Online Model Updating with Analog Aggregation in Wireless Edge Learning

Juncheng Wang*, Min Dong†, Ben Liang*, Gary Boudreau‡, and Hatem Abou-zeid‡
*Department of Electrical and Computer Engineering, University of Toronto, Canada,
†Department of Electrical, Computer, and Software Engineering, Ontario Tech University, Canada, ‡Ericsson Canada, Canada

***Abstract*—We consider federated learning in a wireless edge network, where multiple power-limited mobile devices collaboratively train a global model, using their local data with the assistance of an edge server. Exploiting over-the-air computation, the edge server updates the global model via analog aggregation of the local models over noisy wireless fading channels. Unlike existing works that separately optimize computation and communication at each step of the learning algorithm, in this work, we jointly optimize the training of the global model and the analog aggregation of local models over time. Our objective is to minimize the accumulated training loss at the edge server, subject to individual long-term transmit power constraints at the mobile devices. We propose an efficient algorithm, termed Online Model Updating with Analog Aggregation (OMUAA), to adaptively update the local and global models based on the time-varying communication environment. The trained model of OMUAA is channel- and power-aware, and it is in closed form with low computational complexity. We study the mutual impact between model training and analog aggregation over time, to derive performance bounds on the computation and communication performance metrics. Simulation results based on real-world image classification datasets and typical Long-Term Evolution network settings demonstrate substantial performance gain of OMUAA over the known best alternatives.**

I. Introduction

In wireless edge networks, mobile devices collect an enormous amount of data that can be used to train machine learning models. This motivates new machine leaning technologies at the edge servers and devices, collectively call *edge learning* [1]-[4]. However, the migration of learning from central cloud servers to the edge can lead to an explosion of information exchange between edge servers and devices. Thus, the scarcity of communication resources can become a bottleneck for training an accurate machine learning model at the edge. This calls for communication-efficient distributed learning algorithms that integrate techniques from two different areas, *i.e.,* machine learning and communications [5].

As a nascent distributed learning scheme, *federated learning* (FL) allows multiple local devices to collaboratively learn a global model without sending their local data to a central server [6], [7]. In FL, a key operation is to aggregate the local models sent from the local devices into a global model at the server. Toward reducing the communication overhead, the machine learning literature mainly focuses on quantization [8]-[10], sparsification [11]-[13], and local updates [14]-[16]. These approaches assume error-free transmission and ignore the physical wired or wireless communication layer. More recently, with the observation that the global model at the server can be expressed as a weighted sum of the local models, *analog aggregation* of the local models has been proposed, allowing simultaneous wireless transmission by the local devices over a multiple access channel [17]-[26]. Such *over-the-air* computation [27]-[30] reduces latency and bandwidth requirement compared with the conventional orthogonal multiple access.

All existing works on FL with analog aggregation separately optimize model training and wireless transmission [17]-[26]. In contrast, a joint optimization approach would take into fuller account the impact of wireless transmission in the model training process, and vice versa. Furthermore, prior works have focused on offline optimization, by solving one-shot optimization problems, which do not fully account for the changes in the environment over time or any long-term constraints. However, due to the dynamic fluctuation in the wireless channels, both model training and analog aggregation should be channel-aware and online, *i.e.,* adaptive to the unpredictable channel fluctuation over time.

In this work, we aim to develop an *online* algorithm that *jointly* optimizes model training and analog aggregation for FL over noisy wireless fading channels. To achieve this goal, we must address several challenges on multiple fronts. First, noisy wireless channels lead to errors in the analog aggregation of the learning models, and these errors are accumulated and amplified in the iterative steps of model training over time. Second, since the effectiveness of analog communication depends on the transmitted message, when designing the intermediate output models of each iterative step in model training, we must consider both their improvement in learning and their suitability for transmission. Third, the aforementioned tight coupling between model training and analog aggregation must be properly formulated and addressed in a dynamic online setting, where the future wireless communication environment is unpredictable. Finally, we must account for the energy budgets for device communication over time, expressed as long-term transmit power constraints.

Different from the standard offline model training for FL that does not consider the wireless communication layer, our trained models are adaptive to the time-varying channel states.

This work has been funded in part by Ericsson Canada and by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors have provided public access to their code or data at https://github.com/juncheng-wang/INFOCOM2022-OMUAA.

Furthermore, we analyze the mutual impact between computation and communication over time to derive performance bounds for our proposed algorithm. Specifically, the main contributions of this paper are as follows:

- We formulate the above system of FL with analog aggregation over noisy wireless fading channels as an online optimization problem. Our optimization objective is the accumulated training loss at the edge server, subject to individual long-term transmit power constraints at the mobile devices. Thus, we consider both the computation and communication metrics. To the best of our knowledge, joint online optimization of model training and analog aggregation has not been studied in the literature.
- We propose an efficient online algorithm, termed Online Model Updating with Analog Aggregation (OMUAA), which dynamically integrates FL, over-the-air computation, and transmit power allocation over time. The local models yielded by OMUAA are adaptive to the dynamic fluctuation of channel states while accounting for the transmit power budget of the mobile devices. Furthermore, they are in closed forms and thus have low computational complexity.
- We analyze the mutual impact between model training and analog aggregation, and their effect on the performance of OMUAA over time. Our analysis shows that OMUAA achieves $\mathcal{O}((1+\rho^2+\Pi_T\rho)\epsilon)$ optimality gap with $\mathcal{O}(\frac{1}{\epsilon^2})$ convergence time for any approximation level ϵ, and $\mathcal{O}((1+\rho^2)\epsilon)$ long-term power constraint violation with $\mathcal{O}(\frac{1}{\epsilon^3})$ convergence time, where ρ is a measure of channel noise and Π_T represents the accumulated variation of the optimal global models in T iterations over noiseless channels.
- We study the impact of system parameters on the performance of OMUAA, by experimenting with real-world image classification datasets, under typical Long-Term Evolution (LTE) network settings. We demonstrate substantial performance advantage of OMUAA over the known best alternatives under different scenarios.

The rest of this paper is organized as follows. In Section II, we present the related work. Section III describes the system model and problem formulation. In Section IV, we present OMUAA. Performance bounds are provided in Section V. Simulation results are presented in Section VI, followed by concluding remarks in Section VII.

II. Related Work

In this section, we survey existing works on FL in wireless edge networks.

A. FL with Error-Free Wireless Communication

Early works on FL at the edge assume error-free, *i.e.,* digital error-control coded transmission (see [31] and references therein). For example, [32] proposed adaptive global model aggregation under resource constraints for FL. The performance trade-offs between computation and communication were investigated in [33] and [34], using conventional orthogonal multiple access. Differential privacy in federated learning was considered in [35]. None of these solutions are applicable to FL with analog aggregation.

B. FL with Analog Aggregation

To further reduce the communication latency and improve bandwidth efficiency, [17]-[19] exploited the superposition property of a multiple access channel to perform analog model aggregation in FL. In [17], truncated local model parameters were scheduled for aggregation based on the channel condition. Receiver beamforming design was studied in [18] to maximize the number of mobile devices for model aggregation at each iteration. In [19], the convergence of an analog model aggregation algorithm was studied for strongly convex loss functions. Other recent works focused on analog gradient aggregation in FL [20]-[26]. Gradient quantization and sparsification were exploited for compressed analog aggregation in [20] and [21] over static and fading multiple access channels, respectively. The convergence of iterative analog gradient aggregation was studied in [22] and [23] with sparsified and full gradients, respectively. Power allocation was investigated in [24] to achieve differential privacy. Gradient statistics aware power control was proposed in [25] for aggregation error minimization. In [26], the aggregation error caused by noisy channel and gradient compression was minimized through power allocation at each iteration.

The above works all separately optimize model training and analog aggregation at each iteration. In contrast, in this work we propose OMUAA to jointly optimize model training and analog aggregation. Furthermore, we consider an online optimization framework that is adaptive to the unpredictable channel fluctuation over time.

C. Online Convex Optimization and Lyapunov Optimization

Because of the dynamic nature of iterative model training and analog aggregation over time-varying channels, a part of our solution resembles existing concepts of online convex optimization (OCO) [36] and Lyapunov optimization [37]. These techniques have been applied to solve various online problems in wireless systems. The standard OCO framework mainly concerns delayed information feedback, which is inherently different from the joint online optimization framework of this work. In particular, [38] proved that no OCO algorithm can simultaneously provide $\mathcal{O}(\epsilon)$ optimality gap and $\mathcal{O}(\epsilon)$ long-term time-varying constraint violation, which OMUAA can achieve (see Section V). Under the standard iterative Lyapunov optimization framework, an upper bound of the weighted sum of loss and constraint functions is minimized at each iteration. However, for machine learning tasks, this often means finding the optimal model, which is difficult in general. Furthermore, the standard Lyapunov optimization requires centralized implementation, which does not apply to FL based on local data.

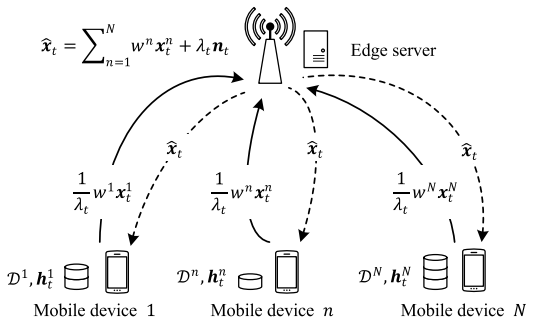

Fig. 1. An illustration of federated learning at wireless edge.

III. System Model and Problem Formulation

A. Learning Objective

We consider a wireless edge network comprising N mobile devices and an edge server as shown in Fig. 1. Each mobile device n collects its local training dataset denoted by $\mathcal{D}^n$. The i-th data sample in $\mathcal{D}^n$ is represented by $(\mathbf{u}^{n,i}, v^{n,i})$, where $\mathbf{u}^{n,i}$ is a data feature vector and $v^{n,i}$ is the true label for this data sample. Based on the local training datasets $\{\mathcal{D}^n\}$, the objective of learning is to train a global model $\mathbf{x} \in \mathbb{R}^d$, which predicts the true labels of data feature vectors.

We define a sample-wise convex and differentiable training loss function $l(\mathbf{x}; \mathbf{u}^{n,i}, v^{n,i}) : \mathbb{R}^d \to \mathbb{R}$ associated with every data sample. The training loss function is generally defined to represent the training error. For example, it can be defined as the logistic regression to measure the prediction accuracy on data feature vector $\mathbf{u}^{n,i}$ with respect to (w.r.t.) its true label $v^{n,i}$.

The *local* training loss function $f^n(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ for each mobile device n is defined as the averaged training loss incurred by the local dataset $\mathcal{D}^n$, given by

$$f^n(\mathbf{x}) = \frac{1}{|\mathcal{D}^n|} \sum_{i=1}^{|\mathcal{D}^n|} l(\mathbf{x}; \mathbf{u}^{n,i}, v^{n,i}) \tag{1}$$

where $|\mathcal{D}^n|$ is the cardinality of $\mathcal{D}^n$. Let $\mathcal{D} = \bigcup_{n=1}^{N} \{\mathcal{D}^n\}$ denote the global dataset with $|\mathcal{D}| = \sum_{n=1}^{N} |\mathcal{D}^n|$. The *global* training loss function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$f(\mathbf{x}) = \sum_{n=1}^{N} w^n f^n(\mathbf{x}) \tag{2}$$

where $w^n = \frac{|\mathcal{D}^n|}{|\mathcal{D}|}$ is the weight on mobile device n, and we have $\sum_{n=1}^{M} w^n = 1$. This is equivalent to the averaged training loss incurred by the global dataset $\mathcal{D}$. The learning objective is to find an optimal global model $\mathbf{x}^\star$ that solves the following optimization problem

$$\min_{\mathbf{x}} \quad f(\mathbf{x}). \tag{3}$$

In traditional centralized machine learning, the edge server would compute $\mathbf{x}^\star$ after collecting all the local training datasets $\{\mathcal{D}^n\}$. However, such a centralized approach is undesirable, as it incurs a large amount of communication overhead, and it can cause privacy issues.

B. Federated Learning with Over-the-Air Analog Aggregation

The standard FL scheme can be seen as an iterative distributed learning process with an aim to approach $\mathbf{x}^\star$ [6], [7]. It alternates between local and global model updates. At the t-th iteration, each mobile device n updates its local model, denoted by $\mathbf{x}_t^n \in \mathbb{R}^d$. The edge server computes the weighted sum of the local models to update its global model. The original FL does not consider the physical wired or wireless communication layer. Thus, under the idealized *noiseless* scenario, the global model would be computed at the edge server as

$$\mathbf{x}_t = \sum_{n=1}^{N} w^n \mathbf{x}_t^n. \tag{4}$$

In the wireless environment, (4) may be efficiently computed over the air, *i.e.,* through analog aggregation over a multiple access channel [27]-[30]. Such analog aggregation scheme exploits the superposition property of a multiple access channel to compute the target function over the air through concurrent transmission of distributed data. It was originally proposed for analog network coding [27] and was recently extended to FL [17]-[26] assuming perfect synchronization. We make the same assumption in this work. Further studies on relaxing the synchronization requirement in analog aggregation can be found in [29] and [30], which are outside the scope of this work.

Note that the local model $\mathbf{x}_t^n$ cannot be directly transmitted to the edge server, since its values can be too large or too small, resulting in very high transmit power or severe noise pollution. Furthermore, due to the noisy and fading nature of wireless channels, the local models $\{\mathbf{x}_t^n\}$ need to be carefully pre-processed at the mobile devices in order to recover the desired global model $\mathbf{x}_t$ in (4) at the edge server. Let $\mathbf{s}_t^n \in \mathbb{C}^d$ be the transmitted signal vector generated by mobile device n at the t-th iteration, which carries the information of $\mathbf{x}_t^n$. Each entry of $\mathbf{s}_t^n$ is sent using one orthogonal channel that is created through division by frequency or time.[1]

We model the channel between the N mobile devices and the edge server as a noisy wireless fading multiple access channel. Let $\mathbf{h}_t^n = [h_t^{n,1}, \ldots, h_t^{n,d}]^T \in \mathbb{C}^d$ be the channel state vector between mobile device n and the edge server at the t-th iteration. As in [17], [20]-[22], [25], we assume a block fading channel model, where $\mathbf{h}_t^n$ over iteration t is independent and identically distributed (i.i.d.). The distribution of $\mathbf{h}_t^n$ is unknown and can be arbitrary. We assume the local channel state information (CSI) is available at each mobile device [17]-[26]. We note that this channel model is suitable for either single-antenna or multi-antenna communication.

The received signal vector $\mathbf{y}_t \in \mathbb{C}^d$ at the edge server is given by

$$\mathbf{y}_t = \sum_{n=1}^{N} \mathbf{h}_t^n \circ \mathbf{s}_t^n + \mathbf{z}_t = \frac{1}{\lambda_t} \sum_{n=1}^{N} w^n \mathbf{x}_t^n + \mathbf{z}_t. \tag{5}$$

[1]The proposed method and analysis in this work can be easily extended to any form of orthogonal channels. Later in Section VI, we divide the system bandwidth over both frequency and time under typical LTE settings.

where $\mathbf{a}\circ\mathbf{b}$ represents entry-wise product, $\mathbf{z}_t \in \mathbb{C}^d$ is the noise vector, and

$$\mathbf{s}_t^n = \frac{1}{\lambda_t} w^n \mathbf{b}_t^n \circ \mathbf{x}_t^n \tag{6}$$

is the transmitted signal vector with λ_t being a power-scaling factor and $\mathbf{b}_t^n = [\frac{h_t^{n,1}}{|h_t^{n,1}|^2}, \ldots, \frac{h_t^{n,d}}{|h_t^{n,d}|^2}]^T \in \mathbb{C}^d$ being the entry-wise channel inversion vector w.r.t. $\mathbf{h}_t^n$. The design of a common λ_t among the N mobile devices at each iteration t was studied in [17], [18], [20], [21], [23]-[26], and is outside the scope of this paper. An important special case is when λ_t is fixed over all iterations t. This can save a large amount of communication overhead, between the mobile devices and the edge server, that is required to agree on a common λ_t at each iteration t before the signal transmission. We assume λ_t depends on the underlying system states.

The edge server scales $\mathbf{y}_t$ and recovers a *noisy* version of the global model $\mathbf{x}_t$ in (4), given by

$$\hat{\mathbf{x}}_t = \Re\{\lambda_t \mathbf{y}_t\} = \mathbf{x}_t + \lambda_t \mathbf{n}_t \tag{7}$$

where $\Re\{\mathbf{a}\}$ denotes the real part of complex vector $\mathbf{a}$ and $\mathbf{n}_t \triangleq \Re\{\mathbf{z}_t\}$.[2] The edge server then broadcasts $\hat{\mathbf{x}}_t$ to all the N mobile devices. As in [17]-[26], we assume that the edge server uses coded digital communication in a separate downlink channel, such that $\hat{\mathbf{x}}_t$ can be received by all the mobile devices in an error-free fashion, before the next iteration.

In the standard error-free FL, the local model $\mathbf{x}_t^n$ is updated via local gradient descent, given by

$$\mathbf{x}_t^n = \mathbf{x}_{t-1} - \alpha \nabla f^n(\mathbf{x}_{t-1}) \tag{8}$$

where $\alpha > 0$ is a step-size parameter. This is equivalent to solving the following optimization problem:

$$\min_{\mathbf{x}} \quad \langle \nabla f^n(\mathbf{x}_{t-1}), \mathbf{x} - \mathbf{x}_{t-1} \rangle + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{x}_{t-1}\|^2 \tag{9}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle$ represents the inner product of vectors $\mathbf{a}$ and $\mathbf{b}$. All existing works on FL with analog aggregation [17]-[26] adopt the above local model updating scheme by simply replacing $\mathbf{x}_{t-1}$ with the received noisy version $\hat{\mathbf{x}}_{t-1}$, and then they *separately* optimize the analog aggregation at each iteration t. In this work, we consider a joint optimization approach to account for the impacts of analog aggregation, including communication error, channel fading, and power allocation, on model training.

Remark 1. One may implement stochastic gradient descent by sampling a batch dataset $\mathcal{B}_t^n \subseteq \mathcal{D}^n$ at each iteration t [8]-[16]. In this work, we focus on the aggregation error caused by the noisy wireless fading channels, and therefore we will initially assume as an example full gradient descent using a fixed local dataset $\mathcal{D}^n$ at each iteration t. This is commonly adopted in prior works [17]-[25]. Our performance analysis can be easily extended to the case of batch gradient descent. Furthermore, in Section VI, we numerically evaluate the proposed algorithm with batch datasets.

[2]We use the real part of the channels to transmit the real vector $\mathbf{x}_t$. The derivations can be easily extended to utilize both the real and imaginary parts of the channels by separating $\mathbf{x}_t$ into half as in [21], without major technical alternation.

C. Problem Formulation

We aim to jointly optimize model training and analog aggregation over time. Due to the time-varying channel states, our objective is to minimize the time-averaged global loss, *i.e.,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{f(\hat{\mathbf{x}}_t)\} \tag{10}$$

where T is the total number of iterations and the expectation is taken over the randomness of the channel states. In steady state, the accumulated training loss $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{f(\hat{\mathbf{x}}_t)\}$ over T iterations approaches the training loss $\mathbb{E}\{f(\hat{\mathbf{x}}_T)\}$ at the T-th iteration.

We assume the following long-term transmit power constraint at each mobile device n:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{\|\mathbf{s}_t^n\|^2\right\} \le \bar{P}^n, \quad \forall n \tag{11}$$

where $\bar{P}^n$ is the average transmit power limit. We also consider possible short-term constraints on the local models, given by $\mathcal{X} = \{\mathbf{x} : -\mathbf{x}_{\max} \preceq \mathbf{x} \preceq \mathbf{x}_{\max}\} \subseteq \mathbb{R}^d$, where $\preceq$ represents entry-wise inequality and $\mathbf{x}_{\max} = x_{\max}\mathbf{1}$ with $x_{\max}$ being the maximum model value and $\mathbf{1}$ being a vector of all 1's.

We aim at selecting a sequence of local models $\{\mathbf{x}_t^n\}$ from $\mathcal{X}$ to minimize the accumulated training loss yielded by the noisy global model $\{\hat{\mathbf{x}}_t\}$ after analog aggregation at the edge server, while ensuring that the individual long-term transmit power constraints at the mobile devices are satisfied. This leads to the following stochastic optimization problem:

$$\begin{aligned} \mathbf{P1}: \quad \min_{\{\mathbf{x}_t^n \in \mathcal{X}\}} \quad & \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{f(\hat{\mathbf{x}}_t)\} \\ \text{s.t.} \quad & \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{g_t^n(\mathbf{x}_t^n)\} \le 0, \quad \forall n \end{aligned} \tag{12}$$

where

$$g_t^n(\mathbf{x}) = \frac{(w^n)^2}{\lambda_t^2} \|\mathbf{b}_t^n \circ \mathbf{x}\|^2 - \bar{P}^n, \tag{13}$$

the expectation is taken over the randomness of the channel states. From (6) and (13), it is easy to see that (12) is equivalent to (11).

Note that **P1** is a stochastic optimization problem due to the random channel states. In **P1**, the training loss $f(\hat{\mathbf{x}}_t)$ is determined by the noisy global model $\hat{\mathbf{x}}_t$ aggregated over the air from the local models $\{\mathbf{x}_t^n\}$. The long-term transmit power violation $g_t^n(\mathbf{x}_t^n)$ depends on both the local channel state $\mathbf{h}_t^n$ and the local model $\mathbf{x}_t^n$. Because of the need for signal amplification at the receiver, as shown in (7), a small transmit power amplifies the channel noise in analog aggregation, which in turn deteriorates the model training. Due to such coupling of model training and analog aggregation caused by

Algorithm 1 OMUAA: Mobile device n's algorithm

1: Initialize $\mathbf{x}_1^n = \hat{\mathbf{x}}_1 = \mathbf{0}$ and the virtual queue $Q_1^n = 0$.
For each t, do the following:
2: Update local model $\mathbf{x}_t^n$ by solving $\mathbf{P2}^n$ via (17).
3: Update local virtual queue Q_t^n via (14).
4: Transmit signals $\mathbf{s}_t^n$ in (6) to the edge server.

Algorithm 2 OMUAA: Edge server's algorithm

1: Initialize and broadcast step-size parameters $\alpha, \gamma > 0$.
For each t, do the following:
2: Receive signals $\mathbf{y}_t$ in (5) over the air.
3: Update noisy global model $\hat{\mathbf{x}}_t$ in (7)
4: Broadcast $\hat{\mathbf{x}}_t$ to all mobile devices.

wireless fading channels, solving **P1** requires simultaneous consideration for computation and communication. Furthermore, compared with the one-shot optimization problem (3) considered in standard FL, the additional long-term transmit power constraints in (12) of **P1** require a more complicated online algorithm, especially since the channel state varies over time. In this work, without needing to know the channel distribution, we aim to develop an online algorithm based on the local channel state $\mathbf{h}_t^n$ at each mobile device n, to compute a solution $\{\mathbf{x}_t^n\}$ to **P1**.

IV. Online Model Updating with Analog Aggregation (OMUAA)

In this section, we present the design details of OMUAA. Existing algorithms for FL in wireless networks alternatingly optimize model training and wireless transmission at each iteration. In contrast, OMUAA jointly optimize model training and analog aggregation, while considering the mutual impact between them over time. The local models yielded by OMUAA are adaptive to the time-varying channel states. Furthermore, the local models can be obtained in closed-forms with low computational complexity. In the following, we present OMUAA algorithms at the mobile devices and the edge server.

We first introduce a virtual queue Q_t^n at each mobile device n to account for the long-term transmit power constraint (12) in **P1**. It has the following updating rule:

$$Q_t^n = \max\{Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), 0\}, \quad \forall n, \quad \forall t. \tag{14}$$

The role of Q_t^n is similar to a Lagrangian multiplier for **P1** or a backlog queue for the long-term constraint violation, which is a technique used in Lyapunov optimization [37]. However, we note that, although a part of our performance bound analysis for OMUAA borrows techniques from Lyapunov drift analysis, as explained in Section II, OMUAA is structurally different from Lyapunov optimization.

Using the virtual queue in (14), we convert **P1** into solving a per-iteration optimization problem at each mobile device n, given by

$$\mathbf{P2}^n: \min_{\mathbf{x}\in\mathcal{X}} \quad \langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x} - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2 + \gamma Q_{t-1}^n g_t^n(\mathbf{x})$$

where $\alpha, \gamma > 0$ are step-size parameters. Note that $\mathbf{P2}^n$ is a per-device per-iteration optimization problem using the current local CSI $\mathbf{h}_t^n$ and the virtual queue length Q_{t-1}^n and is subject to the short-term constraints only. Compared with the original **P1**, the long-term transmit power constraint is converted to the third term of the objective function in $\mathbf{P2}^n$, which is a penalization term on $g_t^n(\mathbf{x})$. Note that different from problem (9), which does not consider the communication noise, the local gradient $\nabla f^n(\hat{\mathbf{x}}_{t-1})$ in $\mathbf{P2}^n$ is evaluated using the noisy global model $\hat{\mathbf{x}}_{t-1}$.

In OMUAA, we perform local model updates on $\{\mathbf{x}_t^n\}$ by solving $\mathbf{P2}^n$. Note that the long-term transmit power constraint function $g_t^n(\mathbf{x})$ is convex and the feasible set $\mathcal{X}$ is affine. Furthermore, due to the regularization term $\frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2$, $\mathbf{P2}^n$ is a strongly convex optimization problem. In the following, we present a closed-form solution to $\mathbf{P2}^n$.

We observe that the gradient of the objective function of $\mathbf{P2}^n$ w.r.t. $\mathbf{x}$ is

$$\nabla f^n(\hat{\mathbf{x}}_{t-1}) + \frac{1}{\alpha}(\mathbf{x} - \hat{\mathbf{x}}_{t-1}) + \boldsymbol{\theta}_t^n \circ \mathbf{x} \tag{15}$$

where $\boldsymbol{\theta}_t^n \in \mathbb{R}^d$ with the i-th entry given by

$$\theta_t^{n,i} = \frac{2\gamma Q_{t-1}^n (w^n)^2}{\lambda_t^2 |h_t^{n,i}|^2}. \tag{16}$$

The optimal solution to $\mathbf{P2}^n$ can be obtained by setting the gradient in (15) to zero to solve for $\mathbf{x}$ and then projecting it onto the affine set $\mathcal{X}$. Thus, the local model update $\mathbf{x}_t^n$ can be computed in a closed form, given by

$$\mathbf{x}_t^n = \left[(\mathbf{1} + \boldsymbol{\theta}_t^n)^{-1} \circ (\hat{\mathbf{x}}_{t-1} - \alpha \nabla f^n(\hat{\mathbf{x}}_{t-1}))\right]_{-\mathbf{x}_{\max}}^{\mathbf{x}_{\max}} \tag{17}$$

where $\mathbf{a}^{-1}$ is the entry-wise inverse operator and $[\mathbf{x}]_{\mathbf{a}}^{\mathbf{b}} = \min\{\mathbf{b}, \max\{\mathbf{x}, \mathbf{a}\}\}$ is the entry-wise projection operator. Note that the minimization in $\mathbf{P2}^n$ is entry-wise in $\mathbf{x}_t^n$.

Compared with the standard local gradient descent update for error-free FL in (8), the local model update in (17) is scaled entry-wise by a factor $\frac{1}{1+\theta_t^{n,i}}$ that depends on the ratio of the long-term transmit power constraint violation measured by Q_{t-1}^n and the individual channel power $|h_t^{n,i}|^2$. The local model $\mathbf{x}_t^n$ is updated roughly the same as the error-free case (*i.e.,* model update is scaled close to 1) when the channels are strong, but its values decrease when the queue length Q_{t-1}^n is relatively large compared with the channel gain. Therefore, the local model update by OMUAA is both channel- and power-aware. In Section V, we will show that the update sequence $\{\mathbf{x}_t^n\}$ further satisfies the long-term transmit power constraint.

To summarize, in OMUAA, each mobile device n first initializes the local models $\mathbf{x}_1^n = \hat{\mathbf{x}}_1 = \mathbf{0}$ and the local virtual queue $Q_1^n = 0$. At each iteration t, after obtaining its own local CSI $\mathbf{h}_t^n$, each mobile device n updates $\mathbf{x}_t^n$ by solving $\mathbf{P2}^n$ via (17) and then updates Q_t^n via (14). The mobile device then transmits signals $\mathbf{s}_t^n$ in (6) to the edge server. We summarize the mobile device n's algorithm in Algorithm 1.

At each iteration t, the edge server receives signals $\mathbf{y}_t$ in (5) through analog aggregation of the signals $\{\mathbf{s}_t^n\}$ transmitted

by the N mobile devices. The edge server recovers a noisy global model $\hat{\mathbf{x}}_t$ in (7), which is then broadcasted to all mobile devices. We summarize the edge server's algorithm in Algorithm 2. The choice of step-size parameters α, γ will be discussed in Section V, after deriving the performance bounds.

Remark 2. The computational complexity of calculating the local gradient $\nabla f^n(\hat{\mathbf{x}}_{t-1})$ in (17) depends on the machine learning task. Compared with the local model update for FL in (8), the additional computational complexity in (17) is in computing the virtual queue Q^n_{t-1} and the factor $\boldsymbol{\theta}^n_t$, both are in the order of $\mathcal{O}(d)$. Therefore, the local model update in (17) has low computational complexity.

V. Performance Bounds

In this section, we derive the performance bounds of OMUAA. We develop new techniques, particularly to account for the mutual impact of model training and analog aggregation over time. We first state the following assumptions, which are required for our mathematical analysis but are easily satisfied in practical systems.

Assumption 1. The loss function $f^n(\mathbf{x})$ has bounded gradient $\nabla f^n(\mathbf{x})$: $\exists\ D > 0$ s.t.,

$$\|\nabla f^n(\mathbf{x})\| \leq D, \quad \forall \mathbf{x} \in \mathbb{R}^d, \quad \forall n. \tag{18}$$

Assumption 2. The constraint function $g^n_t(\mathbf{x})$ is bounded: $\exists\ G > 0$, s.t.,

$$|g^n_t(\mathbf{x})| \leq G, \quad \forall \mathbf{x} \in \mathcal{X}, \quad \forall n, \quad \forall t. \tag{19}$$

Assumption 3. The communication noise $\mathbf{n}_t$ is bounded: $\exists\ \rho > 0$, s.t.,

$$\|\mathbf{n}_t\| \leq \rho, \quad \forall t. \tag{20}$$

A. Bound for the Accumulated Training Loss

Define $L^n_t \triangleq \frac{1}{2}(Q^n_t)^2$ as a quadratic Lyapunov function and $\Delta^n_t \triangleq L^n_t - L^n_{t-1}$ as the corresponding Lyapunov drift for each mobile device n. We first provide an upper bound on Δ^n_t in the following lemma. The proof follows from the virtual queue dynamics in (14) and is omitted for brevity.

Lemma 1. The Lyapunov drift is upper bounded as follows:

$$\Delta^n_t \leq \frac{1}{2}G^2 + Q^n_{t-1} g^n_t(\mathbf{x}^n_t), \quad \forall n. \tag{21}$$

We also require the following lemma from [36, Lemma 2.8].

Lemma 2. (Lemma 2.8, [36]) Let $\mathcal{X} \in \mathbb{R}^d$ be a nonempty convex set. Let $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ be a $\frac{1}{\alpha}$-strongly convex function over $\mathcal{X}$ w.r.t. a norm $\|\cdot\|$. Let $\mathbf{z} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x})\}$. Then, for any $\mathbf{y} \in \mathcal{X}$, we have $f(\mathbf{z}) \leq f(\mathbf{y}) - \frac{1}{2\alpha}\|\mathbf{y} - \mathbf{z}\|^2$.

For i.i.d. channel state $\mathbf{h}_t$, there exists a stationary randomized optimal global solution $\mathbf{x}^\star_t$ to **P1** over noiseless channels, which depends only on the (unknown) distribution of $\mathbf{h}^n_t$, and achieves the minimum objective value (*i.e.,* the minimum accumulated training loss) of **P1**, denoted by $f^\star$ [37]. Using the results in Lemmas 1 and 2, the following theorem provides an upper bound on the accumulated training loss by OMUAA over noisy channels.

Theorem 3. For any $\alpha, \gamma > 0$, regardless of the channel distribution, the accumulated training loss yielded by OMUAA is upper bounded by

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{f(\hat{\mathbf{x}}_t)\} \leq f^\star + \frac{D^2\alpha}{2} + \frac{G^2\gamma}{2} + \frac{R^2 + \rho^2\Lambda_{2,T} + 2R\rho\Lambda_T}{2\alpha T} + \frac{2R + \lambda_{\max}\rho}{T}\left(D + \frac{\Pi_T}{\alpha}\right) \tag{22}$$

where $R = \sqrt{d}x_{\max}$, $\lambda_{\max} = \max\{\lambda_t, \forall t\}$, $\Lambda_T = \sum_{t=1}^T \mathbb{E}\{\lambda_t\}$, $\Lambda_{2,T} = \sum_{t=1}^T \mathbb{E}\{\lambda_t^2\}$, and $\Pi_T = \sum_{t=1}^T \mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}^\star_{t+1}\|\}$.[3]

Proof: The objective function in $\mathbf{P2}^n$ is $\frac{1}{\alpha}$-strongly convex over $\mathcal{X}$ w.r.t. Euclidean norm $\|\cdot\|$ due to the regularization term $\frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2$. Since $\mathbf{x}^n_t$ minimizes $\mathbf{P2}^n$ over $\mathcal{X}$ for any t, from Lemma 2, we have

$$\begin{aligned}
&\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\|^2 + \gamma Q^n_{t-1} g^n_t(\mathbf{x}^n_t) \\
&\quad \leq \langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}^\star_t - \hat{\mathbf{x}}_{t-1}\rangle + \gamma Q^n_{t-1} g^n_t(\mathbf{x}^\star_t) \\
&\quad + \frac{1}{2\alpha}\left(\|\mathbf{x}^\star_t - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2\right).
\end{aligned} \tag{23}$$

Now, we bound the third term on the right-hand side (RHS) of (23). We have

$$\begin{aligned}
&\|\mathbf{x}^\star_t - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2 \\
&\overset{(a)}{\leq} \|\mathbf{x}^\star_t - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}^\star_{t+1} - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}^\star_t - \mathbf{x}^\star_{t+1}\|^2 \\
&\quad + 2\|\mathbf{x}^\star_{t+1} - \hat{\mathbf{x}}_t\|\|\mathbf{x}^\star_t - \mathbf{x}^\star_{t+1}\| + \|\mathbf{x}^\star_t - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2 \\
&\overset{(b)}{\leq} \psi_t + 2\|\mathbf{x}^\star_{t+1} - \hat{\mathbf{x}}_t\|\pi_t + \phi^n_t
\end{aligned} \tag{24}$$

where (a) is because $\|\mathbf{a}+\mathbf{b}\|^2 \geq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\|\mathbf{b}\|$; and (b) follows from defining $\psi_t \triangleq \|\mathbf{x}^\star_t - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}^\star_{t+1} - \hat{\mathbf{x}}_t\|^2$, $\pi_t \triangleq \|\mathbf{x}^\star_t - \mathbf{x}^\star_{t+1}\|$, and $\phi^n_t \triangleq \|\mathbf{x}^\star_t - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2$.

Substituting (24) into (23), adding $f^n(\hat{\mathbf{x}}_{t-1})$ on both sides, applying the first order condition of convexity

$$f^n(\hat{\mathbf{x}}_{t-1}) + \langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}^\star_t - \hat{\mathbf{x}}_{t-1}\rangle \leq f^n(\mathbf{x}^\star_t)$$

to its RHS, and rearranging terms, we have

$$\begin{aligned}
&f^n(\hat{\mathbf{x}}_{t-1}) - f^n(\mathbf{x}^\star_t) \\
&\quad \leq -\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\rangle - \frac{1}{2\alpha}\|\mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\|^2 \\
&\quad + \gamma Q^n_{t-1} g^n_t(\mathbf{x}^\star_t) - \gamma Q^n_{t-1} g^n_t(\mathbf{x}^n_t) \\
&\quad + \frac{1}{2\alpha}(\psi_t + 2\|\mathbf{x}^\star_{t+1} - \hat{\mathbf{x}}_t\|\pi_t + \phi^n_t).
\end{aligned} \tag{25}$$

We now bound the RHS of (25). Completing the square and noting that $\nabla f(\mathbf{x})$ is bounded in (18), we have

$$\begin{aligned}
&-\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\rangle - \frac{1}{2\alpha}\|\mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\|^2 \\
&\quad \leq \frac{\alpha}{2}\|\nabla f^n(\hat{\mathbf{x}}_{t-1})\|^2 \leq \frac{D^2\alpha}{2}.
\end{aligned} \tag{26}$$

[3]Note that Π_T is the accumulated variation of the optimal global model over noiseless channels.

From $\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$, the definition of $\hat{\mathbf{x}}_t$ in (7), $\mathcal{X}$ being bounded, *i.e.,*

$$\|\mathbf{x}\| \leq R, \quad \forall x \in \mathcal{X}, \tag{27}$$

where $R = \sqrt{d}x_{\max}$, and $\mathbf{n}_t$ being bounded in (20), we have

$$\|\mathbf{x}^\star_{t+1} - \hat{\mathbf{x}}_t\| \leq \|\mathbf{x}^\star_{t+1}\| + \|\mathbf{x}_t\| + \|\lambda_t \mathbf{n}_t\| \leq 2R + \lambda_{\max}\rho. \tag{28}$$

Substituting (21), (26), and (28) into (25), multiplying both sides by w^n, summing over n, and taking expectation, we have

$$\begin{aligned}
&\mathbb{E}\{f(\hat{\mathbf{x}}_{t-1})\} - \mathbb{E}\{f(\mathbf{x}^\star_t)\} \\
&\leq \frac{D^2\alpha}{2} + \gamma\left(\sum_{n=1}^N w^n \mathbb{E}\{Q^n_{t-1} g^n_t(\mathbf{x}^\star_t)\} + \frac{1}{2}G^2 - \sum_{n=1}^N w^n \mathbb{E}\{\Delta^n_t\}\right) \\
&\quad + \frac{1}{2\alpha}\left(\mathbb{E}\{\psi_t\} + 2(2R+\lambda_{\max}\rho)\mathbb{E}\{\pi_t\} + \sum_{n=1}^N w^n \mathbb{E}\{\phi^n_t\}\right). \qquad (29)
\end{aligned}$$

From $\mathbf{x}^\star_t$ being independent of $Q^n_{t-1} \geq 0$, and $\mathbb{E}\{g^n_t(\mathbf{x}^\star_t)\} \leq 0$, we have $\mathbb{E}\{Q^n_{t-1} g^n_t(\mathbf{x}^\star_t) | Q^n_{t-1}\} = Q^n_{t-1}\mathbb{E}\{g^n_t(\mathbf{x}^\star_t)\} \leq 0$. It then follows from the iterated law of expectation that $\mathbb{E}\{Q^n_{t-1} g^n_t(\mathbf{x}^\star_t)\} = \mathbb{E}\{\mathbb{E}\{Q^n_{t-1} g^n_t(\mathbf{x}^\star_t)|Q^n_{t-1}\}\} \leq 0$. Substituting it into (29) and summing from $t = 2$, we have

$$\begin{aligned}
&\sum_{t=1}^{T-1} \mathbb{E}\{f(\hat{\mathbf{x}}_t)\} - \sum_{t=2}^{T} \mathbb{E}\{f(\mathbf{x}^\star_t)\} \\
&\leq \frac{D^2\alpha}{2}T + \frac{G^2\gamma}{2}T - \gamma\sum_{t=2}^{T}\sum_{n=1}^{N} w^n \mathbb{E}\{\Delta^n_t\} + \frac{1}{2\alpha}\sum_{t=2}^{T}\mathbb{E}\{\psi_t\} \\
&\quad + \frac{2R+\lambda_{\max}\rho}{\alpha}\sum_{t=2}^{T}\mathbb{E}\{\pi_t\} + \frac{1}{2\alpha}\sum_{t=2}^{T}\sum_{n=1}^{N} w^n \mathbb{E}\{\phi^n_t\}. \qquad (30)
\end{aligned}$$

We now bound the RHS of (30). From the definition of Δ_t, $Q^n_1 = 0$, and $Q^n_t \geq 0, \forall t$, we have

$$-\sum_{t=2}^{T}\mathbb{E}\{\Delta^n_t\} = \frac{1}{2}\mathbb{E}\{(Q^n_1)^2\} - \frac{1}{2}\mathbb{E}\{(Q^n_T)^2\} \leq 0. \tag{31}$$

Noting that ψ_t is a telescoping term, $\hat{\mathbf{x}}_1 = \mathbf{0}$ by initialization, and $\|\mathbf{x}^\star_t\| \leq R, \forall t$, we have

$$\sum_{t=2}^{T}\mathbb{E}\{\psi_t\} = \mathbb{E}\{\|\mathbf{x}^\star_2 - \hat{\mathbf{x}}_1\|^2\} - \mathbb{E}\|\mathbf{x}^\star_{T+1} - \hat{\mathbf{x}}_T\|^2\} \leq R^2. \tag{32}$$

For the last term on the RHS of (30), we have

$$\begin{aligned}
&\sum_{t=2}^{T}\sum_{n=1}^{N} w^n \mathbb{E}\{\phi^n_t\} = \sum_{t=2}^{T}\sum_{n=1}^{N} w^n \mathbb{E}\{\|\mathbf{x}^\star_t - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2\} \\
&\overset{(a)}{\leq} \sum_{t=2}^{T}\sum_{n=1}^{N} w^n (\mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}_t\|^2\} - \mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2\}) \\
&\quad + \sum_{t=2}^{T}\mathbb{E}\{\|\lambda_t \mathbf{n}_t\|^2\} + 2\sum_{t=2}^{T}\mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}_t\|\|\lambda_t \mathbf{n}_t\|\} \\
&\overset{(b)}{\leq} \sum_{t=2}^{T}\mathbb{E}\{\|\lambda_t \mathbf{n}_t\|^2\} + 2\sum_{t=2}^{T}\mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}_t\|\|\lambda_t \mathbf{n}_t\|\} \\
&\overset{(c)}{\leq} \rho^2\Lambda_{2,T} + 2R\rho\Lambda_T \qquad (33)
\end{aligned}$$

where (a) follows form $\|\mathbf{a}+\mathbf{b}\|^2 \leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{a}\|\|\mathbf{b}\|$, (b) is because of the separate convexity of Euclidean norm and the definition of $\mathbf{x}_t$ in (4) such that for any t

$$\begin{aligned}
&\sum_{n=1}^{N} w^n (\mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}_t\|^2\} - \mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2\}) \\
&\quad \leq \sum_{n=1}^{N} w^n \left(\sum_{j=1}^{N}(w^j \mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}^j_t\|^2\}) - \mathbb{E}\{\|\mathbf{x}^\star_t - \mathbf{x}^n_t\|^2\}\right) = 0,
\end{aligned}$$

and (c) follows from $\mathbf{n}_t$ and $\mathcal{X}$ being bounded in (20) and (27), respectively, and the definitions of $\Lambda_{2,T}$ and Λ_T.

Substituting (31)-(33) into (30), noting that $f(\hat{\mathbf{x}}_T) - f(\mathbf{x}^\star_1) \leq \sum_{n=1}^N w^n \langle \nabla f(\hat{\mathbf{x}}_T), \hat{\mathbf{x}}_T - \mathbf{x}^\star_1\rangle \leq D(2R + \lambda_{\max}\rho)$, and from the definition of Π_T and $f^\star$, we have (22). ■

Theorem 3 provides a general bound for the accumulated training loss by OMUAA, for any values of step-size parameters α, γ, and power-scaling factors $\{\lambda_t\}$. The following corollary provides the accumulated training loss by OMUAA when α, γ, and $\{\lambda_t\}$ take specific values. It follows by substituting the corresponding α, γ, and $\{\lambda_t\}$ into the general bound in (22).

Corollary 4. For any $\epsilon > 0$, set $\alpha = \gamma = \epsilon$ and $\lambda_t = \epsilon^2, \forall t$. The accumulated training loss yielded by OMUAA is upper bounded by

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{f(\hat{\mathbf{x}}_t)\} \leq f^\star + \mathcal{O}((1+\rho^2+\Pi_T\rho)\epsilon), \ \forall T \geq \frac{1}{\epsilon^2}. \tag{34}$$

Corollary 4 provides an upper bound on the objective value of **P1** in (34), *i.e.,* the accumulated training loss yielded by the noisy global model. It indicates that for all $T \geq \frac{1}{\epsilon^2}$, the accumulate training loss produced by OMUAA over noisy channels is within $\mathcal{O}((1+\rho^2+\Pi_T\rho)\epsilon)$ to the optimum achieved over noiseless channels. Note that Π_T can be small when the channel state does not vary too drastically over time. In particular, when the channel is static, we have $\Pi_T = 0$.

B. Bound for the Long-Term Transmit Power

We now proceed to provide a performance bound on the individual long-term transmit power constraint violation at each mobile device by OMUAA.

Theorem 5. For any $\alpha, \gamma > 0$, the violation of each individual long-term transmit power constraint is upper bounded by

$$\frac{1}{T}\sum_{t=1}^{T} g^n_t(\mathbf{x}^n_t) \leq \frac{G}{T} + \frac{\alpha\gamma G^2 + 2\alpha DR + (R+\lambda_{\max}\rho)^2}{2\alpha\gamma\bar{P}^n T}, \ \forall n. \tag{35}$$

Proof: Since $\mathbf{x}^n_t$ minimizes the objective function of P2^n over $\mathcal{X}$ for any n, we have

$$\begin{aligned}
&\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x}^n_t - \hat{\mathbf{x}}_{t-1}\|^2 + \gamma Q^n_{t-1} g^n_t(\mathbf{x}^n_t) \\
&\overset{(a)}{\leq} \langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), -\hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\hat{\mathbf{x}}_{t-1}\|^2 - \gamma Q^n_{t-1}\bar{P}^n \qquad (36)
\end{aligned}$$

where (a) follows from $g_t^n(\mathbf{0}) = -\bar{P}^n$. Rearranging terms of (36), we have

$$\gamma Q_{t-1}^n g_t^n(\mathbf{x}_t^n) \leq -\gamma Q_{t-1}^n \bar{P}^n - \langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n \rangle + \frac{1}{2\alpha}\|\hat{\mathbf{x}}_{t-1}\|^2 \overset{(a)}{\leq} -\gamma Q_{t-1}^n \bar{P}^n + DR + \frac{(R+\lambda_{\max}\rho)^2}{2\alpha} \quad (37)$$

where (a) follows from $\nabla f(\mathbf{x})$, $\mathbf{n}_t$, and $\mathcal{X}$ being bounded in (18), (20), and (27), respectively.

Substituting (37) into (21), we have

$$\Delta_t^n \leq -Q_{t-1}^n \bar{P}^n + \frac{G^2}{2} + \frac{DR}{\gamma} + \frac{(R+\lambda_{\max}\rho)^2}{2\alpha\gamma}.$$

Therefore, a sufficient condition for $\Delta_t^n < 0$ is

$$Q_{t-1}^n > \frac{\alpha\gamma G^2 + 2\alpha DR + (R+\lambda_{\max}\rho)^2}{2\alpha\gamma\bar{P}^n}. \quad (38)$$

If (38) holds, we have $Q_t^n < Q_{t-1}^n$, *i.e.,* the virtual queue decreases; otherwise, there is a maximum increase from Q_{t-1}^n to Q_t^n since $Q_t^n - Q_{t-1}^n \leq g_t^n(\mathbf{x}_t^n) \leq G$. Therefore, the virtual queue is bounded for all t by

$$Q_t^n \leq G + \frac{\alpha\gamma G^2 + 2\alpha DR + (R+\lambda_{\max}\rho)^2}{2\alpha\gamma\bar{P}^n}. \quad (39)$$

From the virtual queue dynamics in (14), we have $Q_t^n \geq Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), \forall t$. Summing it from $t = 2$, we have $\sum_{t=2}^T g_t^n(\mathbf{x}_t^n) = \sum_{t=2}^T Q_t^n - Q_{t-1}^n = Q_T^n - Q_1^n = Q_T^n$. Noting that $g_1^n(\mathbf{x}_1^n) = -\bar{P}^n < 0$, we have $\frac{1}{T}\sum_{t=1}^T g_t^n(\mathbf{x}_t^n) \leq \frac{Q_T^n}{T}$. Substituting the virtual queue bound in (39) into this inequality, we have (35). ∎

From the general bound on the violation of individual long-term transmit power constraint provided in Theorem 5, which is for any step-size parameters α, γ, and power-scaling factors $\{\lambda_t\}$, we can derive the following corollary for some specific values of α, γ, and $\{\lambda_t\}$.

Corollary 6. For any $\epsilon > 0$, set $\alpha = \gamma = \epsilon$ and $\lambda_t = \epsilon^2, \forall t$. The individual long-term transmit power constraint violations yielded by OMUAA is upper bounded by

$$\frac{1}{T}\sum_{t=1}^T g_t^n(\mathbf{x}_t^n) \leq \mathcal{O}((1+\rho^2)\epsilon), \quad \forall n, \quad \forall T \geq \frac{1}{\epsilon^3}. \quad (40)$$

In addition to the upper bound on the accumulated training loss in (34), Corollary 6 shows that for each mobile device n, OMUAA guarantees that the deviation from its average transmit power limit $\bar{P}^n$ is within $\mathcal{O}((1+\rho^2)\epsilon)$ if $T \geq \frac{1}{\epsilon^3}$. Finally, we point out that the i.i.d. condition of the channel state can be relaxed to Markovian evolution over time, and similar performance bounds can be obtained by using the extension method discussed in [37].

VI. Simulation Results

To complement the theoretical performance guarantees of OMUAA provided in Section V, we evaluate the performance of OMUAA in edge learning based on real-world image classification datasets, under typical LTE network settings.

A. Simulation Setup

We consider a wireless edge network with one edge server and $N = 10$ mobile devices. Following typical LTE specifications [39], we set noise power spectral density $N_0 = -174$ dBm/Hz and noise figure $N_F = 10$ dB. We consider an orthogonal frequency-division multiplexing system with $S = 500$ subcarriers, each with bandwidth $B_W = 15$ kHz. The fading channel from mobile device n to the edge server at the t-th iteration is modeled as $\mathbf{h}_t^n \sim \mathcal{N}(\mathbf{0}, \beta^n\mathbf{I})$, with β^n representing the large-scale fading variation consisting of the path-loss and shadowing. We model $\beta^n[\text{dB}] = -31.54 - 33\log_{10}(r) - \varphi^n$ [39], where $r = 100$ m is the distance to the edge server, and $\varphi^n \sim \mathcal{N}(0, \sigma_\phi^2)$ is the shadowing with $\sigma_\phi^2 = 8$ dB. We assume each channel is i.i.d. over iteration t. We use a fixed power-scaling factor $\lambda_t = \lambda$ in all simulations.

We use the MNIST dataset [40] for model training and testing. The training dataset $\mathcal{D}$ consists of $|\mathcal{D}| = 6 \times 10^4$ data samples and the test dataset $\mathcal{E}$ has $|\mathcal{E}| = 1 \times 10^4$ data samples. Each data sample $(\mathbf{u}, v)$ represents a labeled image of size 28×28 pixels, *i.e.,* $\mathbf{u} \in \mathbb{R}^{784}$, with $J = 10$ different labels, *i.e.,* $v \in \{1, \ldots, J\}$. We consider the cross-entropy loss for multinomial logistic regression

$$l(\mathbf{x}; \mathbf{u}, v) = -\sum_{j=1}^J 1\{v = j\} \log \frac{\exp(\langle \mathbf{x}[j], \mathbf{u}\rangle)}{\sum_{k=1}^J \exp(\langle \mathbf{x}[k], \mathbf{u}\rangle)} \quad (41)$$

where $\mathbf{x} = [\mathbf{x}[1]^T, \ldots, \mathbf{x}[J]^T]^T$ with $\mathbf{x}[j] \in \mathbb{R}^{784}$ being the model for label j. The entire model $\mathbf{x}$ is thus of dimension $d = 7840$ and is transmitted over $M = \lceil \frac{d}{S} \rceil = 16$ transmission frames over time at each iteration t. We assume the same average transmit power limit at the mobile devices, *i.e.,* $\bar{P}^n = M\bar{P}, \forall n$. We consider non-i.i.d. data distribution, where the local dataset $\mathcal{D}^n$ at each mobile device n only contains data samples of label n. Therefore, the mobile devices do not share data samples of the same labels. We assume each mobile device n samples a batch dataset $\mathcal{B}_t^n \subset \mathcal{D}^n$ consisting of $|\mathcal{B}_t^n| = 20$ data samples at each iteration t. Therefore, the weight of each mobile device n is $w^n = \frac{1}{N}$.

Our performance metrics are the time-averaged test accuracy over $\mathcal{E}$

$$\bar{A}(T) = \frac{1}{T|\mathcal{E}|}\sum_{t=1}^T\sum_{i=1}^{|\mathcal{E}|} 1\left\{\text{argmax}_j \left\{\frac{\exp(\langle \hat{\mathbf{x}}_t[j], \mathbf{u}^i\rangle)}{\sum_{k=1}^J \exp(\langle \hat{\mathbf{x}}_t[k], \mathbf{u}^i\rangle)}\right\} = v^i\right\},$$

and the time-averaged training loss over $\{\mathcal{B}_t^n\}$

$$\bar{f}(T) = \frac{1}{T}\sum_{t=1}^T\sum_{n=1}^N \frac{1}{|\mathcal{B}_t^n|}\sum_{i=1}^{|\mathcal{B}_t^n|} w^n l(\hat{\mathbf{x}}_t; \mathbf{u}_t^{n,i}, v_t^{n,i}).$$

B. Performance Comparison

We compare OMUAA with the following schemes.[4]

- *Error-free FL:* We run the FL scheme that alternates local model update in (8) and global model aggregation in (4) over noiseless channels with batch datasets. This scheme provides a performance upper bound for OMUAA.

[4]We use the same step-size parameters in these schemes as OMUAA.

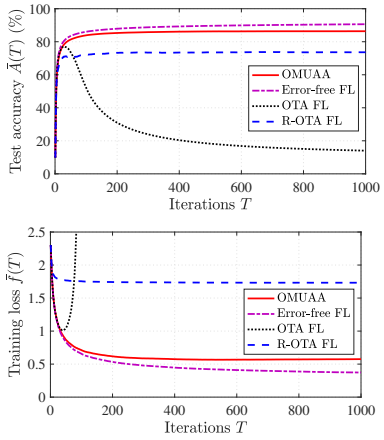

Fig. 2. Test accuracy $\bar{A}(T)$ and training loss $\bar{f}(T)$ vs. iterations T.

- *OTA FL:* We adopt the transmit power control scheme in [20], [21], which are the best existing alternatives that consider over-the-air (OTA) FL with long-term transmit power constraints.[5] In [20] and [21], a time-varying power-scaling factor λ_t is used in (6) to set the transmit power at each mobile device n around a predefined transmit power limit P_t at each iteration t. Since different strategies to set P_t achieve nearly the same performance as shown in [20], we set $P_t = M\bar{P}, \forall t$ as in [21].
- *R-OTA FL:* Based on OTA FL, we add a regularization term $\kappa\|\mathbf{x}\|^2$ to $l(\mathbf{x};\mathbf{u},v)$ in (41), where κ is a tunable parameter. This regularization scheme was adopted in [22]-[24]. We have optimized κ in the presented results.

Fig. 2 shows $\bar{A}(T)$ and $\bar{f}(T)$ versus T with $\bar{P} = 16$ dBm. Despite the presence of communication noise, OMUAA converges quickly and achieves better classification performance compared with OTA FL and R-OTA FL. We observe that the performance of OTA FL deteriorates as T increases. This is because OTA FL uses the power-scaling factor λ_t for transmit power control, which magnifies the communication error $\lambda_t\mathbf{n}_t$ in the global model $\hat{\mathbf{x}}_t$ in (7) when λ_t is large. Since $\hat{\mathbf{x}}_t$ is further used in the training process at the next iteration, there will be severe communication error propagation in the learning process. Adding a regularization term as in R-OTA FL helps minimize $\|\mathbf{x}_t\|^2$ and thus prevents λ_t from being too large. We observe that, with properly tuned κ, R-OTA FL substantially outperforms OTA FL. In comparison, the virtual queue in OMUAA serves as automatically-tuned regularization on minimizing $\|\mathbf{x}_t\|^2$ in the model training process over time. This leads to better performance than OTA FL and R-OTA FL.

In Fig 3, we compare the steady-state test accuracy $\bar{A}$ and training loss $\bar{f}$ among OMUAA, OTA FL, and R-OTA FL with different values of the average transmit power limit $\bar{P}$. The test accuracy $\bar{A}$ yielded by OTA FL and R-OTA FL decreases drastically as $\bar{P}$ decreases. The training loss $\bar{f}$ for OTA FL is not plotted in Fig. 3, as it is much larger than those plotted. Over a wide range of $\bar{P}$, OMUAA significantly outperforms

[5]The gradient sparsification and quantization techniques considered in [20] and [21] are orthogonal to the OMUAA design. Therefore, in our simulation, we assume the full gradient is sent to the edge server.

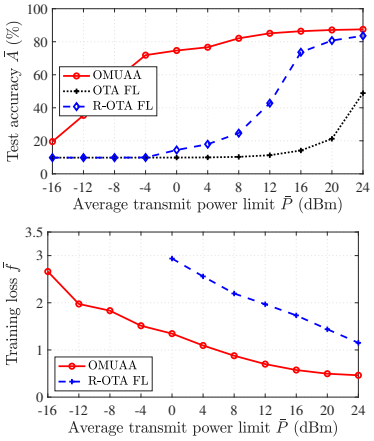

Fig. 3. The impact of average transmit power limit $\bar{P}$. The $\bar{f}$ plot for OTA FL is not included as its value of $\bar{f}$ is much larger than those of OMUAA and R-OTA FL.

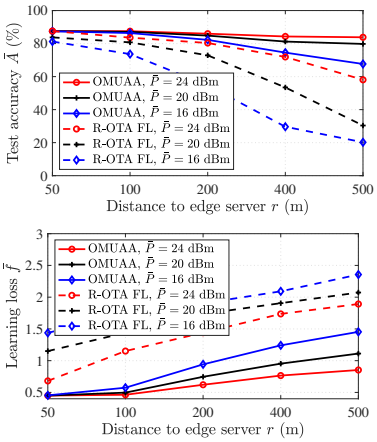

Fig. 4. The impact of distance to edge server r.

the other two schemes. Fig. 4 shows the impact of r, the distance of mobile devices to the edge server, on $\bar{A}$ and $\bar{f}$. The test accuracy of OMUAA is more robust to r than that of R-OTA FL. Furthermore, the performance gain of OMUAA over R-OTA FL becomes more substantial as r increases.

VII. Conclusions

We consider FL in wireless edge networks with analog aggregation over noisy wireless fading multiple access channels. We propose an efficient OMUAA algorithm to minimize the accumulated training loss over time at the edge server, subject to individual long-term transmit power constraints at the mobile devices. OMUAA depends only on the current local CSI, without needing to know the channel distribution. The local models yielded by OMUAA are channel- and power-aware, and are in closed forms with low computational complexity. Our analysis considers the mutual impact between model training and analog aggregation over time to provide performance guarantees on both the computation and communication performance metrics. Simulation results based on realistic LTE network settings and real-word image classification datasets show substantial performance advantage of OMUAA over the known best alternatives under different scenarios.

REFERENCES

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, pp. 2322–2358, 2017.

[2] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, pp. 1738–1762, Aug. 2019.

[3] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, pp. 2204–2239, Nov. 2019.

[4] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, pp. 19–25, Jan. 2020.

[5] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *J. Amer. Stat. Assoc.*, vol. 19, pp. 2322–2358, Feb. 2018.

[6] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. NIPS Workshop on Private Multi-Party Mach. Learn.*, 2016.

[7] 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. Intel. Conf. Artif. Intell. Statist. (AISTATS)*, 2017.

[8] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Info. Proc. Sys. (NIPS)*, 2017.

[9] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Adv. Neural Info. Proc. Sys. (NIPS)*, 2017.

[10] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "SignSGD: Compressed optimisation for non-convex problems," in *Proc. Intel. Conf. Mach. Learn. (ICML)*, 2018.

[11] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. Conf. Empirical Methods in Natural Lang. Process. (EMNLP)*, 2017.

[12] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. Adv. Neural Info. Proc. Sys. (NIPS)*, 2018.

[13] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Proc. Adv. Neural Info. Proc. Sys. (NIPS)*, 2018.

[14] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. Intel. Conf. Learn. Represent. (ICLR)*, 2019.

[15] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019.

[16] T. Lin, S. U. Stich, and M. Jaggi, "Don't use large mini-batches, use local SGD," in *Proc. Intel. Conf. Learn. Represent. (ICLR)*, 2020.

[17] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, pp. 491–506, Jan. 2020.

[18] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, pp. 2022–2035, Mar. 2020.

[19] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "COTAF: Convergent over-the-air federated learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2020.

[20] M. M. Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, Mar. 2020.

[21] M. M. Amiri and D. Gunduz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, pp. 3546–3557, May 2020.

[22] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, Apr. 2020.

[23] H. Guo, A. Liu, and V. K. N. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *Internet Things J.*, vol. 8, pp. 197–210, Jan. 2021.

[24] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, pp. 170–185, Jan. 2021.

[25] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. wireless commun.*, Mar. 2021.

[26] J. Zhang, N. Li, and M. Dedeoglu, "Federated learning over wireless networks: A band-limited coordinated descent approach," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2021.

[27] S. Katti, S. Gollakota, and D. Katabi, "Embracing wireless interference: Analog network coding," in *Proc. ACM SIGCOMM*, 2007.

[28] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3498–3516, Oct. 2007.

[29] M. Goldenbaum and S. Stanczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Trans. Commun.*, vol. 61, pp. 3863–3877, Sep. 2013.

[30] O. Abari, H. Rahul, D. Katabi, and M. Pant, "Airshare: Distributed coherent transmission made seamless," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2015.

[31] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts*, vol. 22, pp. 2031–2063, 2020.

[32] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2018.

[33] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2019.

[34] 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, pp. 269–283, Jan. 2021.

[35] K. Wei, J. Li, C. Ma, M. Ding, C. Chen, S. Jin, Z. Han, and H. Vincent Poor, "Low-latency federated learning over wireless channels with differential privacy," *IEEE J. Sel. Areas Commun.*, Nov. 2021.

[36] S. Shalev-Shwartz, "Online learning and online convex optimization," *Found. Trends Mach. Learn.*, vol. 4, pp. 107–194, Feb. 2012.

[37] M. J. Neely, *Stochastic Network Optimization with Application on Communication and Queueing Systems*. Morgan & Claypool, 2010.

[38] S. Mannor, J. N. Tsitsiklis, and J. Y. Yu, "Online learning with sample path constraints," *J. Mach. Learn. Res.*, vol. 10, pp. 569–590, Mar. 2009.

[39] H. Holma and A. Toskala, *WCDMA for UMTS - HSPA evolution and LTE*. John Wiely & Sons, 2010.

[40] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/