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*—Diversified Global Models and Anchor Gradients, Adaptive Central Acceleration**

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 data set.

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 still based on the unique shared global model and the personality is obtained by the trade-off between the global model and the diversified local needs [7] [8] [9]. This paper 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 successful participation of all machines. Therefore, the intrinsic requirements 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 client selection, which involves the selection of a subset of individual machines for local updates.

Nonetheless, 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 client selection. Additionally, 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 client selection.

The main contribution of this paper 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 mitigating the high variance in global model updates and accelerating 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

L. Zhao, L. Cai, and W-S. Lu are with Dept. of Electrical & Computer Engineering, University of Victoria, 3800 Finnerty Road, Victoria, BC, V8P 5C2, Canada. Corresponding author: Lin Cai (E-mail: cai@ece.uvic.ca). This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and Compute Canada.

models 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 paper is organized as follows. The related work 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 towards individual machines with diversified objectives. Solely optimizing the accuracy of the global model tends to have negative impact on its capacity to personalize [13] [3]. 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 trade-off 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 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 personal information of individual machines [18] [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 [23] [14]. These works 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 client data, a data-sharing strategy is also proposed by [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 trade-off 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 still based on the unique shared global model and the personality is obtained by the trade-off between the global model and the diversified local needs. Therefore, it will be 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 ensuring convergence to a high-quality global model with a diverse set of individual machine updates is still one open issue.

In this work, we propose the adaptive central acceleration on multiple global models generated by model-based and gradient-based grouping to provide the diversified training procedure. Diversified global models can not only promotes fairness and inclusivity in the training process, but also can avoid the trade-off between the performance and the heterogeneous local demands. Except mitigating the impact of individual machine heterogeneity, the diversified global models based approach can also enhance the system's resilience to attacks from malicious individual machines since the grouping mechanism by individual machines helps prevent malicious actors from significantly influencing the overall performance.

III. Federated Learning 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 data set for the i-th individual machine with n_i local training samples for $i = 1, 2, \cdots, |E|$. There is no overlap among different local data sets, 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 i-th local objective function, which is the average of the empirical loss over all local samples $\{F_k(w)\}_{k\in P_i}$:

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

TABLE I
NOTATIONS IN FEDERATED LEARNING ARCHITECTURE

NOTATION	DESCRIPTION
E	The set of all individual machines
P_i	The local data set for the i-th individual machine
n_i	The number of local samples in individual machine i
n	The total number of samples in set E
$f(\boldsymbol{w})$,	The global objective function
$f_i(\boldsymbol{w})$	The local objective function of machine i
$\nabla f(\boldsymbol{w}_g)$	Global gradient w.r.t. $\boldsymbol{w}_g$
$\nabla f_i(\boldsymbol{w})$	Local gradient of machine i
$\boldsymbol{g}_i(\boldsymbol{w})$	The local gradient estimation of machine i
$\Delta \boldsymbol{w}_g$	The updating in global model
$\boldsymbol{w}^*$	The optimal global model
d	The number of model parameters
L_i	The largest eigenvalue of $\nabla^2 \boldsymbol{f_i}(\boldsymbol{w})$
μ_i	The smallest eigenvalue of $\nabla^2 \boldsymbol{f_i}(\boldsymbol{w})$

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 is summarized in Table I.

B. Federated Learning Architecture with Reduced Variance

With both stochastic client 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 $\boldsymbol{g_i}(\boldsymbol{w})$ to represent the local gradient from individual training sample or a batch of the training samples which is an unbiased estimation of $\nabla f_i(\boldsymbol{w})$. The variance of local stochastic gradient $\boldsymbol{g_i}(\boldsymbol{w})$ can be analyzed as follows. According to Jensen's inequality, we obtain the upper bound of the variance of the sampled local gradient as

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

which can be written as

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

where $\boldsymbol{w}^*$ denotes the optimal model. According to Appendix A.1 and (4), we can obtain

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

where L_i refers to the largest eigenvalue of $\nabla^2 \boldsymbol{f_i}(\boldsymbol{w})$. Due to the heterogeneous local data sets 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 the global model, $\Delta \boldsymbol{w}_g$ to denote the updating in global model. And similarly to the analysis in Appendix A.1, the global objective function can be upper bounded by

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

and lower bounded by

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

where the global gradient is defined as

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

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

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

There are discrepancy 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

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

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 $E[\boldsymbol{g_i}(\boldsymbol{w}) - \boldsymbol{g_i}(\boldsymbol{w_g})] = \boldsymbol{0}$, which leads to

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

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

$$\text{Var}(\boldsymbol{g_i}(\boldsymbol{w}) - \boldsymbol{g_i}(\boldsymbol{w_g}) + E[\boldsymbol{g}(\boldsymbol{w_g})]) = \text{Var}(\boldsymbol{g_i}(\boldsymbol{w}) - \boldsymbol{g_i}(\boldsymbol{w_g})),$$

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

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

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

$$\text{Var}(\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i(\boldsymbol{w}_g)) \leq \mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i(\boldsymbol{w}_g)||^2]. \tag{12}$$

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

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

Since the stochastic local gradients are unbiased to the full local gradients , and according to Appendix A.1, we obtain

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

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

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

But as long as $\boldsymbol{w} \neq \boldsymbol{w_g}$, 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

At the beginning of the federated training, without the knowledge of others, 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 r-th 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

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

where n^r denotes the number of samples from all individual machines in subset S^r and n_i for individual machine i. Note that, 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 data sets 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 improve the training efficiency, individual machines need to collaborate based on their preference. The grouping procedure is illustrated in Fig. 1. And all notations defined in algorithms design are listed in Table II.

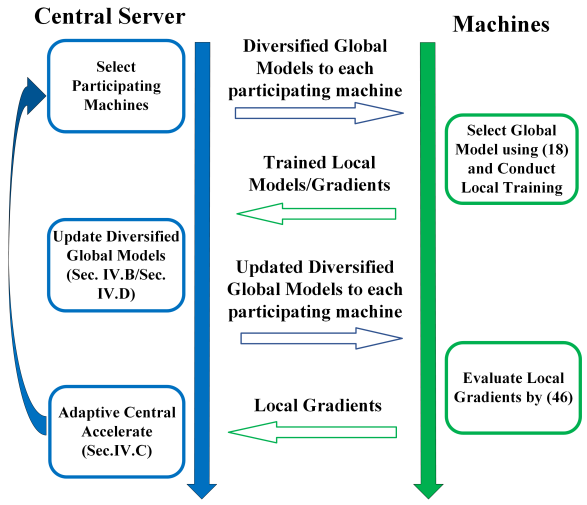


Fig. 1. Illustration of Grouping Procedure with Adaptive Central Acceleration.

A. Local Training with Reduced Gradient Variance

The central server randomly select a subset S^r of machines at the beginning of the r-th round. And the C global models $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^C$ are distributed to each participated machine to initialize the local training.

TABLE II
Notations in Algorithm Design

NOTATION	DESCRIPTION
T	The threshold of federated round to begin grouping mechanism
S^r	The participated individual machine subset in the r-th round
n^r	The total number of samples in subset S^r
C	The number of diversified global models
α_l^i	Local learning rates of machine i in l-th 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
$\hat{\boldsymbol{w}}_{i,K}^r$	The trained local model from machine i in round r
$\hat{\boldsymbol{w}}_{c,r-1}$	The global model representation of group c in round r
S_c^r	The subset of machines of group c in round r
$\boldsymbol{g}_c^{r-1}$	The anchor gradient of group c in round r
$\boldsymbol{\Lambda}_i$	Local gradient adjustment matrix for machine i
$\boldsymbol{d}_{i,k-1}^r$	The k-th local updating direction for machine i in round r
$\bar{\boldsymbol{w}}_i^{r-1}$	The mean-centered trained local model of machine i
$\boldsymbol{W}_{r-1}$	The subset of the mean-centered local models in S^r
$\nabla \bar{f}_i(\hat{\boldsymbol{w}}_{i,K}^r)$	The mean-centered local gradient of machine i
$\boldsymbol{G}_{r-1}$	The subset of the mean-centered local gradients in S^r
c_i^l	The group that machine i prefers in the l-th round of grouping updating
$\hat{\boldsymbol{w}}_{i,0}^{r-1}$	The initialized local model for machine i in round r
$\tilde{\boldsymbol{w}}_i^{r-1}$	The transformed trained local model of machine i
$\tilde{\boldsymbol{w}}_{c,r-1}^l$	The c-th global model in l-th grouping updating
$\hat{\boldsymbol{w}}_{c,r-1}^*$	The c-th global model after grouping update in round r
$\tilde{\boldsymbol{g}}_{c,r-1}^l$	The c-th anchor gradient in l-th grouping updating
$\tilde{\boldsymbol{g}}_{c,r-1}^*$	The c-th anchor gradient after grouping update in round r
$S_{c,l}^r$	The machine subset in group c in l-th grouping updating
$\hat{\boldsymbol{w}}_c^r$	The c-th central accelerated global model
$\boldsymbol{A}_r$	The scaling diagonal matrix for model aggregation in round r

Furthermore, at the beginning of r-th round, the central server manages C diversified global models, denoted by $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^C$ and the local gradients collected from the individual machines in subset S^r for all global models in $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^C$. The central server collects these local gradients to obtain multiple anchor gradients $\{\boldsymbol{g}_c^{r-1}\}_{c=1}^C$ w.r.t. the groups $\{S_c^r\}_{c=1}^C$ as

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

The anchor gradient $\{\boldsymbol{g}_c^{r-1}\}_{c=1}^C$ is distributed to the selected individual machines in S^r to reduce the variance in local training.

At the beginning of the local training, 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 data sets and select the global model based on their own preference as

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

The individual machines in S_c^r conduct K local stochastic updates based on their own local data sets. To enforce the auxiliary local gradient to be of the correct magnitude, it is scaled carefully by the number of non-zero features of the samples. The number of samples in the local data set of

individual machine i with nonzero j-th feature is denoted by n_i^j. After going through their local data sets, individual machines send the number of local nonzero j-th feature $\{n_i^j\}_{i\in E}$ to the center, and the center generates the number of samples with nonzero j-th feature over all local data sets as

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

The variance between the gradient w.r.t. 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

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

where

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

The local updating direction is designed as

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

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 data sets. The local iteration number, K, should be re-visited 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 $\alpha_l^i = \frac{\alpha_l}{n_i}$ to neutralize the data size difference. The local model update of the i-th 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

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

B. Model-based Grouping Mechanism

We first propose to group the machines for collaboration by measuring their current local models' similarities. To relieve the communication burden, the grouping procedure is conducted in the central server. At the r-th round, the central server collects the local models from the randomly selected individual machines $\{\hat{\boldsymbol{w}}_{i,K}^{r-1}\}_{i\in S^r}$, which represent the corresponding individual machines to select the collaborative partners where collaboration groups are formed. The collaboration groups are formed by grouping based on the similarity, which is interpreted as preference of individual machines.

The objective in the center is formulated as increasing the local models' similarity within the same group, which can be formulated as

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

The mean-centered local model parameters of machine i in the current round is 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

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

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 machines 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

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

We can define $\boldsymbol{H}_{r-1} = \{\boldsymbol{h}_c\}_{c=1}^C$ to refer the assignment of machines into different groups, where the c-th column $\boldsymbol{h}_c$ is the indicator vector of the local models in $\boldsymbol{W}_{r-1}$ which belong to the c-th group scaled by $\frac{1}{\sqrt{|S_c^r|}}$. 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

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

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 $\boldsymbol{t}_C = \{\sqrt{\frac{|S_c^r|}{|S^r|}}\}_{c=1}^C$, and we can get

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

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 $c = 1, \cdots, C-1$ according to connectivity analysis [26]. We define $\hat{\boldsymbol{Q}}_{\boldsymbol{r-1}}$ as the matrix involve the first $C-1$ columns in $\boldsymbol{Q}_{r-1}$, and according to the property of trace, we can obtain

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

Since

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

(30) can be rewritten as

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

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 [27]. 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 our global 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 and the transformation between $\hat{\boldsymbol{Q}}_{r-1}^*$ and $\boldsymbol{H}_{r-1}^*$ is hard to obtain [26]. 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.

Since our target is not only to grouping the local models, but also to apply the grouping information to update the diversified global models. The grouping information in the lower-dimensional subspace need 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 an orthogonal matrix, and $\hat{\boldsymbol{\Lambda}}_{r-1} = \text{diag}\{\lambda_c\}_{c=1}^{C-1}$, and the decomposition of $\boldsymbol{W}_{r-1}^T\boldsymbol{W}_{r-1}$ related to the first $C-1$ principle components can be written as

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

According to (37), the more representative local models used for grouping defined as $\hat{\boldsymbol{\Lambda}}_{r-1}^{\frac{1}{2}}\hat{\boldsymbol{V}}_{r-1}^T$ can be transformed to

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

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

The initial grouping of the trained local models comes from individual machines' preference of the previous global models $\{\hat{\boldsymbol{w}}_{c,r-1}\}_{c=1}^C$, which are projected to the same lower-dimensional space as

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

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

In the grouping procedure, each machine measures its preference of the current global models by the Euclidean distance between its local model and global models. The preference of individual machines is obtained by

$$\left\{c_i^l = \underset{c=1,\cdots,C}{\text{argmin}} \quad ||\tilde{\boldsymbol{w}}_i^{r-1} - \tilde{\boldsymbol{w}}_{c,r-1}^l||\right\}_{i\in S^r}, \tag{40}$$

where c_i^l denotes the group that machine i prefers in the l-th round of grouping updating. All individual machines with the same preference $\{c_i^l = c\}_{i\in S^r}$ of the global models form group $S_{c,l}^r$. To obtain the most powerful representation of different grouping, the diversified multiple global models should be updated as

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

in the l-th round of grouping updating in the center server. When the grouping updating in the center server converged, i.e.,

$$\sum_{c=1}^{C}\sum_{i\in S_{c,l}^r} ||\tilde{\boldsymbol{w}}_i^{r-1} - \tilde{\boldsymbol{w}}_{c,r-1}^l|| \le \tilde{\epsilon}, \tag{42}$$

where $\tilde{\epsilon} = 10^{-4}$. We define $\{\tilde{\boldsymbol{w}}_{c,r-1}^* = \tilde{\boldsymbol{w}}_{c,r-1}^l\}_{c=1}^C$ and $\{S_c^r = S_{c,l}^r\}_{c=1}^C$. To conduct the diversified global model updating in the original space which is used for the local training, the following procedure need to be conducted in the center server as

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

Then, we can regard the final grouping results denoted by $\{\hat{\boldsymbol{w}}_{c,r-1}^*\}_{c=1}^C$ as the current optimal representations of the preference of the selected individual machines in multiple groups denoted by $\{S_c^r\}_{c=1}^C$, and the adaptive acceleration based on $\{\hat{\boldsymbol{w}}_{c,r-1}^*\}_{c=1}^C$ are designed in Section IV-C.

C. Adaptive Central Acceleration on Diversified Global Models

For each group, new global models $\{\hat{\boldsymbol{w}}_{c,r-1}^*\}_{c=1}^C$ are scaled based on whether the specific feature appears in one local data set or not, since the updating to the gradient related to the features that are less collected with individual machines should be enhanced. The scaling diagonal matrix for model aggregation is defined as

$$\boldsymbol{A}_r = \text{diag}\left(\left\{\frac{|S^r|}{\sum_{i\in E}1_{n_i^j\neq 0}}\right\}_{j=1,\cdots,q}\right), \tag{44}$$

to enhance the updating related to the features less collected by machines. The multiple global models are updated by

$$\left\{\boldsymbol{w}_c^r = \hat{\boldsymbol{w}}_{c,r-1} + \boldsymbol{A}_r(\hat{\boldsymbol{w}}_{c,r-1}^* - \hat{\boldsymbol{w}}_{c,r-1})\right\}_{c=1}^C. \tag{45}$$

which is distributed to the current participated individual machines in S^r to collect the local gradients as

$$\left\{\boldsymbol{g}(\boldsymbol{w}_c^r) = \sum_{i\in S^r}\frac{n_i}{n^r}\nabla f_i(\boldsymbol{w}_c^r)\right\}_{c=1}^C. \tag{46}$$

Then, the acceleration procedure for multiple 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 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 global gradient. We initialize $\hat{\boldsymbol{m}}_c^0 = \mathbf{0}$ and the estimation of the first moment of the anchor gradients in C groups can be evaluated recursively using

$$\{\hat{\boldsymbol{m}}_c^r = \beta_1 \hat{\boldsymbol{m}}_c^r + (1-\beta_1)\boldsymbol{g}(\hat{\boldsymbol{w}}_c^r)\}_{c=1}^C. \tag{47}$$

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

$$\left\{\hat{\boldsymbol{v}}_c^r = \beta_2 \hat{\boldsymbol{v}}_c^{r-1} + (1-\beta_2)\boldsymbol{g}(\boldsymbol{w}_c^r)^2\right\}_{c=1}^C. \tag{48}$$

The central updating in the r-th round is designed as

$$\left\{\hat{\boldsymbol{w}}_c^r = \boldsymbol{w}_c^r - \alpha_g^0(1-\beta_1)\cdot\sqrt{\frac{1-\beta_2^r}{1-\beta_2}}\frac{\hat{\boldsymbol{m}}_c^r}{\sqrt{\hat{\boldsymbol{v}}_c^r+\epsilon}}\right\}_{c=1}^C, \tag{49}$$

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 gradient. As such, 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 machine 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. After the central acceleration of the multiple global models, $\{\hat{\boldsymbol{w}}_c^r\}_{c=1}^C$ are the initialized diversified model parameters for the next round.

D. Gradients-based Grouping Mechanism

We also propose the local gradients based grouping which is designed via the training dynamics. The central server initializes C groups, but the grouping is based on the similarity of the local gradients. The selected individual machines conduct K local iterations as shown in (23) and (24).

The objective in the center is formulated as increasing the similarity of the gradient information within the same group, which can be formulated as

$$\underset{\{S_c^r\}_{c=1}^C}{\text{minimize}} \quad \frac{1}{2}\sum_{c=1}^C \frac{1}{|S_c^r|}\sum_{i,j\in S_c^r}\|\nabla f_i(\hat{\boldsymbol{w}}_{i,K}^r) - \nabla f_j(\hat{\boldsymbol{w}}_{j,K}^r)\|^2. \tag{50}$$

After collecting the local gradient information by $\left\{\nabla f_i(\hat{\boldsymbol{w}}_{i,K}^r)\right\}_{i\in S^r}$, the central server conducts mean-centering local gradients from $\{\hat{\boldsymbol{g}}_i^r\}_{i\in S^r}$ as

$$\left\{\nabla \bar{f}_i(\hat{\boldsymbol{w}}_{i,K}^r) = \nabla f_i(\hat{\boldsymbol{w}}_{i,K}^r) - \frac{1}{|S^r|}\sum_{j\in S^r}\nabla f_j(\hat{\boldsymbol{w}}_{i,K}^r)\right\}_{i\in S^r}. \tag{51}$$

The objective (50) can be rewritten as

$$\underset{\{S_c^r\}_{c=1}^C}{\text{maximize}} \quad \sum_{c=1}^C \frac{1}{|S_c^r|}\sum_{i,j\in S_c^r}\nabla \bar{f}_i(\hat{\boldsymbol{w}}_{i,K}^r)^T\nabla \bar{f}_j(\hat{\boldsymbol{w}}_{j,K}^r), \tag{52}$$

since

$$\|\nabla f_i(\hat{\boldsymbol{w}}_{i,K}^r) - \nabla f_j(\hat{\boldsymbol{w}}_{j,K}^r)\|^2 = \|\nabla \bar{f}_i(\hat{\boldsymbol{w}}_{i,K}^r) - \nabla \bar{f}_j(\hat{\boldsymbol{w}}_{j,K}^r)\|^2. \tag{53}$$

Following the conclusion in (36), the solution of (52) is related to the $C-1$ largest eigenvalues of $\boldsymbol{G}_{r-1}^T\boldsymbol{G}_{r-1}$, where

$$\boldsymbol{G}_{r-1}^T\boldsymbol{G}_{r-1} = \left\{\nabla \bar{f}_i(\hat{\boldsymbol{w}}_{i,K}^r)^T\nabla \bar{f}_j(\hat{\boldsymbol{w}}_{j,K}^r)\right\}_{i,j\in S^r}, \tag{54}$$

and its eigen-decomposition is obtained by

$$\boldsymbol{G}_{r-1}^T\boldsymbol{G}_{r-1} = \tilde{\boldsymbol{V}}_{r-1}^T\tilde{\boldsymbol{\Lambda}}_{r-1}\tilde{\boldsymbol{V}}_{r-1}. \tag{55}$$

Based on the above analysis, the gradient based grouping is conducted via the lower-dimensional local gradients as $\tilde{\boldsymbol{\Lambda}}_{r-1}^{\frac{1}{2}}\tilde{\boldsymbol{V}}_{r-1}^T$, which can be transformed to the projection of all mean-centered local gradients in $\boldsymbol{G}_{r-1}$ on the subspace spanned by $\tilde{\boldsymbol{U}}_{r-1}$ as following:

$$\tilde{\boldsymbol{G}}_{r-1} = \tilde{\boldsymbol{U}}_{r-1}^T\boldsymbol{G}_{r-1} = \tilde{\boldsymbol{\Lambda}}_{r-1}^{\frac{1}{2}}\tilde{\boldsymbol{V}}_{r-1}^T = \left\{\nabla \tilde{f}_i(\hat{\boldsymbol{w}}_{i,K}^r)\right\}_{i\in S^r}, \tag{56}$$

where

$$\nabla \tilde{f}_i(\hat{\boldsymbol{w}}_{i,K}^r) = \tilde{\boldsymbol{U}}_{r-1}^T\nabla \bar{f}_i(\hat{\boldsymbol{w}}_{i,K}^r). \tag{57}$$

The initial grouping of the trained local models come from individual machines' preference of the previous global gradients

$$\left\{\hat{\boldsymbol{g}}_{c,r-1} = \sum_{i\in S^r}\frac{n_i}{n^r}\nabla f_i(\hat{\boldsymbol{w}}_c^{r-1})\right\}_{c=1}^C. \tag{58}$$

which are projected to the same lower-dimensional space as

$$\left\{\tilde{\boldsymbol{g}}_{c,r-1}^0 = \tilde{\boldsymbol{U}}_{r-1}^T\hat{\boldsymbol{g}}_{c,r-1}\right\}_{c=1}^C. \tag{59}$$

Following the initial grouping among the chosen machines, a crucial phase entails the computation of new anchor gradients. These new anchors serve as group representations of the global mode updating directions. This process of recalculating anchor gradients is geared towards refining the grouping mechanism, aiming to generate more potent representations of these gradients in $\tilde{\boldsymbol{G}}_{r-1}$.

In the grouping procedure, each individual machine measures its affinity towards the current anchor gradients by quantifying the Euclidean distance between its local gradient and the multiple global gradients and its preference can be represented by

$$\left\{c_i^l = \underset{c=1,\cdots,C}{\text{argmin}} \quad ||\nabla \tilde{f}_i(\hat{\boldsymbol{w}}_{i,K}^r) - \tilde{\boldsymbol{g}}_{c,r-1}^l||\right\}_{i\in S^r}, \tag{60}$$

where c_i^l denotes the group that machine i prefers in the l-th round of grouping updating. All individual machines with the same preference $\{c_i^l = c\}_{i\in S^r}$ of the global gradients form group $S_{c,l}^r$. Individual machines can be assembled into distinct groups based on their shared gradient preferences which refer to their visions of the model updating directions. To obtain the most powerful representation of different grouping, the diversified multiple global gradients should be updated as

$$\left\{\tilde{\boldsymbol{g}}_{c,r-1}^l = \frac{1}{|S_{c,l}^r|}\sum_{i\in S_{c,l}^r}\nabla \tilde{f}_i(\hat{\boldsymbol{w}}_{i,K}^r)\right\}_{c=1}^C, \tag{61}$$

in the l-th round of grouping updating in the center server.

The iterative sequence of updating progressively guides the anchor gradients towards their optimal states. This sequence persists until no further adjustments are observed, i.e.,

$$\sum_{c=1}^{C}\sum_{i\in S_{c,l}^{r}} ||\nabla \tilde{f}_i(\hat{\boldsymbol{w}}_{i,K}^{r}) - \tilde{\boldsymbol{g}}_{c,r-1}^{l}|| \leq \tilde{\epsilon}, \tag{62}$$

where $\tilde{\epsilon} = 10^{-4}$. We define $\{\tilde{\boldsymbol{g}}_{c,r-1}^{*} = \tilde{\boldsymbol{g}}_{c,r-1}^{l}\}_{c=1}^{C}$ and $\{S_c^r = S_{c,l}^r\}_{c=1}^{C}$. And the optimal reflections of the global updating preference are obtained by

$$\left\{\Delta \boldsymbol{w}_g^c = \boldsymbol{A}_r(\tilde{\boldsymbol{U}}_{r-1}\tilde{\boldsymbol{g}}_{c,r-1}^{*} + \frac{1}{|S^r|}\sum_{j\in S^r}\nabla f_j(\hat{\boldsymbol{w}}_{i,K}^{r}))\right\}_{c=1}^{C}, \tag{63}$$

which are the representations of C different visions to update the global models as

$$\left\{\boldsymbol{w}_c^r = \hat{\boldsymbol{w}}_{c,r-1} + \Delta \boldsymbol{w}_g^c\right\}_{c=1}^{C}. \tag{64}$$

Then, it follows the same adaptive central acceleration procedure as shown in (46)-(49) to obtain the diversified global models which are distributed to the participating machines for local updating in next round.

V. Experiments

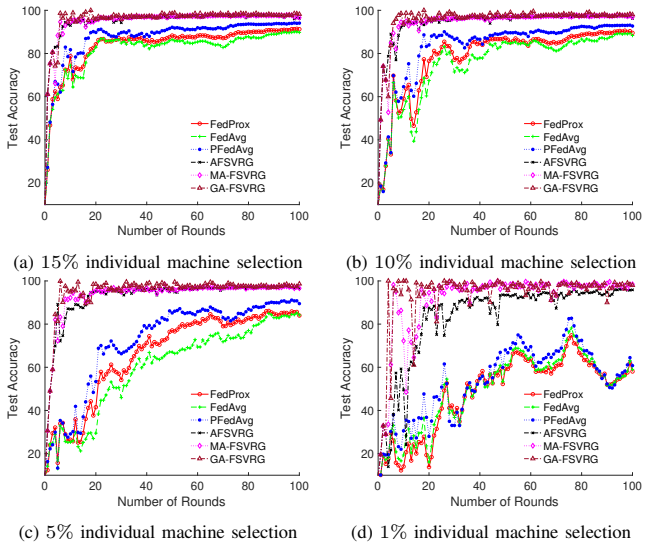
A. Local Training and Test Datasets Design

In our experiments, based on the MNIST data set containing 10 categories for classification, we design the highly heterogeneous local data sets to simulate the diversified local demands. To construct the heterogeneous local training data sets, each individual machine only possess samples belongs to 1 or 2 categories. To evaluate the performance of the diversified global models after each training round, we compare the test accuracy with 10000 test samples.

We also applied the histogram of gradient (HoG) method [28] to each sample. HoG utilized a block size of 7 and stride of 3, resulting in 64 blocks per sample. We evenly divided all possible gradient angles from 0 to 2π into 9 bins, and the magnitudes of gradients associated with each block were assigned to the corresponding bins based on their associated angles. This transformation reduced the original 784 features to 576 HoG features for each sample.

First, we compare the performance of both MA-FSVRG and GA-FSVRG with the baselines including FedAvg, FedProx, AFSVRG, and the personalized FL (PFedAvg) on the test accuracy with decreasing participated individual machines from 20% to 1% of 400 machines in total. For the adaptive central acceleration, we focus on the impact of the second-order moments of the anchor gradient, i.e., $\beta_1 = 0$ and $\beta_2 = 0.999$. For the proposed algorithms with diversified global models and anchor gradients, there are 4 different groups for individual machines to vote. And the grouping mechanism begins after 4 federated rounds.

As shown in Fig. 2, the proposed methods can outperform all baselines with the reduced machine participation. The benefit of diversified global models and anchor gradients focus on the fast convergence to higher test accuracy with limited



(a) 15% individual machine selection (b) 10% individual machine selection

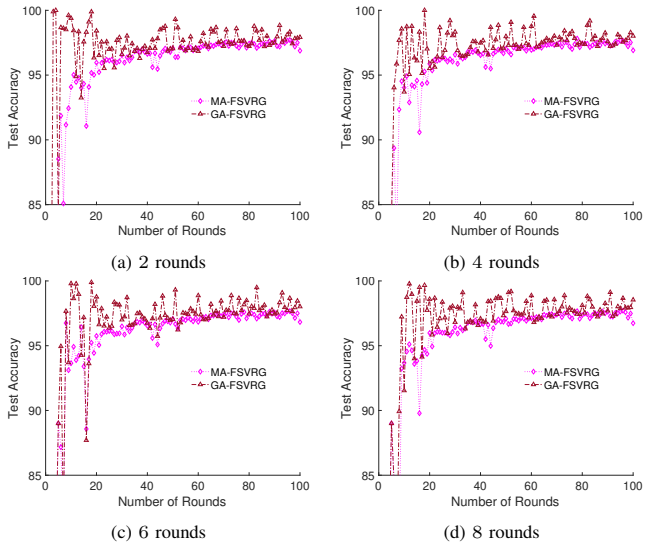
(c) 5% individual machine selection (d) 1% individual machine selection

Fig. 2. Performance comparison with test accuracy by decreasing participated individual machines where $\beta_1 = 0.0$, $\beta_2 = 0.999$ and maintain 4 diversified global models, threshold is 4 rounds.

machine participation in the federated training procedure. GA-FSVRG achieves 98.68% test accuracy at the 6th round with 15% participated machines where PFedAvg only achieves 82.94% test accuracy and both FedAvg and FedProx achieve 72.5%. The performance of MA-FSVRG is similar to GA-FSVRG with 20% participated machines. But, the advantage of using diversified anchor gradients for global model updating show up with limited participated machines. As shown in Fig. 2(d), the performance of GA-FSVRG can obviously outperform all other algorithms with only 1% machine participation. GA-FSVRG can still achieve 99% test accuracy at the 8th round where AFSVRG only achieves 40.2% test accuracy and jump to 59.25% test accuracy at the 9th round, and the performance of PFedAvg, FedProx, FedAvg are much worse by achieving 23.24%, 14.58%, and 17.71%, respectively. The performance of MA-FSVRG is more stable than that of GA-FSVRG with limited machine participation as shown in Fig. 2(d) GA-FSVRG even drop to 89.98% test accuracy at 90th round which is even worse than AFSVRG which achieves 94.85%. However the achieved test accuracy of MA-FSVRG will always above 96.56% after 70 rounds.

Second, since each machine is still exploring its model updating directions at the beginning of 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.

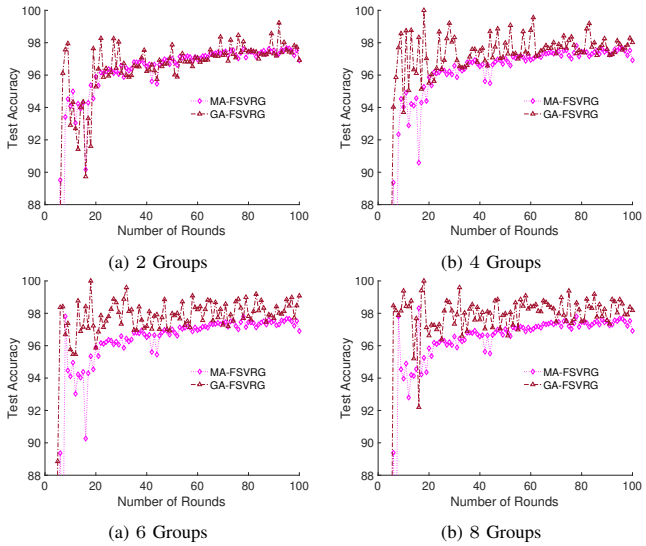
We focus on the performance comparison of the proposed MA-FSVRG and GA-FSVRG. As shown in Fig. 3 with 10% machines selected, both MA-FSVRG and GA-FSVRG are initialized with 2 to 8 federated rounds before the grouping mechanism to generate diversified global models and anchor gradients, respectively. Larger threshold to start the diversified global updating, smaller the overall variance on the achieved



(a) 2 rounds (b) 4 rounds

(c) 6 rounds (d) 8 rounds

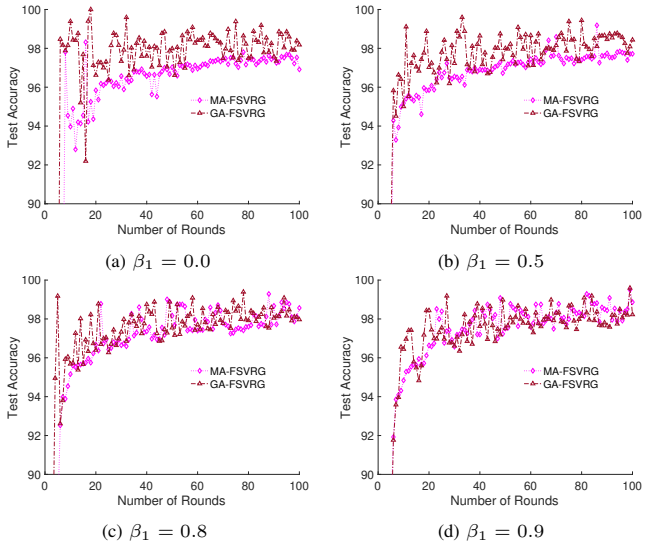
Fig. 3. Performance comparison with test accuracy by different thresholds of the initial rounds where $\beta_1 = 0.0$, $\beta_2 = 0.999$ and maintain 4 diversified global models.



(a) 2 Groups (b) 4 Groups

(a) 6 Groups (b) 8 Groups

Fig. 4. 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 4 rounds.



(a) $\beta_1 = 0.0$ (b) $\beta_1 = 0.5$

(c) $\beta_1 = 0.8$ (d) $\beta_1 = 0.9$

Fig. 5. Performance comparison with test accuracy by different β_1 where $\beta_2 = 0.999$ with 10% individual machine selection and threshold is 4 rounds with 4 groups.

test accuracy, especially with MA-FSVRG whose performance is greatly impacted by the quality of the local models. With different threshold 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 larger threshold as 8 rounds shown in Fig. 3(d), the performance of GA-FSVRG can be stabilized to higher lever of test accuracy by the increased threshold. The threshold can be adjusted by different applications which have different tolerance of the variance in performance. Larger threshold can guarantee a more stable initial training procedure especially with limited machine participation.

One of the important parameter 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. 4. With 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 can not 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 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.

Last, 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. 5, we check the first-order moment parameter β_1 from the set $\{0.0, 0.5, 0.8, 0.9\}$. With increasing β_1, 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 a same level after 20 rounds, as shown in Fig. 5(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.

But the advantage of the first-order moments becomes to disadvantage in diversified global model for the grouping mechanism based on anchor gradients except the stabler performance. The first-order moments is 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 eliminate the variance of anchor gradients which also reduces the diversity of the anchor gradients.

VI. Conclusions and Discussions

In this paper, we proposed the adaptive central accelerated FL with diversified global updating to tackle the challenges in heterogeneous demands of vary individual machines. It can not only show 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. In our future work, we will extend the horizontal diversified global models into diversified hierarchical global models which can bring different levels of services to diversified individual machines.

Appendix A

A. Properties of Objective Functions

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

$$\hat{f}_i(\boldsymbol{w}+\Delta\boldsymbol{w}) = f_i(\boldsymbol{w}) + \nabla f_i(\boldsymbol{w})^T\Delta\boldsymbol{w}. \tag{65}$$

Since

$$f_i(\boldsymbol{w}+\Delta\boldsymbol{w}) - f_i(\boldsymbol{w}) = \int_0^1 f_i(\boldsymbol{w}+\tau\Delta\boldsymbol{w})d\tau. \tag{66}$$

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

$$\boldsymbol{z} = \boldsymbol{w} + \tau\Delta\boldsymbol{w} \tag{67}$$

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

$$\int_0^1 df_i(\boldsymbol{z}) = \int_0^1 \nabla f_i(\boldsymbol{z})^T\Delta\boldsymbol{w}d\tau. \tag{68}$$

Combining (65), (66), and (68), we obtain

$$\begin{aligned} f_i(\boldsymbol{w}+\Delta\boldsymbol{w}) &= \hat{f}_i(\boldsymbol{w}+\Delta\boldsymbol{w}) \\ &+ \int_0^1 (\nabla f_i(\boldsymbol{z}) - \nabla f_i(\boldsymbol{w}))^T\Delta\boldsymbol{w}d\tau. \end{aligned} \tag{69}$$

Since

$$\begin{aligned} &\left|\int_0^1 (\nabla f_i(\boldsymbol{z}) - \nabla f_i(\boldsymbol{w}))^T\Delta\boldsymbol{w}d\tau\right| \\ &\leq \int_0^1 |(\nabla f_i(\boldsymbol{z}) - \nabla f_i(\boldsymbol{w}))^T\Delta\boldsymbol{w}|d\tau, \end{aligned} \tag{70}$$

and

$$\begin{aligned} &|(\nabla f_i(\boldsymbol{z}) - \nabla f_i(\boldsymbol{w}))^T\Delta\boldsymbol{w}| \\ &\leq ||(\nabla f_i(\boldsymbol{z}) - \nabla f_i(\boldsymbol{w}))||_2 \cdot ||\Delta\boldsymbol{w}||_2, \end{aligned} \tag{71}$$

and by the Lipschitz continuous gradient assumption, we have

$$||(\nabla f_i(\boldsymbol{z}) - \nabla f_i(\boldsymbol{w}))||_2 \leq L_i||\boldsymbol{z}-\boldsymbol{w}||_2 = L_i||\tau\Delta\boldsymbol{w}||_2. \tag{72}$$

Combining (69)-(72), we can upper bound $f_i(\boldsymbol{w}+\Delta\boldsymbol{w})$ by

$$f_i(\boldsymbol{w}+\Delta\boldsymbol{w}) \leq \hat{f}_i(\boldsymbol{w}+\Delta\boldsymbol{w}) + \frac{L_i}{2}||\Delta\boldsymbol{w}||^2. \tag{73}$$

Similarly, we can lower bound $f_i(\boldsymbol{w}+\Delta\boldsymbol{w})$ by

$$f_i(\boldsymbol{w}+\Delta\boldsymbol{w}) \geq \hat{f}_i(\boldsymbol{w}+\Delta\boldsymbol{w}) + \frac{\mu_i}{2}||\Delta\boldsymbol{w}||^2, \tag{74}$$

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

B. Convergence of Local Updating

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

$$\mathbb{E}_r[f(\boldsymbol{w}^r) - f(\boldsymbol{w}^*)] \leq \varphi^r(f(\boldsymbol{w}^0) - f(\boldsymbol{w}^*)), \tag{75}$$

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

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

$$f(\boldsymbol{w}) = \frac{1}{|E|}\sum_{i=1}^{|E|} f_i(\boldsymbol{w}), \tag{76}$$

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

$$\mathbb{E}_r[f_i(\boldsymbol{w})_{i\in E}] = f(\boldsymbol{w}), \tag{77}$$

which leads to

$$\mathbb{E}_r[\nabla f_i(\boldsymbol{w})_{i\in E}] = \nabla f(\boldsymbol{w}). \tag{78}$$

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

$$\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w})] = \nabla f_i(\boldsymbol{w}), \tag{79}$$

where the sample space is the local data sets. With the random selection over r rounds and combining (78) and (79), we can obtain the following expectation

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w})]] = \mathbb{E}_r[\nabla f_i(\boldsymbol{w})] = \nabla f(\boldsymbol{w}). \tag{80}$$

Notice that the $\boldsymbol{w}$ in (77)-(80) are general definition, both the local model $\hat{\boldsymbol{w}}^r_{i,k-1}$ and the global model $\boldsymbol{w}^{r-1}$ can be regarded as special cases of $\boldsymbol{w}$.

According to (80) and set $\boldsymbol{w} = \hat{\boldsymbol{w}}^r_{i,k-1}$, we get

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1})]] = \nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}), \tag{81}$$

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

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1})]] = \nabla f(\boldsymbol{w}^{r-1}), \tag{82}$$

which leads to

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})]] = \nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}). \tag{83}$$

Using (**??**) and (83), we have the expected distance from the current local model $\hat{\boldsymbol{w}}^r_{i,k}$ to the optimal model $\boldsymbol{w}^*$ w.r.t. round r as

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k} - \boldsymbol{w}^*||^2]] = \mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k-1} - \boldsymbol{w}^*||^2]] \\
&- 2\alpha_l \mathbb{E}_r[\mathbb{E}[(\hat{\boldsymbol{w}}^r_{i,k-1} - \boldsymbol{w}^*)]]^T \nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}) \\
&+ \alpha_l^2 \mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]].
\end{aligned} \tag{84}$$

First, we analyze the expected upper bound of the third term in the right-hand side of (84) w.r.t. the round number r. We apply the Cauchy-Schwarz inequality to write

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]] \\
&\le 2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]] \\
&+ 2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^*) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]].
\end{aligned} \tag{85}$$

To continue to provide upper bound of the right-hand side of (85), we have the following analysis. By writing

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}) = f(\boldsymbol{w}^*) - f(\boldsymbol{\beta}) + f(\boldsymbol{\beta}) - f(\boldsymbol{w}), \tag{86}$$

where $\boldsymbol{\beta}$ is an auxiliary variable to be specified shortly and using (73), we can write

$$\begin{aligned}
&f(\boldsymbol{w}^*) - f(\boldsymbol{w}) \le \nabla f(\boldsymbol{w}^*)^T(\boldsymbol{w}^* - \boldsymbol{w}) \\
&+ (\nabla f(\boldsymbol{w}^*) - \nabla f(\boldsymbol{w}))^T(\boldsymbol{w} - \boldsymbol{\beta}) + \frac{L}{2}||\boldsymbol{\beta} - \boldsymbol{w}||^2.
\end{aligned} \tag{87}$$

If we let

$$\boldsymbol{\beta} = \boldsymbol{w} - \frac{1}{L}(\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{w}^*)), \tag{88}$$

and substitute (88) into (87), we obtain

$$\begin{aligned}
&f(\boldsymbol{w}^*) - f(\boldsymbol{w}) \le \nabla f(\boldsymbol{w}^*)^T(\boldsymbol{w}^* - \boldsymbol{w}) \\
&- \frac{1}{2L}||(\nabla f(\boldsymbol{w}^*) - \nabla f(\boldsymbol{w}))||^2.
\end{aligned} \tag{89}$$

Using (80), we have

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

and combining with (13), we have

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1})||^2]] \\
&\le ||\nabla f(\hat{\boldsymbol{w}}^r_{i,k-1}) - \nabla f(\boldsymbol{w}^*)||^2.
\end{aligned} \tag{90}$$

Using (89) and (90), we obtain

$$2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]] \le 4L(f(\hat{\boldsymbol{w}}^r_{i,k-1}) - f(\boldsymbol{w}^*)). \tag{91}$$

Using (80), we have

$$\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)]] = \nabla f(\boldsymbol{w}^{r-1}), \tag{92}$$

which leads to

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^*) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]] \\
&= \mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*) - \mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)]]||^2]] \\
&= \mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]] \\
&- ||\mathbb{E}_r[\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)]]||^2 \\
&\le \mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^{r-1}) - \boldsymbol{g}_i(\boldsymbol{w}^*)||^2]].
\end{aligned} \tag{93}$$

Similar to the conclusion in (91), using (93), we have

$$\begin{aligned}
&2\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\boldsymbol{w}^*) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]] \\
&\le 4L(f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)).
\end{aligned} \tag{94}$$

Combining (85), (91) and (94), we have

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\boldsymbol{g}_i(\hat{\boldsymbol{w}}^r_{i,k-1}) - \boldsymbol{g}_i(\boldsymbol{w}^{r-1}) + \nabla f(\boldsymbol{w}^{r-1})||^2]] \\
&\le 4L(f(\hat{\boldsymbol{w}}^r_{i,k-1}) - f(\boldsymbol{w}^*) + f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)).
\end{aligned} \tag{95}$$

Second, we derive an upper bound for the second term in the right-hand side of (84). According to (7), we have

$$\nabla f(\hat{\boldsymbol{w}}^r_{i,k-1})^T(\boldsymbol{w}^* - \hat{\boldsymbol{w}}^r_{i,k-1}) \le f(\boldsymbol{w}^*) - f(\hat{\boldsymbol{w}}^r_{i,k-1}), \tag{96}$$

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

$$\begin{aligned}
&- 2\alpha_l \mathbb{E}_r[\mathbb{E}[\nabla f(\hat{\boldsymbol{w}}^r_{i,k-1})^T(\hat{\boldsymbol{w}}^r_{i,k-1} - \boldsymbol{w}^*)]] \\
&\le 2\alpha_l \mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^*) - f(\hat{\boldsymbol{w}}^r_{i,k-1})]].
\end{aligned} \tag{97}$$

Combining (95) and (97), (84) can be upper bounded by

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k} - \boldsymbol{w}^*||^2]] \\
&\le \mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k-1} - \boldsymbol{w}^*||^2]] - 2\alpha_l \mathbb{E}_r[\mathbb{E}[f(\hat{\boldsymbol{w}}^r_{i,k-1}) - f(\boldsymbol{w}^*)]] \\
&+ 4\alpha_l^2 L(f(\hat{\boldsymbol{w}}^r_{i,k-1}) - f(\boldsymbol{w}^*) + f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)),
\end{aligned} \tag{98}$$

which leads to the conclusion that

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k} - \boldsymbol{w}^*||^2]] \le \mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,k-1} - \boldsymbol{w}^*||^2]] \\
&- 2\alpha_l(1 - 2\alpha_l L)\mathbb{E}_r[\mathbb{E}[f(\hat{\boldsymbol{w}}^r_{i,k-1}) - f(\boldsymbol{w}^*)]] \\
&+ 4\alpha_l^2 L(f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)).
\end{aligned} \tag{99}$$

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

$$\begin{aligned}
&\mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,K} - \boldsymbol{w}^*||^2]] \\
&+ 2\alpha_l(1 - 2\alpha_l L)K\mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^r) - f(\boldsymbol{w}^*)]] \\
&\le \mathbb{E}_r[\mathbb{E}[||\hat{\boldsymbol{w}}^r_{i,0} - \boldsymbol{w}^*||^2]] + 4\alpha_l^2 LK\mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)]].
\end{aligned} \tag{100}$$

Since $\hat{\boldsymbol{w}}_{i,0}^r = \boldsymbol{w}^{r-1}$, and using (7), we can write

$$f(\boldsymbol{w}^{r-1}) \geq f(\boldsymbol{w}^*) + \nabla f(\boldsymbol{w}^*)^T(\boldsymbol{w}^{r-1} - \boldsymbol{w}^*) + \frac{\mu}{2}||\boldsymbol{w}^{r-1} - \boldsymbol{w}^*||^2, \tag{101}$$

which leads to

$$||\boldsymbol{w}^{r-1} - \boldsymbol{w}^*||^2 \leq \frac{2}{\mu}(f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)), \tag{102}$$

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

$$\mathbb{E}_r[\mathbb{E}[f(\boldsymbol{w}^r) - f(\boldsymbol{w}^*)]] = \mathbb{E}_r[f(\boldsymbol{w}^r) - f(\boldsymbol{w}^*)], \tag{103}$$

combining (100), (102) and (103), we obtain

$$\begin{aligned} &\mathbb{E}_r[f(\boldsymbol{w}^r) - f(\boldsymbol{w}^*)] \\ &\leq \left(\frac{1}{\mu\alpha_l(1-2\alpha_l L)K} + \frac{2\alpha_l L}{1-2\alpha_l L}\right) \\ &\mathbb{E}_r[f(\boldsymbol{w}^{r-1}) - f(\boldsymbol{w}^*)], \end{aligned} \tag{104}$$

which leads to

$$\mathbb{E}_r[f(\boldsymbol{w}^r) - f(\boldsymbol{w}^*)] \leq \varphi^r(f(\boldsymbol{w}^0) - f(\boldsymbol{w}^*)), \tag{105}$$

where

$$\varphi = \frac{1}{\mu\alpha_l(1-2\alpha_l L)K} + \frac{2\alpha_l L}{1-2\alpha_l L}. \tag{106}$$

(106) indicates that the proposed algorithm shall converge in the sense of (75) as long as the step length α_l and the number of local iterations K are selected to make the φ in (106) strictly less than one. There exists a variety of such selections. Provided below is a pair of α_l and K which, for a given target $\varphi < 1$, reduces the number of local iterations K to minimum.

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

$$\hat{\varphi} = p + q \tag{107}$$

with $p > 0$ and $q > 0$. We select the step-length α_l such that the second term of (106) satisfies

$$q = \frac{2\alpha_l L}{1-2\alpha_l L}, \tag{108}$$

i.e.,

$$\alpha_l = \frac{q}{2(1+q)L}. \tag{109}$$

With α_l given by (109), we select the smallest integer K such that the first term of (106) satisfies

$$\frac{1}{\mu\alpha_l(1-2\alpha_l L)K} \leq p, \tag{110}$$

i.e.,

$$K \geq \frac{1}{\mu\alpha_l(1-2\alpha_l L)p}. \tag{111}$$

From (109), we see that α_l is a concave function that monotonically increases with q. Concerning the selection of appropriate $\{\alpha_l, K\}$, we use $p = \hat{\varphi} - q$ and (109) to write the lower bound of K in (111) as a function of q:

$$b(q) = \frac{1}{\mu\alpha_l(1-2\alpha_l L)p} = \frac{2L(1+q)^2}{\mu q(\hat{\varphi}-q)}. \tag{112}$$

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

$$q^* = \frac{\hat{\varphi}}{2+\hat{\varphi}}, \tag{113}$$

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

$$\alpha_l^* = \frac{\hat{\varphi}}{4L(1+\hat{\varphi})}, \tag{114}$$

and

$$K^* = \left\lceil \frac{8L(1+\hat{\varphi})}{\mu\hat{\varphi}^2} \right\rceil, \tag{115}$$

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

REFERENCES

[1] S. Zhou and G. Y. Li, "Federated learning via inexact admm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[2] Y. Xue and V. Lau, "Riemannian low-rank model compression for federated learning with over-the-air aggregation," *IEEE Transactions on Signal Processing*, 2023.

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

[4] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2031–2044, 2020.

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

[7] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[8] Q. Wang, H. Yin, T. Chen, J. Yu, A. Zhou, and X. Zhang, "Fast-adapting and privacy-preserving federated recommender system," *The VLDB Journal*, pp. 1–20, 2021.

[9] D. Nawara and R. Kashef, "Iot-based recommendation systems–an overview," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2020, pp. 1–7.

[10] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1, pp. 83–112, 2017.

[11] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in neural information processing systems*, vol. 27, 2014.

[12] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, 2013.

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

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

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

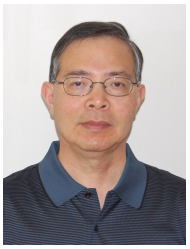
[16] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "Fedhealth: A federated transfer learning framework for wearable healthcare," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 83–93, 2020.

[17] I. Kevin, K. Wang, X. Zhou, W. Liang, Z. Yan, and J. She, "Federated transfer learning based cross-domain prediction for smart manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 4088–4096, 2021.

[18] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.
[19] H. Zhu, H. Zhang, and Y. Jin, "From federated learning to federated neural architecture search: a survey," *Complex & Intelligent Systems*, vol. 7, pp. 639–657, 2021.
[20] C. He, E. Mushtaq, J. Ding, and S. Avestimehr, "Fednas: Federated deep learning via neural architecture search," 2021.
[21] C. T Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 394–21 405, 2020.
[22] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
[23] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.
[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
[25] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.
[26] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.
[27] K. Fan, "On a theorem of weyl concerning eigenvalues of linear transformations i," *Proceedings of the National Academy of Sciences*, vol. 35, no. 11, pp. 652–655, 1949.
[28] W.-S. Lu, "Handwritten digits recognition using pca of histogram of oriented gradient," in *2017 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. IEEE, 2017, pp. 1–5.

Lei Zhao (S'17) received the B.S. and M.A.Sc. degrees in computer science and technology from Xidian University, Xi'an, China, in 2015 and 2018, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Victoria, Victoria, B.C., Canada, in 2023. He is currently an instructor at the Department of Electrical & Computer Engineering at the University of Victoria, Victoria, B.C., Canada.

Lin Cai (S'00-M'06-SM'10-F'20) has been with the Department of Electrical & Computer Engineering at the University of Victoria since 2005, and she is currently a Professor. She is an NSERC E.W.R. Steacie Memorial Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, an IEEE Fellow, and a Member of the Royal Society of Canada's College of New Scholars, Artists and Scientists. 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.



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