Tailored Federated Learning With Adaptive Central Acceleration on Diversified Global Models

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

***Abstract*— We consider a setting engaging in collaborative learning with other machines where each individual machine has its own interests. How to effectively collaborate among machines with diverse requirements to maximize the profits of each participant poses a challenge in federated learning (FL). Our studies are motivated by the observation that in FL the global model attempts to acquire knowledge from each individual machine, while aggregating all local models into one optimal solution may not be desirable for some machines. To effectively leverage the knowledge of others while obtaining the customized solution for individual machine, we propose the accelerated federated training procedures with diversified global models. Based on the federated stochastic variance reduced gradient (FSVRG) framework, we propose the model-based grouping mechanism with adaptive central acceleration (MA-FSVRG) and gradients-based grouping mechanism with adaptive central acceleration (GA-FSVRG) to tackle the challenges of heterogeneous demands. The simulation results demonstrate the advantages of the proposed MA-FSVRG and GA-FSVRG over the state-of-the-art FL baselines. MA-FSVRG exhibits greater stability in performance and significant cost savings in local computation expenses compared to GA-FSVRG. On the other hand, GA-FSVRG attains higher test accuracy and faster convergence speed, particularly in scenarios with limited individual machine participation.**

***Index Terms*— Adaptive central acceleration, diversified global models, gradient-based grouping, model-based grouping.**

Nomenclature

E	Set of all individual machines.
P_i	Local dataset for the ith individual machine.
n_i	Number of local samples in individual machine i.
n	Total number of samples in set E.
$f(\boldsymbol{w})$	Global objective function.
$f_i(\boldsymbol{w})$	Local objective function of machine i.
$\triangledown f(\boldsymbol{w}_g)$	Global gradient with respect to $\boldsymbol{w}_g$.
$\triangledown f_i(\boldsymbol{w})$	Local gradient of machine i.
$\boldsymbol{g}_i(\boldsymbol{w})$	Local gradient estimation of machine i.
$\Delta \boldsymbol{w}_g$	Updating in global model.
$\boldsymbol{w}^*$	Optimal global model.
d	Number of model parameters.
L_i	Largest eigenvalue of $\triangledown^2 f_i(\boldsymbol{w})$.
μ_i	Smallest eigenvalue of $\triangledown^2 f_i(\boldsymbol{w})$.
T	Threshold of rounds to begin grouping mechanism.
S^r	Participated subset in the rth round.
n^r	Total number of samples in subset S^r.
C	Number of diversified global models.
α_l^i	Learning rate of machine i in lth local updating.
$\hat{\boldsymbol{m}}_c^r$	Estimation of the anchor gradients' first moments.
$\hat{\boldsymbol{v}}_c^r$	Estimation of the anchor gradients' second moments.
$\hat{\boldsymbol{w}}_{i,K}^r$	Trained local model from machine i in round r.
$\hat{\boldsymbol{w}}_{c,r-1}$	Global model representation of group c in round r.
S_c^r	Subset of machines of group c in round r.
$\boldsymbol{g}_c^{r-1}$	Anchor gradient of group c in round r.
$\boldsymbol{\Lambda}_i$	Local gradient adjustment matrix for machine i.
$\boldsymbol{d}_{i,k-1}^r$	kth updating direction for machine i in round r.
$\bar{\boldsymbol{w}}_i^{r-1}$	Mean-centered trained local model of machine i.
$\boldsymbol{W}_{r-1}$	Subset of the mean-centered local models in S^r.
$\triangledown \bar{f}_i(\hat{\boldsymbol{w}}_{i,K}^r)$	Mean-centered local gradient of machine i.
$\boldsymbol{G}_{r-1}$	Subset of the mean-centered local gradients in S^r.
c_i^l	Group that machine i prefers in the lth round of grouping updating.
$\hat{\boldsymbol{w}}_{i,0}^{r-1}$	Initialized local model for machine i in round r.
$\tilde{\boldsymbol{w}}_i^{r-1}$	Transformed trained local model of machine i.
$\tilde{\boldsymbol{w}}_{c,r-1}^l$	cth global model in lth grouping updating.
$\hat{\boldsymbol{w}}_{c,r-1}^*$	cth global model after grouping in round r.
$\tilde{\boldsymbol{g}}_{c,r-1}^l$	cth anchor gradient in lth grouping updating.
$\tilde{\boldsymbol{g}}_{c,r-1}^*$	cth anchor gradient after grouping in round r.
$S_{c,l}^r$	Subset in group c in lth grouping updating.
$\hat{\boldsymbol{w}}_c^r$	cth central accelerated global model.
$\boldsymbol{A}_r$	Scaling factors for model aggregation in round r.

Received 30 November 2023; revised 9 July 2024; accepted 26 October 2024. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and Compute Canada. *(Corresponding author: Lin Cai.)*

The authors are with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P 5C2, Canada (e-mail: cai@ece.uvic.ca).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2024.3487873, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2024.3487873

I. Introduction

IN FEDERATED learning (FL), a number of individual machines or organizations work together to train a model under the coordination of a central server. A key aspect of FL is that the training data are kept decentralized, ensuring privacy and security [1], [2]. By leveraging the collective power of individual machines and their local data, FL enables model training while mitigating privacy risks and reducing costs compared to centralized machine learning methods. Existing research in FL primarily focuses on harnessing the strengths of individual machines to train a single global model [3], where each individual machine contributes its computing resources to conduct model training on its local dataset.

However, current FL often neglects the heterogeneous needs of individual machines. In some cases, a single global model may not adequately satisfy the different requirements of different machines [4]. A critical issue is to obtain personalized FL models tailored for diversified individual demands [5], [6]. The existing personalized FL research is still based on the unique shared global model and the personality is obtained by the tradeoff between the global model and the diversified local needs [7], [8] [9]. This article proposes to enhance FL by accommodating diverse requisites of individual machines. We address highly varied local demands within FL and empower individual machines to make decisions aligned with their own interests.

On the other hand, due to the dynamic collaboration environments, it is hard to guarantee the successful participation of all machines. Therefore, the intrinsic requirement for FL is to converge faster. The training speed of FL is also impacted by the communication between the central server and individual machines during the training procedure.

While performing multiple local updates on individual machines before communicating with the server can significantly reduce communication costs [1], heterogeneous local datasets can result in higher variance as the number of local updates increases. One alternative approach to effectively mitigate communication costs is the utilization of stochastic individual machine selection, which involves the selection of a subset of individual machines for local updates.

Nonetheless, the random selection of participating machines causes increased variance of the stochastic gradient, which in turn leads to slow convergence and hence requests more iterations [10]. To mitigate this challenge, combining with variance reduction techniques have been introduced [11], [12], we compensate this negative aspect of stochastic individual machine selection by a momentum-based model update acceleration mechanism at the central server, which can achieve lower computational complexity and communication cost for individual machines, and achieve higher test accuracy.

Central acceleration expedites convergence, with communication being scaled down in each round proportional to the ratio of selected individual machines to the total count, thanks to the efficacy of stochastic individual machine selection. In addition, the computational load for a given FL task is significantly lower compared to traditional FL. This reduction is an outcome of the central acceleration combined with stochastic individual machine selection.

The main contributions of this article are as follows. First, we propose a comprehensive approach combining federated stochastic variance reduced gradient (FSVRG) with adaptive central acceleration on diversified global models, aiming to mitigate the high variance in global model updates and accelerate the convergence speed. Second, for the training process, we devise two strategies, i.e., the model-based grouping mechanism (MA-FSVRG) which applies the local model information to generate diversified global models, and the gradients-based grouping mechanism (GA-FSVRG) applies the local gradient information to generate diversified global models. Compared with MA-FSVRG, GA-FSVRG can converge faster to higher test accuracy but with higher variance in the performance and higher local computation cost, where MA-FSVRG achieves more stable performance with less cost. The experiment results show that the proposed algorithms can converge faster and achieve higher accuracy compared with the state-of-the-art baseline algorithms.

The rest of this article is organized as follows. The related works are summarized in Section II. Section III formulates the FL architecture with stochastic variance reduced gradient. The accelerated FSVRG with diversified global models method is proposed in Section IV. Simulation results are presented in Section V followed by the concluding remarks in Section VI.

II. Related Work

In standard FL, a central server aggregates model updates from all participating individual machines, and the model is updated based on a weighted average of these updates [6]. However, this approach can be biased toward individual machines with diversified objectives. Solely optimizing the accuracy of the global model tends to have a negative impact on its capacity to personalize [3], [13]. Personalized FL is an area of research that aims to tailor the FL process to each individual machine, allowing the model to be personalized based on their specific data and preferences [7], [8] [9]. The recent research efforts in personalized FL mainly focus on the tradeoff between the collaborative optimization and and model generalization [14].

Several works apply knowledge transfer into FL [15], [16]. The basic idea of federated transfer learning is to transfer the globally-shared model to distributed devices for further personalization in order to mitigate the statistical heterogeneity inherent in FL [17]. Based on the idea that lower layers of deep networks focus on learning common and low-level features and model parameters in higher layers learn more specific features, model parameters in lower layers of the global model can be shared with all individual machines, while the model parameters in higher layers should be fine-tuned with local data [16]. The shared layers are trained in a collaborative manner using the existing FL method, while the personalization layers are trained locally thereby enabling to capture of personal information of individual machines [18]. In another way, neural architecture search is utilized to find personalized neural network architectures for each individual machine, enhancing the model's performance on individual data distributions [19], [20]. Furthermore, personalized regularization terms are introduced in the local training to enhance personalized FL [21].

A model-agnostic meta-learning approach [22] is proposed to adapt the global model to each individual machine's data distribution, achieving better personalized performance [14], [23]. These works are based on the assumption that local domains are related which makes knowledge transfer possible [24]. However, how to measure the relations among highly heterogeneous individual machines is still an open challenge. To alleviate the highly unbalanced distribution of individual machine data, a data-sharing strategy is also proposed by Zhao et al. [25] where a small amount of global data containing a uniform distribution over classes from the center is distributed to individual machines. However, directly distributing the global data to individual machines will impose great privacy leakage risk, this approach is required to make a tradeoff between data privacy protection and performance improvement.

No matter the federated transfer learning or federated meta-learning, their aim is to learn a shared model of the same or similar tasks across individual machines. However, they are still based on the unique shared global model and the personality is obtained by the tradeoff between the global model and the diversified local needs. Therefore, it will be a challenge to enhance the training efficiency, i.e., speed up the convergence speed with superb performance. The fast convergence speed is critical for FL since individual machines, e.g., IoT devices, are frequently offline or on slow and expensive connections. However, the existing works need to improve the performance of the shared global model to enhance the overall FL performance, which means there will be few individual machines available for personalization. How to manage communication overhead and ensure convergence to a high-quality global model with a diverse set of individual machine updates is still one open issue.

SCAFFOLD [26] addresses individual machine drift caused by heterogeneous data distributions in FL. It employs control variates to correct individual machine updates, ensuring alignment with the global model direction. This method reduces variance in updates and enhances stability and convergence speed. By using control variates, SCAFFOLD significantly lowers variance in individual machine updates, leading to more stable and efficient convergence. It offers strong convergence guarantees, demonstrating lower error rates compared to traditional FL methods. However, the extent of variance reduction achieved by SCAFFOLD may vary depending on factors such as the sparsity of the data and the heterogeneity of edge devices.

MimeSVRG builds on MIME by introducing a variance reduction mechanism to ensure more accurate and stable gradient updates in FL [27]. It addresses the common issue of gradient estimation variance due to data heterogeneity across individual machines, which can lead to slow convergence and poor model performance. The server selects a subset of individual machines, sends them the current global model and optimizer state, and each individual machine performs local updates. Unlike standard FL methods, MimeSVRG includes an additional step where individual machines compute a control variate, a reference gradient based on a subset of the individual machine's data, to adjust local gradient estimates and reduce variance. This incorporation of SVRG [12] into the FL framework achieves more stable and faster convergence, especially beneficial in environments with high data heterogeneity. However, MimeSVRG introduces additional computational overhead and higher communication costs, which can be a significant drawback for individual machines with limited resources, such as mobile or IoT devices. The performance benefits are also dependent on the data distribution, and in less heterogeneous environments, the extra costs may not be justified.

LoSAC [28] introduces a novel approach to federated optimization by using delayed gradients to estimate the global full gradient on individual machines. These gradients are updated with the latest information at each local iteration and aggregated from participating individual machines, improving estimation accuracy while maintaining low computational complexity. LoSAC performs multiple local iterations to enhance communication efficiency and incorporates local second-order information during model updates to reduce the variance of stochastic gradients. This leads to faster convergence and improved robustness, especially in scenarios with nonconvex and ill-conditioned objective functions. However, LoSAC requires significant local memory to store various local dataset partitions, improving performance by allowing better local estimation of the global full gradient and reducing the impact of non-IID data. This memory requirement can be a significant drawback for individual machines with limited resources, making it less feasible for all FL applications.

Our proposed methods, MA-FSVRG and GA-FSVRG, aim to tackle the challenges of heterogeneity and diversified requirements in FL by focusing on adaptive central acceleration with model-based and gradient-based grouping mechanisms. This approach reduces variance in local updates and enhances computational efficiency and convergence speed through diversified global models. Unlike SCAFFOLD, which uses control variates to mitigate individual machine drift, our methods introduce adaptive strategies that dynamically adjust to the diversity of local data distributions, making them particularly robust in scenarios with significant non-IID data. Similarly, while MimeSVRG relies on global optimizer states and SVRG-style corrections, our methods leverage adaptive central acceleration to provide a more comprehensive solution to the challenges in FL including enhanced computational efficiency, improving stability and applicability in real-world environments. Compared to LoSAC, which focuses on using delayed gradients and multiple local iterations for efficient communication and accurate gradient estimation, our methods emphasize adaptive strategies that not only reduce variance but also dynamically adjust to diverse local data distributions, further enhancing performance and convergence speed. Overall, MA-FSVRG and GA-FSVRG offer a robust and efficient solution for FL by addressing the unique challenges presented by heterogeneous data distributions.

III. FL Architecture With Stochastic Variance Reduced Gradient

A. Problem Setup

Considering a distributed learning system, there are n training samples in total. We use E to denote the set of all individual machines, and use P_i to denote the local dataset for the ith individual machine with n_i local training samples for $i = 1, 2, \dots, |E|$. There is no overlap among different local datasets, i.e., $P_i \cap P_j = \emptyset$ whenever $i \neq j$. The optimization problem in an FL objective is 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 $f_i(\boldsymbol{w})$ represents the ith local objective function, which is the average of the empirical loss over all local samples $\{F_k(w)\}_{k\in P_i}$

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

The global objective function $f(\boldsymbol{w})$ is the weighted average of the local objective functions. The traditional FL goal is to find the optimal global model $\boldsymbol{w}^*$ to minimize overall losses of all machines. All notations in the FL architecture formulation are summarized in Nomenclature.

B. FL Architecture With Reduced Variance

With both stochastic individual machine selection in each federated round and stochastic local model updating, we need to reduce the variance to improve the learning efficiency. In the stochastic local updating, we use $\boldsymbol{g_i(w)}$ to represent the local gradient from individual training samples or a batch of the training samples which is an unbiased estimation of $\nabla f_i(\boldsymbol{w})$. The variance of local stochastic gradient $\boldsymbol{g_i(w)}$ can be analyzed as follows. According to Jensen's inequality, we obtain the upper bound of the variance of the sampled local gradient as

$$E\left[\left|\left|\boldsymbol{g}_i(\boldsymbol{w})\right|\right|^2\right] \le ||\nabla f_i(\boldsymbol{w})||^2 \tag{3}$$

which can be written as

$$E\left[\boldsymbol{\left|\left|g_i(w)\right|\right|^2}\right] \le \left|\left|\nabla f_i(\boldsymbol{w}) - \nabla f_i\left(\boldsymbol{w}^*\right)\right|\right|^2 \tag{4}$$

where $\boldsymbol{w}^*$ denotes the optimal model. According to Appendix in the Supplementary Material and (4), we can obtain

$$E\left[\boldsymbol{\left|\left|g_i(w)\right|\right|^2}\right] \le L_i^2\left|\left|\boldsymbol{w} - \boldsymbol{w}^*\right|\right|^2 \tag{5}$$

where L_i refers to the largest eigenvalue of $\nabla^{\boldsymbol{2}}\boldsymbol{f_i(w)}$. Due to the heterogeneous local datasets and diversified objectives, the local gradients in FL optimization are biased, and it will be very hard to make efficient progress in the training procedure.

We use $\boldsymbol{w_g}$ to refer to the global model, $\Delta\boldsymbol{w}_g$ to denote the updating in the global model. And similar to the analysis in Appendix in the Supplementary Material, the global objective function can be upper bounded by

$$\begin{aligned} f\left(\boldsymbol{w}_g + \Delta\boldsymbol{w}_g\right) \le{} & f\left(\boldsymbol{w}_g\right) + \nabla f\left(\boldsymbol{w}_g\right)^T\left(\Delta\boldsymbol{w}_g\right) \\ & + \sum_{i=1}^{|E|}\frac{n_i}{2n}L_i\boldsymbol{\left|\left|\Delta w_g\right|\right|^2} \end{aligned} \tag{6}$$

and lower bounded by

$$\begin{aligned} f\left(\boldsymbol{w}_g + \Delta\boldsymbol{w}_g\right) \ge{} & f\left(\boldsymbol{w}_g\right) + \nabla f\left(\boldsymbol{w}_g\right)^T\left(\Delta\boldsymbol{w}_g\right) \\ & + \sum_{i=1}^{|E|}\frac{n_i}{2n}\mu_i\boldsymbol{\left|\left|\Delta w_g\right|\right|^2} \end{aligned} \tag{7}$$

where the global gradient is defined as

$$\nabla f\left(\boldsymbol{w}_g\right) = \sum_{i=1}^{|E|}\frac{n_i}{n}\nabla f_i\left(\boldsymbol{w}_g\right).$$

We define $\boldsymbol{g}(\boldsymbol{w}_g) = \nabla f(\boldsymbol{w}_g;\xi)$ as an unbiased stochastic gradient of the global objective function $f(\boldsymbol{w}_g)$ with the variance bounded by σ^2. With the smoothness assumption, we can obtain

$$\left|\left|\boldsymbol{g}\left(\boldsymbol{w}_g\right)\right|\right| \le \sum_{i=1}^{|E|}\frac{n_i}{n}L_i\left|\left|\boldsymbol{w}_g - \boldsymbol{w}^*\right|\right|. \tag{8}$$

There are discrepancies between the global model and the local models. To reduce the updating variance, we modify the local updating of individual machine i should follow the guidance from:

$$\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i\left(\boldsymbol{w_g}\right) + \boldsymbol{g}\left(\boldsymbol{w_g}\right) \tag{9}$$

to force the local gradient to be unbiased for the local training procedure. The stochastic update in machine i yields an unbiased estimate of the global gradient $\nabla f(\boldsymbol{w}_g)$ since $E[\boldsymbol{g_i(w) - g_i(w_g)}] = \boldsymbol{0}$, which leads to

$$\nabla f\left(\boldsymbol{w_g}\right) \approx E\left[\boldsymbol{g_i(w) - g_i(w_g)} + E\left[\boldsymbol{g(w_g)}\right]\right]. \tag{10}$$

And the variance of the global gradient estimation from (10) can be rewritten as

$$\mathrm{Var}\left(\boldsymbol{g_i(w) - g_i\left(w_g\right)} + E\left[\boldsymbol{g\left(w_g\right)}\right]\right) = \mathrm{Var}\left(\boldsymbol{g_i(w) - g_i\left(w_g\right)}\right)$$

where $E[\boldsymbol{g(w_g)}]$ is a constant during the local updating. Since we can rewrite

$$\begin{aligned} \mathrm{Var}\left(\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right) ={} & \mathbb{E}\left[\left|\left|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right|\right|^2\right] \\ & - \left|\left|\mathbb{E}\left[\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right]\right|\right|^2 \end{aligned} \tag{11}$$

the variance of the estimated global gradient can be upper bounded by

$$\mathrm{Var}\left(\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right) \le \mathbb{E}\left[\left|\left|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right|\right|^2\right]. \tag{12}$$

Furthermore, according to Jensen's inequality, we can obtain

$$\mathbb{E}\left[\left|\left|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right|\right|^2\right] \le \left|\left|\mathbb{E}\left[\boldsymbol{g}_i(\boldsymbol{w})\right] - E\left[\boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right]\right|\right|^2. \tag{13}$$

Since the stochastic local gradients are unbiased to the full local gradients, and according to the Appendix in the Supplementary Material, we obtain

$$\left|\left|\mathbb{E}\left[\boldsymbol{g}_i(\boldsymbol{w})\right] - \mathbb{E}\left[\boldsymbol{g}_i\left(\boldsymbol{w}_g\right)\right]\right|\right|^2 \le L_i^2\left|\left|\boldsymbol{w} - \boldsymbol{w}_g\right|\right|^2. \tag{14}$$

Combining (11)–(14), we can obtain the upper bound of the variance

$$\mathrm{Var}\left(\boldsymbol{g_i(w) - g_i\left(w_g\right)} + E\left[\boldsymbol{g\left(w_g\right)}\right]\right) \le L_i^2\boldsymbol{\left|\left|w - w_g\right|\right|^2}. \tag{15}$$

But as long as $\boldsymbol{w \ne w_g}$, the variance will still exist.

Given the inherent variability of local optimization procedures within practical FL scenarios involving heterogeneous local datasets, it is common for these procedures to converge to diverse solutions. Consequently, the local gradients in FL optimization tend to exhibit bias, presenting a significant challenge when attempting to achieve efficient training progress. Thus, we are motivated to investigate and propose central acceleration on diversified global models to fill the gap.

IV. Adaptive Central Acceleration on Diversified Global Models

In this section, we present two innovative algorithms: model-based and gradients-based grouping mechanisms with adaptive central acceleration (MA-FSVRG and GA-FSVRG), designed to enhance FL with diversified local requirements. MA-FSVRG generates diversified global models based on the similarity of local model parameters, leveraging the inherent similarity among models to improve the overall learning process. This method offers stability and consistency, with parameters changing gradually over federated iterations, providing a steady and predictable measure of progress. GA-FSVRG, on the other hand, generates diversified global models based on the similarity of local gradient information, focusing on the direction and magnitude of updates made by individual machines during training. Utilizing gradient information provides a dynamic and immediate measure of the learning process, allowing for quicker adaptation to changes in data distribution.

In both MA-FSVRG and GA-FSVRG, the process begins with the central server initializing multiple diversified global models. During each training round, a subset of individual machines is selected to participate in federated training. The central server broadcasts the current diversified global models to these selected machines. Upon receiving the models, each machine computes the full local gradients of all diversified models and its preferred global model based on its own data. It then sends the full local gradients and its preference information back to the central server, which uses this data to generate the anchor gradients for each global model. The central server transmits the corresponding anchor gradient to each machine based on its preference, allowing the machines to perform variance-reduced local updates to their copies of the preferred global models. The machines then send their updated models or local gradient update information back to the central server. The central server uses this information from all participating machines to refine the diversified global models. After refining the group representations, the central server applies adaptive acceleration to all diversified global models. The updated models are then redistributed to the individual machines in the next round.

A. Local Training With Reduced Gradient Variance

At the beginning of the federated training, without the knowledge of others, an individual's preference is not obvious. Therefore, we define a threshold T, and when the federated training number $r \leq T$, we randomly pick a subset of individual machines to construct a subset $S^r \subseteq [E]$ with size $|S^r| = S$ at the beginning of the rth round. The global mode $\boldsymbol{w}^{r-1}$ is distributed to the selected individual machines in S^r. The selected individual machines in S^r evaluate their full local gradients and transmit $\{\nabla f_i(\boldsymbol{w}^{r-1})\}_{i \in S^r}$ to central server to aggregate the current anchor gradient as

$$\boldsymbol{g}\left(\boldsymbol{w}^{r-1}\right) = \sum_{i \in S^r} \frac{n_i}{n^r} \nabla f_i\left(\boldsymbol{w}^{r-1}\right) \tag{16}$$

where n^r denotes the number of samples from all individual machines in subset S^r and n_i for individual machine i. With random individual machine selection, the anchor gradient $\boldsymbol{g}(\boldsymbol{w}^{r-1})$ obtained by an entire data pass of all the selected local datasets is an unbiased estimation of the full global gradient $\nabla f(\boldsymbol{w}^{r-1})$. After the threshold T, i.e., $r > T$, to continue improving the training efficiency, individual machines need to collaborate based on their preference. All notations defined in the algorithm design are listed in Nomenclature.

The central server randomly selects a subset S^r of machines at the beginning of the rth round. The central server manages C diversified global models, denoted by $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^{C}$ which are distributed to each participated machine to initialize the local training. First, the local gradients collected from the individual machines in subset S^r for all global models in $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^{C}$ are collected by the central server to obtain multiple anchor gradients $\{\boldsymbol{g}_c^{r-1}\}_{c=1}^{C}$ with respect to the groups $\{S_c^r\}_{c=1}^{C}$ as

$$\left\{\boldsymbol{g}_c^{r-1} = \sum_{i \in S^{r-1}} \frac{n_i}{n^{r-1}} \nabla f_i\left(\hat{\boldsymbol{w}}_{c,r-1}\right)\right\}_{c=1}^{C}. \tag{17}$$

Furthermore, the individual machines in S^r first initialize their starting point by evaluating the received multiple global models from the central server by their own local datasets and select the global model based on their own preference as

$$\left\{\hat{\boldsymbol{w}}_{i,0}^{r-1} = \underset{\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^{C}}{\operatorname{argmin}} \, f_i\left(\hat{\boldsymbol{w}}_{c,r-1}\right)\right\}_{i \in S^r}. \tag{18}$$

This individual preference information is also collected by the central server, based on which the central server transmits the corresponding anchor gradients from $\{\boldsymbol{g}_c^{r-1}\}_{c=1}^{C}$ to the selected individual machines in S^r to reduce the variance in local training. Then, the individual machines in S_c^r conduct K local stochastic updates based on their own local datasets.

To enforce the auxiliary local gradient to be of the correct magnitude, it is scaled carefully by the number of nonzero features of the samples. The number of samples in the local dataset of individual machine i with nonzero jth feature is denoted by n_i^j. After going through their local datasets, individual machines send the number of local nonzero jth feature $\{n_i^j\}_{i \in E}$ to the central server, and the central server generates the number of samples with nonzero jth feature over all local datasets as

$$\hat{n}^j = \sum_{i \in E} n_i^j. \tag{19}$$

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

$$\Delta \boldsymbol{g}_{i,k-1}^{r} = \boldsymbol{\Lambda}_i \left[\boldsymbol{g}_i\left(\hat{\boldsymbol{w}}_{i,k-1}^{r}\right) - \boldsymbol{g}_i\left(\hat{\boldsymbol{w}}_{i,0}^{r-1}\right)\right] \tag{20}$$

where

$$\boldsymbol{\Lambda}_i = \operatorname{diag}\left(\left\{\frac{\hat{n}^j \cdot n_i}{n \cdot n_i^j}\right\}_{j=1,\ldots,q}\right). \tag{21}$$

The local updating direction is designed as

$$\boldsymbol{d}_{i,k-1}^{r} = -\left(\Delta \boldsymbol{g}_{i,k-1}^{r} + \boldsymbol{g}_c^{r-1}\right). \tag{22}$$

The data available locally may be quite different in size. The auxiliary local gradient and the aggregation step need to be carefully tuned considering the large variance of the size of local datasets. The local iteration number, K, should be revisited when the local data sizes are different. The iteration number should be related to the number of steps to pass through all local samples. Each individual machine needs to make roughly the same progress in the same round. By setting the same local iteration number, the local learning rate is designed as $\alpha_l^i = (\alpha_l / n_i)$ to neutralize the data size difference. The local model update of the ith individual machine is now formulated as

$$\hat{\boldsymbol{w}}_{i,k}^{r-1} = \hat{\boldsymbol{w}}_{i,k-1}^{r-1} + \alpha_l^i \boldsymbol{d}_{i,k-1}^{r}. \tag{23}$$

Carrying (23) K times iterations leads to a formula below for the local model update at individual machine i as

$$\left\{\hat{\boldsymbol{w}}_{i,K}^{r} = \hat{\boldsymbol{w}}_{i,0}^{r-1} + \sum_{k=1}^{K} \alpha_l^i \boldsymbol{d}_{i,k-1}^{r}\right\}_{i \in S^r}. \tag{24}$$

The steps for local training with reduced gradient variance after the threshold T, i.e., $r > T$, are summarized in Procedure 1.

Procedure 1 Local Training With Variance Reduction

```
Require: $S^r$, $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^{C}$, $K$, $\alpha_l$
Central server distributes $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^{C}$ to machines in $S^r$
for each machine $i \in S^r$ do
    Compute full local gradients $\{\nabla f_i(\hat{\boldsymbol{w}}_{c,r-1})\}_{c=1}^{C}$
    Initialize local model with preference via (18)
    Transmit $\{\nabla f_i(\hat{\boldsymbol{w}}_{c,r-1})\}_{c=1}^{C}$ and local preference to central server
end for
Central server computes $\{\boldsymbol{g}_c^{r-1}\}_{c=1}^{C}$ via (17)
for each machine $i \in S^r$ do
    Compute scaling matrix $\boldsymbol{\Lambda}_i$ via (21)
    for each local iteration $k = 1$ to $K$ do
        Compute local update direction $\boldsymbol{d}_{i,k-1}^{r}$ via (22)
        Update to local model $\hat{\boldsymbol{w}}_{i,k}^{r-1}$ via (23)
    end for
end for
```

B. Model-Based Grouping Mechanism

The model-based grouping mechanism generates diversified global models by leveraging the inherent similarity among local model parameters, thereby enhancing the overall learning process. One notable advantage of this mechanism is its stability and consistency, as parameters change gradually over federated iterations, offering a steady and predictable measure of the model's progress. Furthermore, the direct relevance of model parameters to the state of the model simplifies performance assessment and necessary adjustments. To relieve the communication burden, the grouping procedure is conducted in the central server.

At the rth round, the central server collects the trained local models from the selected machines, i.e., $\{\hat{\boldsymbol{w}}_{i,K}^{r-1}\}_{i\in S^r}$. The objective in the central server is formulated as increasing the local models' similarity within the same group, which can be formulated as

$$\underset{\{S_c^r\}_{c=1}^{C}}{\text{minimize}} \quad \frac{1}{2}\sum_{c=1}^{C}\frac{1}{|S_c^r|}\sum_{i,j\in S_c^r}\left\|\hat{\boldsymbol{w}}_{i,K}^{r-1}-\hat{\boldsymbol{w}}_{j,K}^{r-1}\right\|^2. \tag{25}$$

The mean-centered local model parameters of machine i in the current round are denoted by

$$\left\{\bar{\boldsymbol{w}}_i^{r-1}=\hat{\boldsymbol{w}}_{i,K}^{r-1}-\frac{1}{|S^r|}\sum_{j\in S^r}\hat{\boldsymbol{w}}_{j,K}^{r-1}\right\}_{i\in S^r}. \tag{26}$$

According to the fact that

$$\left\|\hat{\boldsymbol{w}}_{i,K}^{r-1}-\hat{\boldsymbol{w}}_{j,K}^{r-1}\right\|^2=\left\|\bar{\boldsymbol{w}}_i^{r-1}-\bar{\boldsymbol{w}}_j^{r-1}\right\|^2 \tag{27}$$

the objective (25) can be rewritten as

$$\underset{\{S_c^r\}_{c=1}^{C}}{\text{maximize}} \quad \sum_{c=1}^{C}\frac{1}{|S_c^r|}\sum_{i,j\in S_c^r}\bar{\boldsymbol{w}}_i^{r-1T}\bar{\boldsymbol{w}}_j^{r-1}. \tag{28}$$

The solution of (28) is the assignment of trained local models into different groups, with the arrangement that all local models within the same group are adjacent to each other. We use matrix $\boldsymbol{W}_{r-1}$ to denote the subset of the mean-centered local models $\{\bar{\boldsymbol{w}}_i^{r-1}\}_{i\in S^r}$ where we can obtain

$$\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}=\left\{\bar{\boldsymbol{w}}_i^{r-1T}\bar{\boldsymbol{w}}_j^{r-1}\right\}_{i,j\in S^r}. \tag{29}$$

We define $\boldsymbol{H}_{r-1}=\{\boldsymbol{h}_c\}_{c=1}^{C}$ to refer to the assignment of trained local models into different groups, where the cth column $\boldsymbol{h}_c$ is the indicator vector of the local models in $\boldsymbol{W}_{r-1}$ which belong to the cth group scaled by $(1/(|S_c^r|)^{1/2})$. By this definition, the column vectors in $\boldsymbol{H}_{r-1}$ are orthonormal since $\boldsymbol{h}_i^T\boldsymbol{h}_i=1$ and $\boldsymbol{h}_i^T\boldsymbol{h}_j=0$ when $i\neq j$, which leads to $\boldsymbol{H}_{r-1}^T\boldsymbol{H}_{r-1}=\boldsymbol{I}_C$. Then, the objective (28) is equal to

$$\underset{\boldsymbol{H}_{r-1}}{\text{maximize}} \quad \text{Tr}\left(\boldsymbol{H}_{r-1}^T\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}\boldsymbol{H}_{r-1}\right). \tag{30}$$

To solve (30), we define a linear transformation on the group selection matrix $\boldsymbol{H}_{r-1}$ denoted by $\boldsymbol{Q}_{r-1}=\boldsymbol{H}_{r-1}\boldsymbol{T}_{r-1}$, where $\boldsymbol{T}_{r-1}$ is orthogonal matrix and the last column of $\boldsymbol{T}_{r-1}$ is defined as $\boldsymbol{t}_C=\{((|S_c^r|/|S^r|))^{1/2}\}_{c=1}^{C}$, then we have

$$\boldsymbol{H}_{r-1}\boldsymbol{t}_C=\sqrt{\frac{1}{|S^r|}}\boldsymbol{e} \tag{31}$$

where $\boldsymbol{e}$ is the all-one vector. The rest columns in $\boldsymbol{T}_{r-1}$ should satisfy

$$\boldsymbol{e}^T\boldsymbol{H}_{r-1}\boldsymbol{t}_c=0 \tag{32}$$

for $c=1,\ldots,C-1$ according to connectivity analysis [29]. We define $\hat{\boldsymbol{Q}}_{\boldsymbol{r-1}}$ as the matrix involving the first $C-1$ columns in $\boldsymbol{Q}_{r-1}$, and according to the property of trace, we can obtain

$$\begin{aligned}&\text{Tr}\left(\boldsymbol{H}_{r-1}^T\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}\boldsymbol{H}_{r-1}\right)\\&=\text{Tr}\left(\hat{\boldsymbol{Q}}_{r-1}^T\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}\hat{\boldsymbol{Q}}_{r-1}\right)+\frac{1}{|S^r|}\boldsymbol{e}^T\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}\boldsymbol{e}.\end{aligned} \tag{33}$$

Since

$$\boldsymbol{W}_{r-1}\boldsymbol{e}=\sum_{i\in S^r}\hat{\boldsymbol{w}}_{i,K}^{r-1}-\frac{1}{|S^r|}\sum_{j\in S^r}\hat{\boldsymbol{w}}_{j,K}^{r-1}=0. \tag{34}$$

Equation (30) can be rewritten as

$$\underset{\hat{\boldsymbol{Q}}_{r-1}}{\text{maximize}} \quad \text{Tr}\left(\hat{\boldsymbol{Q}}_{r-1}^T\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}\hat{\boldsymbol{Q}}_{r-1}\right) \tag{35}$$

with the condition that $\hat{\boldsymbol{Q}}_{r-1}^T\hat{\boldsymbol{Q}}_{r-1}=\boldsymbol{I}_{C-1}$, and (32), we can obtain the optimal solution to (35) as

$$\hat{\boldsymbol{Q}}_{r-1}^*=\hat{\boldsymbol{V}}_{r-1}\boldsymbol{R} \tag{36}$$

where $\boldsymbol{R}$ is an arbitrary $(C-1)\times(C-1)$ orthogonal matrix, and $\hat{\boldsymbol{V}}_{r-1}$ is the collection of the $C-1$ eigenvectors of $\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}$ corresponding to the $C-1$ largest eigenvalues according to [30]. Then, the objective of (35) is equivalent to $\sum_{c=1}^{C-1}\lambda_c$ where $\{\lambda_c\}_{c=1}^{C-1}$ are the $C-1$ largest eigenvalue of $\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}$.

However, there is a challenge to apply the solution in (36) for models grouping, since $\hat{\boldsymbol{Q}}_{r-1}^*$ is a transformation version of $\boldsymbol{H}_{r-1}^*$ which is the indicator matrix for the grouping results and the transformation between $\hat{\boldsymbol{Q}}_{r-1}^*$ and $\boldsymbol{H}_{r-1}^*$ is hard to obtain [29]. But the analysis given above justifies the advantages of using the principle components and eigenvalues of $\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}$ to conduce the grouping.

Our objective is not solely to group the local models but also to utilize the grouping information for updating the diversified global models. Consequently, the grouping information in the lower dimensional subspace needs to be transformed back to update the diversified global model. We use $\hat{\boldsymbol{U}}_{r-1} \in R^{d\times(C-1)}$ to denote the matrix consists of $C-1$ eigenvectors of $\boldsymbol{W}^T_{r-1}\boldsymbol{W}_{r-1}$, and $\hat{\boldsymbol{\Lambda}}_{r-1} = \mathrm{diag}\{\lambda_c\}^{C-1}_{c=1}$. The decomposition of $\boldsymbol{W}^T_{r-1}\boldsymbol{W}_{r-1}$ related to the first $C-1$ principle components can be written as

$$\hat{\boldsymbol{V}}_{r-1}\hat{\boldsymbol{\Lambda}}_{r-1}\hat{\boldsymbol{V}}^T_{\boldsymbol{r-1}} = \hat{\boldsymbol{V}}_{r-1}\hat{\boldsymbol{\Lambda}}^{\frac{1}{2}}_{r-1}\hat{\boldsymbol{U}}^T_{r-1}\hat{\boldsymbol{U}}_{r-1}\hat{\boldsymbol{\Lambda}}^{\frac{1}{2}}_{r-1}\hat{\boldsymbol{V}}^T_{r-1}. \tag{37}$$

According to (37), the local models in lower dimensional subspace used for grouping defined as $\hat{\boldsymbol{\Lambda}}^{(1/2)}_{r-1}\hat{\boldsymbol{V}}^T_{r-1}$ can be reformulated as

$$\tilde{\boldsymbol{W}}_{r-1} = \hat{\boldsymbol{U}}^T_{r-1}\boldsymbol{W}_{r-1} = \hat{\boldsymbol{\Lambda}}^{\frac{1}{2}}_{r-1}\hat{\boldsymbol{V}}^T_{r-1} = \sum_{i=1}^{C-1}\lambda_i^{\frac{1}{2}}\boldsymbol{u_i}\boldsymbol{v_i^T} \tag{38}$$

which are the projections of the participated mean-centered local models on the directions represented by each column in $\hat{\boldsymbol{U}}_{r-1}$. Then, we based on $\tilde{\boldsymbol{W}}_{r-1} = \{\tilde{\boldsymbol{w}}^{r-1}_i\}_{i\in S^r}$ to group the local models from participated machines.

The initial group representatives come from the previous accelerated global models $\{\hat{\boldsymbol{w}}_{c,r-1}\}^C_{c=1}$, which are projected to the lower dimensional space as

$$\left\{\tilde{\boldsymbol{w}}^0_{c,r-1} = \hat{\boldsymbol{U}}^T_{r-1}\hat{\boldsymbol{w}}_{c,r-1}\right\}^C_{c=1}. \tag{39}$$

After the initial grouping for all selected machines, it is needed to recalculate C new group representations resulting from the current trained local model set $\tilde{\boldsymbol{W}}_{r-1}$.

In the grouping procedure, each local model in $\tilde{\boldsymbol{W}}_{r-1}$ measures its Euclidean distance to the current group representatives, and the group selection is obtained by

$$\left\{c^l_i = \operatorname*{argmin}_{c=1,\ldots,C} \left\|\tilde{\boldsymbol{w}}^{r-1}_i - \tilde{\boldsymbol{w}}^l_{c,r-1}\right\|\right\}_{i\in S^r} \tag{40}$$

where c^l_i denotes the group of local model from machine i in the lth round of grouping updating. All local models with the same group selection $\{c^l_i = c\}_{i\in S^r}$ form group $S^r_{c,l}$. To achieve the most influential representation within different groups, the diversified multiple global models are updated as follows:

$$\tilde{\boldsymbol{w}}^l_{c,r-1} = \frac{1}{|S^r_{c,l}|}\sum_{i\in S^r_{c,l}}\tilde{\boldsymbol{w}}^{r-1}_i \tag{41}$$

in the lth round of grouping updating in the central server. When the grouping update in the central server converged, i.e.,

$$\sum_{c=1}^{C}\sum_{i\in S^r_{c,l}}\left\|\tilde{\boldsymbol{w}}^{r-1}_i - \tilde{\boldsymbol{w}}^l_{c,r-1}\right\| \leq \tilde{\epsilon} \tag{42}$$

where $\tilde{\epsilon} = 10^{-4}$, we define $\{\tilde{\boldsymbol{w}}^*_{c,r-1} = \tilde{\boldsymbol{w}}^l_{c,r-1}\}^C_{c=1}$ and $\{S^r_c = S^r_{c,l}\}^C_{c=1}$. To perform the updating of the diversified global model in the original space utilized for local training, the following procedure needs to be executed in the central server as follows:

$$\left\{\hat{\boldsymbol{w}}^*_{c,r-1} = \hat{\boldsymbol{U}}_{r-1}\tilde{\boldsymbol{w}}^*_{c,r-1} + \frac{1}{|S^r|}\sum_{j\in S^r}\hat{\boldsymbol{w}}^{r-1}_{j,K}\right\}^C_{c=1}. \tag{43}$$

The steps for model-based grouping mechanism after the threshold T, i.e., $r > T$, are summarized in Procedure 2.

Procedure 2 Model-Based Grouping Mechanism

```
Require: $C$, $S^r$, $\{\hat{\boldsymbol{w}}_{c,r-1}\}^C_{c=1}$
Conduct local training via Procedure (1)
Transmit $\{\hat{\boldsymbol{w}}^{r-1}_{i,K}\}_{i\in S^r}$ to the central server
Compute mean-centered trained local models via (26)
Construct the mean-centered similarity matrix $\boldsymbol{W}^T_{r-1}\boldsymbol{W}_{r-1}$ via (29) and its decomposition via (37)
Compute local model projections $\tilde{\boldsymbol{W}}_{r-1}$ via (38)
Initialize group representations via (39)
repeat
    for each machine $i \in S^r$ do
        Assign to group via (40)
    end for
    for each group $c$ do
        Update group representation via (41)
    end for
until Convergence
Update diversified global models via (43)
Output: diversified global models $\{\hat{\boldsymbol{w}}^*_{c,r-1}\}^C_{c=1}$
```

Following this process, the final grouping results, represented by $\hat{\boldsymbol{w}}c, r-1^{*}c = 1^{C}$, can be viewed as the current optimal representations of the trained local models from the participating machines. These models are organized into multiple groups denoted by $S^r_c c = 1^C$. The adaptive acceleration based on $\hat{\boldsymbol{w}}c, r-1^{*C}_{c=1}$ is further detailed in Section IV-D.

C. Gradient-Based Grouping Mechanism

We introduce another novel approach, i.e., local gradients-based grouping, carefully crafted to leverage training dynamics to generate diversified global models. In this method, the central server initiates C groups, with the grouping determined by the similarity of local gradients. Each machine selected in this process performs K local iterations, as illustrated in (23) and (24), and then evaluate and transmit the full local gradients w.r.t the trained local models $\{\nabla f_i(\hat{\boldsymbol{w}}^r_{i,K})\}_{i\in S^r}$ to the central server.

The central objective is framed as enhancing the coherence of gradient information within a given group, which can be expressed as

$$\operatorname*{minimize}_{\{S^r_c\}^C_{c=1}} \ \frac{1}{2}\sum_{c=1}^{C}\frac{1}{|S^r_c|}\sum_{i,j\in S^r_c}\left\|\nabla f_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right) - \nabla f_j\left(\hat{\boldsymbol{w}}^r_{j,K}\right)\right\|^2 \tag{44}$$

to ensure a focused and efficient collaboration based on the distinctive features of local gradients. Following the gathering of local gradient information denoted by $\{\nabla f_i(\hat{\boldsymbol{w}}^r_{i,K})\}_{i\in S^r}$, the central server computes mean-centered local gradients as

$$\left\{\nabla\bar{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right) = \nabla f_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right) - \frac{1}{|S^r|}\sum_{j\in S^r}\nabla f_j\left(\hat{\boldsymbol{w}}^r_{i,K}\right)\right\}_{i\in S^r}. \tag{45}$$

The objective (44) can be reformulated as

$$\operatorname*{maximize}_{\{S^r_c\}^C_{c=1}} \ \sum_{c=1}^{C}\frac{1}{|S^r_c|}\sum_{i,j\in S^r_c}\nabla\bar{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)^T\nabla\bar{f}_j\left(\hat{\boldsymbol{w}}^r_{j,K}\right) \tag{46}$$

according to

$$\begin{aligned}&\left\|\nabla f_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)-\nabla f_j\left(\hat{\boldsymbol{w}}^r_{j,K}\right)\right\|^2\\&\quad=\left\|\nabla \bar{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)-\nabla \bar{f}_j\left(\hat{\boldsymbol{w}}^r_{j,K}\right)\right\|^2.\end{aligned}\tag{47}$$

Following the conclusion in (36), the solution of (46) is related to the $C-1$ largest eigenvalues of $\boldsymbol{G}^T_{r-1}\boldsymbol{G}_{r-1}$, where

$$\boldsymbol{G}^T_{r-1}\boldsymbol{G}_{r-1}=\left\{\nabla \bar{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)^T\nabla \bar{f}_j\left(\hat{\boldsymbol{w}}^r_{j,K}\right)\right\}_{i,j\in S^r}\tag{48}$$

and its eigen-decomposition is acquired through

$$\boldsymbol{G}^T_{r-1}\boldsymbol{G}_{r-1}=\tilde{\boldsymbol{V}}^T_{r-1}\tilde{\boldsymbol{\Lambda}}_{r-1}\tilde{\boldsymbol{V}}_{r-1}.\tag{49}$$

The grouping based on gradients is executed using lower dimensional local gradients, represented as $\tilde{\boldsymbol{\Lambda}}^{(1/2)}_{r-1}\tilde{\boldsymbol{V}}^T_{r-1}$. This can be translated into the projection of all mean-centered local gradients in $\boldsymbol{G}_{r-1}$ onto the subspace spanned by $\tilde{\boldsymbol{U}}_{r-1}$ as follows:

$$\tilde{\boldsymbol{G}}_{r-1}=\tilde{\boldsymbol{U}}^T_{r-1}\boldsymbol{G}_{r-1}=\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}_{r-1}\tilde{\boldsymbol{V}}^T_{r-1}=\left\{\nabla \tilde{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)\right\}_{i\in S^r}\tag{50}$$

where

$$\nabla \tilde{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)=\tilde{\boldsymbol{U}}^T_{r-1}\nabla \bar{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right).\tag{51}$$

The initial grouping centers are derived from anchor gradients with respect to the prior diversified global models, i.e., $\{\hat{\boldsymbol{w}}^{r-1}_c\}^C_{c=1}$, which encompass the current local datasets from machines in S^r. This formulation is expressed as

$$\left\{\hat{\boldsymbol{g}}_{c,r-1}=\sum_{i\in S^r}\frac{n_i}{n^r}\nabla f_i\left(\hat{\boldsymbol{w}}^{r-1}_c\right)\right\}^C_{c=1}\tag{52}$$

where the lower dimensional representation is indicated by

$$\left\{\tilde{\boldsymbol{g}}^0_{c,r-1}=\tilde{\boldsymbol{U}}^T_{r-1}\hat{\boldsymbol{g}}_{c,r-1}\right\}^C_{c=1}.\tag{53}$$

After establishing the initial grouping centers, it becomes essential to recalculate the diversified anchor gradients with the goal of generating more robust representations within $\tilde{\boldsymbol{G}}_{r-1}$. These representations encapsulate the updating directions for the diversified global model.

During the grouping procedure, each local gradient in $\tilde{\boldsymbol{G}}_{r-1}$ evaluates its affinity with the current anchor gradients by quantifying the Euclidean distance to these multiple anchors. The selection of its group can be expressed as

$$\left\{c^l_i=\underset{c=1,\ldots,C}{\operatorname{argmin}}\left\|\nabla \tilde{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)-\tilde{\boldsymbol{g}}^l_{c,r-1}\right\|\right\}_{i\in S^r}\tag{54}$$

where c^l_i denotes the group for the local gradient from machine i in the lth round of grouping updating. All local gradients in $\tilde{\boldsymbol{G}}_{r-1}$ with identical assignments $\{c^l_i=c\}_{i\in S^r}$ to the anchor gradients constitute group $S^r_{c,l}$. In the same group, the local gradients share similar perspectives on the directions for global model updates.

To emerge as the most influential representatives within distinct groups, the diversified anchor gradients undergo updating as follows:

$$\left\{\tilde{\boldsymbol{g}}^l_{c,r-1}=\frac{1}{|S^r_{c,l}|}\sum_{i\in S^r_{c,l}}\nabla \tilde{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)\right\}^C_{c=1}\tag{55}$$

Procedure 3 Gradient-Based Grouping Mechanism

```
Require: $C$, $S^r$, $\{\hat{\boldsymbol{g}}_{c,r-1}\}^C_{c=1}$
Conduct local training via Procedure (1)
Transmit $\{\nabla f_i(\hat{\boldsymbol{w}}^r_{i,K})\}_{i\in S^r}$ to the central server
Compute mean-centered local gradients via (45)
Compute $\boldsymbol{G}^T_{r-1}\boldsymbol{G}_{r-1}$ and its decomposition via (48)-(49)
Compute gradient projections $\{\nabla \tilde{f}_i(\hat{\boldsymbol{w}}^r_{i,K})\}_{i\in S^r}$ via (50)
Initialize group representations via (53)
repeat
    for each machine $i \in S^r$ do
        Assign to group via (54)
    end for
    for each group $c$ do
        Update group representation via (55)
    end for
until Convergence
Compute diversified global updates via (57)
Update diversified global models via (58)
Output: diversified global models $\{\hat{\boldsymbol{w}}^*_{c,r-1}\}^C_{c=1}$
```

in the lth round of grouping updating in the central server. The iterative updating sequence progressively steers the anchor gradients toward their optimal states, which continues until no further adjustments are observed, i.e., until

$$\sum_{c=1}^{C}\sum_{i\in S^r_{c,l}}\left\|\nabla \tilde{f}_i\left(\hat{\boldsymbol{w}}^r_{i,K}\right)-\tilde{\boldsymbol{g}}^l_{c,r-1}\right\|\le\tilde{\epsilon}\tag{56}$$

where $\tilde{\epsilon}=10^{-4}$. We define $\{\tilde{\boldsymbol{g}}^*_{c,r-1}=\tilde{\boldsymbol{g}}^l_{c,r-1}\}^C_{c=1}$ and $\{S^r_c=S^r_{c,l}\}^C_{c=1}$. The optimal reflections of the diversified global updating are obtained by

$$\left\{\Delta\boldsymbol{w}^c_g=\boldsymbol{A}_r\left(\tilde{\boldsymbol{U}}_{r-1}\tilde{\boldsymbol{g}}^*_{c,r-1}+\frac{1}{|S^r|}\sum_{j\in S^r}\nabla f_j\left(\hat{\boldsymbol{w}}^r_{i,K}\right)\right)\right\}^C_{c=1}\tag{57}$$

which are the representations of C different visions to update the global models, and the diversified global models are generated as

$$\left\{\hat{\boldsymbol{w}}^*_{c,r-1}=\hat{\boldsymbol{w}}_{c,r-1}+\Delta\boldsymbol{w}^c_g\right\}^C_{c=1}.\tag{58}$$

The steps for gradient-based grouping mechanism after the threshold T, i.e., $r>T$, are summarized in Procedure 3. Subsequently, it follows the same adaptive central acceleration procedure outlined in Section IV-D to acquire the accelerated diversified global models. These models are then distributed to the participating machines for local updating in the next round.

D. Adaptive Central Acceleration on Diversified Global Models

Adaptive central acceleration is employed following both the model-based and gradient-based grouping mechanisms. This approach aims to further refine the learning process by dynamically adjusting the diversified central models' influence. By adaptively accelerating the diversified central models, this method ensures that the global model converges more efficiently, and ultimately enhancing the robustness and accuracy.

First, we scale the diversified global models $\{\hat{\boldsymbol{w}}^*_{c,r-1}\}_{c=1}^{C}$ according to the presence of specific features in individual local datasets. This scaling is employed to enhance the updating of gradients related to features that are less frequently collected by individual machines, and the scaling diagonal matrix for model aggregation is defined as

$$\boldsymbol{A}_r = \mathrm{diag}\left(\left\{\frac{|S^r|}{\sum_{i\in E} 1_{n_i^j \neq 0}}\right\}_{j=1,\ldots,q}\right). \tag{59}$$

The diversified global models are first updated by

$$\left\{\boldsymbol{w}_c^r = \hat{\boldsymbol{w}}_{c,r-1} + \boldsymbol{A}_r\left(\hat{\boldsymbol{w}}^*_{c,r-1} - \hat{\boldsymbol{w}}_{c,r-1}\right)\right\}_{c=1}^{C} \tag{60}$$

and then distributed to the current participated individual machines in S^r to collect the local gradients and calculate the anchor gradients as

$$\left\{\boldsymbol{g}\left(\boldsymbol{w}_c^r\right) = \sum_{i\in S^r}\frac{n_i}{n^r}\nabla f_i\left(\boldsymbol{w}_c^r\right)\right\}_{c=1}^{C}. \tag{61}$$

Then, the acceleration procedure for diversified global models $\{\boldsymbol{w}_c^r\}_{c=1}^{C}$ are conducted to obtain the updated global models $\{\hat{\boldsymbol{w}}_c^r\}_{c=1}^{C}$ based on both the current and past anchor gradient information with appropriate weights to ensure significant and lasting models. We define the exponential decay rate for the estimation of the first moment of the global gradient as a heavy-ball style momentum parameter β_1, and the decay rate of the per-coordinate exponential moving average of the squared gradients as β_2.

We define $\hat{\boldsymbol{m}}_c^r$ as an estimate of the first moment of the anchor gradients in the group c and initialize $\hat{\boldsymbol{m}}_c^0 = \boldsymbol{0}$. The estimation of the first moment of the anchor gradients in C groups can be evaluated recursively using

$$\left\{\hat{\boldsymbol{m}}_c^r = \beta_1\hat{\boldsymbol{m}}_c^r + (1-\beta_1)\boldsymbol{g}\left(\hat{\boldsymbol{w}}_c^r\right)\right\}_{c=1}^{C}. \tag{62}$$

We define $\hat{\boldsymbol{v}}_c^r$ as an estimate of the second moment of the anchor gradients in the group c, and $\boldsymbol{g}(\boldsymbol{w}_c^r)^2$ is a vector obtained by componentwise squaring vector $\boldsymbol{g}(\boldsymbol{w}_c^r)$. In practice, we initialize $\hat{\boldsymbol{v}}_c^0 = \boldsymbol{0}$ and the estimated second moment is evaluated recursively using

$$\left\{\hat{\boldsymbol{v}}_c^r = \beta_2\hat{\boldsymbol{v}}_c^{r-1} + (1-\beta_2)\boldsymbol{g}\left(\boldsymbol{w}_c^r\right)^2\right\}_{c=1}^{C}. \tag{63}$$

The central acceleration of the diversified global models in the rth round is designed as

$$\left\{\hat{\boldsymbol{w}}_c^r = \boldsymbol{w}_c^r - \alpha_g^0(1-\beta_1)\cdot\sqrt{\frac{1-\beta_2^r}{1-\beta_2}}\frac{\hat{\boldsymbol{m}}_c^r}{\sqrt{\hat{\boldsymbol{v}}_c^r}+\epsilon}\right\}_{c=1}^{C} \tag{64}$$

where ϵ is a small positive scalar to avoid ill-conditioning. We typically set $\alpha_g^0 = 0.02$, and the decay rates β_1 and β_2 weigh the importance of the past moments relative to the present anchor gradient. Thus, they are always set in the range $(0, 1)$, whose actual values are influential on how quickly the model is updated 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 individual machines is very small in each group, since larger values of β_1 help to pick up a consistent velocity in the direction leading to a promising global model updating. The steps for the adaptive central acceleration of diversified global models after the threshold T, i.e., $r > T$, are summarized in Procedure 4. After the central acceleration of the diversified global models, $\{\hat{\boldsymbol{w}}_c^r\}_{c=1}^{C}$ are the initialized diversified global models for the next round.

Procedure 4 Adaptive Central Acceleration on Diversified Global Models

```
Require: $C$, $S^r$, $n_i$, $n^r$, $\beta_1$, $\beta_2$, $\alpha_g^0$, $\epsilon$, $\{\hat{\boldsymbol{w}}^*_{c,r-1}\}_{c=1}^{C}$, $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^{C}$
Initialize $\hat{\boldsymbol{m}}_c^0 = \boldsymbol{0}$ and $\hat{\boldsymbol{v}}_c^0 = \boldsymbol{0}$ for all $c$
for each round $r$ do
    Calculate the scaling diagonal matrix $\boldsymbol{A}_r$ via (59)
    Update the diversified global models via (60)
    Distribute $\{\boldsymbol{w}_c^r\}_{c=1}^{C}$ to the machines in $S^r$
    Collect the local gradients and calculate the anchor gradients via (61)
    Compute $\{\hat{\boldsymbol{m}}_c^r\}_{c=1}^{C}$ via (62) and $\{\hat{\boldsymbol{v}}_c^r\}_{c=1}^{C}$ via (63)
    Conduct diversified global models acceleration via (64)
    Set $\{\hat{\boldsymbol{w}}_c^r\}_{c=1}^{C}$ as the initialized diversified models for the next round training in Procedure (2) or Procedure (3)
end for
```

E. Algorithm Analysis

The direct manipulation of the current state of the model makes it computationally efficient and straightforward to implement. However, using model parameters to generate diversified global models has its downsides, such as slow adaptation to changes in data distribution and potential overfitting to specific local data distributions. This approach can also be storage-intensive, particularly for large models. MA-FSVRG is preferred in scenarios where stability and consistency are crucial, such as in applications with relatively stable data distributions over time and limited computational resources.

Grouping based on gradients can enhance the coherence of updates, improving the collaboration among participating individual machines and reducing communication overhead, as gradients typically require less bandwidth to transmit than full model parameters. However, gradients can be volatile and noisy, requiring careful management to ensure stable and effective training. Implementing algorithms based on gradient information can also be more complex and pose a risk of gradient leakage, potentially revealing sensitive details about the training data. GA-FSVRG is preferred in scenarios where quick adaptation to changing data distributions and communication efficiency are essential, such as in real-time applications or environments with highly dynamic data.

V. Experiments

A. Local Training and Test Datasets Design

To simulate diversified local demands and reflect realistic FL scenarios with non-IID data distributions, we designed highly heterogeneous local datasets. Each individual machine possesses samples belonging to only one or two categories, ensuring significant variation in local data. We applied the histogram of gradients (HoGs) method to each sample for feature extraction [31]. Using a block size of 7 and a stride of 3, the HoG method resulted in 64 blocks per sample. All possible gradient angles from 0 to 2π were evenly divided into nine bins. The magnitudes of gradients within each block were assigned to the corresponding bins based on their angles, reducing the original 784 features to 576 HoG features for each sample. This transformation provided a more compact and informative representation.

For our experiments, we employed a softmax regression model to handle multiclass classification tasks. At the start of the rth round, the central server maintains $C = \{2, 4, 6, 8\}$ diversified global models. The server also collects local gradients from individual machines in subset S^r for all global models, where S^r is sampled from $\{20\%, 15\%, 10\%, 5\%\}$ with 400 individual machines using the MNIST dataset and from $\{40\%, 30\%, 20\%, 10\%\}$ with 1000 individual machines using the CIFAR-10 dataset. The local iteration number K is adjusted considering the large variance in the size of local datasets. The local learning rate for each machine i is designed as $\alpha_l^i = (\alpha_l/n_i)$ to neutralize the data size difference where n_i is the number of samples on machine i, with $\alpha_l = 12$ for the MNIST dataset and $\alpha_l = 0.08$ for the CIFAR-10 dataset. L_2 regularization with a coefficient of 0.01 is applied to prevent overfitting. To evaluate the performance of the diversified global models after each training round, we compare the test accuracy of the trained global models with the entire test dataset from MNIST and CIFAR-10.

In our performance evaluation, aside from test accuracy, we distinguish between macro-averaged and micro-averaged metrics, each providing unique insights into algorithm effectiveness. Macro-averaging computes the metric for each class individually and then averages these values, treating all classes equally regardless of their dataset frequency. This method is valuable for assessing performance across diverse class distributions, emphasizing the algorithm's ability to handle minority classes effectively. In contrast, micro-averaging aggregates contributions from all classes to compute a single metric, prioritizing performance on more frequent classes. Our study shows that MA-FSVRG and GA-FSVRG consistently achieve high scores in both metrics, demonstrating their robustness and reliability in maintaining accuracy and efficiency across varied conditions and outperforming other methods. We have provided detailed comparison to the state-of-the-art FL algorithms FedProx [6], Personalized FedAvg (PFedAvg) [23], SCAFFOLD [26], MimeSVRG [27], and LoSAC [28].

B. Experimental Results on MNIST Dataset

Fig. 1(a)–(d) shows the test accuracy of algorithms on the MNIST dataset across different participation rates. With 20% participation as shown in Fig. 1(a), MA-FSVRG achieves approximately 98.2% accuracy, while GA-FSVRG reaches 98.5%, outperforming FedProx and PFedAvg. With 15% participation, MA-FSVRG and GA-FSVRG maintain high accuracies around 98.0% and 98.3%, respectively, demonstrating robust performance compared to SCAFFOLD and MimeSVRG as shown in Fig. 1(b). With 10% participation as shown in Fig. 1(c), MA-FSVRG and GA-FSVRG achieve accuracies of approximately 97.8% and 98.1%, respectively, surpassing FedProx and LoSAC. As shown in Fig. 1(d), MA-FSVRG and GA-FSVRG achieve accuracies of around 97.5% and 97.8%, respectively, showing resilience compared to FedProx and PFedAvg with 5% participation. MA-FSVRG and GA-FSVRG consistently outperform other algorithms across varying participation rates in test accuracy, highlighting their effectiveness and adaptability in FL environments.

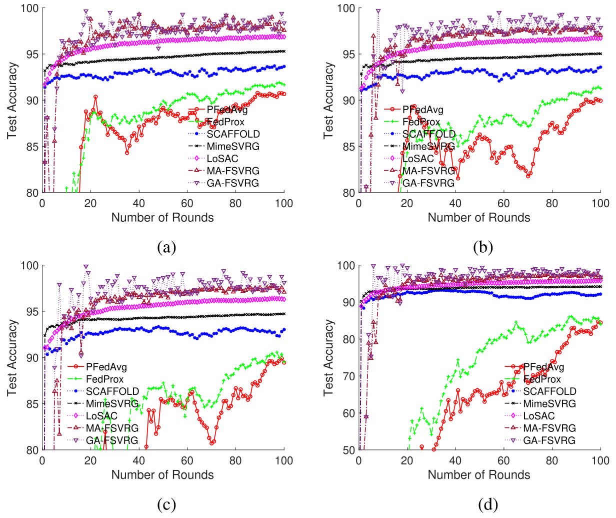


Fig. 1. Performance comparison with test accuracy by decreasing participated individual machines where $\beta_1 = 0.0$, $\beta_2 = 0.999$, and maintain four diversified global models, a threshold is four rounds; (a) 20% individual machine selection; (b) 15% individual machine selection; (c) 10% individual machine selection; (d) 5% individual machine selection.

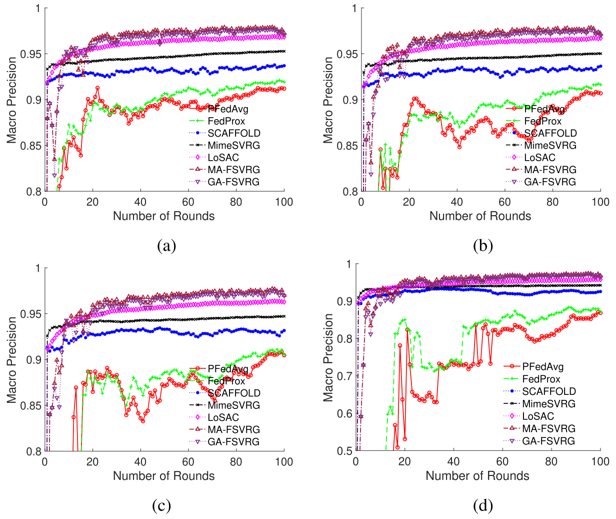


Fig. 2. Performance comparison with macro-averaged precision by decreasing participated individual machines where $\beta_1 = 0.0$, $\beta_2 = 0.999$, and maintain four diversified global models, a threshold is four rounds; (a) 20% individual machine selection; (b) 15% individual machine selection; (c) 10% individual machine selection; (d) 5% individual machine selection.

Fig. 2(a)–(d) displays the macro-averaged precision for the algorithms under varying machine participation rates of 20%, 15%, 10%, and 5%, respectively. In Fig. 2(a), with 20% participation, MA-FSVRG achieves a precision of approximately 0.96, while GA-FSVRG reaches 0.965. FedProx and PFedAvg exhibit lower precision values of about 0.94 and 0.945, respectively. Fig. 2(b) shows that with 15% participation, MA-FSVRG and GA-FSVRG maintain high precisions of around 0.958 and 0.963, respectively, compared to approximately 0.945 for SCAFFOLD and MimeSVRG. At a participation rate of 10% as shown in Fig. 2(c), MA-FSVRG and GA-FSVRG achieve precisions of about 0.955 and 0.96, respectively, while FedProx and LoSAC achieve about 0.94 and 0.943. As shown in Fig. 2(d), with a 5% participation rate, MA-FSVRG and

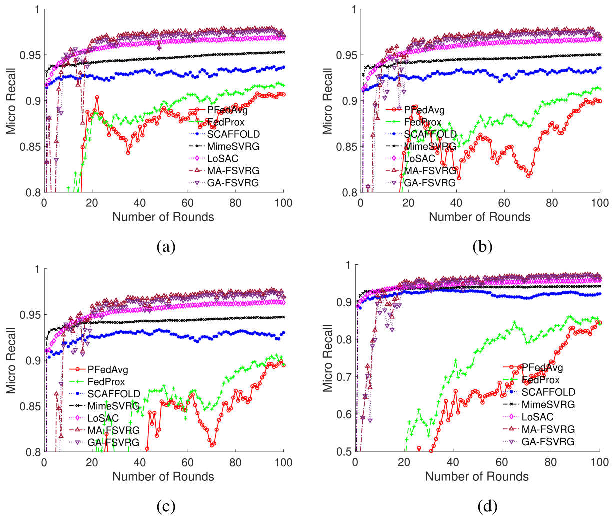


Fig. 3. Performance comparison with micro-averaged recall by decreasing participated individual machines where $\beta_1 = 0.0$, $\beta_2 = 0.999$, and maintain four diversified global models, a threshold is four rounds; (a) 20% individual machine selection; (b) 15% individual machine selection; (c) 10% individual machine selection; (d) 5% individual machine selection.

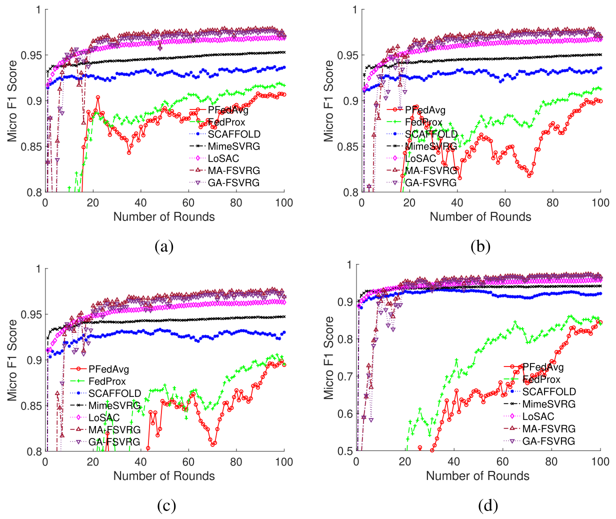


Fig. 4. Performance comparison with micro-averaged $F1$ score by decreasing participated individual machines where $\beta_1 = 0.0$, $\beta_2 = 0.999$, and maintain four diversified global models, a threshold is four rounds; (a) 20% individual machine selection; (b) 15% individual machine selection; (c) 10% individual machine selection; (d) 5% individual machine selection.

GA-FSVRG attain precisions of approximately 0.95 and 0.955, respectively, higher than FedProx and PFedAvg, which reach about 0.935 and 0.94. MA-FSVRG and GA-FSVRG consistently achieve superior precision even under conditions of limited machine participation, making them well-suited for practical deployment in FL environments with resource constraints. Their reliable high precision across different participation rates underscores their effectiveness and resilience. Not only do MA-FSVRG and GA-FSVRG outperform other methods in terms of precision, but they also demonstrate notable adaptability to fluctuating participation rates. This adaptability is crucial for maintaining high-quality performance in FL settings where machine involvement can vary significantly.

Fig. 3(a)–(d) shows the micro-averaged recall for all algorithms across machine participation rates of 20%, 15%, 10%, and 5%, respectively. Micro-averaged recall evaluates the proportion of correctly identified true positives out of all actual positive instances across all participating machines. These higher recall values indicate the ability of MA-FSVRG and GA-FSVRG to accurately identify positive instances. In Fig. 3(a), with 20% participation, MA-FSVRG achieves a recall of approximately 0.97, with GA-FSVRG reaching about 0.975, both highlighting their superior recall performance compared to FedProx and PFedAvg at 0.95 and 0.955, respectively. With 15% participation as shown in Fig. 3(b), MA-FSVRG and GA-FSVRG maintain high recalls of around 0.968 and 0.973, while SCAFFOLD and MimeSVRG achieve lower recalls of approximately 0.955 and 0.958, which demonstrates a significant performance gap, underscoring the robustness and efficiency of MA-FSVRG and GA-FSVRG with fewer participating machines. Similarly, in Fig. 3(c) with 10% participation, MA-FSVRG and GA-FSVRG lead with recalls of about 0.965 and 0.97, compared to FedProx and LoSAC at about 0.95 and 0.955. Even at a minimal 5% participation rate as shown in Fig. 3(d), MA-FSVRG and GA-FSVRG achieve recalls of approximately 0.96 and 0.965, surpassing FedProx and PFedAvg at 0.945 and 0.95. This consistent improvement in recall at lower participation rates underscores the robustness and reliability of MA-FSVRG and GA-FSVRG, showcasing their ability to maintain high recall performance across varying participation rates.

Fig. 4(a)–(d) displays the micro-averaged $F1$ scores for all algorithms under different machine participation rates of 20%, 15%, 10%, and 5%, respectively. The micro-averaged $F1$ score metric integrates both precision and recall, providing a comprehensive evaluation of the algorithms' effectiveness in achieving balanced performance across distributed data settings. In Fig. 4(a), with 20% participation, MA-FSVRG achieves an $F1$ score of about 0.965, while GA-FSVRG reaches approximately 0.97. These values indicate the superior performance of MA-FSVRG and GA-FSVRG compared to other baselines such as FedProx and PFedAvg, which attain lower $F1$ scores of about 0.945 and 0.95, respectively. The higher $F1$ scores of MA-FSVRG and GA-FSVRG reflect their balanced precision and recall, demonstrating their effectiveness in maintaining overall model performance. With 15% participation as shown in Fig. 4(b), MA-FSVRG and GA-FSVRG maintain high $F1$ scores of around 0.963 and 0.968, respectively, while SCAFFOLD and MimeSVRG achieve lower scores of approximately 0.955 and 0.957. This performance gap underscores the robustness and efficiency of MA-FSVRG and GA-FSVRG even with fewer participating machines. Similarly, with a reduced participation rate of 10%, MA-FSVRG and GA-FSVRG continue to lead with $F1$ scores of around 0.96 and 0.965, respectively, compared to FedProx and LoSAC at about 0.945 and 0.95 as shown in Fig. 4(c). Even at a minimal 5% participation rate as shown in Fig. 4(d), MA-FSVRG and GA-FSVRG achieve $F1$ scores of about 0.955 and 0.96, surpassing FedProx and PFedAvg at 0.94 and 0.945, which highlights the robustness and reliability of MA-FSVRG and GA-FSVRG.

C. Experimental Results on CIFAR-10 Dataset

Fig. 5(a)–(d) depicts the macro-averaged recalls for all algorithms across different machine participation rates of 40%, 30%, 20%, and 10%, respectively, using the CIFAR-10 dataset. Macro-averaged recall evaluates the average recall across all classes, providing insight into the model's overall ability to correctly identify relevant instances. In Fig. 5(a), with 40%

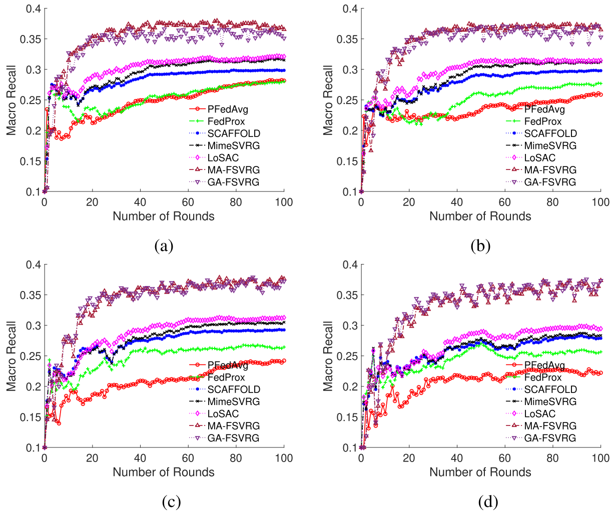


Fig. 5. Performance comparison with macro-averaged recall by decreasing participated individual machines where $\beta_1 = 0.9$, $\beta_2 = 0.999$, and maintain three diversified global models, a threshold is four rounds; (a) 40% individual machine selection; (b) 30% individual machine selection; (c) 20% individual machine selection; (d) 10% individual machine selection.

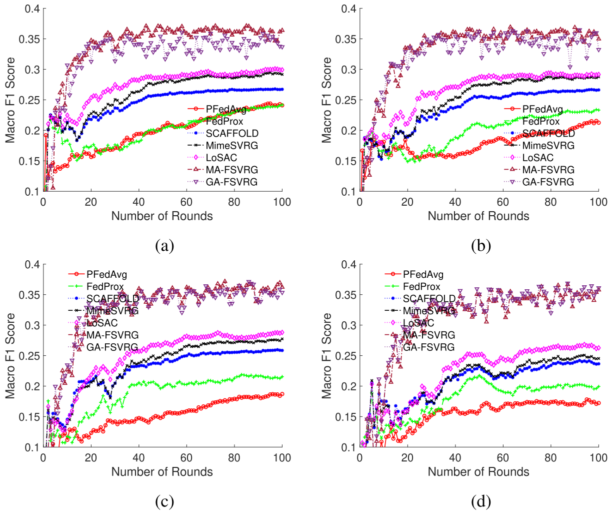


Fig. 6. Performance comparison with macro-averaged $F1$ score by decreasing participated individual machines where $\beta_1 = 0.9$, $\beta_2 = 0.999$, and maintain three diversified global models, a threshold is four rounds; (a) 40% individual machine selection; (b) 30% individual machine selection; (c) 20% individual machine selection; (d) 10% individual machine selection.

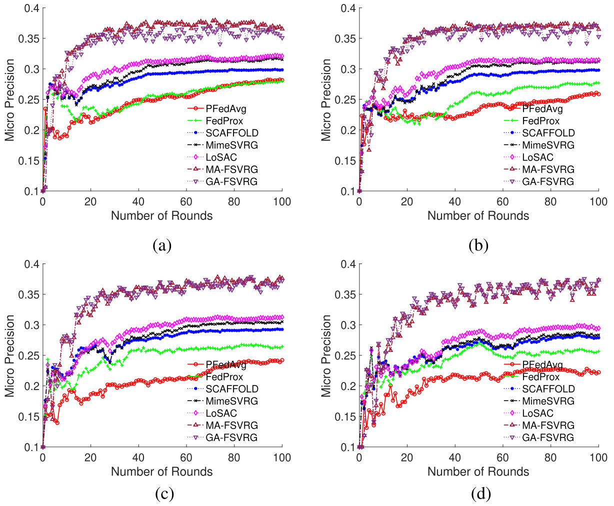
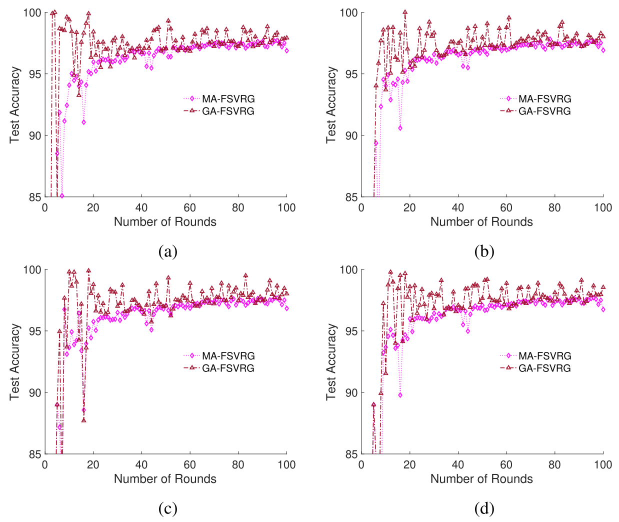


Fig. 7. Performance comparison with micro-averaged precision by decreasing participated individual machines where $\beta_1 = 0.9$, $\beta_2 = 0.999$, and maintain three diversified global models, a threshold is four rounds; (a) 40% individual machine selection; (b) 30% individual machine selection; (c) 20% individual machine selection; (d) 10% individual machine selection.

participation, MA-FSVRG and GA-FSVRG achieve recalls of approximately 0.87 and 0.88, surpassing other algorithms like FedProx and PFedAvg with recalls around 0.84 and 0.85, and SCAFFOLD and MimeSVRG with approximately 0.82 and 0.83. Moving to 30% participation in Fig. 5(b), MA-FSVRG and GA-FSVRG maintain recalls of around 0.85 and 0.86, outperforming FedProx and PFedAvg which achieve about 0.82 and 0.83, respectively. With 20% participation in Fig. 5(c), MA-FSVRG and GA-FSVRG achieve recalls of approximately 0.83 and 0.84, compared to FedProx and PFedAvg with about 0.80 and 0.81. This consistency highlights the robustness of MA-FSVRG and GA-FSVRG in maintaining high recall rates despite fewer participating machines. As shown in Fig. 5(d) even with 10% participation, MA-FSVRG and GA-FSVRG achieve recalls of about 0.80 and 0.81, respectively, compared to FedProx and PFedAvg with approximately 0.77 and 0.78. The significant difference in recall at this low participation rate emphasizes the reliability of MA-FSVRG and GA-FSVRG in capturing relevant data points despite limited machine involvement, where maintaining high macro-averaged recall in FL across varying participation rates is essential for robust deployment in real-world scenarios.

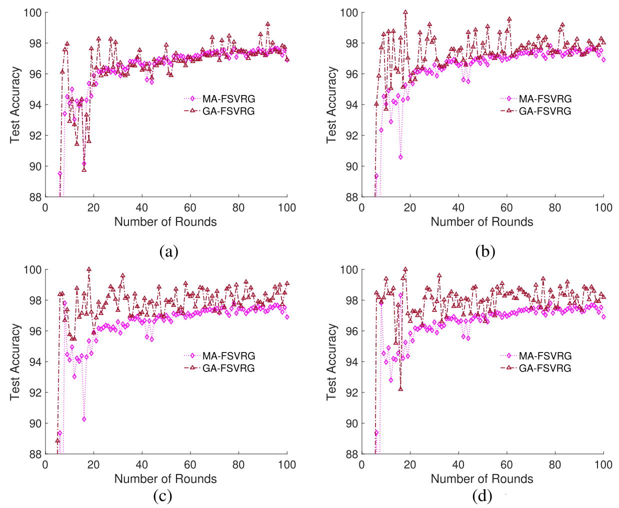
The macro-averaged $F1$ score combines precision and recall across all classes, providing a comprehensive measure of the model's ability to balance true positives and negatives. As shown in Fig. 6(a), with 40% participation, MA-FSVRG and GA-FSVRG achieve macro-averaged $F1$ scores of approximately 0.86 and 0.87, respectively, outperforming FedProx and PFedAvg with $F1$ scores around 0.83 and 0.84, and SCAFFOLD and MimeSVRG with about 0.81 and 0.82, using the CIFAR-10 dataset. With 30% participation, as shown in Fig. 6(b), MA-FSVRG and GA-FSVRG achieve $F1$ scores of around 0.84 and 0.85, surpassing FedProx and PFedAvg at approximately 0.81 and 0.82, demonstrating their resilience and consistent high performance as participation decreases. In Fig. 6(c), MA-FSVRG and GA-FSVRG maintain their lead with $F1$ scores around 0.82 and 0.83, compared to FedProx and PFedAvg with about 0.79 and 0.80 with 20% participation, showcasing their adaptability in varied FL scenarios. As shown in Fig. 6(d) where even with 10% participation, MA-FSVRG and GA-FSVRG achieves $F1$ scores of approximately 0.79 and 0.80, surpassing FedProx and PFedAvg with scores around 0.76 and 0.77, highlighting their reliability and effectiveness even with limited machine participation, making them suitable for practical deployment in FL environments with resource constraints.

As shown in Fig. 7(a), MA-FSVRG and GA-FSVRG achieve micro-averaged precisions of approximately 0.86 and 0.87, respectively, outperforming FedProx and PFedAvg with precisions around 0.83 and 0.84, and SCAFFOLD and MimeSVRG with about 0.81 and 0.82, using the CIFAR-10 dataset with 40% participation. These results highlight the superior micro-averaged precision of MA-FSVRG and GA-FSVRG in correctly identifying relevant instances while minimizing false positives. With 30% participation, as shown in Fig. 7(b), MA-FSVRG and GA-FSVRG achieve micro-averaged precisions of around 0.84 and 0.85, surpassing



(a) (b) (c) (d)

Fig. 8. Performance comparison with test accuracy by different thresholds of the initial rounds where $\beta_1 = 0.0$, $\beta_2 = 0.999$, and maintain four diversified global models. (a) Two rounds. (b) Four rounds. (c) Six rounds. (d) Eight rounds.



(a) (b) (c) (d)

Fig. 9. Performance comparison with test accuracy by diversified anchor gradients where $\beta_1 = 0.0$, $\beta_2 = 0.999$ with 10% individual machine selection and threshold is four rounds. (a) Two groups. (b) Four groups. (c) Six groups. (d) Eight groups.

FedProx and PFedAvg at approximately 0.81 and 0.82, demonstrating their robustness and consistent high precision despite reduced participation. In Fig. 7(c), MA-FSVRG and GA-FSVRG maintain their lead with precisions of around 0.82 and 0.83, compared to FedProx and PFedAvg with about 0.79 and 0.80 with 20% participation. With only 10% participation, Fig. 7(d) shows MA-FSVRG and GA-FSVRG achieving precisions of approximately 0.79 and 0.80, outperforming FedProx and PFedAvg with scores around 0.76 and 0.77, underscoring their reliability and effectiveness in FL environments with limited resources. The high micro-averaged precisions of MA-FSVRG and GA-FSVRG across varying participation rates highlight their superior precision in correctly identifying relevant instances while minimizing false positives in dynamic FL settings.

D. Impact of Different Thresholds of Grouping

Since each machine is still exploring its model updating directions at the beginning of the federated training procedure, and due to the limited machine participation, it is hard to collect effective information of the global model updating at the beginnings. Thus, we conduct the following experiment to explore the impact of different thresholds of federated rounds before the development of the diversified global updating.

We focus on the performance comparison of the proposed MA-FSVRG and GA-FSVRG. As shown in Fig. 8 with 10% machines selected, both MA-FSVRG and GA-FSVRG are initialized with 2–8 federated rounds before the grouping mechanism to generate diversified global models and anchor gradients, respectively. The larger the threshold to start the diversified global updating, the smaller the overall variance in the achieved test accuracy, especially with MA-FSVRG whose performance is greatly impacted by the quality of the local models. With different thresholds to start the grouping mechanism, GA-FSVRG can outperform MA-FSVRG in the achieved test accuracy, but MA-FSVRG is better with stabler performance. By increasing the threshold to conduct the diversified global updating, the convergence speed of MA-FSVRG is increased, however, the convergence speed of GA-FSVRG is slightly deteriorated at the early training stage. However, with a larger threshold as 8 rounds shown in Fig. 8(d), the performance of GA-FSVRG can be stabilized to a higher level of test accuracy by the increased threshold. The threshold can be adjusted by different applications which have different tolerance of the variance in performance. The larger threshold can guarantee a more stable initial training procedure, especially with limited machine participation.

E. Impact of Number of Diversified Global Models

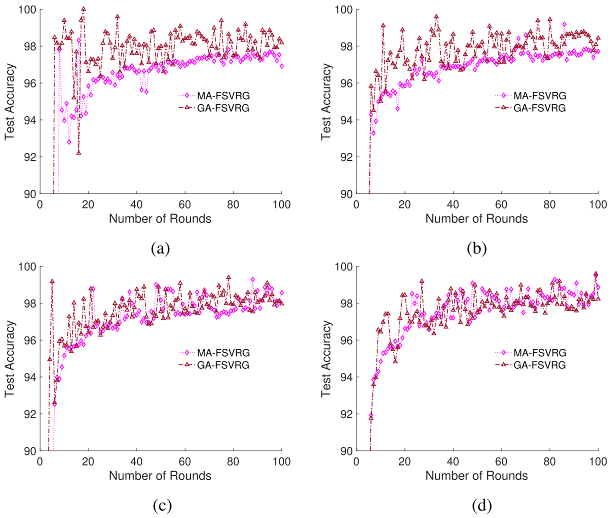
One of the important parameters in diversified global updating is the number of groups the central server managed which leads to diversified global model acceleration. We evaluate the performance of the proposed methods with increasing number of global models from 2 to 8 as shown in Fig. 9. With the increasing number of diversified global models, the performance of GA-FSVRG is extensively improved, however, the variance of the performance is also increased. The performance of MA-FSVRG is much smoother compared with that of GA-FSVRG, but it cannot achieve the same test accuracy level as GA-FSVRG. Furthermore, the performance gap between MA-FSVRG and GA-FSVRG on the test accuracy enlarged with the increasing number of groups. The convergence speed for both MA-FSVRG and GA-FSVRG are accelerated with increasing group numbers thanks to the multiple global models acceleration in the central server. Although it can achieve higher test accuracy at the early stage with more groups in the central server, the high-quality performance is not stable. After 30 rounds, the influence of increased group number is limited on MA-FSVRG, but is large on the performance of GA-FSVRG. Therefore, GA-FSVRG has the advantage to converge faster and achieve higher test accuracy compared with MA-FSVRG, and the performance of MA-FSVRG is stabler.

F. Impact of Central Adaptive Parameters

Finally, we explore the impact of different parameters β_1 for the first-order moments of the anchor gradients on different diversified global model updating in the central server. As shown in Fig. 10, we check the first-order moment parameter β_1 from the set $\{0.0, 0.5, 0.8, 0.9\}$. With increasing β_1,

TABLE I
CENTRAL COMPUTATION COST (S)

Participation Diversity	MA-FSVRG				GA-FSVRG			
	10%	20%	30%	40%	10%	20%	30%	40%
2 **Global Models**	10.14	20.13	29.31	39.30	10.66	20.74	31.15	41.32
4 **Global Models**	20.27	38.27	56.88	74.94	20.69	40.33	59.69	79.01
6 **Global Models**	29.19	57.32	84.67	110.64	29.63	58.16	86.84	113.04
8 **Global Models**	38.75	74.31	111.52	147.30	38.62	75.05	111.47	147.73



(a)

(b)

(c)

(d)

Fig. 10. Performance comparison with test accuracy by different β_1 where $\beta_2 = 0.999$ with 10% individual machine selection and threshold is four rounds with four groups. (a) $\beta_1 = 0.0$. (b) $\beta_1 = 0.5$. (c) $\beta_1 = 0.8$. (d) $\beta_1 = 4$ 0.9.

the performance of MA-FSVRG is improved a lot. Although smaller β_1 helps to improve the convergence at the early stage of FL training for both MA-FSVRG and GA-FSVRG, the variance of the performance is also larger compared with that with larger β_1. However, the increasing β_1 also causes more variance into the performance after 30 rounds, where the drawback impact on MA-FSVRG is greater than that of GA-FSVRG.

GA-FSVRG achieves the best performance compared with MA-FSVRG with $\beta_1 = 0$, where the test accuracy can be stabled to around 98% during the first 10 federated rounds. However, when the first-order moment parameter β_1 increases to 0.9, the performance of GA-FSVRG can converge to the same level after 20 rounds, as shown in Fig. 10(d), which is bouncing around below 98% during the first 20 rounds. However, the influence of the first-order moments on MA-FSVRG has more advantages compared that of GA-FSVRG. It is obvious that smaller β_1 can achieve stabler performance with MA-FSVRG, but the test accuracy performance is improved with the increasing value of β_1.

However, the advantage of the first-order moments becomes to disadvantage in the diversified global model for the grouping mechanism based on anchor gradients except the stabler performance. The first-order moments are useful with unified global model federated training thanks to its ability to enhance the global updating trends to avoid the jitters during the training procedure. However, this ability is harmful to diversified global model updating for GA-FSVRG, due to that first-order moments enhanced the updating trends by eliminating the variance of anchor gradients which also reduces the diversity of the anchor gradients.

G. Central Server Computation Time Analysis

The central computation cost for the proposed MA-FSVRG and GA-FSVRG algorithms is measured using an experimental platform equipped with an 8-core CPU, a 14-core GPU, and 16 GB of RAM. This analysis provides insights into the scalability of these algorithms under varying conditions of model diversity and individual machine participation rates. The results are summarized in Table I, which reports the central server computation times in seconds for different scenarios.

When analyzing MA-FSVRG, we observe its behavior concerning model diversity and individual machine participation rates. Initially, with two global models, the computation time scales linearly from 10.14 s at 10% individual machine participation to 39.30 s at 40% participation. This trend becomes more pronounced as the number of global models increases, i.e., with four models, computation time rises notably from 20.27 to 74.94 s. The pattern continues with six and eight global models, where computation times escalate from 29.19 to 110.64 s and from 38.75 to 147.30 s, respectively. Regarding individual machine participation, increasing from 10% to 40% results in approximately a fourfold increase in computation time across all levels of model diversity. This underscores the direct impact of individual machine engagement on the computational workload of the central server.

Similar to MA-FSVRG, GA-FSVRG demonstrates consistent behaviors in terms of model diversity and individual machine participation rates. With two global models, computation times range from 10.66 s at 10% participation to 41.32 s at 40% participation. Increasing the number of global models to four leads to computation times escalating from 20.69 to 79.01 s. For six and eight global models, the computation times span from 29.63 to 113.04 s and from 38.62 to 147.73 s, respectively. The impact of individual machine participation is similarly pronounced in GA-FSVRG, showing a proportional increase in computation time from 10% to 40% participation rates.

VI. CONCLUSION AND DISCUSSIONS

In this article, we proposed the adaptive central accelerated FL with diversified global updating to tackle the challenges in heterogeneous demands of various individual machines. It can not only show a faster convergence rate in the training procedure but also can achieve higher test accuracy compared with the state-of-the-art FL baseline algorithms. Two different diversified global updating methods are proposed, i.e., MA-FSVRG and GA-FSVRG, where MA-FSVRG can achieve stabler performance with cheaper local computing cost and GA-FSVRG can converge faster to a higher test accuracy.

Regardless of the promising performance of the current work, there are several future research issues beckoning further investigation. Our future research should focus on reducing communication overhead through efficient protocols like gradient compression and quantization, and addressing synchronous communication inefficiencies by exploring asynchronous training strategies. While we have streamlined our current study to focus on our main contributions, we acknowledge the importance of investigating the defensive capabilities of diverse global models. To address the feasibility of this defense mechanism, future research could focus on the investigation of the integration of adaptive defense mechanisms that leverage model diversity to detect and isolate compromised models.

References

[1] S. Zhou and G. Y. Li, "Federated learning via inexact ADMM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9699–9708, Aug. 2023.

[2] Y. Xue and V. Lau, "Riemannian low-rank model compression for federated learning with over-the-air aggregation," *IEEE Trans. Signal Process.*, vol. 71, pp. 2172–2187, 2023.

[3] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, Jun. 2021.

[4] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4615–4625.

[5] J. Sun, T. Chen, G. B. Giannakis, Q. Yang, and Z. Yang, "Lazily aggregated quantized gradient innovation for communication-efficient federated learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2031–2044, Apr. 2020.

[6] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[7] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9587–9603, Dec. 2023.

[8] Q. Wang, H. Yin, T. Chen, J. Yu, A. Zhou, and X. Zhang, "Fast-adapting and privacy-preserving federated recommender system," *VLDB J.*, vol. 31, pp. 877–896, Oct. 2021.

[9] D. Nawara and R. Kashef, "IoT-based recommendation systems—An overview," in *Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS)*, Sep. 2020, pp. 1–7.

[10] M. Schmidt, N. L. Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Math. Program.*, vol. 162, nos. 1–2, pp. 83–112, Mar. 2017.

[11] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[12] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.

[13] Y. Deng, M. Mahdi Kamani, and M. Mahdavi, "Adaptive personalized federated learning," 2020, *arXiv:2003.13461*.

[14] Y. Jiang, J. Konečný, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," 2019, *arXiv:1909.12488*.

[15] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, "A secure federated transfer learning framework," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 70–82, Jul. 2020.

[16] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "FedHealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Jul./Aug. 2020.

[17] K. I. Wang, X. Zhou, W. Liang, Z. Yan, and J. She, "Federated transfer learning based cross-domain prediction for smart manufacturing," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4088–4096, Jun. 2022.

[18] M. Ghuhan Arivazhagan, V. Aggarwal, A. Kumar Singh, and S. Choudhary, "Federated learning with personalization layers," 2019, *arXiv:1912.00818*.

[19] H. Zhu, H. Zhang, and Y. Jin, "From federated learning to federated neural architecture search: A survey," *Complex Intell. Syst.*, vol. 7, no. 2, pp. 639–657, Apr. 2021.

[20] C. He, E. Mushtaq, J. Ding, and S. Avestimehr, "FedNAS: Federated deep learning via neural architecture search," in *Proc. CVPR Workshop Neural Architecture Search Beyond Represent. Learn.*, 2020.

[21] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21394–21405.

[22] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.

[23] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," 2020, *arXiv:2002.07948*.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Jan. 2009.

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

[26] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, Jul. 2020, pp. 5132–5143.

[27] S. Praneeth Karimireddy et al., "Mime: Mimicking centralized stochastic algorithms in federated learning," 2020, *arXiv:2008.03606*.

[28] H. Chen, H. Wang, Q. Yao, Y. Li, D. Jin, and Q. Yang, "LoSAC: An efficient local stochastic average control method for federated optimization," *ACM Trans. Knowl. Discovery Data*, vol. 17, no. 4, pp. 1–28, May 2023.

[29] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 29.

[30] K. Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations I," *Proc. Nat. Acad. Sci. USA*, vol. 35, no. 11, pp. 652–655, Nov. 1949.

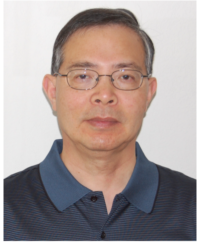
[31] W.-S. Lu, "Handwritten digits recognition using PCA of histogram of oriented gradient," in *Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process. (PACRIM)*, Aug. 2017, pp. 1–5.

Lei Zhao (Member, IEEE) 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 the Ph.D. degree in electrical and computer engineering from the University of Victoria, Victoria, BC, Canada, in 2023.

He is currently a Post-Doctoral Fellow with the Department of Electronics and Computer Engineering, University of Victoria. His research interests include federated learning and optimization with applications in finance.

Lin Cai (Fellow, IEEE) has been with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada, since 2005, and she is currently a Professor. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia traffic and the Internet of Things.

Prof. Cai is an NSERC E.W.R. Steacie Memorial Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Royal Society of Canada Fellow. She has been elected to serve on the board of the IEEE Vehicular Technology Society from 2019 to 2024 and as its VP in Mobile Radio. She has been a Board Member of the IEEE Women in Engineering from 2022 to 2024 and the IEEE Communications Society from 2024 to 2026. She served as an Associate Editor-in-Chief for IEEE Transactions on Vehicular Technology and a Distinguished Lecturer for the IEEE VTS Society and the IEEE Communications Society.



Wu-Sheng Lu (Life Fellow, IEEE) received the B.Sc. degree in mathematics from Fudan University, Shanghai, China, in 1964, and the M.S. degree in electrical engineering and the Ph.D. degree in control science from the University of Minnesota, Minneapolis, MN, USA, in 1983 and 1984, respectively.

Since 1987, he has been with the University of Victoria, Victoria, BC, Canada, and is now a 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).