Uplink Over-the-Air Aggregation for Multi-Model Wireless Federated Learning

Chong Zhang*, Min Dong[†], Ben Liang*, Ali Afana[‡], Yahia Ahmed[‡]

*Dept. of Electrical and Computer Engineering, University of Toronto, Canada, [‡]Ericsson Canada, Canada [†]Dept. of Electrical, Computer and Software Engineering, Ontario Tech University, Canada

Abstract—We propose an uplink over-the-air aggregation (OAA) method for wireless federated learning (FL) that simultaneously trains multiple models. To maximize the multi-model training convergence rate, we derive an upper bound on the optimality gap of the global model update, and then, formulate an uplink joint transmit-receive beamforming optimization problem to minimize this upper bound. We solve this problem using the block coordinate descent approach, which admits low-complexity closed-form updates. Simulation results show that our proposed multi-model FL with fast OAA substantially outperforms sequentially training multiple models under the conventional singlemodel approach.

I. INTRODUCTION

Federated learning (FL) [1] is a widely recognized method for multiple devices to collaboratively train machine learning models. However, FL in the wireless environment, usually with a base station (BS) taking the role of a parameter server, suffers from degraded performance due to limited wireless resources and signal distortion. This necessitates efficient communication design to effectively support wireless FL.

Most existing works on wireless FL have focused on training only a single model [2]–[8]. Various design schemes have been proposed to improve the communication efficiency of wireless FL, including transmission design of the downlink [2], uplink [3]–[5], and combined downlink-uplink [6]–[8]. However, in practice a system often needs to train multiple models. Directly using the conventional single-model FL schemes, to train the models sequentially one at a time, can cause substantial latency.

Simultaneously training multiple models in FL has recently been considered in [9], [10]. Under error-free communication, it was shown in [9] that multi-model FL can substantially improve the training convergence rate. Later, considering noisy downlink and uplink wireless channels, [10] proposed a multigroup multicast beamforming method to facilitate the downlink transmission of global models from the BS to the devices. However, [10] used the conventional orthogonal multiple access design for uplink model aggregation, which can consume large bandwidth and incur high latency as the number of devices becomes large. While over-the-air aggregation (OAA) has recently become popular for uplink design in single-model FL due to its bandwidth efficiency over orthogonal schemes [3]-[5], it has not been considered in multi-model FL, due to the substantial design challenges from additional inter-model interference and high computational complexity.

In this paper, we propose a computationally efficient uplink OAA method for multi-model wireless FL. Aiming to maximize the FL convergence rate, we derive an upper bound on the optimality gap of the FL global model update, capturing the impact of noisy transmission and inter-model interference. We then show that the minimization of this upper bound leads to a joint transmit-receive beamforming design to minimize the sum of inverse received SINRs subject to some power budget at the BS and devices. We solve this problem using block coordinate descent (BCD) and derive closed-form solutions to each subproblem, leading to a low-complexity design. Simulation under typical wireless network settings shows that the proposed multi-model FL design with fast OAA substantially outperforms the conventional single-model approach that sequentially trains one model at a time.

II. SYSTEM MODEL AND PROBLEM STATEMENT

A. Multi-Model FL System

We consider an FL system consisting of a server and K worker devices that collaboratively train M ML models. Let $\mathcal{K}_{\text{tot}} \triangleq \{1, \ldots, K\}$ denote the total set of devices and $\mathcal{M} \triangleq \{1, \ldots, M\}$ the set of models. Let $\boldsymbol{\theta}_m \in \mathbb{R}^{D_m}$ be the parameter vector of model m, which has D_m parameters.

Each device k holds local training datasets for all M models, with the dataset for model m being $S_m^k \triangleq \{(\mathbf{s}_{m,i}^k, v_{m,i}^k) : 1 \le i \le S_m^k\}$, where $\mathbf{s}_{m,i}^k \in \mathbb{R}^b$ is the *i*-th data feature vector and $v_{m,i}^k$ is the corresponding label. The local training loss function representing the training error at device k for model m is defined as $F_m^k(\boldsymbol{\theta}_m) = \frac{1}{S_m^k} \sum_{i=1}^{S_m^k} L_m(\boldsymbol{\theta}_m; \mathbf{s}_{m,i}^k, v_{m,i}^k)$, where $L_m(\cdot)$ is the sample-wise training loss for model m. The global training loss function for model m is a weighted average of $F_m^k(\boldsymbol{\theta}_m)$'s, given by $F_m(\boldsymbol{\theta}_m) = \frac{1}{\sum_{k=1}^K S_m^k} \sum_{k=1}^K S_m^k F_m^k(\boldsymbol{\theta}_m)$. The learning objective is to find optimal $\boldsymbol{\theta}_m^*$ that minimizes $F_m(\boldsymbol{\theta}_m)$ for each model $m \in \mathcal{M}$.

For multi-model FL, we consider the K devices train the M models simultaneously, and the model updates are exchanged with the server via multiple rounds of downlinkuplink wireless communication. In each communication round, each model is trained by a subset of devices. For simplicity, we assume $K/M \in \mathbb{N}$. We consider the round robin scheduling approach for device-model assignment [9], [10]. Specifically, we define a frame consisting of M communication rounds. At the beginning of each frame, the K devices are randomly partitioned into M equal-sized groups, denoted by $\mathcal{K}_1, \ldots, \mathcal{K}_M$. These device groups remain unchanged within a frame. For each communication round t within the frame, each device group i is assigned to train model $\hat{m}(i, t)$, given by

$$\hat{m}(i,t) = [(M+i - [t \mod M] - 1) \mod M] + 1.$$
(1)

Fig. 1 shows an example of the round-robin device-model assignment within a frame of three communication rounds for M = 3 models.

The iterative multi-model FL training procedure in round t, which is in frame $n = \lfloor t/M \rfloor$, is as follows:

- Downlink broadcast: The server broadcasts the current M global model parameter vectors $\theta_{m,t}$'s to their respective assigned device group.
- Local model update: Device k ∈ K_i performs local training of its assigned model m̂(i, t) using the corresponding local dataset S^k_{m̂(i,t)}. Suppose m̂(i,t) = μ. Device k divides S^k_μ into mini-batches, and applies the standard mini-batch stochastic gradient descent (SGD) algorithm with J iterations to generate the updated local model based on the received version of the global model θ̂^k_{μ,t}. In particular, let θ^{k,τ}_{μ,t} denote the local model update by device k ∈ K_i at iteration τ ∈ {0,...,J − 1}, with θ^{k,0}_{μ,t} = θ̂^k_{μ,t}, and let B^{k,τ}_{μ,t} ⊆ S^k_μ denote the mini-batch used at iteration τ. The local model update is given by

$$\boldsymbol{\theta}_{\mu,t}^{k,\tau+1} = \boldsymbol{\theta}_{\mu,t}^{k,\tau} - \frac{\eta_n}{|\mathcal{B}_{\mu,t}^{k,\tau}|} \sum_{(\mathbf{s},v)\in\mathcal{B}_{\mu,t}^{k,\tau}} \nabla L_{\mu}(\boldsymbol{\theta}_{\mu,t}^{k,\tau};\mathbf{s},v) \quad (2)$$

where η_n is the learning rate in frame n, and ∇L_{μ} is the gradient of the sample-wise training loss function for model μ w.r.t. $\boldsymbol{\theta}_{\mu,t}^{k,\tau}$.

• Uplink aggregation: The K devices send their updated local models $\theta_{m,t}^{k,J}$'s to the server using the uplink transmission. The server aggregates $\theta_{m,t}^{k,J}$, $k \in \mathcal{K}_i$, received from device group *i* to generate the global model $\theta_{m,t+1}$ for each $m \in \mathcal{M}$ for the next round t + 1.

B. Wireless Communication Model

We consider a practical wireless communication system where the server is hosted by a BS. Assume the BS is equipped with N antennas, and each device has a single antenna.

We assume downlink broadcast of M models to their respective device groups uses orthogonal channels among groups. The BS uses the multicast beamforming technique [11], [12] to send the model update $\theta_{\hat{m}(i,t),t}$ to its assigned device group *i*. Device *k* in group *i* then obtains an estimate of $\theta_{\hat{m}(i,t),t}$ [10]:

$$\hat{\boldsymbol{\theta}}_{\hat{m}(i,t),t}^{k} = \boldsymbol{\theta}_{\hat{m}(i,t),t} + \mathbf{n}_{k,t}^{\text{dl}}$$
(3)

where $\mathbf{n}_{k,t}^{\text{dl}} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{d}}^2 \mathbf{I})$ is the post-processed receiver noise due to the noisy downlink channel.

For uplink transmission and local model aggregation, we consider OAA to conserve system bandwidth. In particular, the *K* devices send their local model updates $\theta_{m,t}^{k,J}$'s to the BS simultaneously over a common uplink channel. The BS uses receive beamforming to aggregate the local models $\theta_{\hat{m}(i,t),t}^{k,J}$.

	I	\mathcal{K}_1	\mathcal{K}_2	\mathcal{K}_3
One frame	Round 0	Model 1	Model 2	Model 3
	Round 1	Model 3	Model 1	Model 2
	Round 2	Model 2	Model 3	Model 1

Fig. 1. Device-model round robin scheduling for M = 3 models.

 $k \in \mathcal{K}_i$, received from device group *i*, for i = 1, ..., M. Due to the analog nature of OAA, the devices must send the values of $\theta_{m,t}^{k,J}$'s directly under their transmit power budget.

In this paper, we focus on the uplink OAA design via joint transmit-receiver beamforming, aiming to maximize the learning performance of multi-model wireless FL in terms of the training convergence rate. Note that the downlink received models are noisy versions of $\theta_{m,t}$'s due to the noisy wireless channel, while the uplink received models are also distorted versions of $\theta_{m,t}^{k,J}$'s due to the noisy wireless channels and the inter-group interference. These errors further propagate in the model updates over the subsequent communication rounds, degrading the learning performance. Thus, an effective uplink OAA design must capture such errors generated in the complex interaction between learning and communication.

III. UPLINK OAA FOR MULTI-MODEL FL

A. Uplink Aggregation Framework

We propose an uplink aggregation framework where the devices simultaneously send the multiple local model updates $\theta_{m,t}^{k,J}$'s to the BS via the common uplink channel. Recall that $\theta_{m,t}^{k,J} \in \mathbb{R}^{D_m}$. For efficient transmission, we convert $\theta_{m,t}^{k,J}$ into an equivalent complex vector representation $\tilde{\theta}_{m,t}^{k,J}$, whose real and imaginary parts respectively contain the first and second halves of the elements in $\theta_{m,t}^{k,J}$. That is, $\tilde{\theta}_{m,t}^{k,J} = \tilde{\theta}_{m,t}^{k,Jre} + j\tilde{\theta}_{m,t}^{k,Jim} \in \mathbb{C}^{\frac{D_m}{2}}$, where $\tilde{\theta}_{m,t}^{k,Jre}$ contains the first $\frac{D_m}{2}$ elements in $\theta_{m,t}^{k,J}$ and $\tilde{\theta}_{m,t}^{k,Jim}$ contains the rest $\frac{D_m}{2}$ elements.

We assume the uplink channels remain unchanged within one frame. Let $\mathbf{h}_{k,n} \in \mathbb{C}^N$ denote the channel from device kto the BS in frame n, which is assumed known perfectly at the BS. Let $a_{k,n} \in \mathbb{C}$ be the transmit beamforming weight at device k in frame n for sending its local model update. Let $D_{\max} \triangleq \max_{m \in \mathcal{M}} D_m$. Under perfect synchronization, all Kdevices simultaneously send their respective normalized complex model updates, $\frac{\hat{\theta}_{m,t}^{k,J}}{\|\hat{\theta}_{m,t}^{k,J}\|}$'s, to the BS using $\frac{D_{\max}}{2}$ channel uses in round t. For model $\hat{m}(i,t) = m$ with $D_m < D_{\max}$, a random position is set for all $k \in \mathcal{K}_i$. Each device $k \in \mathcal{K}_i$ uses this position for $\hat{\theta}_{m,t}^{k,J}$ within $\frac{D_{\max}}{2}$ channel uses and applies zero padding to the rest of positions. Thus, the equivalent signal vector at this device k is $\hat{\theta}_{m,t}^{k,J} = [\mathbf{0}^H, (\hat{\theta}_{m,t}^{k,J})^H, \mathbf{0}^H]^H$ with length $\frac{D_{\max}}{2}$. The received signal vector $\mathbf{v}_{l,t} \in \mathbb{C}^N$ at the BS in channel use l is given by

$$\mathbf{v}_{l,t} = \sum_{i=1}^{M} \sum_{k \in \mathcal{K}_i} \mathbf{h}_{k,n} a_{k,n} \frac{\bar{\theta}_{ml,t}^{k,J}}{\|\tilde{\theta}_{m,t}^{k,J}\|} + \mathbf{u}_{l,t}^{\text{ul}}, \quad l = 1, \dots, \frac{D_{\max}}{2}$$

where $\mathbf{u}_{l,t}^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \sigma_{u}^{2}\mathbf{I})$ is the receiver noise vector with i.i.d. zero mean and variance σ_{u}^{2} .

The BS applies receive beamforming to $\mathbf{v}_{l,t}$'s for over-theair aggregation of $\tilde{\boldsymbol{\theta}}_{m,t}^{k,J}$, $k \in \mathcal{K}_i$, from each group *i*. Let $\mathbf{w}_{i,n}^{\text{ul}} \in \mathbb{C}^N$ be the unit-norm receive beamforming vector at the BS for group *i* in frame *n*, with $\|\mathbf{w}_{i,n}^{\text{ul}}\|^2 = 1$. For device $k \in \mathcal{K}_i$, and assume $\hat{m}(i,t) = m$, its effective channel after the BS receive beamforming is given by $\alpha_{k,t}^{\text{ul}} \triangleq \frac{(\mathbf{w}_{i,n}^{\text{ul}})^H \mathbf{h}_{k,n} a_{k,n}}{\|\tilde{\boldsymbol{\theta}}_{m,t}^{k,J}\|}$. Thus, the corresponding post-processed received signal vector for $\tilde{\boldsymbol{\theta}}_{m,t}^{k,J}$ over the $\frac{D_m}{2}$ channel uses is given by

$$\mathbf{z}_{m,t} = \sum_{k \in \mathcal{K}_i} \alpha_{k,t}^{\text{ul}} \tilde{\boldsymbol{\theta}}_{m,t}^{k,J} + \sum_{j \neq i} \sum_{q \in \mathcal{K}_j} (\mathbf{w}_{i,n}^{\text{ul}})^H \mathbf{h}_{q,n} a_{q,n} \frac{\bar{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{',g,J}}{\|\tilde{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{q,J}\|} + \mathbf{n}_{m,t}^{\text{ul}}$$

where $\bar{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{'q,J} \in \mathbb{C}^{\frac{D_m}{2}}$ is the portion of other (zero-padded) model $\bar{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{q,J}$ sent by device $q \in \mathcal{K}_j$ that aligns with the location of $\tilde{\boldsymbol{\theta}}_{m,t}^{k,J}$ in $\bar{\boldsymbol{\theta}}_{m,t}^{k,J}$, and $\mathbf{n}_{m,t}^{\text{ul}}$ is the post-processed receiver noise with the *l*-th element being $(\mathbf{w}_{i,n}^{\text{ul}})^H \mathbf{u}_{l,t}^{\text{ul}}$, for $l = 1, \ldots, \frac{D_m}{2}$.

We consider uplink joint transmit-receive beamforming, where $\{a_{k,n}\}_{k \in \mathcal{K}_i}$ and $\mathbf{w}_{i,n}^{ul}$ are designed jointly for each device group *i* in frame *n*. For OAA to be effective, the local models $\tilde{\theta}_{m,t}^{k,J}$'s need to be added coherently. Thus, the transmit and receive beamforming design should ensure that the resulting effective channels $\alpha_{k,t}^{ul}$'s, for $k \in \mathcal{K}_i$ in group *i*, are phase aligned. Thus, we set the transmit beamforming weight $a_{k,n} = \sqrt{p_{k,n}} \frac{\mathbf{h}_{k,n}^{H} \mathbf{w}_{i,n}^{ul}}{|\mathbf{h}_{k,n}^{H} \mathbf{w}_{i,n}^{ul}|}$, for $k \in \mathcal{K}_i$, where $p_{k,n}$ is the transmit power of this device. The effective channels of all devices in group *i* are then phase aligned to 0 after receive beamforming as

$$\alpha_{k,t}^{\mathrm{ul}} = \frac{(\mathbf{w}_{i,n}^{\mathrm{ul}})^H \mathbf{h}_{k,n} a_{k,n}}{\|\tilde{\boldsymbol{\theta}}_{m,t}^{k,J}\|} = \frac{\sqrt{p_{k,n}} |\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\mathrm{ul}}|}{\|\tilde{\boldsymbol{\theta}}_{m,t}^{k,J}\|}, \ k \in \mathcal{K}_i.$$

Each device is subject to the transmit power budget. Let $D_{\max}P_k^{\mathrm{ul}}$ be the total transmit power budget for sending the entire normalized local model in $\frac{D_{\max}}{2}$ channel uses at device k, where $2P_k^{\mathrm{ul}}$ denotes the average transmit power budget per channel use for sending two elements. Then, for transmitting $\frac{\tilde{\theta}_{m,t}^{k,J}}{\|\tilde{\theta}_{m,t}^{k,j}\|}$, we have the transmit power constraint $p_{k,n} \leq D_{\max}P_k^{\mathrm{ul}}$.

After receive beamforming, the BS further scales $\mathbf{z}_{m,t}$ to obtain the complex equivalent global model update $\tilde{\theta}_{m,t+1}$ for the next round t + 1, where $m = \hat{m}(i, t)$:

$$\tilde{\boldsymbol{\theta}}_{m,t+1} = \frac{\mathbf{z}_{m,t}}{\sum_{k \in \mathcal{K}_i} \alpha_{k,t}^{ul}} = \sum_{k \in \mathcal{K}_i} \rho_{k,t} \tilde{\boldsymbol{\theta}}_{m,t}^{k,J} + \tilde{\mathbf{n}}_{m,t}^{ul}$$
$$+ \frac{1}{\sum_{k \in \mathcal{K}_i} \alpha_{k,t}^{ul}} \sum_{j \neq i} \sum_{q \in \mathcal{K}_j} \frac{\mathbf{h}_{q,n}^H \mathbf{w}_{j,n}^{ul} (\mathbf{w}_{i,n}^{ul})^H \mathbf{h}_{q,n}}{|\mathbf{h}_{q,n}^H \mathbf{w}_{j,n}^{ul}|} \cdot \frac{\sqrt{p_{q,n}} \bar{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{\prime q,J}}{\|\tilde{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{q,J}\|}$$
(4)

where $\rho_{k,t} \triangleq \frac{\alpha_{k,t}^{\text{ul}}}{\sum_{q \in \mathcal{K}_i} \alpha_{q,t}^{\text{ul}}}$ is the weight with $\sum_{k \in \mathcal{K}_i} \rho_{k,t} = 1$, and $\tilde{\mathbf{n}}_{m,t}^{\text{ul}} \triangleq \frac{\mathbf{n}_{m,t}^{\text{ul}}}{\sum_{k \in \mathcal{K}_i} \alpha_{k,t}^{\text{ul}}}$ is the post-processed receiver noise at the BS. The weight $\rho_{k,t}$ represents the uplink processing effect including the device transmission and BS receiver processing. Let $\hat{\theta}_{m,t}$ and $\tilde{\mathbf{n}}_{k,t}^{\text{dl}}$ denote the equivalent complex representations of $\theta_{m,t}$ and $\mathbf{n}_{k,t}^{\text{dl}}$ in (3), respectively, for $m = \hat{m}(i,t)$. For local model update in (2), $\Delta \tilde{\theta}_{m,t}^k \triangleq \tilde{\theta}_{m,t}^{k,J} - \tilde{\theta}_{m,t}^{k,0}$ is the equivalent complex representation of the local model difference after the local training at device $k \in \mathcal{K}_i$ in round t. Using (3) and (4), we obtain the global model update $\tilde{\theta}_{m,t+1}$ from $\tilde{\theta}_{m,t}$ as

$$\tilde{\boldsymbol{\theta}}_{m,t+1} = \tilde{\boldsymbol{\theta}}_{m,t} + \sum_{k \in \mathcal{K}_i} \rho_{k,t} \Delta \tilde{\boldsymbol{\theta}}_{m,t}^k + \sum_{k \in \mathcal{K}_i} \rho_{k,t} \tilde{\mathbf{n}}_{k,t}^{\mathrm{dl}} + \tilde{\mathbf{n}}_{m,t}^{\mathrm{ul}} \\ + \frac{1}{\sum_{k \in \mathcal{K}_i} \alpha_{k,t}^{\mathrm{ul}}} \sum_{j \neq i} \sum_{q \in \mathcal{K}_j} \frac{\mathbf{h}_{q,n}^H \mathbf{w}_{j,n}^{\mathrm{ul}} (\mathbf{w}_{i,n}^{\mathrm{ul}})^H \mathbf{h}_{q,n}}{|\mathbf{h}_{q,n}^H \mathbf{w}_{j,n}^{\mathrm{ul}}|} \cdot \frac{\sqrt{p_{q,n}} \bar{\boldsymbol{\theta}}_{m(j,t),t}^{'q,J}}{\|\tilde{\boldsymbol{\theta}}_{m(j,t),t}^{q,J}\|}$$
(5)

Finally, the real-valued global model update $\theta_{m,t+1}$ can be recovered from its complex version as $\theta_{m,t+1} = [\Re \mathfrak{e} \{ \tilde{\theta}_{m,t+1} \}^T, \Im \mathfrak{m} \{ \tilde{\theta}_{m,t+1} \}^T]^T$.

B. Multi-Model FL Convergence Analysis under Uplink OAA

Our objective is to design uplink joint transmit-receive beamforming to minimize the maximum expected optimality gap to $\boldsymbol{\theta}_m^{\star}$ among all M models after S frames, subject to the transmitter power budget. In particular, let $S \triangleq \{0, \ldots, S-1\}$. Let $\mathbf{p}_n \triangleq [\mathbf{p}_{1,n}^T, \ldots, \mathbf{p}_{M,n}^T]^T$, with $\mathbf{p}_{i,n} \in \mathbb{R}^{\frac{K}{M}}$ being the power vector containing $p_{k,n}, k \in \mathcal{K}_i$ of group i in frame n. Also, let $\mathbf{w}_n^{\text{ul}} \triangleq [(\mathbf{w}_{1,n}^{\text{ul}})^H, \ldots, (\mathbf{w}_{M,n}^{\text{ul}})^H]^H \in \mathbb{C}^{MN}$ denote the BS receive beamforming vector for all M groups in frame n. Our optimization problem can be formulated as

$$\begin{aligned} \mathcal{P}_{o}: \min_{\{\mathbf{w}_{n}^{\text{ul}}, \mathbf{p}_{n}\}_{n=0}^{S-1}} \max_{m \in \mathcal{M}} \mathbb{E}[\|\boldsymbol{\theta}_{m,SM} - \boldsymbol{\theta}_{m}^{\star}\|^{2}] \\ \text{s.t.} \ p_{k,n} \leq D_{\max}P_{k}^{\text{ul}}, \ k \in \mathcal{K}_{\text{tot}}, \ n \in \mathcal{S}, \\ \|\mathbf{w}_{i,n}^{\text{ul}}\|^{2} = 1, \quad i \in \mathcal{M}, \ n \in \mathcal{S} \end{aligned}$$

where $\mathbb{E}[\cdot]$ is taken w.r.t. receiver noise and mini-batch local data samples at each device. Problem \mathcal{P}_o is a stochastic optimization problem with a min-max objective. To tackle this challenging problem, we develop a more tractable upper bound on $\mathbb{E}[\|\boldsymbol{\theta}_{m,SM} - \boldsymbol{\theta}_m^{\star}\|^2]$ by analyzing the convergence rate of the global model update.

We make the following assumptions on the local loss functions, the local model updates, and the divergence of the global and local loss gradients. They are commonly used in the convergence analysis of FL training [2], [6], [9].

Assumption 1. The local loss function $F_m^k(\cdot)$ is *L*-smooth and λ -strongly convex, $\forall m \in \mathcal{M}, \forall k \in \mathcal{K}_{tot}$.

Assumption 2. Bounded local model parameters: $\|\hat{\boldsymbol{\theta}}_{m,t}^{k,J}\|^2 \leq r$, for some r > 0, $\forall m \in \mathcal{M}$, $\forall k \in \mathcal{K}_{tot}$, $\forall t$.

Assumption 3. Bounded gradient divergence of loss functions: $\mathbb{E}[\|\nabla F_m(\boldsymbol{\theta}_{m,t}) - \sum_{k=1}^{K} c_k \nabla F_m^k(\boldsymbol{\theta}_{m,t}^{k,\tau})\|^2] \leq \phi$ and $\mathbb{E}[\|\nabla F_m^k(\boldsymbol{\theta}_{m,t}^{k,\tau}) - \nabla F_m^k(\boldsymbol{\theta}_{m,t}^{k,\tau}, \mathcal{B}_{m,t}^{k,\tau})\|^2] \leq \delta$ for some $\phi \geq 0$, $\delta \geq 0, 0 \leq c_k \leq 1, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}_{\text{tot}}, \forall \tau, \forall t.$

Based on (5), we first obtain the per-model global update equation over frames. Let device group \hat{i} be the group that trains model m in communication round t in frame n. The

device-model assignment between \hat{i} and m is given in (1). Summing both sides of (5) over M rounds in frame n, and subtracting the optimal $\tilde{\theta}_m^{\star}$ (complex version of θ_m^{\star}) from both sides, we obtain

$$\tilde{\boldsymbol{\theta}}_{m,(n+1)M} - \tilde{\boldsymbol{\theta}}_{m}^{\star} = \tilde{\boldsymbol{\theta}}_{m,nM} - \tilde{\boldsymbol{\theta}}_{m}^{\star} + \sum_{t=nM}^{(n+1)M-1} \sum_{k \in \mathcal{K}_{\hat{i}}} \rho_{k,t} \Delta \tilde{\boldsymbol{\theta}}_{m,t}^{k} + \tilde{\mathbf{e}}_{m,n}$$

where $\tilde{\mathbf{e}}_{m,n}$ is the accumulated error term in (5) over M rounds in frame n, given by

$$\begin{split} \tilde{\mathbf{e}}_{m,n} &\triangleq \sum_{t=nM}^{(n+1)M-1} \sum_{k \in \mathcal{K}_{\hat{i}}} \rho_{k,t} \tilde{\mathbf{n}}_{k,t}^{\mathrm{dl}} + \sum_{t=nM}^{(n+1)M-1} \tilde{\mathbf{n}}_{m,t}^{\mathrm{ul}} \\ &+ \sum_{t=nM}^{(n+1)M-1} \sum_{j \neq \hat{i}} \sum_{q \in \mathcal{K}_j} \frac{\mathbf{h}_{q,n}^{H} \mathbf{w}_{j,n}^{\mathrm{ul}} (\mathbf{w}_{\hat{i},n}^{\mathrm{ul}})^{H} \mathbf{h}_{q,n}}{|\mathbf{h}_{q,n}^{H} \mathbf{w}_{j,n}^{\mathrm{ul}}| \sum_{k \in \mathcal{K}_{\hat{i}}} \alpha_{k,t}^{\mathrm{ul}}} \cdot \frac{\sqrt{p_{q,n}} \bar{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{'q,J}}{\|\tilde{\boldsymbol{\theta}}_{\hat{m}(j,t),t}^{q,J}\|}. \end{split}$$

By Assumption 2, we can further bound $\mathbb{E}[\|\tilde{\mathbf{e}}_{m,n}\|^2]$ as

$$\mathbb{E}\left[\|\tilde{\mathbf{e}}_{m,n}\|^{2}\right] \leq rMK \sum_{i=1}^{M} \frac{\sum_{j \neq i} \sum_{q \in \mathcal{K}_{j}} p_{q,n} |\mathbf{h}_{q,n}^{H} \mathbf{w}_{i,n}^{ul}|^{2} + \tilde{\sigma}_{u}^{2}}{(\sum_{k \in \mathcal{K}_{i}} \sqrt{p_{k,n}} |\mathbf{h}_{k,n}^{H} \mathbf{w}_{i,n}^{ul}|)^{2}} + 2K\tilde{\sigma}_{d}^{2}$$
(6)

where $\tilde{\sigma}_{\rm d}^2 \triangleq \sigma_{\rm d}^2 D_{\rm max}/2$ and $\tilde{\sigma}_{\rm u}^2 \triangleq \sigma_{\rm u}^2 D_{\rm max}/2$.

Using the above, we obtain an upper bound on $\mathbb{E}[\|\boldsymbol{\theta}_{m,SM} - \boldsymbol{\theta}_m^*\|^2]$ below. The proof is omitted due to space limitation.

Proposition 1. Under Assumptions 1–3 and for $\eta_n < \frac{1}{\lambda}$, $\forall n$, the expected optimality gap after *S* frames is bounded by

$$\mathbb{E}[\|\boldsymbol{\theta}_{m,SM} - \boldsymbol{\theta}_{m}^{\star}\|^{2}] \leq \Gamma_{m} \prod_{n=0}^{S-1} G_{n} + \Lambda + \sum_{n=0}^{S-2} H(\mathbf{w}_{n}^{\text{ul}}, \mathbf{p}_{n}) \prod_{s=n+1}^{S-1} G_{s} + H(\mathbf{w}_{S-1}^{\text{ul}}, \mathbf{p}_{S-1}), \quad m \in \mathcal{M}$$
(7)

where $\Gamma_m \triangleq \mathbb{E}[\|\boldsymbol{\theta}_{m,0} - \boldsymbol{\theta}_m^{\star}\|^2], G_n \triangleq 4(1 - \eta_n \lambda)^{2JM}, \Lambda \triangleq \sum_{n=0}^{S-2} C_n(\prod_{s=n+1}^{S-1} G_s) + C_{S-1}$ with $C_n \triangleq 4\eta_n^2 J^2(M^2 \phi + K^2 \delta) + 8K \tilde{\sigma}_d^2$, and $H(\mathbf{w}_n^{ul}, \mathbf{p}_n) \triangleq 4rMK \sum_{i=1}^M \frac{\sum_{j \neq i} \sum_{q \in \mathcal{K}_j} p_{q,n} |\mathbf{h}_{q,n}^H \mathbf{w}_{i,n}^{ul}|^2 + \tilde{\sigma}_u^2}{(\sum_{k \in \mathcal{K}_i} \sqrt{p_{k,n}} |\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{ul}|)^2}.$

C. Uplink Joint Transmit-Receive Beamforming Design

We now replace the objective function in \mathcal{P}_o with the more tractable upper bound in (7). Omitting the first two constant terms in (7) that do not depend on the beamforming design, we arrive at the following equivalent optimization problem:

$$\mathcal{P}_{1}: \min_{\left\{\mathbf{w}_{n}^{\mathrm{ul}}, \mathbf{p}_{n}\right\}_{n=0}^{S-1}} \sum_{n=0}^{S-2} H(\mathbf{w}_{n}^{\mathrm{ul}}, \mathbf{p}_{n}) \prod_{s=n+1}^{S-1} G_{s} + H(\mathbf{w}_{S-1}^{\mathrm{ul}}, \mathbf{p}_{S-1})$$

s.t. $p_{k,n} \leq D_{\max} P_{k}^{\mathrm{ul}}, \quad k \in \mathcal{K}_{\mathrm{tot}}, \ n \in \mathcal{S},$
 $\|\mathbf{w}_{i,n}^{\mathrm{ul}}\|^{2} = 1, \quad i \in \mathcal{M}, \ n \in \mathcal{S}.$

By Proposition 1, for $\eta_n < \frac{1}{\lambda}$, we have $G_n > 0$, $\forall n$. Thus, \mathcal{P}_1 can be decomposed into S subproblems, one for each frame n, given by

$$\mathcal{P}_{2,n}: \min_{\mathbf{w}_{n}^{\mathrm{ul}},\mathbf{p}_{n}} \sum_{i=1}^{M} \frac{\sum_{j\neq i} \sum_{q\in\mathcal{K}_{j}} p_{q,n} |\mathbf{h}_{q,n}^{H} \mathbf{w}_{i,n}^{\mathrm{ul}}|^{2} + \tilde{\sigma}_{u}^{2}}{(\sum_{k\in\mathcal{K}_{i}} \sqrt{p_{k,n}} |\mathbf{h}_{k,n}^{H} \mathbf{w}_{i,n}^{\mathrm{ul}}|)^{2}}$$
s.t. $p_{k,n} \leq D_{\max} P_{k}^{\mathrm{ul}}, \quad k \in \mathcal{K}_{\mathrm{tot}},$

$$\|\mathbf{w}_{i,n}^{\mathrm{ul}}\|^2 = 1, \quad i \in \mathcal{M}.$$

Problem $\mathcal{P}_{2,n}$ is a multi-user joint uplink transmit power allocation and receive beamforming problem with a complicated objective function of $\{\mathbf{w}_n^{ul}, \mathbf{p}_n\}$. To make the problem amenable for a solution, we consider an upper bound of the objective function. Let $\mathbf{f}_{k,n} \triangleq \mathbf{h}_{k,n}/\tilde{\sigma}_u$. Since $(\sum_{k \in \mathcal{K}_i} \sqrt{p_{k,n}} |\mathbf{f}_{k,n}^H \mathbf{w}_{i,n}^{ul}|)^2 \ge \sum_{k \in \mathcal{K}_i} p_{k,n} |\mathbf{f}_{k,n}^H \mathbf{w}_{i,n}^{ul}|^2$, we replace the objective function in $\mathcal{P}_{2,n}$ by an upper bound and arrive at the following problem:

$$\mathcal{P}_{3,n}: \min_{\mathbf{w}_{n}^{ul},\mathbf{p}_{n}} \sum_{i=1}^{M} \frac{\sum_{j\neq i} \sum_{q\in\mathcal{K}_{j}} p_{q,n} |\mathbf{f}_{q,n}^{H} \mathbf{w}_{i,n}^{ul}|^{2} + 1}{\sum_{k\in\mathcal{K}_{i}} p_{k,n} |\mathbf{f}_{k,n}^{H} \mathbf{w}_{i,n}^{ul}|^{2}}$$

s.t. $p_{k,n} \leq D_{\max} P_{k}^{ul}, \quad k \in \mathcal{K}_{\text{tot}},$
 $\|\mathbf{w}_{i,n}^{ul}\|^{2} = 1, \quad i \in \mathcal{M}.$

We note that in the objective function, the *i*th term in the summation is the inverse of SINR for the aggregated local models received from group *i*, and the objective function is the sum of inverse SINRs of all M groups. For this jointly non-convex problem $\mathcal{P}_{3,n}$, we apply BCD to solve it, *i.e.*, alternatingly updates the BS receive beamforming \mathbf{w}_n^{ul} and the device powers in \mathbf{p}_n . The two subproblems are given below:

1) Updating \mathbf{w}_n^{ul} : Given \mathbf{p}_n , $\mathcal{P}_{3,n}$ can be further decomposed into M subproblems, one for each beamformer $\mathbf{w}_{i,n}^{\text{ul}}$ as

$$\mathcal{P}_{3,n,i}^{\text{wsub}}: \min_{\mathbf{w}_{i,n}^{\text{ul}}} \frac{(\mathbf{w}_{i,n}^{\text{ul}})^{H} \left(\sum_{j \neq i} \sum_{q \in \mathcal{K}_{j}} p_{q,n} \mathbf{f}_{q,n} \mathbf{f}_{q,n}^{H} + \mathbf{I}\right) \mathbf{w}_{i,n}^{\text{ul}}}{(\mathbf{w}_{i,n}^{\text{ul}})^{H} \left(\sum_{k \in \mathcal{K}_{i}} p_{k,n} \mathbf{f}_{k,n} \mathbf{f}_{k,n}^{H}\right) \mathbf{w}_{i,n}^{\text{ul}}}$$

s.t. $\|\mathbf{w}_{i,n}^{\text{ul}}\|^{2} = 1,$

which is a generalized eigenvalue problem. The optimal solution $\mathbf{w}_{i,n}^{\text{ul}}$ can be obtained in closed-form, which is the generalized eigenvector corresponding to the smallest generalized eigenvalue. We omit the detail due to space limitation.

2) Updating \mathbf{p}_n : Let $\mathbf{g}_{ij,n}$ be a $\frac{K}{M} \times 1$ vector containing $\{g_{iq,n} \triangleq |\mathbf{f}_{q,n}^H \mathbf{w}_{i,n}^{ul}|^2, q \in \mathcal{K}_j\}$ of group $j \in \mathcal{M}$. Given $\{\mathbf{w}_{i,n}^{ul}\}$, we can rewrite $\mathcal{P}_{3,n}$ as

$$\mathcal{P}_{3,n}^{\text{psub}}: \min_{\{\mathbf{p}_{i,n}\}_{i=1}^{M}} \sum_{i=1}^{M} \frac{\sum_{j \neq i} \mathbf{g}_{ij,n}^{T} \mathbf{p}_{j,n} + 1}{\mathbf{g}_{ii,n}^{T} \mathbf{p}_{i,n}}$$

s.t. $p_{k,n} \leq D_{\max} P_{k}^{\text{ul}}, \quad k \in \mathcal{K}_{\text{tot}}.$

We propose to update $\mathbf{p}_{1,n}, \ldots, \mathbf{p}_{M,n}$ sequentially using BCD. Given $\mathbf{p}_{j,n}, \forall j \in \mathcal{M}, j \neq i, \mathcal{P}_{3,n}^{\text{psub}}$ is convex w.r.t. $\mathbf{p}_{i,n}$ for group *i*, for which the optimal $\mathbf{p}_{i,n}$ can be obtained in closed-form via the KKT conditions. Specifically, let $\bar{\mathbf{p}}_{i}^{\text{ul}}$ be the vector containing $\{D_{\max}P_{k}^{\text{ul}}, k \in \mathcal{K}_{i}\}$ of group *i*. Let $a_{i,n}^{\min} \triangleq \min_{k \in \mathcal{K}_{i}} \left(\frac{\sum_{j \neq i} \mathbf{g}_{j,n}^{T} \mathbf{p}_{j,n} + 1}{\sum_{j \neq i} \frac{g_{j,n}}{\mathbf{g}_{j,n}^{T} \mathbf{p}_{j,n}}} g_{ik,n}\right)^{1/2}$ and let $k' \in \mathcal{K}_{i}$ be the corresponding index that achieves a^{\min} . Thus, the

be the corresponding index that achieves $a_{i,n}^{\min}$. Thus, the optimal $\mathbf{p}_{i,n}$ is given by

$$p_{k,n} = \begin{cases} D_{\max} P_k^{\text{ul}} & \text{for } k \in \mathcal{K}_i, k \neq k' \\ D_{\max} P_{k'}^{\text{ul}} - \frac{\left[\mathbf{g}_{ii,n}^T \bar{\mathbf{p}}_i^{\text{ul}} - a_{i,n}^{\min}\right]^+}{g_{ik',n}} & \text{for } k = k' \end{cases}$$

where $[a]^+ = \max\{a, 0\}$. Thus, $\mathbf{p}_{i,n}$ is updated sequentially using the above solution.

IV. SIMULATION RESULTS

We consider image classification under the current cellular system setting with 10 MHz bandwidth and 2 GHz carrier frequency. The maximum transmit powers at the BS and devices are 47 dBm and 23 dBm, respectively. We assume the devices use 1 MHz bandwidth for uplink transmission. Each channel is generated as $\mathbf{h}_{k,t} = \sqrt{G_k} \bar{\mathbf{h}}_{k,t}$, where $\bar{\mathbf{h}}_{k,t} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$, and the path gain $G_k[dB] = -136.3 - 35 \log_{10} d_k - \psi_k$, with the BS-device distance $d_k \in (0.02, 0.5)$ in kilometers and the shadowing random variable ψ_k having the standard deviation 8 dB. Noise power spectral density is -174 dBm/Hz, and we assume noise figure $N_F = 8$ dB and 2 dB at the device and BS receivers, respectively. We set N = 64 and K = 12.

We use the MNIST dataset for the multi-model training and testing. It consists of 60,000 training samples and 10,000 test samples. We train three types of convolutional neural networks: i) Model A: an $8 \times 3 \times 3$ ReLU convolutional layer, a 2×2 max pooling layer, and a softmax output layer with 13,610 parameters. ii) Model B: a $6 \times 4 \times 4$ ReLU convolutional layer, a 2×2 max pooling layer, a ReLU fullyconnected layer with 22 units, and a softmax output layer with 19,362 parameters. ii) Model C: an $8 \times 3 \times 3$ ReLU convolutional layer, a 2×2 max pooling layer, a ReLU fullyconnected layer with 20 units, and a softmax output layer with 27, 350 parameters. We use the 10,000 test samples to measure the test accuracy of each global model update $\theta_{m,t}$ at round t. The training samples are randomly and evenly distributed across devices, with the local dataset size $S_k = 60,000/K$ samples at device k. For the local training using SGD, we set J = 100, the mini-batch size $|\mathcal{B}_{m,t}^{k\tau}| = 600/K, \forall k, \tau, m, t$, and the learning rate $\eta_n = 0.1$, $\forall n$. All results are obtained by taking the current best test accuracy and averaging over 20 channel realizations.

We denote our proposed method as MultiModel. We also consider two schemes for comparison: i) Ideal: Multi-model FL via (5) with error-free downlink and uplink. It serves as a performance upper bound for all schemes. ii) SeqnModel: Sequentially train each model using the single-model FL with all K devices by the uplink beamforming scheme that maximizes the aggregated SNR provided in [8]. Fig. 2-Top Left shows the test accuracy vs. M models, after T = 30rounds, where all models are from Model A. We see that MultiModel substantially outperforms the sequential model training for all M values. We also consider mixed model types. We set M = 3, one from each of Models A, B, and C. Fig. 2-Top Right, Bottom Left, and Bottom Right show the test accuracy over round t for Models A, B, and C, respectively. We see that MultiModel outperforms the sequential training using the single-model-based scheme for all models.

V. CONCLUSION

This paper considers uplink transmission design for multimodel wireless FL. We design uplink beamforming for send-



Fig. 2. Top Left: Test accuracy vs. M (Model A). Rest of figures: Test accuracy vs. communication round t: Top Right – Model A; Bottom Left – Model B; Bottom Right – Model C (90% confidence intervals are shown).

ing multiple models simultaneously to the BS via OAA to maximize the FL training performance. We utilize an upper bound on the optimality gap of the global multi-model update to formulate the joint uplink transmit-receive beamforming problem and apply BCD to solve it with closed-form iteration updates. Simulation results demonstrate substantial performance advantage of the proposed multi-model scheme over the conventional single-model sequential training.

REFERENCES

- B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.*, Apr. 2017, pp. 1273–1282.
- [2] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1422–1437, Mar. 2022.
- [3] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via overthe-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.
- [4] N. Zhang and M. Tao, "Gradient statistics aware power control for overthe-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, Aug. 2021.
- [5] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 227–242, Jan. 2022.
- [6] W. Guo, R. Li, C. Huang, X. Qin, K. Shen, and W. Zhang, "Joint device selection and power control for wireless federated learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2395–2410, Aug. 2022.
- [7] Z. Wang, Y. Zhou, Y. Shi, and W. Zhuang, "Interference management for over-the-air federated learning in multi-cell wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2361–2377, Aug. 2022.
- [8] C. Zhang, M. Dong, B. Liang, A. Afana, and Y. Ahmed, "Joint downlink-uplink beamforming for wireless multi-antenna federated learning," in *Proc. WiOpt*, Aug. 2023.
- [9] N. Bhuyan, S. Moharir, and G. Joshi, "Multi-model federated learning with provable guarantees," in *Proc. Int. Conf. Perform. Eval. Method*ologies Tools, Nov. 2022, pp. 207–222.
- [10] C. Zhang, M. Dong, B. Liang, A. Afana, and Y. Ahmed, "Multimodel wireless federated learning with downlink beamforming," in *Proc. ICASSP*, Apr. 2024.
- [11] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.
- [12] M. Dong and Q. Wang, "Multi-group multicast beamforming: Optimal structure and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 3738–3753, May 2020.