Fast-Convergent Federated Learning With Adaptive Weighting

Hongda Wu, *Student Member, IEEE*, and Ping Wang, *Senior Member, IEEE*

***Abstract*—Federated learning (FL) enables resource-constrained edge nodes to collaboratively learn a global model under the orchestration of a central server while keeping privacy-sensitive data locally. The non-independent-and-identically-distributed (non-IID) data samples across participating nodes slow model training and impose additional communication rounds for FL to converge. In this paper, we propose `Federated Adaptive Weighting (FedAdp)` algorithm that aims to accelerate model convergence under the presence of nodes with non-IID dataset. We observe the implicit connection between the node contribution to the global model aggregation and data distribution on the local node through theoretical and empirical analysis. We then propose to assign different weights for updating the global model based on node contribution adaptively through each training round. The contribution of participating nodes is first measured by the angle between the local gradient vector and the global gradient vector, and then, weight is quantified by a designed non-linear mapping function subsequently. The simple yet effective strategy can reinforce positive (suppress negative) node contribution dynamically, resulting in communication round reduction drastically. Its superiority over the commonly adopted Federated Averaging (`FedAvg`) is verified both theoretically and experimentally. With extensive experiments performed in Pytorch and PySyft, we show that FL training with `FedAdp` can reduce the number of communication rounds by up to 54.1% on MNIST dataset and up to 45.4% on FashionMNIST dataset, as compared to `FedAvg` algorithm.**

***Index Terms*—Federated learning, communication efficiency, mobile edge computing, Internet of Things.**

I. Introduction

THE RAPID advancement of edge devices (e.g., Internet of Things (IoT), mobile phones) is constantly generating an unprecedented amount of data [1]. These devices are currently equipped with enhanced sensors, computing, and communication capability. Coupled with the rise of Deep Learning (DL) [2], the edge devices unfold the countless opportunities for various tasks of modern society, e.g., road congestion prediction [3] and environmental monitoring [4].

In the traditional cloud-centric approaches, data generated and collected by edge devices is uploaded and processed in a data center. It is predicted that the data generation rate will exceed the capacity of today's Internet in the near future [5], Mobile Edge Computing (MEC) has naturally been proposed to incorporate the data processing outside the cloud [6], [7]. With computing and storage capability, MEC systems generally consist of end-edge-server architecture. Multiple edge servers are capable of performing large-scale distributed tasks involving local processing and remote execution under the coordination of a remote cloud. MEC approaches compromise training efficiency and communication cost by bringing model training towards where the data is generated. However, computation offloading task and data processing at the edge server still involves the transmission of sensitive data.

In either centralized cloud training or MEC approaches, collecting data for model training is unrealistic from a privacy, security, regulatory, or necessity perspective. In order to maintain privacy-sensitive data and to facilitate collaborative machine learning (ML) among distributed nodes, Federated Learning (FL) has emerged as an attractive paradigm, where local nodes collaboratively train a task model under the orchestration of a central server without accessing end-user data [8], [9]. In FL, local nodes cooperatively train an ML model required by the central server by utilizing their local data. Through transferring local model updates to the central server for model aggregation and acquiring a global model for local training rather than sending raw data, user data privacy is well protected. As such, FL features from conventional approaches in data acquisition, storage, and training. FL has been deployed by major service providers and plays an important role in supporting privacy-sensitive applications, including computer vision, natural language processing, and medical database [10].

Even though good convergence performance of FL approach is shown, owing to limited connectivity of wireless networks, the availability of local nodes and straggler of participating nodes, communication cost becomes a critical bottleneck in FL context since generally several iterations are involved for model converging [8]–[10]. Another fundamental challenge for FL is strongly non-independent-and-identically-distributed (non-IID) and highly skewed data across local nodes. The presence of non-IID data significantly degrades the performance of federated learning, which makes model training take more rounds to converge, and the variance caused by non-IID data brings instability to the training process [11]–[13]. Since the completion time of federated learning is largely impacted by the communication time, how to reduce the communication round for model convergence in FL, especially for participating nodes with non-IID datasets, is urgent to be addressed.

Manuscript received November 30, 2020; revised March 20, 2021; accepted May 20, 2021. Date of publication May 27, 2021; date of current version December 9, 2021. This work was supported by the Canada NSERC Discovery under Grant RGPIN-2019-06375. The associate editor coordinating the review of this article and approving it for publication was M. Pan. *(Corresponding author: Ping Wang.)*

The authors are with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, ON M3J 1P3, Canada (e-mail: hwu1226@eecs.yorku.ca; pingw@yorku.ca).

Digital Object Identifier 10.1109/TCCN.2021.3084406

In this paper, to surmount the slow convergence of vanilla Federated Averaging (FedAvg) [8] under the presence of non-IID dataset, we propose Federated Adaptive Weighting (FedAdp) algorithm that aims to improve the performance of federated learning through assigning distinct weight for participating node to update the global model. We observe that nodes with heterogeneous datasets make different contributions to the global model aggregation. Therefore, our main intuition is to measure the contribution of the participating node based on the gradient information from local nodes then assign different weights accordingly and adaptively at each communication round for global model aggregation. According to node contribution, the proposed adaptive weighting strategy is capable of reducing the expected training loss of FL in each communication round under the presence of non-IID nodes, which accelerates the model convergence. Our main contributions in this paper are as follows:

- We identify the presence of nodes with non-independent-and-identically-distributed (non-IID) data distributions slows the convergence speed of federated learning. In addition, we analyze the convergence bound of gradient-descent based federated learning from a theoretical perspective and derive the convergence bound that incorporates the non-IID data distribution across participating nodes and weighting strategy for model updating.
- We observe the implicit connection between data distribution on a node and the contribution from that node to the global model aggregation, measured at the central server-side by inferring gradient information of participating nodes. The convergence bound is lowered, and the convergence speed is accelerated by a carefully designed weighting strategy, which is formalized as Federated Adaptive Weighting (FedAdp), that assigns different weights to nodes for global model aggregation in each round of communication.
- We empirically evaluate the performance of the proposed weighting algorithm via extensive experiments using different real datasets with different learning objectives (i.e., convex and non-convex loss function). Our experimental results have shown that FL training with FedAdp can drastically reduce the communication rounds compared with the commonly adopted FedAvg algorithm.

The rest of this paper is organized as follows. Section II discusses the related works. Section III provides the preliminaries of federated learning and the impact of non-IID data on FL. In Section IV, the convergence analysis and the proposed weighting algorithm are presented. Experimental results are shown in Section V. Section VI presents the conclusion.

II. Related Work

Generally, the FL algorithm adopts synchronous aggregation and selects a subset of nodes randomly to participate in each round randomly to avoid long-tailed waiting time due to the network uncertainty and straggler. To boost convergence and reduce the communication rounds, tuning the number of local updates [8], [13], [14], [15], and selecting appropriate nodes for FL training [12], [16], [17] are the usually adopted approaches.

In particular, McMahan *et al.* [8] presented the vanilla Federated Averaging (FedAvg) algorithm, which increases the number of local updates instead of updating the local model one time at each round. Li *et al.* [13] proposed to allow participating nodes to perform a variable number of local updates, rather than applying the same amount of workload for each node [8], to consequently overcome the heterogeneity of the system. Similar to [13], authors in [15] also posed local accuracy for participating nodes, based on limited computing resources on nodes, as an index to steer the number of local updates performed. Different from [13], [15], the work in [14] exposed an analytical model to dynamically adapt the number of local updates between two consecutive global aggregations in real-time to minimize the learning loss under a fixed resource budget of the edge computing system. Regarding the node selection, Nishio and Yonetani [16] proposed FedCS algorithm to do node selection intentionally rather than randomly, based on the resource conditions of local nodes. Authors in [17] utilized gradient information to do node selection. The node whose inner product between its gradient vector and the global gradient vector is negative will be excluded from FL training.

To handle the non-IID data distribution, Zhao *et al.* [11] quantified the weight divergence by earth mover's distance between data distribution on nodes and population distribution. However, the strategy of pushing a small set of uniformly distributed data to participating nodes in [11] violates the privacy concern of FL and imposes extra communication cost. It was proposed in [12] that communication rounds can be reduced effectively by selecting nodes based on their uploaded model weights, which profile the data distribution on those nodes. In contrast, Wang *et al.* [18] proposed to identify the irrelevant update caused by different data distribution at the node side. The communication cost is accordingly reduced by precluding these nodes with irrelevant updates before updates transmission. However, local nodes are required to check the relevance in each round using the global model kept in the previous round, which is in contravention of FL and brings computational burdens to local nodes.

Regarding the weighting strategy, authors in [19] proposed to assign different weights for global model aggregation adaptively by considering the time difference when the model update is done in a layerwise asynchronous manner. Chai *et al.* [20] designed a tier-based FL system by dividing the participating nodes into tiers according to their responding time and devised to adaptively assign weights to different tiers for model aggregation since there exists different updating frequency across tiers. Both methods in [19], [20] aim to weigh the local update along with different communication rounds.

To enhance the convergence of FL with the presence of non-IID nodes, different from [11], [12] that measure model weight, we find out that nodes contribute differently to the global model aggregation owing to their different data distribution, and there exists an implicit connection between data distribution and gradient information. In this paper, we propose to measure the node contribution quantitatively by the angle between the local gradient of each participating node and the global gradient across all participating nodes at the server-side.

With the quantified contribution, the weight for aggregating the global model can be devised discriminatively across the nodes and adaptively in each round according to node contribution. The proposed adaptive weighting strategy can effectively speed up the convergence of FL in the presence of non-IID data. Different from [17], [18], our method does not impose additional communication and computation burden to local nodes. Besides, our adaptive weighting strategy is done in each communication round, which is orthogonal with the methods proposed in [19], [20].

III. Preliminaries

In this section, we briefly introduce key ingredients behind the recent method for federated learning, FedAvg, and show how non-IID data impacts model convergence.

A. Standard Federated Learning

In general, federated learning methods [8], [10] are designed to handle the consensus learning task in a decentralized manner, where a central server coordinates the global learning objective and multiple devices training the local model with locally collected data. In particular, assume that we have *N* local nodes with dataset $\mathcal{D}_1, \ldots, \mathcal{D}_i, \ldots, \mathcal{D}_N$ and we define $D_i \triangleq |\mathcal{D}_i|$ as the number of data samples owned by each node, where $|\cdot|$ denotes the Cardinality of sets. FL methods aim to minimize:

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) \triangleq \sum_{i=1}^{N} \psi_i F_i(\mathbf{w}), \tag{1}$$

where **w** is global model weight, $\psi_i = D_i / \sum_{i'=1}^{N} D_{i'}$ is the weight for aggregation in FL training, and global objective function *F*(**w**) is surrogated by using local objective function $F_i(\mathbf{w})$, which is defined, as an example, in the context of *C*-class classification problem thereinafter. In particular, *C*-class classification problem is defined over a feature space $\mathcal{X}$ and a label space $\mathcal{Y} = [C]$, where $[C] = \{1, \ldots, C\}$. For each labeled data sample {**x**, *y*}, predicted probability vector $\widetilde{\mathbf{y}}$ is achieved by using mapping function $f : \mathcal{X} \rightarrow \widetilde{\mathcal{Y}}$, where $\widetilde{\mathcal{Y}} = \{\widetilde{\mathbf{y}} | \sum_{j=1}^{C} \widetilde{y}_j = 1, \widetilde{y}_j \geq 0, \forall j \in [C]\}$. As such, $F_i(\mathbf{w})$ commonly measures the local empirical risk over possibly different data distribution $p^{(i)}$ of node *i*, which is defined by using cross entropy for *C*-class classification as follow,

$$\begin{aligned}\min_{\mathbf{w}} F_i(\mathbf{w}) &\triangleq \mathbb{E}_{\mathbf{x}, y \backsim p^{(i)}} \left[-\sum_{j=1}^{C} \mathbb{1}_{y=j} \log f_j(\mathbf{x}, \mathbf{w}) \right] \\ &= -\sum_{j=1}^{C} p^{(i)}(y=j) \mathbb{E}_{\mathbf{x}|y=j} \left[\log f_j(\mathbf{x}, \mathbf{w}) \right], \end{aligned} \tag{2}$$

where $f_j(\mathbf{x}, \mathbf{w})$ denotes the probability that the data sample **x** is classified as the *j*-th class given model **w**, and $p^{(i)}(y=j)$ denotes the data distribution on node *i* over class $j \in [C]$.

In general federated learning setting (e.g., FedAvg), the participating nodes perform local training with the same training configuration (e.g., optimizer, learning rate, etc). At each communication round *t*, a subset of the nodes $\mathcal{S}_t, |\mathcal{S}_t| = K \ll N$ are selected and global model $\mathbf{w}(t-1)$ in previous iteration is sent to the selected nodes. Each of the participating nodes *i* performs stochastic gradient descent (SGD) training to optimize its local objective $F_i(\mathbf{w})$:

$$\mathbf{w}_i(t) = \mathbf{w}(t-1) - \eta \nabla F_i(\mathbf{w}(t-1)), \tag{3}$$

where η is the learning rate and $\nabla F_i(\cdot)$ is the gradient at node *i*. (3) gives a general principle of SGD optimization. $\mathbf{w}_i(t)$ could be the result after one or several local updates of SGD (e.g., $\tau = 1$ in FedSGD [8] or $\tau > 1$ in FedAvg [8], [14] with τ denoting the number of local updates between two consecutive global rounds). Hereinafter, SGD is applied to mini-batch data samples with size $\bar{B}$. As such, local model is updated by $\tau = \frac{D_i}{B} E$ times, where D_i and E are the number of training samples on node *i* and the number of local training epochs, respectively.

The nodes then communicate their local model updates $\Delta_i(t) = \mathbf{w}_i(t) - \mathbf{w}(t-1)$ to the central server,[1] which aggregates them and updates the global model accordingly,

$$\begin{aligned}\Delta(t) &= \sum_{i=1}^{|\mathcal{S}_t|} \psi_i \Delta_i(t) \\ \mathbf{w}(t) &= \mathbf{w}(t-1) + \Delta(t). \end{aligned} \tag{4}$$

B. FedAvg for Non-IID Data

The independent and identically distributed (IID) sampling condition of training data is important that the stochastic gradient is an unbiased estimate of the full gradient [14]. FedAvg is shown to be effective, given that the data distribution across different nodes is the same as centrally collected data. However, the data distribution determined by usage patterns across local nodes is typically non-IID, i.e., $p^{(i)}$ is different across participating nodes.

Since local objective $F_i(\mathbf{w})$ is closely related with data distribution $p^{(i)}$, a large number of local updates lead the model towards optima of its local objective $F_i(\mathbf{w})$ as opposed to the global objective *F*(**w**). The inconsistency between local models $\mathbf{w}_i$ and global model **w** is accumulated along with local training, leading to more communication rounds before training converges. As such, local training with multiple local updates potentially hurts convergence and even leads to divergence with the presence of non-IID data [8], [11].

We conduct an experiment to demonstrate the impact of non-IID data on model convergence. We train a two-layer CNN model with the same neural network architecture in [8] using Pytorch on the MNIST dataset (containing 60,000 samples with 10 classes) until the model achieves 95% test accuracy. 10 nodes are selected, each with 600 samples that are selected based on their label criteria. If a node is at *IID setting*, 600 samples are randomly selected over the whole training set. If a node is at *x-class non-IID setting*, 600 samples are randomly selected over a subset, which is composed of *x* class data samples. Each class of the *x*-class is selected at random

[1] Typically there are two ways for nodes to upload their local model to the server, either by uploading model parameters **w**(*t*) or by uploading the model difference $\Delta_i(t)$. Although the same amount of data are to be sent in both ways, conveying $\Delta_i(t)$ is proven to be more amenable for compression [9].

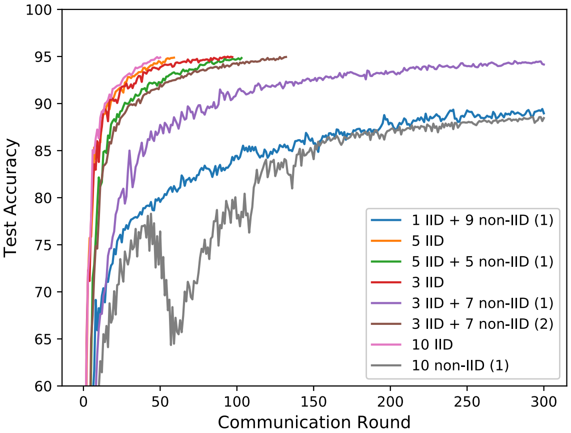

Fig. 1. Test accuracy over communication rounds of `FedAvg` with heterogeneous data distribution over participating nodes. X IID + Y non-IID (1) (or (2)) represents X nodes are at *IID setting* and Y nodes are at *1-class (or 2-class) non-IID setting.*

and can be overlapped. The skewness of datasets is measured and reflected by the value of x.

We use the same notations for `FedAvg` algorithm as [8]: $\bar{B}$, the local minibatch size, and E, the number of local training epochs. In this experiment, $\bar{B} = 32$, $E = 1$, $\eta = 0.01$ and learning rate decay of 0.995 per communication round. We can conclude from Fig. 1:

- Model convergence highly depends on IID nodes. The presence of non-IID nodes imposes variance to model training, which slows the convergence of FL (e.g., 5 IID case converges faster than 5 IID + 5 non-IID (1) case).
- The skewness of data affects model convergence. With the participation of the non-IID node, the model converges much slower when the skewness of the dataset increases (e.g., 3 IID + 7 non-IID (2) case converges much faster than 3 IID + 7 non-IID (1) case).

IV. Federated Adaptive Weighting

In this section, we develop our methodology for improving the convergence rate of federated learning. We first analyze the convergence property of federated learning (Section IV-A). The theoretical analysis on the expected decrease of FL loss in each round of training reveals that gradient information and data distribution impact the convergence. The experimental result shows the diversity of node contribution in reducing the FL loss in each round (Section IV-B), measured by the local gradient of each node and the global gradient from participating nodes. This motivates us to assign weight adaptively according to node contribution for global model aggregation. Finally, we theoretically prove that assigning weight based on node contribution adaptively leads to accelerating model convergence and formally present the methodology of the proposed `FedAdp` algorithm (Section IV-C).

A. Convergence Analysis

For theoretical analysis of federated learning algorithms, we employ the following typical assumptions in our analysis (see, e.g., [11], [13], [14], [17]).

Assumption 1 (β-Lipschitz Smoothness): $F_i(\mathbf{w})$ is β-Lipschitz smoothness for each of the participating nodes $i \in \mathcal{S}_t$, i.e., $\|\nabla F_i(\mathbf{w}) - \nabla F_i(\mathbf{w}')\| \leq \beta\|\mathbf{w} - \mathbf{w}'\|$ for any two parameter vectors $\mathbf{w}$, $\mathbf{w}'$.

Based on Assumption 1, the definition of $F(\mathbf{w})$, and triangle inequality, we can easily get the following lemma.

Lemma 1: $F(\mathbf{w})$ is β-Lipschitz smoothness.

Assumption 2 (Bounded Local Dissimilarity):[2] For any participating node i, the dissimilarity between local objective and global objective at $\mathbf{w}$ is bounded by A and B, i.e., $A\|\nabla F(\mathbf{w})\| \leq \|\nabla F_i(\mathbf{w})\| \leq B\|\nabla F(\mathbf{w})\|$.

Here $\nabla F(\mathbf{w})$ is the gradient of the global objective that is defined as $\nabla F(\mathbf{w}) = \sum_{i=1}^{|\mathcal{S}|}(D_i/\sum_{i'=1}^{|\mathcal{S}|} D_{i'})\nabla F_i(\mathbf{w})$ in FL context. The local dissimilarity in assumption 2 can be seen as a metric that reveals the data heterogeneity when the same training configuration (e.g., learning rate, batch size, training epoch, etc.) across participating nodes is held. As a sanity check, when all the local data samples are the same, we have $A = B = 1$.

Theorem 1: With loss function $F_i(\mathbf{w})$ satisfying Assumptions 1-2 and supposing $\mathbf{w}(t)$ is not a stationary solution, the expected decrease in the global loss function between two consecutive rounds satisfies,

$$F(\mathbf{w}(t+1)) \leq F(\mathbf{w}(t)) - \eta\mathbb{E}_{i|t}\left[\left(\frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|} - \frac{B\beta\eta}{2}\right) \cdot\frac{A^2}{B}\|\nabla F(\mathbf{w}(t))\|^2\right], \tag{5}$$

where the expectation $\mathbb{E}_{i|t}$ refers to the weighting strategy of the participating node $i \in \mathcal{S}_t$ for global model aggregation. $\langle\cdot\rangle$ is the inner product operation and $\|\cdot\|$ denotes the $\ell 2$ norm of a vector.

The proof of Theorem 1 is presented in Appendix-A. Theorem 1 provides a bound on how rapid the decrease of the global FL loss can be expected. Based on Theorem 1, we have the following corollary and remarks.

Corollary 1: The convergence upper bound of FL after T global rounds is given by,

$$F(\mathbf{w}(T)) \leq F(\mathbf{w}(0)) - \eta\sum_{t=0}^{T-1}\mathbb{E}_{i|t}\left[\left(\frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|} - \frac{B\beta\eta}{2}\right) \cdot\frac{A^2}{B}\|\nabla F(\mathbf{w}(t))\|^2\right]. \tag{6}$$

Remark 1: The decrease of FL loss between two consecutive global rounds shows a dependency on learning rate η, the bounded local dissimilarity of participating nodes, the correlation between the local gradient and the global gradient $\frac{\langle\nabla F(\mathbf{w}(t)),\nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}$, and the weight strategy $\mathbb{E}_{i|t}$ that weighs participating nodes for the global model aggregation in each global round.

[2] Similar assumption has made in FL context, for example in [13], [14], [17]. In [13], [17], the dissimilarity across local gradients is imposed by an upper bound to capture the impact of data heterogeneity on FL convergence, and an analogous definition named gradient divergence is also presented in [14]. By tracking the divergence of gradients on each participating node, we observe that the dissimilarity can be further bounded by a lower bound as shown in Assumption 2.

Remark 2: The local gradient, which is correlated with minimizing the local objective, may not align with the direction of approaching the optimal of the global objective. The correlation $\frac{\langle \nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\| \|\nabla F_i(\mathbf{w}(t))\|}$ between the local gradient and the global gradient is a metric to measure their alignment level. From Theorem 1, we can see this metric also indicates how much each node contributes to reducing FL loss in each round.

Remark 3: The FL loss $F(\mathbf{w}(t+1))$ is negatively associated with the bound gap in Assumption 2, meaning that as bound gap $[A, B]$ grows larger, the bound weakens, and the convergence exacerbates. Intuitively, the root cause of dissimilarity is the divergence of local gradients across participating nodes with heterogeneous datasets, which can be intentionally regularized by a properly designed weighting strategy.

An immediate suggestion from Theorem 1 is that to improve the convergence of FL, one can reduce the FL loss by increasing $\mathbb{E}_{i|t}[\cdot]$ in each global round. This motivates us to measure node contribution quantitatively through the correlation $\frac{\langle \nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\| \|\nabla F_i(\mathbf{w}(t))\|}$ between the local gradient and the global gradient and assign larger weights to the nodes with higher contribution to enlarge the expected decrease of FL loss in each global round.

B. Measurement of Node Contribution

In FL, the direction of minimizing local objective $F_i(\mathbf{w})$ might not align with the direction of minimizing $F(\mathbf{w})$. In particular, it can be deduced from (3) that the gradient on different nodes may be tremendously diverse, especially for participating nodes with heterogeneous datasets. As such, the contribution from participating nodes for global aggregation is different. Empirically, we note that if the data distribution on a node is highly skewed, the gradient may highly deviate from or even in the opposite direction to the global gradient, causing a negative effect on the global aggregation.

Instead of assigning weight for participating nodes based on the size of datasets as in FedAvg [8], we measure the contribution of participating nodes based on the correlation between local gradient and global gradient. Particularly, we quantify the contribution of each node at each global round based on *angle* $\theta_i(t)$, that is defined as:

$$\theta_i(t) = \arccos \frac{\langle \nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\| \|\nabla F_i(\mathbf{w}(t))\|}. \quad (7)$$

From (7), we can see that when the angle $\theta_i(t)$ is small, it means the local gradient $\nabla F_i(\mathbf{w}(t))$ has a similar direction to the global gradient, thereby positively contributing to the global aggregation. In contrast, when $\theta_i(t)$ is large, e.g., larger than $\pi/2$, the local gradient $\nabla F_i(\mathbf{w}(t))$ has an opposite direction to the global gradient, thereby negatively contributing to the global aggregation.

To restrain the instability caused by randomness presented in instantaneous angle $\theta_i(t)$ at each round, we use so-called *smoothed angle* $\widetilde{\theta}_i(t)$ as a substitution, which is the averaged angle over previous training rounds and is defined as:

$$\widetilde{\theta}_i(t) = \begin{cases} \theta_i(t) & t = 1 \\ \frac{t-1}{t}\widetilde{\theta}_i(t-1) + \frac{1}{t}\theta_i(t) & t > 1. \end{cases} \quad (8)$$

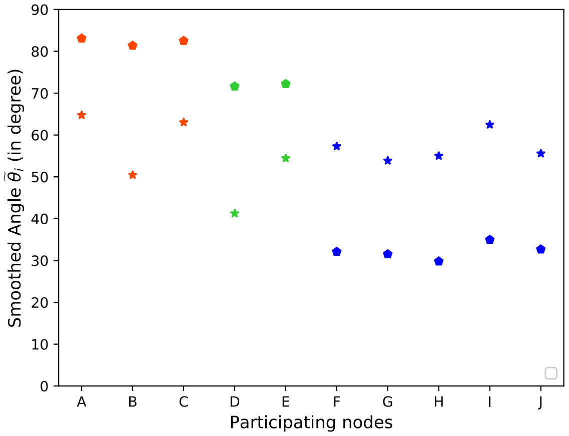

Fig. 2. The smoothed angle $\widetilde{\theta}_i$ of participating node at different training round, where star and pentagon sign denote the angle at communication round 1 and communication round 15, respectively. Nodes with different data distribution are marked with different colors.

By using *smoothed angle* $\widetilde{\theta}_i(t)$, the angle difference across nodes uniquely depends on the data distribution. Intuitively, the angle $\widetilde{\theta}_i(t)$ will be larger as the dissimilarity between data distribution on node i and population distribution grows. Also, the smoothed angle is capable of quantifying the degree of data dissimilarity among the local nodes.

We conduct an experiment to illustrate how data distribution can be reflected by angle. Under the same training model in Section III-B, we randomly assign i) 3 nodes with *1-class non-IID setting* (i.e., node "A", "B", "C"), ii) 2 nodes with *2-class non-IID setting* (i.e., nodes "D" and "E"), and iii) the rest of 5 nodes with *IID setting*.

As shown in Fig. 2, the smoothed angle between the local gradient and the global gradient is full of randomness at the beginning of FL training. Along with the training, smoothed angle $\widetilde{\theta}_i$ shows diversity across the participating nodes due to the impact of data heterogeneity on local training. To be more specific, for those nodes with 1-class non-IID setting, the data samples from which are highly skewed since the label space $\mathcal{Y}$ is extremely limited. Due to the limited richness of data samples on node i, the direction for minimizing its local objective $F_i(\mathbf{w})$, which is reflected by $\nabla F_i(\mathbf{w})$, will be far away from the direction for minimizing the overall objective $F(\mathbf{w})$, which is reflected by $\nabla F(\mathbf{w}) = \sum_{i=1}^{|\mathcal{S}|}(D_i/\sum_{i'=1}^{|\mathcal{S}|} D_{i'})\nabla F_i(\mathbf{w})$, resulting a greater θ_i as defined by (7). As shown in Fig. 2, the gradient from the node with extremely skewed data (e.g., node "A", "B", "C") is nearly orthogonal with the global gradient after 15 communication rounds, which barely brings a contribution to the global model. If we ignore the discrepancy of node contribution and average local update according to the size of datasets, as in FedAvg, it slows model convergence.

C. Federated Adaptive Weighting (FedAdp)

Provided the diverse node contribution from participating nodes, the weighting strategy affects Theorem 1 through the expectation $\mathbb{E}_{i|t}[\frac{\langle \nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\| \|\nabla F_i(\mathbf{w}(t))\|}]$ consequently. To accelerate the convergence rate, we seek to lower the upper bound of the expected loss in each communication round, which reveals to assign different weights $\widetilde{\psi}_i$ to different nodes for the global model aggregation. As such, the corresponding objective is formally stated as enlarging $\mathbb{E}_{i|t}[\frac{\langle\nabla F(\mathbf{w}(t)),\nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}] = \sum_i^{|\mathcal{S}_t|}\frac{\langle\nabla F(\mathbf{w}(t)),\nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}\cdot\widetilde{\psi}_i(t)$ via designing $\widetilde{\psi}_i$ under the inherent constrain $\sum_i^{|\mathcal{S}_t|}\widetilde{\psi}_i(t) = 1, \quad \widetilde{\psi}_i(t) \geq 0 \ \forall i, t$.

Considering the node contribution is measured by (7), a natural weighting design aiming to enlarge the expectation should follow the criterion that nodes with higher contribution deserve higher weights for aggregation in each global round. We characterize the contribution-regulated weighting strategy for the global aggregation in each global round adaptively as `Federated Adaptive` Weighting (`FedAdp`).

Assigning adaptive weight for updating the global model in the proposed `FedAdp` algorithm includes two steps:

1) Non-Linear Mapping Function: We design a non-linear mapping function to first quantify the contribution of each node based on angle information. Inspired by the sigmoid function, we use a variant of *Gompertz function* [21], which is a non-linear decreasing function defined as

$$f\left(\widetilde{\theta}_i(t)\right) = \alpha\left(1 - e^{-e^{-\alpha(\widetilde{\theta}_i(t)-1)}}\right), \tag{9}$$

where $\widetilde{\theta}_i(t)$ is the *smoothed angle* in *radian*, e denotes the exponential constant and α is a constant as explained in the following.

The designed mapping function has several properties that are important for the subsequent weight calculation:

- $\lim_{\widetilde{\theta}_i(t)\to\pi/2} f(\widetilde{\theta}_i(t)) = \epsilon$, where $\epsilon \propto \frac{1}{\alpha}$ is constant;
- $\lim_{0\to\widetilde{\theta}_i(t)\to\upsilon} f(\widetilde{\theta}_i(t)) = \alpha$, where $\upsilon \propto \alpha$ is a constant;

α controls the decreasing rate of $f(\widetilde{\theta}_i(t))$ from α to ϵ as $\widetilde{\theta}_i(t)$ increases from υ to $\pi/2$. For example, a small $\alpha \in \mathbb{Z}^+$ indicates a lower decreasing rate of $f(\widetilde{\theta}_i(t))$ that decreases from α to $\epsilon \propto \frac{1}{\alpha}$ as $\widetilde{\theta}_i(t)$ increases from $\upsilon \propto \alpha$ to $\pi/2$. As α increases, the gap between small angle and large angle is amplified (e.g., $f(\widetilde{\theta}_i(t))$ changes within a relatively large range $[\alpha, \epsilon]$ as $\widetilde{\theta}_i(t)$ increases within range $[\alpha, \pi/2]$), so is the difference of contribution from those nodes. However, keeping increasing α is not consistently effective to distinguish the difference of contributions from nodes. Since υ is proportional to α, a large α narrows the boundary $[\upsilon, \frac{\pi}{2}]$ where the node contribution should be considered, making the contribution of nodes whose angle lays between $[0, \upsilon]$ indistinguishable. The choice of α is empirically verified in Section V-B.

2) Weighting: After getting the contribution mapped using the smoothed angle from each node, we use *Softmax function* to finally calculate the weight of participating nodes for global model aggregation as follows:

$$\widetilde{\psi}_i(t) = \begin{cases} \dfrac{e^{f(\widetilde{\theta}_i(t))}}{\sum_{i'=1}^{|\mathcal{S}_t|} e^{f\left(\widetilde{\theta}_{i'}(t)\right)}} & D_m = D_n, \ \forall m, n \in \mathcal{S}_t \\ \dfrac{D_i e^{f(\widetilde{\theta}_i(t))}}{\sum_{i'=1}^{|\mathcal{S}_t|} D_{i'} e^{f\left(\widetilde{\theta}_{i'}(t)\right)}} & D_m \neq D_n, \exists m, n \in \mathcal{S}_t. \end{cases} \tag{10}$$

From the first line of (10), if all the participating nodes have the same size of data samples, the proposed `FedAdp` algorithm will assign weight solely based on their contribution quantified by $e^{f(\widetilde{\theta}_i(t))}$. From the 2nd line of (10), `FedAdp` will assign weight based on both the contribution and the data size.

Algorithm 1 `Federated Adaptive` Weighting (`FedAdp`)

```
procedure FEDERATED OPTIMIZATION
Input: node set $\mathcal{S}, E, B, T, \eta$,
  Server initializes global model $\mathbf{w}(0)$, global update $\Delta(0)$,
    smoothed angle $\widetilde{\theta}_i(0), i \in \mathcal{S}$
  for $t = 1, \ldots, T-1$ do
    for node $i \in \mathcal{S}_t$ in parallel do
      $\Delta_i(t) \leftarrow$ LOCAL UPDATE $(i, \mathbf{w}_i(t-1))$
    $\mathbf{w}(t) \leftarrow$ GLOBAL UPDATE
                    $(\Delta_1(t)\ \Delta_2(t), \cdots, \Delta_{|\mathcal{S}_t|}(t))$
procedure LOCAL UPDATE
Input: node index $i$, model $\mathbf{w}_i(t-1)$
  Calculate local updates for $\tau = D_i\frac{E}{B}$ times of SGD with
    step-size $\eta$ on $F_i(\mathbf{w})$ and obtain $\mathbf{w}_i(t)$ using (3)
  Calculate the model difference $\Delta_i(t) = \mathbf{w}_i(t) - \mathbf{w}(t-1)$
  return $\Delta_i(t)$
procedure GLOBAL UPDATE
Input: local update $\Delta_1(t), \Delta_2(t), \cdots, \Delta_{|\mathcal{S}_t|}(t)$
  Calculate the global gradient
    $\nabla F(\mathbf{w}(t)) = \sum_{i=1}^{|\mathcal{S}_t|}(D_i/\sum_{i'=1}^{|\mathcal{S}_t|} D_{i'})\nabla F_i(\mathbf{w}(t))$, where
    $\nabla F_i(\mathbf{w}(t)) = -\Delta_i(t)/\eta$
  Calculate instantaneous angle $\theta_i(t)$ by (7)
  Update smoothed angle $\widetilde{\theta}_i(t)$ by (8)
  Calculate weight for model aggregation by (9), (10)
  Update global model $\mathbb{E}_{i,t}\left[\widetilde{\psi}_i(t)\mathbf{w}_i(t-1)\right]$
  return $\mathbf{w}(t)$
```

Remark 4: Different from `FedAvg`, where the weight for aggregation is solely proportional to the size of local datasets (e.g., $\psi_i = D_i/\sum_{i'=1}^{|\mathcal{S}_t|} D_{i'}$), `FedAdp` takes both the data size and the node contribution into consideration when assigning weights for model aggregation.

The reason for adopting the Softmax function is twofold: i) The output of the Softmax function is a *normalized value* with a larger angle corresponding to a smaller weight. ii) Using the Softmax function, each node's contribution can be reinforced or suppressed, depending on the smoothed angle between its gradient and the global gradient.

The complete procedures of the proposed `FedAdp` algorithm are presented in Algorithm 1 and `FedAdp` with adaptive weighting strategy leads to the following theorem.

Theorem 2: `FedAdp` with weight design $\widetilde{\psi}_i$ achieves a tighter bound on FL loss decrease in Theorem 1 than `FedAvg` with weight ψ_i.

The proof of Theorem 2 is presented in Appendix-B.

Compared to `FedAvg`, `FedAdp` adopts a simple yet effective strategy that measures the node contribution by quantifying the correlation between the local gradient and the global gradient. Weight for the global model updates can be adaptively assigned based on node contribution rather than evenly averaging, which results in greater FL loss reduction in each

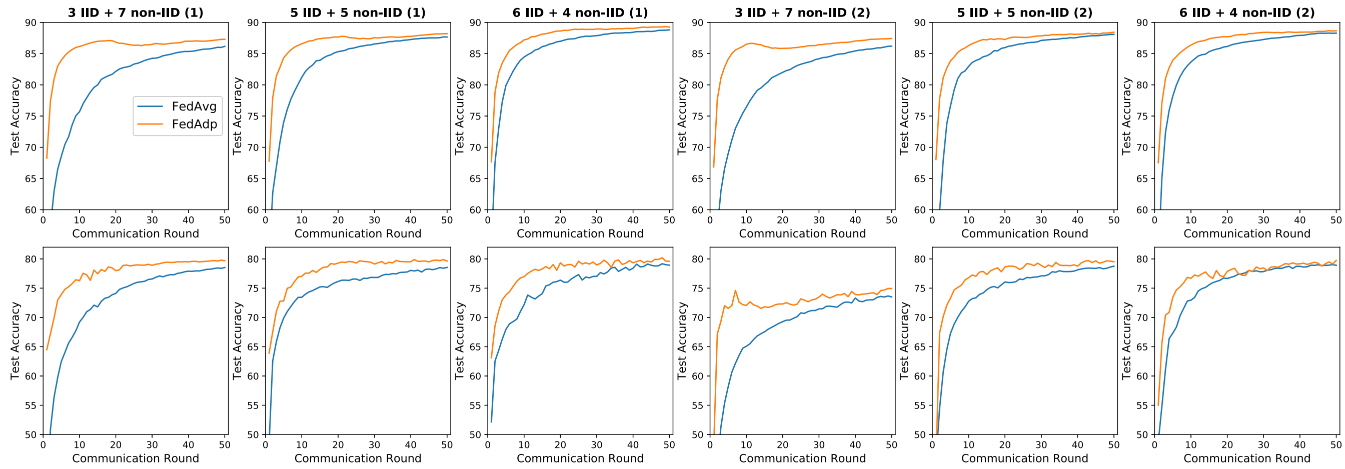

Fig. 3. Test accuracy over communication rounds of `FedAdp` and `FedAvg` with heterogeneous data distribution over participating nodes using MLR model. Upper and lower subplots correspond to training performance on MNIST and FashionMNIST datasets, respectively.

global round and accelerates model convergence consequently, as confirmed by our experimental results.

V. Evaluation and Analysis

To evaluate the performance of our proposed adaptive weighting algorithm, we implemented `FedAdp` with PyTorch framework and PySyft library, and studied the image classification task. We evaluated `FedAdp` by training typical convex and non-convex learning models on two datasets: MNIST and FashionMNIST. Similar to the experiment in Section III-B, when the different degree of skewness of non-IID dataset is presented, we first investigated how `FedAdp` outperforms `FedAvg` [8] by assigning adaptive weight for model aggregation. Note that our proposed algorithm is not limited by the presence of the IID dataset and can be applied to a general scenario with data heterogeneity as verified in Section IV-A. Then, the choice of α for non-linear mapping in `FedAdp` is discussed in Section IV-B. Finally, by tracking the divergence of gradients on participating nodes, we showed `FedAdp` alleviates the impact brought by the data heterogeneity, compared to `FedAvg`, which is beneficial to reducing the FL loss in each round and accelerating FL model convergence as discussed in Section IV-C. We briefly describe our experiment settings as follows.

We consider Multinomial Logistic Regression[3] (MLR) model and CNN model[4] to represent convex and non-convex learning objective, respectively. we use the number of communication rounds for the FL model to reach a target testing accuracy as a performance metric. Unless otherwise specified, the target accuracy is set to 95% for training on MNIST, and 80% for training on FashionMNIST. The number of participating nodes $|\mathcal{S}_t| = 10$, $D_i = 600$, $\bar{B} = 50$ for MLR and $\bar{B} = 32$ for CNN, $E = 1$, $T = 300$, $\eta = 0.01$, decay rate $= 0.995$, the constant in non-linear mapping function $\alpha = 5$. The skewness of the dataset is measured by *x-class non-IID*. The dataset for nodes is generated in the same way as in Section III-B.

A. Data Heterogeneity

We investigate the different number of non-IID nodes with different skewness levels of non-IID data to testify the efficiency of `FedAdp`. For non-IID data, two skewness cases that $x = 1, 2$ are considered. We plot the test accuracy vs. the communication rounds of federated learning in Fig. 3 and Fig. 4 when MLR and CNN models are adopted, respectively.

1) MLR Model: Given the learning capability of MLR is limited, instead of setting a target accuracy, we simply train a model over 50 global rounds. We plot the test accuracy vs. the communication rounds of federated learning algorithms in Fig. 3. From Fig. 3, we can tell `FedAdp` always outperforms `FedAvg` when the nodes with non-IID dataset are present. In addition, `FedAdp` converges very fast in the early training stage, and the superiority of `FedAdp` is more prominent when the proportion of nodes with non-IID datasets is larger. It is noted that the gap between `FedAdp` and `FedAvg` over 50 global rounds is not conspicuous because of the simplicity of the MLR model. Different weighting strategies will not make much difference when the model is reaching its learning capability. In contrast, the weighting strategy will consistently impact the FL training process when a more complex neural network model is applied, as shown in the following experiment.

2) CNN Model: We plot the test accuracy vs. the communication rounds of federated learning in Fig. 4. From Fig. 4, we can tell `FedAdp` always outperforms `FedAvg` when the nodes with non-IID dataset are present. In particular, `FedAdp` converges very fast in the early training stage since the gradient divergence is more obvious in the initial rounds, which makes the effect of assigning adaptive weight for updating the global model even more significant.

To measure the effectiveness of `FedAdp`, we count the number of communication rounds needed to reach a target

[3] For MLR model, the input is a flattened 784-dimensiona (28 × 28) image, and the output is a class label between 0 and 9. Note that MLR model can be extended to strongly-convex setting by adding regularlization term [22].

[4] The CNN has 7 layers with the following structure: 5 × 5 × 32 Convolutional → 2 × 2 MaxPool → 5 × 5 × 64 Convolutional → 2 × 2 MaxPool → 1024 × 512 Fully connected → 512 × 10 Fully connected → Softmax (1,663,370 total parameters). All Convolutional and Fully connected layers are mapped by ReLu activation. The configuration is similar to [8].

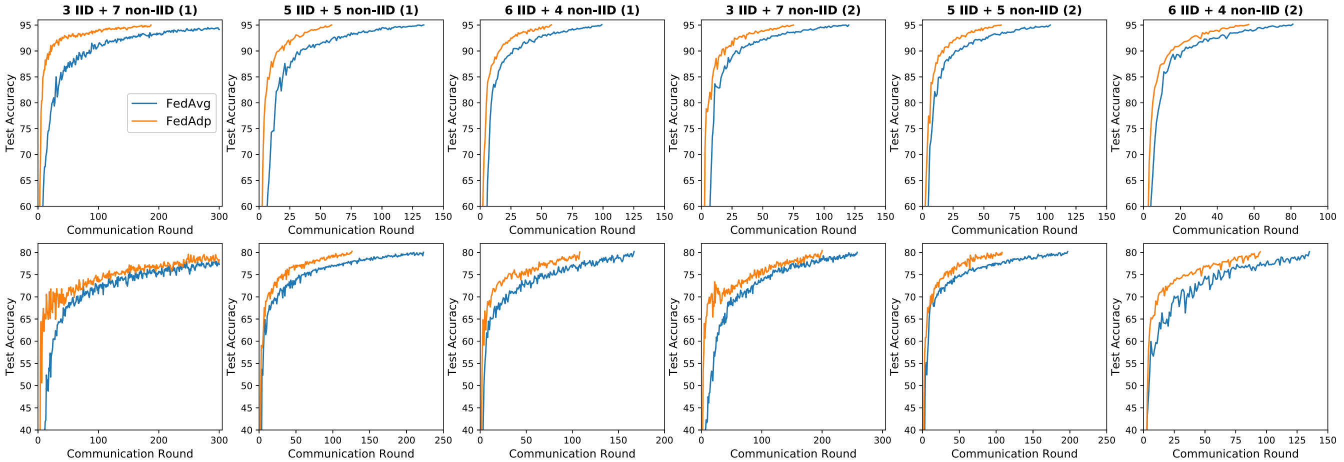

Fig. 4. Test accuracy over communication rounds of `FedAdp` and `FedAvg` with heterogeneous data distribution over participating nodes using CNN model. Upper and lower subplots correspond to training performance on MNIST and FashionMNIST datasets, respectively.

TABLE I
NUMBER OF COMMUNICATION ROUNDS TO REACH A TARGET ACCURACY FOR FEDADP, VERSUS FEDAVG [8], WITHIN 300 ROUNDS. N/A REFERS THAT ALGORITHMS CAN NOT REACH TARGET ACCURACY BEFORE TERMINATION WHERE THE HIGHEST TEST ACCURACY IS SHOWN

	MNIST 95% ACCURACY		
	1-CLASS NON-IID		
	3 IID + 7 non-IID	5 IID + 5 non-IID	6 IID + 4 non-IID
FedAvg	N/A (94.48%)	133	99
FedAdp	**187**	**61**	**58**
	2-CLASS NON-IID		
FedAvg	120	104	81
FedAdp	**75**	**59**	**52**
	FASHION MNIST 80% ACCURACY		
	1-CLASS NON-IID		
	3 IID + 7 non-IID	5 IID + 5 non-IID	6 IID + 4 non-IID
FedAvg	N/A (77.31%)	222	167
FedAdp	N/A (**79.5%**)	**125**	**107**
	2-CLASS NON-IID		
FedAvg	258	196	134
FedAdp	**207**	**107**	**94**

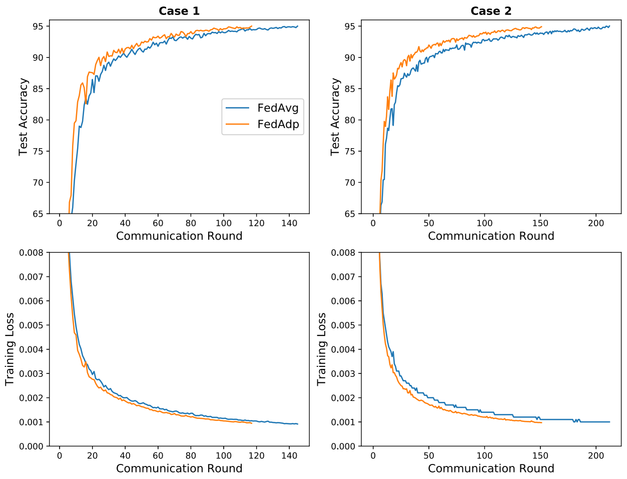

Fig. 5. FL training performance over communication rounds when `FedAdp` is adopted considering general heterogeneous data distribution over participating nodes. The top row and bottom row represent the test accuracy and training loss over the communication round, respectively.

accuracy when `FedAdp` is adopted. Each entry in Table I shows the number of communication rounds necessary to achieve a test accuracy of 95% for CNN on MNIST and 80% for FashionMNIST. The bold number indicates the better result achieved by `FedAdp`, as compared to `FedAvg`. `FedAdp` decreases the number of communication rounds by up to 54.1% and 43.2% for the MNIST task when non-IID nodes are at 1-class and 2-class non-IID setting, respectively. For the FashionMNIST task, the corresponding decreases are up to 43.7% and 45.4%, respectively. In the cases when the target accuracy is not reachable before 300 rounds, `FedAdp` always terminates with higher testing accuracy.

Previously, two extremely skewness cases that $x = 1, 2$ are considered, while the superiority of the proposed weighting strategy is not limited to extreme cases. To verify the proposed weighting strategy in a more general data heterogeneity case, we consider the CNN model for the MNIST dataset in the following two cases.

- *Case 1:* The number of classes of data samples owned by node *i*, denoted by x_i, is randomly selected from the set $\{1, 2, \ldots, 10\}$ without overlapping. Whereafter, the data samples on each node are randomly selected from the x_i-subset of the training dataset.
- *Case 2:* For half of the nodes, their x_i (i.e., the number of classes of data samples) is selected following the uniform distribution $\mathcal{U}(1, 5)$, whereas for the other half, x_i follows the uniform distribution $\mathcal{U}(6, 10)$. The data samples on each node are randomly selected from the x_i-subset of the training dataset.

From Fig. 5, we can see `FedAdp` outperforms `FedAvg` in both cases. In both cases, the convergence performance is worse than the result in Fig. 4 because the number of IID nodes is small and the local dissimilarity is greater in these two cases. However, it is clear by measuring node contribution, `FedAdp` is more rapid in reducing FL loss in each global round thus accelerating model convergence, even without the participation of IID nodes.

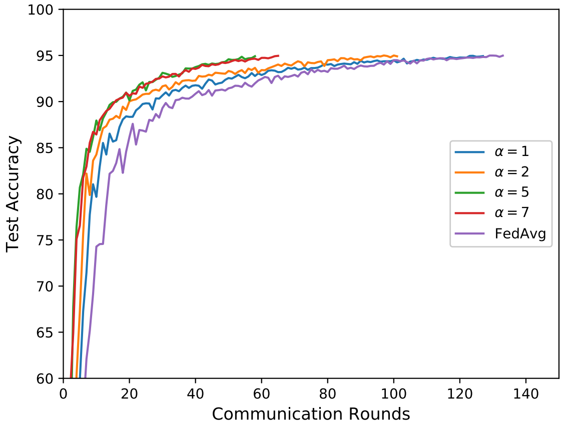

Fig. 6. Effect of setting α on federated learning performance. Data heterogeneity setting is 5 IID + 5 non-IID (1) and CNN model is adopted.

B. Choosing α

One natural question is how to determine α for non-linear function. A large α may increase the convergence by emphasizing the difference of contribution from participating nodes, which hastens model convergence in the initial training stage. Meanwhile, since υ is proportional to α, a large α also narrows the boundary $[\upsilon, \frac{\pi}{2}]$ where the node contribution should be considered, making the contribution of nodes whose angle lays between $[0, \upsilon]$ indistinguishable.

We heuristically choose $\alpha \in \mathbb{Z}^+$ in the ascending order. From Fig. 6, increasing α leads to faster convergence since the gap between small angle and large angle is amplified, so is the difference of contribution from those nodes. However, a larger α is not always effective, especially after the initial training stage. Empirically, the best α is 5 for our experimental setting.

C. Divergence Measurement

Finally, in Fig. 7, we take one experimental case as an example to demonstrate the divergence of local gradients, which captures the overall data heterogeneity of participating nodes. In particular, we track the divergence of gradients over all participating nodes, which is measured by $\sum_i^{\mathcal{S}_t} \frac{1}{|\mathcal{S}_t|} \|\nabla F(\mathbf{w}) - \nabla F_i(\mathbf{w})\|$. Empirically, we observe that our proposed weighting strategy leads to smaller divergence among participating nodes, and the smaller the divergence, the smaller the FL loss. As $\mathbf{w}(t)$ is not a stationary solution along with the training, aggregation by `FedAdp` is seen as a regularization process that restrains the local weight $\mathbf{w}_i(t+1)$ trained by skewed datasets from being deviatory, which lowers the model divergence and consequently accelerates the convergence.

VI. Conclusion

In this paper, we have presented our design of `FedAdp` algorithm that assigns nodes with different weights for updating the global model in each round adaptively to reduce the communication rounds of FL training in the presence of non-IID data. We argue that non-IID data exacerbates the model divergence and observe the nodes with non-IID data make a smaller (or even negative) contribution to the global model aggregation than the nodes with IID data. We have proposed to measure the node contribution based on the angle between local gradient and global gradient and designed a non-linear mapping function to quantify node contribution. We have designed an adaptive weighting strategy that assigns weight proportional to node contribution instead of according to the size of local datasets. The simple yet effective strategy is able to reinforce positive (suppress negative) node contribution dynamically, leading to a significant communication round reduction. Its performance superiority over `FedAvg` is verified both theoretically and experimentally. We have shown that FL training with `FedAdp` has reduced the communication rounds by up to 54.1% on the MNIST dataset and up to 45.4% on the FashionMNIST dataset compared to `FedAvg`.

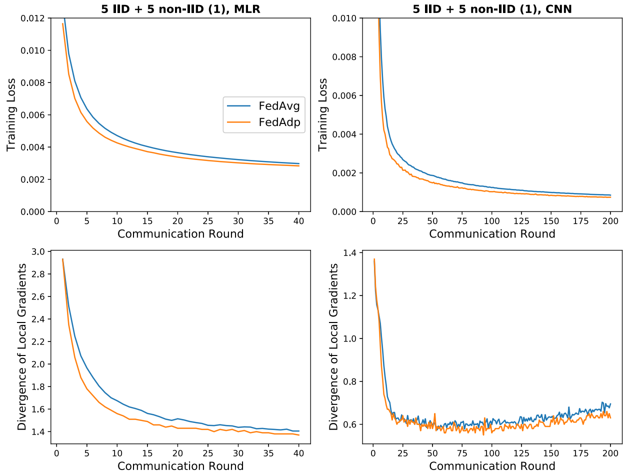

Fig. 7. The connection between the model test loss and the divergence across local gradients. The proposed weighting strategy `FedAdp` gives an impact on alleviating the divergence brought by nodes with skewed datasets. (1) Top row: the training loss on the MNIST dataset under one data heterogeneity setting (5 IID + 5 non-IID (1)). (2) Bottom row: the corresponding divergence measurement.

Appendix A
Proof of Theorem 1

From the β-Lipschitz smoothness of $F(\mathbf{w})$ in Lemma 1 and Taylor expansion, we have

$$F(\mathbf{w}(t+1)) \leq F(\mathbf{w}(t)) + \langle \nabla F(\mathbf{w}(t)), \mathbf{w}(t+1) - \mathbf{w}(t) \rangle + \frac{\beta}{2} \|\mathbf{w}(t+1) - \mathbf{w}(t)\|^2. \quad \text{(A1)}$$

The last two terms on the right hand side of the above inequality are bounded respectively as:

- *Bounding* $\|\mathbf{w}(t+1) - \mathbf{w}(t)\|^2$*:* By the definition of the global aggregation for $\mathbf{w}(t+1)$, we have

$$\|\mathbf{w}(t+1) - \mathbf{w}(t)\| = \mathbb{E}_{i|t}[\|\mathbf{w}_i(t+1) - \mathbf{w}(t)\|]. \quad \text{(A2)}$$

By following SGD optimization, for each term within the expectation in the right hand side of A2, we have

$$\mathbf{w}_i(t+1) = \mathbf{w}(t) - \eta \nabla F_i(\mathbf{w}(t)). \quad \text{(A3)}$$

Therefore,

$$\begin{aligned}\|\mathbf{w}(t+1)-\mathbf{w}(t)\|^2 &= \Big(\mathbb{E}_{i|t}[\|\mathbf{w}_i(t+1)-\mathbf{w}(t)\|]\Big)^2\\ &= \eta^2\Big(\mathbb{E}_{i|t}[\|\nabla F_i(\mathbf{w}(t))\|]\Big)^2\\ &\overset{1}{\leq} \eta^2\mathbb{E}_{i|t}\Big[\|\nabla F_i(\mathbf{w}(t))\|^2\Big], \end{aligned} \tag{A4}$$

where inequality 1 holds by Cauchy-Schwarz inequality.

- *Bounding* $\langle\nabla F(\mathbf{w}(t)), \mathbf{w}(t+1)-\mathbf{w}(t)\rangle$*:* Again, by the definition of the global aggregation for $\mathbf{w}(t+1)$ and A3 we have

$$\begin{aligned}&\langle\nabla F(\mathbf{w}(t)), \mathbf{w}(t+1)-\mathbf{w}(t)\rangle\\ &\quad = -\eta\mathbb{E}_{i|t}[\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle]. \end{aligned}\tag{A5}$$

The expectation term in A5 can be further rewritten as

$$\begin{aligned}&\mathbb{E}_{i|t}[\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle]\\ &\quad = \mathbb{E}_{i|t}\Bigg[\frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}\\ &\qquad\quad \cdot\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|\Bigg]\\ &\quad \overset{2}{\geq} \mathbb{E}_{i|t}\Bigg[\frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}\cdot\frac{\|F_i(\mathbf{w}(t))\|^2}{B}\Bigg], \end{aligned}\tag{A6}$$

where inequality 2 comes from Assumptions 2 that local dissimilarity is upper bounded by B.

Plugging A6 into A5, then the last two terms on the right hand side of A1 are expressed as

$$\begin{aligned}&\langle\nabla F(\mathbf{w}(t)), \mathbf{w}(t+1)-\mathbf{w}(t)\rangle + \frac{\beta}{2}\|\mathbf{w}(t+1)-\mathbf{w}(t)\|^2\\ &\leq -\eta\mathbb{E}_{i|t}\Bigg[\frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}\cdot\frac{\|F_i(\mathbf{w}(t))\|^2}{B}\Bigg]\\ &\quad + \frac{\beta\eta^2}{2}\mathbb{E}_{i|t}\Big[\|\nabla F_i(\mathbf{w}(t))\|^2\Big]\\ &\overset{3}{\leq} -\eta\mathbb{E}_{i|t}\Bigg[\bigg(\frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}-\frac{B\beta\eta}{2}\bigg)\\ &\qquad\quad \cdot\frac{A^2}{B}\|\nabla F(\mathbf{w}(t))\|^2\Bigg], \end{aligned}\tag{A7}$$

where inequality 3 holds because of Assumptions 2 that local dissimilarity is lower bounded by A.

Finally, Theorem 1 is proved by substituting A7 into A1.

APPENDIX B
PROOF OF THEOREM 2

We consider the general case that participating nodes have a different number of data samples. For node i with data size D_i, we create D_i virtual nodes, each with a unit sample size. Hereinafter, we use index $(i,j), j\in\{1,\ldots,D_i\}$ to denote the j-th virtual node split from the participating node $i, i\in\mathcal{S}_t$, where the gradient information is kept on virtual nodes as on the participating node (e.g., $\nabla F_{i,j}(\mathbf{w}(t) = \nabla F_i(\mathbf{w}(t)), \theta_{i,j} = \theta_i$). As such, all virtual nodes split by node i share the same weight (i.e., $\widetilde{\psi}_{i,j}(t) = \widetilde{\psi}_{i,k}(t), \forall j,k\in\{1,\ldots,D_i\}$), where $\widetilde{\psi}_{i,j}(t)$ denotes the weight for virtual node (i,j). The weight of node i is $\widetilde{\psi}_i(t) = \sum_{j=1}^{D_i}\widetilde{\psi}_{i,j}(t) = D_i\widetilde{\psi}_{i,j}(t)$.

From (7), $\theta_{i,j} = \theta_i$ monotonically decreases with $\frac{\langle\nabla F(\mathbf{w}(t)),\nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}$. From (9), $f(\cdot)$ is a decreasing function of θ. Thus, by that $\widetilde{\psi}_{i,j}(t) = \frac{e^{f(\widetilde{\theta}_{i,j}(t))}}{\sum_{i'=1}^{|\mathcal{S}_t|}D_{i'}e^{f(\widetilde{\theta}_{i'}(t))}}$, we can see $\widetilde{\psi}_{i,j}(t)$ monotonically increases with $\frac{\langle\nabla F(\mathbf{w}(t)),\nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}$. Therefore, generic $\widetilde{\psi}_{i,j}(t)$ satisfies the following criterion

$$\begin{aligned}&\widetilde{\psi}_{i,j}(t) \propto \frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}\\ &\widetilde{\psi}_{i,j}(t)\geq 0\ \ \forall i,j,t\\ &\sum_{i=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_i}\widetilde{\psi}_{i,j}(t) = \sum_{i=1}^{|\mathcal{S}_t|}\widetilde{\psi}_i(t) = 1, \end{aligned}\tag{B1}$$

with the corresponding bound of the expected loss being

$$\begin{aligned}&F(\mathbf{w}(t+1))\leq F(\mathbf{w}(t))\\ &\quad -\eta\sum_{i}^{|\mathcal{S}_t|}\bigg(\frac{\langle\nabla F(\mathbf{w}(t)), \nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}\widetilde{\psi}_i(t)-\frac{B\beta\eta}{2}\bigg)\\ &\quad \cdot\frac{A^2}{B}\|\nabla F(\mathbf{w})\|^2. \end{aligned}\tag{B2}$$

where $\widetilde{\psi}_i(t)$ is defined as in (10).

In order to compare the expected loss achieved by `FedAdp` and `FedAvg`, one can simply measure the expectation term in (5). We use $u_{i,j}$ to denote the contribution from virtual node *j* of participating node *i* for model aggregation. In each global round, we sort the contribution from all the virtual nodes that is measured by the correlation $\frac{\langle\nabla F(\mathbf{w}(t)),\nabla F_i(\mathbf{w}(t))\rangle}{\|\nabla F(\mathbf{w}(t))\|\|\nabla F_i(\mathbf{w}(t))\|}$ between the local gradient and the global gradient in descending order, that is $u_{1,1} = u_{1,2} = \cdots = u_{1,D_1}\geq u_{2,1} = u_{2,2} = \cdots = u_{2,D_2}\geq\cdots\geq u_{|\mathcal{S}_t|,1} = u_{|\mathcal{S}_t|,2} = \cdots = u_{|\mathcal{S}_t|,D_{|\mathcal{S}_t|}}$. Apparently, the weight assigned to virtual node in `FedAdp` should follow the same order $\widetilde{\psi}_{1,1} = \widetilde{\psi}_{1,2} = \cdots = \widetilde{\psi}_{1,D_1}\geq\widetilde{\psi}_{2,1} = \widetilde{\psi}_{2,2} = \cdots = \widetilde{\psi}_{2,D_2}\geq\cdots\geq\widetilde{\psi}_{|\mathcal{S}_t|,1} = \widetilde{\psi}_{|\mathcal{S}_t|,2} = \cdots = \widetilde{\psi}_{|\mathcal{S}_t|,D_{|\mathcal{S}_t|}}$, with $\sum_i\sum_j\widetilde{\psi}_{i,j} = 1$. As such, by Chebyshev's inequality [23], we have the following hold for any $u_{m,j}$, $u_{n,j'}$,

$$\begin{aligned}&\bar{\psi}(u_{m,j}-u_{n,j'})\Bigg(\frac{\widetilde{\psi}_{m,j}}{\bar{\psi}_{m,j}}-\frac{\widetilde{\psi}_{n,j'}}{\bar{\psi}_{n,j'}}\Bigg)\geq 0\\ &\bar{\psi}\Big[u_{m,j}\widetilde{\psi}_{m,j}\bar{\psi}_{n,j'}+u_{n,j'}\widetilde{\psi}_{n,j'}\bar{\psi}_{m,j}\Big]\\ &\quad\geq\bar{\psi}\Big[u_{m,j}\widetilde{\psi}_{n,j'}\bar{\psi}_{m,j}+u_{n,j'}\widetilde{\psi}_{m,j}\bar{\psi}_{n,j'}\Big], \end{aligned}\tag{B3}$$

where $\bar{\psi} = \bar{\psi}_{m,j} = \bar{\psi}_{n,j'} = \frac{1}{D}$ denotes the weight of `FedAvg` for all virtual nodes with $D = \sum_i^{|\mathcal{S}_t|}D_i$.

Adding all the D^2 inequalities, we have

$$\begin{aligned}&\bar{\psi}\Bigg[\sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m}\sum_{n=1}^{|\mathcal{S}_t|}\sum_{j'=1}^{D_n}u_{m,j}\widetilde{\psi}_{m,j}\bar{\psi}_{n,j'}+u_{n,j'}\widetilde{\psi}_{n,j'}\bar{\psi}_{m,j}\Bigg]\\ &\geq\bar{\psi}\Bigg[\sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m}\sum_{n=1}^{|\mathcal{S}_t|}\sum_{j'=1}^{D_n}u_{m,j}\widetilde{\psi}_{n,j'}\bar{\psi}_{m,j}+u_{n,j'}\widetilde{\psi}_{m,j}\bar{\psi}_{n,j'}\Bigg]\end{aligned}$$

$$\sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m} u_{m,j}\widetilde{\psi}_{m,j} \underbrace{\sum_{n=1}^{|\mathcal{S}_t|}\sum_{j'=1}^{D_n} \bar{\psi}_{n,j'}}_{=1} + \sum_{n=1}^{|\mathcal{S}_t|}\sum_{j'=1}^{D_n} u_{n,j'}\widetilde{\psi}_{n,j'}$$

$$\times \underbrace{\sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m} \bar{\psi}_{m,j}}_{=1}$$

$$\geq \sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m} u_{m,j}\bar{\psi}_{m,j} \underbrace{\sum_{n=1}^{|\mathcal{S}_t|}\sum_{j'=1}^{D_n} \widetilde{\psi}_{n,j'}}_{=1} + \sum_{n=1}^{|\mathcal{S}_t|}\sum_{j'=1}^{D_n} u_{n,j'}\bar{\psi}_{n,j'}$$

$$\times \underbrace{\sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m} \widetilde{\psi}_{m,j}}_{=1}$$

$$2 \cdot \sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m} u_{m,j}\widetilde{\psi}_{m,j} \geq 2 \cdot \sum_{m=1}^{|\mathcal{S}_t|}\sum_{j=1}^{D_m} u_{m,j}\bar{\psi}_{m,j}$$

$$\underbrace{\sum_{m} u_m\widetilde{\psi}_m}_{\texttt{FedAdp}} \overset{4}{\geq} \underbrace{\sum_{m} u_m\psi_m}_{\texttt{FedAvg}} . \tag{B4}$$

where $u_m = u_{m,1} = \cdots = u_{m,D_m}$. Inequality 4 holds because $\widetilde{\psi}_m = \widetilde{\psi}_{m,j} \cdot D_m$ and $\psi_m = \bar{\psi}_{m,j} \cdot D_m$ with $\widetilde{\psi}_m$ and ψ_m denoting the weight for model aggregation in FedAdp and FedAvg, respectively. The equality 4 holds when $u_i = u_j, \forall i, j \in \mathcal{S}_t$.

Due to the greater expectation term in (5). FedAdp results in greater decrease of FL loss in each global round, as compared to FedAvg. This completes the proof.

References

[1] K. L. Lueth. (Aug. 2019). *State of the IoT 2018: Number of IoT Devices Now at 7B-Market Accelerating*. [Online]. Available: https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] O.-E.-K. Aktouf, T. Zhang, J. Gao, and T. Uehara, "Testing location-based function services for mobile applications," in *Proc. IEEE Symp. Serv. Orient. Syst. Eng. (SOSE)*, 2015, pp. 308–314.

[4] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 32–39, Nov. 2011.

[5] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[6] Z. Xiong, Y. Zhang, D. Niyato, P. Wang, and Z. Han, "When mobile blockchain meets edge computing," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 33–39, Aug. 2018.

[7] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869–904, 2nd Quart., 2020.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist. Conf. (AISTATS)*, 2017, pp. 1273–1282.

[9] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016. [Online]. Available: arXiv:1610.05492.

[10] W. Y. B. Lim *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.

[11] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018. [Online]. Available: arXiv:1806.00582.

[12] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2020, pp. 1698–1707.

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2018. [Online]. Available: arXiv:1812.06127.

[14] S. Wang *et al.*, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[15] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2019, pp. 1387–1395.

[16] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, 2019, pp. 1–7.

[17] H. T. Nguyen, V. Sehwag, S. Hosseinalipour, C. G. Brinton, M. Chiang, and H. V. Poor, "Fast-convergent federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 201–218, Jan. 2021.

[18] L. Wang, W. Wang, and B. Li, "CMFL: Mitigating communication overhead for federated learning," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Dallas, TX, USA, 2019, pp. 954–964.

[19] Y. Chen, X. Sun, and Y. Jin, "Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4229–4238, Oct. 2020.

[20] Z. Chai, Y. Chen, L. Zhao, Y. Cheng, and H. Rangwala, "FedAT: A communication-efficient federated learning method with asynchronous tiers under non-IID data," 2020. [Online]. Available: arXiv:2010.05958.

[21] M. N. Gibbs and D. J. C. MacKay, "Variational Gaussian process classifiers," *IEEE Trans. Neural Netw.*, vol. 11, no. 6, pp. 1458–1464, Nov. 2000.

[22] C. T. Dinh *et al.*, "Federated learning over wireless networks: Convergence analysis and resource allocation," 2019. [Online]. Available: arXiv:1910.13067.

[23] A. W. Marshall and I. Olkin, "Multivariate Chebyshev inequalities," *Ann. Math. Stat.*, vol. 31, pp. 1001–1014, Dec. 1960.

Hongda Wu (Student Member, IEEE) received the M.A.Sc. degree in electrical engineering from the Communication University of China in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Science, York University, Canada. His research interests include federated learning, reinforcement learning, wireless network, and the Internet of Things.

Ping Wang (Senior Member, IEEE) received the Bachelor and Master degrees in electrical and computer engineering from the Huazhong University of Science and Technology, in 1994 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in 2008. She joined York University as an Associate Professor in August 2018. Prior to that, she worked with Nanyang Technological University, Singapore, from 2008 to July 2018. Her research interests are mainly in wireless communication networks, cloud computing, and the Internet of Things. Her scholarly works have been widely disseminated through top-ranked IEEE journals/conferences and received the Best Paper Awards from IEEE Wireless Communications and Networking Conference in 2012 and 2020, from IEEE Communication Society: Green Communications and Computing Technical Committee in 2018, and from IEEE International Conference on Communications in 2007.