t-1$ as below

$$\hat{\beta}(t) = \frac{\hat{\alpha}(t-1) - \sum_{d=1}^{D} \hat{g}_d^2(t-1)}{\sum_{d=1}^{D} \hat{g}_d^2(t-1)}, \tag{31}$$

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 $\beta(t)$ based on the received gradient norm at time block t and historical aggregated gradient (line 9 and line 16). Then the optimal transmit power and denoising factor are obtained based on (21) and (22), respectively (line 10). Third, the edge server informs the optimal transmit power 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).

Algorithm 1 FL Process with Adaptive Power Control

Initialize $\boldsymbol{w}(0)$ in edge server, $\hat{\beta}(1)$;
for time block $t = 1, ..., T$ **do**
  Edge server broadcasts the global model $\boldsymbol{w}(t)$ to all edge devices $k \in \mathcal{K}$;
  for each device $k \in \mathcal{K}$ in parallel **do**
    $\boldsymbol{g}_k(t) = \nabla L_{k,t}^{SGD}(\boldsymbol{w}(t))$;
    $B_k(t) = \sqrt{\sum_d g_{k,d}^2(t)}$;
    Upload $B_k(t)$ to edge server;
  end for
  Edge server estimates $\hat{\alpha}(t)$ based on (30);
  Edge server obtains the optimal transmit power $\boldsymbol{p}^*(t)$ based on (21) and the optimal denoising factor $\eta^*(t)$ based on (22);
  Edge server sends $p_k^*(t)$ to device k for all $k \in \mathcal{K}$;
  for each device $k \in \mathcal{K}$ in parallel **do**
    Transmit gradient $\boldsymbol{g}_k(t)$ with power $p_k^*(t)$ to edge server using AirComp;
  end for
  Edge server receives $\boldsymbol{y}(t)$ and recovers $\hat{\boldsymbol{g}}(t)$ based on (7);
  Edge server estimates $\hat{\beta}(t+1)$ based on (31);
  Edge server updates global model $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \gamma\hat{\boldsymbol{g}}(t)$;
end for
Edge server returns $\boldsymbol{w}(T+1)$;

V. Experimental Results

In this section, we provide experimental results to validate the performance of the proposed power control for AirComp-based FL over fading channels.

A. *Experiment Setup*

We conducted experiments on a simulated environment where the number of edge devices is $K = 10$ if not specified otherwise. 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 [17]:* this is the power control scheme given in [17], 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 misalignment error only in problem $\mathcal{P}_1$.
- *Full power transmission:* all devices transmit with full power P_k and the edge server applies the optimal denoising factor in (22), where $l^* = K$.

2) Datasets: We evaluate the training of convolutional neural network on three datasets: MNIST, CIFAR-10 and SVHN. MNIST dataset consists of 10 categories ranging from digit 0 to 9 and a total of 70000 labeled data samples (60000 for training and 10000 for testing). CIFAR-10 dataset includes 60000 color images (50000 for training and 10000 for testing) of 10 different types of objects. SVHN is a real-world image dataset for developing machine learning and object recognition algorithms with minimal requirement on data preprocessing and formatting, which includes 99289 labeled data samples (73257 for training and 26032 for testing).

3) Data Distribution: To study the impact of the SMCV of gradient β for optimal transmit power, we simulate two types of dataset 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 equal shards, and randomly assign 2 shards to each device.

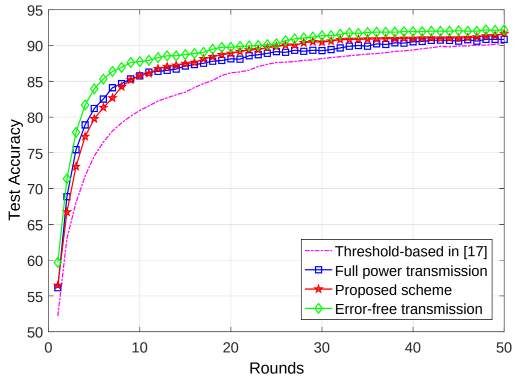
4) Training and Control Parameters: In all our experiments, the number of local update steps between two global aggregations is 1. Local batch size of each edge device is 10. The gradient descent step size is $\gamma = 0.01$.

B. Experimental Results

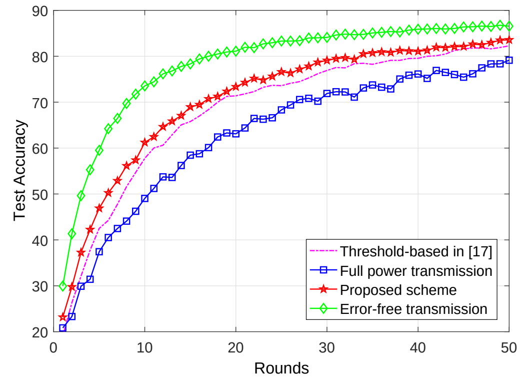
Fig. 6 compares the test accuracy for the three considered datasets with IID dataset partition and non-IID dataset partition, respectively, where the average received SNR is set as 10dB for all devices. It is observed that the model accuracy of the proposed scheme is better than threshold-based power control in [17] and full power transmission. From Figs. 2-4, we know that the averaged gradient SMCV $\beta(t)$ in the IID dataset partition is less than that in the non-IID dataset partition and gradient SMCV $\beta(t)$ increases over iterations. Threshold-based power control in [17] has significant accuracy loss in the IID partition or at the beginning of training. 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, threshold-based power control in [17] scheme that considers the individual misalignment error only is much inferior. Besides, full power transmission has significant accuracy loss in the non-IID partition or at the end of training. 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. 7 illustrates the test accuracy for MNIST with non-IID data partition at the average received SNR $=$ 5db. It is observed that the overall performance of the proposed scheme is still better than two baseline approaches at low SNR region. In specific, full power transmission performs better than threshold-based power control in [17] scheme. This is mainly because when the noise variance is large, full power transmission can strongly suppress noise error that dominates the MSE.

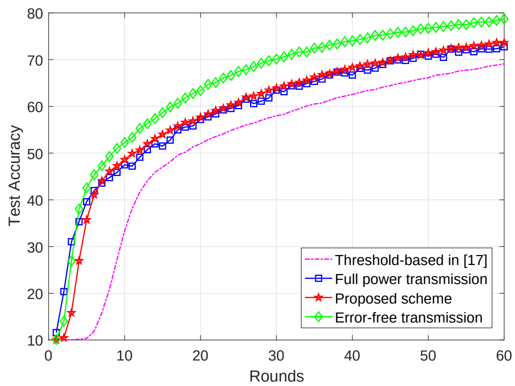
Finally, Fig. 8 compares the test accuracy of different power control schemes at varying number of devices K. Here, MNIST dataset with non-IID partition is used, the average received SNR of all the devices is set as $\text{SNR}_k = 10$dB and the results are averaged over 50 model trainings. First, it is observed that the test accuracy achieved by all the four schemes increases as K increases, due to the fact that the edge server can aggregate more data for averaging. Second, the proposed scheme considerably outperforms both of threshold-based power control in [17] and full power transmission throughout the whole regime of K. Full power transmission approaches threshold-based power control in [17] when K is small (i.e., $K = 4$ in Fig. 8), but



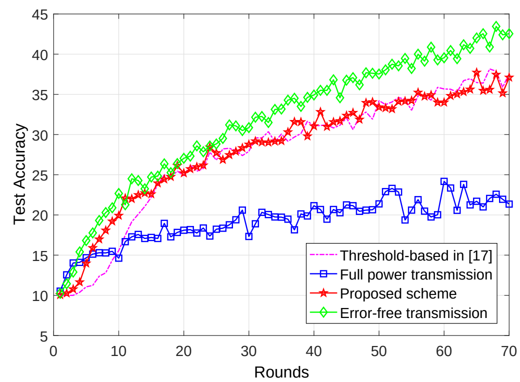
(a) MNIST dataset with IID partition.



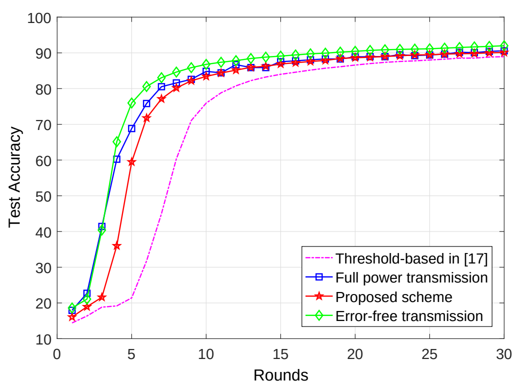
(b) MNIST dataset with non-IID partition.



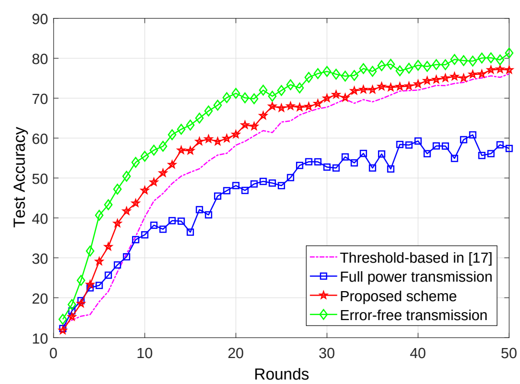
(c) CIFAR-10 dataset with IID partition.



(d) CIFAR-10 dataset with non-IID partition.



(e) SVHN dataset with IID partition.



(f) SVHN dataset with non-IID partition.

Fig. 6. Performance comparison on different dataset partition, for $K = 10$ and $\text{SNR}_k = 10\text{dB}, \forall k \in \mathcal{K}$.

the performance compromises as K increases, due to the lack of power adaptation to reduce the misalignment error.

VI. Conclusion

This work studied the power control optimization problem for the over-the-air federated learning over fading channels by taking the gradient statistics into account. The optimal power

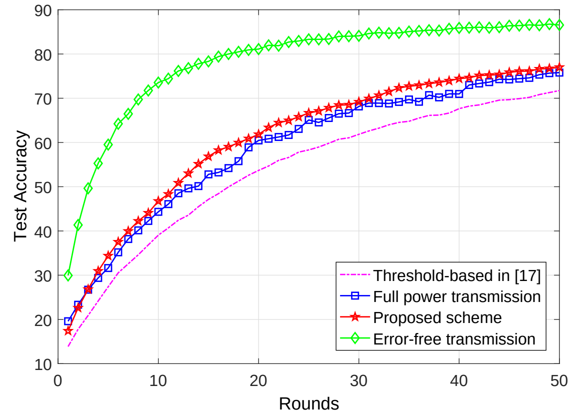


Fig. 7. Performance comparison for MNIST dataset with non-IID partition at the average received SNR $= 5$db.

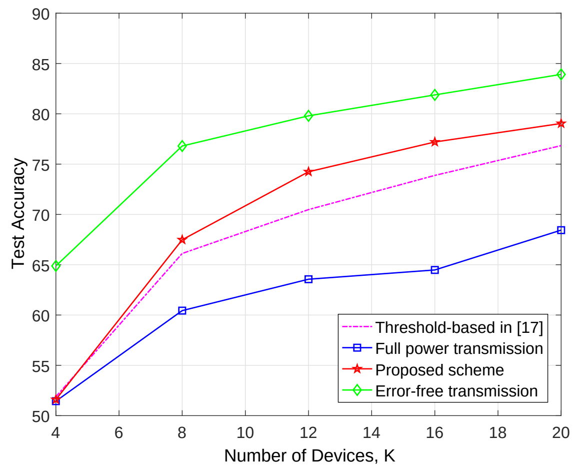


Fig. 8. Performance comparison over the number of devices, where MNIST dataset is non-IID partition and $\text{SNR}_k = 10\text{dB}, \forall k \in \mathcal{K}$.

control policy is derived in closed form when the first- and second-order gradient statistics are known. It is shown that the optimal transmit power on each device decreases with gradient SMCV and increases with noise variance. In the special cases where β approaches infinity and zero, the optimal transmit power reduces to threshold-based power control in [17] and full power transmission, respectively. We propose an adaptive power control algorithm that dynamically adjusts the transmit power in each iteration based on the estimation results. Experimental results show that our proposed adaptive power control scheme outperforms the existing schemes.

Appendix A
Proof of Lemma 1

We prove this Lemma by contradiction. Suppose the optimal denoising factor $\eta^* \leq C_1^2$. Both the individual and composite signal misalignment errors can be forced to zero by letting $p_k^* = \frac{\eta B_k^2}{|h_k|^2}, \forall k \in \mathcal{K}$. The problem $\mathcal{P}_1$ can thus be expressed as

$$\min_{\eta \leq C_1^2} \frac{D\sigma_n^2}{\eta}. \tag{32}$$

It is obvious that the optimal solution of the above problem is $\eta^* = C_1^2$. Thus, it must hold that $\eta \geq C_1^2$ for problem $\mathcal{P}_1$.

Appendix B
Proof of Lemma 2

To prove this lemma, we need to prove that if $p_{k_2}^* = P_{k_2}$ for some device k_2, we shall have $p_{k_1}^* = P_{k_1}, \forall k_1 < k_2$. We prove this by contradiction. Let $\boldsymbol{p}^* = [p_1^*, ..., p_K^*]$ denote the optimal transmit power to the problem $\mathcal{P}_1$. We assume that there are two devices $k_1 < k_2$ satisfying $p_{k_1}^* < P_{k_1}$ and $p_{k_2}^* = P_{k_2}$. Then there always exists a modified transmit power $\boldsymbol{p}^{'} = [p_1^*, ..., p_{k_1-1}^*, p_{k_1}^{'}, p_{k_1+1}^*, ..., p_{k_2-1}^*, p_{k_2}^{'}, p_{k_2+1}^*, ..., p_K^*]$, where $p_{k_1}^{'} = P_{k_1}$ and $p_{k_2}^{'} < P_{k_2}$, satisfying $\frac{\sqrt{p_{k_1}^*}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}} + \frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} = \frac{\sqrt{p_{k_1}^{'}}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}} + \frac{\sqrt{p_{k_2}^{'}}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}}$. The resulting MSE obtained by $\boldsymbol{p}^{'}$ only differs from the minimum MSE by $\boldsymbol{p}^*$ in the individual misalignment error term. The difference is given by

$$\begin{aligned}
&\mathrm{MSE}(\boldsymbol{p}^*) - \mathrm{MSE}(\boldsymbol{p}^{'})\\
=&\frac{\beta\alpha}{\beta+1}\left[\left(\frac{\sqrt{p_{k_1}^*}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}} - 1\right)^2 + \left(\frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} - 1\right)^2\right.\\
&\left. - \left(\frac{\sqrt{p_{k_1}^{'}}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}} - 1\right)^2 - \left(\frac{\sqrt{p_{k_2}^{'}}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} - 1\right)^2\right] &\text{(33a)}\\
=&\frac{\beta\alpha}{\beta+1}\left[\left(\frac{\sqrt{p_{k_1}^*}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}}\right)^2 + \left(\frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}}\right)^2 - \left(\frac{\sqrt{p_{k_1}^{'}}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}}\right)^2 - \left(\frac{\sqrt{p_{k_2}^{'}}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}}\right)^2\right] &\text{(33b)}\\
=&\frac{\beta\alpha}{\beta+1}\left(\frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} - \frac{\sqrt{p_{k_2}^{'}}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}}\right)\left(\frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} + \frac{\sqrt{p_{k_2}^{'}}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} - \frac{\sqrt{p_{k_1}^*}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}} - \frac{\sqrt{p_{k_1}^{'}}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}}\right) &\text{(33c)}\\
=&\frac{2\beta\alpha}{\beta+1}\left(\frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} - \frac{\sqrt{p_{k_2}^{'}}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}}\right)\left(\frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} - \frac{\sqrt{p_{k_1}^{'}}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}}\right) &\text{(33d)}\\
\geq&0. &\text{(33e)}
\end{aligned}$$

The inequality in (33e) holds as $p_{k_2}^* = P_{k_2} > p_{k_2}^{'}$ and $\frac{\sqrt{p_{k_2}^*}|h_{k_2}|}{\sqrt{\eta^*}B_{k_2}} \geq \frac{\sqrt{p_{k_1}^{'}}|h_{k_1}|}{\sqrt{\eta^*}B_{k_1}}$ by the aggregation capability ranking in (13). This indicates that $\boldsymbol{p}^{'}$ is a better solution than $\boldsymbol{p}^*$, which contradicts the assumption. Therefore, for all pairs of two devices $k_1 < k_2$, if $p_{k_2}^* = P_{k_2}$, we must have $p_{k_1}^* = P_{k_1}$. Lemma 2 is proved.

Appendix C
Proof of Lemma 3

To prove this lemma is equal to proving that if the global optimal solution p^* of problem $\mathcal{P}_1$ lies in subregion $\mathcal{M}_l$, the optimal solution $\tilde{\boldsymbol{p}}_l^*$ of the sub-problem defined in the relaxed subregion $\tilde{\mathcal{M}}_l$ should also lie $\mathcal{M}_l$. We prove this by contradiction. Assume that $\boldsymbol{p}^* \in \mathcal{M}_l$ but $\tilde{\boldsymbol{p}}_l^* \notin \mathcal{M}_l$. The derivative of (14a) w.r.t. p_k and η shows that $\tilde{\boldsymbol{p}}_l^*$ is the only local optimal solution in $\tilde{\mathcal{M}}_l$. As $\boldsymbol{p}^*$ is the optimal solution in $\mathcal{M}_l$ and cannot be on the boundary of $\mathcal{M}_l$, $\boldsymbol{p}^*$ is also a local optimal solution in $\mathcal{M}_l \subseteq \tilde{\mathcal{M}}_l$. It contradicts that $\tilde{\boldsymbol{p}}_l^*$ is the only local optimal solution in $\tilde{\mathcal{M}}_l$. Therefore, if the global optimal transmit power $\boldsymbol{p}^*$ is in the l-th subregion $\mathcal{M}_l$, $\tilde{\boldsymbol{p}}_l^*$ must be in $\mathcal{M}_l$. The proof of Lemma 3 is completed.

APPENDIX D
PROOF OF THEOREM 1

To complete the proof, we need to show that with l^* defined in (20), the optimal transmit power is $\tilde{\boldsymbol{p}}_{l^*}^*$ in the l^*-th relaxed subregion $\tilde{\mathcal{M}}_{l^*}$. By Lemma 3, for all l in $\mathcal{L}$, since $\tilde{\boldsymbol{p}}_l^*$ is also in subregion $\mathcal{M}_l$, $\tilde{\boldsymbol{p}}_l^*$ is the optimal transmit power in the subregion $\mathcal{M}_l$. Therefore, the candidate transmit power $\tilde{\boldsymbol{p}}_{l^*}^*$ with the smallest value V_{l^*} is the optimal transmit power of the problem $\mathcal{P}_1$. According to (17) and (19), the optimal transmit power and denoising factor are the forms given in Theorem 1.

APPENDIX E
PROOF OF LEMMA 4

When $\boldsymbol{p}^*$ lies in subregion $\mathcal{M}_l$, $\boldsymbol{p}^*$ is equal to $\tilde{\boldsymbol{p}}_l^*$. Thus, to prove the sufficiency of this Lemma is equal to prove that the optimal transmit power $\boldsymbol{p}^*$ holds

$$C_{l^*} \leq \frac{\sqrt{p_k^*}|h_k|}{B_k} < C_{l^*+1}, \forall k \in \{l^*+1, ..., K\}. \tag{34}$$

Based on (21), $p_k^* < P_k$, for $k \in \{l^*+1, ..., K\}$, and when $k \in \{l^*+1\}$, we have $\frac{\sqrt{p_{l^*+1}^*}|h_{l^*+1}|}{B_{l^*+1}} < C_{l^*+1}$. Since $\frac{\sqrt{p_k^*}|h_k|}{B_k}, \forall k \in \{l^*+1, ..., K\}$ is the same, we have the inequality $\frac{\sqrt{p_k^*}|h_k|}{B_k} < C_{l^*+1}$, for $k \in \{l^*+1, ..., K\}$. We prove $C_{l^*} \leq \frac{\sqrt{p_k^*}|h_k|}{B_k}, \forall k \in \{l^*+1, ..., K\}$ by contradiction. We assume that $C_{l^*} > \frac{\sqrt{p_{l^*+1}^*}|h_{l^*+1}|}{B_{l^*+1}}$. As $p_k^* = P_k$, for $k \in \{1, ..., l^*\}$, we have $\frac{\sqrt{p_{l^*}^*}|h_{l^*}|}{B_{l^*}} = C_{l^*} > \frac{\sqrt{p_{l^*+1}^*}|h_{l^*+1}|}{B_{l^*+1}}$. Then there always exists a modified transmit power $\boldsymbol{p}' = [p_1^*, ..., p_{l^*-1}^*, p_{l^*}', p_{l^*+1}', p_{l^*+2}^*, ..., p_K^*]$ where transmit power p_{l^*}' and p_{l^*+1}' satisfy $\frac{\sqrt{p_{l^*}'}|h_{l^*}|}{B_{l^*}} = \frac{\sqrt{p_{l^*+1}'}|h_{l^*+1}|}{B_{l^*+1}} = \frac{1}{2}(\frac{\sqrt{p_{l^*}^*}|h_{l^*}|}{B_{l^*}} + \frac{\sqrt{p_{l^*+1}^*}|h_{l^*+1}|}{B_{l^*+1}})$. The difference between MSE of the transmit power $\boldsymbol{p}^*$ and $\boldsymbol{p}'$ is given by

$$\begin{aligned} &\mathrm{MSE}(\boldsymbol{p}^*) - \mathrm{MSE}(\boldsymbol{p}') \\ =&\frac{\beta\alpha}{\beta+1}\left[\left(\frac{\sqrt{p_{l^*}^*}|h_{l^*}|}{\sqrt{\eta^*}B_{l^*}} - 1\right)^2 + \left(\frac{\sqrt{p_{l^*+1}^*}|h_{l^*+1}|}{\sqrt{\eta^*}B_{l^*+1}} - 1\right)^2\right. \\ &\left. - \left(\frac{\sqrt{p_{l^*}'}|h_{l^*}|}{\sqrt{\eta^*}B_{l^*}} - 1\right)^2 - \left(\frac{\sqrt{p_{l^*+1}'}|h_{l^*+1}|}{\sqrt{\eta^*}B_{l^*+1}} - 1\right)^2\right] \end{aligned} \tag{35a}$$

$$=\frac{\beta\alpha}{\beta+1}\left[\left(\frac{\sqrt{p_{l^*}^*}|h_{l^*}|}{\sqrt{\eta^*}B_{l^*}}\right)^2+\left(\frac{\sqrt{p_{l^*+1}^*}|h_{l^*+1}|}{\sqrt{\eta^*}B_{l^*+1}}\right)^2-\left(\frac{\sqrt{p_{l^*}^{'}}|h_{l^*}|}{\sqrt{\eta^*}B_{l^*}}\right)^2-\left(\frac{\sqrt{p_{l^*+1}^{'}}|h_{l^*+1}|}{\sqrt{\eta^*}B_{l^*+1}}\right)^2\right] \tag{35b}$$

$$=\frac{\beta\alpha}{2\left(\beta+1\right)}\left[\left(\frac{\sqrt{p_{l^*}^*}|h_{l^*}|}{\sqrt{\eta^*}B_{l^*}}-\frac{\sqrt{p_{l^*+1}^*}|h_{l^*+1}|}{\sqrt{\eta^*}B_{l^*+1}}\right)^2\right]>0. \tag{35c}$$

This indicates that $\boldsymbol{p}^{'}$ is a better solution than $\boldsymbol{p}^*$, which contradicts the assumption. We have the inequality $C_{l^*} \leq \frac{\sqrt{p_k^*}|h_k|}{B_k}$, for $k \in \{l^*+1, ..., K\}$. Thus, the sufficiency of this Lemma has been proved.

To prove the necessity of this Lemma, we prove the inverse negative proposition of it: if the optimal transmit power $\boldsymbol{p}^*$ does not lie in subregion $\mathcal{M}_l$, i.e., $l \in \{1, ..., l^*-1, l^*+1, ..., K\}$, $\tilde{\boldsymbol{p}}_l^*$ satisfies: $\frac{\sqrt{\tilde{p}_{l,k}^*}|h_k|}{B_k} \geq C_{l+1}$ or $\frac{\sqrt{\tilde{p}_{l,k}^*}|h_k|}{B_k} < C_l$ for all $k \in \{l+1, ..., K\}$.

First, we prove that $\tilde{\boldsymbol{p}}_l^*$, for $\forall l \in \{1, ..., l^*-1\}$ holds

$$\frac{\sqrt{\tilde{p}_{l,k}^*}|h_k|}{B_k} \geq C_{l+1}, \forall k \in \{l+1, ..., K\}. \tag{36}$$

We prove this by contradiction. We assume that $\exists l \in \{1, ..., l^*-1\}$, $\frac{\sqrt{\tilde{p}_{l,k}^*}|h_k|}{B_k} < C_{l+1}, \forall k \in \{l+1, ..., K\}$. As $C_{l+1} \leq C_k$, for $\forall k \in \{l+1, ..., K\}$ given in (13), $\frac{\sqrt{\tilde{p}_{l,k}^*}|h_k|}{B_k} < C_k$, for $\forall k \in \{l+1, ..., K\}$, $\tilde{p}_{l,k}^* < P_k$, for $\forall k \in \{l+1, ..., K\}$, i.e., $\tilde{\boldsymbol{p}}_l^*$ is a feasible transmit power. Note that $\tilde{\boldsymbol{p}}_l^*$ is the optimal transmit power in $\tilde{\mathcal{M}}_l$ and $\boldsymbol{p}^*$ is the optimal transmit power in $\mathcal{M}_{l^*}$. As the constraint of $\tilde{\boldsymbol{p}}_l^*$ is less restrictive than the constraint of $\boldsymbol{p}^*$, i.e., $\mathcal{M}_{l^*} \subseteq \tilde{\mathcal{M}}_l$, the feasible transmit power $\tilde{\boldsymbol{p}}_l^*$ is better than the optimal transmit power $\boldsymbol{p}^*$, which contradicts the assumption. We have the inequality (36).

Second, we prove that $\tilde{\boldsymbol{p}}_l^*$, for $\forall l \in \{l^*+1, ..., K\}$ holds

$$\frac{\sqrt{\tilde{p}_{l,k}^*}|h_k|}{B_k} < C_l, \forall k \in \{l+1, ..., K\}. \tag{37}$$

We prove this by contradiction. When $l = l^*+1$, we assume that $\frac{\sqrt{\tilde{p}_{l,k}^*}|h_k|}{B_k} \geq C_l, \forall k \in \{l+1, ..., K\}$. Let $\boldsymbol{p}^{avg} = [\tilde{p}_{l,1}^*, ..., \tilde{p}_{l,l-1}^*, p_l^{avg}, ..., p_K^{avg}]$ denote a modified $\tilde{\boldsymbol{p}}_l^*$, where $\frac{\sqrt{p_l^{avg}}|h_l|}{B_l} = ... = \frac{\sqrt{p_K^{avg}}|h_K|}{B_K} = \frac{1}{K-l+1}\sum_{i=l}^{K}\frac{\sqrt{\tilde{p}_{l,i}^*}|h_i|}{B_i}$. Using a proof similar to that of (35c), we can prove that $\text{MSE}(\boldsymbol{p}^{avg}) \leq \text{MSE}(\tilde{\boldsymbol{p}}_l^*)$. Let $\boldsymbol{p}^{fes} = [\tilde{p}_{l,1}^*, ..., \tilde{p}_{l,l-1}^*, p_l^{fes}, ..., p_K^{fes}]$ denote a modified $\tilde{\boldsymbol{p}}_l^*$, where $\frac{\sqrt{p_k^{fes}}|h_k|}{B_k} = C_l,$, for $k \in \{l, ..., K\}$. The derivative of (14a) w.r.t. p_k for all $k \in \{l+1, ..., K\}$ shows that the MSE increases with p_k when $p_k \geq p_k^*$. Note that $\boldsymbol{p}^{avg}$ and $\boldsymbol{p}^{fes}$ are in the l^*-th relaxed subregion $\tilde{\mathcal{M}}_{l^*}$ and $p_k^{avg} \geq p_k^{fes} > p_k^*$, for $k \in \{l, ..., K\}$, $\text{MSE}(\boldsymbol{p}^{fes}) \leq \text{MSE}(\boldsymbol{p}^{avg})$ and $\text{MSE}(\boldsymbol{p}^{fes}) \leq \text{MSE}(\tilde{\boldsymbol{p}}_l^*)$. The transmit power $\boldsymbol{p}^{fes} \in \tilde{\mathcal{M}}_l$ is better than the optimal transmit power $\tilde{\boldsymbol{p}}_l^* \in \tilde{\mathcal{M}}_l$, which contradicts that $\tilde{\boldsymbol{p}}_l^*$ is optimal solution in $\tilde{\mathcal{M}}_l$. We have the inequality (37) when $l = l^*+1$ and it can be extended to when $l \in \{l^*+2, ..., K\}$ by mathematical induction.

Thus, the necessity of this Lemma has been proved. We complete the proof of Lemma 4.

Appendix F
Proof of Lemma 5

We first prove that the upper boundary $\mathcal{U}_l$ is equal to the lower boundary $\mathcal{L}_{l+1}$, for $l \in \{1, ..., K-1\}$. Then we prove that the optimal transmit power function $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is continuous at $\mathcal{U}_l = \mathcal{L}_{l+1}$, for $l \in \{1, ..., K-1\}$.

For all $(\beta, \sigma_n^2) \in \mathcal{L}_{l+1}$, the corresponding optimal transmit power $\tilde{\boldsymbol{p}}^*_{l+1}(\beta, \sigma_n^2) = (\frac{C_{l+1}B_k}{|h_k|})^2$, i.e., $\frac{\sqrt{\tilde{p}^*_{l+1,k}(\beta,\sigma_n^2)}|h_k|}{B_k} = C_{l+1}$, for $k \in \{l+2, ..., K\}$. Based on (17) and (19), we have

$$\sqrt{\tilde{\eta}^*_{l+1}(\beta, \sigma_n^2)} = \frac{\sum_{i=1}^{l+1} C_i + (\beta+K-l-1)C_{l+1}}{\beta+K} \tag{38a}$$

$$\Leftrightarrow \quad \frac{\frac{\beta\alpha}{\beta+1}\sum\limits_{i=1}^{l+1} C_i^2 + \frac{\beta\alpha}{(\beta+K-l-1)(\beta+1)}\left(\sum\limits_{i=1}^{l+1} C_i\right)^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l-1)(\beta+1)}\sum\limits_{i=1}^{l+1} C_i} = \frac{\sum_{i=1}^{l+1} C_i + (\beta+K-l-1)C_{l+1}}{\beta+K} \tag{38b}$$

$$\Leftrightarrow \quad \frac{\frac{\beta\alpha}{\beta+1}\sum\limits_{i=1}^{l+1} C_i^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l-1)(\beta+1)}\sum\limits_{i=1}^{l+1} C_i} = \frac{(\beta+K-l-1)C_{l+1}}{\beta+K} \tag{38c}$$

$$\Leftrightarrow \quad \frac{\beta\alpha}{\beta+1}\sum_{i=1}^{l+1} C_i^2 + D\sigma_n^2 = \frac{\beta\alpha}{\beta+1}C_{l+1}\sum_{i=1}^{l+1} C_i \tag{38d}$$

$$\Leftrightarrow \quad \frac{\beta\alpha}{\beta+1}\sum_{i=1}^{l} C_i^2 + D\sigma_n^2 = \frac{\beta\alpha}{\beta+1}C_{l+1}\sum_{i=1}^{l} C_i \tag{38e}$$

$$\Leftrightarrow \quad \frac{\frac{\beta\alpha}{\beta+1}\sum\limits_{i=1}^{l} C_i^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l)(\beta+1)}\sum\limits_{i=1}^{l} C_i} = \frac{(\beta+K-l)C_{l+1}}{\beta+K} \tag{38f}$$

$$\Leftrightarrow \quad \frac{\frac{\beta\alpha}{\beta+1}\sum\limits_{i=1}^{l} C_i^2 + \frac{\beta\alpha}{(\beta+K-l)(\beta+1)}\left(\sum\limits_{i=1}^{l} C_i\right)^2 + D\sigma_n^2}{\frac{\beta(\beta+K)\alpha}{(\beta+K-l)(\beta+1)}\sum\limits_{i=1}^{l} C_i} = \frac{\sum_{i=1}^{l} C_i + (\beta+K-l)C_{l+1}}{\beta+K} \tag{38g}$$

$$\Leftrightarrow \quad \sqrt{\tilde{\eta}^*_l(\sigma_n^2, \beta)} = \frac{\sum_{i=1}^{l} C_i + (\beta+K-l)C_{l+1}}{\beta+K} \tag{38h}$$

$$\Leftrightarrow \quad \frac{\sqrt{\tilde{p}^*_{l,k}(\beta, \sigma_n^2)}|h_k|}{B_k} = C_{l+1}, \forall k \in \{l+1, ..., K\} \tag{38i}$$

$$\Leftrightarrow \quad \tilde{p}^*_{l,k}(\beta, \sigma_n^2) = (\frac{C_{l+1}B_k}{|h_k|})^2, \forall k \in \{l+1, ..., K\}, \tag{38j}$$

then we have $(\sigma_n^2, \beta) \in \mathcal{U}_l$ by the definition of $\mathcal{U}_l$, i.e., $\mathcal{L}_{l+1} \subseteq \mathcal{U}_l$. For all $(\beta, \sigma_n^2) \in \mathcal{U}_l$, we have $(\sigma_n^2, \beta) \in \mathcal{L}_{l+1}$, i.e., $\mathcal{U}_l \subseteq \mathcal{L}_{l+1}$ by reversing the derivation of (38). Therefore, we have $\mathcal{L}_{l+1} = \mathcal{U}_l$, for $l \in \{1, ..., K-1\}$.

For all $(\beta, \sigma_n^2)_0 \in \mathcal{L}_{l+1} = \mathcal{U}_l$, define $(\beta, \sigma_n^2)_+ \in \mathcal{X}_{l+1}$ and $(\beta, \sigma_n^2)_- \in \mathcal{X}_l$ as two points infinitely close to $(\beta, \sigma_n^2)_0$, respectively. Then the left limitation of $\boldsymbol{p}^*((\beta, \sigma_n^2)_0)$ is given by

$$\begin{aligned} \lim_{(\beta,\sigma_n^2)_- \to (\beta,\sigma_n^2)_0} \boldsymbol{p}^*((\beta, \sigma_n^2)_-) &= \lim_{(\beta,\sigma_n^2)_- \to (\beta,\sigma_n^2)_0} \tilde{\boldsymbol{p}}_l^*((\beta, \sigma_n^2)_-) \\ &= \tilde{\boldsymbol{p}}_l^*((\beta, \sigma_n^2)_0) = \tilde{\boldsymbol{p}}_{l+1}^*((\beta, \sigma_n^2)_0) = \boldsymbol{p}^*((\beta, \sigma_n^2)_0). \end{aligned} \tag{39}$$

Similarly, the right limitation of $\boldsymbol{p}^*((\beta, \sigma_n^2)_0)$ is also equal to $\boldsymbol{p}^*((\beta, \sigma_n^2)_0)$. Therefore, the optimal transmit power function $\boldsymbol{p}^*(\beta, \sigma_n^2)$ is continuous at $\mathcal{U}_l = \mathcal{L}_{l+1}$, for $l \in \{1, ..., K-1\}$. We complete the proof of Lemma 5.

Appendix G
Proof of Theorem 3

For any denoising factor $\eta \geq \frac{1}{K^2}\Big(\sum_{k\in\mathcal{K}} C_k\Big)^2$, it must hold that composite signal alignment $\sum_{k\in\mathcal{K}} G_k(p_k, \eta) \leq K$ for problem $\mathcal{P}_3$. Therefore, for minimizing composite signal misalignment error $\Bigg(\sum_{k\in\mathcal{K}} G_k(p_k, \eta) - K\Bigg)^2$, all the devices should always transmit with full power, i.e., $p_k^* = P_k, \forall k \in \mathcal{K}$. The problem $\mathcal{P}_3$ can be expressed as

$$\min_{\eta\geq 0} \alpha \left(\sum_{k\in\mathcal{K}} G_k(p_k, \eta) - K \right)^2 + \frac{D\sigma_n^2}{\eta} \tag{40}$$

This is a unary quadratic function about $\frac{1}{\sqrt{\eta}}$, thus it is easy to derive an optimal solution, i.e.,

$$\eta^* = \left(\frac{\alpha\big(\sum_{i\in\mathcal{K}} C_i\big)^2 + D\sigma_n^2}{\alpha K \sum_{i\in\mathcal{K}} C_i} \right)^2. \tag{41}$$

We complete the proof of Theorem 3.

References

[1] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning in fading channels," in *Proc. IEEE ICC Workshops*, 2020, pp. 1–6.

[2] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

[3] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 440–445.

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

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

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

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

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

[15] M. M. Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," in *Proc. IEEE ISIT*, July 2019, pp. 1432–1436.

[16] ——, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2020.

[17] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimal power control for over-the-air computation," in *Proc. IEEE GLOBECOM*, Dec. 2019, pp. 1–6.