Sparse Federated Learning with Hierarchical Personalization Models

Xiaofeng Liu, Qing Wang, Yunfeng Shao, Yinchuan Li

Abstract—Federated learning (FL) can achieve privacy-safe and reliable collaborative training without collecting users' private data. Its excellent privacy security potential promotes a wide range of FL applications in Internet-of-Things (IoT), wireless networks, mobile devices, autonomous vehicles, and cloud medical treatment. However, the FL method suffers from poor model performance on non-i.i.d. data and excessive traffic volume. To this end, we propose a personalized FL algorithm using a hierarchical proximal mapping based on the moreau envelop, named sparse federated learning with hierarchical personalized models (sFedHP), which significantly improves the global model performance facing diverse data. A continuously differentiable approximated ℓ_1-norm is also used as the sparse constraint to reduce the communication cost. Convergence analysis shows that sFedHP's convergence rate is state-of-the-art with linear speedup and the sparse constraint only reduces the convergence rate to a small extent while significantly reducing the communication cost. Experimentally, we demonstrate the benefits of sFedHP compared with the FedAvg, HierFAVG (hierarchical FedAvg), and personalized FL methods based on local customization, including FedAMP, FedProx, Per-FedAvg, pFedMe, and pFedGP.

Index Terms—Federated learning (FL), machine learning, privacy preservation, non-i.i.d. data, cloud computing.

I. INTRODUCTION

MACHINE learning methods have proliferated in real-life applications thanks to the tremendous number of labeled samples [1]. Typically, these samples collected on users' devices, such as mobile phones, are expected to send to a centralized server with mighty computing power to train a deep model [2]. However, users are often reluctant to share personal data due to privacy concerns, which motivates the emergence of federated learning (FL) [3]. Federated Averaging (FedAvg) [3] is known as the first FL algorithm to build a global model for different clients while protecting their data locally. Moreover, FL has been used in the Internet of Things (IoT), wireless networks, mobile devices, autonomous vehicles, and cloud medical treatment for its excellent potential in privacy security [4]–[13].

Unfortunately, the distribution of local data stored in different clients varies greatly, and FedAvg performs unwise when meeting none independent and identically distributed (non-i.i.d.) data. In particular, generalization errors of the global model increase significantly with the data's statistical diversity

increasing [14], [15]. To address this problem, personalized FL based on multi-task learning [16] and personalization layers [17], and local customization [18]–[22] have been proposed. Most personalized federated learning algorithms prioritize the personalized model performance of individual clients, while overlooking the performance of the global model. However, this approach contradicts the fundamental purpose of federated learning, which aims to build a high-quality global model. Disregarding the global model's performance may lead to a suboptimal model and hinder the inclusion of new clients. Additionally, frequent communication between clients and the server is typically necessary in federated learning to ensure convergence performance, which can be hampered by high latency and limited bandwidth.

To address these challenges, we introduce a novel hierarchical personalized federated learning framework. Our approach includes a personalized edge server that minimizes differences between models during global model aggregation, leading to significant improvements in global model performance and allowing for personalized user models. Furthermore, our hierarchical personalized federated learning architecture covers the client-edge-cloud, reducing direct communication between clients and the cloud server and greatly decreasing communication overhead.

A. Main Contributions

Our main contributions in this paper are summarized as follows:

(1) We propose a noval personalized FL framework, named sparse federated learning with hierarchical personalization models (`sFedHP`). Our approach employs a hierarchical proximal mapping technique based on the moreau envelop. This method separates the optimization of client and edge models from that of the global model, promoting personalized models for clients and edge servers while keeping them close to the reference model. As a result, our approach enhances the performance of the global model on non-i.i.d data.

(2) The hierarchical architecture of `sFedHP` reduces direct communication between clients and the cloud server, leading to a significant decrease in communication overhead. The continuously differentiable approximated ℓ_1-norm constraints in `sFedHP` with sparse version generate sparse models, which further reduce communication costs.

(3) We present the convergence analysis of `sFedHP` by exploiting the convexity-preserving and smoothness-enabled properties of the loss function, which characterizes two notorious issues (client-sampling and client-drift errors) in FL [23].

Xiaofeng Liu and Qing Wang are with School of Electrical and Information Engineering, Tianjin University, Tianjin, China (e-mail: xiaofengliull@tju.edu.cn, wangq@tju.edu.cn). This work was completed during Xiaofeng Liu's internship at the Huawei Noah's Ark Lab for Advanced Study.

Yinchuan Li and Yunfeng Shao are with Huawei Noah's Ark Lab, Beijing, China (e-mail: liyinchuan@huawei.com, shaoyunfeng@huawei.com)

Corresponding author: Yinchuan Li.

With carefully tuned hyperparameters, theoretical analysis shows that sFedHP's convergence rate is state-of-the-art with linear speedup.

(4) We empirically evaluate the performance of `sFedHP` using different datasets that capture the statistical diversity of clients' data. We show that `sFedHP` obtained the state-of-the-art performance while greatly reducing the number of parameters by 80%. Moreover, `sFedHP` with non-sparse version outperforms FedAvg, the hierarchical FedAvg [24], and other local customization based personalized FL methods [18], [20]–[22] in terms of global model.

B. Organization and Main Notation

The remainder of the paper is organized as follow. Section II undertake a complete literature review to illustrate the existing research findings. The problem formulation and algorithm are formulated in Section III. Section IV shows the convergence analysis with some important Lemmas and Theorems. Section V presents the experimental results, followed by some analysis. Finally, conclusion and discussions are given in Section VI.

The main notations used are listed below.

Definition	Notation
Number of edge servers	N
Number of clients for each edge server	J
Global model of the cloud server	$\boldsymbol{w}$
Edge global model of i-th edge server	$\boldsymbol{w}_i$
Edge personalized model of i-th edge server	$\boldsymbol{\varphi}_i$
Local edge model of i-th edge server and j-th client	$\boldsymbol{\varphi}_{i,j}$
Local personalized model	$\boldsymbol{\theta}_{i,j}$
Expected loss over the data of the i-th server and j-th client	$\ell_{i,j}$
Training data of j-th client in i-th server	$Z_{i,j}$
Hyperparameter for sparsity	γ_1, γ_2
Hyperparameter for personalization	λ_1, λ_2
Hyperparameter for aggregation	β
Global training round	T
Local training round	R
Number of edges to aggregate	S
Learning rate for updating $\boldsymbol{w}_i$	η_1
Learning rate for updating $\boldsymbol{\theta}_{i,j}$	η_2
Expectation function	$\mathbb{E}(\cdot)$
Same-order infinitesimal function	$\mathcal{O}(\cdot)$

II. RELATED WORK

Traditional machine learning methods require local training data to be uploaded to a central server for centralized training. However, in practical scenarios like training on mobile phone data, sensor data in the Internet of Things, and data from external companies like banks, uploading local data poses a significant privacy risk. Thus, participants are often hesitant to expose their local data during training. To address these issues, McMahan et al. proposed a federated learning framework and its model aggregation algorithm, FedAvg [3]. They also developed a network protocol for federated learning in 2019 [25]. The framework involves multiple participants and a central server, where participants use their local data for training and upload model updates to a parameter server. The global model is obtained by aggregating these updates, enabling multi-party collaborative machine learning while protecting local data.

Federated learning has been successfully applied in various scenarios, including smart apps such as Google Keyboard, Siri Speech Classifier, and QuickType Keyboard, as well as cross-silo applications such as drug discovery, financial risk prediction, and smart manufacturing.

In federated learning, data is supposed to be independent and identically distributed, meaning each user's mini-batch sample must be statistically the same as a sample drawn uniformly from the entire training data set of all users. However, due to differences in devices, participants, enterprises, and scenarios, statistical diversity often exists among users, resulting in non-i.i.d. data. Non-i.i.d. data includes skewed feature distribution, skewed label distribution, different features of the same label, and different labels of the same feature [26]. Recent research on non-i.i.d. data in federated learning has primarily focused on skewed label distribution, where different users have different label distributions. To construct non-i.i.d. training samples, researchers partition existing "flat" datasets based on labels [27]. However, global models trained on non-i.i.d. datasets often struggle to generalize well [28]. To address this problem, various personalized federated learning methods have been proposed, including local customization based methods [18]–[22], personalized layer based methods [17], knowledge distillation based methods [29], [30], multi-task learning based methods [31], and lifelong learning based methods [32].

In particular, local customization based methods fine-tune the local model to customize a personalized model. Fedprox [18] achieves good personalization through ℓ_2-norm regularization. pFedMe [19] can decouple personalized model optimization from the global-model learning in a bi-level problem stylized based on Moreau envelopes. Per-FedAvg [21] sets up an initial meta-model that can be updated effectively after one more gradient descent step. FedAMP and HeurFedAMP [20] utilize the attentive message passing to facilitate more collaboration with similar customers. pFedGP [22] is proposed as a solution to PFL based on Gaussian processes with deep kernel learning. The proposed `sFedHP` is a local customization based personalized FL method, so we compare the performance of `sFedHP` with other local customization based personalized FL methods, including [18]–[22].

However, personalized federated learning algorithms often prioritize the performance of individual client models over the global model's performance, which contradicts the fundamental objective of federated learning to develop a high-quality global model and hinder the inclusion of new clients. Additionally, due to the need for frequent communication between clients and the cloud server to ensure convergence performance, FL suffers from high latency and limited bandwidth. As a result, we propose `sFedHP`, a communication-friendly personalized federated learning scheme that produces a high-performance global model.

III. SPARSE FEDERATED LEARNING WITH HIERARCHICAL PERSONALIZED MODELS (SFEDHP)

A. `sFedHP`: Problem Formulation

Consider a client-edge-cloud framework with one cloud server, N edge servers and NJ clients, i.e., each edge server
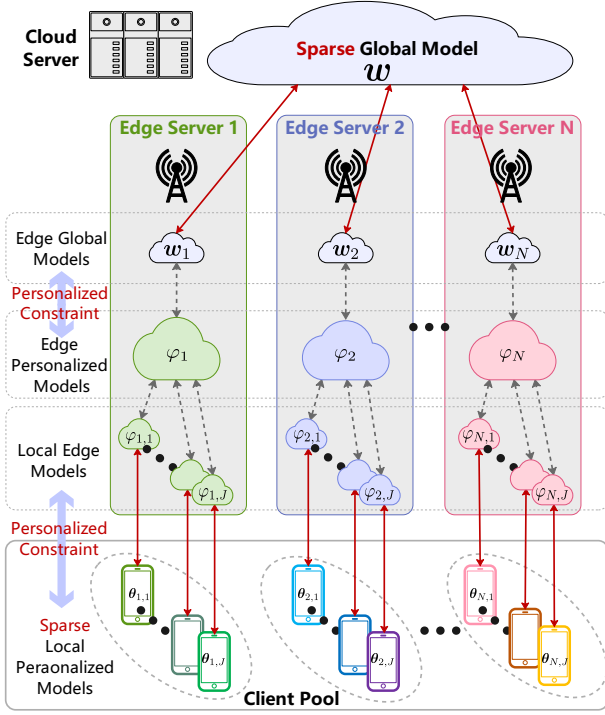
Fig. 1. Sparse client-edge-cloud FL with hierarchical personalized models.

collects data from J clients as showed in Fig. 1. In `sFedHP`, we aims to find a *sparse global-model w* based on local *sparse personalized models* $\boldsymbol{\theta}_{i,j}$, $i = 1,...,N$, $j = 1,...,J$, which can greatly reduce the model size and communication costs, by minimizing

$$\text{sFedHP:} \quad \min_{\boldsymbol{w}\in\mathbb{R}^d}\left\{F(\boldsymbol{w}) \triangleq \frac{1}{N}\sum_{i=1}^N F_i(\boldsymbol{w})\right\}, \quad (1)$$

with

$$F_i(\boldsymbol{w}) \triangleq \frac{1}{J}\sum_{j=1}^J F_{i,j}(\boldsymbol{w}), \quad (2)$$

$$F_{i,j}(\boldsymbol{w}) = \min_{\boldsymbol{\varphi}_{i,j},\boldsymbol{\theta}_{i,j}\in\mathbb{R}^d} \ell_{i,j}\left(\boldsymbol{\theta}_{i,j}\right) + \frac{\lambda_1}{2}\left\|\boldsymbol{\theta}_{i,j} - \boldsymbol{\varphi}_{i,j}\right\|_2^2$$
$$+\gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j})+\frac{\lambda_2}{2}\left\|\boldsymbol{\varphi}_{i,j} - \boldsymbol{w}\right\|_2^2 + \gamma_2\phi_\rho(\boldsymbol{w}), \quad (3)$$

where $\ell_{i,j}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ denotes the expected loss over the data distribution of the i-th edge server and the j-th client; λ_1 and λ_2 denoting regularization parameters that control the strength of $\boldsymbol{w}$ to the personalized model; $\boldsymbol{\varphi}_{i,j}$ is the local edge model of the i-th edge server and the j-th client; γ_1 and γ_2 denoting weight factors that control the sparsity level; and $\phi_\rho(\boldsymbol{x})$ being a twice continuously differentiable approximation for $\|\boldsymbol{x}\|_1$ [33], which is given by

$$\phi_\rho(\boldsymbol{x}) = \rho\sum_{n=1}^d \log\cosh\left(\frac{x_n}{\rho}\right), \quad (4)$$

where x_n denotes the n-th element in $\boldsymbol{x}$ and ρ is a weight parameter, which controls the smoothing level. Note that we use $\phi_\rho(\cdot)$ instead of $\|\cdot\|_1$ to exploit the sparsity in $\boldsymbol{w}$ and $\boldsymbol{\theta}$.

`sFedHP` utilizes the moreau envelope twice to partition the optimization problem into three stages, enabling independent optimization of the global model $\boldsymbol{w}$, client models $\boldsymbol{\theta}_{i,j}$, and edge server models $\boldsymbol{\varphi}_{i,j}$. The first application of the moreau envelope decouples the optimization of the global model from that of the edge server models, while the second separates the optimization of the client models from that of the edge server models. Leveraging the moreau envelope allows us to achieve personalized models for clients and edge servers while maintaining proximity to the reference model, resulting in enhanced performance on non-i.i.d data. To encourage sparsity in the global model $\boldsymbol{w}$, we permit clients to discover their own sparse models $\boldsymbol{\theta}_{i,j}$ using continuously differentiable approximated ℓ_1-norm constraints within a reasonable distance from the reference point $\boldsymbol{w}$.

In the first stage, the optimal local personalized model $\hat{\boldsymbol{\theta}}_{i,j}$ is obtained by minimizing (3) w.r.t. the data distribution of the i-th edge server and j-th client in the third stage as the following,

$$\hat{\boldsymbol{\theta}}_{i,j}(\boldsymbol{\varphi}_{i,j}) = \arg\min_{\boldsymbol{\theta}_{i,j}\in\mathbb{R}^d} L_{i,j}(\boldsymbol{\theta}_{i,j}) + \frac{\lambda_1}{2}\left\|\boldsymbol{\theta}_{i,j} - \boldsymbol{\varphi}_{i,j}\right\|_2^2, \quad (5)$$

where $L_{i,j}(\boldsymbol{\theta}_{i,j}) \triangleq \ell_{i,j}(\boldsymbol{\theta}_{i,j}) + \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j})$.

In the second stage, the edge personalized model $\boldsymbol{\varphi}_i$ and the local edge model $\boldsymbol{\varphi}_{i,j}$ are determined by minimizing (3) w.r.t. the client models of the i-th edge server,

$$\hat{\boldsymbol{\varphi}}_i = \frac{1}{J}\sum_{i=1}^J \hat{\boldsymbol{\varphi}}_{ij}, \quad (6)$$

$$\hat{\boldsymbol{\varphi}}_{ij} = \arg\min_{\boldsymbol{\varphi}_{ij}\in\mathbb{R}^d} \frac{\lambda_1}{2}\left\|\hat{\boldsymbol{\theta}}_{i,j} - \boldsymbol{\varphi}_{ij}\right\|_2^2 + \frac{\lambda_2}{2}\left\|\boldsymbol{\varphi}_{ij} - \boldsymbol{w}_i\right\|_2^2,$$

In the third stage, $\boldsymbol{w}$ is determined by utilizing the sparse model aggregation from multiple edges.

Assumption 1. *(Strong convexity and smoothness) Assume that $\ell(\boldsymbol{w}) : \mathbb{R}^d \to \mathbb{R}$ is (a) μ-strongly convex or (b) nonconvex and L-smooth on $\mathbb{R}^d$, then we respectively have the following inequalities*

$$(a) \ \ell(\boldsymbol{w}) \geq \ell(\bar{\boldsymbol{w}}) + \langle\nabla\ell(\bar{\boldsymbol{w}}), \boldsymbol{w} - \bar{\boldsymbol{w}}\rangle + \frac{\mu}{2}\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|_2^2,$$

$$(b) \ \|\nabla\ell(\boldsymbol{w}) - \nabla\ell(\bar{\boldsymbol{w}})\|_2 \leq L\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|_2.$$

Assumption 2. *(Bounded variance) The variance of stochastic gradients (sampling noise) in each client is bounded by*

$$\mathbb{E}_{\mathcal{Z}}\left[\left\|\nabla\tilde{\ell}_{i,j}\left(\boldsymbol{w}; Z_{i,j}\right) - \nabla\ell_{i,j}(\boldsymbol{w})\right\|_2^2\right] \leq \gamma_\ell^2,$$

where $Z_{i,j}$ is the training data randomly drawn from the distribution of client i and edge server j.

Assumption 3. *(Bounded diversity) The diversity of client's data distribution is bounded by*

$$\frac{1}{NJ}\sum_{i,j=1}^{N,J}\|\nabla\ell_{i,j}(\boldsymbol{w}) - \nabla\ell(\boldsymbol{w})\|_2^2 \leq \sigma_\ell^2.$$

While Assumption 1, 2 and 3 are widely used in FL gradient calculation and convergence analysis [19], [21], [23], [34].

B. `sFedHP`: Algorithm

The pseudocode for `sFedHP` is outlined in Algorithm 1. During the t-th communication round, the cloud server broadcasts the global model $\boldsymbol{w}^t$ to all edge servers. Subsequently, the edge servers and their clients conduct R rounds of iterative training and upload their edge global model $\boldsymbol{w}_i^t$ to aggregate a new global model $\boldsymbol{w}^{t+1}$. As previously mentioned, `sFedHP` decomposes the optimization process into three stages.

In the first level, the local personalized model $\hat{\boldsymbol{\theta}}_{i,j}$ of the i-th edge server and the j-th client is determined by solving (5), whose parameters are sparse and can reduce the communication load between clients and edge servers. Note that (5) can be easily solved by many first-order approaches, for example, Nesterov's accelerated gradient descent, based on the gradient. However, calculate the exact $\nabla \ell_{i,j}(\boldsymbol{\theta}_{i,j})$ requires the distribution of $Z_{i,j}$, we hence use the unbiased estimate by sampling a mini-batch of data $\mathcal{D}_{i,j}$,

$$\nabla \check{\ell}_{i,j}(\boldsymbol{\theta}_{i,j}, \mathcal{D}_{i,j}) = \frac{1}{|\mathcal{D}_{i,j}|} \sum_{Z_{i,j} \in \mathcal{D}_{i,j}} \nabla \ell_{i,j}(\boldsymbol{\theta}_{i,j}, Z_{i,j}). \quad (7)$$

such that $\mathbb{E}[\nabla \check{\ell}_{i,j}(\boldsymbol{\theta}_{i,j}, \mathcal{D}_{i,j})] = \nabla \ell_{i,j}(\boldsymbol{\theta}_{i,j})$. Thus, we solve the following minimization problem instead of solving (5) to obtain an approximated local personalized model

$$\check{\boldsymbol{\theta}}_{i,j}(\boldsymbol{\varphi}_{i,j}^{t,r}) = \arg \min_{\boldsymbol{\theta}_{i,j} \in \mathbb{R}^d} H(\boldsymbol{\theta}; \boldsymbol{\varphi}_{i,j}^{t,r}, \mathcal{D}_{i,j}), \quad (8)$$

where $H(\boldsymbol{\theta}_{i,j}; \boldsymbol{\varphi}_{i,j}^{t,r}, \mathcal{D}_{i,j}) = \check{\ell}_{i,j}(\boldsymbol{\theta}_{i,j}, \mathcal{D}_{i,j}) + \gamma_1 \phi_\rho(\boldsymbol{\theta}_{i,j}) + \frac{\lambda_1}{2} \|\boldsymbol{\theta}_{i,j} - \boldsymbol{\varphi}_{i,j}^{t,r}\|_2^2$. Similarly, (8) can be solved by Nesterov's accelerated gradient descent, we let the iteration goes until the condition

$$\|\nabla H(\check{\boldsymbol{\theta}}_{i,j}; \boldsymbol{\varphi}_{i,j}^{t,r}, \mathcal{D}_{i,j})\|_2^2 \leq \nu, \quad (9)$$

is reached, where ν is an accuracy level.

In the second level, after clients' local personalized models are updated, the local edge models are determined by

$$\check{\boldsymbol{\varphi}}_{i,j}^{t,r+1} = (\lambda_1 \check{\boldsymbol{\theta}}_{i,j}^{t,r} + \lambda_2 \boldsymbol{w}_i^{t,r})/(\lambda_1 + \lambda_2), \quad (10)$$

where $\check{\boldsymbol{\varphi}}_i = \frac{1}{J} \sum_{i=1}^J \check{\boldsymbol{\varphi}}_{i,j}$ is the current edge peraonalized model of the i-th edge server.

In the third level, once edge personalized models are updated, the edge global models can be updated by stochastic gradient descent as follows

$$\boldsymbol{w}_i^{t,r+1} = \boldsymbol{w}_i^{t,r} - \eta_1 \nabla F_i(\boldsymbol{w}_i^{t,r}), \quad (11)$$
$$= \boldsymbol{w}_i^{t,r} - \eta_1 [\lambda_2 (\boldsymbol{w}_i^{t,r} - \check{\boldsymbol{\varphi}}_i^{t,r+1}) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}_i^{t,r})],$$

where η_1 is a learning rate and $\boldsymbol{w}_i^{t,r}$ is the edge global model of i-th edge server at the global round t and edge round r. The global model $\boldsymbol{w}^{t+1}$ is determined by utilizing the edge global model $\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^{t,R}$ aggregation from multiple edges. Moreover, similarly with [19], [23], an additional parameter β is used for global-model update to improve the convergence performance.

Note that communication between the clients and edge servers is more efficient than between the clients and the cloud server. The latter's communication is high cost and latency because the distance is relatively long.

Algorithm 1 `sFedHP`: Sparse Federated Learning with Hierarchical Personalization Models Algorithm

---

Input: hyperparameters in (3) $\{\lambda_1, \lambda_2, \gamma_1, \gamma_2, \rho\}$
    communication ronds $\{T, R\}$
    init parameters $\boldsymbol{w}^0$
    client learning rate η_1, server learning rate η_2
Cloud server executes:
  for $t = 0, 1, ..., T-1$ **do**
    for $i = 1, 2, ..., N$ **in parallel do**
      $\boldsymbol{w}_i^{t+1} \leftarrow$ EdgeUpdate$(i, \boldsymbol{w}^t)$
      $\mathcal{S}^t \leftarrow$ (random set of S edge servers)
      $\boldsymbol{w}^{t+1} = (1-\beta)\boldsymbol{w}^t + \frac{\beta}{S} \sum_{i \in S^t} \boldsymbol{w}_i^{t+1}$
EdgeUpdate$(i, \boldsymbol{w}^t)$**:**
  $\boldsymbol{\varphi}_i^{t,0} = \boldsymbol{w}_i^{t,0} = \boldsymbol{w}^t$
  for $r = 0, 1, ..., R-1$ **do**
    for $j = 1, ..., J$ **in parallel do**
      Update the local personalized model $\boldsymbol{\theta}_{i,j}$:
      $\mathcal{D}_{i,j} \leftarrow$ (sample a mini-batch with size $|\mathcal{D}|$)
      $\check{\boldsymbol{\theta}}_{i,j}^{t,r}(\boldsymbol{\varphi}_i^{t,r}) = \arg \min_{\boldsymbol{\theta}_{i,j} \in \mathbb{R}^d} H(\boldsymbol{\theta}_{i,j}; \boldsymbol{\varphi}_i^{t,r}, \mathcal{D}_{i,j})$
      Update the local edge model $\boldsymbol{\varphi}_{i,j}$ according to (6):
      $\check{\boldsymbol{\varphi}}_{i,j}^{t,r+1} = (\lambda_1 \check{\boldsymbol{\theta}}_{i,j}^{t,r} + \lambda_2 \boldsymbol{w}_i^{t,r})/(\lambda_1 + \lambda_2)$
    Update the edge personalized model $\boldsymbol{\varphi}_i$:
    $\boldsymbol{\varphi}_i^{t,r+1} = \frac{1}{J} \sum_{j=1}^J \check{\boldsymbol{\varphi}}_{i,j}^{t,r+1}$
    $\boldsymbol{w}_i^{t,r+1} = \boldsymbol{w}_i^{t,r} - \eta_1 [\lambda_2 (\boldsymbol{w}_i^{t,r} - \boldsymbol{\varphi}_i^{t,r+1}) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}_i^{t,r})]$
  Return $\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^{t,R}$ to the cloud server

---

IV. CONVERGENCE ANALYSIS

A. Convergence Theorems

In this section, we present the convergence or `sFedHP`. We first prove Theorem 1 under Assumptions 1-3, which presents the smoothness and strong convexity proporties of $F_{i,j}(\boldsymbol{w})$.

Theorem 1. *If $\ell_{i,j}$ is convex or nonconvex with L-Lipschitz $\nabla \ell_{i,j}$, then $F_{i,j}$ is L_F-smooth with $L_F = \lambda_2 + \frac{\gamma_2}{\rho}$ (with the condition that $\lambda_2 > 4L + \frac{4\gamma_1}{\rho}$ for nonconvex L-smooth $\ell_{i,j}$) and if $\ell_{i,j}$ is μ-strongly convex, then $F_{i,j}$ is μ_F-strongly convex with $\mu_F = \frac{\lambda_1 \lambda_2 \mu}{\lambda_1 \mu + \lambda_2 \mu + \lambda_1 \lambda_2}$.*

Proof. We first prove some interesting character of $\phi_\rho(\boldsymbol{x})$ in (4), which is convex smooth approximation to $\|\boldsymbol{x}\|_1$ [33] and

$$0 \leq \nabla^2 \phi_\rho(x) = \frac{1}{\rho}\left(1 - (\tanh(x/\rho))^2\right) \leq \frac{1}{\rho},$$

which yields $\nabla \phi_\rho(x) - \nabla \phi_\rho(\bar{x}) \leq \frac{1}{\rho}(x - \bar{x})$, then we have

$$\|\nabla \phi_\rho(\boldsymbol{x}) - \nabla \phi_\rho(\bar{\boldsymbol{x}})\|_2$$
$$= \sqrt{\sum_{n=1}^d \left(\tanh(x_n/\rho) - \tanh(\bar{x}_n/\rho)\right)^2}$$
$$\leq \frac{1}{\rho}\sqrt{\sum_{n=1}^d (x_n - \bar{x}_n)^2} = \frac{1}{\rho}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2.$$

Hence, ϕ_ρ is convex function with $\frac{1}{\rho}$-Lipschitz $\nabla \phi_\rho$.

Let $L_{i,j}(\boldsymbol{\theta}_{i,j}) = \ell_{i,j}(\boldsymbol{\theta}_{i,j}) + \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j})$. On the one hand if $\ell_{i,j}$ is μ-strong convex. For $\alpha \in [0,1]$, we hence have

$$\ell_{i,j}\left(\alpha\boldsymbol{\theta}_{i,j} + (1-\alpha)\bar{\boldsymbol{\theta}}_{i,j}\right) \quad (12)$$
$$\leq \alpha\ell_{i,j}(\boldsymbol{\theta}_{i,j}) + (1-\alpha)\ell_{i,j}(\bar{\boldsymbol{\theta}}_{i,j}) - \frac{\mu}{2}\alpha(1-\alpha)\|\boldsymbol{\theta}_{i,j} - \bar{\boldsymbol{\theta}}_{i,j}\|_2.$$

In addition, by noting that ϕ_ρ is convex function, we have

$$\phi_\rho\left(\alpha\boldsymbol{\theta}_{i,j} + (1-\alpha)\bar{\boldsymbol{\theta}}_{i,j}\right) \leq \alpha\phi_\rho(\boldsymbol{\theta}_{i,j}) + (1-\alpha)\phi_\rho(\bar{\boldsymbol{\theta}}_{i,j}).$$
$$(13)$$

Combining (12) and (13) yields

$$L_{i,j}\left(\alpha\boldsymbol{\theta}_{i,j} + (1-\alpha)\bar{\boldsymbol{\theta}}_{i,j}\right)$$
$$\leq \alpha L_i(\boldsymbol{\theta}_{i,j}) + (1-\alpha)L_i(\bar{\boldsymbol{\theta}}_{i,j}) - \frac{\mu}{2}\alpha(1-\alpha)\|\boldsymbol{\theta}_{i,j} - \bar{\boldsymbol{\theta}}_{i,j}\|_2,$$

which shows that if $\ell_{i,j}$ is μ-strong convex, $L_{i,j}(\boldsymbol{\theta}_{i,j}) = \ell_{i,j}(\boldsymbol{\theta}_{i,j}) + \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j})$ is μ-strong convex.

On the other hand, if $\ell_{i,j}$ is nonconvex with L-Lipschitz $\nabla\ell_{i,j}$, we have

$$\|\nabla L_{i,j}(\boldsymbol{\theta}_{i,j}) - \nabla L_{i,j}(\bar{\boldsymbol{\theta}}_{i,j})\|_2$$
$$= \|\nabla\ell_{i,j}(\boldsymbol{\theta}_{i,j}) - \nabla\ell_{i,j}(\bar{\boldsymbol{\theta}}_{i,j}) + \gamma_1\nabla\phi_\rho(\boldsymbol{\theta}_{i,j}) - \gamma_1\nabla\phi_\rho(\bar{\boldsymbol{\theta}}_{i,j})\|_2$$
$$\leq \|\nabla\ell_{i,j}(\boldsymbol{\theta}_{i,j}) - \nabla\ell_{i,j}(\bar{\boldsymbol{\theta}}_{i,j})\|_2 + \gamma_1\|\nabla\phi_\rho(\boldsymbol{\theta}_{i,j}) - \nabla\phi_\rho(\bar{\boldsymbol{\theta}}_{i,j})\|_2$$
$$\leq (L + \frac{\gamma_1}{\rho})\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|_2,$$

which shows that if $\ell_{i,j}$ is L-smooth, then $L_{i,j}(\boldsymbol{\theta}_{i,j}) = \ell_{i,j}(\boldsymbol{\theta}_{i,j}) + \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j})$ is $L + \frac{\gamma_1}{\rho}$-smooth.

Let $g(\boldsymbol{y}_{i,j}) = M_{L_{i,j}}^{\boldsymbol{\theta}_{i,j}}(\boldsymbol{y}_{i,j}) = \min_{\boldsymbol{\theta}_{i,j}\in\mathbb{R}^d} L_{i,j}(\boldsymbol{\theta}_{i,j}) + \frac{\lambda_1}{2}\|\boldsymbol{y}_{i,j} - \boldsymbol{\theta}_{i,j}\|_2^2$, which is the famous Moreau envelope [34], thus if $L_{i,j}$ is convex or nonconvex but $(L + \frac{\gamma_1}{\rho})$-smooth, then $g(\boldsymbol{y}_{i,j})$ is L_g-smooth with $L_g = \lambda_1$ (with the condition that $\lambda_1 > 2(L + \frac{\gamma_1}{\rho})$ for nonconvex $L + \frac{\gamma_1}{\rho}$-smooth). Therefore if $L_{i,j}$ is μ-strong convex, $g(\boldsymbol{y}_{i,j})$ is μ_g-strong convex wiht $\mu_g = \frac{\lambda_1\mu}{\lambda_1+\mu}$.

Let $G_{i,j}(\boldsymbol{w}) = M_{g_{i,j}}^{\boldsymbol{y}_{i,j}}(\boldsymbol{y}_{i,j}) = \min_{\boldsymbol{y}_{i,j}\in\mathbb{R}^d} g_{i,j}(\boldsymbol{y}_{i,j}) + \frac{\lambda_2}{2}\|\boldsymbol{w} - \boldsymbol{y}_{i,j}\|_2^2$, similarly we have if $g_{i,j}$ is convex or nonconvex with L_g-smooth, then $G_{i,j}(\boldsymbol{w})$ is L_G-smooth with $L_G = \lambda_2$ (with the condition that $\lambda_2 > 2L_g$ for nonconvex L_g-smooth). Therefore if $g_{i,j}$ is μ_g-strong convex, yeilds $G_{i,j}(\boldsymbol{w})$ is μ_G-strong convex with

$$\mu_G = \frac{\lambda_2\mu_g}{\lambda_2 + \mu_g} = \frac{\lambda_1\lambda_2\mu}{\lambda_1\mu + \lambda_2\mu + \lambda_1\lambda_2}.$$

Furthermore, recall that

$$F_{i,j}(\boldsymbol{w}) = \min_{\boldsymbol{y}_{i,j}, \boldsymbol{\theta}_{i,j}\in\mathbb{R}^d} \ell_{i,j}(\boldsymbol{\theta}_{i,j}) + \frac{\lambda_1}{2}\|\boldsymbol{\theta}_{i,j} - \boldsymbol{y}_{i,j}\|_2^2$$
$$+ \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j}) + \frac{\lambda_2}{2}\|\boldsymbol{y}_{i,j} - \boldsymbol{w}\|_2^2 + \gamma_2\phi_\rho(\boldsymbol{w})$$
$$= \min_{\boldsymbol{y}_{i,j}\in\mathbb{R}^d} g_{i,j}(\boldsymbol{y}_{i,j}) + \frac{\lambda_2}{2}\|\boldsymbol{w} - \boldsymbol{y}_{i,j}\|_2^2 + \gamma_2\phi_\rho(\boldsymbol{w})$$
$$= G(\boldsymbol{w}) + \gamma_2\phi_\rho(\boldsymbol{w}).$$

Using the character of ϕ_ρ mentioned before, we have the similarly couclusion that if G_i is μ_G-strongly convex, $F_{i,j}$ is μ_F-strong convex with $\mu_F = \mu_G = \frac{\lambda_1\lambda_2\mu}{\lambda_1\mu+\lambda_2\mu+\lambda_1\lambda_2}$, and if G_i is L_G-smooth, $F_{i,j}$ is L_F-smooth with $L_F = L_G + $

$\frac{\gamma_2}{\rho} = \lambda_2 + \frac{\gamma_2}{\rho}$. Finally, combining the first-order condition $\lambda_1(\hat{\boldsymbol{y}}_{i,j} - \hat{\boldsymbol{\theta}}_{i,j}) + \lambda_2(\hat{\boldsymbol{y}}_{i,j} - \boldsymbol{w}) = 0$, we complete the proof. $\square$

For unique solution $\boldsymbol{w}^*$ to sFedHP, we have the following convergence theorems, Theorems 2 and 3, based on Assumptions 1-3 and Theorem 1.

Theorem 2. *(Strongly convex sFedHP's convergence). Let Assumptions 1(a) and 2 hold, there exists an $\tilde{\eta} \leq \min[\frac{1}{\mu_F}, \frac{1}{3L_F(3+128L_F/(\mu_F\beta))}]$, such that*

$$(a)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*)\right] \leq \frac{\Delta_0}{\tilde{\eta}T} + M_1 + \tilde{\eta}M_2,$$

$$(b)\frac{1}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left[\left\|\tilde{\theta}_{i,j}^T(\boldsymbol{w}^T) - \boldsymbol{w}^*\right\|_2^2\right]$$
$$\leq \mathcal{O}\left(\frac{L_F+\lambda^2}{\lambda^2\mu_F}\right)\mathbb{E}[F(\boldsymbol{w}_T) - F^*] + \mathcal{O}(D_1),$$

where $\Delta_0 \triangleq \|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2$, $M_1 = \frac{8\delta^2\bar{\lambda}^2}{\mu_F}$, $M_2 = 6(1 + 64\frac{L_F}{\mu_F\beta})\sigma_{F,1}^2 + \frac{128L_F\delta^2\bar{\lambda}^2}{\mu_F\beta}$, $\lambda \triangleq \frac{\lambda_1\lambda_2}{\sqrt{\lambda_1^2+\lambda_2^2}}$, $\bar{\lambda} \triangleq \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}$ and $D_1 = \sigma_{F,1}^2/\lambda^2 + \gamma_2^2 d_s^2/\bar{\lambda}^2 + \delta^2$ with $\sigma_{F,1}^2 \triangleq \frac{1}{NJ}\sum_{i,j=1}^{N,J}\|\nabla F_{i,j}(\boldsymbol{w}^*)\|_2^2$, δ being the sampling noise and d_s denoting the number of non-zero elements in the global-model $\boldsymbol{w}$.

Theorem 3. *(Nonconvex and smooth sFedHP's convergence). Let Assumptions 1(b), 2 and 3 hold, there exists an $\tilde{\eta} \leq \min[\frac{1}{588L_F\lambda^2)}, \frac{\beta}{2L_F}]$ with $\lambda \leq \sqrt{16(L + \frac{\gamma_1}{\rho})^2 + 1}$, such that*

$$(a)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^t)\right\|_2^2\right] \leq 4(\frac{\Delta_F}{\tilde{\eta}T} + G_1 + \tilde{\eta}^1 G_2)$$

$$(b)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{NJ}\sum_{i,j=1}^{N}\mathbb{E}\left[\left\|\tilde{\theta}_{i,j}(\boldsymbol{w}^t) - \boldsymbol{w}^t\right\|_2^2\right]$$
$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^t)\right\|_2^2\right] + \mathcal{O}(D_2),$$

where $\Delta_F \triangleq F(\boldsymbol{w}_0) - F^*$, $G_1 = 3\beta\bar{\lambda}^2\delta^2$, $G_2 = \frac{3L_F(49R\sigma_{F,2}^2 + 16\delta^2\bar{\lambda}^2)}{R}$ and $D_2 = \sigma_{F,2}^2/\lambda^2 + \gamma_2^2 d_s^2/\bar{\lambda}^2 + \delta^2$ with $\sigma_{F,2}^2 \triangleq \frac{\lambda^2}{\lambda^2-16(L+\frac{\gamma_1}{\rho})^2}\sigma_\ell^2$.

Theorems 2(a) and 3(a) show the convergence results of the global-model. Δ_0 and Δ_F denote the initial error which can be reduced linearly. Theorem 2(b) and 3(b) show the convergence of personalized client models in average to a ball of center $\boldsymbol{w}^*$ and radius $\mathcal{O}(D_1)$ and $\mathcal{O}(D_2)$, respectively, for μ-strongly convex and nonconvex. Note that, our sFedHP obtain sparsity with a convergence speed cost $\mathcal{O}(\gamma_2^2 d_s^2/\bar{\lambda}^2)$, the experimental results show that a good sparsity can be obtained by only using a tiny γ_2, yields $\gamma_2^2 d_s^2/\bar{\lambda}^2 \ll \{\sigma_{F,1}^2/\lambda^2, \sigma_{F,2}^2/\lambda^2\}$, i.e., the convergence speed cost of the sparse constraints can be omitted. Note that, according to our experiments, limiting the learning rate does not increase the training time. This technique is both useful and common in Federated Learning analysis. Similar to the works in [19], [23], we have applied

this technique to our theorem analysis. The proofs of Theorems 2 and 3 are presented in Appendix.

B. Some Important Lemmas

In this subsection, we present some important lemmas, which are well used in the proof of Theorems 2 and 3 ,to help better understand the conclusion of the these convergence theorem.

We re-write the local update in (11) as follow:

$$\boldsymbol{w}_i^t = \boldsymbol{w}_i^{t,r} - \eta_1 \underbrace{\lambda_2 \left[\left(\boldsymbol{w}_i^{t,r} - \tilde{\boldsymbol{y}}_i^{t,r+1} \right) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}_i^{t,r}) \right]}_{\triangleq g_i^{t,r}},$$

which implies

$$\eta_1 \sum_{r=0}^{R-1} g_i^{t,r} = \sum_{r=0}^{R-1} \left(\boldsymbol{w}_i^{t,r} - \boldsymbol{w}_{i,r+1}^t \right) = \boldsymbol{w}^t - \boldsymbol{w}_i^{t,R},$$

where $g_i^{t,r}$ can be interpreted as the biased estimate of $\nabla F_i(\boldsymbol{w}_i^{t,r})$ since $\mathbb{E}(g_i^{t,r}) \neq \nabla F_i(\boldsymbol{w}_i^{t,r})$. Then, we re-write the global update as

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \underbrace{\eta_1 \beta R}_{\triangleq \tilde{\eta}} \underbrace{\frac{1}{SR} \sum_{i \in \mathcal{S}^t} \sum_{r=0}^{R-1} g_i^{t,r}}_{\triangleq g^t},$$

where $\tilde{\eta}$ and g^t can be respectively considered as the step size and approximate stochastic gradient of the global update, which cause drift error in one-step update of the global model formulated in Lemma 1.

Lemma 1 (One-step global update). *Let Assumption 1(b) holds. We have*

$$\mathbb{E} \left[\left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^* \right\|_2^2 \right]$$
$$\leq \mathbb{E} \left[\left\| \boldsymbol{w}^t - \boldsymbol{w}^* \right\|_2^2 \right] - \tilde{\eta} \left(2 - 6L_F \tilde{\eta} \right) \mathbb{E} \left[F \left(\boldsymbol{w}^t \right) - F \left(\boldsymbol{w}^* \right) \right]$$
$$+ \frac{\tilde{\eta} \left(3\tilde{\eta} + 1/\mu_F \right)}{NR} \sum_{i,r}^{N,R} \mathbb{E} \left[\left\| g_i^{t,r} - \nabla F_i \left(\boldsymbol{w}^t \right) \right\|_2^2 \right]$$
$$+ 3\tilde{\eta}^2 \frac{1}{NJ} \sum_{i,j=1}^{NJ} \mathbb{E} \left[\left\| \nabla F_{i,j} \left(\boldsymbol{w}^t \right) - \nabla F \left(\boldsymbol{w}^t \right) \right\|_2^2 \right],$$

where $\sum_{i,r}^{N,R}$ *and* $\sum_{i,j}^{NJ}$ *are respectively used as alternatives for* $\sum_{i=1,r=0}^{N,R-1}$ *and* $\sum_{i,j=1}^{NJ}$.

The various parts of Lemma 1 are discussed in detail in the following Lemmas.

Lemma 2 (Bounded diversity of $\boldsymbol{\theta}_{i,j}$ w.r.t. mini-batch sampling). *For the solution* $\tilde{\boldsymbol{\theta}}_{i,j}(\boldsymbol{y}_{i,j}^{t,r})$ *to (8), we have*

$$\mathbb{E} \left[\left\| \tilde{\boldsymbol{\theta}}_{i,j} \left(\boldsymbol{y}_{i,j}^{t,r} \right) - \hat{\boldsymbol{\theta}}_{i,j} \left(\boldsymbol{y}_{i,j}^{t,r} \right) \right\|_2^2 \right] \leq \delta^2,$$

with

$$\delta^2 \triangleq \begin{cases} \frac{2}{(\lambda_1+\mu)^2} \left(\frac{\gamma_l^2}{|\mathcal{D}|} + \nu \right), & \text{if } A.1(a) \text{ holds;} \\ \frac{2}{(\lambda_1-L-\frac{\gamma_1}{\rho})^2} \left(\frac{\gamma_l^2}{|\mathcal{D}|} + \nu \right), & \text{if } A.1(b) \text{ holds.} \end{cases}$$

where $\lambda_1 > L + \gamma_1/\rho$ *is required for* $A.1(b)$ *case.*

Lemma 3 (Bounded edge server drift error). *If* $\tilde{\eta} \leq \frac{\beta}{2L_F}$, *we have*

$$\frac{1}{NR} \sum_{i,r=1}^{N,R} \mathbb{E} \left[\left\| g_i^{t,r} - \nabla F_i \left(\boldsymbol{w}^t \right) \right\|_2^2 \right] \leq 2\bar{\lambda}^2 \delta^2 +$$
$$\frac{16 L_F \tilde{\eta}}{\beta} \left(3 \frac{1}{NJ} \sum_{i,j=1}^{NJ} \mathbb{E} \left[\left\| \nabla F_{i,j} \left(\boldsymbol{w}^t \right) \right\|_2^2 \right] + \frac{2\bar{\lambda}^2 \delta^2}{R} \right),$$

where $g_i^{t,r} = \lambda_2 \left[\left(\boldsymbol{w}_i^{t,r} - \tilde{\boldsymbol{y}}_i^{t,r+1} \right) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}_i^{t,r}) \right]$ *and* $\bar{\lambda} = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}$.

Lemma 4 (Bounded diversity of F_i w.r.t. distributed training). *Let* $\boldsymbol{w}^*$ *denote the optimal solution for* (1), *we have*

$$\frac{1}{NJ} \sum_{i=1,j=1}^{N,J} \left\| \nabla F_{i,j}(\boldsymbol{w}) - \nabla F(\boldsymbol{w}) \right\|_2^2 \leq$$
$$\begin{cases} 4L_F \left(F(\boldsymbol{w}) - F \left(\boldsymbol{w}^* \right) \right) + 2\sigma_{F,1}^2, & \text{if } A.1(a) \text{ holds;} \\ \frac{16(L+\frac{\gamma_1}{\rho})^2}{\lambda^2-16(L+\frac{\gamma_1}{\rho})^2} \left\| \nabla F(\boldsymbol{w}) \right\|_2^2 + 2\sigma_{F,2}^2, & \text{if } A.1(b) \text{ holds.} \end{cases}$$

where $\lambda > 4(L + \gamma_1/\rho)$ *is required for* $A.1(b)$ *case,* $\sigma_{F,1}^2 \triangleq \frac{1}{NJ} \sum_{i,j=1}^{N,J} \left\| \nabla F_{i,j} \left(\boldsymbol{w}^* \right) \right\|_2^2$, $\lambda = \frac{\lambda_1 \lambda_2}{\sqrt{\lambda_1^2 + \lambda_2^2}}$ *and* $\sigma_{F,2}^2 \triangleq \frac{\lambda^2}{\lambda^2-16(L+\frac{\gamma_1}{\rho})^2} \sigma_\ell^2$.

Lemma 2 and Lemma 3 show the diversity and drift errors caused by mini-batch training strategy are bounded. While, Lemma 4 shows the effection w.r.t. distributed training is bounded. By combining Lemma 1-4, we can get the recurrence relation between $\boldsymbol{w}^t - \boldsymbol{w}^*$ and $F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*)$ in Theorems 2 and 3. The proofs of Lemma 1-4 are presented in Appendix.

V. EXPERIMENTAL RESULTS

A. Performance Comparison

To empirically highlight the performance of the proposed method, we first compare sFedHP in non-saprsity setting with FedAvg [3], hierarchical FedAvg (HierFAVG) [24] and local customization personalized FL methods, including Fedprox [18], Per-FedAvg [21], pFedMe [19], HeurFedAMP [20], and pFedGP [22] based on MNIST [35], Fashion-MNIST (FMNIST) [36] and CIFAR-10 [37] datasets. For non-i.i.d. setups, we follow the strategy in [19] for $N = 20$ clients and assign each client a unique local data with only 5 out of 10 labels. For client-edge-cloud framework, we set 4 edge servers and each edge server manages 5 clients. All 20 clients are selected to generate the global model in following experiments.

Furthermore, we use two different dataset split settings to validate the performance of the above algorithms. In Setting 1, we used 200, 200, 100 training samples and 800, 800, 400 test samples in each class for MNIST, FMNIST and CIFAR-10 datasets respectively. In Setting 2, 900, 900, 450 training samples and 300, 300, 150 test samples are applied on MNIST, FMNIST and CIFAR-10 datasets for each class. We use the DNN model with two hidden layers of size [500, 200] for MNIST/FMNIST datasets and the VGG model for CIFAR-10

dataset with "[16, 'M', 32, 'M', 64 'M', 128, 'M', 128, 'M']" cfg setting. Each experiment is run at least 3 times to obtain statistical reports.

Table I shows the global model performance of each algorithm. We maintain $|\mathcal{D}| = 20$, $T = 800$, $K = 5$, and $\eta_1 = \eta_2 = 0.05$ across all algorithms and fine-tune other fundamental hyperparameters. We found that hierarchical federated learning methods, such as sFedHP and HierFAVG, significantly outperformed other federated learning methods, including FedAvg, FedProx, and pFedMe, in the device-cloud structure on non-i.i.d. data. The proposed algorithm, sFedHP, outperforms the comparative algorithms concerning the global model in all settings by more than 1% (Setting 1 on MNIST), 2% (Setting 2 on MNIST), 2% (Setting 1 on FMNIST), 3% (Setting 2 on FMNIST), 19% (Setting 1 on CIFAR-10), and 3% (Setting 2 on CIFAR-10).

We present more detailed results for fine-tuning hyperparameters, including methods without a global model, in Tables II to V. Here, E represents the number of edge servers in the hierarchical framework, and α represents the proportion of the client model that does not interact with the global model in HeurFedAMP. For HeurFedAMP, we tune $\alpha \in [0.1, 0.5]$, where α represents the proportion of the client model that does not interact with the global model. For Fedprox, we tune $\lambda \in \{0.001, 0.01, 0.1, 1\}$ following the setting in [18]. For pFedMe and sFedHP, we use the same setting as in [19] with $\lambda \in \{5, 15, 20, 25, 30\}$. For pFedGP, we use the same setting as in [22] for other basic hyperparameters. We find that sFedHP outperforms all compared algorithms in the global model (GM) and achieves state-of-the-art results in the personalized model (PM).

B. More results

We compare the global-model between sFedHP, pFedMe [19] and hierarchical FedAvg [24] on μ-strongly convex and nonconvex situations. Similarly with [19], in each round of local training, the client uses $K = 5$

TABLE I
THE GLOBAL MODEL PERFORMANCE OF DIFFERENT ALGORITHMS ON MNIST, FMNIST, AND CIFAR-10. WE MAINTAIN $|\mathcal{D}| = 20$, $T = 800$, $K = 5$, AND $\eta_1 = \eta_2 = 0.05$ ACROSS ALL ALGORITHMS AND FINE-TUNE OTHER HYPERPARAMETERS. BEST RESULTS ARE BOLDED.

Dataset	Method	Setting 1 (%)	Setting 2 (%)
MNIST	FedAvg [3]	91.63±0.12	94.31±0.08
	HierFAVG [24]	93.22±0.03	95.05±0.07
	Fedprox [18]	93.43±0.15	90.04±0.08
	pFedMe [19]	89.34±0.04	93.04±0.14
	sFedHP (Ours)	**94.93**±0.03	**97.66**±0.06
FMNIST	FedAvg [3]	79.11±0.09	84.48±0.06
	HierFAVG [24]	83.55±0.15	85.48±0.08
	Fedprox [18]	78.67±0.04	84.33±0.10
	pFedMe [19]	80.57±0.07	85.14±0.11
	sFedHP (Ours)	**85.64**±0.04	**89.29**±0.05
CIFAR-10	FedAvg [3]	42.36±0.19	81.74±0.17
	HierFAVG [24]	52.76±0.21	87.28±0.13
	Fedprox [18]	39.10±0.23	72.98±0.35
	pFedMe [19]	58.66±0.15	90.31±0.22
	sFedHP (Ours)	**78.44**±0.08	**94.09**±0.10

TABLE II
RESULTS OF ALL OF ALGORITHMS ON ON MNIST WITH SETTING 1.

Method	η_1	η_2	λ	K	E	α	Acc.(%) PM	GM
Fedavg [3]	0.05	-	-	5	-	-	-	91.63
HierFAVG [24]	0.05	-	-	5	4	-	-	93.22
HeurFedAMP [20]	0.05	0.05	-	5	-	0.1	97.11	-
	0.05	**0.05**	**-**	**5**	-	**0.5**	**98.55**	-
Fedprox [18]	0.05	-	0.001	5	-	-	-	93.43
PerFedavg [21]	0.05	-	-	5	-	-	56.5	-
pFedMe [19]	0.05	0.05	5	5	-	-	98.8	88.83
	0.05	**0.05**	**15**	**5**	-	-	**98.92**	**89.34**
	0.05	0.05	20	5	-	-	98.91	89.06
	0.05	0.05	25	5	-	-	97.6	80.12
	0.05	0.05	30	5	-	-	52.07	10.41
pFedGP [22]	0.05	-	-	5	-	-	98.12	-
sFedHP (Ours)	0.05	0.05	5	5	4	-	98.24	93.06
	0.05	0.05	15	5	4	-	96.55	91.85
	0.05	**0.05**	**20**	**5**	**4**	-	**97.83**	**94.93**
	0.05	0.05	25	5	4	-	96.63	95.39
	0.05	0.05	30	5	4	-	96.23	95.63

TABLE III
RESULTS OF ALL OF ALGORITHMS ON MNIST WITH SETTING 2.

Method	η_1	η_2	λ	K	E	α	Acc.(%) PM	GM
Fedavg [3]	0.05	-	-	5	-	-	-	94.31
HierFAVG [24]	0.05	-	-	5	4	-	-	95.05
HeurFedAMP [20]	0.05	0.05	-	5	-	0.1	97.65	-
	0.05	**0.05**	**-**	**5**	-	**0.5**	**98.8**	-
Fedprox [18]	0.05	-	0.001	5	-	-	-	90.04
PerFedavg [21]	0.05	-	-	5	-	-	78.15	-
pFedMe [19]	0.05	0.05	5	5	-	-	99.17	91.14
	0.05	0.05	15	5	-	-	99.38	91.60
	0.05	**0.05**	**20**	**5**	-	-	**99.39**	**93.04**
	0.05	0.05	25	5	-	-	98.72	89.47
	0.05	0.05	30	5	-	-	52.08	10.53
pFedGP [22]	0.05	-	-	5	-	-	99.18	-
sFedHP (Ours)	0.05	0.05	5	5	4	-	99.01	94.69
	0.05	0.05	15	5	4	-	98.04	94.06
	0.05	**0.05**	**20**	**5**	**4**	-	**99.08**	**97.66**
	0.05	0.05	25	5	4	-	98.63	97.87
	0.05	0.05	30	5	4	-	98.32	97.94

TABLE IV
RESULTS OF ALL OF ALGORITHMS ON FMNIST WITH SETTING 1.

Method	η_1	η_2	λ	K	E	α	Acc.(%) PM	GM
Fedavg [3]	0.05	-	-	5	-	-	-	79.11
HierFAVG [24]	0.05	-	-	5	4	-	-	83.55
HeurFedAMP [20]	0.05	0.05	-	5	-	0.1	95.00	-
	0.05	**0.05**	**-**	**5**	-	**0.5**	**98.28**	-
Fedprox [18]	0.05	-	0.001	5	-	-	-	78.67
PerFedavg [21]	0.05	-	-	5	-	-	96.95	-
pFedMe [19]	0.05	0.05	5	5	-	-	98.84	79.76
	0.05	0.05	15	5	-	-	98.80	80.22
	0.05	0.05	20	5	-	-	98.81	79.66
	0.05	0.05	25	5	-	-	98.90	80.19
	0.05	**0.05**	**30**	**5**	-	-	**98.94**	**80.57**
pFedGP [22]	0.05	-	-	5	-	-	98.88	-
sFedHP (Ours)	0.05	0.05	5	5	4	-	98.55	84.01
	0.05	0.05	15	5	4	-	97.61	82.75
	0.05	**0.05**	**20**	**5**	**4**	-	**98.42**	**85.64**
	0.05	0.05	25	5	4	-	95.86	86.51
	0.05	0.05	30	5	4	-	91.94	86.86

TABLE V
RESULTS OF ALL OF ALGORITHMS ON FMNIST WITH SETTING 2.

Method	η_1	η_2	λ	K	E	α	Acc.(%) PM	GM
Fedavg [3]	0.05	-	-	5	-	-	-	84.48
HierFAVG [24]	0.05	-	-	5	4	-	-	85.48
HeurFedAMP [20]	0.05	0.05	-	5	-	0.1	97.43	-
	0.05	**0.05**	**-**	**5**	**-**	**0.5**	**98.97**	**-**
Fedprox [18]	0.05	-	0.001	5	-	-	-	84.33
PerFedavg [21]	0.05	-	-	5	-	-	96.47	-
pFedMe [19]	0.05	0.05	5	5	-	-	99.12	82.21
	0.05	0.05	15	5	-	-	99.14	82.18
	0.05	0.05	20	5	-	-	99.14	84.42
	0.05	**0.05**	**25**	**5**	**-**	**-**	**99.17**	**85.14**
	0.05	0.05	30	5	-	-	99.16	84.74
pFedGP [22]	0.05	-	-	5	-	-	99.20	-
sFedHP (Ours)	0.05	0.05	5	5	4	-	98.93	85.73
	0.05	0.05	15	5	4	-	98.5	83.7
	0.05	**0.05**	**20**	**5**	**4**	**-**	**98.98**	**89.29**
	0.05	0.05	25	5	4	-	96.73	89.47
	0.05	0.05	30	5	4	-	93.35	89.05

TABLE VI
RESULTS OF ALL OF ALGORITHMS ON CIFAR-10 WITH SETTING 1.

Method	η_1	η_2	λ	K	E	α	Acc.(%) PM	GM
Fedavg [3]	0.05	-	-	5	-	-	-	42.36
HierFAVG [24]	0.05	-	-	5	4	-	-	52.76
HeurFedAMP [20]	0.05	0.05	-	5	-	0.1	71.36	-
	0.05	**0.05**	**-**	**5**	**-**	**0.5**	**82.26**	**-**
Fedprox [18]	0.05	-	0.001	5	-	-	-	39.10
PerFedavg [21]	0.05	-	-	5	-	-	77.20	-
pFedMe [19]	0.05	0.05	5	5	-	-	85.46	34.75
	0.05	0.05	15	5	-	-	85.45	43.26
	0.05	0.05	20	5	-	-	87.33	48.25
	0.05	0.05	25	5	-	-	88.36	52.39
	0.05	**0.05**	**30**	**5**	**-**	**-**	**88.45**	**58.66**
pFedGP [22]	0.05	-	-	5	-	-	85.14	-
sFedHP (Ours)	0.05	0.05	5	5	4	-	78.00	47.91
	0.05	0.05	15	5	4	-	75.53	68.48
	0.05	**0.05**	**20**	**5**	**4**	**-**	**83.95**	**78.44**
	0.05	0.05	25	5	4	-	76.98	65.83
	0.05	0.05	30	5	4	-	76.24	66.39

TABLE VII
RESULTS OF ALL OF ALGORITHMS ON CIFAR-10 WITH SETTING 2.

| Method	η_1	η_2	λ	K	E	α	Acc.(%) PM	GM
Fedavg [3]	0.05	-	-	5	-	-	-	81.74
HierFAVG [24]	0.05	-	-	5	4	-	-	87.28
HeurFedAMP [20]	0.05	0.05	-	5	-	0.1	83.51	-
	0.05	**0.05**	**-**	**5**	**-**	**0.5**	**88.13**	**-**
Fedprox [18]	0.05	-	0.001	5	-	-	-	72.98
PerFedavg [21]	0.05	-	-	5	-	-	84.92	-
pFedMe [19]	0.05	0.05	5	5	-	-	89.22	51.70
	0.05	0.05	15	5	-	-	90.08	62.06
	0.05	0.05	20	5	-	-	92.08	71.77
	0.05	0.05	25	5	-	-	93.17	86.95
	0.05	**0.05**	**30**	**5**	**-**	**-**	**93.40**	**90.31**
pFedGP [22]	0.05	-	-	5	-	-	85.98	-
sFedHP (Ours)	0.05	0.05	5	5	4	-	88.44	83.17
	0.05	0.05	15	5	4	-	88.86	87.62
	0.05	0.05	20	5	4	-	92.39	92.42
	0.05	0.05	25	5	4	-	93.16	93.27
	0.05	**0.05**	**30**	**5**	**4**	**-**	**93.98**	**94.09**
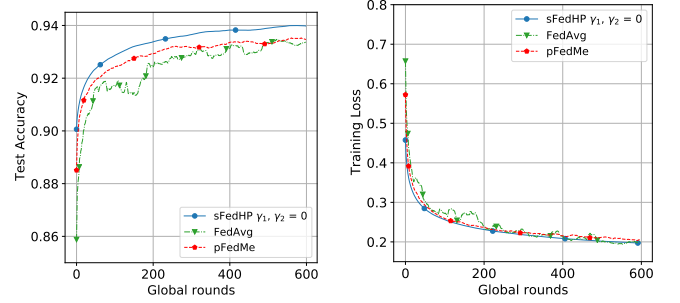
Fig. 2. μ-Strongly Convex Performance comparison of sFedHP, FedAvg and pFedMe in MNIST ($\eta = 0.05, R = 20, |\mathcal{D}| = 20, \beta = 1$ for all experiments, $\lambda_1 = \lambda_2 = \lambda = 25$ for sFedHP and pFedMe)

gradient-based iterations to obtain an approximated optimal local model, i.e., solve (8) in sFedHP. An ℓ_2-regularized multinomial logistic regression model (MLR) is used for μ-strong convex situation, while a deep neural netwrok (DNN) with two hidden layers of size [500, 200] is used for nonconvex situation. In our experiments, $\frac{3}{4}$ datasets are for training and the others are for testing. All experiments were conducted on a NVDIA Quadro RTX 6000 environment, and the code based on PyTorch is available online.

We use both sparsity and non-sparsity settings for sFedHP. In the sparsity setting, we set $\gamma_1 = \gamma_2 = 0.001$ at the beginning, while set γ_1 and γ_2 to a tiny number respectively when the client model sparsity is lower than 0.2 and after training 100 global rounds. Since according to our theoretical results, a tiny γ_1, γ_2 can reduce the convergence cost. In non-sparsity setting, we set $\gamma_1 = \gamma_2 = 0$.

Fig. 2 shows the performance of sFedHP, hierarchical FedAvg and pFedMe in μ-strongly convex situation. Since the MLR model is already lightweight enough, we set $\gamma_1 = \gamma_2 = 0$ for sFedHP. In Fig. 2, we can see that both sFedHP and pFedMe perform better than hierarchical FedAvg, while sFedHP obtain higher test accuracy and convergence speed, which shows that the hierarchical personalization scheme is more suitable for solving statistical diversity problem.

Fig. 3 shows the performance of sFedHP, hierarchical FedAvg and pFedMe in nonconvex situation. We test sparsity and non-sparsity settings for sFedHP in accuracy, model sparsity and the accumulative communication. We set 20% sparsity, the proportion of non-zero parameters, as a lower bound to preserve performance. According to Fig. 3, the proposed sFedHP can reduce the communication cost while achieve good test accuracy simultaneously. Specific, sFedHP obtain similar performance in a global model with pFedMe while reducing communication costs by 80% and obtain higher performance when using a non-sparsity setting.

Remark 1. *The accumulated communication cost in Fig. 3 is calculated based on one client in the client-cloud FL or one edge server in the client-edge-cloud FL since the client and the edge server are nearby between which low-cost communication is possible. Consider the DNN model with 79510 parameters, whose non-zero parameters are quantized in 64 bits while zero parameters are quantized in 1 bit. 79510*
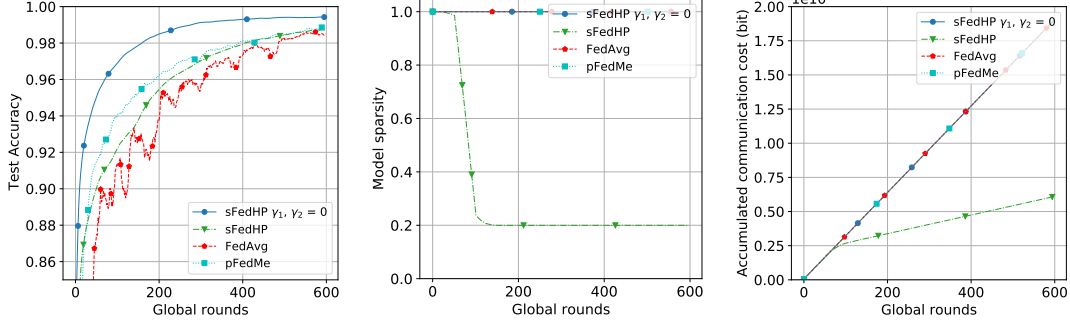
Fig. 3. Nonconvex Performance comparison of `sFedHP`, FedAvg and pFedMe in MNIST ($\eta = 0.05, R = 20, |\mathcal{D}| = 20, \beta = 1$ for all experiments, $\lambda_1 = \lambda_2 = \lambda = 25$ for `sFedHP` and pFedMe)
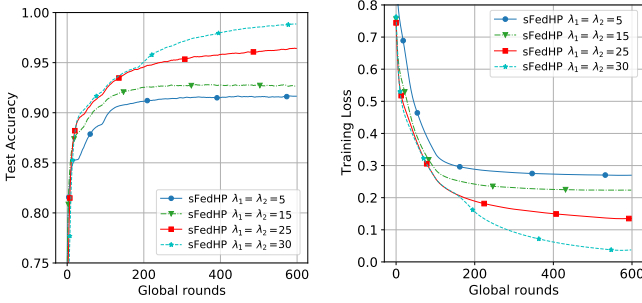

Fig. 4. Effect of λ on the convergence of `sFedHP` on MNIST ($\rho = 6 \times 10^{-5}, \gamma_1 = \gamma_2 = 0.008, R = 20, K = 5, \beta = 1$).
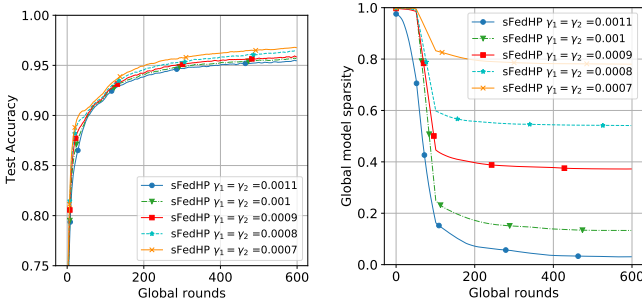

Fig. 5. Effect of γ on the convergence of `sFedHP` on MNIST ($\rho = 6 \times 10^{-5}, R = 20, \lambda_1 = \lambda_2 = 25, K = 5, \beta = 1$).

bits location parameters are needed to mark the positions of zero parameters in sFedHP. This section aims to demonstrate the excellent sparsity of sFedHP qualitatively. We focused on analyzing the ability of sFedHP to process non-i.i.d. data under different settings. More rigorous communication cost analysis needs to cooperate with the encoding and decoding process.

In summary, `sFedHP` performs better that pFedMe and hierarchical FedAvg in test accuracy, convergence rate, and communication cost in μ-strongly convex and nonconvex situations. Furthermore, the hierarchical personalization scheme is more suitable for solving statistical diversity problems.

C. Effect of hyperparameters

We empirically study the effect of different hyperparameters in `sFedHP`.

Effects of λ: According to Fig. 4, properly increasing λ can effectively improve the test accuracy and convergence rate for `sFedHP`. And we find that an oversize λ_1 and λ_2 may cause gradient explosion.

Effects of γ: Fig. 5 shows the relationship of γ with the model sparsity and the convergence speed, we can see that increasing γ will reduce the communication speed and global accuracy.

VI. CONCLUSION

In this paper, we propose `sFedHP` as a sparse hierarchical personalized FL algorithm that can greatly remiss the statistical diversity issue to improve the FL performance and reduce the communication cost in FL. Our approach uses an approximated ℓ_1-norm and the hierarchical proximal mapping to generate the loss function. The hierarchical proximal mapping enables the personalized edge-model, and client-model optimization can be decomposed from the global-model learning, which allows `sFedHP` parallelly optimizes the personalized edge-model and client-model by solving a tri-level problem. Theoretical results present that the sparse constraint in `sFedHP` only reduces the convergence speed to a small extent. Experimental results further demonstrate that `sFedHP` outperforms the client-edge-cloud hierarchical FedAvg and many other personalized FL methods based on local customization under different settings.

APPENDIX
PROOF OF THE RESULTS

Proof of Lemma 1

Proof. First, we have

$$\mathbb{E}\left[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2\right] = \mathbb{E}\left[\|\boldsymbol{w}^t - \tilde{\eta}g^t - \boldsymbol{w}^*\|_2^2\right] \quad (14)$$

$$= \mathbb{E}\left[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|_2^2\right] - 2\tilde{\eta}\mathbb{E}\left[\langle g^t, \boldsymbol{w}^t - \boldsymbol{w}^*\rangle\right] + \tilde{\eta}^2\mathbb{E}\left[\|g^t\|_2^2\right]$$

and the second term of (14) is as follow

$$
\begin{aligned}
&-\mathbb{E}\left\langle g^t, \boldsymbol{w}^T - \boldsymbol{w}^*\right\rangle \\
&= -\frac{1}{NR}\sum_{i,r}^{N,R}(\langle g_i^{t,r} - \nabla F_i\left(\boldsymbol{w}^t\right), \boldsymbol{w}^t - \boldsymbol{w}^*\rangle \\
&\qquad\qquad + \langle \nabla F_i\left(w^t\right), \boldsymbol{w}^t - \boldsymbol{w}^*\rangle) \\
&\overset{(a)}{\leq} \frac{1}{2NR}\sum_{i,r}^{N,R}\left(\frac{\left\|g_i^{t,r} - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2}{\mu_F} + \mu_F\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2\right) \\
&\qquad\qquad - \frac{1}{NJ}\sum_{i,j=1}^{NJ}\left(\langle\nabla F_{i,j}\left(\boldsymbol{w}^t\right), \boldsymbol{w}^t - \boldsymbol{w}^*\rangle\right) \\
&\overset{(b)}{\leq} \frac{1}{2NR}\sum_{i,r}^{N,R}\left(\frac{\left\|g_i^{t,r} - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2}{\mu_F} + \mu_F\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2\right) \\
&\qquad\qquad + \left(F\left(\boldsymbol{w}^*\right) - F\left(\boldsymbol{w}^t\right) - \frac{\mu_F}{2}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2\right) \\
&= \frac{1}{2NR}\sum_{i,r}^{N,R}\left(\frac{\left\|g_i^{t,r} - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2}{\mu_F}\right) + F\left(\boldsymbol{w}^*\right) - F\left(\boldsymbol{w}^t\right).
\end{aligned}
\tag{15}
$$

where (a) follows by the Peter Paul inequality and the Jensens inequality and (b) is due to Theorem 1.

From equations (18) and (19) in [19], we have

$$
\begin{aligned}
\left\|g^t\right\|^2 &\leq \frac{3}{NR}\sum_{i,r}^{N,R}\left\|g_i^{t,r} - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2 \\
&\quad + 3\left\|\frac{1}{S}\sum_{i\in\mathcal{S}^t}\nabla F_i\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2 \\
&\quad + 6L_F\left(F\left(\boldsymbol{w}^t\right) - F\left(\boldsymbol{w}^*\right)\right).
\end{aligned}
\tag{16}
$$

Taking expectation of the second term of (16) w.r.t edge server sampling, we have

$$
\begin{aligned}
&\mathbb{E}_{S_t}\left\|\frac{1}{S}\sum_{i\in\mathcal{S}^t}\nabla F_i\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2 \\
&\overset{(a)}{\leq} \frac{1}{SJ}\sum_{j=1}^{J}\mathbb{E}_{S_t}\left(\sum_{i\in\mathcal{S}^t}\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right) \\
&\overset{(b)}{=} \frac{1}{SJ}\sum_{i,j}^{NJ}\left(\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right)\mathbb{E}_{S_t}[\mathbb{I}_{i\in S_t}] \\
&\leq \frac{1}{NJ}\sum_{i,j}^{NJ}\left(\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right)
\end{aligned}
\tag{17}
$$

where (a) is due to the Jensen's inequality; $\mathbb{I}_X$ in (b) is an indicator function of an event X; and $\mathbb{E}_{S_t}[\mathbb{I}_{i\in S_t}] = \frac{S}{N}$.

By substituting (15), (16), (17) into (14), we finish the proof of Lemma 1. $\qquad\square$

Proof of Lemma 2

Proof. By noting that the last term of (8) only w.r.t. $\boldsymbol{y}_{i,j}^{t,r}$ we have

$$
\tilde{\boldsymbol{\theta}}_{i,j}(\boldsymbol{y}_{i,j}^{t,r}) = \arg\min_{\boldsymbol{\theta}_{i,j}\in\mathbb{R}^d}\tilde{\ell}_{i,j}\left(\boldsymbol{\theta}_{i,j}\right) + \frac{\lambda_1}{2}\left\|\boldsymbol{\theta}_{i,j} - \boldsymbol{y}_{i,j}^{t,r}\right\|_2^2 + \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j}).
$$

$$
\hat{\boldsymbol{\theta}}_{i,j}(\boldsymbol{y}_{i,j}^{t,r}) = \arg\min_{\boldsymbol{\theta}_{i,j}\in\mathbb{R}^d}\hat{\ell}_{i,j}\left(\boldsymbol{\theta}_{i,j}\right) + \frac{\lambda_1}{2}\left\|\boldsymbol{\theta}_{i,j} - \boldsymbol{y}_{i,j}^{t,r}\right\|_2^2 + \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j}).
$$

Denote $h_{i,j}\left(\boldsymbol{\theta}_{i,j}; \boldsymbol{y}_{i,j}^{t,r}\right) = \ell_{i,j}\left(\boldsymbol{\theta}_{i,j}\right) + \gamma_1\phi_\rho(\boldsymbol{\theta}_{i,j}) + \frac{\lambda_1}{2}\left\|\boldsymbol{\theta}_{i,j} - \boldsymbol{y}_{i,j}^{t,r}\right\|_2^2$. Then, $h_i\left(\boldsymbol{\theta}_{i,j}; \boldsymbol{y}_{i,j}^{t,r}\right)$ is $(\lambda_1 + \mu)$-strongly convex when $A.1(a)$ holds and $(\lambda_1 - L - \frac{\gamma_1}{\rho})$-strongly convex when $A.1(b)$ and $\lambda_1 > L + \frac{\gamma_1}{\rho}$ holds. Follow the proof in [19], we obtain the result in Lemma 2. $\qquad\square$

Proof of Lemma 3

Proof.

$$
\begin{aligned}
&\mathbb{E}\left[\left\|g_i^{t,r} - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2\right] \\
&\overset{(a)}{\leq} 2\mathbb{E}\left[\left\|g_i^{t,r} - \nabla F_i\left(\boldsymbol{w}_i^{t,r}\right)\right\|_2^2 + \left\|\nabla F_i\left(\boldsymbol{w}_i^{t,r}\right) - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2\right] \\
&\overset{(b)}{=} \frac{2}{J}\sum_{j=1}^{J}(\bar{\lambda}^2\mathbb{E}\left[\left\|\tilde{\boldsymbol{\theta}}_{i,j}^{t,r+1} - \hat{\boldsymbol{\theta}}_{i,j}^{t,r+1}\right\|_2^2\right] + L_F^2\mathbb{E}\left[\left\|\boldsymbol{w}_i^{t,r} - \boldsymbol{w}^t\right\|_2^2\right]) \\
&\leq 2\left(\bar{\lambda}^2\delta^2 + L_F^2\mathbb{E}\left[\left\|\boldsymbol{w}_i^{t,r} - \boldsymbol{w}^t\right\|_2^2\right]\right)
\end{aligned}
\tag{18}
$$

where (a) follows by Jensen's inequality; (b) is due to Theorem 1 and the first-order conditions $\lambda_1(\hat{\boldsymbol{y}}_{i,j} - \hat{\boldsymbol{\theta}}_{i,j}) + \lambda_2(\hat{\boldsymbol{y}}_{i,j} - \boldsymbol{w}) = 0$ from (3) with $\bar{\lambda} = \frac{\lambda_1\lambda_2}{\lambda_1 + \lambda_2}$. Moreover, the last term is bounded by

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{w}_i^{t,r} - \boldsymbol{w}^t\right\|_2^2\right] \\
&\overset{(a)}{\leq} \frac{8\tilde{\eta}}{\beta L_F}\left(3\mathbb{E}\frac{1}{J}\sum_{j=1}^{J}\left[\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right)\right\|_2^2\right] + \frac{2\bar{\lambda}^2\delta^2}{R}\right)
\end{aligned}
$$

where (a) follows by our proof of Lemma 2 in [38]. Thus we finish the proof of Lemma 3. $\qquad\square$

Proof of Lemma 4

Proof. We first prove case (a), we have

$$
\begin{aligned}
&\frac{1}{NJ}\sum_{i=1,j=1}^{N,J}\|\nabla F_{i,j}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|_2^2 \\
&\overset{(a)}{=} \frac{1}{NJ}\sum_{i,j=1}^{N,j}\|\nabla F_{i,j}(\boldsymbol{w})\|_2^2 - \left\|\frac{1}{NJ}\sum_{p,q=1}^{N}\nabla F_{i,j}(\boldsymbol{w})\right\|_2^2 \\
&\overset{(b)}{\leq} \frac{2}{NJ}\sum_{i,j=1}^{N,J}\|\nabla F_{i,j}(\boldsymbol{w}) - \nabla F_{i,j}(\boldsymbol{w}^*)\|_2^2 + \|\nabla F_{i,j}(\boldsymbol{w}^*)\|_2^2 \\
&\overset{(c)}{\leq} 4L_F\left(F(\boldsymbol{w}) - F(\boldsymbol{w}^*)\right) + \frac{2}{NJ}\sum_{i,j=1}^{N}\|\nabla F_{i,j}(\boldsymbol{w}^*)\|_2^2
\end{aligned}
$$

where (a) follows by $\mathbb{E}\left[\|\boldsymbol{x} - \mathbb{E}\left[\boldsymbol{x}\right]\|_2^2\right] = \mathbb{E}\left[\|\boldsymbol{x}\|_2^2\right] - \mathbb{E}\left[\|\boldsymbol{x}\|_2\right]^2$; (b) follows by the Jensen's inequality; (c) follows by Theorem 1, thus F_i is μ_F-strong convex and L_F-smooth

with $\nabla F(\boldsymbol{w}^*) = 0$.

We next prove case (b). For the minimization problem in (3), we have the first-order conditions $\lambda_1(\hat{\boldsymbol{y}}_{i,j} - \hat{\boldsymbol{\theta}}_{i,j}) + \lambda_2(\hat{\boldsymbol{y}}_{i,j} - \boldsymbol{w}) = 0$ and $\nabla \ell_{i,j}(\hat{\boldsymbol{\theta}}_{i,j}) - \lambda_1(\hat{\boldsymbol{y}}_{i,j} - \hat{\boldsymbol{\theta}}_{i,j}) + \gamma_1 \nabla \phi_\rho(\hat{\boldsymbol{\theta}}_{i,j}) = 0$, then we have

$$
\begin{aligned}
\nabla F_{i,j}(\boldsymbol{w}) &= \lambda_2(\boldsymbol{w} - \hat{\boldsymbol{y}}_{i,j}) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}) \\
&= \lambda_1(\hat{\boldsymbol{y}}_{i,j} - \hat{\boldsymbol{\theta}}_{i,j}) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}) \\
&= \nabla \ell_{i,j}(\hat{\boldsymbol{\theta}}_{i,j}) + \gamma_1 \nabla \phi_\rho(\hat{\boldsymbol{\theta}}_{i,j}) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}) \\
&= \nabla L_{i,j}(\hat{\boldsymbol{\theta}}_{i,j}) + \gamma_2 \nabla \phi_\rho(\boldsymbol{w}), \quad (19)
\end{aligned}
$$

where $L_{i,j}(\cdot) = \ell_{i,j}(\cdot) + \gamma_1 \phi_\rho(\cdot)$.

Hence, we have

$$
\begin{aligned}
&\|\nabla F_{i,j}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|_2^2 \\
&\overset{(a)}{\leq} 2 \left\| \nabla L_{i,j}\left(\hat{\boldsymbol{\theta}}_{i,j}\right) - \frac{1}{NJ} \sum_{p,q=1}^{NJ} \nabla L_{p,q}\left(\hat{\boldsymbol{\theta}}_{i,j}\right) \right\|_2^2 \\
&\quad + 2 \left\| \frac{1}{NJ} \sum_{p,q=1}^{NJ} \left(\nabla L_{p,q}\left(\hat{\boldsymbol{\theta}}_{i,j}\right) - \nabla L_{p,q}\left(\hat{\boldsymbol{\theta}}_{p,q}\right) \right) \right\|_2^2,
\end{aligned}
$$

where (a) follows by (19) and the Jensen's inequality. Taking the average over the number of clients, we have

$$
\begin{aligned}
&\frac{1}{NJ} \sum_{i,j=1}^{N,J} \|\nabla F_{i,j}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|_2^2 \\
&\overset{(a)}{\leq} 2\sigma_\ell^2 + \frac{2}{N^2 J^2} \sum_{i,j=1}^{N,J} \sum_{p,q=1}^{N,J} \left\| \nabla L_{p,q}\left(\hat{\boldsymbol{\theta}}_{i,j}\right) - \nabla L_{p,q}\left(\hat{\boldsymbol{\theta}}_{p,q}\right) \right\|_2^2 \\
&\overset{(b)}{\leq} 2\sigma_\ell^2 + \frac{2(L + \frac{\gamma_1}{\rho})^2}{N^2 J^2} \sum_{i,j=1}^{N,J} \sum_{p,q=1}^{N,J} \left\| \hat{\boldsymbol{\theta}}_{i,j} - \hat{\boldsymbol{\theta}}_{p,q} \right\|_2^2, \quad (20)
\end{aligned}
$$

where (a) follows by Assumption 3 and the Jensen's inequality; (b) is due to the $(L + \frac{\gamma_1}{\rho})$-smoothness of $L_{i,j}(\cdot)$; and the last term

$$
\begin{aligned}
&\sum_{i,j=1}^{N,J} \sum_{p,q=1}^{N,J} \left\| \hat{\boldsymbol{\theta}}_{i,j} - \hat{\boldsymbol{\theta}}_{p,q} \right\|_2^2 \\
&\overset{(a)}{\leq} 4 \sum_{i,j=1}^{N,J} \sum_{p,q=1}^{N,J} \left(\left\| \hat{\boldsymbol{\theta}}_{i,j} - \hat{\boldsymbol{y}}_{i,j} - \frac{\gamma_2}{\lambda_1} \nabla \phi_\rho(\boldsymbol{w}) \right\|_2^2 \right. \\
&\quad + \left\| \hat{\boldsymbol{\theta}}_{p,q} - \hat{\boldsymbol{y}}_{p,q} - \frac{\gamma_2}{\lambda_1} \nabla \phi_\rho(\boldsymbol{w}) \right\|_2^2 \\
&\quad + \left\| \hat{\boldsymbol{y}}_{i,j} - \boldsymbol{w} - \frac{\gamma_2}{\lambda_2} \nabla \phi_\rho(\boldsymbol{w}) \right\|_2^2 \\
&\quad + \left. \left\| \hat{\boldsymbol{y}}_{p,q} - \boldsymbol{w} - \frac{\gamma_2}{\lambda_2} \nabla \phi_\rho(\boldsymbol{w}) \right\|_2^2 \right) \\
&\overset{(b)}{=} \frac{8NJ(\lambda_1^2 + \lambda_2^2)}{\lambda_1^2 \lambda_2^2} \sum_{i,j=1}^{N,J} \|\nabla F_{i,j}(\boldsymbol{w})\|_2^2, \quad (21)
\end{aligned}
$$

where (a) is due to the Jensen's inequality; (b) follows by (19) and re-arranging the terms. Substituting (21) back to (20) we have

$$
\begin{aligned}
&\frac{1}{NJ} \sum_{i,j=1}^{N,J} \|\nabla F_{i,j}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|_2^2 \\
&\overset{(a)}{\leq} 2\sigma_\ell^2 + \frac{16(L + \frac{\gamma_1}{\rho})^2(\lambda_1^2 + \lambda_2^2)}{\lambda_1^2 \lambda_2^2} \left(\|\nabla F(\boldsymbol{w})\|_2^2 \right. \\
&\quad \left. + \frac{1}{NJ} \sum_{i,j=1}^{N,J} \|\nabla F_{i,j}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|_2^2 \right) \\
&\overset{(b)}{\leq} \frac{16(L + \frac{\gamma_1}{\rho})^2}{\lambda^2 - 16(L + \frac{\gamma_1}{\rho})^2} \|\nabla F(\boldsymbol{w})\|_2^2 + \frac{2\lambda^2}{\lambda^2 - 16(L + \frac{\gamma_1}{\rho})^2} \sigma_\ell^2,
\end{aligned}
$$

where (a) follows by the equation $\mathbb{E}\left[\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\|_2^2\right] = \mathbb{E}\left[\|\boldsymbol{x}\|_2^2\right] - \mathbb{E}[\|\boldsymbol{x}\|_2]^2$; (b) is by re-arranging the terms with setting $\lambda = \frac{\lambda_1 \lambda_2}{\sqrt{\lambda_1^2 + \lambda_2^2}}$. $\qquad \square$

Proof of Theorem 2

Completing the proof of Theorem 1:

Proof. Let $\tilde{\eta} \leq \min\left\{ \frac{1}{\mu_F}, \frac{1}{3L_F(3 + 128L_F/(\mu_F \beta))} \right\}$ hold, combining Lemma 3 and Lemma 4 with Lemma 1 yields

$$
\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|_2^2\right] \\
&\overset{(a)}{\leq} \Delta_t - \tilde{\eta}\mathbb{E}\left[F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*)\right] + \tilde{\eta}M_1 + \tilde{\eta}^2 M_2,
\end{aligned}
$$

where Δ_t denotes $\mathbb{E}\left[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|_2^2\right]$, $M = 2 - \tilde{\eta}L_F(18 + \frac{768L_F}{\mu_F \beta})$, $M_1 = \frac{8\delta^2 \bar{\lambda}^2}{\mu_F}$, $M_2 = 6(1 + 64\frac{L_F}{\mu_F \beta})\sigma_{F,1}^2 + \frac{128L_F \delta^2 \bar{\lambda}^2}{\mu_F \beta}$, and (a) is due to $\tilde{\eta} \leq \frac{1}{3L_F(3 + 128L_F/(\mu_F \beta))}$, then

$$
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*)\right] \leq \frac{\Delta_0}{\tilde{\eta}T} + M_1 + \tilde{\eta}M_2. \quad (22)
$$

Then, the proof of part (a) is finished.

We next prove the part (b), we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \tilde{\theta}_{i,j}^T - \boldsymbol{w}^* \right\|_2^2\right] \\
&\overset{(a)}{\leq} 5\mathbb{E}\left(\left\| \tilde{\theta}_{i,j}^T - \hat{\theta}_{i,j}^T \right\|_2^2 + \left\| \hat{\theta}_{i,j}^T - \hat{\boldsymbol{y}}_{i,j}^T - \frac{\gamma_2}{\lambda_1} \nabla \phi_\rho(\boldsymbol{w}^T) \right\|_2^2 \right. \\
&\quad + \left\| \hat{\boldsymbol{y}}_{i,j}^T - \boldsymbol{w}^T - \frac{\gamma_2}{\lambda_2} \nabla \phi_\rho(\boldsymbol{w}^T) \right\|_2^2 + \left\| \boldsymbol{w}^T - \boldsymbol{w}^* \right\|_2^2 \\
&\quad + \left. \left\| \frac{\gamma_2(\lambda_1 + \lambda_2)}{\lambda_1 \lambda_2} \nabla \phi_\rho(\boldsymbol{w}^T) \right\|_2^2 \right) \\
&\overset{(b)}{\leq} 5(\delta^2 + \frac{\gamma_2^2}{\bar{\lambda}^2} \mathbb{E}\left[\|\nabla \phi_\rho(\boldsymbol{w}^T)\|_2^2\right] + \mathbb{E}\left[\|\boldsymbol{w}^T - \boldsymbol{w}^*\|_2^2\right]) \\
&\quad + \frac{10}{\lambda^2} \mathbb{E}\left[\|\nabla F_{i,j}(\boldsymbol{w}^T) - \nabla F_{i,j}(\boldsymbol{w}^*)\|_2^2 + \|\nabla F_{i,j}(\boldsymbol{w}^*)\|_2^2\right] \\
&\overset{(c)}{\leq} 5(\delta^2 + \frac{\gamma_2^2}{\bar{\lambda}^2} \mathbb{E}\left[\|\nabla \phi_\rho(\boldsymbol{w}^T)\|_2^2\right] + \mathbb{E}\left[\|\boldsymbol{w}^T - \boldsymbol{w}^*\|_2^2\right]) \\
&\quad + \frac{10}{\lambda^2} \mathbb{E}\left[L_F \|\boldsymbol{w}^T - \boldsymbol{w}^*\|_2^2 + \|\nabla F_{i,j}(\boldsymbol{w}^*)\|_2^2\right], \quad (23)
\end{aligned}
$$

where (a) and (b) are due to the Jensen's inequality and (19); and (c) is due to Theorem 1. In addition, the second term of (23) is bounded by

$$\mathbb{E}\left[\left\|\nabla\phi_\rho(\boldsymbol{w}^T)\right\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^d\left(\frac{\sinh\left(|w_n^T|/\rho\right)\,w_n^T}{\cosh\left(|w_n^T|/\rho\right)|w_n^T|}\right)^2\right] \le d_s^2, \quad (24)$$

where $\frac{\sinh(|w_n^T|/\rho)}{\cosh(|w_n^T|/\rho)} \in (-1,1)$, w_n^T is the n-th element of $\boldsymbol{w}^T$ and d_s denotes the number of non-zero elements in $\boldsymbol{w}$. Due to the μ_F-strong convex of F, we have

$$\mathbb{E}\left[\left\|\boldsymbol{w}^T - \boldsymbol{w}^*\right\|_2^2\right] \le \frac{2\left(\mathbb{E}\left[F\left(\boldsymbol{w}_T\right) - F\left(\boldsymbol{w}^*\right)\right]\right)}{\mu_F}. \quad (25)$$

By substituting (22), (24) and (25) into (23) we complete the proof. $\qquad\square$

Proof of Theorem 3

Proof. We first prove part (a). Similar with the proof in [19], we rewrite the recursive formula due to the L-smooth of $F(\cdot)$

$$\mathbb{E}\left[F\left(\boldsymbol{w}^{t+1}\right) - F\left(\boldsymbol{w}^t\right)\right]$$

$$\le \mathbb{E}\left[\left\langle\nabla F\left(\boldsymbol{w}^t\right), \boldsymbol{w}^{t+1} - \boldsymbol{w}^t\right\rangle\right] + \frac{L_F}{2}\mathbb{E}\left[\left\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\right\|_2^2\right]$$

$$= -\tilde{\eta}\mathbb{E}\left[\left\langle\nabla F\left(\boldsymbol{w}^t\right), g^t\right\rangle\right] + \frac{\tilde{\eta}^2 L_F}{2}\mathbb{E}\left[\left\|g^t\right\|_2^2\right]$$

$$= -\tilde{\eta}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] - \tilde{\eta}\mathbb{E}\left[\left\langle\nabla F\left(\boldsymbol{w}^t\right), g^t - \nabla F\left(\boldsymbol{w}^t\right)\right\rangle\right]$$
$$+ \frac{\tilde{\eta}^2 L_F}{2}\mathbb{E}\left[\left\|g^t\right\|_2^2\right]$$

$$\overset{(b)}{\le} -\tilde{\eta}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] + \frac{\tilde{\eta}}{2}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right]$$
$$+ \frac{\tilde{\eta}}{2}\mathbb{E}\left\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2 + \frac{\tilde{\eta}^2 L_F}{2}\mathbb{E}\left[\left\|g^t\right\|_2^2\right]$$

$$\overset{(c)}{\le} -\frac{\tilde{\eta}(1-3L_F\tilde{\eta})}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}_t)\right\|_2^2\right]$$
$$+ \frac{\tilde{\eta}(1+3L_F\tilde{\eta})}{2}\mathbb{E}\left\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla F_i\left(\boldsymbol{w}^t\right)\right\|_2^2$$
$$+ \frac{3\tilde{\eta}^2 L_F}{2}\mathbb{E}\left\|\frac{1}{S}\sum_{i\in\mathcal{S}^t}\nabla F_i\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2$$

$$\overset{(d)}{\le} -\tilde{\eta}G\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right] + \tilde{\eta}G_1 + \tilde{\eta}^2 G_2.$$

where (b) is due to the Cauchy-Swartz inequalities; (c) is due to the Jensen's inequality; (d) follows by $1+3L_F\tilde{\eta} \le 3\beta$ since $\tilde{\eta} \le \frac{\beta}{2L_F}$, and $G_1 = 3\beta\bar{\lambda}^2\delta^2$, $G_2 = \frac{3L_F\left(49R\sigma_{F,2}^2+16\delta^2\bar{\lambda}^2\right)}{R}$.

Next, let $\lambda^2 - 16(L+\frac{\gamma_1}{\rho})^2 \ge 1$ and $\tilde{\eta} \le \frac{1}{588L_F\lambda^2}$ hold, we have

$$G = \frac{1}{2} - \tilde{\eta}L_F\left(\frac{3}{2} + \frac{24(L+\frac{\gamma_1}{\rho})^2}{\lambda^2 - 16(L+\frac{\gamma_1}{\rho})^2} + \frac{72\lambda^2}{\lambda^2 - 16(L+\frac{\gamma_1}{\rho})^2}\right)$$
$$\ge \frac{1}{2} - \frac{147\tilde{\eta}L_F\lambda^2}{2} \ge \frac{1}{4},$$

we hence have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2\right]$$

$$\le 4\left(\frac{\mathbb{E}\left[F\left(\boldsymbol{w}^0\right) - F\left(\boldsymbol{w}^T\right)\right]}{\tilde{\eta}T} + G_1 + \tilde{\eta}^1 G_2\right).$$

$$\le 4\left(\frac{\Delta_F}{\tilde{\eta}T} + G_1 + \tilde{\eta}^1 G_2\right),$$

where $\Delta_F \triangleq F(\boldsymbol{w}^0) - F(W^*)$.

We next prove part (b), we have

$$\frac{1}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left[\left\|\tilde{\boldsymbol{\theta}}_{i,j}^t\left(\boldsymbol{w}^t\right) - \boldsymbol{w}^t\right\|_2^2\right]$$

$$\overset{(a)}{\le} \frac{4}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left(\left\|\tilde{\theta}_{i,j}^t - \hat{\theta}_{i,j}^t\right\|_2^2 + \left\|\hat{\theta}_{i,j}^t - \hat{\boldsymbol{y}}_{i,j}^t - \frac{\gamma_2}{\lambda_1}\nabla\phi_\rho(\boldsymbol{w^t})\right\|_2^2\right.$$
$$\left. + \left\|\hat{\boldsymbol{y}}_{i,j}^t - \boldsymbol{w}^t - \frac{\gamma_2}{\lambda_2}\nabla\phi_\rho(\boldsymbol{w^t})\right\|_2^2 + \left\|\frac{\gamma_2(\lambda_1+\lambda_2)}{\lambda_1\lambda_2}\nabla\phi_\rho(\boldsymbol{w^t})\right\|_2^2\right)$$

$$\overset{(b)}{\le} 4\left(\delta^2 + \frac{\lambda_1^2+\lambda_2^2}{\lambda_1^2\lambda_2^2}\frac{1}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left[\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right)\right\|_2^2\right]\right.$$
$$\left. + \frac{\gamma_2^2(\lambda_1+\lambda_2)^2}{\lambda_1^2\lambda_2^2}\frac{1}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left[\left\|\nabla\phi_\rho(\boldsymbol{w}^t)\right\|_2^2\right]\right)$$

$$= 4\left(\delta^2 + \frac{1}{\lambda^2}\frac{1}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left[\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right)\right\|_2^2\right] + \frac{\gamma_2^2}{\bar{\lambda}^2}d_s^2\right)$$
$$(26)$$

where (a) is due to the Jensen's inequality, (b) is due to (19). And the last term is bounded by

$$\frac{1}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left[\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right)\right\|_2^2\right]$$

$$\overset{(a)}{=} \frac{1}{NJ}\sum_{i,j=1}^{NJ}\mathbb{E}\left[\left\|\nabla F_{i,j}\left(\boldsymbol{w}^t\right) - \nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2 + \left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right]$$

$$\overset{(b)}{\le} \sigma_{F,2}^2 + \frac{\lambda^2}{\lambda^2 - 16(L+\frac{\gamma_1}{\rho})^2}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right]$$

$$\overset{(c)}{\le} \sigma_{F,2}^2 + \lambda^2\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|_2^2\right]$$
$$(27)$$

where (a) is due to the equation $\mathbb{E}\left[\left\|\boldsymbol{x} - \mathbb{E}\left[\boldsymbol{x}\right]\right\|_2^2\right] = \mathbb{E}\left[\left\|\boldsymbol{x}\right\|_2^2\right] - \mathbb{E}\left[\left\|\boldsymbol{x}\right\|_2\right]^2$, (b) follows by Lemma 4, (c) is by setting $\lambda^2 - 16(L+\frac{\gamma_1}{\rho})^2 \le 1$. By substituting (27) into (26), we finished the proof. $\qquad\square$

REFERENCES

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

[2] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

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

[4] W. Zhang, Q. Lu, Q. Yu, Z. Li, Y. Liu, S. K. Lo, S. Chen, X. Xu, and L. Zhu, "Blockchain-based federated learning for device failure detection in industrial iot," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5926–5937, 2021.

[5] B. Ghimire and D. B. Rawat, "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8229–8249, 2022.

[6] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys Tutorials*, vol. 22, no. 4, pp. 2167–2191, 2020.

[7] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2021.

[8] J. Le, J. Lei, N. Mu, H. Zhang, K. Zeng, and X. Liao, "Federated continuous learning with broad network architecture," *IEEE Transactions on Cybernetics*, vol. 51, no. 8, pp. 3874–3888, 2021.

[9] S. R. Pokhrel and J. Choi, "Federated learning with blockchain for autonomous vehicles: Analysis and design challenges," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4734–4746, 2020.

[10] M. A. Ferrag, O. Friha, L. Maglaras, H. Janicke, and L. Shu, "Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis," *IEEE Access*, vol. 9, pp. 138 509–138 542, 2021.

[11] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1759–1799, 2021.

[12] Y. Wang, G. Gui, H. Gacanin, B. Adebisi, H. Sari, and F. Adachi, "Federated learning for automatic modulation classification under class imbalance and varying noise condition," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 1, pp. 86–96, 2021.

[13] Z. He, J. Yin, Y. Wang, G. Gui, B. Adebisi, T. Ohtsuki, H. Gacanin, and H. Sari, "Edge device identification based on federated learning and network traffic feature engineering," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 4, pp. 1898–1909, 2021.

[14] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

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

[16] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3710–3722, 2021.

[17] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.

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

[19] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with moreau envelopes," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21 394–21 405.

[20] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," in *Proc. of Association for the Advancement of Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7865–7873.

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

[22] I. Achituve, A. Shamsian, A. Navon, G. Chechik, and E. Fetaya, "Personalized federated learning with gaussian processes," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8392–8406.

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

[24] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *Proc. of IEEE International Conference on Communications*. IEEE, 2020, pp. 1–6.

[25] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[26] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-iid data quagmire of decentralized machine learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4387–4398.

[27] 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.

[28] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[30] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2351–2363, 2020.

[31] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. of Advances in neural information processing systems*, 2017, pp. 4424–4434.

[32] N. Shoham, T. Avidor, A. Keren, N. Israel, D. Benditkis, L. Mor-Yosef, and I. Zeitak, "Overcoming forgetting in federated learning on non-iid data," *arXiv preprint arXiv:1910.07796*, 2019.

[33] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere i: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[34] A. Beck, *First-order methods in optimization*. SIAM, 2017.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[36] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[37] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Master's thesis, University of Tront*, 2009.

[38] X. Liu, Y. Li, Q. Wang, X. Zhang, Y. Shao, and Y. Geng, "Sparse personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.