Joint Online Optimization of Model Training and Analog Aggregation for Wireless Edge Learning

Juncheng Wang, *Member, IEEE*, Ben Liang, *Fellow, IEEE*, Min Dong, *Senior Member, IEEE*,
Gary Boudreau, *Senior Member, IEEE*, and Hatem Abou-Zeid, *Member, IEEE*

***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 the 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 incurring 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. Furthermore, we consider a variant of OMUAA with double regularization on both the local and global models, termed OMUAA-DR, and show that it can significantly reduce the convergence time to reach long-term transmit power constraints. In addition, we extend both OMUAA and OMUAA-DR to enable analog gradient aggregation, while preserving their performance bounds. Simulation results based on real-world image classification datasets and typical wireless network settings demonstrate substantial performance gain of OMUAA and OMUAA-DR over the known best alternatives.**

***Index Terms*—Federated learning, wireless edge network, over-the-air computation, online optimization, long-term constraint**

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 learning technologies at the edge servers and devices, collectively call *edge learning* [2]-[5]. 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 [6].

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 [7], [8]. 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 [9]-[11], sparsification [12]-[14], and local updates [15]-[17]. 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 [18]-[27]. Such *over-the-air* computation can be traced back at least to *analog network coding* [28]-[31], which takes advantage of the superposition property of wireless channels, reducing communication overhead 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 [18]-[27]. 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 per-iteration 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,

J. Wang was with the University of Toronto and is now with Hong Kong Baptist University (e-mail: jcwang@comp.hkbu.edu.hk). B. Liang is with the University of Toronto (e-mail: liang@ece.utoronto.ca). M. Dong is with the Ontario Tech University (e-mail: min.dong@ontariotechu.ca). G. Boudreau is with Ericsson Canada (e-mail: gary.boudreau@ericsson.com). H. Abou-Zeid was with Ericsson Canada and is now with the University of Calgary (e-mail: hatem.abouzeid@ucalgary.ca).

This work was supported in part by Ericsson Canada and by the Natural Sciences and Engineering Research Council of Canada.

A preliminary version of this work was presented in IEEE International Conference on Computer Communications (INFOCOM), 2022 [1] (DOI: 10.1109/INFOCOM48880.2022.9796860).

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 heterogeneous energy budgets for device communication over time, expressed as individual long-term transmit power constraints.

Different from the standard per-iteration model training for FL that does not consider the wireless communication layer, our trained models are adaptive to the time-varying channel states. Furthermore, we analyze the mutual impact between computation and communication over time to derive performance bounds for our proposed algorithms. 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 individual transmit power budgets 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 transmit 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 further consider a variant of OMUAA with double regularization, termed OMUAA-DR, to update the current local models based on both the previous local and global models. It captures useful information from both the local and global models to further minimize the accumulated training loss and long-term transmit power constraint violation. We analyze the impact of double regularization in OMUAA-DR, and show that it achieves an improved $\mathcal{O}(\frac{1}{\epsilon})$ convergence time in long-term transmit power constraint violation, while maintaining the same $\mathcal{O}(\frac{1}{\epsilon^2})$ convergence time in training loss. Such improved performance is achieved with an additional step-size parameter, some extra memory to store the local model, and a Lipschitz continuity assumption on the constraint function.
- We extend both OMUAA and OMUAA-DR to enable analog gradient aggregation, where the mobile devices transmit the gradient or the model difference instead of the model itself to the edge server. Our analysis shows that when the step size is small or when the gradient itself is small, analog gradient aggregation can yield better learning performance than analog model aggregation. Furthermore, our derived performance bounds still hold for analog gradient aggregation.
- We study the impact of system parameters on the performance of our proposed algorithms, by experimenting with real-world image classification datasets, under typical wireless network settings. We demonstrate substantial performance advantage of OMUAA and OMUAA-DR over the known best alternatives for both convex logistic regression and non-convex neural network training.

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 and its performance bounds. Then, we discuss the OMUAA-DR variant and study its performance in Section V. In Section VI, we extend both OMUAA-DR and OMUAA to enable analog gradient aggregation. Simulation results are presented in Section VII, followed by concluding remarks in Section VIII.

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 communication, *i.e.,* digital error-control coded transmission (see [32] and references therein). For example, [33] proposed adaptive global model aggregation under resource constraints for FL. The performance trade-offs between computation and communication were investigated in [34] and [35], using conventional orthogonal multiple access. Differential privacy in federated learning was considered in [36]. FL with source coding for quantized transmission was investigated in [37], [38]. None of these solutions are applicable to FL with analog aggregation. The above works all adopt the orthogonal digital transmission approach to communicate the local models to the edge server. The model parameters are transmitted separately from each mobile device through subchannels (*e.g.,* time, in frequency, or code) to the edge server, which can lead to high communication overhead. Furthermore, digital communication requires source coding to convert the model parameters into bits and channel coding to losslessly convey these bits over the noisy channel.

B. FL with Analog Aggregation

To further reduce the communication overhead, [18]-[20] exploited the superposition property of a multiple access channel to allow simultaneous model transmissions from the mobile devices, without any source coding or channel coding. It has been shown that such analog aggregation approach can lead to superior performance in FL [18], [21], [22]. In [18], truncated local model parameters were scheduled for aggregation based on the channel condition. Receiver beamforming design was studied in [19] to maximize the number of mobile devices for model aggregation at each iteration. In [20], 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 [21]-[27]. Gradient quantization and sparsification were exploited for compressed analog aggregation in [21] and [22] over static and fading multiple access channels, respectively. The convergence of iterative analog gradient aggregation was studied in [23] and [24] with sparsified and full gradients, respectively. Power allocation was investigated in [25] to achieve differential privacy. Gradient statistics aware power control was proposed in [26] for aggregation error minimization. In [27], 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 and OMUAA-DR 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) [39], especially OCO with long-term constraints [40]-[47]. In [40], a saddle-point-typed OCO algorithm was proposed, which achieves a time averaged objective value that is $\mathcal{O}(\epsilon)$ worse than the best fixed offline decision when the time horizon $T \geq \frac{1}{\epsilon^2}$, and a time averaged violation on the long-term time-invariant constraints that is $\mathcal{O}(\epsilon)$ when $T \geq \frac{1}{\epsilon^4}$. A follow-up work [41] provided trade-off between the static regret and constraint violation. A virtual-queue based algorithm in [42] reduced the convergence time of the long-term time-invariant constraints to $\mathcal{O}(\frac{1}{\epsilon})$. Virtual-queue based algorithms were also proposed for OCO with independent and identically distributed (i.i.d.) long-term constraints in [43] and [44], based on the standard gradient descent approach and general mirror descent approach, respectively. The saddle-point-typed and virtual-queue-typed algorithms were modified in [45] and [46], to provide performance bounds with respect to (w.r.t.) a dynamic online decision sequence while dealing with long-term time-varying constraints. Distributed saddle-point-typed OCO with long-term constraints was considered in [47] for error-free communication and in [48] for noisy analog aggregation. However, the long-term constraint function is assumed to be fixed over time in [47], [48], which does not apply to our joint online optimization problem with long-term time-varying power constraint functions. Furthermore, the performance analysis in [47], [48] focuses on the static regret by comparing with the best fixed benchmark, while ours is on the more attractive dynamic regret by comparing with the per-slot optimal benchmark.

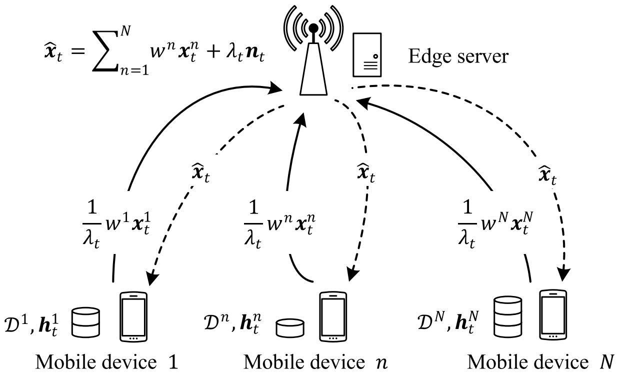

$\hat{\boldsymbol{x}}_t = \sum_{n=1}^{N} w^n \boldsymbol{x}_t^n + \lambda_t \boldsymbol{n}_t$
Edge server
$\hat{\boldsymbol{x}}_t$
$\hat{\boldsymbol{x}}_t$
$\hat{\boldsymbol{x}}_t$
$\frac{1}{\lambda_t} w^1 \boldsymbol{x}_t^1$
$\frac{1}{\lambda_t} w^n \boldsymbol{x}_t^n$
$\frac{1}{\lambda_t} w^N \boldsymbol{x}_t^N$
$\mathcal{D}^1, \boldsymbol{h}_t^1$
$\mathcal{D}^n, \boldsymbol{h}_t^n$
$\mathcal{D}^N, \boldsymbol{h}_t^N$
Mobile device 1
Mobile device n
Mobile device N

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

The above works [40]-[47] on OCO with long-term constraints mainly concern delayed information feedback, which is inherently different from the joint online optimization framework of this work for FL with analog aggregation. In particular, [49] proved that no OCO algorithm can simultaneously provide $\mathcal{O}(\epsilon)$ optimality gap and $\mathcal{O}(\epsilon)$ long-term time-varying constraint violation due to feedback delay, which OMUAA and OMUAA-DR can achieve (see Sections IV-B and V-B).

A part of our solution also resembles Lyapunov optimization [50], which uses the system state and queueing information to implicitly learn the system variations and update the online decisions accordingly without needing to know the system statistics. However, 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.

III. System Model and Problem Formulation

A. Learning Objective

We consider a wireless edge network where N mobile devices are connected to the same 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 cross-entropy for logistic regression, to measure the prediction accuracy on data feature vector $\mathbf{u}^{n,i}$ w.r.t. its true label $v^{n,i}$ (see Section VII-B).

In general, the learning objective is to find a global model $\mathbf{x}^\star$ that minimizes the following *global* training loss function

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

where $\mathcal{D} = \bigcup_{n=1}^{N} \{\mathcal{D}^n\}$ is the global dataset and $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$. This is equivalent to the averaged training loss incurred by the global dataset $\mathcal{D}$. 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. In FL, with the assistant of the edge server, N mobile devices cooperate to minimize (1) based on local datasets.

In practical systems, the local training datasets $\{\mathcal{D}^n\}$ can be very large in size while mobile devices often have limited computation capacities. To address this issue, each mobile device n can sample a *batch* dataset $\mathcal{B}_t^n \subseteq \mathcal{D}^n$ for model training at each iteration t [9]-[17]. The *local* training loss function $f_t^n(\mathbf{x})$ at iteration t incurred by the batch dataset $\mathcal{B}_t^n$ is given by

$$f_t^n(\mathbf{x}) = \frac{1}{|\mathcal{B}_t^n|} \sum_{i=1}^{|\mathcal{B}_t^n|} l(\mathbf{x}; \mathbf{u}_t^{n,i}, v_t^{n,i}) \tag{2}$$

where $|\mathcal{B}_t^n|$ is the cardinality of $\mathcal{B}_t^n$. We assume the batch size $|\mathcal{B}_t^n|$ is fixed over time for each mobile device n. Let $\mathcal{B}_t = \bigcup_{n=1}^{N} \{\mathcal{B}_t^n\}$ denote the sampled global dataset at iteration t. Our *global* training loss function $f_t(\mathbf{x})$ at iteration t incurred by $\mathcal{B}_t$ is defined as

$$f_t(\mathbf{x}) = \sum_{n=1}^{N} w^n f_t^n(\mathbf{x}) \tag{3}$$

where $w^n = \frac{|\mathcal{B}_t^n|}{|\mathcal{B}_t|}$ is the weight on mobile device n, and we have $\sum_{n=1}^{N} w^n = 1$.[1]

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 $\mathbf{x}^{\star}$ [7], [8]. 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 [28]-[31]. 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 [28] and was recently extended to FL [18]-[27] 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 [30] and [31], 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.[2]

We model the channel between the N mobile devices and the edge server as a noisy wireless fading multiple access channel. Let $\mathbf{h}_t^n = [h_t^{n,1}, \dots, h_t^{n,d}]^T \in \mathbb{C}^d$ be the channel state vector between mobile device n and the edge server at the t-th iteration. Mobile devices that are far away from the edge server generally have weak channel states over time. We assume the local channel state information (CSI) is available at each mobile device [18]-[27]. We note that this channel model is suitable for either single-antenna or multi-antenna communication.

The transmitted signals from the mobile devices carried by the noisy wireless fading multiple access channel are summed over the air due to the superposition property of wireless channels. The received signal vector $\mathbf{y}_t \in \mathbb{C}^d$ at the edge server is given by

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

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 $\mathbf{b}_t^n = [\frac{h_t^{n,1}}{|h_t^{n,1}|^2}, \dots, \frac{h_t^{n,d}}{|h_t^{n,d}|^2}]^T \in \mathbb{C}^d$ being the entry-wise channel inversion vector w.r.t. $\mathbf{h}_t^n$.[3] The design of a common λ_t among the N mobile devices at each iteration t was studied in [18], [19], [21], [22], [24]-[27], 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.

[1] Most prior works on FL with analog aggregation [18]-[26], as well as the preliminary version of our work in [1], consider only the simplified scenario where each mobile device n use the entire local dataset $\mathcal{D}^n$ at each iteration t. In that case, the global training loss function is fixed over time as in (1).

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

[3] When the channel power is small, we can add some constant at the denominator of the channel inversion vector to avoid using too much transmit power or causing numerical problems. Using this regularized channel inversion method will incur some additional noise to the global model $\hat{\mathbf{x}}_t$ in (7). However, such noise can be treated as one part of the receiver noise $\mathbf{n}_t$ and does not impact our performance analysis later.

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 $\mathbf{n}_t \triangleq \Re\{\mathbf{z}_t\}$.[4] The edge server then broadcasts $\hat{\mathbf{x}}_t$ to all the N mobile devices. As in [18]-[27], 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.

To summarize, over-the-air model aggregation has the following two major steps:

- At each iteration t, after obtaining the local models $\{\mathbf{x}_t^n\}$, the N mobile devices generate the transmitted signal vectors $\{\mathbf{s}_t^n\}$ in (6) that carry the information of $\{\mathbf{x}_t^n\}$, and then transmit $\{\mathbf{s}_t^n\}$ simultaneously over a noisy multiple access channel to the edge server.
- At each iteration t, after receiving the signal vector $\mathbf{y}_t$ in (5), the edge server performs scaling to recover a noisy global model $\hat{\mathbf{x}}_t$ in (7).

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

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

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

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

where $\langle \mathbf{a}, \mathbf{b} \rangle$ represents the inner product of vectors $\mathbf{a}$ and $\mathbf{b}$, and $\|\cdot\|$ denotes Euclidean norm. All existing works on FL with analog aggregation [18]-[27] 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 online optimization approach to account for the impacts of analog aggregation, including communication error, channel fading, and power allocation, on model training over time.

C. Problem Formulation

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

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

where T is the total number of iterations and the expectation is taken over the randomness of the channel states and batch datasets. We note that, as the training process goes on until reaching the steady state, *i.e.,* as $T \to \infty$, the accumulated training loss $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{f_t(\hat{\mathbf{x}}_t)\}$ over T iterations approaches the final 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:

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

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 and batch datasets. 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 and batch datasets. In **P1**, the training loss $f_t(\hat{\mathbf{x}}_t)$ over batch dataset $\mathcal{B}_t$ 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 wireless fading channels, solving **P1** requires simultaneous consideration for computation and communication. Furthermore, compared with the one-shot optimization problem that minimizes (1) with full dataset $\mathcal{D}$, the additional long-term transmit power constraints in (12) of **P1** require a more complicated online algorithm, especially since the channel state and batch dataset varies over time. In this work, without needing to know the channel or batch dataset distribution, we aim to develop an online algorithm based on the local channel state $\mathbf{h}_t^n$ and batch dataset $\mathcal{B}_t^n$ 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

[4] We use the real part of the channels to transmit the real vector $\mathbf{x}_t$. The derivation can be easily extended to utilize both the real and imaginary parts of the channels by separating $\mathbf{x}_t$ into half as in [22], without major technical alternation.

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.

A. OMUAA Algorithm

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 [50]. However, we note that, although a part of our performance bound analysis for OMUAA borrows techniques from Lyapunov drift analysis, OMUAA is structurally different from Lyapunov optimization as explained in Section II.

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

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

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 $\mathbf{h}_t^n$, batch dataset $\mathcal{B}_t^n$, 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 $\nabla f_t^n(\hat{\mathbf{x}}_{t-1})$ in $\mathbf{P2}^n$ is evaluated using the noisy global model $\hat{\mathbf{x}}_{t-1}$. Note that one of the main novelties of this work is to design the local model $\mathbf{x}_t^n$ as a part of the analog aggregation. When we update $\mathbf{x}_t^n$, we jointly optimize the training loss and transmit power over time. In this sense, our method is a joint online optimization of model training and analog aggregation.

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 and therefore can be solved efficiently using standard optimization tools. 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

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

Algorithm 1 OMUAA: Mobile device n's algorithm

```
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:
    Update local model $\mathbf{x}_t^n$ by solving $\mathbf{P2}^n$ via (17).
    Update local virtual queue $Q_t^n$ via (14).
    Transmit signals $\mathbf{s}_t^n$ in (6) to the edge server.
```

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

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

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 $\mathbf{x}_t^n$ and therefore $\mathbf{x}_t^n$ can be computed in parallel.

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 $\frac{1}{1+\alpha\theta_t^{n,i}}$ 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 IV-B, we will show that the update sequence $\{\mathbf{x}_t^n\}$ further satisfies individual long-term transmit power constraints.

To summarize, OMUAA has the following two components:

- 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 $\mathbf{h}_t^n$ and batch dataset $\mathcal{B}_t^n$, 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 and γ will be discussed in Section IV-B, after we derive the performance bounds for OMUAA.

Remark 1. The computational complexity of calculating the local batch gradient $\nabla f_t^n(\hat{\mathbf{x}}_{t-1})$ 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_{t-1}^n and the factor $\boldsymbol{\theta}_t^n$, both

Algorithm 2 OMUAA: Edge server's algorithm

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

are in the order of $\mathcal{O}(d)$. Therefore, the local model update in (17) has low computational complexity.

B. Performance Bounds of OMUAA

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. Specifically, we require the gradient of the loss function, the output of the constraint function, and the communication noise to be upper bounded by some constants. These are mild assumptions that are easily satisfied in practical systems.

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

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

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

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

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

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

1) Bound for the Accumulated Training Loss: Define $L_t^n \triangleq \frac{1}{2}(Q_t^n)^2$ as a quadratic Lyapunov function and $\Delta_t^n \triangleq L_t^n - L_{t-1}^n$ as the corresponding Lyapunov drift for each mobile device n. We first provide an upper bound on Δ_t^n in the following lemma.

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

$$\Delta_t^n \le \frac{1}{2}G^2 + Q_{t-1}^n g_t^n(\mathbf{x}_t^n), \quad \forall n, \quad \forall t. \tag{21}$$

Proof: From the virtual queue updating rule in (14), we have

$$\begin{aligned}\Delta_t^n &= \frac{1}{2}\left[(\max\{Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), 0\})^2 - (Q_{t-1}^n)^2\right]\\ &\le \frac{1}{2}\left[(Q_{t-1}^n + g_t^n(\mathbf{x}_t^n))^2 - (Q_{t-1}^n)^2\right]\\ &= \frac{1}{2}[g_t^n(\mathbf{x}_t^n)]^2 + Q_{t-1}^n g_t^n(\mathbf{x}_t^n) \overset{(a)}{\le} \frac{1}{2}G^2 + Q_{t-1}^n g_t^n(\mathbf{x}_t^n)\end{aligned}$$

where (a) follows from $g_t^n(\mathbf{x})$ being bounded in (19). ■

We also require the following lemma from [39, Lemma 2.8].

Lemma 2. [39, Lemma 2.8] 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 $f(\mathbf{z}) \le f(\mathbf{y}) - \frac{1}{2\alpha}\|\mathbf{y} - \mathbf{z}\|^2$.

We consider a block fading channel model, where $\mathbf{h}_t$ over iteration t are i.i.d. [18], [21]-[23], [26]. The sampled batch datasets $\mathcal{B}_t$ are assumed to be i.i.d. over iteration t as in the standard stochastic gradient descent approach. The distributions of $\mathbf{h}_t^n$ and $\mathcal{B}_t^n$ are unknown and can be arbitrary. We assume the power-scaling factor λ_t depends on the underlying system states. For i.i.d. channel state $\mathbf{h}_t$ and batch dataset $\mathcal{B}_t$, there exists a stationary randomized optimal global solution $\mathbf{x}_t^\star$ to **P1** over noiseless channels, which depends only on the (unknown) distributions of $\mathbf{h}_t$ and $\mathcal{B}_t$, and achieves the minimum objective value (*i.e.,* the minimum accumulated training loss) of **P1**, denoted by $f^\star$ [50]. Using Lemmas 1 and 2, the following theorem provides an upper bound on the accumulated training loss by OMUAA.

Theorem 1. For any $\alpha, \gamma > 0$, regardless of the channel and batch dataset distributions, the accumulated training loss yielded by OMUAA is upper bounded by

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{f_t(\hat{\mathbf{x}}_t)\} \le f^\star + \frac{D^2\alpha}{2} + \frac{G^2\gamma}{2} + \frac{R^2 + \rho^2\Lambda_{2,T} + 4R\rho\Lambda_T}{2\alpha T}\\ + \frac{(2R + \lambda_{\max}\rho)\Pi_T}{\alpha T} + \frac{f_T(\hat{\mathbf{x}}_T) - f_1(\mathbf{x}_1^\star)}{T}\end{aligned} \tag{22}$$

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 $\Pi_T = \sum_{t=1}^{T}\mathbb{E}\{\|\mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\|\}$ is the accumulated variation of the optimal global model over noiseless channels.

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 $\mathbf{x}_t^n$ minimizes the objective of $\mathbf{P2}^n$ over $\mathcal{X}$, from Lemma 2, we have

$$\begin{aligned}&\langle\nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\|^2 + \gamma Q_{t-1}^n g_t^n(\mathbf{x}_t^n)\\ &\quad\le \langle\nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\rangle + \gamma Q_{t-1}^n g_t^n(\mathbf{x}_t^\star)\\ &\quad\quad + \frac{1}{2\alpha}\left(\|\mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2\right).\end{aligned} \tag{23}$$

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

$$\begin{aligned}&\|\mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2\\ &\overset{(a)}{\le} \|\mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}_{t+1}^\star - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\|^2\\ &\quad + 2\|\mathbf{x}_{t+1}^\star - \hat{\mathbf{x}}_t\|\|\mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\| + \|\mathbf{x}_t^\star - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2\\ &\overset{(b)}{\le} \psi_t + 2\|\mathbf{x}_{t+1}^\star - \hat{\mathbf{x}}_t\|\pi_t + \phi_t^n\end{aligned} \tag{24}$$

where (a) is because $\|\mathbf{a}+\mathbf{b}\|^2 \ge \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\|\mathbf{b}\|$; and (b) follows from defining $\psi_t \triangleq \|\mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}_{t+1}^\star - \hat{\mathbf{x}}_t\|^2$, $\pi_t \triangleq \|\mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\|$, and $\phi_t^n \triangleq \|\mathbf{x}_t^\star - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2$.

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

$$f_t^n(\hat{\mathbf{x}}_{t-1}) + \langle\nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\rangle \le f_t^n(\mathbf{x}_t^\star) \tag{25}$$

to the first term on the RHS of (23), and rearranging the terms on both sides, we have

$$f_t^n(\hat{\mathbf{x}}_{t-1}) - f_t^n(\mathbf{x}_t^\star)$$

$$
\begin{aligned}
&\leq -\langle \nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\rangle - \frac{1}{2\alpha}\|\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\|^2 \\
&\quad + \gamma Q_{t-1}^n g_t^n(\mathbf{x}_t^\star) - \gamma Q_{t-1}^n g_t^n(\mathbf{x}_t^n) \\
&\quad + \frac{1}{2\alpha}(\psi_t + 2\|\mathbf{x}_{t+1}^\star - \hat{\mathbf{x}}_t\|\pi_t + \phi_t^n). \qquad (26)
\end{aligned}
$$

We now bound the RHS of (26). Completing the square and noting that $\nabla f(\mathbf{x})$ is bounded in (18), we have

$$
\begin{aligned}
&-\langle \nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\rangle - \frac{1}{2\alpha}\|\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\|^2 \\
&= -\left\|\sqrt{\frac{\alpha}{2}}\nabla f_t^n(\hat{\mathbf{x}}_{t-1}) + \frac{\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}}{\sqrt{2\alpha}}\right\|^2 + \frac{\alpha}{2}\|\nabla f_t^n(\hat{\mathbf{x}}_{t-1})\|^2 \\
&\leq \frac{\alpha}{2}\|\nabla f_t^n(\hat{\mathbf{x}}_{t-1})\|^2 \leq \frac{D^2\alpha}{2}. \qquad (27)
\end{aligned}
$$

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.,*

$$
\|\mathbf{x}\| \leq R, \quad \forall x \in \mathcal{X}, \qquad (28)
$$

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

$$
\|\mathbf{x}_{t+1}^\star - \hat{\mathbf{x}}_t\| \leq \|\mathbf{x}_{t+1}^\star\| + \|\mathbf{x}_t\| + \|\lambda_t \mathbf{n}_t\| \leq 2R + \lambda_{\max}\rho. \quad (29)
$$

Substituting (21), (27), and (29) into the RHS of (26), multiplying both sides by w^n, summing over $n = 1$ to N, and taking expectation, we have

$$
\begin{aligned}
&\mathbb{E}\{f_t(\hat{\mathbf{x}}_{t-1})\} - \mathbb{E}\{f_t(\mathbf{x}_t^\star)\} \\
&\leq \frac{D^2\alpha}{2} + \gamma\left(\sum_{n=1}^N w^n \mathbb{E}\{Q_{t-1}^n g_t^n(\mathbf{x}_t^\star)\} + \frac{1}{2}G^2 - \sum_{n=1}^N w^n \mathbb{E}\{\Delta_t^n\}\right) \\
&\quad + \frac{1}{2\alpha}\left(\mathbb{E}\{\psi_t\} + 2(2R + \lambda_{\max}\rho)\mathbb{E}\{\pi_t\} + \sum_{n=1}^N w^n \mathbb{E}\{\phi_t^n\}\right). (30)
\end{aligned}
$$

From $\mathbf{x}_t^\star$ being independent of $Q_{t-1}^n \geq 0$, and $\mathbb{E}\{g_t^n(\mathbf{x}_t^\star)\} \leq 0$, we have $\mathbb{E}\{Q_{t-1}^n g_t^n(\mathbf{x}_t^\star)|Q_{t-1}^n\} = Q_{t-1}^n \mathbb{E}\{g_t^n(\mathbf{x}_t^\star)\} \leq 0$. It then follows from the iterated law of expectation that $\mathbb{E}\{Q_{t-1}^n g_t^n(\mathbf{x}_t^\star)\} = \mathbb{E}\{\mathbb{E}\{Q_{t-1}^n g_t^n(\mathbf{x}_t^\star)|Q_{t-1}^n\}\} \leq 0$. Further note that the batch datasets $\mathcal{B}_t$ are i.i.d. over iterations, we have $\mathbb{E}\{f_t(\mathbf{x})\} = \mathbb{E}\{f_{t-1}(\mathbf{x})\}, \forall \mathbf{x} \in \mathbb{R}^d, \forall t$ and therefore $\mathbb{E}\{f_t(\hat{\mathbf{x}}_{t-1})\} = \mathbb{E}\{f_{t-1}(\hat{\mathbf{x}}_{t-1})\}$. Substituting them into the RHS of (30) and summing it over $t = 2$ to T, we have

$$
\begin{aligned}
&\sum_{t=1}^{T-1}\mathbb{E}\{f_t(\hat{\mathbf{x}}_t)\} - \sum_{t=2}^{T}\mathbb{E}\{f_t(\mathbf{x}_t^\star)\} \\
&\leq \frac{D^2\alpha}{2}T + \frac{G^2\gamma}{2}T - \gamma\sum_{t=2}^T\sum_{n=1}^N w^n\mathbb{E}\{\Delta_t^n\} + \frac{1}{2\alpha}\sum_{t=2}^T\mathbb{E}\{\psi_t\} \\
&\quad + \frac{2R + \lambda_{\max}\rho}{\alpha}\sum_{t=2}^T\mathbb{E}\{\pi_t\} + \frac{1}{2\alpha}\sum_{t=2}^T\sum_{n=1}^N w^n\mathbb{E}\{\phi_t^n\}. \qquad (31)
\end{aligned}
$$

We now bound the RHS of (31). From the definition of Δ_t, $Q_1^n = 0$, and $Q_t^n \geq 0, \forall t$, we have

$$
-\sum_{t=2}^T \mathbb{E}\{\Delta_t^n\} = \frac{1}{2}\mathbb{E}\{(Q_1^n)^2\} - \frac{1}{2}\mathbb{E}\{(Q_T^n)^2\} \leq 0. \qquad (32)
$$

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

$$
\sum_{t=2}^T \mathbb{E}\{\psi_t\} = \mathbb{E}\{\|\mathbf{x}_2^\star - \hat{\mathbf{x}}_1\|^2\} - \mathbb{E}\|\mathbf{x}_{T+1}^\star - \hat{\mathbf{x}}_T\|^2\} \leq R^2. \quad (33)
$$

For the last term on the RHS of (31), we have

$$
\begin{aligned}
&\sum_{t=2}^T\sum_{n=1}^N w^n\mathbb{E}\{\phi_t^n\} = \sum_{t=2}^T\sum_{n=1}^N w^n\mathbb{E}\{\|\mathbf{x}_t^\star - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2\} \\
&\overset{(a)}{\leq} \sum_{t=2}^T\sum_{n=1}^N w^n(\mathbb{E}\{\|\mathbf{x}_t^\star - \mathbf{x}_t\|^2\} - \mathbb{E}\{\|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2\}) \\
&\quad + \sum_{t=2}^T\mathbb{E}\{\|\lambda_t\mathbf{n}_t\|^2\} + 2\sum_{t=2}^T\mathbb{E}\{\|\mathbf{x}_t^\star - \mathbf{x}_t\|\|\lambda_t\mathbf{n}_t\|\} \\
&\overset{(b)}{\leq} \sum_{t=2}^T\mathbb{E}\{\|\lambda_t\mathbf{n}_t\|^2\} + 2\sum_{t=2}^T\mathbb{E}\{\|\mathbf{x}_t^\star - \mathbf{x}_t\|\|\lambda_t\mathbf{n}_t\|\} \\
&\overset{(c)}{\leq} \rho^2\Lambda_{2,T} + 4R\rho\Lambda_T \qquad (34)
\end{aligned}
$$

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

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

and (c) follows from $\mathbf{n}_t$ and $\mathcal{X}$ being bounded in (20) and (28), respectively, and the definitions of $\Lambda_{2,T}$ and Λ_T.

Substituting (32)-(34) into (31) and from the definition of Π_T and $f^\star$, we have (22). ■

Theorem 1 provides a general bound for the accumulated training loss by OMUAA, for any values of step-size parameters α and 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 1. 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

$$
\frac{1}{T}\sum_{t=1}^T\mathbb{E}\{f_t(\hat{\mathbf{x}}_t)\} \leq f^\star + \mathcal{O}((1+\rho^2+\Pi_T\rho)\epsilon), \quad \forall T \geq \frac{1}{\epsilon^2}. (35)
$$

Corollary 1 provides an upper bound on the objective value of **P1** in (35), *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 and batch dataset do not vary too drastically over time. In particular, when the channel is static and the batch dataset is fixed (*e.g.,* full batch dataset $\mathcal{D}$), we have $\Pi_T = 0$. The impact of noise on the accumulated training loss is quantified by its upper bound ρ in (20). As T approaches infinity (*i.e.,* ϵ becomes infinitely small), the accumulated training loss yielded by OMUAA is guaranteed to converge.

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

$$\frac{1}{T}\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq \frac{G}{T} + \frac{\alpha\gamma G^2 + 2\alpha DR + (R+\lambda_{\max}\rho)^2}{2\alpha\gamma \bar{P}^n T}, \ \forall n. \quad (36)$$

Proof: Since $\mathbf{x}_t^n$ minimizes the objective of P2^n over $\mathcal{X}$, which contains $\mathbf{0}$, we have

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

where (a) follows from $g_t^n(\mathbf{0}) = -\bar{P}^n$. Rearranging terms of (37), we have

$$\gamma Q_{t-1}^n g_t^n(\mathbf{x}_t^n) \leq -\gamma Q_{t-1}^n \bar{P}^n - \langle \nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n\rangle + \frac{1}{2\alpha}\|\hat{\mathbf{x}}_{t-1}\|^2$$
$$\overset{(a)}{\leq} -\gamma Q_{t-1}^n \bar{P}^n + DR + \frac{(R+\lambda_{\max}\rho)^2}{2\alpha} \quad (38)$$

where (a) follows from $\nabla f_t(\mathbf{x})$, $\mathbf{n}_t$, and $\mathcal{X}$ being bounded in (18), (20), and (28), respectively.

Substituting (38) into the second term on the RHS of (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

$$Q_{t-1}^n > \frac{\alpha\gamma G^2 + 2\alpha DR + (R+\lambda_{\max}\rho)^2}{2\alpha\gamma\bar{P}^n}. \quad (39)$$

If (39) holds, we have $Q_t^n < Q_{t-1}^n$, *i.e.,* the virtual queue decreases; otherwise, the increment from Q_{t-1}^n to Q_t^n is upper bounded, since $Q_t^n - Q_{t-1}^n \leq g_t^n(\mathbf{x}_t^n) \leq G$. It follows that, the virtual queue is upper bounded for all t by

$$Q_t^n \leq G + \frac{\alpha\gamma G^2 + 2\alpha DR + (R+\lambda_{\max}\rho)^2}{2\alpha\gamma\bar{P}^n}. \quad (40)$$

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 over $t=2$ to T, 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 upper bound in (40) into this inequality, we have (36). ■

From Theorem 2, which is for any step-size parameters α and γ, and power-scaling factors $\{\lambda_t\}$, we have the following corollary for some specific values of α, γ, and $\{\lambda_t\}$.

Corollary 2. 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

$$\frac{1}{T}\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq \mathcal{O}((1+\rho^2)\epsilon), \quad \forall n, \quad \forall T \geq \frac{1}{\epsilon^3}. \quad (41)$$

Corollary 2 implies 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 $T \geq \frac{1}{\epsilon^3}$. As T approaches infinity, the long-term power constraint violation goes to zero.

Remark 2. The convergence analysis for FL in the machine learning literature, such as [9]-[17], mainly focuses on bounding the training loss. Different from these convergence analysis, we analyze the mutual impact between model training and analog aggregation over time, to provide performance bounds on both the accumulated training loss and the long-term transmit power. In particular, when we bound the training loss by OMUAA in Theorem 1, we need to consider the additional impact of the long-term transmit power constraints on the training loss, which is not considered in [9]-[17]. Furthermore, we need to provide a convergence analysis on the long-term transmit power, which is given in Theorem 2. Therefore, our performance analysis requires new techniques to simultaneously bound the accumulated training loss and the long-term transmit power.

V. Online Model Updating with Analog Aggregation and Double Regularization

In this section, we propose a variant of the OMUAA algorithm with double regularization on both the local and global models, together with a new constraint penalty on the transmit power violation. While OMUAA-DR requires an additional step-size parameter and some extra space to store the local model, it significantly reduces the convergence time of long-term transmit power violation, from $\mathcal{O}(\frac{1}{\epsilon^3})$ yielded by OMUAA to $\mathcal{O}(\frac{1}{\epsilon})$, while keeping the the same $\mathcal{O}(\frac{1}{\epsilon^2})$ convergence time of training loss.

A. OMUAA-DR Algorithm

We use double regularization on the previous local model $\frac{1}{2\beta}\|\mathbf{x} - \mathbf{x}_{t-1}^n\|^2$ and global model $\frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2$ to update the new local model $\mathbf{x}_t^n$ at each mobile device n, where $\alpha, \beta > 0$ are step-size parameters. The double regularization approach prevents the $\mathbf{x}_t^n$ from deviating too far away from either $\mathbf{x}_{t-1}^n$ or $\hat{\mathbf{x}}_{t-1}$, since both of them help to minimize the training loss and transmit power violation.

Furthermore, for the aforementioned double regularization to be effective, we require a new virtual queue $\tilde{Q}_t^n$ at each mobile device n for the individual long-term transmit power constraints (12) in **P1**. Its updating rule is given by

$$\tilde{Q}_t^n = \max\{-\gamma g_t^n(\mathbf{x}_t^n), \tilde{Q}_{t-1}^n + \gamma g_t^n(\mathbf{x}_t^n)\}, \quad \forall n, \quad \forall t \quad (42)$$

where $\gamma > 0$ is a step-size parameter.[5]

In OMUAA-DR, we solve the following per-iteration optimization problem at each mobile device n, subject to the short-term constraints only:

$$\mathbf{P3}^n : \min_{\mathbf{x}\in\mathcal{X}} \ \langle \nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x} - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2$$

[5] Compared with the standard virtual queue updating rule (14) in OMUAA, the maximum in (42) is taken over the negative constraint violation $-\gamma g_t^n(\mathbf{x}_t^n)$ instead of zero. Such negative lower bound was first proposed in [51] for centralized OCO with time-invariant long-term constraints. Here, we apply it to distributed online learning with model averaging, and the long-term constraints are time-varying, so our algorithm and analysis are substantially different from those of [51].

Algorithm 3 OMUAA-DR: Mobile device n's algorithm

```
Initialize $\mathbf{x}_1^n = \hat{\mathbf{x}}_1 = \mathbf{0}$, $g_1^n(\cdot) \equiv 0$, and $\tilde{Q}_1^n = 0$.
For each $t$, do the following:
    Update local model $\mathbf{x}_t^n$ by solving $\mathbf{P3}^n$ via (45).
    Update local virtual queue $\tilde{Q}_t^n$ via (42).
    Transmit signals $\mathbf{s}_t^n$ in (6) to the edge server.
```

$$+ [\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\gamma g_t^n(\mathbf{x}) + \frac{1}{2\beta}\|\mathbf{x} - \mathbf{x}_{t-1}^n\|^2.$$

Compared with the per-iteration optimization problem $\mathbf{P2}^n$ in OMUAA, we introduce a new constraint penalty $\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)$ and double regularization on the previous local model and global models.

In the following, we show that $\mathbf{P3}^n$ has a closed-form solution. Taking the gradient of the objective function of $\mathbf{P3}^n$ w.r.t. $\mathbf{x}$, we have

$$\nabla f_t^n(\hat{\mathbf{x}}_{t-1}) + \frac{1}{\alpha}(\mathbf{x} - \hat{\mathbf{x}}_{t-1}) + \frac{1}{\beta}(\mathbf{x} - \mathbf{x}_{t-1}^n) + \tilde{\boldsymbol{\theta}}_t^n \circ \mathbf{x} \quad (43)$$

where the i-th entry of $\tilde{\boldsymbol{\theta}}_t^n$ is given by

$$\tilde{\theta}_t^{n,i} = \frac{2\gamma[\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)](w^n)^2}{\lambda_t^2 |h_t^{n,i}|^2}. \quad (44)$$

Setting the gradient in (43) to zero to solve for $\mathbf{x}$ and then projecting it onto $\mathcal{X}$, we have a closed-form local model update for $\mathbf{x}_t^n$, given by

$$\mathbf{x}_t^n = \Bigg[\left(\frac{\alpha+\beta}{\beta}\mathbf{1} + \alpha\tilde{\boldsymbol{\theta}}_t^n\right)^{-1} \circ \left(\hat{\mathbf{x}}_{t-1} + \frac{\alpha}{\beta}\mathbf{x}_{t-1}^n - \alpha\nabla f_t^n(\hat{\mathbf{x}}_{t-1})\right)\Bigg]_{-\mathbf{x}_{\max}}^{\mathbf{x}_{\max}}. \quad (45)$$

To summarize, in OMUAA-DR, each mobile device n first initializes the models $\mathbf{x}_1^n = \hat{\mathbf{x}}_1 = \mathbf{0}$, the local constraint function $g_1^n(\cdot) \equiv 0$, and the local virtual queue $\tilde{Q}_1^n = 0$. At the t-th iteration, each mobile device n solves $\mathbf{P3}^n$ for its local model $\mathbf{x}_t^n$ via (45) and then updates its local virtual queue $\tilde{Q}_t^n$ via (42). Then, each mobile device n transmit signals $\mathbf{s}_t^n$ in (6) to the edge server. We summarize mobile device n's algorithm in Algorithm 3. The edge server's algorithm is the same as Algorithm 2. The choice of step-size parameters α, β, and γ will be discussed in Section V-B, after we derive the performance bounds for OMUAA-DR.

Remark 3. Compared with the local model update (17) in OMUAA, (45) also depends on the previous local model $\mathbf{x}_{t-1}^n$ due to the additional regularization $\frac{1}{2\beta}\|\mathbf{x} - \mathbf{x}_{t-1}^n\|^2$. Furthermore, the step-size parameters α and β tune the relative weights of the previous local model $\mathbf{x}_{t-1}^n$ and global model $\hat{\mathbf{x}}_{t-1}$ on the new local model update. Therefore, OMUAA-DR requires some additional memory to store the previous local model $\mathbf{x}_{t-1}^n$ and transmit power violation $g_{t-1}^n(\mathbf{x}_{t-1}^n)$, as well as an additional step-size parameter β. However, the computational complexity of OMUAA-DR is the same as OMUAA.

B. Performance Bounds of OMUAA-DR

In this section, we derive the performance bounds of OMUAA-DR, taking into account its new constraint penalty with double regularization. Compared with OMUAA, the main technical challenges in analyzing the new constraint penalty with double regularization are to consider its impacts on both the model training and analog aggregation over time, to show an improved convergence time of the long-term transmit power, while maintaining the same convergence time of the accumulated training loss. Specifically, the new virtual queue updating rule in (42) will be shown to provide new virtual queue properties and a new Lyapunov drift upper bound. The double regularization $\frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2 + \frac{1}{2\beta}\|\mathbf{x} - \mathbf{x}_{t-1}^n\|^2$ in $\mathbf{P3}^n$ will provide some additional freedom to construct new telescoping sums. The new virtual queue together with the double regularization help cancel undesired leftover terms in the performance bounds of OMUAA-DR for an improved convergence behavior.

We require the following additional assumption on the Lipschitz continuity of the transmit power constraint function. Note that the value of the Lipschitz continuity constant can be large when the channel power is small. However, it is still usually bounded in practice.

Assumption 4. The constraint function $g_t^n(\mathbf{x})$ is L-Lipschitz continuous: $\exists\ L > 0$, s.t.,

$$|g_t^n(\mathbf{x}) - g_t^n(\mathbf{y})| \le L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \quad \forall n, \forall t. \quad (46)$$

The following lemma provides bounds on the local virtual queue $\tilde{Q}_t^n$.

Lemma 3. The local virtual queue $\tilde{Q}_t^n$ generated by OMUAA-DR is bounded by the following inequalities:

$$\tilde{Q}_t^n \ge 0, \quad \forall n, \quad \forall t, \quad (47)$$

$$\tilde{Q}_t^n + \gamma g_t^n(\mathbf{x}_t^n) \ge 0, \quad \forall n, \quad \forall t. \quad (48)$$

Proof: We prove (47) by induction. The virtual queue is initialized as $\tilde{Q}_1^n = 0$. Suppose $\tilde{Q}_{t-1}^n \ge 0$ for any $t > 1$. From the virtual queue dynamics in (42), if $g_t^n(\mathbf{x}_t^n) < 0$, we have $\tilde{Q}_t^n \ge -\gamma g_t^n(\mathbf{x}_t^n) \ge 0$; otherwise, we have $\tilde{Q}_t^n \ge \tilde{Q}_{t-1}^n + \gamma g_t^n(\mathbf{x}_t^n) \ge 0$. Combining the above two cases, we have (47).

We have $\tilde{Q}_t^n \ge -\gamma g_t^n(\mathbf{x}_t^n), \forall t > 1$ from the virtual queue dynamics in (42). Further note that $g_1^n(\cdot) \equiv 0$ and $\tilde{Q}_1^n = 0$ by initialization, we have (48). ∎

Define $\tilde{L}_t^n \triangleq \frac{1}{2}(\tilde{Q}_t^n)^2$ as a quadratic Lyapunov function and $\tilde{\Delta}_t^n \triangleq \tilde{L}_t^n - \tilde{L}_{t-1}^n$ as the corresponding Lyapunov drift for each mobile device n. In the following lemma, we provide an upper bound on $\tilde{\Delta}_t^n$.

Lemma 4. The Lyapunov drift is upper bounded as follows:

$$\tilde{\Delta}_t^n \le \gamma\tilde{Q}_{t-1}^n g_t^n(\mathbf{x}_t^n) + \gamma^2[g_t^n(\mathbf{x}_t^n)]^2, \quad \forall n, \quad \forall t. \quad (49)$$

Proof: It is easy to verify that (49) holds for either of the two cases in the virtual queue updating rule (42): i) $\tilde{Q}_t^n = -\gamma g_t^n(\mathbf{x}_t^n)$ and ii) $\tilde{Q}_t^n = \tilde{Q}_{t-1}^n + \gamma g_t^n(\mathbf{x}_t^n)$. ∎

1) Bounding the Accumulated Training Loss: Using Lemmas 2-4, we provide an upper bound on the accumulated training loss by OMUAA-DR over noisy channels in the following theorem.

Theorem 3. For any $\alpha, \gamma > 0$ and $\beta \leq \frac{1}{2\gamma^2 L^2}$, regardless of the channel and batch dataset distributions, the accumulated training loss yielded by OMUAA-D is upper bounded by

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{f_t(\hat{\mathbf{x}}_t)\} \leq f^\star + \frac{D^2\alpha}{2} + \frac{G^2\gamma^2}{2T} + \frac{R^2+\rho^2\Lambda_{2,T}+4R\rho\Lambda_T}{2\alpha T} \\ + \frac{(2R+\lambda_{\max}\rho)\Pi_T}{\alpha T} + \frac{2R\Pi_T}{\beta T} + \frac{\gamma^2\Theta_T}{T} \\ + \frac{R^2}{2\beta T} + \frac{f_T(\hat{\mathbf{x}}_T) - f_1(\mathbf{x}_1^\star)}{T}.\end{aligned}$$

where $\Theta_T \triangleq \sum_{t=1}^{T} \max_{\mathbf{x}\in\mathcal{X},n} \mathbb{E}\left\{[g_t^n(\mathbf{x}) - g_{t-1}^n(\mathbf{x})]^2\right\}$ is the accumulated variation of the transmit power constraint function over time.

Proof: The objective function in $\mathbf{P3}^n$ is $(\frac{1}{\alpha} + \frac{1}{\beta})$-strongly convex over $\mathcal{X}$ w.r.t. Euclidean norm $\|\cdot\|$ due to the double regularization. Since $\mathbf{x}_t^n$ minimizes the objective of $\mathbf{P3}^n$ over $\mathcal{X}$, from Lemma 2, we have

$$\begin{aligned}&\langle\nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\|^2 \\ &\quad + [\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\gamma g_t^n(\mathbf{x}_t^n) + \frac{1}{2\beta}\|\mathbf{x}_t^n - \mathbf{x}_{t-1}^n\|^2 \\ &\leq \langle\nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\rangle + [\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\gamma g_t^n(\mathbf{x}_t^\star) \\ &\quad + \frac{1}{2\alpha}(\|\mathbf{x}_t^\star - \hat{\mathbf{x}}_{t-1}\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2) \\ &\quad + \frac{1}{2\beta}(\|\mathbf{x}_t^\star - \mathbf{x}_{t-1}^n\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2).\end{aligned} \tag{50}$$

To bound the last term on the RHS of (50), we have

$$\begin{aligned}&\|\mathbf{x}_t^\star - \mathbf{x}_{t-1}^n\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_t^n\|^2 \\ &\quad = \|\mathbf{x}_t^\star - \mathbf{x}_{t-1}^n\|^2 - \|\mathbf{x}_{t+1}^\star - \mathbf{x}_t^n + \mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\|^2 \\ &\quad \overset{(a)}{\leq} \|\mathbf{x}_t^\star - \mathbf{x}_{t-1}^n\|^2 - \|\mathbf{x}_{t+1}^\star - \mathbf{x}_t^n\|^2 - \|\mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\|^2 \\ &\qquad + 2\|\mathbf{x}_{t+1}^\star - \mathbf{x}_t^n\|\|\mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\| \overset{(b)}{\leq} \Psi_t^n + 4R\pi_t.\end{aligned} \tag{51}$$

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 $\Psi_t^n \triangleq \|\mathbf{x}_t^\star - \mathbf{x}_{t-1}^n\|^2 - \|\mathbf{x}_{t+1}^\star - \mathbf{x}_t^n\|^2$.

Substituting (24), (25), (29), and (51) into (50), we have

$$\begin{aligned}&f_t^n(\hat{\mathbf{x}}_{t-1}) - f_t^n(\mathbf{x}_t^\star) \\ &\leq -\langle\nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\rangle - \frac{1}{2\alpha}\|\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\|^2 \\ &\quad + [\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)][\gamma g_t^n(\mathbf{x}_t^\star) - \gamma g_t^n(\mathbf{x}_t^n)] \\ &\quad - \frac{1}{2\beta}\|\mathbf{x}_t^n - \mathbf{x}_{t-1}^n\|^2 + \frac{1}{2\alpha}[\psi_t + 2(2R + \lambda_{\max}\rho)\pi_t + \phi_t^n] \\ &\quad + \frac{1}{2\beta}(\Psi_t^n + 4R\pi_t).\end{aligned} \tag{52}$$

Note that

$$\begin{aligned}&-[\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\gamma g_t^n(\mathbf{x}_t^n) \\ &\overset{(a)}{\leq} -\tilde{\Delta}_t^n + \gamma^2[g_t^n(\mathbf{x}_t^n)]^2 - \gamma^2 g_{t-1}^n(\mathbf{x}_{t-1}^n) g_t^n(\mathbf{x}_t^n) \\ &\overset{(b)}{=} -\tilde{\Delta}_t^n + \frac{\gamma^2}{2}[g_t^n(\mathbf{x}_t^n)]^2 - \frac{\gamma^2}{2}[g_{t-1}^n(\mathbf{x}_{t-1}^n)]^2 \\ &\quad + \frac{\gamma^2}{2}[g_t^n(\mathbf{x}_t^n) - g_t^n(\mathbf{x}_{t-1}^n) + g_t^n(\mathbf{x}_{t-1}^n) - g_{t-1}^n(\mathbf{x}_{t-1}^n)]^2 \\ &\overset{(c)}{\leq} -\tilde{\Delta}_t^n + \Phi_t^n + \gamma^2 L^2\|\mathbf{x}_t^n - \mathbf{x}_{t-1}^n\|^2 + \Omega_t^n\end{aligned} \tag{53}$$

where $\Phi_t^n \triangleq \frac{\gamma^2}{2}[g_t^n(\mathbf{x}_t^n)]^2 - \frac{\gamma^2}{2}[g_{t-1}^n(\mathbf{x}_{t-1}^n)]^2$ and $\Omega_t^n \triangleq \gamma^2[g_t^n(\mathbf{x}_{t-1}^n) - g_{t-1}^n(\mathbf{x}_{t-1}^n)]^2$. Here, (a) follows from (49) in Lemma 4, (b) is because $ab = \frac{1}{2}[a^2 + b^2 - (a-b)^2]$, and (c) follows from $\frac{1}{2}(a+b)^2 \leq a^2 + b^2$ and $g_t^n(\mathbf{x})$ being L-Lipschitz continuous in (46).

Substituting (27) and (53) into (52), multiplying both sides by w^n, summing over $n = 1$ to N, taking expectation, noting that $\mathbb{E}\{f_t(\mathbf{x})\} = \mathbb{E}\{f_{t-1}(\mathbf{x})\}$, and then summing over $t = 2$ to T, on the condition that $\beta \leq \frac{1}{2\gamma^2 L^2}$, we have

$$\begin{aligned}&\sum_{t=1}^{T-1}\mathbb{E}\{f_t(\hat{\mathbf{x}}_t)\} - \sum_{t=2}^{T}\mathbb{E}\{f_t(\mathbf{x}_t^\star)\} \\ &\leq \frac{D^2\alpha}{2}T + \sum_{t=2}^{T}\sum_{n=1}^{N} w^n\mathbb{E}\left\{[\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\gamma g_t^n(\mathbf{x}_t^\star)\right\} \\ &\quad + \sum_{t=2}^{T}\sum_{n=1}^{N} w^n\mathbb{E}\left\{-\tilde{\Delta}_t^n + \Phi_t^n + \Omega_t^n + \frac{1}{2\alpha}\phi_t^n + \frac{1}{2\beta}\Psi_t^n\right\} \\ &\quad + \frac{1}{2\alpha}\sum_{t=2}^{T}\mathbb{E}\{\psi_t\} + \left(\frac{2R+\lambda_{\max}\rho}{\alpha} + \frac{2R}{\beta}\right)\sum_{t=2}^{T}\mathbb{E}\{\pi_t\}.\end{aligned} \tag{54}$$

We now bound the RHS of (54). Note that $\tilde{Q}_t^n + \gamma g_t^n(\mathbf{x}_t^n) \geq 0$ for any t in (48). Then, from $\mathbf{x}_t^\star$ being independent of $\tilde{Q}_{t-1}^n$ and $g_{t-1}^n(\mathbf{x}_{t-1}^n)$, and the iterated law of expectation, we can show that $\mathbb{E}\{[\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\gamma g_t^n(\mathbf{x}_t^\star)\} \leq 0$ for any $t > 1$. Note that $-\mathbb{E}\{\tilde{\Delta}_t^n\}$, $\mathbb{E}\{\Phi_t^n\}$, $\mathbb{E}\{\Psi_t^n\}$, and $\mathbb{E}\{\psi_t\}$ are all telescoping terms such that their sum over $t = 2$ to T are upper bounded by $\frac{1}{2}\mathbb{E}\{(\tilde{Q}_1^n)^2\} = 0$, $\frac{1}{2}\mathbb{E}\{\gamma^2[g_t^n(\mathbf{x}_T^n)]^2\} \leq \frac{1}{2}\gamma^2 G^2$, $\mathbb{E}\{\|\mathbf{x}_2^\star - \mathbf{x}_1^n\|^2\} \leq R^2$, and $\mathbb{E}\{\|\mathbf{x}_2^\star - \hat{\mathbf{x}}_1\|^2\} \leq R^2$, respectively. Then, from (34) and the definitions of Π_T and Θ_T, we have

$$\begin{aligned}&\sum_{t=1}^{T-1}\mathbb{E}\{f_t(\hat{\mathbf{x}}_t)\} - \sum_{t=2}^{T}\mathbb{E}\{f_t(\mathbf{x}_t^\star)\} \\ &\qquad \leq \frac{D^2\alpha}{2}T + \frac{G^2\gamma^2}{2} + \frac{R^2 + \rho^2\Lambda_{2,T} + 4R\rho\Lambda_T}{2\alpha} \\ &\qquad\quad + \left(\frac{2R + \lambda_{\max}\rho}{\alpha} + \frac{2R}{\beta}\right)\Pi_T + \gamma^2\Theta_T + \frac{R^2}{2\beta}.\end{aligned}$$

Adding $f_T(\hat{\mathbf{x}}_T) - f_1(\mathbf{x}_1^\star)$ on both sides of the above inequality, we complete the proof. ■

Theorem 3 provides a general bound on the accumulated training loss yielded by OMUAA-DR. The following corollary provides the accumulated training loss of OMUAA-DR when the step-size parameters α, β, and γ and the power-scaling factors $\{\lambda_t\}$ take specific values.

Corollary 3. For any $\epsilon > 0$, set $\alpha = \epsilon$, $\gamma^2 = \frac{1}{\epsilon}$, $\beta = \frac{\epsilon}{2L^2}$, and $\lambda_t = \epsilon^2, \forall t$. The accumulated training loss yielded by OMUAA-DR is upper bounded by

$$\frac{1}{T}\sum_{t=1}^{T} f_t(\hat{\mathbf{x}}_t) \leq f^\star + \mathcal{O}((1+\rho^2+\Pi_T\rho+\Theta_T)\epsilon), \ \forall T \geq \frac{1}{\epsilon^2}. \tag{55}$$

Corollary 3 states that for all $T \geq \frac{1}{\epsilon^2}$, the accumulated training loss yielded by OMUAA-DR over noisy channels is within $\mathcal{O}((1+\rho^2+\Pi_T\rho+\Theta_T)\epsilon)$ to the optimum achieved under noiseless channels. Compared with the bound in Corollary 1 for OMUAA, (55) has an additional Θ_T term, which measures the accumulated variation of the transmit power constraint functions. Note that Θ_T can be small when the channels are relatively stable over time. Particularly, we have $\Theta_T = 0$ for static channels.

2) Bounding the Long-Term Transmit Power: We now provide an upper bound on the individual long-term transmit power constraint violations by OMUAA-DR in the following theorem.

Theorem 4. For any $\alpha, \beta, \gamma > 0$, the violation of each individual long-term transmit power constraint yielded by OMUAA-DR is upper bounded by

$$\frac{1}{T}\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq \frac{2G}{T} + \frac{2\gamma^2G^2+DR}{\gamma^2\bar{P}^nT} + \frac{(R+\lambda_{\max}\rho)^2}{2\alpha\gamma^2\bar{P}^nT} + \frac{R^2}{2\beta\gamma^2\bar{P}^nT}. \tag{56}$$

Proof: Since $\mathbf{x}_t^n$ minimizes the objective of $\mathbf{P3}^n$ over $\mathcal{X}$, which contains $\mathbf{0}$, and $g_t^n(\mathbf{0}) = -\bar{P}^n, \forall t > 1$, we have

$$\begin{aligned}
&\langle \nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}\|^2 \\
&\quad + [\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\gamma g_t^n(\mathbf{x}_t^n) + \frac{1}{2\beta}\|\mathbf{x}_t^n - \mathbf{x}_{t-1}^n\|^2 \\
&\leq \langle \nabla f_t^n(\hat{\mathbf{x}}_{t-1}), -\hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\hat{\mathbf{x}}_{t-1}\|^2 \\
&\quad - \gamma[\tilde{Q}_{t-1}^n + \gamma g_{t-1}^n(\mathbf{x}_{t-1}^n)]\bar{P}^n + \frac{1}{2\beta}\|\mathbf{x}_{t-1}^n\|^2.
\end{aligned} \tag{57}$$

Rearranging the terms of (57), we have

$$\begin{aligned}
&\gamma\tilde{Q}_{t-1}^n g_t^n(\mathbf{x}_t^n) \\
&\overset{(a)}{\leq} -\gamma\tilde{Q}_{t-1}^n\bar{P}^n + \gamma^2|g_{t-1}^n(\mathbf{x}_{t-1}^n)|\bar{P}^n - \gamma^2 g_{t-1}^n(\mathbf{x}_{t-1}^n)g_t^n(\mathbf{x}_t^n) \\
&\quad - \langle \nabla f_t^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x}_t^n\rangle + \frac{1}{2\alpha}\|\hat{\mathbf{x}}_{t-1}\|^2 + \frac{1}{2\beta}\|\mathbf{x}_{t-1}^n\|^2 \\
&\overset{(b)}{\leq} -\gamma\tilde{Q}_{t-1}^n\bar{P}^n + \gamma^2 G\bar{P}^n + \gamma^2G^2 + DR \\
&\quad + \frac{(R+\lambda_{\max}\rho)^2}{2\alpha} + \frac{R^2}{2\beta}
\end{aligned} \tag{58}$$

where (a) is because $-g_{t-1}^n(\mathbf{x}_{t-1}^n)\bar{P}^n \leq |g_{t-1}^n(\mathbf{x}_{t-1}^n)|\bar{P}^n$, and (b) follows from the bounds on $\nabla f_t^n(\mathbf{x})$, $g_t^n(\mathbf{x})$, $\mathbf{n}_t$, and $\mathcal{X}$ in (18), (19), (20), and (28) respectively.

Substituting (58) into (49) and noting that $[g_t^n(\mathbf{x}_t^n)]^2 \leq G^2$, we have

$$\begin{aligned}
\tilde{\Delta}_t^n \leq &-\gamma\tilde{Q}_{t-1}^n\bar{P}^n + \gamma^2G\bar{P}^n + 2\gamma^2G^2 + DR \\
&+ \frac{(R+\lambda_{\max}\rho)^2}{2\alpha} + \frac{R^2}{2\beta}.
\end{aligned} \tag{59}$$

Thus, a sufficient condition for $\tilde{\Delta}_t^n < 0$ is

$$\tilde{Q}_{t-1}^n > \gamma G + \frac{2\gamma^2G^2+DR}{\gamma\bar{P}^n} + \frac{(R+\lambda_{\max}\rho)^2}{2\alpha\gamma\bar{P}^n} + \frac{R^2}{2\beta\gamma\bar{P}^n}. \tag{60}$$

If (60) holds, we have $\tilde{Q}_t^n < \tilde{Q}_{t-1}^n$; otherwise, the maximum increase from $\tilde{Q}_{t-1}^n$ to $\tilde{Q}_t^n$ is γG since $\gamma g_t^n(\mathbf{x}_t^n) \leq \gamma G$. Therefore the virtual queue is bounded for any $t > 1$ by

$$\tilde{Q}_t^n \leq 2\gamma G + \frac{2\gamma^2G^2+DR}{\gamma\bar{P}^n} + \frac{(R+\lambda_{\max}\rho)^2}{2\alpha\gamma\bar{P}^n} + \frac{R^2}{2\beta\gamma\bar{P}^n}. \tag{61}$$

From the virtual queue dynamics in (42), we have $\tilde{Q}_t^n \geq \tilde{Q}_{t-1}^n + \gamma g_t^n(\mathbf{x}_t^n), \forall t > 1$. Summing it over $t = 2$ to T and rearranging the terms, we have $\gamma\sum_{t=2}^{T} g_t^n(\mathbf{x}_t^n) \leq \sum_{t=2}^{T}(\tilde{Q}_t^n - \tilde{Q}_{t-1}^n) = \tilde{Q}_T^n - \tilde{Q}_1^n = \tilde{Q}_T^n$. Further noting that $g_1^n(\cdot) \equiv 0$ by initialization, we have $\frac{1}{T}\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq \frac{\tilde{Q}_T^n}{\gamma T}$. Substituting the virtual queue upper bound (61) into the above inequality, we have (56). ■

Following Theorem 4, for specific values of the step-size parameters α, β, and γ and the power-scaling factors $\{\lambda_t\}$, we have the following corollary.

Corollary 4. For any $\epsilon > 0$, set $\alpha = \epsilon$, $\gamma^2 = \frac{1}{\epsilon}$, $\beta = \frac{\epsilon}{2L^2}$, and $\lambda_t = \epsilon^2, \forall t$. The individual long-term transmit power constraint violations yielded by OMUAA-DR is upper bounded by

$$\frac{1}{T}\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq \mathcal{O}((1+\rho^2)\epsilon), \quad \forall n, \quad \forall T \geq \frac{1}{\epsilon}. \tag{62}$$

Compared with the transmit power constraint violation bound of OMUAA in Corollary 2, we observe that OMUAA-DR reduces the $\mathcal{O}(\frac{1}{\epsilon^3})$ convergence time of OMUAA to $\mathcal{O}(\frac{1}{\epsilon})$.

Remark 4. Compared with OMUAA, OMUAA-DR requires an additional algorithm parameter β due to its double regularization, and it also requires an additional Assumption 4 on the Lipschitz continuity of the constraint function for its performance analysis. Furthermore, the upper bound on the accumulated training loss yielded by OMUAA-DR in Corollary 3 has an additional Θ_T term that measures the accumulated variation of the transmit power constraint function, compared with the one provided by OMUAA in Corollary 1. However, OMUAA-DR substantially reduces the convergence time of the long-term transmit power constraint violations in Corollary 4. Later in Section VII, we will further show that OMUAA-DR provides better learning performance and faster convergence on the long-term transmit power than OMUAA.

VI. Extension to Analog Gradient Aggregation

In the previous sections, we have considered OMUAA and OMUAA-DR with analog model aggregation, where the edge server receives the aggregated global model from the mobile devices. We now extend both OMUAA and OMUAA-DR to the case of analog gradient aggregation, where the edge server receives the aggregated gradient from the mobile devices to reconstruct the aggregated global model.

A. Over-the-Air Analog Gradient Aggregation

In the presence of channel noise in over-the-air FL, the standard local gradient descent model update in (8) for error-free FL becomes $\mathbf{x}_t^n = \hat{\mathbf{x}}_{t-1} - \alpha\nabla f_t^n(\hat{\mathbf{x}}_{t-1})$, where $\hat{\mathbf{x}}_{t-1}$ is the noisy global model. For analog gradient aggregation, each mobile device n generates its transmitted signals

$$\tilde{\mathbf{s}}_t^n = \frac{1}{\tilde{\lambda}_t} w^n \mathbf{b}_t^n \circ \big(- \alpha \nabla f_t^n(\hat{\mathbf{x}}_{t-1}) \big) \tag{63}$$

by replacing the local model $\mathbf{x}_t^n$ in $\mathbf{s}_t^n$ (6) with the local gradient (or the model difference) $-\alpha \nabla f_t^n(\hat{\mathbf{x}}_{t-1}) = \mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}$.

The received signals at the edge server is given by

$$\tilde{\mathbf{y}}_t = \sum_{n=1}^{N} \mathbf{h}_t^n \circ \tilde{\mathbf{s}}_t^n + \mathbf{z}_t = -\frac{\alpha}{\tilde{\lambda}_t} \sum_{n=1}^{N} w^n \nabla f_t^n(\hat{\mathbf{x}}_{t-1}) + \mathbf{z}_t. \tag{64}$$

Different from analog model aggregation, the edge server needs to keep the previous global model $\hat{\mathbf{x}}_{t-1}$ in order to recover the aggregated global model, given by

$$\begin{aligned}\hat{\mathbf{x}}_t &= \hat{\mathbf{x}}_{t-1} + \Re\{\tilde{\lambda}_t \tilde{\mathbf{y}}_t\} = \hat{\mathbf{x}}_{t-1} - \alpha \sum_{n=1}^{N} w^n \nabla f_t^n(\hat{\mathbf{x}}_{t-1}) + \tilde{\lambda}_t \mathbf{n}_t \\ &= \hat{\mathbf{x}}_{t-1} + \sum_{n=1}^{N} w^n (\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}) + \tilde{\lambda}_t \mathbf{n}_t = \mathbf{x}_t + \tilde{\lambda}_t \mathbf{n}_t. \end{aligned} \tag{65}$$

From (65), we can see that gradient aggregation by transmitting $-\alpha \nabla f_t^n(\hat{\mathbf{x}}_{t-1})$ is equivalent to model difference aggregation by transmitting $\mathbf{x}_t^n - \hat{\mathbf{x}}_{t-1}$ in $\tilde{\mathbf{s}}_t$. Furthermore, comparing (65) with the recovered noisy global model via model aggregation in (7), we can see that their only difference at the edge server side is the power scaling factor $\tilde{\lambda}_t$.

For analog gradient aggregation to have better learning performance than analog model aggregation, the power scaling factor $\tilde{\lambda}_t$ in (65) should to be smaller than λ_t in (7) so that the recovered global model $\hat{\mathbf{x}}_t$ is less impacted by the noise $\mathbf{n}_t$. Under the same transmit power, *i.e.*, $\|\mathbf{s}_t^n\|^2 = \|\tilde{\mathbf{s}}_t^n\|^2$, for $\tilde{\lambda}_t$ to be smaller than λ_t, we need $\alpha \|\mathbf{b}_t^n \circ \nabla f_t^n(\hat{\mathbf{x}}_{t-1})\| < \|\mathbf{b}_t^n \circ \mathbf{x}_t^n\|$. This implies that when the step size α is small or the gradient parameters $\nabla f_t^n(\hat{\mathbf{x}}_{t-1})$ are smaller than the model parameters $\mathbf{x}_t^n$, analog gradient aggregation can lead to smaller error in the recovered global model, which in turn can improve the learning performance.

Remark 5. In practical systems, both gradient aggregation and model aggregation can be useful in different applications. Gradient aggregation can be prone to gradient leakage attack [52], [53], where the attackers utilize the gradient to recover the private local data. Model aggregation makes it difficult for the attackers to reconstruct the gradient, as it requires prior knowledge of the learning rate, previous model, and some other possible advanced mechanisms such as momentum, multi-step gradient descent, and rate decaying, which may not be available to the attackers in practical FL systems [54].

B. Enabling Analog Gradient Aggregation in OMUAA

We now show how to extend both OMUAA and OMUAA-DR to enable analog gradient aggregation.

1) OMUAA (Gradient Aggregation): Since the transmit power at each mobile device n changes from $\|\mathbf{s}_t^n\|^2$ for model aggregation to $\|\tilde{\mathbf{s}}_t^n\|^2$, we redefine the long-term transmit power constraint function in **P1** as

$$\tilde{g}_t^n(\mathbf{x}) = \frac{(w^n)^2}{\tilde{\lambda}_t^2} \|\mathbf{b}_t^n \circ (\mathbf{x} - \hat{\mathbf{x}}_{t-1})\|^2 - \bar{P}^n \tag{66}$$

which is still a convex function of the model $\mathbf{x}$. The virtual queue updating rule for gradient aggregation is the same as the one for model aggregation in (14), except for using the above constraint function $\tilde{g}_t^n(\mathbf{x})$ in stead of $g_t^n(\mathbf{x})$ in (13).

Replacing $g_t^n(\mathbf{x})$ in $\mathbf{P2}^n$ with $\tilde{g}_t^n(\mathbf{x})$ and then taking the gradient of the objective function, we have

$$\nabla f_t^n(\hat{\mathbf{x}}_{t-1}) + \frac{1}{\alpha}(\mathbf{x} - \hat{\mathbf{x}}_{t-1}) + \boldsymbol{\theta}_t^n \circ (\mathbf{x} - \hat{\mathbf{x}}_{t-1}) \tag{67}$$

where $\boldsymbol{\theta}_t^n \in \mathbb{R}^d$ is redefined with the i-th entry $\theta_t^{n,i} = \frac{2\gamma Q_{t-1}^n (w^n)^2}{\tilde{\lambda}_t^2 |h_t^{n,i}|^2}$. Setting the gradient in (67) to zero to solve for $\mathbf{x}$ and then projecting it onto the affine set $\mathbf{x}$, we have a closed-form model update $\mathbf{x}_t^n$, given by

$$\mathbf{x}_t^n = \Big[\hat{\mathbf{x}}_{t-1} - \frac{\alpha}{\mathbf{1} + \alpha \boldsymbol{\theta}_t^n} \circ \nabla f_t^n(\hat{\mathbf{x}}_{t-1}) \Big]_{-\mathbf{x}_{\max}}^{\mathbf{x}_{\max}}. \tag{68}$$

The local model update in (68) can be seen as local gradient descent with entry-wise step size $\frac{\alpha}{1+\theta_t^{n,i}}$ that depends on the ratio of the virtual queue and and the individual channel power.

2) OMUAA-DR (Gradient Aggregation): Replacing the constraint function $g_t^n(\mathbf{x})$ in the virtual queue updating rule (42) and in $\mathbf{P3}^n$ with the new constraint function $\tilde{g}_t^n(\mathbf{x})$ in (66), we can derive a closed-form local model update for $\mathbf{x}_t^n$, given by

$$\begin{aligned}\mathbf{x}_t^n = \Bigg[\Big(\frac{\alpha + \beta}{\beta} \mathbf{1} + \alpha \tilde{\boldsymbol{\theta}}_t^n \Big)^{-1} &\circ \Big((\mathbf{1} + \alpha \tilde{\boldsymbol{\theta}}_t^n) \circ \hat{\mathbf{x}}_{t-1} \\ &+ \frac{\alpha}{\beta} \mathbf{x}_{t-1}^n - \alpha \nabla f_t^n(\hat{\mathbf{x}}_{t-1}) \Big) \Bigg]_{-\mathbf{x}_{\max}}^{\mathbf{x}_{\max}} \end{aligned} \tag{69}$$

where the i-th entry of $\tilde{\boldsymbol{\theta}}_t^n$ is redefined as $\tilde{\theta}_t^{n,i} = \frac{2\gamma [\tilde{Q}_{t-1}^n + \gamma \tilde{g}_{t-1}^n(\mathbf{x}_{t-1}^n)](w^n)^2}{\tilde{\lambda}_t^2 |h_t^{n,i}|^2}$.

Compared with the local model update (45) in the original OMUAA-DR, there is an additional weight $\alpha \tilde{\boldsymbol{\theta}}_t^n$ on the previous local model $\hat{\mathbf{x}}_{t-1}$ caused by the new constraint function $\tilde{g}_t^n(\mathbf{x})$, which can be viewed as an additional regularization term on $\hat{\mathbf{x}}_{t-1}$ in $\mathbf{P3}^n$.

Remark 6. The only difference between analog gradient aggregation and analog model aggregation in OMUAA and OMUAA-DR is the definition of the long-term constraint function. Note that our performance analysis in Theorems 1-4 relies on general assumptions on the output of the constraint function in Assumption 2 and the Lipschitz continuity of the constraint function in Assumption 4. Therefore, both OMUAA (gradient aggregation) and OMUAA-DR (gradient aggregation) yields the same performance bounds as their original versions in Theorem 1-4.

VII. Numerical Performance Evaluation

To complement the theoretical performance guarantees of OMUAA and OMUAA-DR provided in Sections IV-B and V-B, we evaluate the performance of OMUAA and OMUAA-DR in edge learning based on real-world image classification datasets for both convex logistic regression and non-convex neural network training, under common wireless network settings.

A. Simulation Setup

We consider a wireless edge network with one edge server and $N = 10$ mobile devices. We assume an orthogonal frequency-division multiplexing system with $S = 500$ subcarriers, each with bandwidth $B_W = 15$ kHz. Following typical wireless specifications [55], we set noise power spectral density $N_0 = -174$ dBm/Hz and noise figure $N_F = 10$ dB. The fading channel from mobile device n to the edge server at the t-th iteration is modeled as $\mathbf{h}_t^n \sim \mathcal{N}(\mathbf{0}, \xi^n \mathbf{I})$, with ξ^n representing the large-scale fading variation consisting of path-loss and shadowing. We set $\xi^n[\text{dB}] = -31.54 - 33\log_{10}(r) - \varphi^n$ [55], where r is the distance to the edge server, and $\varphi^n \sim \mathcal{N}(0, \sigma_\phi^2)$ is the shadowing term with $\sigma_\phi^2 = 8$ dB. We set $r = 100$ m by default. 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 [56] 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.,* $v \in \{1, \dots, J\}$. 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 $w^n = \frac{1}{N}$.

We compare OMUAA and OMUAA-DR with the following schemes.[6]

- *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 and OMUAA-DR.
- *OTA FL:* We adopt the transmit power control scheme in [21], [22], which are the best existing alternatives that consider over-the-air (OTA) FL with long-term transmit power constraints. In [21] and [22], 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 [21], we set P_t equal to the average transmit power limit at each iteration t as in [22].
- *R-OTA FL:* Based on OTA FL, we add a regularization term $\kappa\|\mathbf{x}\|^2$ to $l(\mathbf{x}; \mathbf{u}, v)$, where κ is a tunable parameter. This regularization scheme was adopted in [23]-[25]. We have optimized κ in the presented results.
- *OTA FL (GA):* Instead of using model aggregation, we adopt the gradient (or model difference) aggregation approach in [20].

B. Convex Loss: Logistic Regression

We consider the cross-entropy loss for multinomial logistic regression, given by $l(\mathbf{x}; \mathbf{u}, v) = -\sum_{j=1}^{J} 1\{v = j\} \log \frac{\exp(\langle \mathbf{x}[j], \mathbf{u}\rangle)}{\sum_{k=1}^{J} \exp(\langle \mathbf{x}[k], \mathbf{u}\rangle)}$, where $\mathbf{x} = [\mathbf{x}[1]^T, \dots, \mathbf{x}[J]^T]^T$ 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.,* $\bar{P}^n = M\bar{P}, \forall n$. Our performance metrics are the time-averaged test accuracy over $\mathcal{E}$ given by $\bar{A}(T) = \frac{1}{T|\mathcal{E}|} \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{E}|} 1\{\arg\max_j \{\frac{\exp(\langle \hat{\mathbf{x}}_t[j], \mathbf{u}^i\rangle)}{\sum_{k=1}^{J} \exp(\langle \hat{\mathbf{x}}_t[k], \mathbf{u}^i\rangle)}\} = v^i\}$, the time-averaged training loss over $\{\mathcal{B}_t^n\}$ given by $\bar{f}(T) = \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} \frac{1}{|\mathcal{B}_t^n|} \sum_{i=1}^{|\mathcal{B}_t^n|} w^n l(\hat{\mathbf{x}}_t; \mathbf{u}_t^{n,i}, v_t^{n,i})$, and the time-averaged transmit power given by $\bar{P}(T) = \frac{1}{TN} \sum_{t=1}^{T} \sum_{n=1}^{N} \|\mathbf{s}_t^n\|^2$.

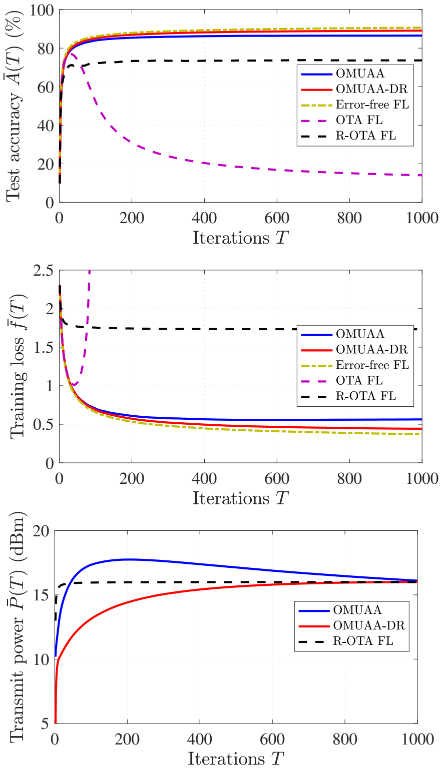


Fig. 2. Test accuracy $\bar{A}(T)$, training loss $\bar{f}(T)$, and transmit power $\bar{P}(T)$ vs. iterations T.

For step-size parameter tuning, we first try several values of the gradient descent step-size parameter for error-free FL. We find that the step-size of value 1×10^{-5} provides error-free FL with the best performance among various trial values. Therefore, for the purpose of comparison among different schemes, we also set $\alpha = 1 \times 10^{-5}$ in OMUAA and use the same step-size in OTA FL, R-OTA FL and OTA FL (GA). For OMUAA-DR, we have studied several combinations of step-size parameters α and β, and present here the case $\alpha = \beta = 2 \times 10^{-5}$. Similarly, we set the step-size parameter $\gamma = 0.02$ and the power-scaling factor $\lambda = 5 \times 10^{-4}$ for both OMUAA and OMUAA-DR after numerical tuning. In practice, α, β, γ, λ may depend on both the learning problem and the communication system. They can be problem-dependent and may be treated as hyper parameters that require tuning.

Fig. 2 shows $\bar{A}(T)$, $\bar{f}(T)$, and $\bar{P}(T)$ versus T with $\bar{P} = 16$ dBm. Despite the presence of communication noise, OMUAA and OMUAA-DR converge quickly and achieve better classification performance compared with OTA FL and

[6] Our codes are available at https://github.com/juncheng-wang/OMUAA.

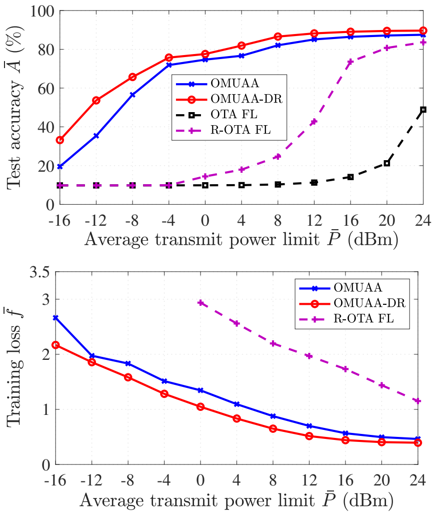


Fig. 3. The impact of average transmit power limit $\bar{P}$. The $\bar{f}$ plot for OTA FL is not included as its value of $\bar{f}$ is much larger than those of OMUAA and R-OTA FL.

R-OTA FL. We observe that the performance of OTA FL deteriorates as T increases. This is because OTA FL relies on the power-scaling factor λ_t for transmit power control, instead of optimizing the local model $\mathbf{x}_t^n$ as how OMUAA and OMUAA-DR do to reduce the transmit power. In the low power region, λ_t yielded by OTA FL becomes large and magnifies the communication error $\lambda_t \mathbf{n}_t$ in the global model $\hat{\mathbf{x}}_t$ in (7). 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 queues in OMUAA and OMUAA-DR serve 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.

Furthermore, Fig. 2 shows that, unlike the case of OMUAA, $\bar{P}(T)$ yielded by OMUAA-DR does not overshoot the long-term transmit power limit $\bar{P}$. It also converges faster due to the new constraint penalty with double regularization in OMUAA-DR. This confirms the results in Corollary 4.

In Fig 3, we compare the steady-state test accuracy $\bar{A}$ and training loss $\bar{f}$ among OMUAA, OMUAA-DR, 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. This is because when $\bar{P}$ is small, the scaled channel noise $\lambda_t \mathbf{n}_t$ becomes large in the noisy global model $\hat{\mathbf{x}}_t$ (7), which in turn causes error propagation in the training process that deteriorates the learning performance. 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 and OMUAA-DR significantly outperforms the other two schemes. Furthermore, OMUAA-DR achieves higher $\bar{A}$ than OMUAA, especially in the low power regime.

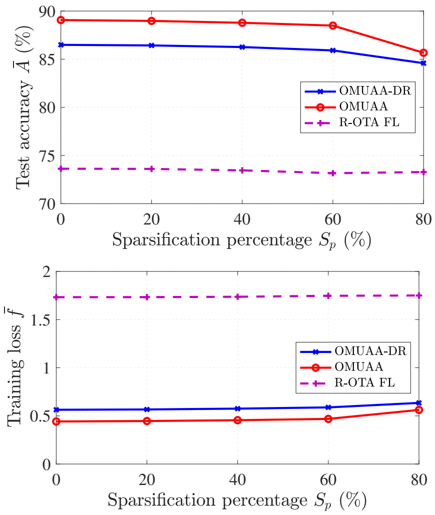


Fig. 4. The impact of sparsification percentage S_p.

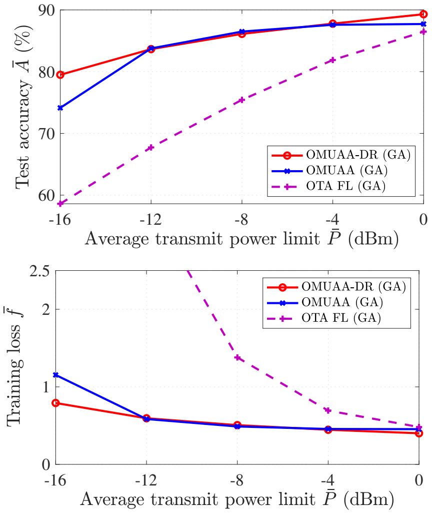


Fig. 5. The impact of average transmit power limit $\bar{P}$ on gradient aggregation.

The sparsification and quantization techniques considered in [21] and [22] are orthogonal to the OMUAA design. However, these techniques can be easily combined with OMUAA and OMUAA-DR to further reduce the amount of required communication resources. Here, we use sparsification as an illustration. Let S_p be the sparsification level in percentage. Specifically, after obtaining the local model $\mathbf{x}_t^n$ via OMUAA or OMUAA-DR, each mobile device n finds the S_p percent of model parameters with the smallest values and sets them to zeros. Fig. 4 shows the impact of the sparsification percentage S_p on $\bar{A}$ and $\bar{f}$ yielded by OMUAA, OMUAA-DR, and R-OTA FL. The average transmit power limit is set to $\bar{P} = 16$ dBm. We observe that OMUAA and OMUAA-DR substantially outperform R-OTA FL for a wide range of S_p values. Furthermore, OMUAA-DR yields better learning performance than OMUAA.

We study the impact of average transmit power limit $\bar{P}$ on the performance of OMUAA-DR (GA), OMUAA (GA), and OTA FL (GA). We can see from Fig. 5 that OMUAA-DR (GA) and OMUAA (GA) substantially outperforms OTA FL (GA), showing the benefit of joint optimization of computation and communication over the separate optimization approach. Compared with the model aggregation performance in Fig. 3, we can see that using gradient aggregation significantly improves the learning performance in the low power region. As explained in Section VI-A, when the step size is small as in our simulation, gradient aggregation yields a smaller power scaling factor than model aggregation, leading to more accurate global models at the edge server and thus better learning performance.

C. Non-Convex Loss: Neural Network Training

The performance bounds derived in Sections IV-B and V-B requires convexity in the loss functions. However, both OMUAA and OMUAA-DR can be directly applied to wireless federated learning with non-convex loss functions. To evaluate their performance in this more general scenario, here we consider a fully connected two-layer neural network with 784 pixel as input, 10 neurons in the hidden layer, and 10 neurons in the output layer, such that the number of parameters $7,940$ is similar to that of logistic regression in Section VII-B. We use the sigmoid activation function in the hidden layer and the softmax activation function in the output layer. In addition to the standard forward and backward propagation for gradient computation in neural network update, OMUAA and OMUAA-DR only require additional computation of the virtual queues (Q_t^n and $\tilde{Q}_t^n$) and transmit power scaling factors ($\boldsymbol{\theta}_t^n$ and $\tilde{\boldsymbol{\theta}}_t^n$). Both of them only has $\mathcal{O}(d)$ computational complexity. Therefore, both OMUAA and OMUAA-DR can be easily merged into the standard neural network update and is computationally efficient for practical neural network training.

To set the step-size parameters, similar to Section VII-B, we find 0.8 gradient descent step-size provides error-free FL with the best performance among various trial values. Therefore, we also set $\alpha = 0.8$ in OMUAA and use the same step-size in OTA FL and R-OTA FL for gradient descent. We set $\gamma = 2$ and $\lambda = 1 \times 10^{-6}$ in OMUAA, and $\alpha = \beta = 0.8$, $\gamma = 1$, and $\lambda = 1 \times 10^{-6}$ in OMUAA-DR.

Fig. 6 shows the time-averaged test accuracy $\bar{A}(T)$, time-averaged cross-entropy loss in the output layer $\bar{f}(T)$, and time-averaged transmit power $\bar{P}(T)$ versus T with $\bar{P} = 8$ dBm. We again observe that both OMUAA and OMUAA-DR converge quickly and substantially outperform OTA-FL and R-OTA FL in neural network training. Compared with OMUAA, the transmit power $\bar{P}(T)$ yielded by OMUAA-DR converges much faster. For neural network training with non-convex loss functions, gradient descent based algorithms such as OMUAA and OMUAA-DR generally converge to some local minima. In our simulation results, we observe that these local minima provide acceptable learning performance. Note that our performance analysis is for general machine learning problems and analog communication systems, and therefore it is based on some upper-bound constants that reflect the properties of the underlying computation and communication systems. The actual system performance yielded by OMUAA and OMUAA-DR is problem dependent and may not necessarily be close to their general theoretical bounds.

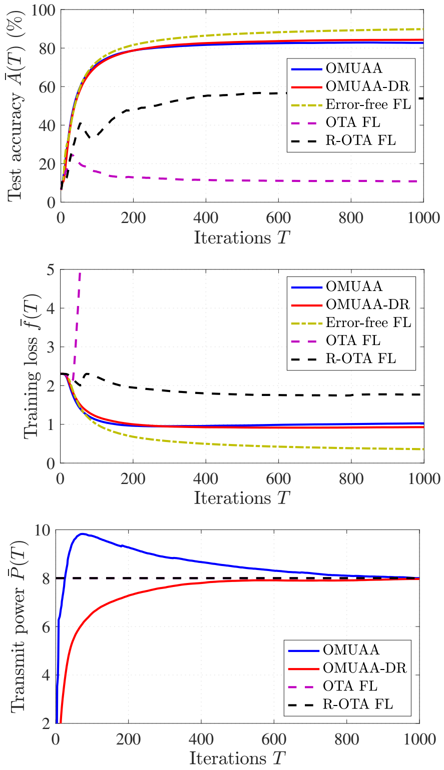


Fig. 6. Test accuracy $\bar{A}(T)$, training loss $\bar{f}(T)$, and transmit power $\bar{P}$ vs. iterations T.

VIII. Conclusions

We consider FL in wireless edge networks with analog aggregation over noisy wireless fading multiple access channels. We propose OMUAA and OMUAA-DR algorithms to minimize the accumulated training loss over time at the edge server, subject to individual long-term transmit power constraints at the mobile devices. Both algorithms depend only on the current local CSI, without needing to know the channel distribution. The local models yielded by OMUAA and OMUAA-DR 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. OMUAA-DR requires an additional step-size parameter and slightly more storage space than OMUAA, but it can substantially reduce the convergence time to reach the long-term transmit power constraint. Simulation results based on realistic wireless network settings and real-word image classification datasets show substantial performance advantage of OMUAA and OMUAA-DR over the known best alternatives for both convex logistic regression and non-convex neural network training under different scenarios.

References

[1] J. Wang, M. Dong, B. Liang, G. Boudreau, and H. Abou-Zeid, "Online model updating with analog aggregation in wireless edge learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2022.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Trans. Signal Process.*, vol. 69, pp. 3796–3811, Jun. 2021.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

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

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

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

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] A. Abdi and F. Fekri, "Reducing communication overhead via CEO in distributed training," in *Proc. IEEE Intel. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, 2019.

[38] R. R. Gajjala, S. Banchhor, A. M. Abdelmoniem, A. Dutta, M. Canini, and P. Kalnis, "Huffman coding based encoding techniques for fast distributed deep learning," in *Proc. Workshop on Distrib. Mach. Learn.*, 2020.

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

[40] M. Mahdavi, R. Jin, and T. Yang, "Trading regret for efficiency: Online convex optimization with long term constraints," *J. Mach. Learn. Res.*, vol. 13, pp. 2503–2528, Sep. 2012.

[41] R. Jenatton, J. Huang, and C. Archambeau, "Adaptive algorithms for online convex optimization with long-term constraints," in *Proc. Intel. Conf. Mach. Learn. (ICML)*, 2016.

[42] H. Yu and M. J. Neely, "A low complexity algorithm with $O(\sqrt{T})$ regret and $O(1)$ constraint violations for online convex optimization with long term constraints," *J. Mach. Learn. Res.*, vol. 21, pp. 1–24, Feb. 2020.

[43] H. Yu, M. J. Neely, and X. Wei, "Online convex optimization with stochastic constraints," in *Proc. Adv. Neural Info. Proc. Sys. (NIPS)*, 2017.

[44] X. Wei, H. Yu, and M. J. Neely, "Online primal-dual mirror descent under stochastic constraints," in *Proc. ACM Meas. Anal. Comput. Syst.*, 2020.

[45] T. Chen, Q. Ling, and G. B. Giannakis, "An online convex optimization approach to proactive network resource allocation," *IEEE Trans. Signal Process.*, vol. 65, pp. 6350–6364, Dec. 2017.

[46] X. Cao, J. Zhang, and H. V. Poor, "A virtual-queue-based algorithm for constrained online convex optimization with applications to data center resource allocation," *IEEE J. Sel. Topics Signal Process.*, vol. 12, pp. 703–716, Aug. 2018.

[47] D. Yuan, A. Proutiere, and G. Shi, "Distributed online optimization with long-term constraints," *IEEE Trans. Automat. Contr.*, vol. 67, pp. 1089–1104, Mar. 2022.

[48] X. Cao and T. Baar, "Distributed constrained online convex optimization over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 70, pp. 3468–3483, Jun. 2022.

[49] 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.

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

[51] H. Yu and M. J. Neely, "A low complexity algorithm with $O(\sqrt{T})$ regret and $O(1)$ constraint violations for online convex optimization with long term constraints," *J. Mach. Learn. Res.*, vol. 21, pp. 1–24, Feb. 2020.

[52] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. Adv. Neural Info. Proc. Sys. (NIPS)*, 2019.

[53] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" in *Proc. Adv. Neural Info. Proc. Sys. (NIPS)*, 2020.

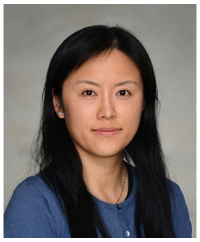
[54] F. Wang, E. Hugh, and B. Li, "More than enough is too much: Adaptive defenses against gradient leakage in production federated learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2023.

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

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

Juncheng Wang (Member, IEEE) received the B.Eng. degree in Electrical Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014, the M.Sc. degree in Electrical and Computer Engineering from the University of Alberta, Edmonton, AB, Canada, in 2017, and the Ph.D. degree in Electrical and Computer Engineering from the University of Toronto, Toronto, ON, Canada, in 2023. He is currently an Assistant Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China. His research interests include communication networks, online learning, distributed computing, and stochastic optimization.

Ben Liang (Fellow, IEEE) received honors-simultaneous B.Sc. (valedictorian) and M.Sc. degrees in Electrical Engineering from Polytechnic University (now the engineering school of New York University) in 1997 and the Ph.D. degree in Electrical Engineering with a minor in Computer Science from Cornell University in 2001. He was a visiting lecturer and postdoctoral research associate at Cornell University in the 2001 - 2002 academic year. He joined the Department of Electrical and Computer Engineering at the University of Toronto in 2002, where he is now Professor and L. Lau Chair in Electrical and Computer Engineering. His current research interests are in networked systems and mobile communications. He is an associate editor for the IEEE Transactions on Mobile Computing and has served on the editorial boards of the IEEE Transactions on Communications, the IEEE Transactions on Wireless Communications, and the Wiley Security and Communication Networks. He regularly serves on the organizational and technical committees of a number of conferences. He is a Fellow of IEEE and a member of ACM and Tau Beta Pi.



Min Dong (Senior Member, IEEE) received the B.Eng. degree from Tsinghua University, Beijing, China, in 1998, and the Ph.D. degree in electrical and computer engineering with a minor in applied mathematics from Cornell University, Ithaca, NY, in 2004. From 2004 to 2008, she was with Qualcomm Research, Qualcomm Inc., San Diego, CA. Since 2008, she has been with Ontario Tech University, where she is currently a Professor in the Department of Electrical, Computer and Software Engineering. She also holds a status-only Professor appointment with the Department of Electrical and Computer Engineering at the University of Toronto. Her research interests include wireless communications, statistical signal processing, learning techniques, optimization and control applications in cyber-physical systems.

Dr. Dong received the Early Researcher Award from the Ontario Ministry of Research and Innovation in 2012, the Best Paper Award at IEEE ICCC in 2012, and the 2004 IEEE Signal Processing Society Best Paper Award. She is a co-author of the Best Student Paper at IEEE SPAWC 2021 and the Best Student Paper of Signal Processing for Communications and Networking at IEEE ICASSP 2016. She is an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE OPEN JOURNAL of SIGNAL PROCESSING. She was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2010-2014) and the IEEE SIGNAL PROCESSING LETTERS (2009-2013). She was also on the Steering Committee of the IEEE TRANSACTIONS ON MOBILE COMPUTING (2019-2021). She was an elected member of the Signal Processing for Communications and Networking (SP-COM) Technical Committee of IEEE Signal Processing Society (2013-2018). She was the lead Co-Chair of the Communications and Networks to Enable the Smart Grid Symposium at the IEEE International Conference on Smart Grid Communications in 2014.

Gary Boudreau (Senior Member, IEEE) received a B.A.Sc. in Electrical Engineering from the University of Ottawa in 1983, an M.A.Sc. in Electrical Engineering from Queens University in 1984 and a Ph.D. in electrical engineering from Carleton University in 1989. From 1984 to 1989 he was employed as a communications systems engineer with Canadian Astronautics Limited and from 1990 to 1993 he worked as a satellite systems engineer for MPR Teltech Ltd. For the period spanning 1993 to 2009 he was employed by Nortel Networks in a variety of wireless systems and management roles within the CDMA and LTE basestation product groups. In 2010 he joined Ericsson Canada where he is currently Director of RAN Architecture and Performance in the North American CTO office. His interests include digital and wireless communications, signal processing and machine learning.

Hatem Abou-Zeid (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Queens University in 2014. From 2015 to 2017, he was at Cisco Systems designing scalable traffic engineering and routing solutions for service provider networks. Prior to joining the University of Calgary, he was at Ericsson Canada for 4 years leading 5G radio access designs and contributing to intellectual property development in the areas of RAN intelligence, low latency communications, spectrum sharing, and interference mitigation. Several wireless access algorithms and traffic engineering techniques that he co-invented and co-developed are deployed in 5G mobile networks worldwide. He is also an Adjunct Professor at Queens University, Ontario Tech University, and Carleton University in Canada.

Dr. Abou-Zeid's research interests are broadly in 6G wireless networking, robust artificial intelligence, and extended reality communications. His work has resulted in 19 patent filings and 60 journal and conference publications in several IEEE flagship venues, and he is a co-author of a Best Paper Award at IEEE ICC 2022. He is an avid supporter of industry-university partnerships, and he served on the Ericsson Government Industry Relations and Talent Development Committees where he directed multiple academic research partnerships. He also served on the TPC of several IEEE Conferences and was the Co-Chair of the IEEE ICC Workshop on Wireless Network Innovations for Mobile Edge Learning, and Corporate Co-Chair of the IEEE LCN Conference 2022. He received the Software Engineering Professor of the Year and Early Research Excellence Awards in 2023 at University of Calgary, was awarded a DAAD RISE Fellowship at Bell Labs, and nominated for an Outstanding Thesis Medal at Queens University.