Energy-Efficient and Fast Controlled Descent for Over-the-Air Assisted Federated Learning

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

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437140

***Abstract*—We propose a novel energy-efficient controlled descent algorithm (EECDA) for over-the-air computation-assisted federated learning. In EECDA, the computing devices transmit their local parameters to the parameter server using amplitude modulation over a common time-frequency resource. As a result, a computation that involves adding the data of multiple users occurs automatically over the wireless channel since the signals superimpose. EECDA adapts the transmit powers of the devices and the amplification at the receiver to minimize the error floor on the optimality gap, which measures the performance of the federated learning algorithm. We derive the transmit powers and receiver amplification in closed-form. This is based on a novel recursive upper bound on the optimality gap that characterizes how wireless channel fades, device transmit powers, receiver amplification, noise variance, and batch selection variance determine the effective learning rate and error floor. For a small total energy, EECDA achieves a markedly lower optimality gap than the conventional minimum mean square error scheme.**

I. Introduction

Federated learning is a distributed and iterative machine learning technique that enables multiple devices to collaborate and train a common model without the need for centralized data storage [1]. It enables each device to train a shared training model with just its own local data. The devices then transmit their local models to a cloud-based central parameter server that aggregates them and then broadcasts the updated global model back to the devices. Federated learning is particularly useful in applications where data privacy and security are of critical concern. It allows for the incorporation of data from multiple sources, enables decentralized learning, and requires a lower communication bandwidth.

Over-the-air computation-assisted federated learning (AEFL) is a recent development that allows for the transmission of large amounts of data from the devices to the parameter server using shared time and frequency resources [2]–[4]. The devices transmit their local parameters to the parameter server using amplitude modulation over a common time-frequency resource. Since the signals superimpose in a wireless channel, a computation that involves adding the data of multiple users occurs automatically over the channel. The parameter server then uses the noisy superimposed signal it receives to update the global model. The efficacy of AEFL depends on the fading of the wireless channel between the device and the parameter server and the transmit powers of the devices.

The energy-efficiency of federated learning, when the devices transmit the model updates over a wireless channel, is investigated in [5]–[7]. In [8], [9], the transmit powers of the devices and the receiver amplification are adapted to minimize the mean square error (MSE) between the parameter aggregated using over-the-air computation and the average of the local parameters. In [10], the device transmit power is optimized subject to a constraint on the total energy consumed.

A. Contributions

We make the following contributions.

- We develop a novel recursive upper bound on the optimality gap, which measures the performance of the federated learning algorithm. It is the difference between the loss function given the current model computed over the datasets of all the devices in an iteration and its minimum value if the data were centrally available at the parameter server. In every iteration, the optimality gap contracts by a factor that depends on the effective learning rate, which is different from the learning rate specified in [9], [10], and increases by an additive error floor. The bound shows that increasing the learning rate leads to faster convergence but also an increase in the error floor.
- The bound has an elegant form that enables us to formulate a causal energy-efficient controlled decent (EECDA) algorithm. It adapts the device transmit powers and the receiver amplification factor to control the rate of convergence of AEFL, and thereby the number of iterations and energy required to converge, while minimizing the error floor contributed by each iteration. We derive closed-form expressions for the transmit powers of the devices and the amplification factor of the receiver.
- EECDA requires a lower sum energy than the conventional min-MSE scheme of [9]. Especially when the devices are power-constrained, the sum energy required for EECDA is 52-80% lower than min-MSE.

B. Organization and Notations

Section II presents the system model. Section III analyses the convergence of AEFL. Section IV specifies EECDA. We present numerical results in Section V, and our conclusions follow in Section VI.

Notations: We denote vectors in bold font. The gradient of a function $F(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ is written as $\nabla F(\boldsymbol{w})$. For a

S. Adhikary and N. B. Mehta are with the Dept. of Electrical Communication Engineering (ECE), Indian Institute of Science (IISc), Bangalore, India (Emails: sayantana@iisc.ac.in, nbmehta@iisc.ac.in).

This work was partially supported by the Qualcomm 6G UR program and the Prime Minister's Research Fellowship.

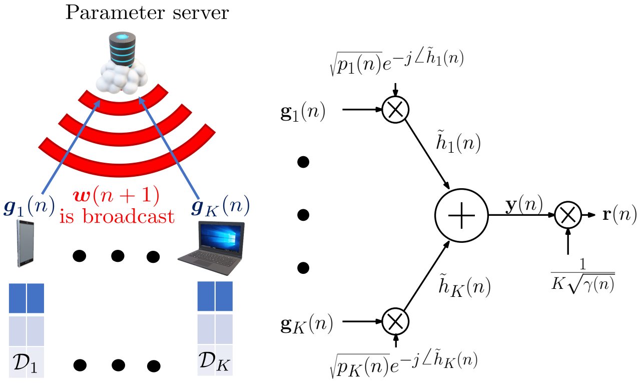

Fig. 1. System model of AEFL.

vector $\boldsymbol{x}$, the Euclidean norm is $\|\boldsymbol{x}\|$ and the transpose is $\boldsymbol{x}^T$. The expectation with respect to a random variable (RV) X is denoted by $\mathbb{E}_X[.]$. Also, $\boldsymbol{y} \sim \mathcal{CN}\left(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_m\right)$ means that $\boldsymbol{y}$ is a Gaussian random vector with mean vector $\boldsymbol{\mu}$ and covariance $\sigma^2\boldsymbol{I}_m$, where $\boldsymbol{I}_m$ denotes the identity matrix of order m. The null set is denoted by $\emptyset$.

II. System Model

Consider a network consisting of $\mathcal{K} = \{1, 2, \ldots, K\}$ computing devices that transmit data over a wireless channel to the parameter server. We consider a supervised learning paradigm in which device k has a local data-set $\mathcal{D}_k = \{1 \leq i \leq D_k : (\boldsymbol{x}_{ki}, \tau_{ki})\}$, where $\boldsymbol{x}_{ki}$ and τ_{ki} are the data and the corresponding ground-truth label for the i^{th} sample of the dataset. Let $f(\boldsymbol{w}, \boldsymbol{x}_{ki}, \tau_{ki})$ be the loss function for sample i of dataset $\mathcal{D}_k$ that measures the prediction error of the learning parameter $\boldsymbol{w} \in \mathbb{R}^M$ on sample $\boldsymbol{x}_{ki}$ relative to its label τ_{ki}.

The goal is to find the optimal parameter $\boldsymbol{w}^*$ that minimizes the average global loss $F(\boldsymbol{w}) = \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{D}_k} f(\boldsymbol{w}, \boldsymbol{x}_{ki}, \tau_{ki})/D$, which is computed over $D = \sum_{k\in\mathcal{K}} D_k$ data points across K devices. Thus, $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^M}\left\{\sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{D}_k} f(\boldsymbol{w}, \boldsymbol{x}_{ki}, \tau_{ki})/D\right\}$.

If the datasets of all the devices are present in the parameter server, then $\boldsymbol{w}^*$ is determined iteratively using the stochastic gradient descent (SGD) algorithm as [11]

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \frac{\eta}{Km_b}\sum_{\substack{k\in\mathcal{K}\\ i\in\mathcal{B}_k(n)}} \nabla f(\boldsymbol{w}, \boldsymbol{x}_{ki}, \tau_{ki}). \quad (1)$$

Here, η is the learning rate and $\mathcal{B}_k(n)$ is a randomly sampled subset of $\mathcal{D}_k$ with cardinality m_b.

A. Over-the-air Computation for Federated learning

Let $\tilde{h}_k(n) = h_k(n)e^{j\angle\tilde{h}_k(n)}$ be the complex baseband fading gain from device k to the parameter server. Device k is assumed to know $\tilde{h}_k(n)$ [8]–[10]. It transmits the local gradient $\boldsymbol{g}_k(n) = \frac{1}{m_b}\sum_{i\in\mathcal{B}_k(n)} \nabla f(\boldsymbol{w}, \boldsymbol{x}_{ki}, \tau_{ki})$ to the parameter server along with the other devices after compensating for the phase of $\tilde{h}_k(n)$, as shown in Fig. 1. The parameter server scales the received signal with $1/\left(K\sqrt{\gamma(n)}\right)$. Assuming time and frequency synchronization, the parameter server receives $\boldsymbol{r}(n)$ which is given by

$$\boldsymbol{r}(n) = \frac{1}{K\sqrt{\gamma(n)}}\sum_{k\in\mathcal{K}} h_k(n)\sqrt{p_k(n)}\boldsymbol{g}_k(n) + \frac{\boldsymbol{\rho}(n)}{K\sqrt{\gamma(n)}}. \quad (2)$$

Here, $p_k(n)$ is the transmit scaling coefficient of device k and $\boldsymbol{\rho}(n) \sim \mathcal{CN}\left(0, \sigma^2\boldsymbol{I}_M\right)$ is the complex Gaussian noise vector. Thus, device k's transmit power is $p_k(n)\|\boldsymbol{g}_k(n)\|^2$. We shall refer to $\gamma(n)$ as the amplification factor in iteration n.

Motivated by SGD, the parameter server updates the global parameter $\boldsymbol{w}(n)$ using $\boldsymbol{r}(n)$ as follows:

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta\boldsymbol{r}(n). \quad (3)$$

III. Convergence Analysis

We make the following assumptions about the loss function and gradients that are typical in the federated learning literature [10], [12]:

1) *Lipschitz-smoothness [13, Ch. 6]:* There exists a constant $L > 0$ such that $F(\boldsymbol{w}) - F(\boldsymbol{w}') \leq \nabla F(\boldsymbol{w}')^T(\boldsymbol{w} - \boldsymbol{w}') + L\|\boldsymbol{w} - \boldsymbol{w}'\|^2/2$.
2) *Polyak-Lojasiewicz Inequality [14]:* Let F^* denote the optimal loss function value. There exists $\delta \geq 0$ such that
$$\|\nabla F(\boldsymbol{w})\|^2 \geq 2\delta\left(F(\boldsymbol{w}) - F^*\right). \quad (4)$$
3) *Finite Variance Batch Selection:* $\boldsymbol{g}_k(n)$ is assumed to be an independent and unbiased estimate of $\nabla F(\boldsymbol{w}(n))$ with finite variance Ψ. It satisfies the following properties:
$$\mathbb{E}_b[\boldsymbol{g}_k(n)] = \nabla F(\boldsymbol{w}(n)), \text{ for } k \in \mathcal{K}, \quad (5)$$
$$\mathbb{E}_b\left[\|\boldsymbol{g}_k(n) - \nabla F(\boldsymbol{w}(n))\|^2\right] \leq \frac{\Psi}{m_b}, \text{ for } k \in \mathcal{K}, \quad (6)$$

where $\mathbb{E}_b[.]$ indicates the expectation over the batch selection.

We now present a novel upper bound on the optimality gap $G(n)$ in iteration n. It is the average of the difference between the global loss $F(\boldsymbol{w}(n))$ at iteration n and F^*.

Result *1:* The optimality gap $G(n+1)$ averaged over the batch selection and noise is upper bounded as:

$$\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[F(\boldsymbol{w}(n+1)) - F^*\right] \leq \alpha(n)\left(\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[F(\boldsymbol{w}(n)) - F^*\right]\right) + \beta(n), \quad (7)$$

where

$$\alpha(n) = 1 - 2\delta\eta_{\text{eff}}\left(1 - \frac{L\eta_{\text{eff}}}{2}\right), \quad (8)$$

$$\beta(n) = \frac{L\eta^2}{2\gamma(n)}\left(\frac{1}{K^2}\frac{\Psi}{m_b}\sum_{k\in\mathcal{K}} h_k^2(n)p_k(n) + \frac{M\sigma^2}{K^2}\right), \quad (9)$$

and

$$\eta_{\text{eff}} = \frac{1}{\sqrt{\gamma(n)}}\frac{\eta}{K}\sum_{k\in\mathcal{K}} h_k(n)\sqrt{p_k(n)}. \quad (10)$$

Proof: The proof is given in Appendix A. ■

Result 1 shows how the transmit powers and $\gamma(n)$ affect the rate of convergence of the algorithm and the error floor it converges to. This can be understood as follows:

- The optimality gap of gradient descent [14] with learning rate η contracts as $G(n+1) \leq$

$[1-2\delta\eta\left(1-L\eta/2\right)]\,G(n)$. Comparing it with (7) and (8) shows that η_{eff} is the effective learning rate.

- $\alpha(n)$ is the factor by which the optimality gap contracts in each iteration. To ensure a contraction, we need $|\alpha(n)| < 1$. The smaller the value of $|\alpha(n)|$, the faster the algorithm converges. From (8), we can show that $\alpha(n)$ achieves the minimum value of $1-\frac{\delta}{L}$ at $\eta_{\text{eff}}=\frac{1}{L}$. Therefore, the feasible range for $\alpha(n)$ is $0 \le 1-\frac{\delta}{L} \le \alpha(n) < 1$. Substituting this in (8), the feasible range for η_{eff} is $0 < \eta_{\text{eff}} < \frac{2}{L}$.
- $\beta(n)$ is the error floor in iteration n. It depends on the batch selection variance Ψ and the receiver noise variance σ^2. The smaller its value, the smaller is the optimality gap when the algorithm converges.

IV. Energy-efficient Controlled Descent

Applying the bound in (7) N times successively yields

$$G(N) \le \left[\prod_{n=0}^{N-1}\alpha(n)\right]G(0)+\sum_{n=0}^{N-1}\left[\prod_{j=n+1}^{N-1}\alpha(j)\right]\beta(n). \quad (11)$$

Since $\alpha(n) < 1$, from (11), we get $\lim_{N\to\infty} G(N) \le \lim_{N\to\infty}\sum_{n=0}^{N-1}\left[\prod_{j=n+1}^{N-1}\alpha(j)\right]\beta(n)$. This is an upper bound on the aggregate error floor.

We choose the transmit powers of the devices and the amplification factor in each iteration to minimize the bound on aggregate error floor subject to two constraints. First, the effective learning rate η_{eff} must be equal to a pre-specified target value η_{tgt}, which can be chosen from the interval $(0, 2/L)$. From (8), we get $\alpha(n) \triangleq \alpha_0 = 1-2\delta\eta_{\text{tgt}}\left(1-L\eta_{\text{tgt}}/2\right)$, for all n. Second, in every iteration, the transmit power for each device should not exceed $P_{\max}$. The constrained optimization problem can be stated as follows:

$$\mathcal{P}_1 : \min_{\substack{\{p_k(n),\forall k\in\mathcal{K},\forall n\},\\ \gamma(n)>0}} \left\{ \lim_{N\to\infty}\sum_{n=0}^{N-1}\alpha_0^{N-n-1}\frac{L\eta^2}{2\gamma(n)} \times\left[\frac{1}{K^2}\frac{\Psi}{m_b}\sum_{k\in\mathcal{K}}h_k^2(n)p_k(n)+\frac{M\sigma^2}{K^2}\right]\right\}, \quad (12a)$$

$$\text{s.t.}\quad \frac{\eta}{K\sqrt{\gamma(n)}}\sum_{k\in\mathcal{K}}h_k(n)\sqrt{p_k(n)}=\eta_{\text{tgt}},\forall n, \quad (12b)$$

$$p_k(n)\|\boldsymbol{g}_k(n)\|^2 \le MP_{\max},\forall k\in\mathcal{K},\forall n. \quad (12c)$$

Notice that $\|\boldsymbol{g}_k(n)\|$ depends on the transmit scaling coefficients of device k that are used in the previous $n-1$ iterations. Hence, $\mathcal{P}_1$ is not separable. It is also non-causal as it requires apriori knowledge of $h_k(0), h_k(1),\cdots,h_k(N),\forall k\in\mathcal{K}$. The additive form of the objective function motivates us to solve the following causal problem $\mathcal{P}_2^{(n)}$ for every n:

$$\mathcal{P}_2^{(n)} : \min_{\substack{\{z_k(n),\forall k\in\mathcal{K}\},\\ \gamma(n)>0}} \left\{\frac{b}{\gamma(n)}+\frac{a}{\gamma(n)}\sum_{k\in\mathcal{K}}\theta_k^2(n)z_k^2(n)\right\}, \quad (13a)$$

$$\text{s.t.}\sum_{k\in\mathcal{K}}\theta_k(n)z_k(n)=\Gamma(n), \quad (13b)$$

$$z_k(n)\le\zeta=\sqrt{MP_{\max}},\forall k, \quad (13c)$$

where

$$\Gamma(n)=K\left(\eta_{\text{tgt}}/\eta\right)\sqrt{\gamma(n)}\ge 0, \quad (14)$$

$\frac{b}{\gamma(n)}+\frac{a}{\gamma(n)}\sum_{k\in\mathcal{K}}\theta_k^2(n)z_k^2(n)$ is the error floor in the n^{th} iteration, $z_k(n)=\sqrt{p_k(n)}\|\boldsymbol{g}_k(n)\|$, $\theta_k(n)=\frac{h_k(n)}{\|\boldsymbol{g}_k(n)\|}$, $a=\frac{L\eta^2\Psi}{2m_bK^2}$, and $b=\frac{L\eta^2M\sigma^2}{2K^2}$. We shall refer to $\theta_k(n)$ as the *metric* of device k in iteration n.

To solve this problem, we first find the optimal $z_k(n)$ for any given $\gamma(n)>0$. Then, we find the optimal $\gamma(n)$. Let $\boldsymbol{z}(n)=\{z_k(n),\ k\in\mathcal{K}\}$. The first step is as follows:

$$\mathcal{P}_2^{(n)} := \min_{\boldsymbol{z}(n)}\left\{\frac{b}{\gamma(n)}+\frac{a}{\gamma(n)}\sum_{k\in\mathcal{K}}\theta_k^2(n)z_k^2(n)\right\}, \quad (15a)$$

$$\text{s.t.}\sum_{k\in\mathcal{K}}\theta_k(n)z_k(n)=\Gamma(n), \quad (15b)$$

$$0\le z_k(n)\le\zeta,\forall k\in\mathcal{K}. \quad (15c)$$

A solution exists so long as the feasible region is non-empty. This happens when

$$\Gamma(n)\le\zeta\sum_{k\in\mathcal{K}}\theta_k(n). \quad (16)$$

Let $[k]$ denote the index of the device with the k^{th} largest metric. Thus, $\theta_{[1]}(n)\ge\cdots\ge\theta_{[K]}(n)$. Let $\mathcal{A}=\{k\in\mathcal{K}: z_k(n)=\zeta\}$ denote the set of devices that are operating at peak power. We shall refer to $\mathcal{A}$ as the *peak power operating* (PPO) set. The solution of $\mathcal{P}_2^{(n)}$ given $\mathcal{A}$ is as follows. We skip the proof due to space constraints.

Result *2:* For a given $\mathcal{A}$, the solution to $\mathcal{P}_2^{(n)}$ is as follows:

$$z_k(n)=\begin{cases}\frac{\Gamma(n)-\zeta\sum_{j\in\mathcal{A}}\theta_j(n)}{\theta_k(n)(K-|\mathcal{A}|)}, & k\notin\mathcal{A},\\ \zeta, & k\in\mathcal{A}.\end{cases} \quad (17)$$

The error floor $\xi_n\left(\mathcal{A}\right)$ is given by

$$\xi_n(\mathcal{A})=\frac{b}{\gamma(n)}+\frac{a}{\gamma(n)}\zeta^2\left(\sum_{k\in\mathcal{A}}\theta_k^2(n)\right) + \frac{a}{\gamma(n)}\frac{\left(\Gamma(n)-\zeta\sum_{k\in\mathcal{A}}\theta_k(n)\right)^2}{K-|\mathcal{A}|}. \quad (18)$$

The following lemma shows an important property of $\xi_n(\mathcal{A})$ for $\mathcal{A}\subset\mathcal{K}$. We skip the proof to conserve space.

Lemma *1:* For any $\Gamma(n)$ and $\mathcal{A}\subset\mathcal{K}$, $\xi_n\left(\mathcal{A}\right)\ge\xi_n\left(\mathcal{A}\setminus\{i\}\right)$, for any $i\in\mathcal{A}$. ■

Thus, $\xi_n\left(\mathcal{A}\right)$ decreases when any device is removed from $\mathcal{A}$.

The feasibility region specified in (16) can be divided into the following three sub-regions:

Region I) $0<\Gamma(n)\le\zeta K\theta_{[K]}(n)$: If $\mathcal{A}=\mathcal{K}$, then $z_k(n)=\zeta,\forall k\in\mathcal{K}$. Hence, from (15b), we get $\Gamma(n)=\zeta\sum_{k\in\mathcal{K}}\theta_k(n)>K\zeta\theta_{[K]}(n)$. Thus, the constraint in (15b) cannot be satisfied. Hence, $\mathcal{A}$ must be a proper subset of $\mathcal{K}$.

Applying Lemma 1 successively, we get $\xi_n\left(\mathcal{A}\right)\ge\xi_n\left(\emptyset\right)$. From (17), $\mathcal{A}=\emptyset$ implies

$$z_k(n)=\frac{\Gamma(n)}{K\theta_k(n)},\forall k\in\mathcal{K}. \quad (19)$$

The solution is feasible since $\sum_{k\in\mathcal{K}} \theta_k(n) z_k(n) = \Gamma(n)$, which satisfies (15b), and $z_k(n) \le \zeta\theta_{[K]}(n)/\theta_k(n) \le \zeta, \forall k \in \mathcal{K}$, which satisfies (15c). Therefore, the optimal PPO set $\mathcal{A}^*$ is $\emptyset$. Using (18) and then (14), the smallest error floor as a function of $\gamma(n)$ is given by

$$\xi_n\left(\emptyset\right) = \frac{b}{\gamma(n)} + \frac{a\Gamma^2(n)}{K\gamma(n)} = \frac{b}{\gamma(n)} + aK\left(\frac{\eta_{\text{tgt}}}{\eta}\right)^2. \tag{20}$$

Region II) $\zeta K\theta_{[K]}(n) < \Gamma(n) < \zeta \sum_{k\in\mathcal{K}} \theta_k(n)$*:* We first note that $\mathcal{A} = \emptyset$ is not a feasible PPO set in this region. This is because it follows from (19) that $z_{[K]}(n) > \zeta$, which violates (15c). Furthermore, $\mathcal{A} = \mathcal{K}$ is also infeasible. It does not satisfy (15c) because $z_k(n) = \zeta, \forall k \in \mathcal{K}$ implies $\Gamma(n) = \zeta \sum_{k\in\mathcal{K}} \theta_k(n)$, which violates $\Gamma(n) < \zeta \sum_{k\in\mathcal{K}} \theta_k(n)$. Thus, at least some, but not all, of the devices operate at peak power. For any feasible $\mathcal{A}$, it follows from Lemma 1 that

$$\xi_n\left(\mathcal{A}\right) \ge \xi_n\left(\emptyset\right) = \frac{b}{\gamma(n)} + aK\left(\frac{\eta_{\text{tgt}}}{\eta}\right)^2. \tag{21}$$

Region III) $\Gamma(n) = \zeta \sum_{k\in\mathcal{K}} \theta_k(n)$*:* From (15b) and (15c), this can only happen when $z_k(n) = \zeta, \forall k \in \mathcal{K}$. Using the Cauchy–Schwarz inequality, we can show that

$$\xi_n\left(\mathcal{K}\right) \ge \frac{b}{\gamma(n)} + \frac{a\Gamma^2(n)}{K\gamma(n)} = \frac{b}{\gamma(n)} + aK\left(\frac{\eta_{\text{tgt}}}{\eta}\right)^2. \tag{22}$$

While we know the optimal powers in closed-form in Regions I and III, it needs to be computed numerically for Region II. However, when $\frac{\Psi}{m_b} \gg M\sigma_z^2$, the optimal solution is given as follows.

Result *3:* The optimal transmit scaling coefficient $p_k(n)$ and post-scaling factor $\gamma(n)$ for $\frac{\Psi}{m_b} \gg M\sigma_z^2$ are given by

$$p_k(n) = MP_{\max}\left(\frac{\theta_{[K]}(n)}{h_k(n)}\right)^2, \tag{23}$$

$$\gamma(n) = MP_{\max}\left(\frac{\eta\theta_{[K]}(n)}{\eta_{\text{tgt}}}\right)^2, \tag{24}$$

and the error floor is equal to

$$\beta(n) = \frac{L\eta_{\text{tgt}}^2}{2K}\left(\frac{\Psi}{m_b} + \frac{1}{K\theta_{[K]}^2(n)}\frac{\sigma^2}{P_{\max}}\right). \tag{25}$$

Proof: The proof is given in Appendix B. ■

Notice that $p_k(n)$ depends on $\theta_{[K]}(n)$. It can be communicated to the parameter server without all the devices having to transmit their metrics using the distributed timer-based scheme [15]. The parameter server, thereafter, broadcasts $\theta_{[K]}(n)$ to all devices.

V. Numerical Results

We benchmark EECDA with the conventional min-MSE scheme [9]. In this scheme, the device transmit powers and the amplification factor in an iteration are chosen to minimize the Euclidean distance between $\boldsymbol{r}(n)$ and the expected gradient:

$$\min_{\substack{\{p_k(n),\forall k\in\mathcal{K},\forall n\},\\ \gamma(n)>0}} \mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\boldsymbol{r}(n) - \frac{1}{K}\sum_{k\in\mathcal{K}} \boldsymbol{g}_k(n)\right\|^2\right], \tag{26a}$$

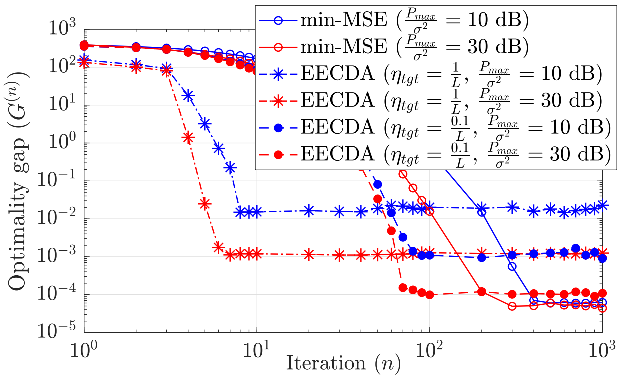

Fig. 2. Optimality gap as a function of the number of iterations for EECDA and min-MSE ($\eta = 0.1/L$).

$$\text{s.t.}\ \frac{1}{M} p_k(n) \|\boldsymbol{g}_k(n)\|^2 \le P_{\max}, \forall k \in \mathcal{K}. \tag{26b}$$

We consider a distributed linear regression problem with the loss function $f\left(\boldsymbol{w}, \boldsymbol{x}, \tau\right) = \left(\boldsymbol{w}^T\boldsymbol{x} - \tau\right)^2/2$. For model training and testing, we employ randomly generated synthetic datasets. The data vector $\boldsymbol{x}$ is of dimension 10. It is Gaussian distributed with mean $\mathbf{0}$ and identity covariance matrix. The label is $\tau = \begin{bmatrix}1 & 2 & \cdots & 10\end{bmatrix}^T \boldsymbol{x} + 0.1\Delta$, where $\Delta \sim \mathcal{CN}\left(0, 1\right)$ represents the dataset noise. A dataset of sample size 10^4 is evenly distributed among $K = 20$ devices without overlap. The batch size is $m_b = 100$ and $\eta = 0.1/L$. The batch average $\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[.\right]$ is computed by taking the average over 10^4 batches and noise realizations. As per [10], the parameter L is the maximum eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}/D$ where $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_K]^T$. The complex channel gain $\tilde{h}_k(n)$ for device k is a circularly symmetric complex Gaussian RV with variance 1, and the channel gains are independent across the devices and iterations. The time taken for a device to transmit its local gradient to the parameter server is T sec. Thus, device k spends an energy $p_k(n)\left\|\boldsymbol{g}_k(n)\right\|^2 \frac{T}{M}$ in iteration n.

Fig. 2 plots the optimality gap of min-MSE and EECDA as a function of the number of iterations. In both algorithms, $G(n)$ decreases as n increases and then saturates to an error floor. As η_{tgt} decreases, the error floor of EECDA decreases. However, it takes more iterations to converge. As $P_{\max}/\sigma^2$ increases, the error floor decreases, as we saw in Result 3. The error floor of min-MSE is lower than that of EECDA. However, min-MSE requires 3.0× and 42.8× more iterations to converge than EECDA with $\eta_{\text{tgt}} = 0.1/L$ and $1/L$, respectively. As $P_{\max}/\sigma^2$ increases or η_{tgt} decreases, the difference between the aggregate error floor of the two schemes diminishes.

We now consider a different simulation in which the schemes terminate when the total energy expenditure of all the devices reaches a pre-specified value E_{tot}. Note that the total energy consumed in the two schemes is different even for the same number of iterations because the device transmit powers are different. Fig. 3 plots the relative optimality gap $G_r(n) = G(n)/F^*$ as a function of $E_{\text{tot}}/(\sigma^2 T)$ for different target learning rates η_{tgt}. As E_{tot} increases, $G_r(n)$ of both

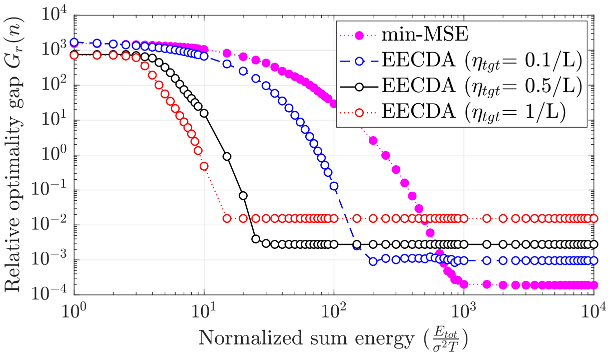

Fig. 3. Relative optimality gap of EECDA and min-MSE as a function of the normalized sum energy ($\eta = 0.1/L$ and $P_{\max}/\sigma^2 = 10$ dB).

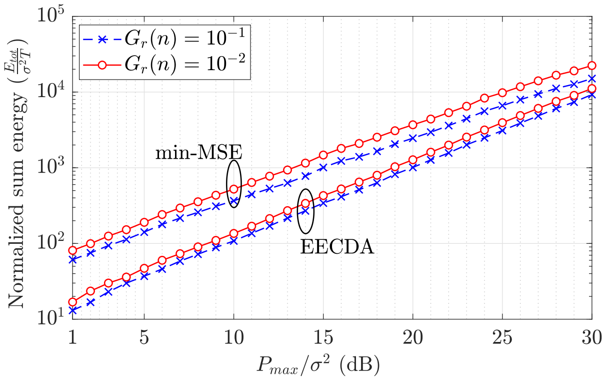

Fig. 4. Normalized sum energy of EECDA and min-MSE as a function of $P_{\max}/\sigma^2$ ($\eta = \eta_{\text{tgt}} = 0.1/L$).

schemes decreases and then converges to an error floor because the number of iterations increases. EECDA has a lower $G_r(n)$ than min-MSE for $E_{\text{tot}}/(\sigma^2 T) \leq 700$. Only for large E_{tot} does min-MSE have a smaller error floor than EECDA. This is because, unlike EECDA, min-MSE can decrease its effective learning rate to achieve a lower error floor when energy sufficient for a large number of iterations is available. As η_{tgt} increases, EECDA converges at a lower sum energy, but the error floor increases, which is consistent with (25).

Fig. 4 plots the normalized sum energy, which the devices expend to achieve a given $G_r(n)$, as a function of $P_{\max}/\sigma^2$. For both schemes, the sum energy increases as $P_{\max}/\sigma^2$ increases. For any $P_{\max}$, EECDA's sum energy is much lower than that of min-MSE. For $P_{\max}/\sigma^2 = 1$ dB, EECDA's sum energy is 80% lower. For $P_{\max}/\sigma^2 = 30$ dB, it is 52% lower.

VI. Conclusions

We developed a novel recursive upper bound on the optimality gap in AEFL. It characterized how the effective learning rate and the error floor depend on the channel fades, transmit powers, noise variance, and batch selection variance. We proposed EECDA, which adapted the transmit powers of the devices and the amplification factor to minimize the error floor while maintaining a target effective learning rate. The contribution of a device was encapsulated by its metric, which was the ratio of its channel power gain and the Euclidean norm of its local gradient. The bound brought out a trade-off between the effective learning rate and the error floor. The minimum error floor in an iteration decreased as the target learning rate η_{tgt} decreased, the maximum output power of a device $P_{\max}$ increased, or the smallest metric $\theta_{[K]}(n)$ among the K devices increased. EECDA required significantly lower sum energy than min-MSE to achieve the same optimality gap.

Appendix

A. Proof of Result 1

The L-smoothness of $F(.)$ and (3) imply that

$$F(\boldsymbol{w}(n+1)) - F(\boldsymbol{w}(n)) \leq -\eta \nabla F(\boldsymbol{w}(n))^T \boldsymbol{r}(n) + \frac{L\eta^2}{2}\left\|\boldsymbol{r}(n)\right\|^2. \tag{27}$$

Taking expectation over the batches and noise, we get

$$\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[F(\boldsymbol{w}(n+1)) - F(\boldsymbol{w}(n))\right] \leq -\eta \nabla F(\boldsymbol{w}(n))^T \times \mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\boldsymbol{r}(n)\right] + \frac{L\eta^2}{2}\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\boldsymbol{r}(n)\right\|^2\right]. \tag{28}$$

Here, we have used $\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\nabla F(\boldsymbol{w}(n))\right] = \nabla F(\boldsymbol{w}(n))$ since $\boldsymbol{w}(n)$ is broadcast by the parameter server and is, thus, known to all the devices.

(a) Computing $\mathbb{E}_b\left[\boldsymbol{r}(n)\right]$: Taking the expectation over the batches and noise (which are independent) and substituting $\mathbb{E}_b\left[\boldsymbol{g}_k(n)\right] = \nabla F(\boldsymbol{w}(n))$ (because $\boldsymbol{g}_k(n)$ is an independent and unbiased estimate of $\nabla F(\boldsymbol{w}(n))$) and $\mathbb{E}_{\boldsymbol{\rho}(n)}\left[\boldsymbol{\rho}(n)\right] = 0$ in (2) yields

$$\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\boldsymbol{r}(n)\right] = \frac{1}{\sqrt{\gamma(n)}}\sum_{k\in\mathcal{K}} c_k(n)\mathbb{E}_b\left[\boldsymbol{g}_k(n)\right], \tag{29}$$

where $c_k(n) = h_k(n)\sqrt{p_k(n)}/K$.

(b) Computing $\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\boldsymbol{r}(n)\right\|^2\right]$: From (2), we get

$$\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\boldsymbol{r}(n)\right\|^2\right] = \frac{1}{\gamma(n)}\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\sum_{k\in\mathcal{K}} c_k(n)\boldsymbol{g}_k(n) + \frac{\boldsymbol{\rho}(n)}{K}\right\|^2\right]. \tag{30}$$

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

$$\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\boldsymbol{r}(n)\right\|^2\right] = \frac{1}{\gamma(n)}\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\left(\sum_{k\in\mathcal{K}} c_k(n)\left(\boldsymbol{g}_k(n) - \nabla F(\boldsymbol{w}(n))\right)\right) + \left(\sum_{k\in\mathcal{K}} c_k(n)\nabla F(\boldsymbol{w}(n)) + \frac{\boldsymbol{\rho}(n)}{K}\right)\right\|^2\right]. \tag{31}$$

Expanding the term inside the expectation, we get

$$\mathbb{E}_{b,\boldsymbol{\rho}(n)}\left[\left\|\boldsymbol{r}(n)\right\|^2\right] = T_1 + T_2$$

$$+ \frac{2}{\gamma(n)} \left[\sum_{k \in \mathcal{K}} c_k(n) \left\{ \mathbb{E}_b \left[\boldsymbol{g}_k(n) \right] - \nabla F \left(\boldsymbol{w}(n) \right) \right\}^T \right]$$

$$\times \left[\sum_{k \in \mathcal{K}} c_k(n) \nabla F \left(\boldsymbol{w}(n) \right) + \frac{\mathbb{E}_{\boldsymbol{\rho}(n)} \left[\Re \left\{ \boldsymbol{\rho}(n) \right\} \right]}{K} \right], \quad (32)$$

where $T_1 = \frac{1}{\gamma(n)} \mathbb{E}_b \left[\left\| \sum_{k \in \mathcal{K}} c_k(n) \left(\boldsymbol{g}_k(n) - \nabla F \left(\boldsymbol{w}(n) \right) \right) \right\|^2 \right]$, $T_2 = \frac{1}{\gamma(n)} \mathbb{E}_{\boldsymbol{\rho}(n)} \left[\left\| \sum_{k \in \mathcal{K}} c_k(n) \nabla F \left(\boldsymbol{w}(n) \right) + \frac{\boldsymbol{\rho}(n)}{K} \right\|^2 \right]$, and $\Re \{ . \}$ denotes the real part.

From (5), we have $\mathbb{E}_b \left[\boldsymbol{g}_k(n) \right] = \nabla F \left(\boldsymbol{w}(n) \right)$. Hence, the third term in (32) is 0. The RV $\boldsymbol{g}_k(n) - \nabla F \left(\boldsymbol{w}(n) \right)$ is zero mean and independent across $k \in \mathcal{K}$. Hence, from (6), we get

$$T_1 \leq \frac{1}{\gamma(n)} \sum_{k \in \mathcal{K}} \left(c_k(n) \right)^2 \frac{\Psi}{m_b}. \quad (33)$$

Similarly, T_2 can be shown to be

$$T_2 = \frac{1}{\gamma(n)} \left(\sum_{k \in \mathcal{K}} c_k(n) \right)^2 \left\| \nabla F \left(\boldsymbol{w}(n) \right) \right\|^2 + \frac{M \sigma^2}{K^2 \gamma(n)}. \quad (34)$$

Substituting (33) and (34) in (32), we get

$$\mathbb{E}_{b, \boldsymbol{\rho}(n)} \left[\left\| \boldsymbol{r}(n) \right\|^2 \right] \leq \frac{1}{\gamma(n)} \sum_{k \in \mathcal{K}} \left(c_k(n) \right)^2 \frac{\Psi}{m_b}$$

$$+ \frac{1}{\gamma(n)} \left[\left(\sum_{k \in \mathcal{K}} c_k(n) \right)^2 \left\| \nabla F \left(\boldsymbol{w}(n) \right) \right\|^2 + \frac{M \sigma^2}{K^2} \right]. \quad (35)$$

Substituting (29) and (35) in (28) and rearranging the terms, we get

$$\mathbb{E}_{b, \boldsymbol{\rho}(n)} \left[F \left(\boldsymbol{w}(n+1) \right) - F \left(\boldsymbol{w}(n) \right) \right] \leq \frac{- \left\| \nabla F \left(\boldsymbol{w}(n) \right) \right\|^2}{\sqrt{\gamma(n)}}$$

$$\times \left[\eta \sum_{k \in \mathcal{K}} c_k(n) - \frac{L \eta^2}{2 \sqrt{\gamma(n)}} \left(\sum_{k \in \mathcal{K}} c_k(n) \right)^2 \right]$$

$$+ \frac{L \eta^2}{2 \gamma(n)} \left[\sum_{k \in \mathcal{K}} \left(c_k(n) \right)^2 \frac{\Psi}{m_b} + \frac{M \sigma^2}{K^2} \right]. \quad (36)$$

From the Polyak-Lojasiewicz inequality, $\left\| \nabla F \left(\boldsymbol{w}(n) \right) \right\|^2 \geq 2 \delta \left(\nabla F \left(\boldsymbol{w}(n) \right) - F^* \right)$. Furthermore, $\nabla F \left(\boldsymbol{w}(n) \right) - F^* = \mathbb{E}_{b, \boldsymbol{\rho}(n)} \left[\nabla F \left(\boldsymbol{w}(n) \right) - F^* \right]$ since the devices know $\boldsymbol{w}(n)$ from the parameter server's broadcast. Hence,

$$\mathbb{E}_{b, \boldsymbol{\rho}(n)} \left[F \left(\boldsymbol{w}(n+1) \right) - F \left(\boldsymbol{w}(n) \right) \right]$$

$$\leq \frac{-2 \delta}{\sqrt{\gamma(n)}} \left[\eta \sum_{k \in \mathcal{K}} c_k(n) - \frac{L \eta^2}{2 \sqrt{\gamma(n)}} \left(\sum_{k \in \mathcal{K}} c_k(n) \right)^2 \right]$$

$$\times \left(\mathbb{E}_{b, \boldsymbol{\rho}(n)} \left[F \left(\boldsymbol{w}(n) \right) \right] - F^* \right) + \beta(n), \quad (37)$$

where $\beta(n) = \frac{L \eta^2}{2 \gamma(n)} \left[\sum_{k \in \mathcal{K}} \left(c_k(n) \right)^2 \frac{\Psi}{m_b} + \frac{M \sigma^2}{K^2} \right]$. Rearranging (37) and substituting $c_k(n) = h_k(n) \sqrt{p_k(n)} / K$ yields (7).

B. Proof of Result 3

The condition $\frac{\Psi}{m_b} \gg M \sigma_z^2$ implies $a \gg b$. Under this condition, comparing the error floor for Region I in (20) with the lower bound of the error floor for Region II in (21) and that for Region III in (22), we observe that the error floor is the lowest if $\Gamma(n)$ is in Region I. In this region, we see from (20) that the error floor is a monotonically decreasing function of $\gamma(n)$. Therefore, the optimal value of $\gamma(n)$ is the largest value it can take in this region and is given as

$$\gamma(n) = \zeta^2 \left(\frac{\eta \theta_{[K]}(n)}{\eta_{\text{tgt}}} \right)^2 \quad (38)$$

Substituting the value of ζ from (13c) in (38) yields (24). Substituting (14) and (24) in (19), we get

$$z_k(n) = \sqrt{M P_{\max}} \frac{\theta_{[K]}(n)}{\theta_k(n)}. \quad (39)$$

Substituting $z_k(n) = \sqrt{p_k(n)} \| \boldsymbol{g}_k(n) \|$ and $\theta_k(n) = \frac{h_k(n)}{\| \boldsymbol{g}_k(n) \|}$ in (39) yields (23). Substituting (23) and (24) in (20) yields (25).

References

[1] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[2] 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, Oct. 2020.

[3] 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, 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, Feb. 2020.

[5] M. H. Mahmoud, A. Albaseer, M. Abdallah, and N. Al-Dhahir, "Federated learning resource optimization and client selection for total energy minimization under outage, latency, and bandwidth constraints with partial or no CSI," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 936–953, Apr. 2023.

[6] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Nov. 2021.

[7] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated deep reinforcement learning for internet of things with decentralized cooperative edge caching," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9441–9455, Apr. 2020.

[8] 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.

[9] 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.

[10] 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.

[11] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2023. [Online]. Available: http://arxiv.org/abs/1602.05629

[12] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, Nov. 2020.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed. Cambridge University Press, 2004.

[14] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition," 2020. [Online]. Available: http://arxiv.org/abs/1608.04636

[15] V. Shah, N. B. Mehta, and R. Yim, "Optimal timer based selection schemes," *IEEE Trans. Wireless Commun.*, vol. 58, no. 6, pp. 1814–1823, Jun. 2010.