Adaptive Central Acceleration with Variance Control for Robust Federated Optimization in Ubiquitous Intelligence

Lei Zhao, Wu-Sheng Lu, *Life Fellow, IEEE,* and Lin Cai, *Fellow, IEEE*

***Abstract*—Federated Learning (FL) in Intelligent Internet of Things (IIoT) environments faces critical challenges, including sparse client participation, non-IID local data distributions, and unreliable communication, which lead to slow convergence and high variance in global updates. To address these issues, we propose Adaptive Central Federated Momentum Optimization (ACFMO), an optimization framework that enhances FL efficiency and stability under constrained participation. ACFMO integrates an adaptive central acceleration mechanism that dynamically adjusts momentum updates based on real-time client availability, preventing instability and ensuring smoother global model updates. Additionally, a variance-controlled local updating strategy refines client contributions, mitigating high variance caused by infrequent and heterogeneous updates. Extensive experiments across diverse FL scenarios demonstrate that ACFMO significantly accelerates convergence, reduces communication overhead, and improves model stability compared to state-of-the-art FL methods, making it particularly well-suited for real-world IIoT deployments where network and computational resources are constrained.**

***Index Terms*—Federated Learning, Adaptive Central Acceleration, Variance-Controlled Updates, Intelligent IoT.**

I. Introduction

The rapid evolution of the Intelligent Internet of Things (IIoT) has ushered in an era of ubiquitous intelligence, where interconnected devices, from simple sensors to advanced autonomous systems, collaborate to deliver adaptive, context-aware services across diverse domains such as healthcare, smart cities, and industrial automation [1]–[4]. For example, wearable medical devices in healthcare continuously monitor patient vital signs to support remote diagnosis and personalized treatment plans, while smart city infrastructures utilize traffic sensors and energy-efficient grid management to optimize resource allocation and public safety [5], [6]. Similarly, autonomous systems, including self-driving vehicles and aerial drones, rely on real-time data processing to make adaptive decisions in highly dynamic environments [7], [8].

Federated Learning (FL) has emerged as a promising paradigm to enable collaborative model training across these distributed devices while preserving data privacy [9]–[11]. By allowing devices to train local models without sharing raw data, FL reduces privacy risks and alleviates communication overhead. However, real-world IIoT deployments introduce significant challenges due to highly heterogeneous data distributions, sparse client participation, and unreliable communication infrastructure [12]. While existing FL research has addressed aspects of these challenges through adaptive aggregation strategies [13], [14], variance reduction techniques [15], and momentum-based optimization [16], these approaches often assume relatively stable participation rates and frequent client updates. In IIoT environments, however, only a small fraction of devices may participate in each round due to intermittent connectivity, power constraints, or network limitations, leading to inefficient utilization of available updates and slowing global convergence.

The impact of these challenges is evident in various IIoT applications. In healthcare, wearable medical devices generate non-IID data due to physiological differences and sensor variations. Existing FL methods struggle to generalize across diverse client populations, particularly when device participation is sparse. Similarly, smart city infrastructures rely on edge devices such as traffic cameras and environmental sensors, where fluctuating network conditions result in unpredictable participation rates. Standard FL aggregation methods fail to effectively handle sparse updates, leading to slow convergence and unstable learning. Autonomous systems further illustrate the difficulty of FL in IIoT, as self-driving vehicles and drones require collaborative learning under strict latency constraints while dealing with diverse sensor observations [17]. Frequent model synchronization is impractical due to bandwidth constraints [18], and infrequent updates can cause global models to diverge.

To address these challenges, we propose Adaptive Central Federated Momentum Optimization (ACFMO), an optimization framework designed to enhance federated learning in IIoT by effectively mitigating variance caused by sparse client participation while accelerating global convergence. At the central server, ACFMO integrates an adaptive momentum-based optimization strategy that dynamically adjusts momentum updates based on real-time participation variability. Unlike conventional FL optimizers that rely on fixed momentum parameters, ACFMO prevents instability caused by fluctuating participation rates, ensuring that the global model remains stable even when only a small fraction of clients contribute updates in each round. This adaptive central acceleration mechanism leverages historical gradients to smooth updates, reducing high-variance effects from infrequent participation and enabling faster, more robust learning.

In addition to central acceleration, ACFMO incorporates a variance-controlled local updating mechanism that enhances the consistency and impact of sparse client updates. Traditional variance reduction techniques assume frequent and balanced updates, which limits their effectiveness under sparse participation. ACFMO addresses this by refining local updates in

L. Zhao, W.-S. Lu, and L. Cai are with Dept. of Electrical & Computer Engineering, University of Victoria, 3800 Finnerty Road, Victoria, BC, V8P 5C2, Canada. *Corresponding author: Lin Cai (E-mail: cai@uvic.ca).

a way that ensures meaningful gradient contributions, even when devices participate infrequently or operate on heterogeneous data. By combining adaptive central acceleration with variance-controlled local updates, ACFMO significantly improves convergence efficiency, reduces communication overhead, and enhances the overall robustness of federated learning systems in dynamic, distributed IIoT environments. Experimental evaluations under various FL scenarios demonstrate that ACFMO achieves faster convergence, better model stability, and lower communication costs compared to existing FL approaches, making it particularly well-suited for real-world IIoT deployments with constrained network and computational resources.

The rest of this paper is structured as follows. Section II reviews related work on federated learning optimization and techniques for improving training efficiency. Section III formulates the federated optimization problem, outlining key challenges in IIoT environments. The design and implementation of ACFMO are presented in Section IV, detailing its adaptive central acceleration mechanism and variance-controlled local updates. Section V provides a comprehensive evaluation of ACFMO through extensive experiments under various FL scenarios. Finally, Section VI concludes the paper with a summary of findings and potential directions for future research.

II. Related Work

FL has emerged as a key paradigm for training shared models across distributed, privacy-sensitive datasets, supporting applications in healthcare, smart cities, and autonomous systems [19]. One of the earliest and most widely adopted FL algorithms is Federated Averaging (FedAvg) [20], which allows clients to perform multiple local updates before communicating with a central server. Under IID local datasets, FedAvg aligns with parallel stochastic gradient descent and provides strong empirical performance with provable asymptotic convergence [21]. However, these assumptions rarely hold in real-world FL scenarios, where client participation is dynamic and local datasets exhibit significant heterogeneity [11].

One of the primary challenges in FL is statistical heterogeneity, where local data distributions vary significantly across clients [10], [22]. Non-IID data can cause training drift during local updates, degrading global model performance and slowing convergence [23]. To mitigate this, researchers have developed methods such as FedProx [22], which introduces a proximal term to regularize local updates, and SCAFFOLD [24], which employs control variates to reduce gradient drift. Additionally, studies on convergence guarantees [25]–[27] and bounded gradient techniques [28], [29] aim to improve stability under non-IID conditions. While these approaches improve learning under data heterogeneity, they do not explicitly address instability caused by sparse client participation, a frequent occurrence in real-world IIoT settings.

Communication efficiency is another critical issue in FL, particularly in resource-constrained environments such as IIoT. Several works [27], [30] have explored adaptive client sampling to reduce communication overhead by selecting a subset of clients per round. Although these techniques improve efficiency, they often assume stable participation patterns and do not mitigate the variance introduced by fluctuating client availability. In highly dynamic settings, where only a small fraction of clients participate in each round, these methods struggle to maintain stable convergence, resulting in performance degradation.

Recent research has also tackled challenges such as model heterogeneity, representation degeneration, and personalization in FL. FedPAC [13] enhances local-global feature alignment using shared feature representations and personalized classifier heads, but its reliance on consistent client participation limits its robustness in dynamic environments. FedDBE [15] introduces a Domain Bias Eliminator (DBE) to reduce domain discrepancies between clients and the server, improving generalization; however, it assumes stable representation spaces and struggles with scalability under resource constraints. FedALA [14] adaptively aggregates global and local models to mitigate statistical heterogeneity, but it incurs additional client-side overhead and does not address update variance due to random participation. FedGH [16] provides a communication-efficient solution for model heterogeneity by training a generalized global prediction header but assumes stable client participation and suffers from high variance in local updates under heterogeneous data distributions.

To address these limitations, we propose ACFMO, an optimization framework designed to enhance FL efficiency in dynamic and heterogeneous environments. ACFMO incorporates an adaptive central acceleration mechanism that leverages momentum-based optimization to stabilize global updates and accelerate convergence, even when client participation is sparse. Additionally, a variance-controlled local updating strategy reduces gradient variance in local contributions, ensuring stable learning despite very limited participation in each round. By effectively balancing convergence speed, stability, and communication efficiency, ACFMO provides a scalable and robust solution for real-world FL deployments in IIoT environments, where device availability fluctuates and resource constraints are prevalent.

III. Motivation and Problem Formulation

A key challenge in FL for IIoT environments is sparse client participation, where only a small subset of devices contribute updates in each round due to power limitations, computational constraints, or network variability. This leads to inefficient model aggregation, increased variance in global updates, and slower convergence. Unlike many existing FL algorithms that assume frequent and stable client participation, real-world IIoT systems exhibit highly unpredictable and sparse updates, making it difficult to ensure steady learning progress.

Additionally, local data heterogeneity further complicates the optimization process, as each IIoT device collects data from distinct, non-IID distributions. This statistical discrepancy results in misaligned local updates, leading to model drift and instability in global aggregation. Without proper variance control mechanisms, conventional FL methods struggle to maintain stable learning, often requiring excessive training

iterations to reach convergence. The interplay between sparse participation and heterogeneous data significantly degrades learning efficiency, making it challenging to train an accurate and robust global model.

To formally capture this problem, we consider a distributed learning system where data is decentralized across IIoT devices. The entire dataset consists of n training samples, denoted as $\{\boldsymbol{x}_k, y_k\}_{k=1}^n$, where $\boldsymbol{x}_k$ and y_k represent input features and corresponding labels. Let E be the set of all IIoT devices, and let P_i represent the local dataset of the i-th device, containing n_i samples for $i = 1, 2, \ldots, |E|$. These local datasets are non-overlapping, i.e., $P_i \cap P_j = \emptyset$ whenever $i \neq j$.

The objective of FL is to collaboratively minimize a global function, formulated as

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\boldsymbol{w}) = \sum_{i=1}^{|E|} \frac{n_i}{n} f_i(\boldsymbol{w}), \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ represents the global model parameters to be optimized, and $f_i(\boldsymbol{w})$ is the local objective function for the i-th IIoT device, defined as

$$f_i(\boldsymbol{w}) = \frac{1}{n_i} \sum_{k\in P_i} F_k(\boldsymbol{w}), \quad i = 1, 2, \ldots, |E|, \tag{2}$$

where $F_k(\boldsymbol{w})$ denotes the loss function associated with the k-th training sample.

The global objective function $f(\boldsymbol{w})$ aggregates the contributions of all IIoT devices, weighted by their dataset sizes. However, due to the challenges of sparse participation and data heterogeneity, optimizing this function in an FL setting is particularly difficult. Sparse client updates introduce high variance in global parameter aggregation, leading to instability in training, while non-IID data distributions cause local models to drift away from the global objective. Addressing these challenges requires a robust FL optimization approach that effectively stabilizes training despite these constraints.

In the following section, we introduce ACFMO, a framework specifically designed to address the challenges of sparse client participation and heterogeneous data distributions in FL. Unlike conventional approaches, ACFMO dynamically adjusts global momentum based on participation variability and regulates local gradient updates to mitigate the impact of infrequent client contributions. This integration ensures stable convergence and improved learning efficiency in non-IID FL settings with limited participation per round.

IV. Adaptive Central Federated Momentum Optimization with Variance-Controlled Updates

A. Challenges of Sparse IIoT Participation in Federated Optimization

In FL systems, sparse IIoT participation is a fundamental challenge in real-world deployments across various applications, including healthcare, smart cities, and autonomous systems. In healthcare, wearable devices such as smartwatches and fitness trackers often operate under limited battery life and intermittent connectivity, resulting in infrequent participation in FL training. Similarly, in smart cities, edge nodes such as traffic monitors and adaptive streetlights may become unavailable due to network congestion or maintenance, further reducing the number of active participants in each round. Autonomous systems, including self-driving vehicles and drones, experience varying computational workloads and mobility constraints, leading to inconsistent engagement in the training process. As a result, only a small and dynamically changing subset of IIoT devices contributes updates in each round, introducing significant variance in model aggregation and slowing global convergence.

At the beginning of the r-th round, the global model $\boldsymbol{w}^{r-1}$ is distributed to a subset of participating IIoTs, denoted as $S^r \subseteq [E]$, where $|S^r| = S$. Each IIoT $i \in S^r$ performs local training using its dataset P_i, consisting of n_i samples. The total number of training samples available in round r is given by $n^r = \sum_{i\in S^r} n_i$. These local datasets, which may include wearable sensor data, traffic patterns, or environmental readings, reflect the diverse and non-IID nature of IIoT data. Each selected IIoT computes its local gradient $\nabla f_i(\boldsymbol{w}^{r-1})$ using its dataset P_i and transmits this to the central server. The server aggregates these gradients to form the global anchor gradient

$$\boldsymbol{g}(\boldsymbol{w}^{r-1}) = \sum_{i\in S^r} \frac{n_i}{n^r} \nabla f_i(\boldsymbol{w}^{r-1}), \tag{3}$$

where $\boldsymbol{g}(\boldsymbol{w}^{r-1})$ represents a weighted average of the gradients from participating IIoTs. The anchor gradient provides an approximation of the full global gradient $\nabla f(\boldsymbol{w}^{r-1})$, which would be computed if all IIoTs participated. Despite being an unbiased estimator, i.e., $E[\boldsymbol{g}(\boldsymbol{w}^{r-1})] = \nabla f(\boldsymbol{w}^{r-1})$, the variance introduced by inconsistent IIoT availability can significantly slow convergence and destabilize the training process.

Dynamic participation reduces the computational and communication load for each round, as only a fraction, $S/|E|$, of the IIoTs are involved. This is particularly advantageous in large-scale FL systems across ubiquitous intelligence applications, where full IIoT participation is infeasible due to bandwidth and resource constraints. However, this setup also poses critical challenges. First, the variability in IIoT contributions exacerbates the already significant impact of statistical heterogeneity, causing divergence in the global model updates. Second, the increased variance in the anchor gradient may require additional iterations to achieve convergence, negating the efficiency gains of partial participation.

To address these challenges, our proposed method introduces an adaptive central acceleration mechanism to stabilize global updates and reduce the effects of variance caused by dynamic IIoT participation. This mechanism leverages momentum-based optimization at the server to mitigate the drift caused by heterogeneous data and inconsistent IIoT availability. Additionally, variance-controlled local updates are employed to further enhance the robustness of the global model aggregation process. These innovations make our approach particularly suited for ubiquitous intelligence applications, where IIoT participation is inherently dynamic, and resource constraints demand scalable and efficient optimization strategies.

B. Adaptive Central Acceleration

To address the variability introduced by dynamic IIoT participation, the central server employs an adaptive central acceleration mechanism that stabilizes global updates and enhances convergence efficiency. The acceleration is achieved by leveraging momentum-based optimization, which integrates historical gradient information into the update process, reducing the impact of inconsistencies in IIoT contributions.

To ensure efficient update acceleration, however, these information must be utilized with care and this is realized with an adaptive strategy detailed below. The central updating in the r-th round is designed as

$$\hat{\boldsymbol{w}}^{r-1} = \boldsymbol{w}^{r-1} - \alpha_g^{r-1} \frac{\hat{\boldsymbol{m}}_{r-1}}{\sqrt{\hat{\boldsymbol{v}}_{r-1} + \epsilon}}, \tag{4}$$

where $\hat{\boldsymbol{m}}_r$ is an exponential moving average of the global gradients

$$\hat{\boldsymbol{m}}_r = \beta_1^{r-1} \boldsymbol{m}_0 + (1-\beta_1) \sum_{i=0}^{r-1} \beta_1^{r-i-1} \boldsymbol{g}(\boldsymbol{w}^i), \tag{5}$$

which is an estimate of the first moment of the global gradient, $\hat{\boldsymbol{v}}_r$ is an estimate of the second moment of the global gradient given by

$$\hat{\boldsymbol{v}}_r = \beta_2^{r-1} \boldsymbol{v}_0 + (1-\beta_2) \sum_{i=0}^{r-1} \beta_2^{r-i-1} \boldsymbol{g}(\boldsymbol{w}^i)^2 \tag{6}$$

where $\boldsymbol{g}(\boldsymbol{w}^i)^2$ is a vector obtained by component-wise squaring vector $\boldsymbol{g}(\boldsymbol{w}^i)$, ϵ is a small positive scalar to avoid ill-conditioning, and

$$\alpha_g^{r-1} = \alpha_g^0 (1-\beta_1) \cdot \sqrt{\frac{1-\beta_2^{r-1}}{1-\beta_2}}, \tag{7}$$

where we usually set $\alpha_g^0 = 0.02$.

In practice, the estimated moments $\hat{\boldsymbol{m}}_r$ and $\hat{\boldsymbol{v}}_r$ are evaluated recursively using

$$\hat{\boldsymbol{m}}_r = \beta_1 \hat{\boldsymbol{m}}_{r-1} + (1-\beta_1) \boldsymbol{g}(\boldsymbol{w}^{r-1}) \tag{8}$$

and

$$\hat{\boldsymbol{v}}_r = \beta_2 \hat{\boldsymbol{v}}_{r-1} + (1-\beta_2) \boldsymbol{g}(\boldsymbol{w}^{r-1})^2 \tag{9}$$

respectively, which can readily be derived from (5) and (6) assuming $\hat{\boldsymbol{m}}_{\boldsymbol{0}} = \boldsymbol{0}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{0}} = \boldsymbol{0}$.

The recursive formulas (8) and (9) reduce the computation required by (4) to minimum. We also remark that the decay rates β_1 and β_2 weigh the importance of the past moments relative to the present gradient. Therefore, they are always set in the range $(0, 1)$, whose actual values are influential on how quickly the model is updated using (4) and hence must be chosen with care. Larger values of β_1 and β_2 tend to yield consistently good and more stable results when the number of selected IIoT S is very small.

The proposed adaptive mechanism ensures efficient and stable updates by leveraging both the first and second moment information of the gradients, enabling the central server to effectively mitigate the gradient drift caused by data heterogeneity and dynamic IIoT participation. By dynamically balancing the influence of historical and current gradients, the mechanism enhances the robustness and convergence speed of the global model, even under highly variable IIoT availability. This innovation is particularly transformative for ubiquitous intelligence applications, where dynamic IIoT participation is inevitable due to resource constraints and connectivity challenges. In healthcare, adaptive acceleration ensures reliable aggregation of fragmented data from wearables, improving remote monitoring and diagnostics [31]. In smart cities, it stabilizes model updates from edge devices with fluctuating connectivity, optimizing urban resource management and public safety. Similarly, in autonomous systems, it enables consistent training despite the mobility-induced variability of participating devices, ensuring robust and scalable learning. These advancements position adaptive central acceleration as a cornerstone for efficient FL in dynamic, real-world environments.

TABLE I
DESCRIPTION OF NOTATION

NOTATION	DESCRIPTION
$f(\boldsymbol{w})$	The global objective function
$f_i(\boldsymbol{w})$	The i-th local objective function
$\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r)$	The local gradient from individual training sample or a batch of the training samples
$\boldsymbol{w}^{r-1}$	The current global model in the r-th round
$\hat{\boldsymbol{w}}^{r-1}$	The global model after central acceleration in the r-th round
$\hat{\boldsymbol{w}}_{i,k}^r$	The i-th local model in the r-th global round and the k-th local iteration
E	The set of total IIoTs
S^r	The randomly selected IIoT subset in the r-th round
α_l, α_g	Local and global learning rates
$\hat{\boldsymbol{m}}_r$	The estimate of the first moment of the global gradient
$\hat{\boldsymbol{v}}_r$	The estimate of the second moment of the global gradient
β_1, β_2	Decay rates to generate $\hat{\boldsymbol{m}}_r$ and $\hat{\boldsymbol{v}}_r$
n_i	The number of local samples in IIoT i
n^r	The total number of samples to calculate the anchor gradient in the r-th round
$\boldsymbol{g}(\boldsymbol{w}^{r-1})$	The anchor gradient in the r-th round
n^j	The number of samples with nonzero j-th feature
n_i^j	The number of samples in the local data set of IIoT i with nonzero j-th feature

C. Adaptive Local Training and Global Aggregation

The accelerated model $\hat{\boldsymbol{w}}^{r-1}$ is distributed to the selected IIoTs in S^r and shared by these local IIoTs as models $\{\hat{\boldsymbol{w}}_{i,0}^r = \hat{\boldsymbol{w}}^{r-1}\}_{i \in S^r}$. The IIoTs in S^r conduct K local stochastic updates based on their own local data sets. The auxiliary local objective function of IIoT i at the r-th global round is given by

$$\tilde{f}_i(\boldsymbol{w}) = f_i(\boldsymbol{w}) - (\nabla f_i(\boldsymbol{w}^{r-1}) - \nabla f(\boldsymbol{w}^{r-1}))^T \boldsymbol{w}. \tag{10}$$

In each local update, IIoT i randomly selects an individual training sample or a batch of the training samples to calculate $\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r)$. The use of current anchor gradient $\boldsymbol{g}(\boldsymbol{w}^{r-1})$ distributed to all the IIoTs in subset S^r forces the local gradient to be unbiased for the local training procedure which is formulated as $\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \boldsymbol{g}(\boldsymbol{w}^{r-1})$.

To enforce the auxiliary local gradient to be of the correct magnitude, it is scaled carefully by the number of non-zero features of the samples. The number of samples in the local data set of IIoT i with nonzero j-th feature is denoted by n_i^j. After going through their local data sets, IIoTs send the number of local nonzero j-th feature $\{n_i^j\}_{i\in E}$ to the central server, and the central server generates the number of samples with nonzero j-th feature over all local data sets as

$$n^j = \sum_{i\in E} n_i^j. \tag{11}$$

The variance between the gradient w.r.t. the current local model $\hat{\boldsymbol{w}}_{i,k-1}^r$ and global model $\hat{\boldsymbol{w}}^{r-1}$ is scaled by diagonal matrix $\boldsymbol{\Lambda}_i$ as

$$\triangle \boldsymbol{g}_{i,k-1}^r = \boldsymbol{\Lambda}_i[\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\hat{\boldsymbol{w}}^{r-1})], \tag{12}$$

where

$$\boldsymbol{\Lambda}_i = \text{diag}\left(\left\{\frac{n^j \cdot n_i}{n \cdot n_i^j}\right\}_{j=1,\cdots,q}\right). \tag{13}$$

The local updating direction is designed as

$$\boldsymbol{d}_{i,k-1}^r = -(\triangle \boldsymbol{g}_{i,k-1}^r + \boldsymbol{g}(\boldsymbol{w}^{r-1})). \tag{14}$$

The local model update of the i-th IIoT is now formulated as

$$\hat{\boldsymbol{w}}_{i,k}^r = \hat{\boldsymbol{w}}_{i,k-1}^r + \alpha_l \boldsymbol{d}_{i,k-1}^r. \tag{15}$$

Carrying (15) iteratively K times leads to a formula below for the local model update at IIoT i

$$\hat{\boldsymbol{w}}_{i,K}^r = \hat{\boldsymbol{w}}^{r-1} + \sum_{k=1}^{K} \alpha_l \boldsymbol{d}_{i,k-1}^r. \tag{16}$$

With (16) accomplished, the IIoTs send $\{\hat{\boldsymbol{w}}_{i,K}^r\}_{i\in S^r}$ to the central server. Then, the drift from the current global model

$$\boldsymbol{\xi}_r = \sum_{i\in S^r} \frac{n_i}{n^r} \sum_{k=1}^{K} \alpha_l \boldsymbol{d}_{i,k-1}^r \tag{17}$$

is scaled based on whether the specific feature appears in one local data set or not. The intuition behind the scaling is that the fewer local datasets a particular feature appears in, the more we aim to amplify the gradient update associated with that feature. The scaling diagonal matrix for model aggregation is defined as

$$\boldsymbol{A}_r = \text{diag}\left(\left\{\frac{|S^r|}{\omega^j}\right\}_{j=1,\cdots,q}\right), \tag{18}$$

where ω^j denotes the number of IIoTs containing data samples with nonzero j-th feature and is given by

$$\omega^j = \sum_{i\in E} 1_{n_i^j \neq 0}. \tag{19}$$

Finally, the global model update is given by

$$\boldsymbol{w}^r = \hat{\boldsymbol{w}}^{r-1} + \frac{\alpha_g^{r-1}}{SK} \boldsymbol{A}_r \boldsymbol{\xi}_r. \tag{20}$$

D. Auxiliary Gradient Analysis

Based on Eq. (10), the stochastic update in IIoT i yields an unbiased estimate of the global gradient $\boldsymbol{\nabla f(w)}$ as

$$\boldsymbol{\nabla} f(\hat{\boldsymbol{w}}^{r-1}) \approx \mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + E[\boldsymbol{g(w^{r-1})}]].$$

Since $\mathbb{E}[\boldsymbol{g(w^{r-1})}]$ is constant during the local updating, we can write

$$\begin{aligned} &\text{Var}(\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \mathbb{E}[\boldsymbol{g(w^{r-1})}]) \\ &= \text{Var}(\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})), \end{aligned} \tag{21}$$

where

$$\begin{aligned} &\text{Var}(\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})) \\ &= \mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})||^2] \\ &- ||\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})]||^2, \end{aligned} \tag{22}$$

and hence

$$\begin{aligned} &\text{Var}(\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})) \\ &\leq \mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})||^2]. \end{aligned} \tag{23}$$

According to Jensen's inequality, we can obtain

$$\begin{aligned} &\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})||^2] \\ &\leq ||\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r)] - E[\boldsymbol{g}_i(\boldsymbol{w}^{r-1})]||^2. \end{aligned} \tag{24}$$

Since the stochastic local gradients are unbiased to the full local gradients in ACFMO, we have

$$\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{g}_i(\boldsymbol{w}^*)] = \boldsymbol{\nabla} f_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{\nabla} f_i(\boldsymbol{w}^*).$$

Thus,

$$\begin{aligned} &||\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r)] - \mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1})]||^2 \\ &= ||\boldsymbol{\nabla} f_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \boldsymbol{\nabla} f_i(\boldsymbol{w}^{r-1})||^2 \\ &\leq L_i^2 ||\hat{\boldsymbol{w}}_{i,k-1}^r - \boldsymbol{w}^{r-1}||^2. \end{aligned} \tag{25}$$

The effectiveness of the central acceleration will be demonstrated through our simulations in Section V and its theoretical validity is supported by a convergence analysis in the Appendix A.

V. Experiments

A. Local Training and Test Datasets Design

To evaluate the performance of our proposed ACFMO, we conducted experiments on the MNIST and CIFAR-10 datasets, which are widely used benchmarks in image classification tasks. These datasets are particularly relevant for ubiquitous intelligence applications that require robust image processing, such as healthcare diagnostics, smart city surveillance, and autonomous vehicle navigation. The non-IID nature of the data distributions in these experiments reflects the real-world challenges faced in these applications, where local datasets collected by distributed devices often vary significantly.

For the MNIST dataset, which contains 10 categories of handwritten digit images, we distributed the training samples among $|E| = 400$ IIoTs following a non-IID distribution to simulate real-world heterogeneity. This setup mimics scenarios like handwritten form recognition in healthcare or education systems, where different devices capture images with varying styles and skewed label distributions. To model non-IID data,

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3546672

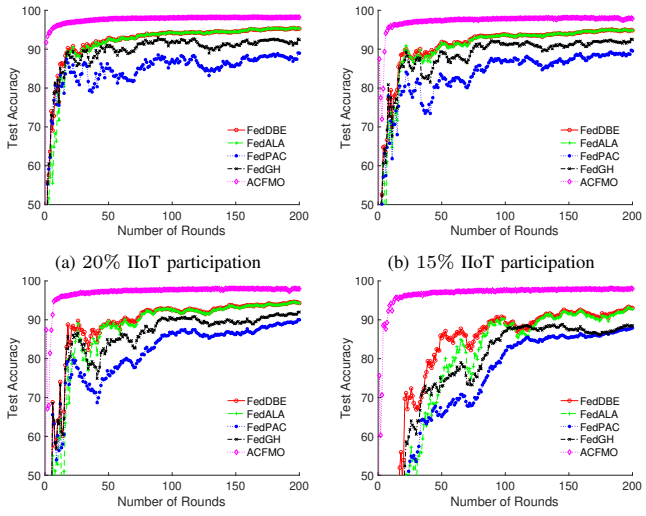
we used a symmetric Dirichlet distribution [32] with a parameter $\xi = 0.5$, which controls the degree of heterogeneity. A smaller ξ increases label skewness, reflecting more extreme variability in local data. For each IIoT i, a random vector $\boldsymbol{\rho}_i$ was drawn from $\boldsymbol{\rho}_i \sim \text{Dir}(\xi)$, and the k-th element of $\boldsymbol{\rho}_i$ determined the proportion of samples in category k assigned to the IIoT. We applied the Histogram of Gradients (HoG) method [33] for feature extraction. Using a block size of 7 and a stride of 3, HoG generated 64 blocks per sample, with gradient angles from 0 to 2π divided into 9 bins. The magnitudes of gradients were assigned to the bins based on their angles, reducing the original 784 features of each MNIST sample to 576 compact and informative HoG features. For the CIFAR-10 dataset, containing 10 categories of natural images, we expanded the IIoT pool to $|E| = 1,000$ and followed a similar Dirichlet-based data distribution. This dataset aligns with use cases such as real-time object recognition in autonomous systems or surveillance applications in smart cities. The dynamic and non-IID data distribution among IIoTs simulated scenarios where devices like drones or cameras collect data with varying perspectives and class distributions.

Both datasets were processed using a softmax regression model for multi-class classification. At each global training round r, the server distributed the global model to a subset of IIoTs, S^r, sampled at proportions $\{20\%, 15\%, 10\%, 5\%\}$ for MNIST and $\{10\%, 20\%, 40\%, 80\%\}$ for CIFAR-10. This dynamic sampling reflects the varying availability of devices in ubiquitous intelligence scenarios. IIoTs performed K local iterations of stochastic gradient updates, with the learning rate for IIoT i set as $\alpha_l^i = \frac{\alpha_l}{n_i}$ to neutralize differences in local data sizes, where $\alpha_l = 0.02$. To prevent overfitting, L_2 regularization with a coefficient of 0.01 was applied.

The trained global model was evaluated after each round using the full test sets of MNIST and CIFAR-10. To provide comprehensive insights into performance, we used both macro-averaged and micro-averaged metrics. Macro-averaging assessed performance across all classes equally, highlighting the model's ability to handle minority classes effectively, while micro-averaging prioritized performance on frequent classes. These metrics are critical for applications like healthcare diagnostics, where minority classes must be identified accurately, and smart city monitoring, where majority classes dominate. We compared ACFMO with state-of-the-art FL algorithms, including FedPAC [13], FedDBE [15], FedALA [14], and FedGH [16]. ACFMO consistently outperformed these methods, achieving higher test accuracy and better robustness under diverse conditions. Its ability to address data heterogeneity and dynamic IIoT participation makes it particularly well-suited for image processing tasks in ubiquitous intelligence applications, where efficient and scalable federated optimization is essential.

B. Experimental Evaluation on the MNIST Dataset

We first evaluate the performance of the proposed ACFMO algorithm against state-of-the-art FL methods, including FedDBE, FedALA, FedPAC, and FedGH, under varying levels of IIoT participation. The evaluation reflects real-world



(a) 20% IIoT participation (b) 15% IIoT participation

(c) 10% IIoT participation (d) 5% IIoT participation

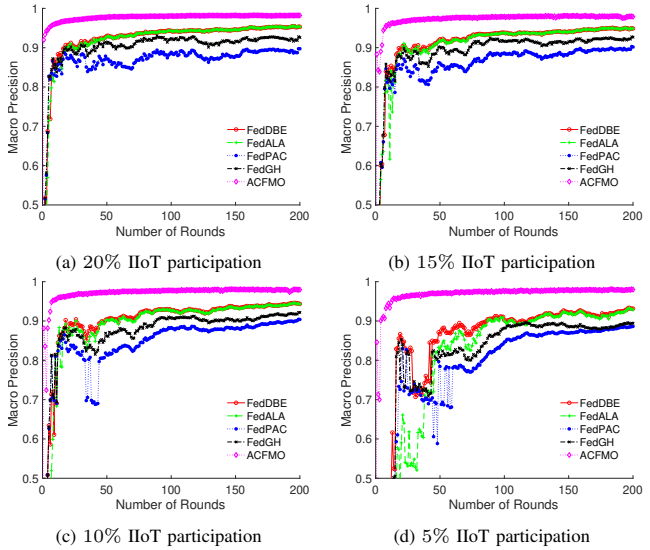
Fig. 1. Performance comparison with test accuracy by decreasing participated IIoTs where $\beta_1 = 0.9$, $\beta_2 = 0.999$.

ubiquitous intelligence scenarios where intelligent devices, such as edge nodes and sensors, contribute dynamically to training, leading to challenges such as reduced participation and high variance in updates. Each algorithm is trained over 200 communication rounds, with test accuracy recorded to assess convergence speed and final accuracy. The momentum-based parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

At 20% participation, all algorithms benefit from sufficient IIoT contributions. ACFMO achieves near-optimal accuracy of 96% within 50 rounds, outperforming all baselines in both convergence speed and accuracy. FedDBE and FedALA perform competitively but require close to 100 rounds to stabilize at around 93 to 94%. FedPAC and FedGH lag noticeably, converging slower and stabilizing below 90% accuracy. When the participation rate drops to 15%, the performance gap widens. ACFMO maintains its efficiency, achieving 96% accuracy within 60 rounds. FedDBE and FedALA continue to converge but stabilize later, reaching slightly lower accuracies of 93 to 94%. In contrast, FedPAC and FedGH show slower convergence and remain below 90%, struggling with the effects of reduced updates and higher gradient variance. At 10% participation, the challenges of dynamic participation and data heterogeneity become more evident. ACFMO demonstrates remarkable robustness, converging to over 95% accuracy within 100 rounds. FedDBE and FedALA degrade noticeably, requiring significantly more rounds to approach 90 to 92%. FedPAC and FedGH exhibit further performance loss, stagnating at 85 to 88% accuracy with unstable convergence. Under the extreme condition of 5% participation, the limitations of baseline methods are amplified. FedDBE and FedALA stabilize at 87 to 89%, while FedPAC and FedGH fail to surpass 85%, exhibiting slow convergence and fluctuating accuracy. In sharp contrast, ACFMO achieves 94% accuracy, highlighting its ability to efficiently handle sparse participation and mitigate the variance introduced by limited IIoT updates.

The performance of ACFMO is next evaluated on the MNIST dataset using Macro-Averaged Precision as the evalu-

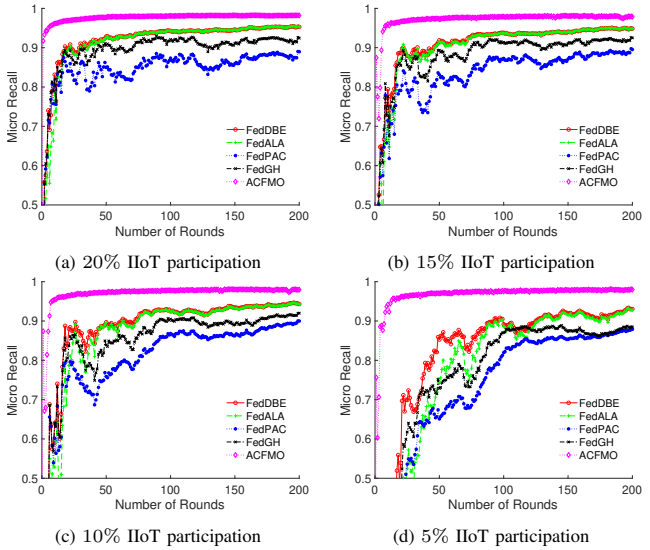
This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3546672



(a) 20% IIoT participation (b) 15% IIoT participation

(c) 10% IIoT participation (d) 5% IIoT participation

Fig. 2. Performance comparison with Macro-Averaged Precision by decreasing participated IIoTs where $\beta_1 = 0.9$, $\beta_2 = 0.999$.



(a) 20% IIoT participation (b) 15% IIoT participation

(c) 10% IIoT participation (d) 5% IIoT participation

Fig. 3. Performance comparison with Micro-Averaged Recall by decreasing participated IIoTs where $\beta_1 = 0.9$, $\beta_2 = 0.999$.

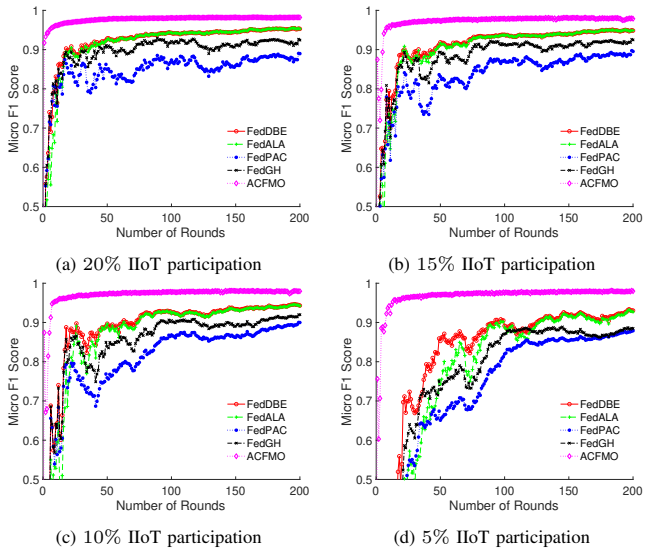
ation metric. This metric treats all classes equally, providing a balanced assessment of performance, particularly under non-IID data distributions where minority classes may otherwise be underrepresented. We assess the results under participation levels of 20%, 15%, 10%, and 5%, reflecting the dynamic and sparse IIoT availability commonly encountered in ubiquitous intelligence applications.

At 20% participation, as shown in Fig. 2(a), ACFMO achieves a precision close to 0.95 within the first 50 rounds. This result demonstrates its ability to rapidly converge while maintaining superior accuracy. In comparison, the baseline methods exhibit slower convergence, stabilizing around 93 to 94% after nearly 100 rounds, while others struggle to exceed 90%. When the participation rate decreases to 15% in Fig. 2(b), the advantage of ACFMO becomes more pronounced. It reaches approximately 94% precision within 60 rounds, maintaining stable and efficient convergence. In contrast, the baseline methods stabilize later, with slight fluctuations, and fail to achieve comparable precision. At 10% participation, illustrated in Fig. 2(c), the challenges of sparse IIoT updates and increased gradient variance become evident. Despite these factors, ACFMO remains robust, converging to approximately 93% precision within 100 rounds. The baselines experience noticeable degradation, with precision values stabilizing below 90%, while their convergence becomes less consistent. In the extreme case of 5% participation, illustrated in Fig. 2(d), the limitations of the baselines are magnified. Precision values plateau around 87 to 88%, with significant fluctuations, highlighting their inability to cope with sparse and heterogeneous updates. In contrast, ACFMO achieves an impressive precision of approximately 92%, demonstrating exceptional stability and resilience under highly limited IIoT participation.

The performance of ACFMO is further assessed using Micro-Averaged Recall on the MNIST dataset. Unlike Macro-Averaged Precision, this metric aggregates contributions across all classes proportionally, emphasizing the global performance of the model, particularly on majority classes. This evaluation provides additional insights into how effectively ACFMO maintains overall model quality under dynamic participation levels, a common scenario in ubiquitous intelligence applications. Each method is evaluated over 200 communication rounds.

At 20% participation, as seen in Fig. 3(a), ACFMO demonstrates fast and stable convergence, reaching a recall value close to 0.95 within the first 50 rounds. This reflects its strong capacity to leverage available IIoT updates efficiently. In contrast, the baseline methods show slower convergence, with some requiring nearly twice the number of rounds to stabilize. FedDBE and FedALA eventually achieve recall values around 0.92 to 0.93, while the remaining methods struggle to cross the 0.90 mark. The benefits of ACFMO become even clearer when participation drops to 15%, as illustrated in Fig. 3(b). Here, ACFMO maintains its upward trend, converging to approximately 0.94 recall within 60 rounds. Although FedDBE and FedALA remain competitive, their convergence slows significantly, stabilizing near 0.91. The other baselines encounter noticeable instability, with recall values plateauing below 0.88. At 10% participation, shown in Fig. 3(c), the limitations of the baseline methods under reduced IIoT updates become more pronounced. ACFMO remains resilient, reaching a stable recall of about 0.93 within 100 rounds, despite the increased gradient variance caused by sparse participation. FedDBE and FedALA converge much later and struggle to exceed 0.90 recall. FedPAC and FedGH, on the other hand, exhibit pronounced fluctuations and stagnate at values between 0.85 and 0.87, highlighting their inability to adapt to fewer updates. In the extreme case of 5% participation, illustrated in Fig. 3(d), the gap between ACFMO and the baseline methods becomes significant. While FedDBE and FedALA plateau at recall values around 0.88 to 0.89, they show noticeable instability across rounds. FedPAC and FedGH fail to deliver meaningful improvements, stabilizing below 0.85. ACFMO, however, achieves a consistent recall of approximately 0.92,

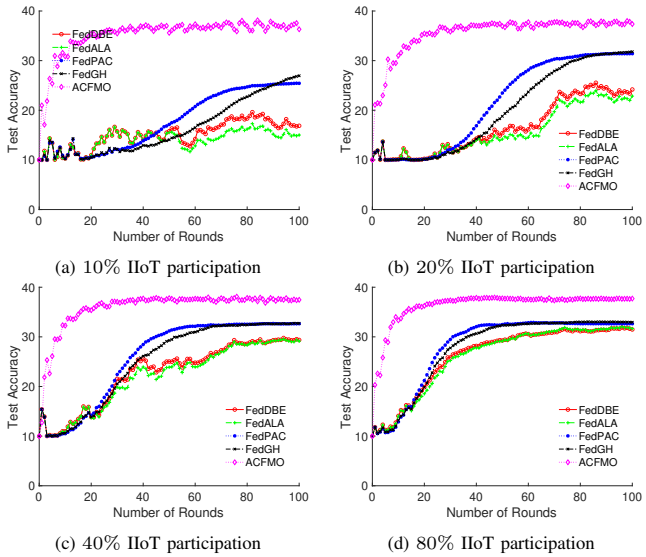
This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3546672



(a) 20% IIoT participation (b) 15% IIoT participation

(c) 10% IIoT participation (d) 5% IIoT participation

Fig. 4. Performance comparison with Micro-Averaged F1 Score by decreasing participated IIoTs where $\beta_1 = 0.9$, $\beta_2 = 0.999$.



(a) 10% IIoT participation (b) 20% IIoT participation

(c) 40% IIoT participation (d) 80% IIoT participation

Fig. 5. Performance comparison with test accuracy by increasing participated IIoTs where $\beta_1 = 0.9$, $\beta_2 = 0.999$.

proving its ability to adapt efficiently and maintain reliable performance even under severe participation sparsity.

Finally, we evaluate the performance of ACFMO on the MNIST dataset using Micro-Averaged F1 Score, which balances precision and recall across all classes. This metric is particularly valuable for assessing the global performance of FL systems, especially under dynamic participation levels.

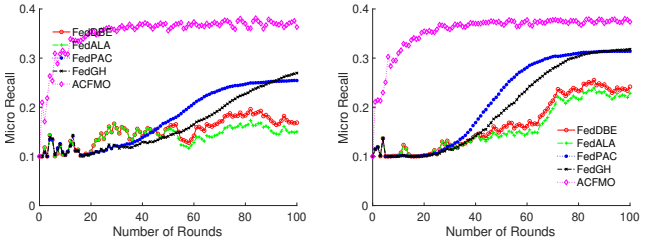
At 20% participation, as shown in Fig. 4(a), ACFMO achieves a Micro-Averaged F1 Score close to 0.95 within the first 50 rounds, demonstrating its ability to converge rapidly and maintain robust performance. In comparison, FedDBE and FedALA stabilize at around 0.92 to 0.93 but require nearly twice the number of rounds to reach this level. FedPAC and FedGH show slower improvements, stabilizing below 0.90, with noticeable instability during the earlier rounds. As participation drops to 15%, illustrated in Fig. 4(b), ACFMO continues to outperform the baselines, converging efficiently to a score of approximately 0.94 within 60 rounds. FedDBE and FedALA remain competitive but exhibit slower convergence, stabilizing near 0.91 to 0.92. FedPAC and FedGH, however, face significant challenges, struggling to exceed 0.88 and displaying greater fluctuations due to reduced IIoT updates and increased gradient variance. At 10% participation, illustrated in Fig. 4(c), the effects of sparse participation become more pronounced. ACFMO demonstrates strong resilience, achieving a Micro-Averaged F1 Score of 0.93 within 100 rounds. In contrast, FedDBE and FedALA stabilize later with scores just above 0.90, while FedPAC and FedGH stagnate at lower scores between 0.85 to 0.87. The high gradient variance and limited updates further amplify the challenges for the baselines, hindering their convergence. In the extreme case of 5% participation, as shown in Fig. 4(d), the degradation in performance for the baseline methods becomes evident. FedDBE and FedALA plateau around 0.88, but their results exhibit considerable instability. FedPAC and FedGH perform the worst, failing to achieve scores above 0.85 and suffering from severe fluctuations throughout the training process. ACFMO, on the other hand, maintains remarkable stability and efficiency, converging to a Micro-Averaged F1 Score of approximately 0.92, even under highly sparse and dynamic participation conditions.

C. Experimental Evaluation on the CIFAR-10 Dataset

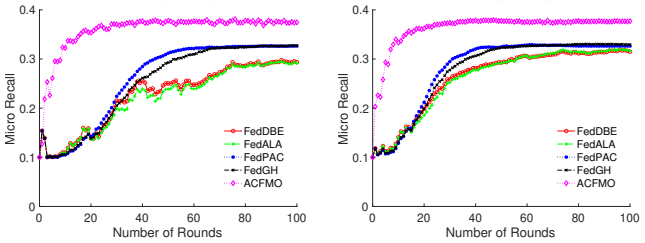
The next experiment evaluates the performance of ACFMO on the CIFAR-10 dataset using Test Accuracy as the evaluation metric. Compared to MNIST, CIFAR-10 presents a greater challenge due to its higher feature complexity and the difficulty of image classification tasks. The participation levels are set at 10%, 20%, 40%, and 80%, simulating realistic IIoT availability scenarios observed in ubiquitous intelligence applications.

At 10% participation, as shown in Fig. 5(a), ACFMO demonstrates exceptional performance, achieving a test accuracy of approximately 38% within 60 rounds. FedGH follows as the closest baseline but stabilizes at a significantly lower accuracy of around 30%. Meanwhile, FedDBE, FedALA, and FedPAC struggle to adapt to this sparse participation setting, plateauing below 25%. The results highlight the superior robustness of ACFMO in handling limited data contributions and increased gradient variance. As the participation rate increases to 20%, illustrated in Fig. 5(b), ACFMO continues to lead, converging quickly to an accuracy of nearly 39% within 50 rounds. FedGH improves its performance under this setting, reaching approximately 34%, but remains behind ACFMO both in speed and accuracy. FedDBE, FedALA, and FedPAC show slight improvements, stabilizing between 28% and 30%. These results emphasize that while higher participation benefits all methods, ACFMO leverages the additional contributions more effectively. At 40% participation, illustrated in Fig. 5(c), the overall performance of all algorithms improves due to greater IIoT involvement. ACFMO reaches an accuracy of 40% within just 40 rounds, maintaining its rapid convergence and stability. FedGH demonstrates competitive performance, stabilizing at around 35%. In contrast, FedDBE,

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3546672

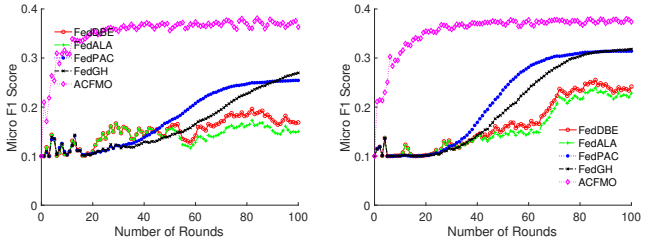


(a) 10% IIoT participation (b) 20% IIoT participation

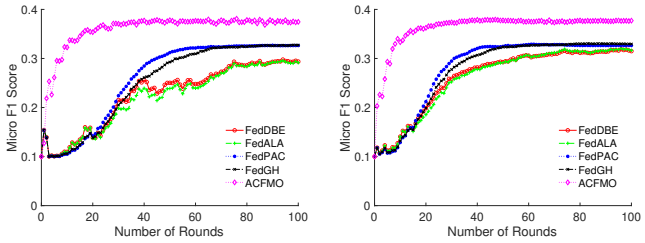


(c) 40% IIoT participation (d) 80% IIoT participation

Fig. 6. Performance comparison with Micro-Averaged Recall by increasing participated IIoTs where $\beta_1 = 0.9$, $\beta_2 = 0.999$.



(a) 10% IIoT participation (b) 20% IIoT participation



(c) 40% IIoT participation (d) 80% IIoT participation

Fig. 7. Performance comparison with Micro-Averaged F1 Score by increasing participated IIoTs where $\beta_1 = 0.9$, $\beta_2 = 0.999$.

FedALA, and FedPAC exhibit slower convergence and settle at lower accuracies between 30% and 32%, reflecting their inefficiency in fully utilizing the increased IIoT updates. Finally, at 80% participation, shown in Fig. 5(d), all methods achieve higher accuracies due to near-complete IIoT availability. ACFMO maintains its performance advantage, achieving approximately 41% accuracy within 30 rounds. FedGH stabilizes at around 37%, narrowing the gap but still unable to match ACFMO's efficiency. FedDBE, FedALA, and FedPAC show moderate improvements, converging between 33% and 35%, but their slower convergence rates and lower accuracies remain evident.

Then we evaluate the performance of ACFMO on the CIFAR-10 dataset using Micro-Averaged Recall, a key metric that aggregates true positive predictions across all classes as shown in Fig. 6. At the lowest participation level of 10%, as shown in Fig. 6(a), ACFMO demonstrates a clear advantage by achieving a recall value of approximately 0.35 within 70 rounds. FedGH follows as the next-best performer but stabilizes around 0.28 after requiring more rounds for convergence. FedDBE, FedALA, and FedPAC, however, struggle to adapt under such sparse conditions, plateauing below 0.2. This result underscores ACFMO's resilience in managing high variance and sparse participation while maintaining retrieval consistency. With an increase in participation to 20%, illustrated in Fig. 6(b), ACFMO achieves a recall of 0.36 within 60 rounds, further reducing its convergence time. FedGH also benefits from the improved participation, stabilizing at approximately 0.30, though it remains behind ACFMO. In comparison, FedDBE, FedALA, and FedPAC show only slight improvements, converging between 0.22 and 0.25. These results highlight ACFMO's efficiency in aggregating IIoT contributions, even under moderately constrained participation. As participation rises to 40%, illustrated in Fig. 6(c), all methods show noticeable improvements, though ACFMO maintains its dominance. It converges rapidly to a recall value of around 0.38 within 50 rounds. FedGH improves to 0.33 but stabilizes more slowly, while FedDBE, FedALA, and FedPAC plateau between 0.25 and 0.28, reflecting their inability to fully capitalize on the increased availability of IIoT updates. At the highest participation level of 80%, shown in Fig. 6(d), ACFMO reaches a recall value of approximately 0.39 within just 40 rounds, maintaining its lead in both convergence speed and overall performance. FedGH performs competitively, achieving a stable recall of around 0.35, while FedDBE, FedALA, and FedPAC remain slower to converge and stabilize between 0.28 and 0.31. Despite the higher participation, ACFMO continues to exhibit superior efficiency and robustness.

The performance of ACFMO is further evaluated on the CIFAR-10 dataset using Micro-Averaged F1 Score. At 10% participation, as shown in Fig. 7(a), ACFMO demonstrates remarkable performance, achieving a Micro-Averaged F1 Score of approximately 0.34 within 70 rounds. FedGH follows as the second-best performer with a score of around 0.28, while FedDBE, FedALA, and FedPAC perform poorly, stabilizing below 0.2. ACFMO's ability to mitigate high variance and imbalance at this sparse participation level highlights its robustness and efficiency. When participation increases to 20%, illustrated in Fig. 7(b), ACFMO maintains its dominance, converging to a score of approximately 0.36 within 60 rounds. FedGH improves its performance to around 0.30, though it continues to lag behind ACFMO. FedDBE, FedALA, and FedPAC show marginal improvements but stabilize between 0.22 and 0.25, reflecting their limited adaptability to moderate participation levels. At 40% participation, illustrated in Fig. 7(c), the overall performance improves across all methods due to greater IIoT contributions. ACFMO achieves a Micro-Averaged F1 Score of nearly 0.38 within 50 rounds, demonstrating both rapid convergence and superior accuracy. FedGH stabilizes at approximately 0.33, while FedDBE, FedALA, and FedPAC plateau between 0.25 and 0.28, highlighting their inefficiency in balancing precision and recall effectively under increased IIoT participation. At the highest participation level of 80%, shown in Fig. 7(d), ACFMO reaches a score of approximately

0.39 within 40 rounds, maintaining its clear advantage over the baselines. FedGH performs relatively well, converging to 0.35, though it remains slower to stabilize. FedDBE, FedALA, and FedPAC exhibit marginal improvements but fail to surpass 0.30, demonstrating their limited capacity to optimize performance in more favorable conditions with heterogeneous data contributions.

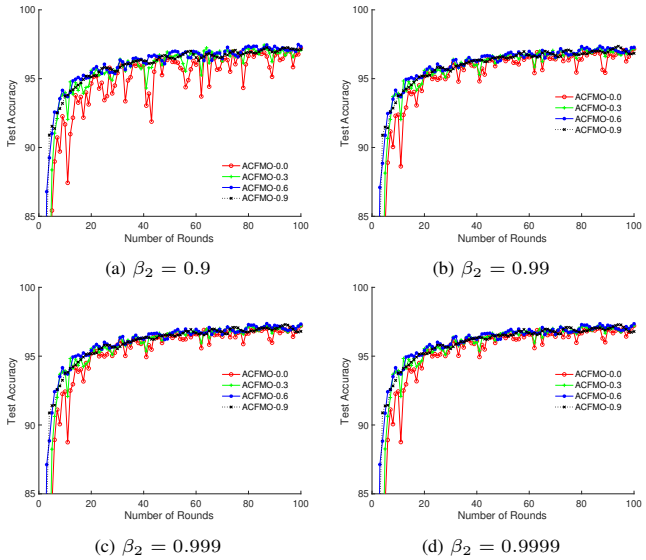
D. Evaluation with Different Momentum Settings

To evaluate the impact of exponential decay rates on ACFMO's performance, we examine different configurations of the first-order gradient momentum decay rate, β_1, using values $\{0.0, 0.3, 0.6, 0.9\}$, referred to as ACFMO-0.0, ACFMO-0.3, ACFMO-0.6, and ACFMO-0.9, respectively. Additionally, we vary the second-order gradient momentum decay rate, β_2, across $\{0.9, 0.99, 0.999, 0.9999\}$ to assess its effect on training stability. The analysis is conducted with 1% IIoT device participation, representing an extreme case of sparse participation, and the results are illustrated in Fig. 8 and Fig. 9.

As shown in Fig. 8, ACFMO-0.0 with $\beta_2 = 0.9$ exhibits the lowest test accuracy and the least stable performance, confirming that the absence of first-order momentum degrades convergence. As β_1 increases from 0.0 to 0.9, test accuracy improves, demonstrating the role of first-order momentum in stabilizing updates and reducing variance. Similarly, increasing β_2 from 0.9 to 0.9999 further enhances accuracy, particularly for higher β_1 values, reinforcing the significance of second-order momentum in maintaining stable learning under sparse participation.

A similar trend is observed in Fig. 9, where higher β_1 and β_2 values result in a smoother decline in test loss. ACFMO-0.0 shows noticeable fluctuations in cross-entropy loss, indicating that the lack of momentum introduces instability in optimization. As β_1 increases, the loss curve becomes smoother, highlighting the advantage of accumulating historical gradients to suppress high-variance updates caused by sparse device participation and non-IID data distributions. Meanwhile, higher β_2 values contribute to greater gradient consistency, reducing fluctuations in the adaptive learning rate and improving convergence stability.

These empirical results validate the theoretical role of momentum hyperparameters in balancing stability and adaptability in FL. The first-order momentum parameter, β_1, controls the exponential moving average of past gradients, filtering noise and promoting smoother updates. When set too low, e.g., $\beta_1 = 0.0$, ACFMO struggles with instability, particularly under high-variance conditions caused by sparse participation. Conversely, excessively high β_1 values slow adaptation, as the model becomes overly reliant on past updates. Our results suggest that an optimal range for β_1 lies between 0.6 and 0.9, depending on participation levels. The second-order momentum parameter, β_2, influences how past squared gradients adjust the adaptive learning rate, affecting convergence dynamics. Lower β_2 values, e.g., $\beta_2 = 0.9$, allow for faster adaptation by making the learning rate more responsive to recent updates, which can be beneficial in environments with highly dynamic participation. However, this comes at the cost of increased variance in updates. In contrast, higher β_2 values, e.g., $\beta_2 = 0.999$ or above, stabilize learning by dampening large gradient fluctuations, making them particularly effective in settings with stable yet heterogeneous data distributions.



(a) $\beta_2 = 0.9$ (b) $\beta_2 = 0.99$

(c) $\beta_2 = 0.999$ (d) $\beta_2 = 0.9999$

Fig. 8. Performance comparison with test accuracy by increasing correction for the second-order gradient momentum with 1% IIoT participation.

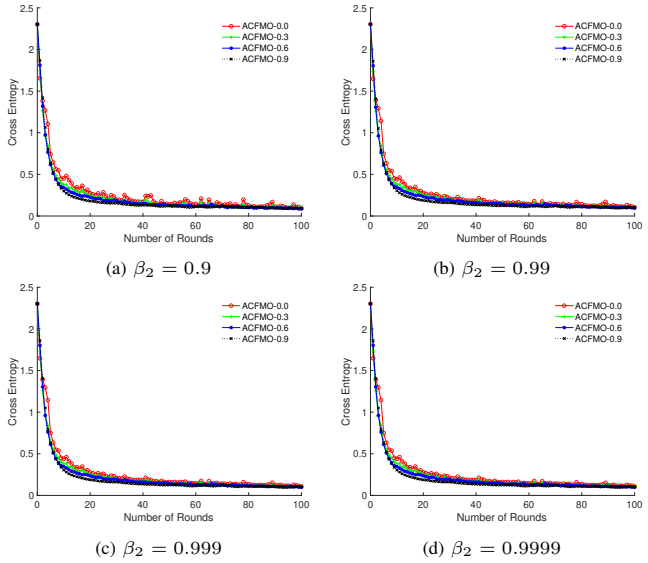
For practical momentum tuning, our results suggest that in high participation and stable environments, setting β_1 between 0.8 and 0.9, along with β_2 close to 0.99 or 0.999, ensures stable updates and minimizes unnecessary fluctuations. In highly dynamic or resource-constrained settings, reducing β_1 to 0.5-0.7 allows the model to remain adaptive to recent gradient updates, while setting β_2 between 0.9 and 0.95 enables faster adaptation to evolving learning conditions without excessive instability. In extreme cases of low participation, such as this experiment with 1% IIoT devices, a combination of $\beta_1 \approx 0.6$ and $\beta_2 \approx 0.99$ provides a balance between adaptability and robustness. These findings highlight ACFMO's ability to handle varying participation rates and data heterogeneity, ensuring that momentum tuning can be effectively leveraged to optimize learning in different FL environments.

VI. Conclusion

This paper proposed ACFMO, an optimization framework designed to enhance the efficiency and stability of FL in IIoT environments. By integrating an adaptive central acceleration mechanism with variance-controlled local updates, ACFMO effectively addresses key challenges such as sparse client participation, local data heterogeneity, and communication constraints. The adaptive momentum-based optimization at the central server stabilizes global updates, while variance-controlled local updates ensure consistent and meaningful contributions from participating devices. Experimental results demonstrate that ACFMO significantly accelerates convergence, reduces communication overhead, and improves model stability, consistently outperforming existing FL methods, even under dynamic participation conditions.

Despite these advantages, ACFMO relies on a central server for model aggregation and momentum-based optimization,

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3546672



(a) $\beta_2 = 0.9$ (b) $\beta_2 = 0.99$

(c) $\beta_2 = 0.999$ (d) $\beta_2 = 0.9999$

Fig. 9. Performance comparison with test loss by increasing correction for the first-order gradient momentum with 1% IIoT participation.

which may limit its applicability in fully decentralized FL frameworks such as peer-to-peer systems. In future work, extending ACFMO to decentralized settings by designing an adaptive aggregation mechanism that eliminates the need for a central server could further enhance its scalability. Additionally, exploring more advanced techniques for handling extreme resource constraints and system heterogeneity would strengthen ACFMO's adaptability across diverse FL applications. By addressing these challenges, ACFMO provides a scalable and efficient solution for real-world IIoT deployments where device availability is dynamic, and communication resources are limited.

APPENDIX A

A. Properties of Objective Functions

The properties of the global objective function depend on the properties of the local objective functions. We start by introducing a linear approximation of the i-th local objective $f_i(\boldsymbol{w})$ in the k-th local updating and r-th global round as

$$\hat{f}_i(\hat{\boldsymbol{w}}_{i,k}^r) = f_i(\hat{\boldsymbol{w}}_{i,k-1}^r) + \nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r)^T(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r). \tag{26}$$

where $\hat{\boldsymbol{w}}_{i,k}^r$ represents the i-th local model in the r-th global round and the k-th local iteration. Since

$$\begin{aligned} &f_i(\hat{\boldsymbol{w}}_{i,k}^r) - f_i(\hat{\boldsymbol{w}}_{i,k-1}^r) \\ &= \int_0^1 f_i(\hat{\boldsymbol{w}}_{i,k-1}^r + \tau(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r))d\tau. \end{aligned} \tag{27}$$

We now define an auxiliary variable $\boldsymbol{z}_k^r$ as

$$\boldsymbol{z}_k^r = \hat{\boldsymbol{w}}_{i,k-1}^r + \tau(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r) \tag{28}$$

where $0 \leq \tau \leq 1$, and notice that

$$\int_0^1 df_i(\boldsymbol{z}_k^r) = \int_0^1 \nabla f_i(\boldsymbol{z}_k^r)^T(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r)d\tau. \tag{29}$$

Combining (26), (27), and (29), we obtain

$$\begin{aligned} &f_i(\hat{\boldsymbol{w}}_{i,k}^r) = \hat{f}_i(\hat{\boldsymbol{w}}_{i,k}^r) \\ &+ \int_0^1 (\nabla f_i(\boldsymbol{z}_k^r) - \nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r))^T(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r)d\tau. \end{aligned} \tag{30}$$

Since

$$\begin{aligned} &\left|\int_0^1 (\nabla f_i(\boldsymbol{z}_k^r) - \nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r))^T(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r)d\tau\right| \\ &\leq \int_0^1 |(\nabla f_i(\boldsymbol{z}_k^r) - \nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r))^T(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r)|d\tau, \end{aligned} \tag{31}$$

and

$$\begin{aligned} &|(\nabla f_i(\boldsymbol{z}_k^r) - \nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r))^T(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r)| \\ &\leq ||(\nabla f_i(\boldsymbol{z}_k^r) - \nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r))||_2 \cdot ||(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r)||_2, \end{aligned} \tag{32}$$

and by the Lipschitz continuous gradient assumption, we have

$$\begin{aligned} &||(\nabla f_i(\boldsymbol{z}_k^r) - \nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r))||_2 \leq L_i||\boldsymbol{z}_k^r - \hat{\boldsymbol{w}}_{i,k-1}^r||_2 \\ &= L_i||\tau(\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r)||_2. \end{aligned} \tag{33}$$

Combining (30)-(33), we can upper bound $f_i(\hat{\boldsymbol{w}}_{i,k}^r)$ by

$$f_i(\hat{\boldsymbol{w}}_{i,k}^r) \leq \hat{f}_i(\hat{\boldsymbol{w}}_{i,k}^r) + \frac{L_i}{2}||\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r||^2. \tag{34}$$

Similarly, we can lower bound $f_i(\hat{\boldsymbol{w}}_{i,k}^r)$ by

$$f_i(\hat{\boldsymbol{w}}_{i,k}^r) \geq \hat{f}_i(\hat{\boldsymbol{w}}_{i,k}^r) + \frac{\mu_i}{2}||\hat{\boldsymbol{w}}_{i,k}^r - \hat{\boldsymbol{w}}_{i,k-1}^r||^2, \tag{35}$$

where μ_i refers to the lower bound of the eigenvalues of the Hessian of the local objective f_i.

With stochastic IIoT participation in each federated round, we need to reduce the variance of the local gradient from individual training sample or a batch of the training samples which is an unbiased stochastic gradient of $f_i(\hat{\boldsymbol{w}}_{i,k-1}^r)$. According to Jensen's inequality, we obtain the upper bound of the variance of the sampled local gradient as

$$E[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}_{i,k-1}^r)||^2] \leq ||\nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r)||^2, \tag{36}$$

which can be written as

$$E[||\boldsymbol{g_i}(\hat{\boldsymbol{w}}_{\boldsymbol{i,k-1}}^{\boldsymbol{r}})||^{\boldsymbol{2}}] \leq ||\nabla f_i(\hat{\boldsymbol{w}}_{i,k-1}^r) - \nabla f_i(\boldsymbol{w}^*)||^2, \tag{37}$$

where $\boldsymbol{w}^*$ denotes the optimal model. The analysis above leads to

$$E[||\boldsymbol{g_i}(\hat{\boldsymbol{w}}_{\boldsymbol{i,k-1}}^{\boldsymbol{r}})||^{\boldsymbol{2}}] \leq L_i^2||\hat{\boldsymbol{w}}_{\boldsymbol{i,k-1}}^{\boldsymbol{r}} - \boldsymbol{w^*}||^{\boldsymbol{2}}. \tag{38}$$

The eigenvalues of the global Hessian $\boldsymbol{\nabla^2} f$ are bounded by

$$\mu = \sum_{i=1}^{|E|} \frac{n_i}{n}\mu_i \leq \lambda_{\min}(\boldsymbol{\nabla^2} f) \leq \lambda_{\max}(\boldsymbol{\nabla^2} f) \leq \sum_{i=1}^{|E|} \frac{n_i}{n}L_i = L. \tag{39}$$

Consequently, if we write a linear approximation of the global objective function as

$$\hat{f}(\hat{\boldsymbol{w}}^r) = \sum_{i=1}^{|E|} \frac{n_i}{n}\hat{f}_i(\hat{\boldsymbol{w}}_{i,k}^r),$$

the global objective function is upper bounded by

$$f(\hat{\boldsymbol{w}}^r) \leq \hat{f}(\hat{\boldsymbol{w}}^r) + \frac{L}{2}||\hat{\boldsymbol{w}}^r - \hat{\boldsymbol{w}}^{r-1}||^2, \tag{40}$$

and is lower bounded by

$$f(\hat{\boldsymbol{w}}^r) \geq \hat{f}(\hat{\boldsymbol{w}}^r) + \frac{\mu}{2}||\hat{\boldsymbol{w}}^r - \hat{\boldsymbol{w}}^{r-1}||^2. \tag{41}$$

B. Convergence Analysis

The convergence analysis below aims to derive an upper bound of the expected distance w.r.t. the global round r between the values of the current objective function and minimum objective function, i.e.,

$$\mathbb{E}_r[f(\boldsymbol{w}^r) - f(\boldsymbol{w}^*)] \leq \varphi^r(f(\boldsymbol{w}^0) - f(\boldsymbol{w}^*)), \tag{42}$$

where $\varphi < 1$ is the reduction rate. There are two parts of randomness in our analysis. First, the local updating is based on unbiased stochastic local sample selection. Second, the IIoTs are randomly participated during each round. In the following analysis, the expectation is based on these two unbiased stochastic processes. With respect to round r, the expectation can be regarded as combining both the random participation of the IIoTs and the randomness in the local minimization in each selected IIoT.

The expectation overs two layers of quantities, i.e., one for random participation of IIoTs and another for the local stochastic gradient iterations can be defined as $\mathbb{E}_r[\mathbb{E}[\cdot]]$, where $\mathbb{E}$ refers to the expectation w.r.t. the local stochastic iterations and $\mathbb{E}_r$ refers to the random participation of IIoTs. For example,

$$f(\boldsymbol{w}) = \frac{1}{|E|}\sum_{i=1}^{|E|} f_i(\boldsymbol{w}), \tag{43}$$

the sample space for random IIoT participation in each round is $\{i, i \in E\}$, and we can obtain the following expectation over r rounds

$$\mathbb{E}_r[f_i(\boldsymbol{w})_{i\in E}] = f(\boldsymbol{w}), \tag{44}$$

which leads to

$$\mathbb{E}_r[\nabla f_i(\boldsymbol{w})_{i\in E}] = \nabla f(\boldsymbol{w}). \tag{45}$$

During the stochastic local updating, $\boldsymbol{g}_i(\boldsymbol{w})$ is the unbiased local gradient estimation, which leads to

$$\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w})] = \nabla f_i(\boldsymbol{w}), \tag{46}$$

where the sample space is the local data sets. With the random participation over r rounds and combining (45) and (46), we can obtain the following expectation

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w})]] = \mathbb{E}_r[\nabla f_i(\boldsymbol{w})] = \nabla f(\boldsymbol{w}). \tag{47}$$

Notice that the $\boldsymbol{w}$ in (44)-(47) are general definition, both the local model $\hat{\boldsymbol{w}}^r_{i,k-1}$ and the global model $\boldsymbol{w}^{r-1}$ can be regarded as special cases of $\boldsymbol{w}$.

According to (47) and set $\boldsymbol{w} = \hat{\boldsymbol{w}}^r_{i,k-1}$, we get

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1})]] = \nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}), \tag{48}$$

and set $\boldsymbol{w} = \boldsymbol{w}^{r-1}$, we obtain

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1})]] = \nabla f(\boldsymbol{w}^{r-1}), \tag{49}$$

which leads to

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})]] = \nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}). \tag{50}$$

Using (15) and (50), we have the expected distance from the current local model $\hat{\boldsymbol{w}}^r_{i,k}$ to the optimal model $\boldsymbol{w}^*$ w.r.t. round r as

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k} - \boldsymbol{w}^*||^2]] = \mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k-1} - \boldsymbol{w}^*||^2]]\\ &- 2\alpha_l\mathbb{E}_r[\mathbb{E}[(\hat{\boldsymbol{w}}^r_{i,k-1} - \boldsymbol{w}^*)]]^T\nabla f(\hat{\boldsymbol{w}}^r_{i,k-1})\\ &+ \alpha_l^2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]].\end{aligned} \tag{51}$$

First, we analyze the expected upper bound of the third term in the right-hand side of (51) w.r.t. the round number r. We apply the Cauchy-Schwarz inequality to write

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]]\\ &\leq 2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]]\\ &+ 2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^*) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]].\end{aligned} \tag{52}$$

To continue to provide upper bound of the right-hand side of (52), we have the following analysis. By writing

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}) = f(\boldsymbol{w}^*) - f(\boldsymbol{\beta}) + f(\boldsymbol{\beta}) - f(\boldsymbol{w}), \tag{53}$$

where $\boldsymbol{\beta}$ is an auxiliary variable to be specified shortly and using (34), we can write

$$\begin{aligned}&f(\boldsymbol{w}^*) - f(\boldsymbol{w}) \leq \nabla f(\boldsymbol{w}^*)^T(\boldsymbol{w}^* - \boldsymbol{w})\\ &+ (\nabla f(\boldsymbol{w}^*) - \nabla f(\boldsymbol{w}))^T(\boldsymbol{w} - \boldsymbol{\beta}) + \frac{L}{2}||\boldsymbol{\beta} - \boldsymbol{w}||^2.\end{aligned} \tag{54}$$

If we let

$$\boldsymbol{\beta} = \boldsymbol{w} - \frac{1}{L}(\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{w}^*)), \tag{55}$$

and substitute (55) into (54), we obtain

$$\begin{aligned}&f(\boldsymbol{w}^*) - f(\boldsymbol{w}) \leq \nabla f(\boldsymbol{w}^*)^T(\boldsymbol{w}^* - \boldsymbol{w})\\ &- \frac{1}{2L}||(\nabla f(\boldsymbol{w}^*) - \nabla f(\boldsymbol{w}))||^2.\end{aligned} \tag{56}$$

Using (47), we have

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)]] = \nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}) - \nabla f(\boldsymbol{w}^*),$$

and combining with (24), we have

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})||^2]]\\ &\leq ||\nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}) - \nabla f(\boldsymbol{w}^*)||^2.\end{aligned} \tag{57}$$

Using (56) and (57), we obtain

$$2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]] \leq 4L(f(\hat{\boldsymbol{w}}^r_{i,k-1}) - f(\boldsymbol{w}^*)). \tag{58}$$

Using (47), we have

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)]] = \nabla f(\boldsymbol{w}^{r-1}), \tag{59}$$

which leads to

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^*) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]]\\ &= \mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*) - \mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)]]||^2]]\\ &= \mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]]\\ &- ||\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)]]||^2\\ &\leq \mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]].\end{aligned} \tag{60}$$

Similar to the conclusion in (58), using (60), we have

$$\begin{aligned}&2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^*) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]]\\ &\leq 4L(f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)).\end{aligned} \tag{61}$$

Combining (52), (58) and (61), we have

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1})-\boldsymbol{g}_i(\boldsymbol{w}^{r-1})+\nabla f(\boldsymbol{w}^{r-1})||^2]]\\&\leq 4L(f(\hat{\boldsymbol{w}}^r_{i,k-1})-f(\boldsymbol{w}^*)+f(\boldsymbol{w}^{r-1})-f(\boldsymbol{w}^*)).\end{aligned}\tag{62}$$

Second, we derive an upper bound for the second term in the right-hand side of (51). According to (41), we have

$$\nabla f(\hat{\boldsymbol{w}}^r_{i,k-1})^T(\boldsymbol{w}^*-\hat{\boldsymbol{w}}^r_{i,k-1})\leq f(\boldsymbol{w}^*)-f(\hat{\boldsymbol{w}}^r_{i,k-1}),\tag{63}$$

and the expected upper bound w.r.t. the round r can be written as

$$\begin{aligned}&-2\alpha_l\mathbb{E}_r[\mathbb{E}[\nabla f(\hat{\boldsymbol{w}}^r_{i,k-1})^T(\hat{\boldsymbol{w}}^r_{i,k-1}-\boldsymbol{w}^*)]]\\&\leq 2\alpha_l\mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^*)-f(\hat{\boldsymbol{w}}^r_{i,k-1})]].\end{aligned}\tag{64}$$

Combining (62) and (64), (51) can be upper bounded by

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k}-\boldsymbol{w}^*||^2]]\\&\leq\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k-1}-\boldsymbol{w}^*||^2]]-2\alpha_l\mathbb{E}_r[\mathbb{E}[f(\hat{\boldsymbol{w}}^r_{i,k-1})-f(\boldsymbol{w}^*)]]\\&+4\alpha_l^2L(f(\hat{\boldsymbol{w}}^r_{i,k-1})-f(\boldsymbol{w}^*)+f(\boldsymbol{w}^{r-1})-f(\boldsymbol{w}^*)),\end{aligned}\tag{65}$$

which leads to the conclusion that

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k}-\boldsymbol{w}^*||^2]]\leq\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k-1}-\boldsymbol{w}^*||^2]]\\&-2\alpha_l(1-2\alpha_lL)\mathbb{E}_r[\mathbb{E}[f(\hat{\boldsymbol{w}}^r_{i,k-1})-f(\boldsymbol{w}^*)]]\\&+4\alpha_l^2L(f(\boldsymbol{w}^{r-1})-f(\boldsymbol{w}^*)).\end{aligned}\tag{66}$$

By summing the inequality over $k=1,\cdots,K$, under the expectation over round r for each k, we can obtain

$$\begin{aligned}&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,K}-\boldsymbol{w}^*||^2]]\\&+2\alpha_l(1-2\alpha_lL)K\mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^r)-f(\boldsymbol{w}^*)]]\\&\leq\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,0}-\boldsymbol{w}^*||^2]]+4\alpha_l^2LK\mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^{r-1})-f(\boldsymbol{w}^*)]].\end{aligned}\tag{67}$$

Since $\hat{\boldsymbol{w}}^r_{i,0}=\boldsymbol{w}^{r-1}$, and using (41), we can write

$$\begin{aligned}f(\boldsymbol{w}^{r-1})\geq f(\boldsymbol{w}^*)+\nabla f(\boldsymbol{w}^*)^T(\boldsymbol{w}^{r-1}-\boldsymbol{w}^*)\\+\frac{\mu}{2}||\boldsymbol{w}^{r-1}-\boldsymbol{w}^*||^2,\end{aligned}\tag{68}$$

which leads to

$$||\boldsymbol{w}^{r-1}-\boldsymbol{w}^*||^2\leq\frac{2}{\mu}(f(\boldsymbol{w}^{r-1})-f(\boldsymbol{w}^*)),\tag{69}$$

Since $\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,K}-\boldsymbol{w}^*||^2]]\geq 0$, and

$$\mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^r)-f(\boldsymbol{w}^*)]]=\mathbb{E}_r[f(\boldsymbol{w}^r)-f(\boldsymbol{w}^*)],\tag{70}$$

combining (67), (69) and (70), we obtain

$$\begin{aligned}&\mathbb{E}_r[f(\boldsymbol{w}^r)-f(\boldsymbol{w}^*)]\\&\leq\left(\frac{1}{\mu\alpha_l(1-2\alpha_lL)K}+\frac{2\alpha_lL}{1-2\alpha_lL}\right)\\&\mathbb{E}_r[f(\boldsymbol{w}^{r-1})-f(\boldsymbol{w}^*)],\end{aligned}\tag{71}$$

which leads to

$$\mathbb{E}_r[f(\boldsymbol{w}^r)-f(\boldsymbol{w}^*)]\leq\varphi^r(f(\boldsymbol{w}^0)-f(\boldsymbol{w}^*)),\tag{72}$$

where

$$\varphi=\frac{1}{\mu\alpha_l(1-2\alpha_lL)K}+\frac{2\alpha_lL}{1-2\alpha_lL}.\tag{73}$$

(73) indicates that the proposed algorithm shall converge in the sense of (42) as long as the step length α_l and the number of local iterations K are selected to make the φ in (73) strictly less than one. There exists a variety of such selections. Provided below is a pair of α_l and K which, for a given target $\varphi<1$, reduces the number of local iterations K to minimum.

Let $\hat{\varphi}<1$ be a given target φ and split it as

$$\hat{\varphi}=p+q\tag{74}$$

with $p>0$ and $q>0$. We select the step-length α_l such that the second term of (73) satisfies

$$q=\frac{2\alpha_lL}{1-2\alpha_lL},\tag{75}$$

i.e.,

$$\alpha_l=\frac{q}{2(1+q)L}.\tag{76}$$

With α_l given by (76), we select the smallest integer K such that the first term of (73) satisfies

$$\frac{1}{\mu\alpha_l(1-2\alpha_lL)K}\leq p,\tag{77}$$

i.e.,

$$K\geq\frac{1}{\mu\alpha_l(1-2\alpha_lL)p}.\tag{78}$$

From (76), we see that α_l is a concave function that monotonically increases with q. Concerning the selection of appropriate $\{\alpha_l,K\}$, we use $p=\hat{\varphi}-q$ and (76) to write the lower bound of K in (78) as a function of q:

$$b(q)=\frac{1}{\mu\alpha_l(1-2\alpha_lL)p}=\frac{2L(1+q)^2}{\mu q(\hat{\varphi}-q)}.\tag{79}$$

Using calculus, it is straightforward to show that $b(q)$ is monotonically decreasing over $(0,\frac{\hat{\varphi}}{2+\hat{\varphi}})$, and monotonically increasing over $(\frac{\hat{\varphi}}{2+\hat{\varphi}},\hat{\varphi})$. Therefore, $b(q)$ admits a unique minimum over $(0,\hat{\varphi})$ at

$$q^*=\frac{\hat{\varphi}}{2+\hat{\varphi}},\tag{80}$$

and hence the best selection $\{\alpha_l^*,K^*\}$ in the sense of smallest number of local iterations is given by

$$\alpha_l^*=\frac{\hat{\varphi}}{4L(1+\hat{\varphi})},\tag{81}$$

and

$$K^*=\left\lceil\frac{8L(1+\hat{\varphi})}{\mu\hat{\varphi}^2}\right\rceil,\tag{82}$$

where $\lceil x\rceil$ denotes the smallest integer larger than or equal to x.

REFERENCES

[1] Y. Cao, S. Xu, J. Liu, and N. Kato, "IRS backscatter enhancing against jamming and eavesdropping attacks," *IEEE Internet of Things Journal*, vol. 10, no. 12, pp. 10 740–10 751, 2023.

[2] B. Mao, Y. Wu, J. Liu, H. Guo, J. Wang, and N. Kato, "Optimizing secrecy rate for federated learning model aggregation with intelligent reflecting surface towards 6g ubiquitous intelligence," *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[3] R. Alt, G. Fridgen, and Y. Chang, "The future of fintech—towards ubiquitous financial services," *Electronic Markets*, vol. 34, no. 1, p. 3, 2024.

[4] L. Zhao, L. Cai, and W.-S. Lu, "Federated learning for data trading portfolio allocation with autonomous economic agents," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[5] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6g: Machine-learning approaches," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 292–307, 2019.

[6] J. Wang, H. Du, D. Niyato, Z. Xiong, J. Kang, B. Ai, Z. Han, and D. I. Kim, "Generative artificial intelligence assisted wireless sensing: Human flow detection in practical communication environments," *IEEE Journal on Selected Areas in Communications*, 2024.

[7] Y. Zhu, B. Mao, and N. Kato, "Intelligent reflecting surface in 6g vehicular communications: A survey," *IEEE Open Journal of Vehicular Technology*, vol. 3, pp. 266–277, 2022.

[8] L. U. Khan, A. Elhagry, M. Guizani, and A. El Saddik, "Edge intelligence empowered vehicular metaverse: Key design aspects and future directions," *IEEE Internet of Things Magazine*, vol. 7, no. 1, pp. 120–126, 2024.

[9] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[11] L. Zhao, L. Cai, and W.-S. Lu, "Tailored federated learning with adaptive central acceleration on diversified global models," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[12] D. Yang, W. Zhang, Q. Ye, C. Zhang, N. Zhang, C. Huang, H. Zhang, and X. Shen, "Detfed: Dynamic resource scheduling for deterministic federated learning over time-sensitive networks," *IEEE Transactions on Mobile Computing*, 2023.

[13] J. Xu, X. Tong, and S.-L. Huang, "Personalized federated learning with feature alignment and classifier collaboration," *arXiv preprint arXiv:2306.11867*, 2023.

[14] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan, "Fedala: Adaptive local aggregation for personalized federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, 2023, pp. 11 237–11 244.

[15] J. Zhang, Y. Hua, J. Cao, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan, "Eliminating domain bias for federated learning in representation space," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] L. Yi, G. Wang, X. Liu, Z. Shi, and H. Yu, "Fedgh: Heterogeneous federated learning with generalized global header," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8686–8696.

[17] Q. Ye, W. Shi, K. Qu, H. He, W. Zhuang, and X. Shen, "Joint ran slicing and computation offloading for autonomous vehicular networks: A learning-assisted hierarchical approach," *IEEE Open Journal of Vehicular Technology*, vol. 2, pp. 272–288, 2021.

[18] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9896–9910, 2018.

[19] T. Zhang and S. Mao, "Energy-efficient federated learning with intelligent reflecting surface," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 2, pp. 845–858, 2021.

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[21] S. U. Stich, "Local sgd converges fast and communicates little," *arXiv preprint arXiv:1805.09767*, 2018.

[22] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[23] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[24] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[25] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[26] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local sgd on identical and heterogeneous data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4519–4529.

[27] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," *arXiv preprint arXiv:2101.11203*, 2021.

[28] 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.

[29] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5693–5700.

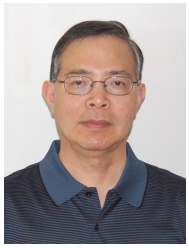
[30] E. Gorbunov, K. P. Burlachenko, Z. Li, and P. Richtárik, "Marina: Faster non-convex distributed learning with compression," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3788–3798.

[31] L. Zhao, L. Cai, and W.-S. Lu, "Collaborative learning of different types of healthcare data from heterogeneous iot devices," *IEEE Internet of Things Journal*, 2023.

[32] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[33] W.-S. Lu, "Handwritten digits recognition using pca of histogram of oriented gradient," in *2017 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. IEEE, 2017, pp. 1–5.

Lei Zhao (S'17) received the B.S. and M.A.Sc. degrees in computer science and technology from Xidian University, Xi'an, China, in 2015 and 2018, respectively, and earned his Ph.D. in Electrical and Computer Engineering from the University of Victoria in 2023. He is currently a Post-Doctoral Fellow in the E&CE Department at the University of Victoria. His research focuses on FL and optimization with applications in finance.



Wu-Sheng Lu (F'99-LF'12) received the B.Sc. degree in Mathematics from Fudan University, Shanghai, China, in 1964, the M.S. degree in electrical engineering, and the Ph.D. degree in control science from the University of Minnesota, Minneapolis, USA, in 1983 and 1984, respectively. Since 1987, he has been with the University of Victoria, Victoria, B.C., Canada, and is now Professor Emeritus. He is the co-author with A. Antoniou of Two-Dimensional Digital Filters (Marcel Dekker, 1992) and Practical Optimization: Algorithms and Engineering Applications (2nd ed., Springer, 2021), and with E. K. P. Chong and S. H. Zak of An Introduction to Optimization (5th ed., Wiley, 2023).

Lin Cai (S'00-M'06-SM'10-F'20) is a Professor with the E&CE Department at the University of Victoria. She is an NSERC Steacie Memorial Fellow, a CAE fellow, a EIC Fellow, an IEEE Fellow, an RSC College member, and a 2020 "Star in Computer Networking and Communications" by N2Women. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting ubiquitous intelligence.