

# MULTI-MODEL WIRELESS FEDERATED LEARNING WITH DOWNLINK BEAMFORMING

*Chong Zhang*<sup>\*</sup>    *Min Dong*<sup>†</sup>    *Ben Liang*<sup>\*</sup>    *Ali Afana*<sup>‡</sup>    *Yahia Ahmed*<sup>‡</sup>

<sup>\*</sup> Department of Electrical and Computer Engineering, University of Toronto, Canada, <sup>‡</sup>Ericsson Canada, Canada

<sup>†</sup> Department of Electrical, Computer and Software Engineering, Ontario Tech University, Canada

## ABSTRACT

This paper studies the design of wireless federated learning (FL) for simultaneously training multiple machine learning models. We consider round robin device-model assignment and downlink beamforming for concurrent multiple model updates. After formulating the joint downlink-uplink transmission process, we derive the per-model global update expression over communication rounds, capturing the effect of beamforming and noisy reception. To maximize the multi-model training convergence rate, we derive an upper bound on the optimality gap of the global model update and use it to formulate a multi-group multicast beamforming problem. We show that this problem can be converted to minimizing the sum of inverse signal-to-interference-plus-noise ratios (SINRs), which can be solved efficiently by projected gradient descent. Simulation shows that our proposed multi-model FL solution outperforms other alternatives, including conventional single-model sequential training and multi-model zero-forcing beamforming.

## 1. INTRODUCTION

Federated learning (FL) [1] is a widely adopted method for multiple devices to collaboratively train a common machine learning (ML) model. In wireless FL, a parameter server, usually the base station (BS), uses wireless communication to exchange model parameters with participating devices [2]. With frequent exchange of a large number of parameters, FL performance degrades in the wireless environment due to signal distortion and limited wireless resource. This necessitates efficient communication design to effectively support FL.

Most existing works on wireless FL have focused on training only a single model [3–13]. Assuming an error-free downlink, [3–8] focused on efficient transmission for the uplink acquisition of local parameters from devices to the BS, including both digital [3] and analog [4–8] schemes. Noisy downlink transmission for FL was studied in [9] with error-free uplink. It was shown in these works that analog transmission can be more efficient than digital for both the downlink and the uplink. Joint noisy downlink-uplink transmission for FL was studied in [10–13], with single-antenna BSs in [10–12] and a multi-antenna BS in [13]. In practice, the parameter server may have multiple models to be trained. Directly using the existing single-model FL schemes may lead to substantial latency, degrading the overall performance of wireless FL.

Simultaneously training multiple models in FL was proposed recently in [14], assuming error-free downlink and uplink transmissions. It was shown that multi-model FL can substantially improve the training convergence rate over the single-model FL approach, which reduces the burden of required computation and communication. However, the idealized system in [14] did not account for the impact of wireless transmission over noisy channels. In wireless multi-model FL, besides noisy downlink and uplink transmis-

sion during model updates, there is also inter-model interference in transmission, which adds substantial challenges toward improving the communication efficiency.

In this paper, we study multi-model wireless FL design for noisy downlink and uplink wireless channels with a multi-antenna BS. We consider analog transmission, downlink beamforming, and round robin model scheduling. Aiming to maximize the training convergence rate, we derive an upper bound on the optimality gap of the FL global model update, which captures the impact of noisy transmission and inter-model interference. We then show that the minimization of this upper bound leads to a downlink multi-group multicast beamforming design to minimize the sum of inverse received SINRs subject to a downlink transmit power budget at the BS, which can be solved using projected gradient descent (PGD). Our simulation results under typical wireless network settings show that the proposed multi-model FL design substantially outperforms the conventional single-model design approach that sequentially trains one model at a time, as well as multi-model training using the popular zero-forcing beamforming scheme.

## 2. SYSTEM MODEL

### 2.1. Multi-Model FL System

We consider FL in a wireless network consisting of a server and  $K$  worker devices that collaboratively train  $M$  global models at the server. Let  $\mathcal{K}_{\text{tot}} = \{1, \dots, K\}$  denote the total set of devices and  $\mathcal{M} = \{1, \dots, M\}$  the set of models. Let  $\theta_m \in \mathbb{R}^{D_m}$  be the parameter vector of model  $m$  with  $D_m$  parameters. Assume each device  $k \in \mathcal{K}_{\text{tot}}$  holds local training datasets for all  $M$  models, with each  $\theta_m$  being locally trained using the dataset for model  $m$  of size  $S_m^k$ , denoted by  $\mathcal{S}_m^k = \{(\mathbf{s}_{m,i}^k, v_{m,i}^k) : 1 \leq i \leq S_m^k\}$ , where  $\mathbf{s}_{m,i}^k \in \mathbb{R}^b$  is the  $i$ -th data feature vector and  $v_{m,i}^k$  is the label for this data sample. The local training loss function representing the training error at device  $k$  for model  $m \in \mathcal{M}$  is defined as  $F_m^k(\theta_m) = \frac{1}{S_m^k} \sum_{i=1}^{S_m^k} L_m(\theta_m; \mathbf{s}_{m,i}^k, v_{m,i}^k)$ , where  $L_m(\cdot)$  is the sample-wise training loss for model  $m$ . The global training loss function for model  $m$  is given by the weighted sum of the local loss functions for model  $m$  over all  $K$  devices:  $F_m(\theta_m) = \frac{1}{\sum_{k=1}^K S_m^k} \sum_{k=1}^K S_m^k F_m^k(\theta_m)$ . The learning objective is to find the optimal  $\theta_m^*$  that minimizes  $F_m(\theta_m)$  for each model  $m \in \mathcal{M}$ .

The  $K$  devices use their respective local training datasets to simultaneously train the  $M$  models and communicate with the server via noisy downlink and uplink wireless channels to exchange the model training information iteratively. At the beginning of each downlink-uplink communication round  $t = 0, 1, \dots$ , the devices are divided into device groups, and the server assigns the training task of each model to a device group. We consider the round robin scheduling approach for efficient device-model assignment [14]. Specifically, we define every  $M$  communication rounds as a *frame*. At the beginning of each frame, the  $K$  devices are partitioned into

$M$  equal-sized groups randomly. Let  $\mathcal{K}_i$  denote the set of devices in device group  $i = 1, \dots, M$ . These device groups remain unchanged within a frame. Each device group  $i$  is assigned to train model  $\hat{m}(i, t)$  at round  $t$  given by

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

Fig. 1 shows an example of the device-model assignment in a frame via the round robin scheduling with  $M = 3$  models.

The iterative multi-model FL training procedure in each downlink-uplink communication round  $t$  is then given as follows:

- **Downlink broadcast:** The server broadcasts each of the current  $M$  global model parameter vectors  $\theta_{m,t}$ 's to its assigned device group via the downlink channel;
- **Local model update:** Device  $k \in \mathcal{K}_i$  in device group  $i$  is scheduled to locally train model  $\hat{m}(i, t)$  using its corresponding local dataset  $\mathcal{S}_{\hat{m}(i,t)}^k$ . In particular, the device divides  $\mathcal{S}_{\hat{m}(i,t)}^k$  into mini-batches for its local model update based on  $\theta_{\hat{m}(i,t),t}$ , where it performs  $J$  iterative local updates and generates the updated local model  $\theta_{\hat{m}(i,t),t}^{k,J}$ ;
- **Uplink aggregation:** The devices send their updated local models  $\theta_{m,t}^{k,J}$ 's to the server via the uplink channels. The server aggregates  $\{\theta_{m,t}^{k,J}\}_{k \in \mathcal{K}_i}$  received from each device group  $i$  to generate updated global model  $\theta_{m,t+1}$ ,  $m \in \mathcal{M}$ , for the next communication round  $t + 1$ .

## 2.2. Wireless Communication Model

We consider a practical wireless communication system where the server is hosted by a BS equipped with  $N$  antennas, and each device has a single antenna. To efficiently send  $M$  global model updates, the BS uses the multi-group multicast beamforming technique [15, 16] to send the  $M$  global model updates  $\theta_{m,t}$ 's to  $M$  device groups simultaneously over a common downlink channel. Also, we consider analog transmission, where the BS sends the values of  $\theta_{m,t}$ 's directly under its transmit power budget. For the uplink aggregation, we consider the orthogonal multiple access technique to efficiently use the system bandwidth for local model aggregation at the BS. For each device group, we consider over-the-air computation via analog aggregation over the multiple access channel. Specifically, the devices in a device group  $i$  send their local models  $\{\theta_{m,t}^{k,J}\}_{k \in \mathcal{K}_i}$  simultaneously over the same uplink channel, while the channels among device groups are orthogonal to each other.

The received model updates over downlink are the distorted noisy versions of  $\theta_{m,t}$ 's, due to the inter-group interference in transmitting  $\theta_{m,t}$ 's and the noisy communication channel. The uplink received model updates are also the distorted noisy versions of  $\theta_{m,t}^{k,J}$ 's due to the noisy channel. These errors in the model updates further propagate over subsequent communication rounds for multi-model training, degrading the learning performance. In this paper, we focus on the communication aspect of FL multi-model training and develop the downlink beamforming design to maximize the learning performance of FL over wireless transmissions.

## 3. MULTI-MODEL DOWNLINK-ULINK TRANSMISSIONS

In this section, we formulate the wireless transmission process in downlink and uplink for the multi-model update using the three steps in a communication round that are mentioned in Section 2.1.

### 3.1. Downlink Broadcast

At the start of round  $t$ , the BS has the current global models, with model  $m$  denoted by  $\theta_{m,t} = [\theta_{m1,t}, \dots, \theta_{mD_m,t}]^T$ . For efficient transmission, we represent  $\theta_{m,t}$  using a complex signal

	$\mathcal{K}_1$	$\mathcal{K}_2$	$\mathcal{K}_3$
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:** An example of round robin scheduling of device-model assignment in a frame for training 3 models.

vector, whose real and imaginary parts respectively contain the first and second half of the elements in  $\theta_{m,t}$ . That is,  $\theta_{m,t} = [(\tilde{\theta}_{m,t}^{\text{re}})^T, (\tilde{\theta}_{m,t}^{\text{im}})^T]^T$ , where  $\tilde{\theta}_{m,t}^{\text{re}} \triangleq [\theta_{m1,t}, \dots, \theta_{m(D_m/2),t}]^T$  and  $\tilde{\theta}_{m,t}^{\text{im}} \triangleq [\theta_{m(D_m/2+1),t}, \dots, \theta_{mD_m,t}]^T$ . Let  $\tilde{\theta}_{m,t}$  denote the equivalent complex vector representation of  $\theta_{m,t}$ . It is given by  $\tilde{\theta}_{m,t} = \tilde{\theta}_{m,t}^{\text{re}} + j\tilde{\theta}_{m,t}^{\text{im}} \in \mathbb{C}^{\frac{D_m}{2}}$ .

With the frame structure, round  $t$  is in frame  $n = \lfloor t/M \rfloor$ . We assume the downlink channel remains unchanged within one frame. Thus, let  $\mathbf{h}_{k,n} \in \mathbb{C}^N$  be the downlink channel vector from the BS to device  $k \in \mathcal{K}_i$ ,  $i = 1, \dots, M$ , in frame  $n$ , which is assumed known perfectly at the BS. Let  $\mathbf{w}_{i,n}^{\text{dl}} \in \mathbb{C}^N$  be the downlink multicast beamforming vector for device group  $i$  in frame  $n$ . The BS uses  $\mathbf{w}_{i,n}^{\text{dl}}$  to send the normalized complex global model  $\frac{\tilde{\theta}_{\hat{m}(i,t),t}}{\|\tilde{\theta}_{\hat{m}(i,t),t}\|}$  to device group  $i$ . Let  $D_{\max} \triangleq \max_{m \in \mathcal{M}} D_m$ . The  $M$  model updates are simultaneously sent using  $D_{\max}$  channel uses. For model  $m$  with  $D_m < D_{\max}$ , the BS randomly sets the position for  $\tilde{\theta}_{m,t}$  within  $D_{\max}$  channel uses and applies zero padding to the rest of positions. Thus, the transmitted signal vector for model  $m$  is  $\tilde{\theta}_{m,t} = [\mathbf{0}^H, \tilde{\theta}_{m,t}^H, \mathbf{0}^H]^H$ . Assume  $\hat{m}(i, t) = m$ . The received signal  $\mathbf{u}_{k,t}$  at device  $k \in \mathcal{K}_i$  corresponding to  $\tilde{\theta}_{m,t}$  is given by

$$\mathbf{u}_{k,t} = (\mathbf{w}_{i,n}^{\text{dl}})^H \mathbf{h}_{k,n} \frac{\tilde{\theta}_{m,t}}{\|\tilde{\theta}_{m,t}\|} + \sum_{j \neq i} (\mathbf{w}_{j,n}^{\text{dl}})^H \mathbf{h}_{k,n} \frac{\tilde{\theta}_{\hat{m}(j,t),t}}{\|\tilde{\theta}_{\hat{m}(j,t),t}\|} + \mathbf{n}_{k,t}^{\text{dl}}$$

where  $n = \lfloor t/M \rfloor$ ,  $\tilde{\theta}_{\hat{m}(j,t),t} \in \mathbb{C}^{\frac{D_m}{2}}$  is the portion of  $\tilde{\theta}_{\hat{m}(j,t),t}$  that aligns with the location of  $\tilde{\theta}_{m,t}$  in  $\tilde{\theta}_{m,t}$  due to zero-padding, and  $\mathbf{n}_{k,t}^{\text{dl}} \sim \mathcal{CN}(\mathbf{0}, \sigma_d^2 \mathbf{I})$  is the receiver additive white Gaussian noise (AWGN) vector. The beamforming vectors  $\{\mathbf{w}_{i,n}^{\text{dl}}\}_{i=1}^M$  are subject to the BS transmit power budget. Let  $D_{\max}P$  be the BS total transmit power budget for sending the entire normalized global models in  $D_{\max}$  channel uses, where  $P$  denotes the average power per channel use. Then, we have the transmit power constraint  $\sum_{i=1}^M \|\mathbf{w}_{i,n}^{\text{dl}}\|^2 \leq D_{\max}P$ . The BS also sends the scaling factor  $\frac{\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}} \|\tilde{\theta}_{m,t}\|}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2}$  to the device via the downlink signaling channel to facilitate this receiver processing. After post-processing  $\mathbf{u}_{k,t}$  using the received scaling factor at device  $k \in \mathcal{K}_i$ , we have

$$\begin{aligned} \hat{\theta}_{m,t}^k &= \frac{\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}} \|\tilde{\theta}_{m,t}\|}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2} \mathbf{u}_{k,t} \\ &= \tilde{\theta}_{m,t} + \sum_{j \neq i} \frac{\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}} (\mathbf{w}_{j,n}^{\text{dl}})^H \mathbf{h}_{k,n}}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2} \cdot \frac{\|\tilde{\theta}_{m,t}\| \tilde{\theta}_{\hat{m}(j,t),t}}{\|\tilde{\theta}_{\hat{m}(j,t),t}\|} + \tilde{\mathbf{n}}_{k,t}^{\text{dl}} \end{aligned} \quad (2)$$

where  $\tilde{\mathbf{n}}_{k,t}^{\text{dl}} \triangleq \frac{\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}} \|\tilde{\theta}_{m,t}\|}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2} \mathbf{n}_{k,t}^{\text{dl}}$  is the post-processed noise vector at device  $k \in \mathcal{K}_i$ . By the equivalence of real and complex signal representations  $\theta_{m,t}$  and  $\tilde{\theta}_{m,t}$ , device  $k \in \mathcal{K}_i$  obtains the estimate of the global model  $\theta_{m,t}$  as

$$\hat{\theta}_{m,t}^k = [\Re\{\hat{\theta}_{m,t}^k\}^T, \Im\{\hat{\theta}_{m,t}^k\}^T]^T. \quad (3)$$

### 3.2. Local Model Update

Device  $k \in \mathcal{K}_i$  is scheduled to perform local model updates on  $\hat{\theta}_{m,t}^k$  in (3) using its local dataset  $\mathcal{S}_m^k$ . We assume each device adopts the standard mini-batch stochastic gradient descent (SGD)

algorithm [17] to perform the local model training. In particular, assume that each device applies  $J$  mini-batch SGD iterations for its local model update in each communication round. Let  $\theta_{m,t}^{k,\tau}$  denote the local model update by device  $k \in \mathcal{K}_i$  at iteration  $\tau \in \{0, \dots, J-1\}$ , with  $\theta_{m,t}^{k,0} = \hat{\theta}_{m,t}^{k,\tau}$  and  $\mathcal{B}_{m,t}^{k,\tau}$  the mini-batch at iteration  $\tau$ , which is a subset of  $\mathcal{S}_{m,t}^k$ . The local model update is given by

$$\begin{aligned}\theta_{m,t}^{k,\tau+1} &= \theta_{m,t}^{k,\tau} - \eta_n \nabla F_m^k(\theta_{m,t}^{k,\tau}; \mathcal{B}_{m,t}^{k,\tau}) \\ &= \theta_{m,t}^{k,\tau} - \frac{\eta_n}{|\mathcal{B}_{m,t}^{k,\tau}|} \sum_{(s,v) \in \mathcal{B}_{m,t}^{k,\tau}} \nabla L_m(\theta_{m,t}^{k,\tau}; s, v)\end{aligned}\quad (4)$$

where  $\eta_n$  is the learning rate in frame  $n$ ,  $\nabla F_m^k$  and  $\nabla L_m$  are the gradients of the corresponding loss functions for model  $m$  w.r.t.  $\theta_{m,t}^{k,\tau}$ . After  $J$  iterations, the device obtains the updated local model  $\theta_{m,t}^{k,J}$ .

### 3.3. Uplink Aggregation

The devices send their updated local models  $\theta_{m,t}^{k,J}$ 's to the BS via the uplink channels. For efficient transmission, we again represent  $\theta_{m,t}^{k,J}$  using a complex vector, similar to downlink transmission. That is, we re-express  $\theta_{m,t}^{k,J} = [(\tilde{\theta}_{m,t}^{k,Jr})^T, (\tilde{\theta}_{m,t}^{k,Jim})^T]^T$ , where  $\tilde{\theta}_{m,t}^{k,Jr}$  and  $\tilde{\theta}_{m,t}^{k,Jim}$  contain the first and second half of elements in  $\theta_{m,t}^{k,J}$ , respectively. The equivalent complex vector representation of  $\theta_{m,t}^{k,J}$  is thus given by  $\tilde{\theta}_{m,t}^{k,J} = \tilde{\theta}_{m,t}^{k,Jr} + j\tilde{\theta}_{m,t}^{k,Jim} \in \mathbb{C}^{\frac{D_m}{2}}$ .

We adopt the orthogonal multiple access technique in uplink to efficiently use the system bandwidth for local model aggregation at the BS. In particular, devices in the same group  $i$  send their local model updates  $\{\tilde{\theta}_{m,t}^{k,J}\}_{k \in \mathcal{K}_i}$  to the BS simultaneously via the same uplink channel. The channels among device groups are orthogonal to each other. Thus, for each model  $m$ , the BS aggregates the received local model updates from the corresponding assigned device group  $i$  via the over-the-air computation [13]. As a result, the BS has the complex equivalent global model update  $\tilde{\theta}_{m,t+1}$  given by

$$\tilde{\theta}_{m,t+1} = \sum_{k \in \mathcal{K}_i} \rho_k \tilde{\theta}_{m,t}^{k,J} + \tilde{\mathbf{n}}_{m,t}^{\text{ul}} \quad (5)$$

where  $\rho_k \in [0, 1]$  is the weight with  $\sum_{k \in \mathcal{K}_i} \rho_k = 1$ , and  $\tilde{\mathbf{n}}_{m,t}^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \sigma_u^2 \mathbf{I})$  is the AWGN at the BS receiver. The weight  $\rho_k$  represents the uplink processing effect including device transmission and BS receiver processing.

For local model update in (4), let  $\Delta \tilde{\theta}_{m,t}^k \triangleq \tilde{\theta}_{m,t}^{k,J} - \tilde{\theta}_{m,t}^{k,0}$  denote the equivalent complex representation of the local model difference after the local training at device  $k \in \mathcal{K}_i$  in round  $t$ . Using (2) and (5), we can express the global model update  $\tilde{\theta}_{m,t+1}$  from  $\tilde{\theta}_{m,t}$  as

$$\begin{aligned}\tilde{\theta}_{m,t+1} &= \tilde{\theta}_{m,t} + \sum_{k \in \mathcal{K}_i} \rho_k \Delta \tilde{\theta}_{m,t}^k + \sum_{k \in \mathcal{K}_i} \rho_k \tilde{\mathbf{n}}_{k,t}^{\text{dl}} + \tilde{\mathbf{n}}_{m,t}^{\text{ul}} \\ &+ \sum_{j \neq i} \sum_{k \in \mathcal{K}_j} \rho_k \frac{\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}} (\mathbf{w}_{j,n}^{\text{dl}})^H \mathbf{h}_{k,n}}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2} \cdot \frac{\|\tilde{\theta}_{m,t}\| \|\tilde{\theta}_{m(j),t}\|}{\|\tilde{\theta}_{m(j),t}\|}.\end{aligned}\quad (6)$$

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

## 4. MULTI-MODEL DOWNLINK BEAMFORMING DESIGN

We consider the transmission design in the multi-model FL system to maximize the training convergence rate. Recall that the BS transmits all  $M$  model updates simultaneously to devices via multicast beamforming. Consider the global model update  $\theta_{m,nM}$  for model  $m$  at the beginning of each frame  $n \in \mathcal{S} \triangleq \{0, \dots, S-1\}$ . We design the downlink beamforming vectors to minimize the maximum expected optimality gap to  $\theta_m^*$  among all  $M$  models after  $S$  frames,

subject to the transmitter power budget. The optimization problem is formulated as

$$\begin{aligned}\mathcal{P}_o : \quad & \min_{\{\mathbf{w}_{i,n}^{\text{dl}}\}} \max_{m \in \mathcal{M}} \mathbb{E}[\|\theta_{m,SM} - \theta_m^*\|^2] \\ \text{s.t.} \quad & \sum_{i=1}^M \|\mathbf{w}_{i,n}^{\text{dl}}\|^2 \leq D_{\max} P, \quad n \in \mathcal{S}\end{aligned}\quad (7)$$

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 first develop a more tractable upper bound on  $\mathbb{E}[\|\theta_{m,SM} - \theta_m^*\|^2]$  by analyzing the convergence rate of the global model update. Then, we propose a downlink multi-group multicast beamforming method to minimize this upper bound.

### 4.1. Convergence Rate Analysis

To analyze the model update convergence rate, we make the following assumptions on the local loss functions, the global model update, and the difference between the global and weighted average of the local loss functions. They are commonly assumed for the convergence analysis of the FL model training [9, 11, 14].

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

**Assumption 2.** Bounded model parameters:  $\|\tilde{\theta}_{m,t}\|^2 \leq r$ , for some  $r > 0$ ,  $\forall m \in \mathcal{M}, \forall t$ . Bounded stochastic gradients and sample-wise loss gradients:  $\mathbb{E}[\|\nabla F_m^k(\theta_m)\|^2] \leq \mu$ ,  $\|\nabla L_m(\theta_m; s_i, v_i)\|^2 \leq \beta_1 \|\nabla F_m^k(\theta_m)\|^2 + \beta_2$ , for some  $\mu > 0$ ,  $\beta_1 \geq 1$  and  $\beta_2 \geq 0$ ,  $\forall m \in \mathcal{M}, \forall k \in \mathcal{K}_{\text{tot}}, \forall t, \forall i$ .

**Assumption 3.** Bounded gradient divergence:  $\mathbb{E}[\|\nabla F_m(\theta_{m,t}) - \sum_{k=1}^K c_k \nabla F_m^k(\theta_{m,t}^{k,\tau})\|^2] \leq \phi$ , for some  $\phi \geq 0$ ,  $0 \leq c_k \leq 1$ ,  $\forall m \in \mathcal{M}, \forall \tau, \forall t$ .

We now analyze the global model convergence rate over frames for each model  $m$ . Based on (6), we first obtain the per-model global update over frames, i.e.,  $\theta_{m,nM}$ . Let device group  $i$  be the group that trains model  $m$  in communication round  $t$  at frame  $n$ . The device-model assignment between  $i$  and  $m$  is given in (1). Summing both sides of (6) over  $M$  rounds in frame  $n$ , and subtracting the complex version of the optimal  $\tilde{\theta}_m^*$  from both sides, we obtain

$$\tilde{\theta}_{m,(n+1)M} - \tilde{\theta}_m^* = \tilde{\theta}_{m,nM} - \tilde{\theta}_m^* + \sum_{t=nM}^{(n+1)M-1} \sum_{k \in \mathcal{K}_i} \rho_k \Delta \tilde{\theta}_{m,t}^k + \tilde{\mathbf{e}}_{m,n}$$

where  $\tilde{\mathbf{e}}_{m,n}$  is the accumulated error term in (6) over  $M$  rounds in frame  $n$ , given by

$$\begin{aligned}\tilde{\mathbf{e}}_{m,n} &\triangleq \sum_{t=nM}^{(n+1)M-1} \sum_{k \in \mathcal{K}_i} \rho_k \tilde{\mathbf{n}}_{k,t}^{\text{dl}} + \sum_{t=nM}^{(n+1)M-1} \tilde{\mathbf{n}}_{m,t}^{\text{ul}} \\ &+ \sum_{t=nM}^{(n+1)M-1} \sum_{k \in \mathcal{K}_j} \rho_k \frac{\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}} (\mathbf{w}_{j,n}^{\text{dl}})^H \mathbf{h}_{k,n}}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2} \cdot \frac{\|\tilde{\theta}_{m,t}\| \|\tilde{\theta}_{m(j),t}\|}{\|\tilde{\theta}_{m(j),t}\|}.\end{aligned}$$

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

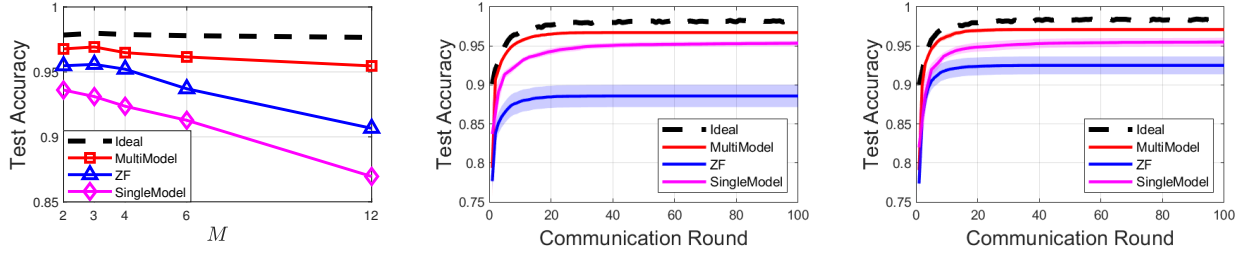
$$\mathbb{E}[\|\tilde{\mathbf{e}}_{m,n}\|^2] \leq rMK \sum_{i=1}^M \sum_{k \in \mathcal{K}_i} \frac{\sum_{j \neq i} |\mathbf{h}_{k,n}^H \mathbf{w}_{j,n}^{\text{dl}}|^2 + \tilde{\sigma}_d^2}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2} + M\tilde{\sigma}_u^2$$

where  $\tilde{\sigma}_d^2 \triangleq \sigma_d^2 D_{\max}/2$ , and  $\tilde{\sigma}_u^2 \triangleq \sigma_u^2 D_{\max}/2$ .

Using the above, we provide an upper bound on  $\mathbb{E}[\|\theta_{m,SM} - \theta_m^*\|^2]$  in Proposition 1 below. Due to the space limitation, the proof is omitted. Part of the proof adopts some techniques in [14, Th. 2].

**Proposition 1.** For the multi-model FL system described in Section 3, under Assumptions 1–3 and for  $\eta_n < \frac{1}{\lambda}$ ,  $\forall n$ , the expected model optimality gap after  $S$  frames is upper bounded by

$$\mathbb{E}[\|\theta_{m,SM} - \theta_m^*\|^2] \leq \Gamma_m \prod_{n=0}^{S-1} G_n + \sum_{n=0}^{S-2} H(\mathbf{w}_n^{\text{dl}}) \prod_{s=n+1}^{S-1} G_s$$



**Fig. 2:** Left: Test accuracy vs.  $M$  (from Model A). Middle & Right: Test accuracy vs. communication round  $t$ : Middle – Model A; Right – Model B.

$$+ H(\mathbf{w}_{S-1}^{\text{dl}}) + \Lambda, \quad m \in \mathcal{M} \quad (8)$$

where  $\Gamma_m \triangleq \mathbb{E}[\|\boldsymbol{\theta}_{m,0} - \boldsymbol{\theta}_m^*\|^2]$ ,  $G_n \triangleq 4(1 - \eta_n \lambda)^2$ ,  $\Lambda \triangleq \sum_{n=0}^{S-2} C_n (\prod_{s=n+1}^{S-1} G_s) + C_{S-1}$  with  $C_n \triangleq 4\eta_n^2 J^2 K^2 (\beta_1 \mu + \beta_2) + 4\eta_n^2 J \phi + 4M\sigma_u^2$ ,  $\mathbf{w}_n^{\text{dl}} \triangleq [(\mathbf{w}_{1,n}^{\text{dl}})^H, \dots, (\mathbf{w}_{M,n}^{\text{dl}})^H]^H$ , and

$$H(\mathbf{w}_n^{\text{dl}}) \triangleq rMK \sum_{i=1}^M \sum_{k \in \mathcal{K}_i} \frac{\sum_{j \neq i} |\mathbf{h}_{k,n}^H \mathbf{w}_{j,n}^{\text{dl}}|^2 + \tilde{\sigma}_d^2}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2}.$$

#### 4.2. Downlink Multi-Group Multicast Beamforming Design

We now replace the objective function in  $\mathcal{P}_o$  with the more tractable upper bound in (8). Note that only  $\Gamma_m \prod_{n=0}^{S-1} G_n$  in (8) depends on  $m$ . Omitting the constant terms  $\Gamma_m \prod_{n=0}^{S-1} G_n + \Lambda$ , we arrive at the following equivalent optimization problem w.r.t. multicast beamforming vectors  $\{\mathbf{w}_n^{\text{dl}}\}$  over  $S$  frames:

$$\mathcal{P}_1 : \min_{\{\mathbf{w}_n^{\text{dl}}\}_{n=0}^{S-1}} \sum_{n=0}^{S-2} H(\mathbf{w}_n^{\text{dl}}) \prod_{s=n+1}^{S-1} G_s + H(\mathbf{w}_{S-1}^{\text{dl}}) \quad \text{s.t.} \quad (7).$$

Note that by Proposition 1,  $G_n > 0$ , for  $\eta_n < \frac{1}{\lambda}$ ,  $n \in \mathcal{S}$ , and  $\prod_{s=n+1}^{S-1} G_s > 0$ . Thus,  $\mathcal{P}_1$  can be decomposed into  $S$  subproblems to solve, one for each frame  $n$ , given by

$$\mathcal{P}_{2,n} : \min_{\mathbf{w}_n^{\text{dl}}} \sum_{i=1}^M \sum_{k \in \mathcal{K}_i} \frac{\sum_{j \neq i} |\mathbf{h}_{k,n}^H \mathbf{w}_{j,n}^{\text{dl}}|^2 + \tilde{\sigma}_d^2}{|\mathbf{h}_{k,n}^H \mathbf{w}_{i,n}^{\text{dl}}|^2} \quad \text{s.t.} \quad \sum_{i=1}^M \|\mathbf{w}_{i,n}^{\text{dl}}\|^2 \leq D_{\max} P.$$

Problem  $\mathcal{P}_{2,n}$  is a multi-group multicast beamforming problem with  $M$  multicast beamforming vectors, one for each device group, to optimize. The objective is a total sum of interference-and-noise-to-signal ratios at the BS receiver as the result of downlink-uplink processing. The family of multicast beamforming problems is non-convex and NP-hard [15, 18]. We propose to use PGD to solve  $\mathcal{P}_{2,n}$ . Since we can obtain the fast closed-form gradient updates, PGD is suitable for solving  $\mathcal{P}_{2,n}$ . Furthermore, it is guaranteed to find an approximate stationary point of  $\mathcal{P}_{2,n}$  in polynomial time [19].

### 5. SIMULATION RESULTS

We consider the image classification task under an LTE system setting. Following the typical LTE specifications, we set system bandwidth 10 MHz and the maximum BS transmit power 47 dBm. The channels are generated i.i.d. as  $\mathbf{h}_{k,t} = \sqrt{G_k} \bar{\mathbf{h}}_{k,t}$  with  $\bar{\mathbf{h}}_{k,t} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ , and  $G_k$  being the path gain from the BS to device  $k$ , modeled as  $G_k[\text{dB}] = -169.2 - 35 \log_{10} d_k - \psi_k$ , where the BS-device distance  $d_k \in (0.02, 0.5)$  km, and  $\psi_k$  represents shadowing with standard deviation 8 dB. Noise power spectral density is  $N_0 = -174$  dBm/Hz, with noise figure  $N_F = 8$  dB and 2 dB at the device and BS receivers, respectively. We use the MNIST

dataset [20] for the multi-model training and testing. It consists of 60,000 training samples and 10,000 test samples. We consider training two types of convolutional neural networks: i) **Model A**: an  $8 \times 3 \times 3$  ReLU convolutional layer, a  $2 \times 2$  max pooling layer, and a softmax output layer with 13,610 parameters. ii) **Model B**: an  $8 \times 3 \times 3$  ReLU convolutional layer, a  $2 \times 2$  max pooling layer, a ReLU fully-connected layer with 20 units, and a softmax output layer with 27,350 parameters. The training samples are randomly and evenly distributed across devices. The local dataset at device  $k$  has  $S_k = 60,000/K$  samples. For the local training using SGD at each device, we set  $\lambda = 3$ ,  $J = 100$ , mini-batch size  $|\mathcal{B}_{m,t}^{k,\tau}| = 600/K, \forall k, \tau, m, t$ , and the learning rate  $\eta_n = 0.2, \forall n$ . All results are obtained by taking the current best test accuracy and averaged over 20 channel realizations.

Besides our proposed method, denoted by MultiModel, we consider three schemes for comparison: i) **Ideal**: Perform multi-model FL via (6) with noise-interference-free downlink/uplink and perfect recovery of model parameters. ii) **ZF**: Perform multi-model FL via (6) using the zero-forcing (ZF) multicast beamformers proposed in [21]. iii) **SingleModel**: Use the single-model FL with downlink multicast beamforming for SNR maximization considered in (31) of [13] to train multiple models sequentially with  $K$  devices.

Fig. 2-Left shows the test accuracy after 30 communication rounds vs. training  $M$  models all from Model A. We set  $(N, K) = (128, 12)$ . Our proposed MultiModel outperforms all other alternatives. Its performance remains roughly unchanged as  $M$  increases and can achieve  $\sim 97\%$  test accuracy after 30 rounds, while other schemes noticeably deteriorate as  $M$  increases. Consider  $M = 2$ , and one from Models A and B each. Fig. 2-Middle and Right show the test accuracy over round  $t$  for Models A and B, respectively. We set  $(N, K) = (128, 10)$ . The shadow area for each curve indicates the 90% confidence interval. MultiModel again outperforms other alternatives for both Models A and B. Between Models A and B, we see that Model B, which is the larger one, achieves slightly higher test accuracy than Model A under both multi-model and single-model training.

### 6. CONCLUSION

Multi-model wireless FL with imperfect transmission/processing over noisy channels is considered in this paper. We formulate the downlink-uplink transmission process and obtain the per-model global update expression in each round. We design downlink beamforming to maximize the FL training performance. Using an upper bound on the optimality gap of the global model update, we optimize downlink multicast beamforming for sending multiple models simultaneously to device groups, which leads to a multi-group multicast beamforming problem for minimizing the sum of the inverse of received SINRs. Simulation results demonstrate the effectiveness of the proposed multi-model method compared with the other alternatives.



## 7. REFERENCES

- [1] 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] 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, no. 1, pp. 19–25, Jan. 2020.
- [3] M. Chen, D. Gndz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, Dec. 2021.
- [4] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.
- [5] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.
- [6] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, Aug. 2021.
- [7] Y. Sun, S. Zhou, Z. Niu, and D. Gndz, "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.
- [8] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4434–4449, Jun. 2022.
- [9] M. M. Amiri, D. Gndz, 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.
- [10] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Trans. Cogn. Commun.*, vol. 8, no. 2, pp. 1253–1268, Jun. 2022.
- [11] 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.
- [12] 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.
- [13] C. Zhang, M. Dong, B. Liang, A. Afana, and Y. Ahmed, "Joint downlink-uplink beamforming for wireless multi-antenna federated learning," in *Proc. Int. Symp. Model. Optim. Mobile Ad hoc Wireless Netw.*, Aug. 2023. Available: <https://arxiv.org/abs/2307.00315>.
- [14] N. Bhuyan, S. Moharir, and G. Joshi, "Multi-model federated learning with provable guarantees," in *Proc. Int. Conf. Perform. Eval. Methodologies Tools*, Nov. 2023, pp. 207–222.
- [15] 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.
- [16] C. Zhang, M. Dong, and B. Liang, "Ultra-low-complexity algorithms with structurally optimal multi-group multicast beamforming in large-scale systems," *IEEE Trans. Signal Process.*, vol. 71, pp. 1626–1641, Apr. 2023.
- [17] S. Bubeck, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3–4, pp. 231–357, 2015.
- [18] 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.
- [19] A. Mokhtari, A. Ozdaglar, and A. Jadbabaie, "Escaping saddle points in constrained optimization," in *Proc. Advances Neural Inf. Process. Syst.*, Dec. 2018, pp. 3629–3639.
- [20] Y. LeCun, C. Cortes, and C. Burges, "The MNIST Database of Handwritten Digits," 1998. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>.
- [21] M. Sadeghi, E. Bjrnson, E. G. Larsson, C. Yuen, and T. L. Marzetta, "Maxmin fair transmit precoding for multi-group multicasting in massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1358–1373, Feb. 2018.