Adaptive Transceiver Design for Wireless Hierarchical Federated Learning

Fangtong Zhou*, Xu Chen†, Hangguan Shan‡, and Yong Zhou*
*School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
†School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510275, China
‡College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, 310027, China
E-mail: *{zhouft, zhouyong}@shanghaitech.edu.cn, †chenxu35@mail.sysu.edu.cn, ‡hshan@zju.edu.cn

2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall) | 979-8-3503-2928-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/VTC2023-Fall60731.2023.10333457

***Abstract*—Deploying federated learning (FL) in wireless networks faces the critical challenge of communication bottlenecks. To address this issue, in this paper, we consider an over-the-air computation (AirComp) assisted hierarchical FL (HFL) framework, where a cloud-edge-device-based three-tier network architecture is constructed to train a global model. We first theoretically characterize the convergence of the AirComp-assisted HFL framework and formulate a combinatorial optimization problem that jointly optimizes the edge interval control and local device transceiver design to minimize the convergence upper bound to boost the overall learning performance and reduce communication cost. We show that the formulated optimization problem can be decoupled into an edge interval control problem and a transceiver design problem, which can be tackled by developing a relaxation and rounding algorithm and an alternating Lyapunov drift-based algorithm, respectively. Extensive simulations demonstrate that our proposed algorithm significantly outperforms the baseline schemes.**

I. Introduction

The ever-growing amount of raw data generated by edge devices can be exploited to distill intelligence for supporting intelligent applications. However, uploading these data to a central server for centralized machine learning (ML) may cause high communication cost and the leakage of user privacy. To this end, federated learning (FL) is proposed as an effective distributed ML paradigm, which can utilize geographically distributed data and computing resources to collaboratively train a statistical model [1]. By exchanging model parameters instead of raw data, FL can enhance data privacy and reduce communication cost [2].

However, deploying FL in wireless networks can be challenging due to random channel fading and limited spectrum resources. To address these challenges, hierarchical FL (HFL) has gained great popularity recently, given its advantage of mitigating the impact of channel fading by allowing devices to transmit their model updates to nearby edge servers and in turn reducing communication distance and delay. The model updates held by edge servers are uploaded to the cloud server after multiple rounds of aggregation and dissemination [3]–[6].

Most existing works on wireless HFL focus on subcarrier assignment [7], power control [8], and interference management [9]. The authors in [10] provide the convergence analysis of HFL in an error-free scenario, and demonstrate that the edge aggregation interval control plays a vital role in determining the convergence bound of FL. A control algorithm is proposed in [11] to balance the tradeoff between local updates and global communication rounds in two-tier FL under the resource budget. In [12], the authors perform the edge interval control in three-tier FL and further minimize the weighted sum of training loss and latency. However, the orthogonal multiple access (OMA) scheme adopted in these works may cause huge communication cost when more devices are accessed to the networks. Over-the-air computation (AirComp), as an emerging technique, can effectively overcome the challenge of limited spectrum resources. It allows concurrent and non-orthogonal transmission from edge devices to the server, which is based on the characteristic of analog waveform superposition, and a specific function of concurrently transmitted signals can be directly received at the server via AirComp without decoding each signal. Such a feature makes the scheme capable of achieving spectrum-efficient model aggregation in wireless FL, as the server only needs the average of local models to update the global model [13]–[16]. To this end, we leverage the AirComp aggregation scheme in HFL, and to the best of our knowledge, there is no existing work analyzing the edge interval control in AirComp-assisted HFL, which is complicated due to the introduction of aggregation noise.

In this paper, to overcome the communication bottleneck and determine an appropriate tradeoff between local iterations and edge aggregations in wireless HFL, we propose an AirComp-assisted HFL framework over wireless networks. We aim at characterizing the convergence of the proposed framework and jointly optimizing the edge interval control and transceiver design to promote the learning performance. The main contributions of this paper are three-folds.

- The convergence of our proposed AirComp-assisted HFL is theoretically derived and analyzed. We formulate the upper bound of the time-averaged norm of global gradients into the function of edge interval, device transmit power, and edge normalizing coefficient.
- A combinatorial optimization problem of edge interval control and local device transceiver design is formulated to minimize the convergence upper bound. To tackle the formulated non-convex combinatorial optimization problem, we decouple the formulated optimization problem into two subproblems of edge interval control and device transceiver design. A relaxation and rounding algorithm and an alternating Lyapunov drift-based algorithm are then developed to solve the above two subproblems, respectively.
- Simulations demonstrate that our proposed edge interval control algorithm can successfully seek an optimal edge interval, and our proposed transceiver design scheme outperforms the baseline channel inversion scheme and full power scheme by 5.45% and 18.73% in test accuracy.

This was supported in part by the National Natural Science Foundation of China under Grants U20A20159, 62001294, 61971286, U21B2029 and U21A20456, in part by the Natural Science Foundation of Shanghai under Grant 23ZR1442800, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010006.

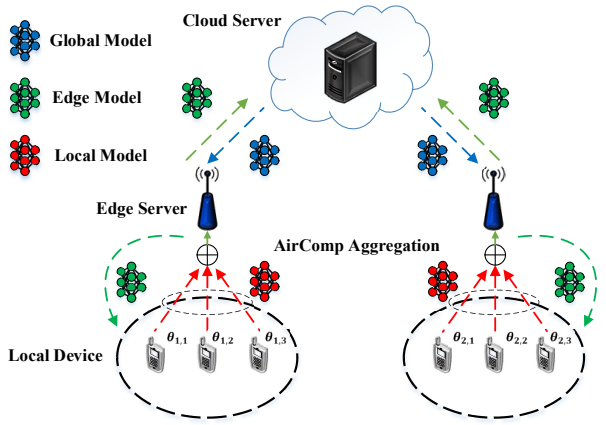


Fig. 1. Illustration of AirComp-assisted HFL.

II. System Model

A. Hierarchical FL Model

Consider a three-tier hierarchical FL framework over a cloud-edge-device architecture, as shown in Fig. 1, where a cloud server is connected to N edge servers and each edge server is associated with its M proximal devices that have non-independent and identically distributed (non-i.i.d.) local datasets of the same size. We denote the i-th edge server as e_i, the j-th device in the i-th cluster as $d_{i,j}$, and the local dataset of device $d_{i,j}$ as $\mathcal{B}_{i,j}$. To train a global model $\boldsymbol{w} \in \mathbb{R}^D$, we minimize the following objective function

$$\min_{\boldsymbol{w}} F(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} F_i(\boldsymbol{w}), \tag{1}$$

with $F_i(\boldsymbol{w}) = \frac{1}{M}\sum_{j=1}^{M} F_{i,j}(\boldsymbol{w})$. Herein, $F_{i,j}(\boldsymbol{w})$ is the empirical loss function over the local dataset $\mathcal{B}_{i,j}$, which can be expressed as

$$F_{i,j}(\boldsymbol{w}) = \frac{1}{|\mathcal{B}_{i,j}|}\sum_{b_{i,j}\in\mathcal{B}_{i,j}} \ell_{i,j}(\boldsymbol{w}, b_{i,j}), \tag{2}$$

where $\ell_{i,j}(\boldsymbol{w}, b_{i,j})$ denotes the sample-wise loss function over sample $b_{i,j}$.

In the hierarchical FL system, we denote each uploading from local device $d_{i,j}$ to edge server e_i as a communication round, and the period from the start of cloud server broadcasting the global model to the end of cloud server receiving and aggregating the edge models as a global communication round. Local devices perform K local iterations in one communication round and edge servers perform edge aggregations for R times in one global communication round. The process of HFL can be expressed as follows:

- At the t-th communication round, a global model $\boldsymbol{w}^t$ is broadcast to each edge server by the cloud server.
- After receiving the global model, each edge server e_i initializes its edge model $\boldsymbol{w}_i^t = \boldsymbol{w}^t$ and broadcasts it to its associated devices.
- At the device side, after receiving the edge model $\boldsymbol{w}_i^t$ at the r-th edge communication round ($r \in \{0, 1, \ldots, R-1\}$), each device in the i-th cluster performs local iteration as follows

$$\boldsymbol{w}_{i,j}^{t,k+1} = \boldsymbol{w}_{i,j}^{t,k} - \eta \boldsymbol{g}_{i,j}^{t,k}, \tag{3}$$

where η is the step size and $\boldsymbol{g}_{i,j}^{t,k}$ is the mini-batch stochastic gradient given by can be expressed as

$$\frac{1}{|\tilde{\mathcal{B}}_{i,j}|}\sum_{\zeta_{i,j}\in\tilde{\mathcal{B}}_{i,j}} \nabla \ell_{i,j}(\boldsymbol{w}_{i,j}^{t,k}, \zeta_{i,j}), \tag{4}$$

with $\tilde{\mathcal{B}}_{i,j}$ representing the mini-batch dataset randomly sampled from the local dataset $\mathcal{B}_{i,j}$ and $\zeta_{i,j}$ denoting a single pair of data and label randomly sampled from the mini-batch dataset $\tilde{\mathcal{B}}_{i,j}$. After performing local iterations for K times, each device uploads the accumulated gradient $\boldsymbol{g}_{i,j}^t = (\boldsymbol{w}_{i,j}^{t,0} - \boldsymbol{w}_{i,j}^{t,K-1})/\eta = \sum_{k=0}^{K-1} \boldsymbol{g}_{i,j}^{t,k}$ to the edge server e_i. Then, edge server e_i performs the model aggregation to update the edge model $\boldsymbol{w}_i^{t+1}$ as follows:

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^t - \eta\frac{1}{M}\sum_{j=1}^{M} \boldsymbol{g}_{i,j}^t, \tag{5}$$

and then broadcasts it to devices in cluster i to start the $(r+1)$-th edge communication round.

- After performing R times of aggregations at the edge side, each edge server transmits its edge model $\boldsymbol{w}_i^{t+R-1}$ to the cloud server for the cloud aggregation. Then, the cloud server updates the global model as $\boldsymbol{w}^{t+R-1}$ and starts the $(t+R)$-th communication round.

B. Model Aggregation via AirComp

We adopt AirComp to achieve low-latency model aggregation between each edge server and its associated devices. This allows all devices $\{d_{i,j}\}_{j=1}^M$ in the i-th cluster to simultaneously transmit their locally-trained accumulated gradients $\{\boldsymbol{g}_{i,j}\}_{j=1}^M$ to edge server e_i. It is assumed that each cluster occupies an orthogonal channel with the same bandwidth and that communication between each edge server and the cloud server is error-free as in [5]. We normalize D-dimensional accumulated local gradient $\boldsymbol{g}_{i,j}^t$ before the uplink transmission to facilitate the power control. In particular, device $d_{i,j}$ calculates mean $\bar{g}_{i,j}^t$ and variance $(\pi_{i,j}^t)^2$ of $\boldsymbol{g}_{i,j}^t$, $\forall\, i, j$, as follows

$$\bar{g}_{i,j}^t = \frac{1}{D}\sum_{d=1}^{D} g_{i,j}^t(d),\ (\pi_{i,j}^t)^2 = \frac{1}{D}\sum_{d=1}^{D}(g_{i,j}^t(d) - \bar{g}_{i,j}^t)^2 \tag{6}$$

where $g_{i,j}^t(d)$ denotes the d-th element of $\boldsymbol{g}_{i,j}^t$. By setting $\bar{g}_i^t = \frac{1}{M}\sum_{j=1}^M \bar{g}_{i,j}^t$ and $(\pi_i^t)^2 = \frac{1}{M}\sum_{j=1}^M (\pi_{i,j}^t)^2$, $\boldsymbol{g}_{i,j}^t$ can be normalized as

$$\boldsymbol{\theta}_{i,j}^t = \frac{\boldsymbol{g}_{i,j}^t - \bar{g}_i^t}{\pi_i^t},\ \forall\, i, j. \tag{7}$$

We assume that $\{\boldsymbol{\theta}_{i,j}^t\}_{j=1}^M$ are independent and have zero mean and unit variance, i.e., $\mathbb{E}[\boldsymbol{\theta}_{i,j}^t(\boldsymbol{\theta}_{i,j}^t)^H] = \boldsymbol{I}_D$ and $\mathbb{E}[\boldsymbol{\theta}_{i,j}^t(\boldsymbol{\theta}_{i,j'}^t)^H] = \boldsymbol{0}$, $\forall\, j \neq j'$. All devices in the i-th cluster transmit their normalized accumulated local gradients $\{\boldsymbol{\theta}_{i,j}^t\}_{j=1}^M$ to edge server e_i simultaneously. In the t-th communication round, the channel coefficient between device $d_{i,j}$ and edge server e_i is denoted as $h_{i,j}^t \in \mathbb{C}$, which is assumed to follow block fading and be known by devices in cluster i. We denote $\varphi_{i,j}^t = \frac{\sqrt{p_{i,j}^t}(h_{i,j}^t)^H}{|h_{i,j}^t|} \in \mathbb{C}$ as the pre-processing coefficient of device $d_{i,j}$ to compensate for the phase distortion caused by channel fading, where $p_{i,j}^t \geq 0$ represents the transmit power of device $d_{i,j}$ in communication round t.

Hence, the aggregation of local gradients received at the edge server e_i is

$$\boldsymbol{q}_i^{t+1} = \sum_{j=1}^{M} h_{i,j}^t \varphi_{i,j}^t \boldsymbol{\theta}_{i,j}^t + \boldsymbol{n}_i^t \tag{8}$$

where $\boldsymbol{n}_i \sim \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_D)$ is the additive white Gaussian noise (AWGN). The estimated function of the received signal can be computed as

$$\check{\boldsymbol{\theta}}_i^{t+1} = \frac{\boldsymbol{q}_i^{t+1}}{\sqrt{\eta_i^t}} = \sum_{j=1}^{M} \frac{\sqrt{p_{i,j}^t}(h_{i,j}^t)^H}{\sqrt{\eta_i^t}} \boldsymbol{\theta}_{i,j}^t + \frac{\boldsymbol{n}_i^t}{\sqrt{\eta_i^t}}, \tag{9}$$

where $\eta_i^t \geq 0$ is the normalizing coefficient we applied at edge server e_i to align the signal amplitude. After de-normalization, we get

$$\begin{aligned} \check{\boldsymbol{g}}_i^{t+1} &= \frac{1}{M}(\pi_i^t \check{\boldsymbol{\theta}}_i^{t+1} + M \bar{g}_i^t) \\ &= \frac{1}{M}\pi_i^t(\check{\boldsymbol{\theta}}_i^{t+1} - \boldsymbol{\theta}_i^{t+1}) + \boldsymbol{g}_i^{t+1}, \end{aligned} \tag{10}$$

where $\boldsymbol{\theta}_i^{t+1} = \sum_{j=1}^{M} \boldsymbol{\theta}_{i,j}^t$ and $\boldsymbol{g}_i^{t+1} = \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{g}_{i,j}^t$. Since only the estimated signals can be transmitted to the cloud server at each communication round, the updated global model can be expressed as

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}(\boldsymbol{g}_{i,j}^t + \boldsymbol{e}_i^t), \tag{11}$$

where $\boldsymbol{e}_i^t = \frac{1}{M}\pi_i^t(\check{\boldsymbol{\theta}}_i^{t+1} - \boldsymbol{\theta}_i^{t+1})$ according to (10) means the aggregation error at communication round t, which will determine the convergence performance.

In addition, we consider that in each global communication round, total local iterations $G = RK$ performed by each device is fixed as in [4], which means that all devices must reach a given local accuracy in each global communication round. Moreover, we consider the total transmit power $P_{i,j}^{\text{tot}}$ in each global communication round and the maximum transmit power $P_{i,j}^{max}$ of device $d_{i,j}$ are given, hence the transmit power $p_{i,j}^t$ has following constraints:

$$p_{i,j}^t \leq P_{i,j}^{\max}, \forall\, i, j, \tag{12}$$

$$\sum_{t=aR}^{aR+R-1} p_{i,j}^t \leq P_{i,j}^{\text{tot}},\ a \in \mathbb{N}. \tag{13}$$

Intuitively, with a fixed number of local iterations G, more frequent edge aggregations (i.e., smaller K) lead to better performance as the detrimental impact of non-i.i.d. dataset can be mitigated (This will be proved in the next section). However, with a fixed total transmit power budget, more frequent edge aggregations lead to small average transmit power $P_{i,j}^{\text{tot}}/R$, which might not be able to compensate for the distortion caused by the unreliable wireless channel.

III. Convergence Analysis and Problem Formulation

We provide the convergence analysis of the AirComp-assisted HFL framework, demonstrate the objective function according to edge interval K, and formulate an optimization problem to adaptively optimize the edge interval K and minimize the transmission distortion in the convergence upper bound.

A. Assumptions

Assumption 1 (Smoothness)**.** *We assume that $\ell_{i,j}(\boldsymbol{w})$ is L-smooth on $\mathbb{R}^D$, i.e.,*

$$\|\nabla \ell_{i,j}(\boldsymbol{w}) - \nabla \ell_{i,j}(\boldsymbol{w}')\|_2 \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|_2. \tag{14}$$

Assumption 2 (Bounded Variance)**.** *The stochastic gradient in each device has a bounded variance, i.e.,*

$$\mathbb{E}_{\zeta_{i,j}}\left[\|\boldsymbol{g}_{i,j}(\boldsymbol{w}, \zeta_{i,j}) - \nabla \ell_{i,j}(\boldsymbol{w})\|\right]^2 \leq \xi^2. \tag{15}$$

Assumption 3 (Bounded Global and Local Diversity)**.** *The dissimilarity between the local, edge, and global loss functions can be constrained as*

$$\frac{1}{M}\sum_{j=1}^{M}\|\nabla \ell_{i,j}(\boldsymbol{w}) - \nabla F_i(\boldsymbol{w})\|^2 \leq \epsilon_i^2, \forall\, i \tag{16}$$

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|^2 \leq \epsilon^2. \tag{17}$$

Assumption 4 (Local Accumulated Gradient Variance Bound)**.** *The variance of D elements of $\boldsymbol{g}_{i,j}$ has the constant upper bound $\Xi \geq 0$, i.e., $\pi_{i,j}^2 \leq \Xi$.*

Remark 1. *Assumption 1 can also be applied to loss functions at edge side (i.e.$\nabla F_i(\cdot)$) and cloud side (i.e.$\nabla F(\cdot)$) [10], Assumptions 2 and 3 are commonly made in the HFL convergence analysis [3]. In Assumption 4, we assume that $\pi_{i,j}^2$ has a non-negative upper bound as in [14] since the value of elements in accumulated local gradients $\boldsymbol{g}_{i,j}$ are finite.*

B. Convergence Analysis

Theorem 1. *Under Assumptions 1, 2, 3, and 4, by letting $\eta \leq \left\{\frac{1}{2KL}, \sqrt{\frac{3}{32L^2K^3}}, \frac{1}{\sqrt{8L^2K^2}}\right\}$, the time-averaged norm of global gradients after T communication rounds has a upper bound, i.e.,*

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2 \leq \frac{f(\boldsymbol{w}^0) - \mathbb{E}[f(\boldsymbol{w}^T)]}{\Omega\eta T} \\ + \frac{\Psi}{\Omega} + \frac{\frac{1}{2} + \eta L}{\Omega T}\frac{D\Xi(NM+1)}{N^2M^2}\sum_{t=0}^{T-1} MSE(t), \end{aligned} \tag{18}$$

where

$$\Omega = \frac{K}{2} - \frac{1}{2} - \frac{\frac{8}{3}\eta^2L^2K^3(1 - 8\eta^2L^2G^2)}{(1 - 12\eta^2L^2G^2)(1 - 4\eta^2L^2K^2)} \geq \frac{K-2}{2}, \tag{19}$$

and $\Psi = \eta LK\xi^2 + KL^2\frac{4\eta^2G(1-\frac{N-1}{N^2})\frac{1}{M}\xi^2 + 6\eta^2G^2\epsilon^2}{1-12\eta^2L^2G^2} + \frac{\frac{2}{3}\eta^2L^2K^3(\xi^2 + \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2 + \epsilon^2)}{(1-12\eta^2L^2G^2)(1-4\eta^2L^2K^2)}$, *which is a function of edge interval K that denotes the variance of stochastic gradient and gradient divergence. The third term is denoted as MSE, we can easily write the instantaneous $MSE(t)$ as*

$$MSE(t) = \sum_{i=1}^{N}\left[\sum_{j=1}^{M}\left(\frac{\sqrt{p_{i,j}^t}|h_{i,j}^t|}{\sqrt{\eta_i^t}} - 1\right)^2 + \frac{\sigma_0^2}{\eta_i^t}\right]. \tag{20}$$

The proof follows a similar idea to that in [10], [14] and is omitted due to space limitation.

Remark 2. *In (18), the first term in the upper bound is caused by the initial optimality gap, the second term is caused by the random mini-batch sampling, and the model diversity with respect to the edge server drift error, while the last term is the time-averaged MSE due to channel fading and receiver noise. The numerator of the second term Ψ monotonically increases with local iteration time K, while K also appears in the denominator and the third term, which makes it hard to determine how K affects the upper bound. As T goes to infinity, the initial optimality gap decreased to 0. As we discussed in the previous section, with a smaller edge interval K, the initial optimality gap decreases while the time-averaged MSE may change in a different trend. Therefore, the convergence of AirComp-assisted HFL is mainly hampered by adjusting the edge interval K and the time-average MSE given in (20), which needs to be minimized to improve the learning performance.*

C. Problem Formulation

Minimizing the upper bound in (18) is challenging since future channel state information (CSI) in each time slot is not known. Hence we use Lyapunov optimization [17] to perform the online power control and edge interval control.

To decompose the long-term constraints (13) into each time slot, we construct a virtual queue $\mathcal{Q}_{i,j}(t)$ to store the power information in each transmission and capture the dynamics of the power state, which can be expressed as

$$\mathcal{Q}_{i,j}(t+1) = \max\{\mathcal{Q}_{i,j}(t) + p_{i,j}^t - \frac{P_{i,j}^{\text{tot}}}{R}, 0\}. \tag{21}$$

By defining $\boldsymbol{S}_i(t) \triangleq [\mathcal{Q}_{i,1}(t), \ldots, \mathcal{Q}_{i,M}(t)]$ as the combined matrix of all virtual queues in cluster i, we define the Lyapunov function as

$$L(\boldsymbol{S}_i(t)) \triangleq \frac{1}{2}\sum_{j=1}^{M} \mathcal{Q}_{i,j}^2(t). \tag{22}$$

With (22), we use

$$\Delta(\boldsymbol{S}_i(t)) \triangleq \mathbb{E}\left[L(\boldsymbol{S}_i(t+1)) - L(\boldsymbol{S}_i(t))|\boldsymbol{S}_i(t)\right] \tag{23}$$

to indicate the Lyapunov drift, i.e., the increment between two time slots. Hence, in communication round t, to optimize the MSE under power constraints, we minimize the following Lyapunov drift-plus-penalty:

$$\sum_{i=1}^{N} \Delta(\boldsymbol{S}_i(t)) + \rho\frac{(\frac{1}{2}+\eta L)D\Xi(NM+1)}{\Omega N^2M^2}\text{MSE}(t), \tag{24}$$

where $\rho \in (0, \infty)$ is a parameter to control the balance between the Lyapunov drift and MSE(t). Based on the property of Lyapunov optimization, we have the following Lemma.

Lemma 1. *With $\mathcal{Q}_{i,j}(0) = 0, \forall i, j$, the Lyapunov drift satisfies*

$$\Delta(\boldsymbol{S}_i(t)) \le \Lambda + \sum_{j=1}^{M} \mathcal{Q}_{i,j}(t)\left(p_{i,j}^t - \frac{P_{i,j}^{\text{tot}}}{R}\right), \tag{25}$$

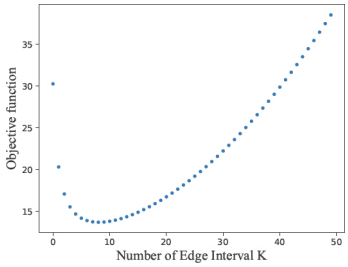


Fig. 2. Objective function with respect to K.

where $\frac{1}{2}\sum_{j=1}^{M}\left(p_{i,j}^t - \frac{P_{i,j}^{\text{tot}}}{R}\right)^2 \le \frac{1}{2}\sum_{j=1}^{M}(P_{i,j}^{\max} - P_{i,j}^{\text{tot}})^2 \le \Lambda$ is a common bound widely used in Lyapunov optimization [18].

Combining the upper bound in (18), (24), and Lemma 1, we formulate a combinatorial optimization problem by jointly optimizing the edge interval K, the transmit power $p_{i,j}^t$, and edge normalizing coefficient η_i^t as follows,

$$\begin{aligned}
\mathcal{P}: \min_{K, p_{i,j}^t, \eta_i^t} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M} \mathcal{Q}_{i,j}(t)(p_{i,j}^t - \frac{P_{i,j}^{\text{tot}}K}{G}) + \frac{2\Psi}{K-2} \\
& + \rho\frac{(1+2\eta L)D\Xi(NM+1)}{KN^2M^2}\text{MSE}(t) && \text{(26a)} \\
\text{s.t.} \quad & \eta_i^t \ge 0, \forall i, && \text{(26b)} \\
& p_{i,j}^t \le P_{i,j}^{\max}, \forall i, j, && \text{(26c)} \\
& K \,|\, G = 0. && \text{(26d)}
\end{aligned}$$

Constraint (26d) denotes that G is an integer multiple of K. To show how the edge interval K affects the performance of AirComp-assisted HFL more intuitively, we fix $p_{i,j}$ and η_i and plot the curve of the objective function with respect to K. As shown in Fig. 2, there exists an optimal K to minimize the objective function, which balances the trade-off between the communication cost introduced by overmuch communication rounds and the model diversity introduced by overmuch local iterations.

IV. Lyapunov Drift-based Alternating Algorithm

In this section, an alternating algorithm is developed to decouple problem (26a) into two subproblems, i.e., edge interval control problem with given transmit power and transceiver design problem with fixed edge interval.

A. Edge Interval Control

At the t-th communication round, with a fixed transmit power $p_{i,j}^t$ and edge normalizing coefficient η_i^t, the problem $\mathcal{P}$ becomes

$$\begin{aligned}
\mathcal{P}_1: \min_{K} \quad & aK + \frac{2\Psi}{K-2} + \rho\frac{b}{K} && \text{(27a)} \\
\text{s.t.} \quad & \text{(26d)}. && \text{(27b)}
\end{aligned}$$

where Ψ has been declared before which is affected by K, $a = -\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\mathcal{Q}_{i,j}(t)P_{i,j}^{\text{tot}}}{G}$, and $b = \frac{(1+2\eta L)D\Xi(NM+1)\text{MSE}}{N^2M^2}$. We adopt relaxation and rounding method [19] to relax the constraint (26d) into $1 \le K \le G$, which makes problem $\mathcal{P}_1$ become convex, then we use the Theorem 3 in [12] to compute K.

B. Transceiver Design

With the edge interval K fixed, problem $\mathcal{P}$ can be expressed as

$$\mathcal{P}_2: \min_{p_j,\eta\geq 0} \quad \rho\sum_{j=1}^{M}\left(\frac{\sqrt{p_j}|h_j|}{\sqrt{\eta}}-1\right)^2+\frac{\sigma_0^2}{\eta}+\sum_{j=1}^{M}\mathcal{Q}_j p_j \tag{28a}$$

$$\text{s.t.} \quad p_j \leq P_j^{\max}, \forall\, j. \tag{28b}$$

We omit i since each cluster is independent of one another and we omit t since we leverage Lyapunov optimization to decompose the long-term constraint (13) into each time slot. To this end, an alternating method is also used to iteratively optimize p_j and η. Firstly, with a given transmit power p_j, problem $\mathcal{P}_2$ becomes

$$\mathcal{P}_{21}: \min_{\eta\geq 0} \Upsilon(\eta) \triangleq \sum_{j=1}^{M}\left(\frac{\sqrt{p_j}|h_j|}{\sqrt{\eta}}-1\right)^2+\frac{\sigma_0^2}{\eta}, \tag{29}$$

by denoting $\Phi = 1/\sqrt{\eta}$, the objective function in (29) can be written as

$$\Upsilon(\Phi) = \sum_{j=1}^{M}\left(\sqrt{p_j}|h_j|\Phi-1\right)^2+(\sigma_0\Phi)^2, \tag{30}$$

which is convex. Therefore, we can solve the optimal η^* by setting the first-order derivative of $\Upsilon(\Phi)$ to zero as follows

$$\eta^* = \left(\frac{\sigma_o^2+\sum_{j=1}^{M}\left(\sqrt{p_j}|h_j|\right)^2}{\sum_{j=1}^{M}\sqrt{p_j}|h_j|}\right)^2. \tag{31}$$

Secondly, with an optimal η^*, the power control problem can be split into M same sub-problems that formulated as

$$\mathcal{P}_{22}: \min_{p_j} \quad \rho\left(\frac{\sqrt{p_j}|h_j|}{\sqrt{\eta^*}}-1\right)^2+\mathcal{Q}_j p_j, \tag{32a}$$

$$\text{s.t.} \quad p_j \leq P_j^{\max}, \tag{32b}$$

where $\mathcal{Q}_j$ is the instantaneous virtual queue value which is given due to the last time slot transmit power $\mathcal{Q}_j(t-1)$. Problem $\mathcal{P}_{22}$ is convex and satisfies the Slater's condition, to this end we can use the KKT conditions to obtain the optimal solution to problem $\mathcal{P}_{22}$ as follows:

$$p_j^* = \min\left\{1\Big/\left[\left(\frac{|h_j|^2}{\eta^*}+\frac{\mathcal{Q}_j}{\rho}\right)\frac{\sqrt{\eta^*}}{|h_j|}\right]^2, P_j^{\max}\right\}. \tag{33}$$

So far, problem $\mathcal{P}_2$ can be tackled by alternately solving problem $\mathcal{P}_{21}$ and $\mathcal{P}_{22}$. Compared with the alternating method used in [14], the proposed Lyapunov drift-based alternating algorithm does not rely on the CSI of all communication rounds, which is more practical. Moreover, by discarding the constraint (13), the proposed algorithm avoids calculating the Lagrangian multiplier using a one-dimensional bisection search as in [14], which saves a lot of computation cost. The complete algorithm that incorporates finding an optimal K^* and calculating p_j^* and η^* is summarized in Algorithm 1.

Algorithm 1: Lyapunov drift-based alternating method

Input: $h_{i,j}^t$, $\mathcal{Q}_{i,j}(t)$, $\forall\, i,j$, G, stopping condition ε.
Output: The optimal edge interval K^*, transmit power $(p_{i,j}^t)^*$, edge normalizing coefficient $(\eta_i^t)^*$, next time slot's virtual queue $\mathcal{Q}_{i,j}(t+1)$, $\forall\, i.j$.

1: **Initialize** $p_{i,j}^0$, η_i^0, $\forall\, i,j$ and $s=0$.
2: With given $p_{i,j}^{t-1}$ and η_i^{t-1}, $\forall\, i,j$, compute K^* by solving (27a).
3: **repeat**
4: $\quad s = s+1$
5: $\quad$ Compute $(\eta_i^{t,s})^*$ according to (31) with given $p_{i,j}^{t,s-1}$ and K^*, $\forall\, i,j$;
6: $\quad$ Solve problem (32a) and obtain the optimal transmit power $(p_{i,j}^{t,s})^*$ via (33);
7: **until** $\frac{\text{MSE}^{s-1}-\text{MSE}^s}{\text{MSE}^s} \leq \varepsilon$;
8: Compute $\mathcal{Q}_{i,j}(t+1)$ via (21).

V. Simulation Results

A. Simulation Setting

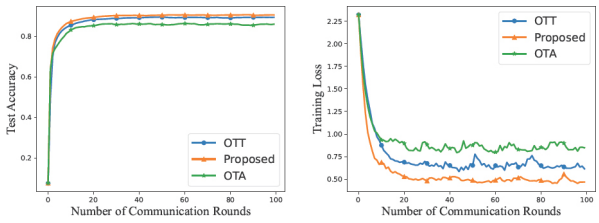
In our experiment, we consider a scene consisting of a cloud server located at $(0,0,100)$ meters, and two edge servers (i.e., $N=2$) located at $(20,0,20)$ and $(-20,0,20)$ meters, respectively. Each edge server manages 10 devices (i.e., $M=10$), which are uniformly distributed within the circles centered at $(140,20,0)$ and $(-140,20,0)$ meters, respectively. Both circles have a radius of 20 meters. The MNIST dataset is used for experiments, and we evenly split the dataset so that each device only contains data with 2 different labels out of 10 in order to construct a data heterogeneous scenario. The large-scale fading is modeled as $T_0(\frac{d}{d_0})^{-\alpha}$, where d represents the distance from each device to its corresponding edge server, α is the path loss exponent, and T_0 means the path loss when $u_0=1$ meter. Besides, Rician fading with Rician factor γ is used to model the small-scale fading. We set $\alpha=3$, $T_0=-30\,\text{dB}$, $P_{i,j}^{\max}=30\,\text{dBm}$, $P_{i,j}^{\text{tot}}=100\,\text{dBm}$, $\gamma=3$, and $\sigma_0^2=-100\,\text{dBm}$. As for the algorithm, we set the Lyapunov drift-based alternating algorithm's convergence accuracy factor to $\varepsilon=10^{-4}$, and the parameter $\rho=0.5$. A deep neural network (DNN) with two hidden layers is used and we set $3/4$ of each device's dataset as the training set and the others are for testing. Besides, we set learning rate $\eta=0.01$, local iterations $G=60$, global communication round $T=100$, and mini-batch size $\left|\tilde{\mathcal{B}}_{i,j}\right|$ is 20.

To illustrate the performance of our proposed edge interval control algorithm, we consider the following two baselines:

- **One Time Training (OTT)**: To reduce the detrimental effect caused by data heterogeneity and make the convergence analysis easier, each edge device only performs one-time local training before uploading the local model to edge servers.
- **One Time Aggregation (OTA)**: To save the communication cost between edge servers and local devices, each edge server only aggregates the local models once before transmitting the aggregated results to the cloud server.

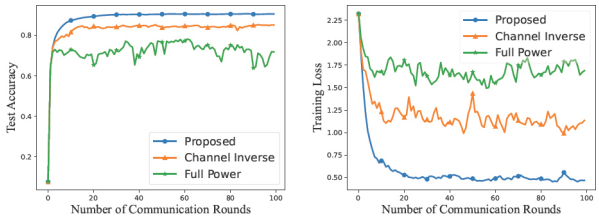
To verify our proposed transceiver design, we consider the following two baselines that do not require full CSI

- **Full Power**: Each edge device takes the maximum power $P_{i,j}^{\text{tot}}K^*/G$ it can use to transmit the signals.
- **Channel Inversion** [13]: With a given truncated coefficient ϱ, the transmit power of device $d_{i,j}$ can be defined



(a) Test Accuracy (b) Training Loss

Fig. 3. Learning performance versus number of communication rounds under different edge interval control schemes.



(a) Test Accuracy (b) Training Loss

Fig. 4. Learning performance versus number of communication rounds under different transceiver design schemes.

as follows:

$$p_{i,j}^t = \begin{cases} \min\left(\bar{P}_{ij}, \dfrac{\eta_i^t}{|h_{i,j}^t|^2}\right), & |h_{i,j}^t|^2 \geq \varrho, \\ 0. & |h_{i,j}^t|^2 < \varrho, \end{cases} \tag{34}$$

whereby $\bar{P}_{ij}$ is denoted as $\frac{P_{i,j}^{\text{tot}} K^*}{G}$, $\eta_i^t = \min_j \left\{ \frac{\sigma_0^2 + \bar{P}_{ij}|h_{i,j}^t|^2}{\sqrt{\bar{P}_{ij}}|h_{i,j}^t|} \right\}$, and ϱ is set to 0.01.

B. Performance Comparison

In Fig. 3, we can see that the learning performance of AirComp-assisted HFL using the edge interval calculated by our proposed algorithm outperforms the baseline schemes. As for OTT scheme, it only considers the data heterogeneity problem while neglecting the huge communication cost it takes, especially with a power budget constraint. As for OTA scheme, although it can utilize all transmit power in one-time aggregation, it faces a great challenge in tackling the performance degradation due to the non-i.i.d. data.

Fig. 4 compares the proposed Lyapunov drift-based transceiver design scheme with other two baselines. All three schemes are able to perform the online transceiver design without full CSI. However, according to the result, we can see that our proposed Lyapunov drift-based transceiver design algorithm achieves a better performance than the two baseline schemes. Since the full power scheme ignores the transceiver design adaptive to the instantaneous channel state, and the truncated channel inverse scheme only sets a truncated threshold but still lacks the specific analysis with respect to the bad channel. On the contrary, our proposed Lyapunov drift-based algorithm not only considers all the channel states and adaptively completes the design of the transceiver but also works with low complexity.

VI. Conclusion

In this paper, we considered an AirComp-assisted HFL framework with a total transmit power budget in each global communication round. We analyzed the convergence and formulated a combinatorial optimization problem to minimize the convergence bound. To this end, a Lyapunov drift-based alternating algorithm is developed to alternately control the edge interval and transceiver design, which achieves better performance compared with different baseline edge interval control schemes and transceiver design schemes.

References

[1] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6g: Vision, principles, and technologies," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 78–85, 2023.

[2] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering*, vol. 149, p. 106854, 2020.

[3] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Hierarchical federated learning with quantization: Convergence analysis and system design," *IEEE Transactions on Wireless Communications*, 2022.

[4] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, "Hfel: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6535–6548, 2020.

[5] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Hierarchical over-the-air federated edge learning," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, May 2022.

[6] F. Zhou, Z. Wang, X. Luo, and Y. Zhou, "Over-the-air computation assisted hierarchical personalized federated learning," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, May 2023.

[7] J. Feng, L. Liu, Q. Pei, and K. Li, "Min-max cost optimization for efficient hierarchical federated learning in wireless edge networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 11, pp. 2687–2700, 2021.

[8] W. Guo, C. Huang, X. Qin, L. Yang, and W. Zhang, "Dynamic clustering and power control for two-tier wireless federated learning," 2022. [Online]. Available: https://arxiv.org/abs/2205.09316

[9] Z. Wang, Y. Zhou, Y. Shi, and W. Zhuang, "Interference management for over-the-air federated learning in multi-cell wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 8, pp. 2361–2377, 2022.

[10] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Local averaging helps: Hierarchical federated learning and convergence analysis," 2020. [Online]. Available: https://arxiv.org/abs/2010.12998

[11] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[12] B. Xu, W. Xia, W. Wen, P. Liu, H. Zhao, and H. Zhu, "Adaptive hierarchical federated learning over wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2070–2083, 2021.

[13] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2019.

[14] Y. Zou, Z. Wang, X. Chen, H. Zhou, and Y. Zhou, "Knowledge-guided learning for transceiver design in over-the-air federated learning," *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 270–285, 2022.

[15] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Letaief, "Federated learning via intelligent reflecting surface," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 808–822, 2021.

[16] Z. Wang, Y. Zhou, Y. Zou, Q. An, Y. Shi, and M. Bennis, "A graph neural network learning approach to optimize ris-assisted federated learning," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[17] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks," *IEEE transactions on Multimedia*, vol. 18, no. 5, pp. 879–892, 2016.

[18] T. Ouyang, Z. Zhou, and X. Chen, "Follow me at the edge: Mobility-aware dynamic service placement for mobile edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 10, pp. 2333–2345, 2018.

[19] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient radio resource allocation for federated edge learning," in *Proc. IEEE Int. Conf. on Commun. Workshops (ICC Wkshps)*, June 2020.