Federated Learning With Controlled Descent Under Fading: Convergence and Energy Implications

Sayantan Adhikary, *Graduate Student Member, IEEE*, and Neelesh B. Mehta, *Fellow, IEEE*

***Abstract*—In over-the-air computation-assisted federated learning (OTA-FL), devices transmit their local models to a parameter server over a shared time-frequency resource. Model aggregation occurs automatically due to the superposition property of the wireless channel. We derive a novel upper bound on the convergence of the optimality gap of OTA-FL that applies to any choice of device transmit powers and receiver scaling. The bound is based on less restrictive assumptions compared to the literature. It leads to the insightful concept of an effective learning rate that captures the dependence of the convergence of OTA-FL on the gains of the channels between the devices and the parameter server. We jointly optimize the transmit powers and the receiver scaling to minimize the error floor implied by the bound while controlling the effective learning rate. This leads to a novel controlled descent algorithm (CDA) and a new variant that adapts the effective learning rate. CDA can be implemented using a low overhead protocol. We benchmark CDA against several transmit power, receiver scaling, and device selection schemes. For both linear regression and multi-class logistic regression, CDA requires fewer iterations and a lower sum energy to achieve a target optimality gap or testing accuracy.**

***Index Terms*—Federated learning, over-the-air computation, convergence, effective learning rate, energy consumption, channel fading, protocol.**

I. Introduction

FEDERATED learning (FL) is a distributed technique that allows multiple devices to collaboratively train a shared or global machine learning model. In this iterative technique, each device computes a local model based on the current global model and its local dataset, which comprises of data and their corresponding labels, and shares this local model with a cloud-based parameter server. The server aggregates the local models it receives, updates the global model, and broadcasts it to all devices. In the classical federated stochastic gradient descent (FedSGD) algorithm, the local model of a device is the local gradient computed over a batch of data-points uniformly selected from the local dataset and the global model is the arithmetic average of the local models [2]. The exchange of models instead of local data improves privacy. However, the latency and the communication overhead of model aggregation increase due to separate uplink transmissions from the devices.

Over-the-air computation-assisted federated learning (OTA-FL) is an aggregation technique that exploits the superposition principle of the wireless channel. The transmissions occur over a shared time and frequency resource. Hence, the latency and bandwidth requirements are markedly lower [3], [4]. However, the channel gains between the devices and the parameter server, the device transmit powers, and the receiver scaling and noise at the parameter server together affect the algorithm's learning performance. For continuous labels, the learning performance is measured in terms of the optimality gap, which is the difference between the minimum value of the loss function if the data were centrally available at the parameter server, and the loss function of the current model computed over all the data points. For discrete labels, the learning performance is measured in terms of the testing accuracy of the global model.

The schemes in the literature control the device transmit powers and the receiver scaling in many different ways. We discuss them below.

- *Channel Inversion (CI) Without Peak Power Constraint [5], [6], [7], [8]:* In CI, the transmit power is proportional to the product of the receiver scaling and the local model power, which is the square of the ℓ_2-norm of the local model. In [5], [6], [7], and [8], the transmit power is set to meet a target signal-to-noise ratio (SNR) at the receiver. However, no peak or average power constraint is imposed. As a result, the devices can consume a large amount of energy, whose average can even be unbounded, when their channels are in a deep fade. In [9], the receiver scaling is inversely proportional to the maximum value of the local model power. However, channel fading is not considered in the theoretical development and all devices transmit with the same power. When fading is present, CI is employed and a device needs to perform sufficiently many local iterations to ensure that its transmit power stays below the peak power.
  In [5], [6], and [7], only a subset of the devices transmit. While this curtails the sum energy consumed in an iteration, it can lead to a poor generalization of the global

Received 21 June 2025; revised 20 November 2025; accepted 22 January 2026. Date of publication 2 February 2026; date of current version 10 February 2026. The work of Sayantan Adhikary was supported in part by the Qualcomm 6G UR Program and in part by the Prime Minister's Research Fellowship. The work of Neelesh B. Mehta was supported in part by the J. C. Bose Fellowship under Grant JCB/2023/000028. An earlier version of this paper was presented in part at the IEEE Global Communications Conference (GLOBECOM) in December 2023 [DOI: 10.1109/GLOBECOM54140.2023.10437140]. The associate editor coordinating the review of this article and approving it for publication was M. Giordani. *(Corresponding author: Sayantan Adhikary.)*

The authors are with the Department of Electrical Communication Engineering (ECE), Indian Institute of Science (IISc), Bengaluru 560012, India (e-mail: adhikarysayantan97@gmail.com; nbmehta@iisc.ac.in).

Digital Object Identifier 10.1109/TWC.2026.3658489

model. In [6], a device transmits a combination of its current and previous local models. In [8], devices not scheduled by the parameter server use a nearby relay to transmit.

- *Scaled-down CI (SCI) [10], [11], [12], [13], [14]:* The device transmit powers are subject to the peak power constraint in [10], [11], [12], [13], and [14]. In [10], the receiver scaling is proportional to the peak power and inversely proportional to the largest local model power among the devices. In [11], [12], [13], and [14], the devices transmit the direction of the local gradient and the receiver scaling is proportional to the smallest channel power gain among the devices. This approach is extended to incorporate multiple antennas at the parameter server in [13] and [14].
- *Truncated Channel Inversion (TCI) [3], [4], [15], [16]:* In [3], [4], and [15], the transmit power is similar to that in CI. However, a device does not transmit if the power gain of the channel between it and the parameter server falls below a pre-specified threshold. The receiver scaling is implicitly assumed to be unity in [3] and is proportional to the inverse of the exponential-integral of the threshold in [4]. However, as mentioned, this approach can lead to a poor generalization of the global model. In [16], all the selected devices in an iteration scale their local models by a common scalar that is chosen to minimize the gradient misalignment error (GME), which is the mean of the squared norm of the difference between the received aggregated model and the desired arithmetic average of the local models.
- *Clipped Channel Inversion (CCI) [17], [18], [19], [20], [21]:* In [17], [18], [19], and [20], the parameter server numerically determines the transmit powers, which are subject to the peak power constraint, and the receiver scaling to minimize the GME. It does so for every realization of the channel gains and informs each device about its transmit power. The optimal solution is such that a subset of the devices transmit at the peak power, while the rest use CI. In [21], the number of data-points that every device selects is adapted to minimize the GME.
- *Regularized Channel Inversion (RCI) [22], [23]:* In [22], the transmit powers in an iteration are numerically determined to minimize the error floor when the optimality gap has converged. We, therefore, refer to it as the minEF algorithm. Constraints are imposed on the peak power of each device and the average sum power consumed by the devices in each iteration. However, the receiver scaling is implicitly assumed to be unity. In minEF, a subset of the devices transmit at the peak power, while the transmit powers of the remaining devices are inversely proportional to their channel power gains plus a regularization constant. In [23], only a subset of the devices are selected based on their local model power and the channel power gain. However, the peak power constraint is not imposed. The learning rate is inversely proportional to the iteration count in [22].

Among the above schemes, [11], [16], [20] account for imperfect channel gains. The above schemes consume different amounts of communication energy to reach a specific optimality gap. This is because the sum energy consumed by the devices in each iteration depends on the scheme. The number of iterations required also depends on the scheme. In [24], the transmit powers and receiver scaling are determined numerically to minimize the ratio of the achievable Shannon rate for model aggregation to the sum power consumed per iteration.

The existing works either pre-suppose the transmit power and receiver scaling [3], [4], [5], [6], [7], [8], [10], [11], [12], [13], [14], [15], [16] or determine these variables to minimize the GME [17], [18], [19], [20], [21]. However, GME does not fully capture the impact of an error in an iteration on future iterations and the overall performance of the algorithm. We instead focus on minimizing the error floor. While minEF in [22] also minimizes an error floor, it assumes that the gradients of the local and global loss functions are upper bounded by a finite constant. However, in practice, this assumption is restrictive and requires choosing a conservatively large, and thus weak, upper bound in the presence of dataset outliers or large model dimensions [25]. Furthermore, the unbiased aggregation constraint imposed in [22] may not have a feasible solution. In such a case, minEF's behavior is undefined.

A. Contributions

We systematically develop a novel joint transmit power adaptation and receiver scaling scheme that guarantees the rate of convergence and minimizes the error floor in the presence of fading and noise. We make the following contributions:

- *Optimality Gap Bound:* We derive a recursive upper bound on the optimality gap that holds for any transmit power and receiver scaling scheme. The bound introduces the notion of the *effective learning rate*, which is analogous to the learning rate in classical gradient descent [26] but in the presence of channel fading. In hindsight, the effective learning rates of [5], [6], [7], [8], [10], [11], [12], [13], and [14] turn out be constants, while those of [3], [4], [15], [16], [17], [18], [19], [20], [21], [22], [23], and [24] are random as they depend on the channel gains. We identify the range of effective learning rates that guarantee convergence.

  The bound reveals a fundamental trade-off between the contraction factor and the residual error, which can be controlled by the choice of the effective learning rate. Unlike [5], [6], [11], [12], [16], [22], [23], our analysis does not require the ℓ_2-norm of the global or local gradients to be bounded. Furthermore, our work is based on the Polyak–Łojasiewicz (PL) inequality, which is more general than strong-convexity assumed in [5], [6], and [23]. Table I brings out the technical differences between our assumptions and those in the literature.
- *Problem Formulation:* Using the bound, we formulate a novel problem in which the transmit power adaptation and receiver scaling are jointly optimized to minimize the error floor while ensuring a target effective learning rate and satisfying the peak transmit power constraint at each device. This formulation ensures unbiased aggregation of the local models. The unbiased aggregation imposed in

TABLE I

KEY TECHNICAL DIFFERENCES IN OBJECTIVE AND ASSUMPTIONS BETWEEN THE LITERATURE AND OUR WORK. HERE, PL REFERS TO THE POLYAK–ŁOJASIEWICZ INEQUALITY

Works	Objective	Assumptions	
		Convexity	Bounded gradients
[3], [4]	Minimize latency of aggregation	-No analysis-	
[5], [6]	Analyze error floor	Strong convexity	Yes
[7]	Schedule devices to balance energy and heterogeneity	-No analysis-	
[8]	Relay to aid devices not selected	-No analysis-	
[10]	Optimize local learning rates to minimize noise	-No analysis-	
[11], [12]	Analyze global gradient	Non-convex	Yes
[13]	Maximize number of devices and target GME	-No analysis-	
[14]	Channel and model-based device scheduling	-No analysis-	
[15]	Analyze global gradient	Non-convex	No
[16]	Minimize GME	Non-convex	Yes
[17]–[21]	Minimize GME	No analysis of optimality gap	
[22]	Minimize error floor	PL	Yes
[23]	Minimize drift of global loss	Strong convexity	Yes
Proposed	Minimize error floor and control effective learning rate	PL	No

TABLE II

KEY DIFFERENCES BETWEEN THE EXISTING WORKS AND OUR APPROACH. HERE, K DENOTES THE NUMBER OF DEVICES

Works	Learning rate	Effective learning rate	Power control	Receiver scaling	Comm. overhead
[3], [4], [15], [16]	Fixed	Random	Presupposes TCI	Fixed	$\mathcal{O}(K)$
[5]–[8]	Fixed	Turns out constant	Presupposes CI	Adapted	$\mathcal{O}(K)$
[11]–[14]	Fixed	Turns out constant	Presupposes SCI	Adapted	$\mathcal{O}(K)$
[10]	Adapted	Turns out constant	Presupposes SCI	Adapted	$\mathcal{O}(K)$
[22], [23]	Fixed / adapted	Random	RCI	Fixed	$\mathcal{O}(K)$
[17]–[21]	Fixed	Random	Jointly adapted (CCI)		$\mathcal{O}(K)$
[24]	Fixed	Random	Jointly adapted		$\mathcal{O}(K)$
Proposed	Fixed	Controlled (Fixed / adapted)	Jointly adapted		$\mathcal{O}(1)$

minEF [22] is a special case of our constraint on the effective learning rate. Furthermore, a feasible solution always exists for our problem, unlike minEF.

- *Algorithm:* We present a novel controlled descent algorithm (CDA), which is based on insights we obtain about the structure of the optimal solution of the above problem and specifies the transmit powers and receiver scaling in closed-form. Compared to the OTA-FL literature, CDA shows that a device's transmit power is determined by a metric that is the ratio between its channel gain and the ℓ_2 norm of its local model. CDA's unique ability to control the effective learning rate leads to two variants, namely CDA-Fixed (CDA-F), which keeps the effective learning rate fixed, and CDA-Adaptive (CDA-A), which adapts the effective learning rate. This is unlike [22], [23] that adapt the learning rate, but whose effective learning rate is still random due to channel fading.
- *Implementation:* We propose a novel communication protocol to implement CDA that requires $\mathcal{O}(1)$ communication overhead to compute the transmit powers at the devices and receiver scaling at the parameter server. This is unlike the GME minimization schemes in [17], [18], [19], [20], and [21] and minEF that require $\mathcal{O}(K)$ overhead because the parameter server needs to inform each device about its transmit power.
- *Performance:* We extensively benchmark CDA-F and CDA-A with several schemes that use SCI, SCI with device selection, TCI, CCI, and RCI. We compare their performance for linear regression and multi-class logistic regression when the devices have heterogeneous local datasets. For multi-class logistic regression, we use the CIFAR-10 dataset on a multi-layer convolutional neural network (CNN) with a non-convex loss function. For both problems, CDA-A requires fewer iterations than all benchmark schemes to reach a target optimality gap or testing accuracy. For a given sum energy budget, it achieves a higher optimality gap or testing accuracy than all benchmark schemes. Furthermore, in linear regression and at low energy budgets, the optimality gap of CDA-F is as low as that of CDA-A.

Table II highlights the differences in the manner in which the transmit power and receiver scaling are controlled and the communication overhead of our approach and the literature.

B. Organization and Notations

Section II presents the system model. Section III analyzes the convergence of OTA-FL. Section IV presents CDA. Section V presents the numerical results and is followed by our conclusions in Section VI.

Notations: We denote vectors in bold font. The gradient of a function $F(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ is denoted by $\nabla F(\boldsymbol{w})$. For a vector $\boldsymbol{x}$, its ℓ_2-norm is $\|\boldsymbol{x}\|$ and its transpose is $\boldsymbol{x}^T$. The notations $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real and imaginary parts, respectively. The expectation with respect to a random variable (RV) X is denoted by $\mathbb{E}_X[\cdot]$. The notation $\boldsymbol{y} \sim \mathcal{CN}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_m)$ means that $\boldsymbol{y}$ is a complex Gaussian random vector with mean vector $\boldsymbol{\mu}$ and covariance $\sigma^2 \boldsymbol{I}_m$, where $\boldsymbol{I}_m$ denotes the identity matrix of size $m \times m$. The notation $\mathcal{A} \subset \mathcal{B}$ means $\mathcal{A}$ is a proper subset of $\mathcal{B}$. The set of non-negative integers is $\mathbb{N} = \{0, 1, \ldots\}$.

II. SYSTEM MODEL

We consider a network consisting of a set $\mathcal{K} = \{1, 2, \ldots, K\}$ of computing devices and a parameter server. Device k has a dataset $\mathcal{D}_k = \{1 \leq i \leq D_k : (\boldsymbol{x}_{ki}, \boldsymbol{\tau}_{ki})\}$, where $\boldsymbol{x}_{ki}$ and $\boldsymbol{\tau}_{ki}$ are the data and the corresponding ground-truth label for the i^{th} sample of the dataset. Let $f(\boldsymbol{w}, \boldsymbol{x}_{ki}, \boldsymbol{\tau}_{ki})$ be the loss function for sample i of dataset $\mathcal{D}_k$ that measures

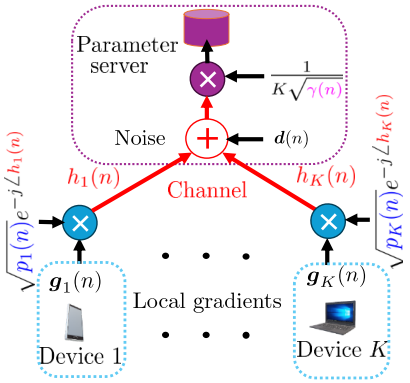

Fig. 1. System model for gradient aggregation in OTA-FL using K devices. The signal superposition in the wireless channel is shown in red ink.

the prediction error of the global model $\boldsymbol{w} \in \mathbb{R}^M$ on sample $\boldsymbol{x}_{ki}$ relative to its label $\boldsymbol{\tau}_{ki}$. Let

$$\boldsymbol{g}_k(n) = \frac{1}{m_b} \sum_{i \in \mathcal{B}_k(n)} \nabla f\left(\boldsymbol{w}, \boldsymbol{x}_{ki}, \boldsymbol{\tau}_{ki}\right), \tag{1}$$

be the local gradient of device k computed over the selected batch $\mathcal{B}_k(n)$, which is a randomly sampled subset of $\mathcal{D}_k$ and is of size m_b. We shall refer to $\boldsymbol{g}_k(n)$ as the local model of device k in iteration n.

The goal of federated learning is to find the optimal global model $\boldsymbol{w}^*$ that minimizes the average global loss $F(\boldsymbol{w}) = \left(\sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{D}_k} f\left(\boldsymbol{w}, \boldsymbol{x}_{ki}, \boldsymbol{\tau}_{ki}\right)\right)/D$ computed over $D = \sum_{k\in\mathcal{K}} D_k$ data points of all devices. Thus,

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathbb{R}^M}{\arg\min} \left\{ \frac{1}{D} \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{D}_k} f\left(\boldsymbol{w}, \boldsymbol{x}_{ki}, \boldsymbol{\tau}_{ki}\right) \right\}. \tag{2}$$

Over-the-Air Computation: Let $h_k(n) = |h_k(n)|\, e^{j\angle h_k(n)}$ be the complex baseband gain of the channel between device k and the parameter server, where $|\cdot|$ denotes the amplitude and $\angle\cdot$ denotes the phase. Device k amplifies $\boldsymbol{g}_k(n)$ by $\sqrt{p_k(n)}$, where $p_k(n)$ is the transmit power coefficient. It then compensates for the phase $\angle h_k(n)$ of the channel, and transmits the elements of $\boldsymbol{g}_k(n)$ sequentially over M symbol durations to the parameter server. This happens over a duration T. Thus, the symbols from the K devices add up coherently at the receiver. The signal vector $\boldsymbol{y}(n)$ received by the parameter server is given by

$$\boldsymbol{y}(n) = \sum_{k\in\mathcal{K}} |h_k(n)| \sqrt{p_k(n)} \boldsymbol{g}_k(n) + \boldsymbol{d}(n), \tag{3}$$

where $\boldsymbol{d}(n) \sim \mathcal{CN}\left(0, \sigma^2 \boldsymbol{I}_M\right)$ is additive white Gaussian noise. The transmit power of device k is $p_k(n) \left\|\boldsymbol{g}_k(n)\right\|^2 / M$.

The parameter server scales $\boldsymbol{y}(n)$ with $1/\left(K\sqrt{\gamma(n)}\right)$, where $\gamma(n) > 0$ is the receiver scaling. It obtains the following noisy estimate $\boldsymbol{r}(n)$ of the average of the local gradients:

$$\boldsymbol{r}(n) = \frac{1}{\sqrt{\gamma(n)}} \left(\sum_{k\in\mathcal{K}} c_k(n) \boldsymbol{g}_k(n) + \frac{\boldsymbol{d}(n)}{K} \right), \tag{4}$$

where $c_k(n) = |h_k(n)| \sqrt{p_k(n)}/K$. This model is illustrated in Fig. 1. The parameter server updates $\boldsymbol{w}(n)$ as follows [16]:

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta \Re\left\{\boldsymbol{r}(n)\right\}. \tag{5}$$

The parameter server then broadcasts $\boldsymbol{w}(n+1)$ to all the devices. This completes iteration n.[1] Device k uses $\boldsymbol{w}(n+1)$ to compute its local gradient $\boldsymbol{g}_k(n+1)$ for the next iteration.

Equation (5) is based on the FedSGD algorithm, which uses the following update rule for $D_1 = D_2 = \cdots = D_K$ and converges to $\boldsymbol{w}^*$ [2], [27]:

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \frac{\eta}{K} \sum_{k\in\mathcal{K}} \boldsymbol{g}_k(n). \tag{6}$$

Here, the parameter η is the learning rate.

Comment: Other algorithms such as federated averaging (FedAvg) have also been explored in the literature [28], [29]. In FedAvg, each device performs multiple local model updates in an iteration to reduce the communication bandwidth required by the system. However, doing so introduces client-drift, where each device's model deviates from the true global model. This drift increases the aggregation error and leads to worse performance than FedSGD [29]. Furthermore, OTA-FL already reduces the communication bandwidth since the devices transmit on the same time-frequency resource. Therefore, we focus on FedSGD in this work. Furthermore, our approach can be applied to FedAvg by specifying the transmit powers and receiver scaling at the end of the multiple local updates.

III. Convergence Analysis and the Concept of Effective Learning Rate

We now analyze the convergence of the learning algorithm in (5) for any choice of transmit powers and receiver scaling. We make the following technical assumptions about the loss function and gradients:

1) *Lipschitz-smoothness [30, Ch. 6]:* There exists a constant $L > 0$ such that

$$F(\boldsymbol{w}) - F(\boldsymbol{w}') \leq \nabla F(\boldsymbol{w}')^T (\boldsymbol{w} - \boldsymbol{w}') + \frac{L}{2} \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2. \tag{7}$$

2) *PL Inequality [26]:* Let F^* denote the optimal loss function value. There exists $\delta \geq 0$ such that

$$\left\|\nabla F(\boldsymbol{w})\right\|^2 \geq 2\delta \left(F(\boldsymbol{w}) - F^*\right). \tag{8}$$

3) *Finite Batch Variance:* $\boldsymbol{g}_k(n)$ is assumed to be an independent and unbiased estimate of $\nabla F(\boldsymbol{w}(n))$ with finite variance Ψ, which we shall refer to as the *batch variance*. $\boldsymbol{g}_k(n)$ satisfies the following properties:

$$\mathbb{E}_{\boldsymbol{b}(n)}\left[\boldsymbol{g}_k(n)\right] = \nabla F\left(\boldsymbol{w}(n)\right), \forall k \in \mathcal{K}, \tag{9}$$

$$\mathbb{E}_{\boldsymbol{b}(n)}\left[\left\|\boldsymbol{g}_k(n) - \nabla F\left(\boldsymbol{w}(n)\right)\right\|^2\right] \leq \frac{\Psi}{m_b}, \forall k \in \mathcal{K}, \tag{10}$$

[1] We ignore the effect of fading and noise for the broadcast of $\boldsymbol{w}(n+1)$, as is also widely assumed in the OTA-FL literature. This is justified because the parameter server, which is typically connected to a power grid, can transmit with a sufficiently high power and employ error correction coding.

where $\boldsymbol{b}(n)$ denotes the batches selected by all the devices in iteration n. Device k selects with uniform probability m_b data points from its dataset.[2]

Comments: We note the following about our assumptions:

- First, strong-convexity, which is assumed in [5], [6], and [23], is a special case of the PL inequality. Furthermore, [31] shows that even non-convex loss functions satisfy the PL inequality locally around the global minima in modern learning models in which the number of model parameters exceeds the number of training samples. Thus, the PL inequality makes the analysis tractable and has practical relevance.
- Second, non-convex loss functions that do not obey the PL inequality are considered in [11], [12], [15], and [16]. However, these works do not analyze the evolution of the optimality gap. They instead show that $\|\nabla F(\boldsymbol{w}(n))\|$ approaches 0, which only guarantees convergence to a stationary point. Furthermore, these works rely on other strong assumptions, such as bounded local gradients. Extending our analysis to such non-convex loss functions, but without assuming bounded local gradients, is an interesting avenue for future work.

Let $\boldsymbol{B}(n)$ denote the batches selected by all the devices up to iteration n and let

$$G(n) \triangleq \mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{d}(n-1),\ldots,\boldsymbol{d}(0)}\left[F\left(\boldsymbol{w}(n)\right) - F^*\right], \tag{11}$$

denote the optimality gap in iteration n averaged over $\boldsymbol{B}(n-1)$ and the noises $\boldsymbol{d}(0),\ldots,\boldsymbol{d}(n-1)$.

Theorem 1: If $0 < \frac{\eta}{K\sqrt{\gamma(n)}}\sum_{k\in\mathcal{K}}|h_k(n)|\sqrt{p_k(n)} < \frac{2}{L}$, then the optimality gap $G(n+1)$ in iteration n is upper bounded as follows:

$$G(n+1) \leq \alpha_n G(n) + \beta_n, \tag{12}$$

where α_n is the contraction factor and β_n is the residual error. They are given by

$$\alpha_n = 1 - 2\delta\eta_{\text{eff}}(n)\left(1 - \frac{L\eta_{\text{eff}}(n)}{2}\right), \tag{13}$$

$$\beta_n = \frac{L\eta^2}{2K^2\gamma(n)}\left(\frac{\Psi}{m_b}\sum_{k\in\mathcal{K}}|h_k(n)|^2 p_k(n) + \frac{M\sigma^2}{2}\right), \tag{14}$$

where $\eta_{\text{eff}}(n)$ is the effective learning rate and is equal to

$$\eta_{\text{eff}}(n) = \frac{\eta}{K\sqrt{\gamma(n)}}\sum_{k\in\mathcal{K}}|h_k(n)|\sqrt{p_k(n)}. \tag{15}$$

Proof: The proof is given in Appendix A. ∎

Interpretation: In the absence of channel fading and noise, we know from [26] that the optimality gap of gradient descent contracts as

$$G(n+1) \leq \left[1 - 2\delta\eta\left(1 - \frac{L\eta}{2}\right)\right]G(n). \tag{16}$$

Comparing (12) and (13) with (16), we see that $\eta_{\text{eff}}(n)$ is equivalent to η. For this reason, we refer to $\eta_{\text{eff}}(n)$ as the effective learning rate. It depends on the learning rate η, the number of devices K, the receiver scaling $\gamma(n)$, and the channel gains $\boldsymbol{h}(n) = [h_1(n),\ldots,h_K(n)]$ and the transmit power coefficients $[p_1(n),\ldots,p_K(n)]$ of all the devices.

Comments: Theorem 1 provides the following insights:

- In (13), α_n is a quadratic polynomial of $\eta_{\text{eff}}(n)$. We can show that $\alpha_n > 0$. To guarantee contraction of the optimality gap, we need $\alpha_n < 1$. Thus, we shall focus on the regime in which $0 < \alpha_n < 1$.
- Using the expressions of α_n and β_n in Theorem 1, we can show that for $\eta_{\text{eff}}(n) < (1/L)$, when $p_k(n)$ increases, α_n decreases but β_n increases. This implies that the optimality gap contracts faster but at the expense of a larger residual error. Hence, the contraction factor and the residual error exhibit an inherent trade-off.
- From (15), we can show that $\eta_{\text{eff}}^2(n) > \left(\eta^2\sum_{k\in\mathcal{K}}|h_k(n)|^2 p_k(n)\right)/\left(K^2\gamma(n)\right)$. Substituting this in (13) yields

$$\beta_n \leq \frac{L\Psi\eta_{\text{eff}}^2(n)}{2m_b} + \frac{LM\eta^2\sigma^2}{4K^2\gamma(n)}. \tag{17}$$

This is unlike classical gradient descent [26], where the residual error is zero. The residual error increases as the batch variance, noise variance, effective learning rate, or model dimension increases. It decreases as the batch size or the number of devices increases.

The following are the differences between our bound in Theorem 1 and the results in the literature:

- The bounds in [5], [15], and [16] are derived for a specific power allocation scheme. On the other hand, our general bound applies to any choice of transmit powers and receiver scaling. Furthermore, the bound in [15] applies only to one-bit quantized transmission and requires the local gradient vector to be Gaussian distributed, which need not be true for small batch sizes.
- The bounds in [5], [6], [11], [12], [16], [22], and [23] consider the ℓ_2-norm of the local and/or global gradients to be bounded in each iteration, which need not be the case if any of the datasets has outliers or the number of global model parameters is large [25]. Furthermore, the effective learning rate when reinterpreted from these bounds turns out to be independent of the transmit powers and the receiver scaling. Therefore, the trade-off between the contraction factor and the residual error is not captured by these bounds.

We now present a sufficient condition on $\eta_{\text{eff}}(n)$ for the recursive upper bound in (12) to converge.

Corollary 1: If $0 < \eta_{\text{eff}}(n) = \eta_0 < (2/L), \forall n$, then

$$\lim_{N\to\infty} G(N) < \infty. \tag{18}$$

Proof: The proof is given in Appendix B. ∎

Note that Corollary 1 specifies a range of values for $\eta_{\text{eff}}(n)$ and not η. Fixing η, as done in [17], [18], [19], [20], [21], [22], and [23], cannot guarantee convergence due to the presence of fading.

[2]We shall see in Section IV-B that while the upper bound is a function of Ψ, the design of CDA does not require the knowledge of Ψ.

Applying the bound in (12) successively for $n = 0, 1, \ldots, N-1$ yields the following upper bound:

$$G(N) \leq \left(\prod_{n=0}^{N-1} \alpha_n\right) G(0) + \sum_{n=0}^{N-2} \left[\prod_{j=n+1}^{N-1} \alpha_j\right] \beta_n + \beta_{N-1}. \tag{19}$$

The initial optimality gap $G(0)$ depends on the initial weights $\boldsymbol{w}(0)$. The aggregate residual error after N iterations is $\sum_{n=0}^{N-2} \left[\prod_{j=n+1}^{N-1} \alpha_j\right] \beta_n + \beta_{N-1}$, which we shall refer to as the *error floor*. Since $0 < \alpha_n < 1, \forall n \in \mathbb{N}$, we get $\left(\prod_{n=0}^{N-1} \alpha_n\right) G(0) \to 0$ as $N \to \infty$. Therefore, the error floor Ξ for large N is given by

$$\Xi = \lim_{N\to\infty} \left(\sum_{n=0}^{N-2} \left[\prod_{j=n+1}^{N-1} \alpha_j\right] \beta_n + \beta_{N-1}\right). \tag{20}$$

IV. Controlled Descent

From (20), we note that Ξ depends on $\{\alpha_n, \forall n \in \mathbb{N}\}$ and $\{\beta_n, \forall n \in \mathbb{N}\}$. These, in turn, depend on the transmit power coefficients $\{p_k(n), \forall n \in \mathbb{N}, \forall k \in \mathcal{K}\}$ of all devices and the receiver scaling $\{\gamma(n), \forall n \in \mathbb{N}\}$. Hence, we must jointly adapt the transmit powers and the receiver scaling to control the effective learning rate and the error floor. We shall minimize the error floor, subject to the following constraints:

1) The effective learning rate $\eta_{\text{eff}}(n)$ must be equal to a target value $\eta_{\text{tgt}}(n)$. From (15), we can compactly write this constraint as $\frac{\eta}{K\sqrt{\gamma(n)}} \sum_{k\in\mathcal{K}} |h_k(n)| \sqrt{p_k(n)} = \eta_{\text{tgt}}(n)$. From (13), this implies that

$$\alpha_n = 1 - 2\delta\eta_{\text{tgt}}(n)\left(1 - \frac{L\eta_{\text{tgt}}(n)}{2}\right), \forall n \in \mathbb{N}. \tag{21}$$

This formulation subsumes the special case where $\eta_{\text{tgt}}(n)$ is a constant. To the best of our knowledge, ours is the first work to explicitly control the effective learning rate.

2) The transmit power for each device should not exceed a peak power of $P_{\max}$. This is motivated by practical limitations on the device's power amplifier output [17], [18], [19], [21], [22]. Thus, $\left(p_k(n)\left\|\boldsymbol{g}_k(n)\right\|^2/M\right) \leq P_{\max}$.

Thus, we pose the following optimization problem:

$$\mathcal{P}_0: \min_{\substack{p_k(n),\forall k\in\mathcal{K},\forall n\in\mathbb{N} \\ \gamma(n)>0,\forall n\in\mathbb{N}}} \{\Xi\}, \tag{22a}$$

$$\text{s.t. } \frac{\eta}{K\sqrt{\gamma(n)}} \sum_{k\in\mathcal{K}} |h_k(n)| \sqrt{p_k(n)} = \eta_{\text{tgt}}(n), \quad \forall n \in \mathbb{N}, \tag{22b}$$

$$0 \leq \frac{1}{M} p_k(n) \left\|\boldsymbol{g}_k(n)\right\|^2 \leq P_{\max}, \quad \forall k \in \mathcal{K}, n \in \mathbb{N}. \tag{22c}$$

Comments: Our design philosophy and problem formulation differ from the literature in the following respects:

- The CCI schemes in [17], [18], [19], [20], and [21] minimize the GME. As mentioned, GME does not fully capture the impact of an error in an iteration on future iterations and the overall convergence behavior of the algorithm. We instead control the effective learning rate and minimize the error floor, which more faithfully captures the convergence behavior of the algorithm over multiple iterations.
- While minEF [22] also minimizes the error floor, it differs from $\mathcal{P}_0$ in three fundamental ways. First, minEF assumes that the ℓ_2-norms of the local and global gradients are bounded. Second, the unbiased aggregation constraint in [22] is a special case of (22b) when one substitutes $\eta_{\text{tgt}}(n) = \eta$ and $\gamma(n) = 1$. Third, minEF does not adapt the receiver scaling. As a result, minEF ends up having a larger residual error. Furthermore, it may not even have a feasible solution for some channel realizations.

Variable Transformations: To simplify the constraints in (22b) and (22c), we use the following variable transformations.

First, we define Γ_n as follows:

$$\Gamma_n = \frac{K\eta_{\text{tgt}}(n)\sqrt{\gamma(n)}}{\eta}. \tag{23}$$

Controlling the receiver scaling is equivalent to controlling Γ_n.

Second, we define $z_{k,n}$ as a scaled version of the transmit signal amplitude of device k in iteration n as follows:

$$z_{k,n} = \sqrt{p_k(n)}\left\|\boldsymbol{g}_k(n)\right\|. \tag{24}$$

Thus, the transmit power of device k is $z_{k,n}^2/M$. From (22c), the peak power constraint is equivalent to

$$z_{k,n} \leq \zeta = \sqrt{MP_{\max}}. \tag{25}$$

Third, we define the variables a_n and b_n as follows:

$$a_n = \frac{L\eta_{\text{tgt}}^2(n)\Psi}{2m_b}, \quad b_n = \frac{ML\eta_{\text{tgt}}^2(n)\sigma^2}{4}. \tag{26}$$

a_n and b_n capture the contribution of the gradient variance and the receiver noise, respectively, to the residual error.

Lastly, we define the metric $\theta_{k,n} \in \mathbb{R}^+$ of device k in iteration n as

$$\theta_{k,n} = \frac{|h_k(n)|}{\left\|\boldsymbol{g}_k(n)\right\|}. \tag{27}$$

As we shall see, this scalar captures all the information pertinent to the device.

With these variable transformations, we get $\beta_n = \frac{b_n}{\Gamma_n^2} + \frac{a_n}{\Gamma_n^2}\sum_{k\in\mathcal{K}} \theta_{k,n}^2 z_{k,n}^2$ and $\eta_{\text{eff}}(n) = \frac{\eta}{K\sqrt{\gamma(n)}} \sum_{k\in\mathcal{K}} \theta_{k,n} z_{k,n}$. From (23) and (27), the constraint in (22b) can be restated as $\sum_{k\in\mathcal{K}} \theta_{k,n} z_{k,n} = \Gamma_n$. From (24), the constraint in (22c) becomes $z_{k,n} \leq \zeta$.

Let $\boldsymbol{z}(n) = \left[z_{1,n}\ z_{2,n}\ \cdots\ z_{K,n}\right]$. Then, $\mathcal{P}_0$ can be restated as follows:

$$\mathcal{P}_1: \min_{\substack{\boldsymbol{z}(n),\forall n\in\mathbb{N}, \\ \Gamma_n>0,\forall n\in\mathbb{N}}} \left\{\lim_{N\to\infty} \left(\sum_{n=0}^{N-2} \left[\prod_{j=n+1}^{N-1} \alpha_j\right] \times \left[\frac{b_n}{\Gamma_n^2} + \frac{a_n}{\Gamma_n^2} \sum_{k\in\mathcal{K}} \theta_{k,n}^2 z_{k,n}^2\right] + \frac{b_{N-1}}{\Gamma_{N-1}^2} + \frac{a_{N-1}}{\Gamma_{N-1}^2} \sum_{k\in\mathcal{K}} \theta_{k,N-1}^2 z_{k,N-1}^2\right)\right\}, \tag{28a}$$

$$\text{s.t.} \sum_{k\in\mathcal{K}} \theta_{k,n} z_{k,n} = \Gamma_n, \forall n \in \mathbb{N}, \tag{28b}$$

$$0 \le z_{k,n} \le \zeta, \forall k \in \mathcal{K}, n \in \mathbb{N}. \tag{28c}$$

$\mathcal{P}_1$ has the following characteristics:

- *Feasibility:* $\mathcal{P}_1$ always has a feasible solution. For example, $\boldsymbol{z}(n) = \left[\zeta\ \zeta\ \cdots\ \zeta\right]^T$ (which corresponds to all devices transmitting at peak power) and $\Gamma_n = \zeta \sum_{k\in\mathcal{K}} \theta_{k,n}, \forall n \in \mathbb{N}$, is feasible. Since the objective function scales as $1/\Gamma_n^2$, it is also clear that the optimal value of Γ_n is strictly bounded away from 0.
- *Non-separability:* The objective function in (28a) has an additive form over the iteration index n. Furthermore, the constraints in (28b) and (28c) are on a per iteration basis. Even so, $\mathcal{P}_1$ is not separable across n because the local gradient of any device in iteration n and, therefore, its norm depend on the transmit power coefficients of all the devices in the previous iterations.
- *Non-causality:* The formulation is non-causal because it requires a priori knowledge of $|h_k(0)|, |h_k(1)|, \cdots$, across all iterations and all devices.

We now address the above challenges of non-separability and non-causality of the problem formulation.

A. Separable and Causal Reformulation

In iteration n, the contribution to the residual error in (28a) is $\left[\prod_{j=n+1}^{N-1} \alpha_j\right] \left(\left(b_n/\Gamma_n^2\right) + \left(a_n/\Gamma_n^2\right) \sum_{k\in\mathcal{K}} \theta_{k,n}^2 z_{k,n}^2 \right)$. To make the problem tractable and causal, we optimize $\boldsymbol{z}(n)$ and Γ_n to minimize the above contribution in iteration n. Thus, $\mathcal{P}_1$ reduces to solving the sequence of sub-problems $\mathcal{P}_2^{(1)}, \ldots, \mathcal{P}_2^{(N-1)}$, where $\mathcal{P}_2^{(n)}$ is the following sub-problem in iteration n:

$$\mathcal{P}_2^{(n)}: \min_{\boldsymbol{z}(n), \Gamma_n > 0} \left\{ \left[\prod_{j=n+1}^{N-1} \alpha_j \right] \left(\frac{b_n}{\Gamma_n^2} + \frac{a_n}{\Gamma_n^2} \sum_{k\in\mathcal{K}} \theta_{k,n}^2 z_{k,n}^2 \right) \right\}, \tag{29a}$$

$$\text{s.t.} \sum_{k\in\mathcal{K}} \theta_{k,n} z_{k,n} = \Gamma_n, \tag{29b}$$

$$0 \le z_{k,n} \le \zeta, \quad \forall k \in \mathcal{K}. \tag{29c}$$

Since $\eta_{\text{tgt}}(j)$ is a constant, it follows from (21) that $\alpha_j = 1 - 2\delta\eta_{\text{tgt}}(j)\left(1 - (L\eta_{\text{tgt}}(j)/2)\right)$ is also a constant. Furthermore, we can show using proof by contradiction that the optimal value of $z_{k,n}$ cannot be negative. Therefore, $\mathcal{P}_2^{(n)}$ is equivalent to the following problem $\mathcal{P}_3^{(n)}$:

$$\mathcal{P}_3^{(n)}: \min_{\boldsymbol{z}(n), \Gamma_n > 0} \left\{ \frac{b_n + a_n \sum_{k\in\mathcal{K}} \theta_{k,n}^2 z_{k,n}^2}{\Gamma_n^2} \right\}, \tag{30a}$$

$$\text{s.t.} \sum_{k\in\mathcal{K}} \theta_{k,n} z_{k,n} = \Gamma_n, \tag{30b}$$

$$z_{k,n} \le \zeta, \quad \forall k \in \mathcal{K}. \tag{30c}$$

Existence of a Solution: The constraints in (30b) and (30c) are linear. Thus, the set of feasible $\boldsymbol{z}(n)$ is compact. A solution exists as long as the feasible set is non-empty. From (30b) and (30c), this is true when $0 < \Gamma_n \le \zeta \sum_{k\in\mathcal{K}} \theta_{k,n}$.

$\mathcal{P}_3^{(n)}$ is a non-convex problem in $\boldsymbol{z}(n)$ and Γ_n. To solve it, we first find the optimal $\boldsymbol{z}(n)$ for any given $\Gamma_n > 0$ and then we find the optimal Γ_n. To characterize the optimal solution, we define the *peak power limited* (PPL) set $\mathcal{A}_n = \{k \in \mathcal{K} : z_{k,n} = \zeta\}$. It denotes the set of devices that transmit at peak power in iteration n. It is intimately related to the optimal $\boldsymbol{z}(n)$ as follows.

Lemma 1: Given $\mathcal{A}_n$ and $\Gamma_n > 0$, the solution of $\mathcal{P}_3^{(n)}$ is

$$z_{k,n} = \begin{cases} \dfrac{\Gamma_n - \zeta \sum_{j\in\mathcal{A}_n} \theta_{j,n}}{\theta_{k,n}\left(K - |\mathcal{A}_n|\right)}, & k \notin \mathcal{A}_n, \\ \zeta, & k \in \mathcal{A}_n. \end{cases} \tag{31}$$

The residual error β_n is given by

$$\beta_n = \frac{b_n}{\Gamma_n^2} + \frac{a_n}{\Gamma_n^2}\zeta^2 \left(\sum_{k\in\mathcal{A}_n} \theta_{k,n}^2 \right) + \frac{a_n}{\Gamma_n^2} \frac{\left(\Gamma_n - \zeta \sum_{k\in\mathcal{A}_n} \theta_{k,n}\right)^2}{K - |\mathcal{A}_n|}. \tag{32}$$

Proof: The proof is given in Appendix C. ∎

From (31), we see that $z_{k,n} \propto 1/\theta_{k,n}$ for devices not in the PPL set. Since $p_k(n) \propto z_{k,n}^2$ and $\theta_{k,n} \propto |h_k(n)|$, it follows that $p_k(n) \propto 1/\left(|h_k(n)|^2\right), \forall k \notin \mathcal{A}_n$. Thus, devices not in the PPL set employ CI.

In (32), we see that the residual error depends on $\mathcal{A}_n$. We henceforth show this dependence by writing β_n as $\beta_n(\mathcal{A}_n)$. Given $\mathcal{A}_n$ and Γ_n, we can get $\boldsymbol{z}(n)$ using (31). Therefore, we optimize $\mathcal{A}_n$ to minimize $\beta_n(\mathcal{A}_n)$. The following lemma presents an important property of $\beta_n(\mathcal{A}_n)$.

Lemma 2: Given $\mathcal{A}_n \subset \mathcal{K}$ and $\Gamma_n > 0$,

$$\beta_n(\mathcal{A}_n) \ge \beta_n(\mathcal{A}_n \setminus \{i\}), \quad \forall i \in \mathcal{A}_n. \tag{33}$$

Proof: The proof is given in Appendix D. ∎

$\beta_n(\mathcal{A}_n)$ decreases when any device is removed from the PPL set $\mathcal{A}_n$. Note that Lemma 2 does not imply that $\mathcal{A}_n$ should be the null set $\emptyset$ because the constraints in (30b) and (30c) must be satisfied.

B. Power Allocation Regions

As we show below, ensuring that the constraints in (30b) and (30c) are met leads to four regions that Γ_n can lie in. The four regions turn out to be functions of the smallest metric $\theta_{[K],n}$, where $[K]$ denotes the index of the device with the smallest metric in iteration n. For each region, we shall compute the optimal PPL set $\mathcal{A}_n^*$, the minimum residual error $\beta_n^*(\mathcal{A}_n^*)$, and $z_{k,n}^*$, which is the optimal value of $z_{k,n}$.

Region Ia) $0 < \Gamma_n < K\zeta\theta_{[K],n}$*:* The optimal solution in this region is as follows:

Lemma 3: For $0 < \Gamma_n < K\zeta\theta_{[K],n}$, the optimal PPL set $\mathcal{A}_n^*$ is $\emptyset$. The optimal value of $z_{k,n}$ is

$$z_{k,n}^* = \frac{\Gamma_n}{K\theta_{k,n}}, \quad \forall k \in \mathcal{K}. \tag{34}$$

Proof: The proof is given in Appendix E. ∎

Since $\mathcal{A}_n^* = \emptyset$, from (32), the minimum residual error is

$$\beta_n^*(\emptyset) = \frac{b_n}{\Gamma_n^2} + \frac{a_n}{K}. \tag{35}$$

From the definitions of a_n and b_n, we see that the optimal residual error in this region depends on the batch variance and the noise variance.

Region Ib) $\Gamma_n = K\zeta\theta_{[K],n}$*:* The optimal solution in this region is as follows:

Lemma 4: For $\Gamma_n = K\zeta\theta_{[K],n}$, the optimal PPL set is $\mathcal{A}_n^* = \{[K]\}$. The optimal value of $z_{k,n}$ is

$$z_{k,n}^* = \frac{\zeta\theta_{[K],n}}{\theta_{k,n}}, \quad \forall k \in \mathcal{K}. \tag{36}$$

Proof: The proof is given in Appendix F. ■

Since $\mathcal{A}_n^* = \{[K]\}$, the minimum residual error is $\beta_n(\{[K]\})$. Substituting (36) in (32) yields

$$\beta_n^*(\{[K]\}) = \frac{b_n}{K^2\zeta^2\theta_{[K],n}^2} + \frac{a_n}{K}. \tag{37}$$

Apart from a_n and b_n, the optimal residual error in this region depends on the weakest metric among all the devices.

Only device $[K]$ transmits at $P_{\max}$. For the other devices, we find upon substituting (23), (24), and (27) in (36) that

$$p_k(n) \propto \frac{1}{|h_k(n)|^2}, \quad \forall k \in \mathcal{K} \setminus \{[K]\}. \tag{38}$$

Thus, *clipped channel inversion* is optimal in Region Ib.

From (35) and (37), we see that the minimum residual error of Region Ia is greater than that of Region Ib because $\Gamma_n < K\zeta\theta_{[K],n}$ in Region Ia. Thus, Region Ia is sub-optimal and can be ignored.

Region II) $K\zeta\theta_{[K],n} < \Gamma_n < \zeta\sum_{k\in\mathcal{K}}\theta_{k,n}$*:* The following lemma provides insights about the optimal solution.

Lemma 5: If $K\zeta\theta_{[K],n} < \Gamma_n < \zeta\sum_{k\in\mathcal{K}}\theta_{k,n}$ then at least one, but not all, of the devices transmit with peak power.

Proof: The proof is given in Appendix G. ■

$\mathcal{A}_n^*$ needs to be determined by evaluating $\beta_n(\mathcal{S})$ over all proper and non-empty subsets of $\mathcal{K}$.

Region III) $\Gamma_n = \zeta\sum_{k\in\mathcal{K}}\theta_{k,n}$*:* From (30b) and (30c), we know that $\Gamma_n \le \zeta\sum_{k\in\mathcal{K}}\theta_{k,n}$, with equality only if $z_{k,n} = \zeta$. Hence, in this region, we must have

$$z_{k,n}^* = \zeta, \quad \forall k \in \mathcal{K}. \tag{39}$$

Thus, all devices transmit at peak power and $\mathcal{A}_n^* = \mathcal{K}$. The minimum residual error is $\beta_n^*(\mathcal{K})$. From (32), it is given by

$$\beta_n^*(\mathcal{K}) = \frac{b_n}{\Gamma_n^2} + \frac{a_n}{\Gamma_n^2}\zeta^2\left(\sum_{k\in\mathcal{K}}\theta_{k,n}^2\right) = \frac{b_n + a_n\zeta^2\sum_{k\in\mathcal{K}}\theta_{k,n}^2}{\zeta^2\left(\sum_{k\in\mathcal{K}}\theta_{k,n}\right)^2}. \tag{40}$$

Therefore, the optimal solution of $\mathcal{P}_3^{(n)}$ can be in Regions Ib, II, and III, depending on the values of a_n and b_n. We need to compute $\beta_n^*(\mathcal{A}_n^*)$ for each of these regions and pick the solution corresponding to the smallest one. However, for $(\Psi/m_b) \gg M\sigma^2$, the solution simplifies as shown below.

Theorem 2: When $(\Psi/m_b) \gg M\sigma^2$, Region Ib yields the minimum residual error among all four regions.

Proof: The proof is relegated to Appendix H. ■

Note that in Region Ib, $z_{k,n}^*$ and Γ_n are not functions of Ψ. Even when $(\Psi/m_b) \approx M\sigma^2$, it is advantageous to apply the solution of Region Ib for the following reasons:

1) In Region II, the optimal transmit powers need to be computed numerically by the parameter server. To do so, the parameter server needs to know the metrics of all the devices. This requires each device to transmit its metric to the parameter server. The parametric server then needs to communicate the computed transmit power to the corresponding device. Thus, the communication overhead at the parameter server and the devices is $\mathcal{O}(K)$.
2) Computing the optimal solution of Region II requires determining the transmit powers and the receiver scaling for all proper and non-empty subsets of $\mathcal{K}$. This entails a large search complexity of $\mathcal{O}\left(2^K\right)$.
3) Computing the minimum residual error in Regions II and III requires the parameter server to know Ψ, which is not the case in Region Ib.

Proposed Scheme: For the above reasons, we use the Ψ-agnostic solution of Region Ib. Substituting (24) and (27) in (36), the transmit power coefficient of device k is given by

$$p_k(n) = MP_{\max}\left(\frac{\theta_{[K],n}}{|h_k(n)|}\right)^2. \tag{41}$$

Recall that $\theta_{[K],n} = \min_{k\in\mathcal{K}}\{\theta_{k,n}\}$. In Region Ib, $\Gamma_n = K\zeta\theta_{[K],n}$. Since $\Gamma_n = K\eta_{\text{tgt}}(n)\sqrt{\gamma(n)}/\eta$ and $\zeta = \sqrt{MP_{\max}}$, the transmit power is given by

$$\gamma(n) = MP_{\max}\left(\frac{\eta\theta_{[K],n}}{\eta_{\text{tgt}}(n)}\right)^2. \tag{42}$$

Substituting the expressions for a_n and b_n from (26) in (37), the minimum residual error in iteration n is given by

$$\beta_n = \frac{L\eta_{\text{tgt}}^2(n)}{2K}\left(\frac{\Psi}{m_b} + \frac{1}{2K\theta_{[K],n}^2}\frac{\sigma^2}{P_{\max}}\right). \tag{43}$$

We shall refer to this scheme as CDA. Only device $[K]$, which has the smallest metric, transmits at the peak power. All the other devices transmit at lower powers. Since the transmit powers and the receiver scaling are known in closed form, CDA entails a computational complexity of $\mathcal{O}(1)$.

C. Implementation Aspects

1) Communication Protocol: From (41), the transmit power coefficient of device k in iteration n depends on $h_k(n)$, which the device knows, and the smallest metric $\theta_{[K],n}$ among all devices, which the device does not know a priori. The following protocol ensures that $\theta_{[K],n}$ is shared with all devices with only $\mathcal{O}(1)$ complexity. At the start of an iteration, the parameter server broadcasts a pilot signal to all the devices, which estimate their respective channel gains. Device k sets a timer that is a monotonically non-decreasing function of $\theta_{k,n}$ [32]. For example, the timer can be proportional to $\theta_{k,n}$. Upon the expiry of its timer, a device transmits its metric to the parameter server. We illustrate this in Fig. 2.

2) Computational Complexity: The complexity of computing the local gradient at a device is $\mathcal{O}(m_b M)$ and updating the global model at the parameter server is $\mathcal{O}(M)$; these complexities are identical for all OTA–FL schemes.

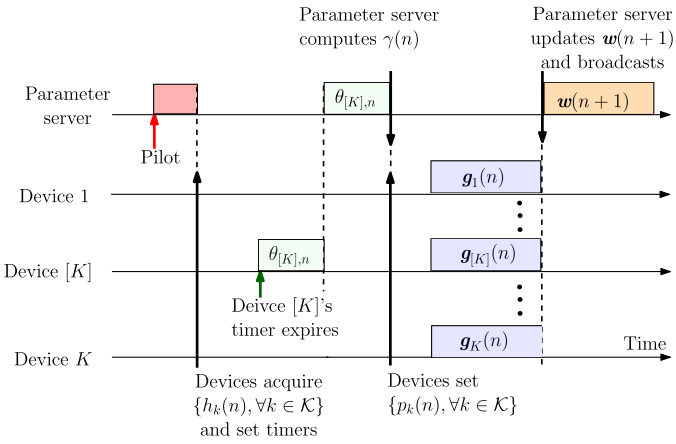

Fig. 2. Timeline of transmissions by the devices and the distributed timer scheme in iteration n.

The difference lies in computing the transmit powers and receiver scaling. In CCI [17], [18], [19], [20] and RCI [22], these are obtained by numerically solving an optimization problem at the parameter server. In contrast, CDA provides closed-form expressions; hence, the computational complexity is $\mathcal{O}(1)$.

3) Hardware and Software Requirements: Each device requires a processor capable of computing local stochastic gradients of complexity $\mathcal{O}(m_b M)$ and a transceiver capable of analog transmission over the shared wireless channel. These requirements are the same as the existing OTA-FL schemes.

D. Convergence Rate of CDA and Two Variants

Given our ability to control the effective learning rate, we consider two adaptations of CDA with: (1) fixed target learning rate (CDA-F), and (2) adaptive target learning rate (CDA-A).

From Corollary 1, CDA-F converges so long as $0 < \eta_{\text{tgt}}(n) < (2/L), \forall n$. When $\eta_{\text{tgt}}(n)$ is fixed at η_0, it follows from (21) that $\alpha_n = 1 - 2\delta\eta_0 + \delta L \eta_0^2 \triangleq \alpha, \forall n \in \mathbb{N}$. Substituting this in (19) yields

$$G(N) \leq \alpha^N G(0) + \sum_{n=0}^{N-1} \alpha^{N-n-1} \beta_n. \tag{44}$$

Thus, the optimality gap of CDA-F decreases as $\mathcal{O}\left(\alpha^N\right)$ rate after N iterations. This is the same as conventional stochastic gradient descent [26], but in the presence of fading.

To lower the error floor further, we adapt $\eta_{\text{eff}}(n)$. We first present a general sufficient condition to guarantee convergence when $\eta_{\text{eff}}(n)$ is a function of n. This will lead us to CDA-A.

Theorem 3: For $P_{\max} \to \infty$, if $\sum_{n=0}^{\infty} \eta_{\text{tgt}}^2(n) < \infty$ and $0 < \eta_{\text{tgt}}(n) < (2/L), \forall n$, then the optimality gap of CDA converges to a finite error floor $\Xi_A < \infty$.

Proof: The proof is given in Appendix I. ■

CDA-A: Consider the following adaptation of $\eta_{\text{tgt}}(n)$:

$$\eta_{\text{tgt}}(n) = \frac{\eta_0}{\left(1 + \frac{n}{R}\right)^2}, \forall n \in \mathbb{N}, \tag{45}$$

where $R > 0$ is a constant. We can show that

$$\sum_{n=0}^{\infty} \eta_{\text{tgt}}^2(n) \leq \eta_{\text{tgt}}^2(0) + \int_0^{\infty} \eta_{\text{tgt}}^2(n)\, dn \leq \eta_0^2 \left(1 + \frac{R}{3}\right). \tag{46}$$

Thus, CDA converges to a finite error floor for any $R > 0$. Substituting (45) in (42), we get the receiver scaling to be

$$\gamma(n) = M P_{\max} \left(\frac{\eta \theta_{[K],n}}{\eta_0}\right)^2 \left(1 + \frac{n}{R}\right)^4, \forall n. \tag{47}$$

We shall refer to this adaptation as CDA-A. The transmit power coefficient $p_k(n)$ is $M P_{\max} \left(\theta_{[K],n}/|h_k(n)|\right)^2$, which is the same as CDA-F. However, the two schemes need not have the same transmit powers. This is because the transmit powers also depend on the local model powers, which evolve differently with n since the effective learning rates are different.

The following corollary provides a sufficient condition for CDA-A to have a lower error floor than CDA-F.

Corollary 2: For any $0 < \eta_0 < (2/L)$, if $R \leq 3((L/\delta) - 1)$ then CDA-A has a lower error floor than CDA-F.

Proof: The proof is given in Appendix J. ■

In practice, CDA-A has a lower error floor even for larger values of R. This is because the derivation in Appendix J uses a weak bound on $\sum_{n=0}^{\infty} \eta_{\text{tgt}}^2(n)$.

E. Total Energy Consumption of CDA

In iteration n, device k expends an energy of $T p_k(n) \left\|\boldsymbol{g}_k(n)\right\|^2 / M$. Substituting the expression for $p_k(n)$ from (41), the energy $e(n)$ consumed by the devices in iteration n is given by

$$e(n) = T P_{\max} \sum_{k \in \mathcal{K}} \frac{\theta_{[K],n}^2}{\theta_{k,n}^2}. \tag{48}$$

V. Numerical Results

The channel power gain is set as $\mathbb{E}\left[|h_k(n)|^2\right] = \varrho d^{-\nu}$ [33], where $\varrho = -61$ dB, the distance d is 50 m, and the pathloss exponent ν is 3.7. The noise variance is $\sigma^2 = -114$ dBmW, which corresponds to a room temperature of 290 K and a bandwidth of 1 MHz. We set $T = 1$ ms. Let $\omega \triangleq \left(P_{\max}/\sigma^2\right) \varrho d^{-\nu}$ denote the fading-averaged maximum receive SNR at the parameter server when a device transmits with peak power.[3] We consider Rayleigh fading, unless mentioned otherwise.

A. Benchmarking Schemes

We benchmark CDA with the following schemes, which use SCI, TCI, CCI, or RCI. The transmit powers and receiver scaling in these schemes are as follows.

1) *SCI [10], [11], [12], [13], [14]:* In iteration n,

$$p_k(n) = \frac{\gamma(n)}{|h_k(n)|^2}. \tag{49}$$

We set $\gamma(n) = \frac{M P_{\max} \min_{k \in \mathcal{K}} \left\{|h_k(n)|^2\right\}}{\max_{k \in \mathcal{K}} \left\{\|\boldsymbol{g}_k(n)\|^2\right\}}$ to ensure that the peak power constraint is always satisfied.

[3]The above settings imply that $\omega = 20$ dB when $P_{\max} = 30$ dBmW.

2) *CI with Device Scheduling [5], [6]*: In the dynamic scheduling with energy constraint (DSEC) scheme of [5] and the gradient and channel-aware dynamic scheduling (GCADS) scheme of [6], we have

$$p_k(n) = \frac{\gamma(n)}{|h_k(n)|^2}, \tag{50}$$

where $\gamma(n)$ is chosen as per SCI. We clip the transmit power at $P_{\max}$ in order to ensure a fair comparison. In DSEC and GCADS, different criteria are used to select the devices that will transmit in an iteration. The criteria consider their past energy expenditure and the energy required for the current transmission. In DSEC, a selected device transmits its current local gradient. In GCADS, a selected device that did not transmit in the previous iteration uses a weighted combination of its current and previous local gradients.

3) *TCI [16]:* In iteration n,

$$p_k(n) = \begin{cases} \dfrac{MP_{\max}\gamma_{\text{th}}}{\xi\,|h_k(n)|^2}, & \text{for } |h_k(n)|^2 \geq \gamma_{\text{th}}, \\ 0, & \text{else}, \end{cases} \tag{51}$$

where ξ is a upper bound on $\|\boldsymbol{g}_k(n)\|^2$ and γ_{th} is a threshold that is numerically optimized. The receiver scaling is $\gamma(n) = MP_{\max}\gamma_{\text{th}}\exp\left(-2\gamma_{\text{th}}\right)/\xi$.[4] We note that the above scheme outperforms [3], [4], [15], which do not optimize γ_{th} to minimize the GME.

4) Minimum GME (min-GME) [17], [18], [19], [20]: In iteration n,

$$p_k(n) = \begin{cases} \dfrac{MP_{\max}}{\|\boldsymbol{g}_k(n)\|^2}, & \text{for } 1 \leq k \leq k^*, \\ \dfrac{\gamma(n)}{|h_k(n)|^2}, & \text{for } k^*+1 \leq k \leq K, \end{cases} \tag{52}$$

where $\gamma(n) \in \mathbb{R}^+$ and $k^* \in \mathbb{N}$ are computed numerically at the parameter server to minimize the GME. This scheme uses CCI.

5) minEF [22]: In iteration n,

$$p_k(n) = \min\left\{\frac{\gamma(n)\hat{\epsilon}_k(n)}{\left(|h_k(n)| + \epsilon_k(n)\right)^2}, \frac{MP_{\max}}{\|\boldsymbol{g}_k(n)\|^2}\right\}, \tag{53}$$

where $\hat{\epsilon}_k(n)$ and $\epsilon_k(n)$ are computed numerically. This scheme uses RCI. Here, $\gamma(n)$ is a constant. We set it to the fading-averaged channel power gain -124 dB to match the pathloss-free model assumed in [22]. We also consider the minEF-Adaptive (minEF-A) variant of [22] where $\eta \propto 1/n$. While [23] also uses RCI, it has a higher error floor as it does not optimize its regularization constant $\epsilon_k(n)$.

We present results for linear regression in Section V-B and for multi-class logistic regression in Section V-C.

B. Linear Regression

The loss function is given by [22]

$$f\left(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{\tau}\right) = \frac{1}{2}\left(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{\tau}\right)^2 + 5 \times 10^{-6}\|\boldsymbol{w}\|^2. \tag{54}$$

[4]We set $\xi = \max_{k\in\mathcal{K}}\left\{\|\boldsymbol{g}_k(n)\|^2\right\}$ because it is the tightest upper bound on $\|\boldsymbol{g}_k(n)\|^2$ for all $k \in \mathcal{K}$.

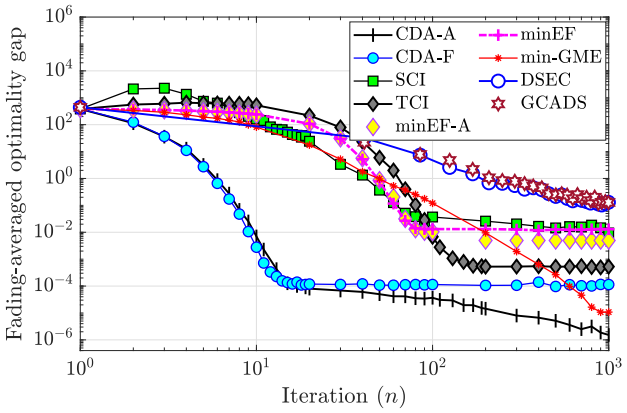

Fig. 3. Fading-averaged optimality gap as a function of the number of iterations ($P_{\max} = 20$ dBmW and $K = 40$).

For model training, we employ a randomly generated synthetic dataset of size 4×10^4. The data vector $\boldsymbol{x}$ is of dimension 10. It is Gaussian distributed with mean $\mathbf{0}$ and an identity covariance matrix. The label is $\boldsymbol{\tau} = \begin{bmatrix}1 & 2 & \cdots & 10\end{bmatrix}^T\boldsymbol{x} + 0.1\boldsymbol{\Omega}$, where $\boldsymbol{\Omega} \sim \mathcal{CN}\left(\mathbf{0}, \boldsymbol{I}_{10}\right)$ is the dataset noise.

Training Dataset Partitioning and Averaging: We create $K = 40$ non-identical partitions of the synthetic dataset and assign them to the devices such that $D_1 : \cdots : D_{40} = 1 : \cdots : 40$ to model heterogeneous local datasets across the devices. Thereafter, device k scales its data and labels by k to create $\mathcal{D}_k$. As per [22], the parameter L is the maximum eigenvalue of $\frac{\boldsymbol{X}^T\boldsymbol{X}}{D} + 10^{-4}\boldsymbol{I}$, where $\boldsymbol{X} = [\boldsymbol{x}_{1,1} \cdots \boldsymbol{x}_{1,D_1} \cdots \boldsymbol{x}_{40,1} \cdots \boldsymbol{x}_{40,D_{40}}]^T$. For our dataset, $\delta = 1.59 \times 10^3$ and $L = 1.69 \times 10^3$. We set $m_b = 100$. We average over 200 batches, channel fades, and noise realizations.

The learning rates for the CI, TCI, min-GME, and minEF schemes are $\eta = 1/L$. For minEF-A, $\eta = 1/\left(L\left(1 + (n/500)\right)\right)$. For CDA-F and CDA-A, we instead set $\eta_{\text{tgt}}(n) = 1/L$ and $\eta_{\text{tgt}}(n) = 1/\left(L\left(1 + (n/100)\right)^2\right)$, respectively. For DSEC and GCADS, the energy budget per device is 2.5 μJ per iteration.

Fig. 3 plots the fading-averaged optimality gaps of all the schemes as a function of the number of iterations n. As n increases, the optimality gaps of CDA-F and CDA-A decrease the fastest. For example, to reach an optimality gap of 10^{-2}, CDA-F and CDA-A require only 9 and 10 iterations, respectively. On the other hand, min-GME requires 202 iterations, TCI requires 97 iterations, minEF-A requires 100 iterations, SCI requires 900 iterations, DSEC and GCADS require more than 1000 iterations, and minEF never achieves the optimality gap. CDA-A achieves the lowest optimality gap among all the schemes. CDA-F also has a lower optimality gap than all the benchmark schemes except for $n > 600$, where min-GME has a lower optimality gap. The error floor of SCI is two orders of magnitude higher than that of CDA-F due to its sub-optimal receiver scaling.[5]

We now compare the schemes for a given energy budget of E_{tot} summed over devices and iterations. Fig. 4 plots the

[5]CDA-A also reaches an error floor. It is equal to 10^{-7} and happens later at $n = 50000$, which is not captured by the x-axis range of the figure.

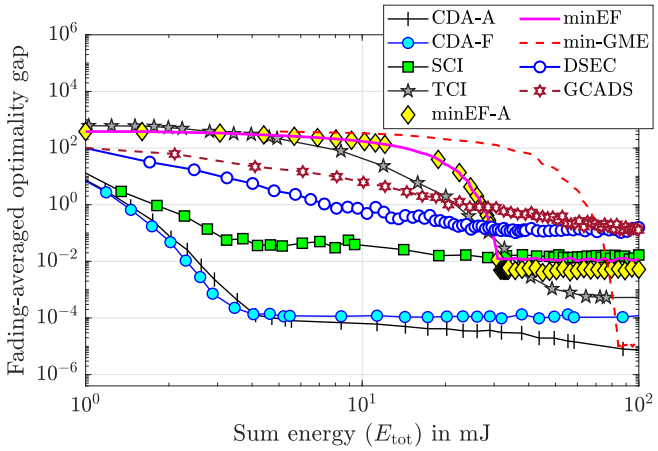

Fig. 4. Fading-averaged optimality gap as a function of the sum energy budget E_{tot} ($P_{\max} = 20$ dBmW and $K = 40$).

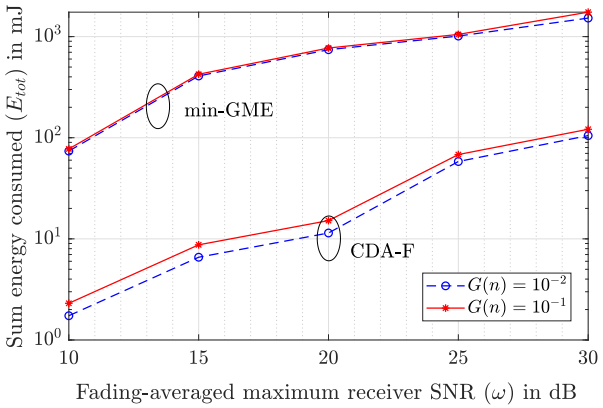

Fig. 5. Sum energy required to reach a target optimality gap of min-GME and CDA-F as a function of the SNR.

optimality gaps of the schemes as a function of E_{tot}. Each scheme consumes a different amount of energy in an iteration as the transmit powers are different. Hence, the number of iterations possible depends on the scheme. For example, $E_{\text{tot}} = 2$ mJ permits 10 iterations of TCI, 19 iterations of SCI, 25 iterations of DSEC, 21 iterations of GCADS, 1 iteration of min-GME, 2 iterations of minEF and minEF-A, and 8 iterations of CDA-F and CDA-A. Furthermore, as we saw above, the schemes converge at different rates. As E_{tot} increases, the optimality gaps of all the schemes decrease and reach error floors. When $E_{\text{tot}} \leq 81.8$ mJ, the optimality gaps of CDA-F and CDA-A are similar and much lower than those of the benchmark schemes. For larger values of E_{tot}, CDA-A has the lowest optimality gap. The optimality gap of min-GME converges to an error floor only for large E_{tot}. The error floor of min-GME is less than that of CDA-F but more than that of CDA-A.

Fig. 5 presents a complementary view of the energies consumed by the schemes. It plots the sum energy that the devices expend to reach optimality gaps of 0.01 and 0.1 as a function of ω. We do not show TCI, SCI, DSEC, GCADS, CDA-A, minEF, and minEF-A to avoid cluttering the figure. The sum energy of CDA-A is close to that of CDA-F. The sum energy increases as ω increases because the device transmit powers increase (see (41)). For any ω, the sum energy of CDA-F is one to two orders of magnitude lower than that of min-GME.

We see in Figs. 4 and 5 that CDA requires less total energy than the benchmark schemes to achieve a given optimality gap. This is due to two reasons. First, by its very design, CDA adapts the transmit powers and the receiver scaling to control the effective learning rate, in the presence of channel fading, and to minimize the residual error in each iteration. This leads to faster convergence. CDA requires fewer iterations than the benchmark schemes. Moreover, we saw in Section IV-B that only one device transmits at the peak power in an iteration in CDA. This results in less sum energy being consumed in each iteration.

C. Multi-Class Logistic Regression

We study the image classification problem for the CIFAR-10 dataset [34] that contains 32×32 colored images of objects that belong to 10 different classes. We use ResNet-18 [35] as our CNN model. It consists of 5 convolutional layers, an average pooling layer, a 512×10 fully connected layer, a rectilinear unit (ReLU) activation layer, and a SoftMax layer that computes the posterior class probabilities of a data-point. We use the cross-entropy between the posterior class probabilities and the one-hot encoded true labels as the non-convex loss function [36]. Unlike linear regression, the loss function does not follow the assumptions in (7) and (8). To prevent over-fitting, we use a drop-out of 0.1 and an ℓ_2-norm regularization with a coefficient of 0.001.

Training Dataset Partitioning and Averaging: We create 10 non-identical and independently distributed partitions of $50,000$ data-points and distribute them across the devices. The number of data points of any particular class across all the devices follows a non-symmetric Dirichlet distribution to model heterogeneous local datasets [37]. We set $m_b = 1024$. We average over 40 batches, channel fades, and noise realizations.

We set $\eta = 0.005$ and the momentum as 0.7 [38] for all the schemes. For minEF-A, $\eta = 0.005/\left(1 + (n/300)\right)$. We set $\eta_{\text{tgt}}(n) = 0.005$ for CDA-F and $\eta_{\text{tgt}}(n) = 0.005/\left((1 + (n/500))^2\right)$ for CDA-A. We measure the testing accuracy over $10,000$ data-points. For DSEC and GCADS, the energy budget per device is 3 mJ per iteration.

Fig. 6 plots the testing accuracy as a function n for all the schemes.[6] The testing accuracy increases as n increases, which implies that the schemes do generalize the CNN model better with more iterations. The testing accuracy of CDA-F and CDA-A increases faster compared to the other schemes. For example, to achieve a testing accuracy of 60%, CDA-F, CDA-A, TCI, SCI, min-GME, minEF, and minEF-A require 77, 48, 286, 285, 107, 170, and 241 iterations, respectively. DSEC and GCADS do not achieve the 60% testing accuracy. CDA-A achieves the highest testing accuracy among all the schemes.

[6]The 95% confidence intervals for all the schemes are less than a percent. We do not show them to avoid clutter. $P_{\max}$ is set higher in this problem to speed up the rate of convergence of all schemes.

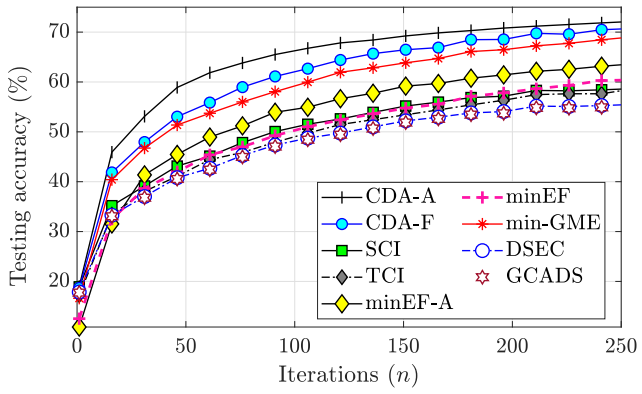

Fig. 6. Testing accuracy as a function of the number of iterations ($P_{\max} = 40$ dBmW and $K = 10$).

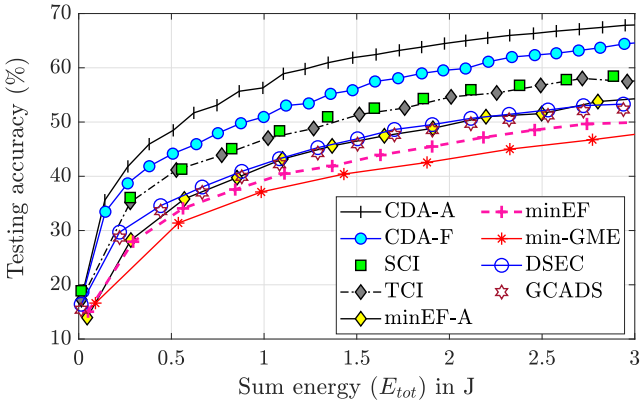

Fig. 7. Testing accuracy as a function of the sum energy budget E_{tot} ($P_{\max} = 40$ dBmW and $K = 10$).

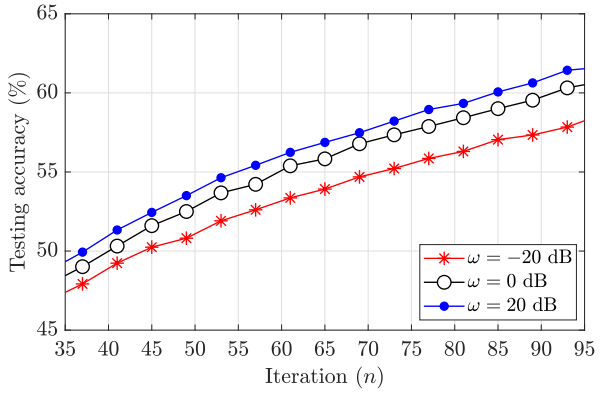

Fig. 8. Zoomed-in view of testing accuracy of CDA-F as a function of the number of iterations for different values of the SNR ω ($K = 10$).

Fig. 7 plots the testing accuracy as a function of the sum energy E_{tot} for all the schemes. As E_{tot} increases, the testing accuracy increases. CDA-A achieves a higher testing accuracy than all the schemes. For example, for a sum energy of 2.5 J, the testing accuracies of CDA-A, CDA-F, SCI, DSEC, GCADS, TCI, min-GME, minEF, and minEF-A are 66.9%, 62.4%, 57.8%, 51.3%, 50.5%, 57.4%, 47.9%, 49.7%, and 53.9%, respectively.

Fig. 8 plots the testing accuracy of CDA-F as a function of the number of iterations over a 40 dB range of the SNR ω.

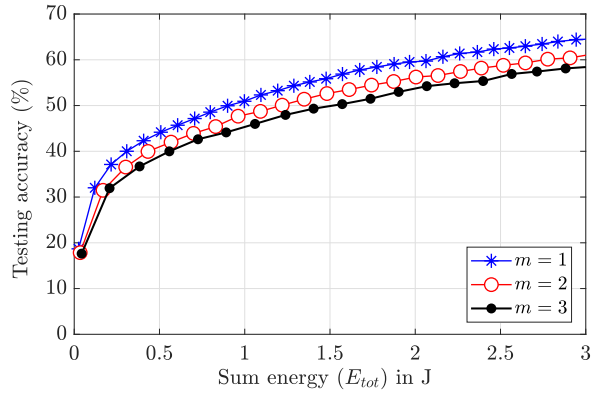

Fig. 9. Testing accuracy of CDA-F as a function of the sum energy consumed by the devices under Nakagami-m fading ($P_{\max} = 40$ dBmW and $K = 10$).

As ω increases, the testing accuracy improves. Even at a very low SNR of -20 dB, the reduction in the testing accuracy of CDA-F is at most 3.6%, which shows the robustness of CDA-F to receiver noise.

Fig. 9 plots the testing accuracy as a function of E_{tot} for the Nakagami-m channel model with $m = 1$ (Rayleigh fading), 2, and 5. The testing accuracy decreases as m increases for a given E_{tot}. This is because more devices transmit closer to the peak power. This can be seen from (41), in which the transmit power of device k is $P_{\max}\left(\theta_{[K],n}/|h_k(n)|\right)^2 \|\boldsymbol{g}_k(n)\|^2$. As m increases, this ratio increases and approaches $P_{\max}\|\boldsymbol{g}_k(n)\|^2/\left\|\boldsymbol{g}_{[K]}(n)\right\|^2$. The testing accuracy as a function of the number of iterations is insensitive of m. We do not show this to conserve space.

VI. Conclusion

We proposed CDA, which jointly adapted the device transmit powers and the receiver scaling to minimize the error floor while controlling the effective learning rate in each iteration. CDA-F kept the effective learning rate fixed, whereas CDA-A adapted it. We provided sufficient conditions for the convergence of both variants. The device-side information relevant for power adaptation was captured by the metric, which was the ratio between the channel power gain and the ℓ_2-norm of the local gradient. Furthermore, the smallest metric among all devices, which was required to set the transmit power, could be communicated using a distributed timer scheme with only $\mathcal{O}(1)$ overhead.

Our numerical results showed that CDA-A converged faster than several benchmark schemes. For the same energy budget, it achieved a lower optimality gap and a higher testing accuracy compared to the benchmark schemes. This advantage stemmed from CDA's ability to control the effective learning rate while minimizing the residual error in every iteration.

An interesting avenue for future work is to investigate the impact of imperfect channel knowledge and imperfect synchronization. Redesigning the proposed controlled descent framework, while avoiding restrictive bounded gradient assumptions, to algorithms such as FedAvg and stochastic controlled averaging is an interesting direction for future work. Characterizing the optimality gap for a general class of

non-convex functions and extending our approach to handle multiple antennas at the parameter server are some other interesting research directions.

APPENDIX

A. Proof of Theorem 1

Substituting $\boldsymbol{w} = \boldsymbol{w}(n+1)$, $\boldsymbol{w}' = \boldsymbol{w}(n)$, and the update of (5) in (7), we get $F(\boldsymbol{w}(n+1)) - F(\boldsymbol{w}(n)) \leq -\eta \nabla F(\boldsymbol{w}(n))^T \Re\{\boldsymbol{r}(n)\} + \left(L\eta^2/2\right) ||\Re\{\boldsymbol{r}(n)\}||^2$.

Taking expectation over the batches and the noise up to iteration n, we get

$$\mathbb{E}_{\boldsymbol{B}(n),\boldsymbol{d}(n),\ldots,\boldsymbol{d}(0)} \left[F\left(\boldsymbol{w}(n+1)\right) - F\left(\boldsymbol{w}(n)\right)\right] \leq -\eta T_1 + \frac{L\eta^2}{2} T_2, \tag{55}$$

where $T_1 = \mathbb{E}_{\boldsymbol{B}(n),\boldsymbol{d}(n),\ldots,\boldsymbol{d}(0)} \left[\nabla F\left(\boldsymbol{w}(n)\right)^T \Re\{\boldsymbol{r}(n)\}\right]$ and $T_2 = \mathbb{E}_{\boldsymbol{B}(n),\boldsymbol{d}(n),\ldots,\boldsymbol{d}(0)} \left[\|\Re\{\boldsymbol{r}(n)\}\|^2\right]$. We derive expressions for T_1 and T_2 below. Let $\boldsymbol{\Delta}(n) = \left[\boldsymbol{d}(0)\ \boldsymbol{d}(1) \cdots \boldsymbol{d}(n)\right]$.

(a) Computing T_1: Since $\boldsymbol{w}(n)$ is a function of $\boldsymbol{B}(n-1)$ and $\boldsymbol{d}(n-1), \ldots \boldsymbol{d}(0)$, and not of $\boldsymbol{b}(n)$ and $\boldsymbol{d}(n)$, we have

$$T_1 = \mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{\Delta}(n-1)} \left[\nabla F\left(\boldsymbol{w}(n)\right)^T \mathbb{E}_{\boldsymbol{b}(n),\boldsymbol{d}(n)} \left[\Re\{\boldsymbol{r}(n)\}\right]\right]. \tag{56}$$

From (4), we get

$$\mathbb{E}_{\boldsymbol{b}(n),\boldsymbol{d}(n)} \left[\Re\{\boldsymbol{r}(n)\}\right] = \frac{1}{\sqrt{\gamma(n)}} \left(\sum_{k\in\mathcal{K}} c_k(n) \mathbb{E}_{\boldsymbol{b}(n)} \left[\boldsymbol{g}_k(n)\right]\right) + \frac{1}{K\sqrt{\gamma(n)}} \mathbb{E}_{\boldsymbol{d}(n)} \left[\Re\{\boldsymbol{d}(n)\}\right]. \tag{57}$$

This is because $\boldsymbol{g}_k(n)$ is independent of $\boldsymbol{d}(n)$. Furthermore, $\boldsymbol{d}(n)$ is a 0-mean random vector independent of $\boldsymbol{b}(n)$. Hence,

$$\mathbb{E}_{\boldsymbol{b}(n),\boldsymbol{d}(n)} \left[\Re\{\boldsymbol{r}(n)\}\right] = \frac{1}{\sqrt{\gamma(n)}} \sum_{k\in\mathcal{K}} c_k(n) \mathbb{E}_{\boldsymbol{b}(n)} \left[\boldsymbol{g}_k(n)\right].$$

Since $\boldsymbol{g}_k(n)$ is an independent and unbiased estimate of $\nabla F\left(\boldsymbol{w}(n)\right)$, we also have $\mathbb{E}_{\boldsymbol{b}(n)} \left[\boldsymbol{g}_k(n)\right] = \nabla F\left(\boldsymbol{w}(n)\right)$. Thus, $\mathbb{E}_{\boldsymbol{b}(n),\boldsymbol{d}(n)} \left[\Re\{\boldsymbol{r}(n)\}\right] = \left(\nabla F\left(\boldsymbol{w}(n)\right)/\sqrt{\gamma(n)}\right) \sum_{k\in\mathcal{K}} c_k(n)$. Substituting this in (56), we get

$$T_1 = \frac{\sum_{k\in\mathcal{K}} c_k(n)}{\sqrt{\gamma(n)}} \mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{\Delta}(n-1)} \left[\|\nabla F\left(\boldsymbol{w}(n)\right)\|^2\right]. \tag{58}$$

(b) Computing T_2: From (4), we get

$$T_2 = \frac{1}{\gamma(n)} \mathbb{E}_{\boldsymbol{B}(n),\boldsymbol{\Delta}(n)} \left[\left\|\sum_{k\in\mathcal{K}} c_k(n) \boldsymbol{g}_k(n) + \frac{\Re\{\boldsymbol{d}(n)\}}{K}\right\|^2\right].$$

Adding and subtracting $\nabla F\left(\boldsymbol{w}(n)\right) \sum_{k\in\mathcal{K}} c_k(n)$, we get

$$T_2 = \frac{1}{\gamma(n)} \mathbb{E}_{\boldsymbol{B}(n),\boldsymbol{\Delta}(n)} \left[\left\|\left(\sum_{k\in\mathcal{K}} c_k(n) \left[\boldsymbol{g}_k(n) - \nabla F\left(\boldsymbol{w}(n)\right)\right]\right) + \left(\nabla F\left(\boldsymbol{w}(n)\right) \sum_{k\in\mathcal{K}} c_k(n) + \frac{\Re\{\boldsymbol{d}(n)\}}{K}\right)\right\|^2\right]. \tag{59}$$

Expanding the quadratic term inside the expectation, we get

$$T_2 = \frac{T_{21} + T_{22}}{\gamma(n)} + \frac{2}{\gamma(n)} \left[\sum_{k\in\mathcal{K}} c_k(n) \left[\mathbb{E}_{\boldsymbol{B}(n)} \left[\boldsymbol{g}_k(n)\right] - \nabla F\left(\boldsymbol{w}(n)\right)\right]^T\right] \times \left[\nabla F\left(\boldsymbol{w}(n)\right) \sum_{k\in\mathcal{K}} c_k(n) + \frac{\mathbb{E}_{\boldsymbol{d}(n)} \left[\Re\{\boldsymbol{d}(n)\}\right]}{K}\right], \tag{60}$$

where

$$T_{21} = \mathbb{E}_{\boldsymbol{B}(n)} \left[\left\|\sum_{k\in\mathcal{K}} c_k(n) \left(\boldsymbol{g}_k(n) - \nabla F\left(\boldsymbol{w}(n)\right)\right)\right\|^2\right], \tag{61}$$

$$T_{22} = \mathbb{E}_{\boldsymbol{d}(n)} \left[\left\|\sum_{k\in\mathcal{K}} c_k(n) \nabla F\left(\boldsymbol{w}(n)\right) + \frac{\Re\{\boldsymbol{d}(n)\}}{K}\right\|^2\right]. \tag{62}$$

Since $\mathbb{E}_{\boldsymbol{b}(n)} \left[\boldsymbol{g}_k(n)\right] = \nabla F\left(\boldsymbol{w}(n)\right)$, the third term in (60) is 0. The RV $\boldsymbol{g}_k(n) - \nabla F\left(\boldsymbol{w}(n)\right)$ is zero mean and independent across $k \in \mathcal{K}$, and its variance is upper bounded by Ψ/m_b. Hence,

$$T_{21} \leq \frac{\Psi}{m_b} \sum_{k\in\mathcal{K}} c_k^2(n). \tag{63}$$

Similarly, we can show that

$$T_{22} = \left(\sum_{k\in\mathcal{K}} c_k(n)\right)^2 \mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{\Delta}(n-1)} \left[\|\nabla F\left(\boldsymbol{w}(n)\right)\|^2\right] + \frac{M\sigma^2}{2K^2}. \tag{64}$$

Substituting (63) and (64) in (60), we get

$$T_2 \leq \frac{1}{\gamma(n)} \frac{\Psi}{m_b} \sum_{k\in\mathcal{K}} c_k^2(n) + \frac{\left(\sum_{k\in\mathcal{K}} c_k(n)\right)^2}{\gamma(n)} \times \mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{\Delta}(n-1)} \left[\|\nabla F\left(\boldsymbol{w}(n)\right)\|^2\right] + \frac{M\sigma^2}{2K^2\gamma(n)}. \tag{65}$$

Substituting (58) and (65) in (55) and rearranging, we get

$$\mathbb{E}_{\boldsymbol{B}(n),\boldsymbol{d}(n),\ldots,\boldsymbol{d}(0)} \left[F\left(\boldsymbol{w}(n+1)\right) - F\left(\boldsymbol{w}(n)\right)\right] \leq \frac{-\mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{\Delta}(n-1)} \left[\|\nabla F\left(\boldsymbol{w}(n)\right)\|^2\right]}{\sqrt{\gamma(n)}} \times \left[\eta \sum_{k\in\mathcal{K}} c_k(n) - \frac{L\eta^2}{2\sqrt{\gamma(n)}} \left(\sum_{k\in\mathcal{K}} c_k(n)\right)^2\right] + \frac{L\eta^2}{2\gamma(n)} \left[\frac{\Psi}{m_b} \sum_{k\in\mathcal{K}} c_k^2(n) + \frac{M\sigma^2}{2K^2}\right]. \tag{66}$$

Since $0 < \frac{\eta}{K\sqrt{\gamma(n)}} \sum_{k\in\mathcal{K}} |h_k(n)| \sqrt{p_k(n)} < \frac{2}{L}$, it can be shown that $\eta \sum_{k\in\mathcal{K}} c_k(n) - \frac{L\eta^2}{2\sqrt{\gamma(n)}} \left(\sum_{k\in\mathcal{K}} c_k(n)\right)^2 > 0$. From (8) we get $\mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{\Delta}(n-1)} \left[\|\nabla F\left(\boldsymbol{w}(n)\right)\|^2\right] \geq 2\delta$

$\left(\mathbb{E}_{\boldsymbol{B}(n-1),\boldsymbol{\Delta}(n-1)}\left[F\left(\boldsymbol{w}(n)\right)\right]-F^{*}\right)$. Substituting this in (66), we get

$$G(n+1)\leq\left[1-\frac{2\delta\eta}{\sqrt{\gamma(n)}}\sum_{k\in\mathcal{K}}c_k(n)+\frac{\delta L\eta^2}{2\gamma(n)}\left(\sum_{k\in\mathcal{K}}c_k(n)\right)^2\right]\times G(n)+\beta_n, \tag{67}$$

where $\beta_n=\frac{L\eta^2}{2\gamma(n)}\left(\frac{\Psi}{m_b}\sum_{k\in\mathcal{K}}c_k^2(n)+\frac{M\sigma^2}{2K^2}\right)$. Substituting $c_k(n)=|h_k(n)|\sqrt{p_k(n)}/K$ in (67) yields (12).

B. Proof of Corollary 1

From (12), we can further upper bound $G(n+1)$ as

$$G(n+1)\leq\alpha_0 G(n)+\beta_{\sup}, \tag{68}$$

where $\alpha_0=1-2\delta\eta_0+\delta L\eta_0^2$, $\beta_{\sup}=\sup_{n\in\mathbb{N}}\{\beta_n\}$, and $\sup_{n\in\mathbb{N}}\{\cdot\}$ denotes supremum.

Consider the map $S:\mathbb{R}\to\mathbb{R}$ where $S(\phi)=\alpha_0\phi+\beta_{\sup}$. Since $\eta_{\text{eff}}(n)<(2/L)$ and $\gamma(n)>0$, from (17), we get $\beta_{\sup}<\infty$. Hence, for any $\phi_1,\phi_2\in\mathbb{R}$, we get $|S(\phi_1)-S(\phi_2)|=\alpha_0|\phi_1-\phi_2|$. Since $\eta_0<(2/L)$, from (13), we get $\alpha_0<1$. Thus, $S(.)$ is a contraction mapping. Therefore, from Banach's fixed-point theorem [39], there exists a fixed-point $\phi^*<\infty$ at which $S(\phi^*)=\phi^*$, and the following iterative update converges to ϕ^*:

$$U(n+1)=S(U(n)). \tag{69}$$

Setting $U(0)=G(0)$, we get $U(1)=\alpha_0G(0)+\beta_{\sup}$. From (68), it follows that $G(1)\leq U(1)$. From (68), we get $G(2)\leq\alpha_0G(1)+\beta_{\sup}\leq\alpha_0U(1)+\beta_{\sup}$. Thus, from (69), we get $G(2)\leq U(2)$. Similarly, we can show that $G(n)\leq U(n),\forall n\in\mathbb{N}$. Since $\lim_{n\to\infty}U(n)=\phi^*<\infty$, we get

$$\lim_{n\to\infty}G(n)\leq\phi^*<\infty. \tag{70}$$

C. Proof of Lemma 1

Let $\boldsymbol{\mu}(n)=\{\mu_{k,n},\ k\in\mathcal{K}\}$. Then, the dual problem of $\mathcal{P}_2^{(n)}$ can be formulated as follows:

$$\Phi:\min_{\boldsymbol{z}(n),\lambda_n,\boldsymbol{\mu}(n)}\left\{\mathcal{L}\left(\boldsymbol{z}(n),\lambda n,\boldsymbol{\mu}(n)\right)\right\}, \tag{71a}$$

$$\text{s.t. }\lambda_n\neq0, \tag{71b}$$

$$\mu_{k,n}\geq0,\forall k\in\mathcal{K}, \tag{71c}$$

where λ_n and $\boldsymbol{\mu}(n)$ are the dual variables and

$$\mathcal{L}\left(\boldsymbol{z}(n),\lambda_n,\boldsymbol{\mu}(n)\right)=\frac{b_n}{\Gamma_n^2}+\frac{a_n}{\Gamma_n^2}\sum_{k\in\mathcal{K}}\theta_{k,n}^2z_{k,n}^2-\lambda_n\left(\sum_{k\in\mathcal{K}}\theta_{k,n}z_{k,n}-\Gamma_n\right)+\sum_{k\in\mathcal{K}}\mu_{k,n}\left(z_{k,n}-\zeta\right). \tag{72}$$

The dual problem Φ in (71) is convex in $\boldsymbol{z}(n)$. Differentiating $\mathcal{L}(.)$ with respect to $z_{k,n}$ and equating it to 0, we get

$$\frac{\partial\mathcal{L}}{\partial z_{k,n}}=\frac{2a}{\Gamma_n^2}\theta_{k,n}^2z_{k,n}-\lambda_n\theta_{k,n}+\mu_{k,n}=0. \tag{73}$$

When the constraint in (30b) is inactive, $\mu_{k,n}=0$. Hence, from (73), $z_{k,n}=\lambda_n\Gamma_n^2/\left(2a\theta_{k,n}\right)$. Furthermore, $k\notin\mathcal{A}_n$ by definition of the PPL set. When the constraint is active, we have $\mu_{k,n}>0$, $z_{k,n}=\zeta$, and $k\in\mathcal{A}_n$. Substituting these in (30b) yields

$$\zeta\sum_{j\in\mathcal{A}_n}\theta_{j,n}+\frac{\lambda_n\Gamma_n^2\left(K-|\mathcal{A}_n|\right)}{2a}=\Gamma_n. \tag{74}$$

Rearranging terms, we get $\left(\lambda_n\Gamma_n^2/2a\right)=\left(\Gamma_n-\zeta\sum_{j\in\mathcal{A}_n}\theta_{j,n}\right)/\left(K-|\mathcal{A}_n|\right)$. Substituting this in $z_{k,n}=\lambda_n\Gamma_n^2/\left(2a\theta_{k,n}\right)$ yields (31).

D. Brief Proof of Lemma 2

From (32), the residual errors $\beta_n(\mathcal{A}_n)$ and $\beta_n(\mathcal{A}_n\setminus\{i\})$ as a function of $\mathcal{A}_n$ are given by

$$\beta_n(\mathcal{A}_n)=\frac{b_n}{\Gamma_n^2}+\frac{a_n\zeta^2}{\Gamma_n^2}\sum_{k\in\mathcal{A}_n}\theta_{k,n}^2+\frac{a_n}{\Gamma_n^2}\frac{\left(\Gamma_n-\zeta\sum_{k\in\mathcal{A}_n}\theta_{k,n}\right)^2}{K-|\mathcal{A}_n|}, \tag{75}$$

$$\beta_n(\mathcal{A}_n\setminus\{i\})=\frac{b_n}{\Gamma_n^2}+\frac{a_n\zeta^2}{\Gamma_n^2}\sum_{k\in\mathcal{A}_n\setminus\{i\}}\theta_{k,n}^2+\frac{a_n}{\Gamma_n^2}\frac{\left(\Gamma_n-\zeta\sum_{k\in\mathcal{A}_n\setminus\{i\}}\theta_{k,n}\right)^2}{K+1-|\mathcal{A}_n|}. \tag{76}$$

After algebraic simplifications, we can show that $\beta_n(\mathcal{A}_n)-\beta_n(\mathcal{A}_n\setminus\{i\})$ is a perfect square and is, thus, positive.

E. Proof of Lemma 3

$\mathcal{A}_n$ cannot be $\mathcal{K}$ for the following reason. If $\mathcal{A}_n=\mathcal{K}$, then all the devices transmit at peak power. Hence, $z_{k,n}=\zeta,\forall k\in\mathcal{K}$. From (30b), we get $\Gamma_n=\zeta\sum_{k\in\mathcal{K}}\theta_{k,n}\geq K\zeta\theta_{[K],n}$. Thus, $\mathcal{A}_n=\mathcal{K}$ violates the upper limit of Γ_n.

Applying Lemma 2, we get $\beta_n\left(\mathcal{A}_n\right)\geq\beta_n\left(\emptyset\right)$. Substituting $\mathcal{A}_n=\emptyset$ in (31) yields $z_{k,n}=\Gamma_n/\left(K\theta_{k,n}\right)$. For $0<\Gamma_n<K\zeta\theta_{[K],n}$, this solution is feasible since $\sum_{k\in\mathcal{K}}\theta_{k,n}z_{k,n}=\Gamma_n$, which satisfies (30b), and $z_{k,n}<\zeta\theta_{[K],n}/\theta_{k,n}<\zeta,\forall k\in\mathcal{K}$, which satisfies (30c). Thus, $\mathcal{A}_n^*=\emptyset$.

F. Proof of Lemma 4

The proof hinges on the following three observations.

First, since $\Gamma_n=K\zeta\theta_{[K],n}$, from (31), we get $z_{[K],n}=\zeta$. Thus, $\mathcal{A}_n^*\neq\emptyset$ since $[K]\in\mathcal{A}_n^*$.

Second, $\mathcal{A}_n=\{[K]\}$ is feasible. This is because, from (31), it follows that $z_{k,n}=\zeta\theta_{[K],n}/\theta_{k,n}$, which satisfies (30c).

Third, $\mathcal{A}_n=\{[i]\}$ for $i\neq K$ is sub-optimal. The logic is as follows. We substitute $\mathcal{A}_n=\{[i]\}$ and $\mathcal{A}_n=\{[K]\}$ in the formula for the residual error in (32) and subtract the resulting expressions. We then get

$$\beta_n\left(\{[i]\}\right)-\beta_n\left(\{[K]\}\right)=\frac{a_n}{\Gamma_n^2}\zeta^2\left(\theta_{[i],n}^2-\theta_{[K],n}^2\right)+\frac{a_n}{\Gamma_n^2}\frac{\left(\Gamma_n-\zeta\theta_{[i],n}\right)^2-\left(\Gamma_n-\zeta\theta_{[K],n}\right)^2}{K-1}. \tag{77}$$

We can show using algebraic simplifications that the above expression is a perfect square. We skip the steps due to space constraints. Therefore, $\beta_n(\{[i]\}) \geq \beta_n(\{[K]\})$.

Lemma 2 then implies that the residual error of any subset of $\mathcal{K}$ that contains $[i]$ for $i \neq K$ is greater than or equal to $\beta_n(\{[K]\})$. Thus, $\mathcal{A}_n^* = \{[K]\}$. From above, we get $z_{k,n}^* = \zeta\theta_{[K],n}/\theta_{k,n}$.

G. Proof of Lemma 5

We first note that $\mathcal{A}_n = \emptyset$ is not a feasible PPL set. This is because, from (34), the lower limit $\Gamma_n > K\zeta\theta_{K,n}$ implies $z_{[K],n} = \Gamma_n/\left(K\theta_{[K],n}\right) > \zeta$. This violates the constraint in (30c).

Furthermore, $\mathcal{A}_n = \mathcal{K}$ is also infeasible. In this case, $z_{k,n} = \zeta, \forall k \in \mathcal{K}$. This implies $\Gamma_n = \zeta\sum_{k\in\mathcal{K}}\theta_{k,n}$, which exceeds the upper limit for Γ_n in Region II.

H. Proof of Theorem 2

From the definition of a_n and b_n in (26), $(\Psi/m_b) \gg M\sigma^2$ implies that $a_n \gg b_n$. The minimum residual errors in the three regions simplify as follows:

- *Region Ib:* From (37), we get $\beta_n^*(\{[K]\}) = a_n/K$.
- *Region II:* From Lemma 5, we get $\mathcal{K} \supset \mathcal{A}_n^* \supset \emptyset$. Thus, from Lemma 2, we get $\beta_n^*(\mathcal{A}_n^*) \geq \beta_n(\emptyset) \geq a_n/K$.
- *Region III:* Using the Cauchy-Schwartz inequality, we have $\left(\sum_{k\in\mathcal{K}}\theta_{k,n}\right)^2 \leq K\sum_{k\in\mathcal{K}}\theta_{k,n}^2$. Substituting this in (40), we get

$$\beta_n^*(\mathcal{K}) \geq \frac{b_n}{\zeta^2\left(\sum_{k\in\mathcal{K}}\theta_{k,n}\right)^2} + \frac{a_n}{K} > \frac{a_n}{K}. \tag{78}$$

Thus, Region Ib yields the lowest residual error.

I. Proof of Theorem 3

Let α_n' denote the contraction factor and β_n' denote the residual error of CDA-A with effective learning rate $\eta_{\text{tgt}}(n)$. From (20), the error floor Ξ_A of CDA-A is given by

$$\Xi_A = \lim_{N\to\infty}\left(\sum_{n=0}^{N-2}\left[\prod_{j=n+1}^{N-1}\alpha_j'\right]\beta_n' + \beta_{N-1}'\right). \tag{79}$$

Since $0 < \eta_{\text{tgt}}(n) < (2/L)$, we get $\alpha_n' < 1$. Therefore, from (79), we get $\Xi_A \leq \lim_{N\to\infty}\sum_{n=0}^{N-1}\beta_n'$.

For $P_{\max} \to \infty$, from (43), we get $\beta_n' = L\Psi\eta_{\text{tgt}}^2(n)/(2Km_b)$. Hence,

$$\Xi_A \leq \frac{L\Psi}{2Km_b}\left(\lim_{N\to\infty}\sum_{n=0}^{N-1}\eta_{\text{tgt}}^2(n)\right) < \infty. \tag{80}$$

J. Proof of Corollary 2

For CDA-F, $\alpha_n = 1 - 2\delta\eta_0 + L\delta\eta_0^2$ when $\eta_{\text{tgt}}(n) = \eta_0$. Furthermore, for $P_{\max} \to \infty$, we get $\beta_n = L\Psi\eta_0^2/(2Km_b)$ from (43). Substituting these in (20) and simplifying, we get the error floor Ξ_F of CDA-F to be

$$\Xi_F = \frac{L\Psi\eta_0}{2Km_b\delta(2-L\eta_0)}. \tag{81}$$

Ξ_F in (81) is greater than the upper bound of Ξ_A in (80) when $\sum_{n=0}^{\infty}\eta_{\text{tgt}}^2(n) \leq \eta_0/(\delta(2-L\eta_0))$. From (46), we get $\sum_{n=0}^{\infty}\eta_{\text{tgt}}^2(n) \leq \eta_0^2(1+(R/3))$. Therefore, $\Xi_A \leq \Xi_F$ when

$$\eta_0^2\left(1+\frac{R}{3}\right) \leq \frac{\eta_0}{\delta(2-L\eta_0)}. \tag{82}$$

It can be shown that this is true for any $0 < \eta_0 < (2/L)$ when $R \leq 3((L/\delta)-1)$.

REFERENCES

[1] S. Adhikary and N. B. Mehta, "Energy-efficient and fast controlled descent for over-the-air assisted federated learning," in *Proc. IEEE Global Commun. Conf.*, Dec. 2023, pp. 5268–5273.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, Apr. 2016, pp. 1273–1282.

[3] 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.

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

[5] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 227–242, Jan. 2022.

[6] J. Du, B. Jiang, C. Jiang, Y. Shi, and Z. Han, "Gradient and channel aware dynamic scheduling for over-the-air computation in federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1035–1050, Apr. 2023.

[7] M. Badi, C. Ben Issaid, A. Elgabli, and M. Bennis, "Balancing energy efficiency and distributional robustness in over-the-air federated learning," 2023, *arXiv:2312.14638*.

[8] F. Zhang, J. Chen, K. Wang, and W. Chen, "Device scheduling for relay-assisted over-the-air aggregation in federated learning," 2023, *arXiv:2312.12417*.

[9] H. Yang, P. Qiu, J. Liu, and A. Yener, "Over-the-air federated learning with joint adaptive computation and power control," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2022, pp. 1259–1264.

[10] C. Xu, S. Liu, Z. Yang, Y. Huang, and K.-K. Wong, "Learning rate optimization for federated learning exploiting over-the-air computation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3742–3756, Dec. 2021.

[11] J. Mao, H. Yang, P. Qiu, J. Liu, and A. Yener, "CHARLES: Channel-quality-adaptive over-the-air federated learning over wireless networks," in *Proc. IEEE 23rd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2022, pp. 1–5.

[12] J. Mao and A. Yener, "Personalized over-the-air federated learning with personalized reconfigurable intelligent surfaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 9076–9080.

[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, Mar. 2020.

[14] X. Ma, H. Sun, Q. Wang, and R. Q. Hu, "User scheduling for federated learning through over-the-air computation," in *Proc. IEEE 94th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2021, pp. 1–5.

[15] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, Mar. 2021.

[16] J. Yao, Z. Yang, W. Xu, D. Niyato, and X. You, "Imperfect CSI: A key factor of uncertainty to over-the-air federated learning," *IEEE Wireless Commun. Lett.*, vol. 12, no. 12, pp. 2273–2277, Dec. 2023.

[17] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7498–7513, Nov. 2020.

[18] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimization, analysis and scaling laws," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5488–5502, Aug. 2020.

[19] Z. Lin, Y. Gong, and K. Huang, "Distributed over-the-air computing for fast distributed optimization: Beamforming design and convergence analysis," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 274–287, Jan. 2023.

[20] Y. Chen, G. Zhu, and J. Xu, "Over-the-air computation with imperfect channel state information," in *Proc. IEEE 23rd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2022, pp. 1–5.
[21] X. An, R. Fan, S. Zuo, H. Hu, H. Jiang, and N. Zhang, "Joint power control and data size selection for over-the-air computation-aided federated learning," *IEEE Internet Things J.*, vol. 11, no. 8, pp. 14031–14046, Apr. 2024.
[22] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 342–358, Jan. 2022.
[23] L. Su and V. K. N. Lau, "Data and channel-adaptive sensor scheduling for federated edge learning via over-the-air gradient aggregation," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1640–1654, Feb. 2022.
[24] Y. Ren, Z. Wang, X. Zhang, and G. Lu, "Energy efficiency maximization for aerial RIS-aided over-the-air federated learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2024, pp. 1–6.
[25] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, Jun. 2012, pp. 1310–1318.
[26] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition," 2016, *arXiv:1608.04636*.
[27] S. J. Reddi et al., "Adaptive federated optimization," in *Proc. ICLR*, May 2020, pp. 1–20.
[28] X. Cao, G. Zhu, J. Xu, and S. Cui, "Transmission power control for over-the-air federated averaging at network edge," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1571–1586, May 2022.
[29] J. Zhu, Y. Shi, Y. Zhou, C. Jiang, W. Chen, and K. B. Letaief, "Over-the-air federated learning and optimization," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 16996–17020, May 2024.
[30] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed., Cambridge, U.K.: Cambridge Univ. Press, 2004.
[31] F. Liu, F. Gao, and Z. Lin, "Loss landscape and PL condition of overparameterized deep neural networks: Beyond the neural tangent kernel regime," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1091–1103, Mar. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9380086
[32] V. Shah, N. B. Mehta, and R. Yim, "Optimal timer based selection schemes," *IEEE Trans. Commun.*, vol. 58, no. 6, pp. 1814–1823, Jun. 2010.
[33] A. Goldsmith, *Wireless Communications*, 1st ed., Cambridge, U.K.: Cambridge Univ. Press, 2005.
[34] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. of Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/∼kriz/learning-features-2009-TR.pdf
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[36] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Aug. 1998.
[37] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, T. N. Hoang, and Y. Khazaeni, "Bayesian nonparametric federated learning of neural networks," in *Proc. ICML*, Jun. 2019, pp. 7252–7261.
[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 1st ed., Cambridge, MA, USA: MIT Press, 2016.
[39] C. Bey, E. Petrov, and R. Salimov, "On three-point generalizations of Banach and edelstein fixed point theorems," 2024, *arXiv:2404.05740*.

Sayantan Adhikary (Graduate Student Member, IEEE) received the Bachelor of Engineering degree in electronics and tele-communications engineering from Jadavpur University, Kolkata, India, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru. In 2019, he was with Samsung Semiconductor India Research. His research interests include design and analysis of distributed fusion and federated learning techniques for wireless network.

Neelesh B. Mehta (Fellow, IEEE) received the B.Tech. degree in electronics and communications engineering from Indian Institute of Technology (IIT) Madras, India, in 1996, and the M.S. and Ph.D. degrees in electrical engineering from California Institute of Technology, Pasadena, CA, USA, in 1997 and 2001, respectively.

He is currently a Professor and the Chair of the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru. His research group works on the design and modeling, performance analysis, and optimization of current and next-generation wireless communication systems. He is a fellow of Indian National Science Academy, Indian Academy of Sciences, Indian National Academy of Engineering, and National Academy of Sciences India. He was a recipient of the J. C. Bose Fellowship, the Shanti Swarup Bhatnagar Award, the Khosla Award, the Vikram Sarabhai Research Award, and the Swarnajayanti Fellowship. He served on the executive editorial committee for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2014 to 2017 and served as the Chair from 2017 to 2018. He also served as the Chair of the journal's steering committee. He has served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE WIRELESS COMMUNICATION LETTERS in the past. He has served on ComSoc's Board of Governors and Nominations and Elections Committee. He currently serves as the Chair of ComSoc's Distinguished Lecturers Selection Committee.