A New Look and Convergence Rate of Federated Multi-Task Learning with Laplacian Regularization

Canh T. Dinh, Tung T. Vu, *Member, IEEE*, Nguyen H. Tran, *Senior Member, IEEE*, Minh N. Dao, Hongyu Zhang, *Senior Member, IEEE*

***Abstract*—Non-Independent and Identically Distributed (non-IID) data distribution among clients is considered as the key factor that degrades the performance of federated learning (FL). Several approaches to handle non-IID data such as personalized FL and federated multi-task learning (FMTL) are of great interest to research communities. In this work, first, we formulate the FMTL problem using Laplacian regularization to explicitly leverage the relationships among the models of clients for multi-task learning. Then, we introduce a new view of the FMTL problem, which in the first time shows that the formulated FMTL problem can be used for conventional FL and personalized FL. We also propose two algorithms FedU and dFedU to solve the formulated FMTL problem in communication-centralized and decentralized schemes, respectively. Theoretically, we prove that the convergence rates of both algorithms achieve linear speedup for strongly convex and sublinear speedup of order $1/2$ for nonconvex objectives. Experimentally, we show that our algorithms outperform the conventional algorithm FedAvg, FedProx, SCAFFOLD, and AFL in FL settings, MOCHA in FMTL settings, as well as pFedMe and Per-FedAvg in personalized FL settings.**

***Index Terms*—Federated multi-task learning, federated learning, personalized learning, Laplacian regularization.**

I. Introduction

Recently, federated learning (FL) has been considered as a promising distributed and privacy-preserving method for building a global model from a massive number of hand-held devices [1]–[4]. FL has a wide range of futuristic applications, such as detecting the symptoms of possible diseases (e.g., stroke, heart attack, diabetes) from wearable devices in healthcare systems [5]–[7], or predicting disaster risks from internet-of-things devices in smart cities [8], [9]. In FL, one of the key challenges is the naturally non-IID data distributions among clients [10], [11]. When the differences among clients' data distributions increase, the generalization error of the FL global model on each client's local data significantly increases [12], [13].

Personalized FL [14], [15] and federated multi-task learning (FMTL) [16] have been proposed as solutions to handle non-IID data distributions among clients. Personalized FL aims to build a global model that is leveraged to find a "personalized model" for each client's local data. Here, the global model is considered as an "agreed point" for each client to start personalizing its model based on its heterogeneous local data distribution. Different from personalized FL, FMTL aims to simultaneously learn separate models, which is motivated by multi-task learning frameworks [17], [18]. Each of these models fits the data distribution of each client. Therefore, FMTL directly addresses the issue stemming from non-IID data distributions without building any global model as personalized FL.

On the other hand, from the aspect of the local data at clients, it is observed that the clients with similar features (e.g, location, time, age, gender) are likely to share similar behaviors. Therefore, although the clients' models are separated, they are normally related to each other. In FMTL, the relationships among the clients' models are captured by a regularization term which is minimized to encourage the clients' models to be mutually impacted. Unfortunately, these relationships have not been clearly taken into consideration in the FMTL problem. Moreover, communication-decentralized and non-convex FMTL algorithms with guaranteed convergence are generally less explored.

The main contributions of this work are as follows:

- We formulate a FMTL problem using Laplacian regularization to explicitly leverage the relationships among the models of clients. We then introduce a new view of the FMTL problem that the formulated FMTL problem can be used not only for the conventional FL but also personalized FL.
- We propose a communication-centralized FMTL algorithm FedU, and its decentralized version dFedU to solve the formulated FMTL problem. We also analyze the convergence rate of FMTL algorithms with both convex and nonconvex objective functions. In particular, FedU and dFedU are proved to achieve a linear speedup (resp. sublinear speedup of order $1/2$) for strongly convex (resp. nonconvex) objective cases.
- We empirically evaluate the performance of FedU and dFedU using real datasets that capture the non-IID data distribution among clients. We show that in terms of local accuracy, FedU and dFedU outperform the traditional algorithm FedAvg in FL settings, the conventional algorithm MOCHA in FMTL settings, as well as pFedMe and Per-FedAvg in personalized FL settings.

C. T. Dinh and N. H. Tran are with the School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (email: {canh.dinh,nguyen.tran}@sydney.edu.au)

T. T. Vu is with Institute of Electronics, Communications, and Information Technology (ECIT), Queen's University Belfast, Belfast BT3 9DT, UK (e-mail: t.vu@qub.ac.uk).

M. N. Dao is with the School of Engineering, Information Technology and Physical Sciences, Federation University, Ballarat, VIC 3353, Australia (e-mail: m.dao@federation.edu.au).

H. Zhang is with the University of Newcastle, Callaghan, NSW 2308, Australia (e-mail: hongyu.zhang@newcastle.edu.au)

II. Related Work

Federated Learning. One of the earliest work of FL is FedAvg [1], which builds the global model based on averaging the local Stochastic Gradient Descent (SGD) updates. Various methods [11], [19]–[22] are introduced to improve the robustness of the global model under non-i.i.d settings. For example, FedProx [19] adds a proximal term to the local objective, therefore addressing the statistical heterogeneity of clients.

Personalized Federated Learning. Several personalized FL approaches have been proposed to tackle the issues sterming form non-IID data in the conventional FL. Mixture methods [13], [23] attempted to combine a local model with the global model, while [24] applied this mixing to jointly learns compact local representations on each client and a global model across all clients. Motivating by creating a well-generalized global model to quickly adapt to client's data after few gradient descent steps, pFedMe [14] used Moreau envelopes, while Per-FedAvg [15] took advances of meta learning approaches: model-agnostic meta-learning [25]. [26] proposed the combination of FedAvg and Reptile [27] to improve FL personalization. A different personalized FL approach to train deep neural networks is FedPer [28]. Clients share a set of base layers with a server and keep personalization layers that adapt quickly to the local data.

Federated Multi-Task Learning. Another approach to deal with the non-IID data distributions at clients is learning separate models each of which fits each local data distribution. In this sense, FMTL was first introduced in [16] where a systems-aware optimization framework MOCHA for handling stragglers and fault tolerance in FL settings is proposed. Besides that, there are also several other works studying FMTL. [29] proposed a framework for generalized total variation minimization, which is useful in FMTL networks. [30] introduced a FMTL algorithm to deal with the issues of accuracy, fairness and robustness in FL. By treating the FL network as a star-shaped Bayesian network, [31] developed a FMTL algorithm using approximated variational inference. [32] focused on a FMTL algorithm for online applications. However, in all these works, the convergence rate of FMTL with nonconvex objectives has not been studied. Moreover, the relations among the problems of FMTL, the standard FL, and personalized FL are not yet investigated in the literature.

III. Federated Multi-Task Learning: A New View

A. The Formulation of the FMTL Problem with Laplacian Regularization

In this work, the goal of FMTL is to fit separate models (i.e., $w_k \in \mathbb{R}^d, \forall k \in \mathcal{N}$) to the local data of clients, taking into account the relationships among these models. For instance, smart-device clients in a mobile network are trying to learn their activities using their personal and private data (e.g., image, text, voice, and sensor data). In FL settings, their data may come from different environments, contexts, and applications, and thus, have non-IID distributions. Despite of this, these clients are likely to behave similarly under similar features or scenarios (e.g., location, time, age). Therefore, there normally exist relationships among the models of clients [33]–[35].

To present the relationships among the models of clients, we consider a connected graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, A\}$, where $\mathcal{N} := \{1, \dots, N\}$ is the set of vertices representing federated learning clients, $\mathcal{E}$ is the set of edges representing relationships among the models of clients, and $A \in \mathbb{R}^N$ is a symmetric, weighted adjacency matrix with $a_{k\ell} := [A]_{k\ell}$. The relationship between clients k and ℓ is presented by $a_{k\ell}$ and reversible, i.e., $a_{k\ell} = a_{\ell k}, \forall k, \ell$. Here, $a_{k\ell} = 0$ means no relationship between the models of clients k and ℓ. The value of $a_{k\ell} > 0$ shows that client k is a neighbor of client ℓ and also determines the strength of the relationship between these two clients' models. Let $D \in \mathbb{R}^N$ be a diagonal matrix in which $[D]_{kk} = \sum_{\ell=1}^{N} a_{k\ell}$. The Laplacian matrix of the graph is thus $L = D - A$.

Let $W = [w_1^T, ..., w_N^T]^T \in \mathbb{R}^{dN}$ be a collective model vector and $\mathcal{L} := L \otimes I_d$ be a Laplacian regularization matrix. Now, we formulate the following FMTL problem:

$$\min_W J(W) = \underbrace{F(W)}_{\text{Global loss}} + \underbrace{\eta R(W)}_{\text{Laplacian regularization}}, \tag{1}$$

where

$$F(W) = \sum_{k=1}^{N} F_k(w_k), \tag{2}$$

$$\mathcal{R}(W) = W^T \mathcal{L} W = \frac{1}{2} \sum_{k=1}^{N} \sum_{\ell \in \mathcal{N}_k} a_{k\ell} ||w_k - w_\ell||^2, \tag{3}$$

$\mathcal{N}_k = \mathcal{N} \setminus \{k\}$, and $\|\cdot\|$ is the Euclidean norm. $F_k(\cdot)$ represents the expected loss function at client k:

$$F_k(w_k) = \mathbb{E}_{\zeta_k}[f_k(w_k; \zeta_k)],$$

where ζ_k is a random data sample drawn from the distribution of client k and $f_k(w_k; \zeta_k)$ is the regularized loss function corresponding to this sample and w_k. The distribution of ζ_k and ζ_ℓ can be distinct when $k \neq \ell$.

Note that in our work, we do not extract the similarity of the existing relationships between the clients by any visualization methods in order to develop our proposed method. Instead, we present the existing relationships among the models of the clients by a Laplacian regularization matrix $\mathcal{L}$ and put it into the Laplacian regularization term in the objective function of the federated multitask-learning problem (1). Theoretically, in (1), $\eta \geq 0$ is a regularization hyperparameter that controls the impact of the models of neighboring clients on each local model. If $\eta = 0$, (1) turns to an individual learning problem where each client learns its local model w_k based on its own local data without collaboration with server or other clients. If $\eta > 0$, minimizing the Laplacian regularization term encourages the models of the neighboring clients to be close to each other. The impacts of the existing relationship between the models of the clients on the performance of our proposed algorithms will be shown in the later section of experiment.

Remark 1. *There are other methods of regularization to encourage the models of the neighboring clients to be close to each other, e.g., using* $||w_k - w_\ell||$ *instead of* $||w_k - w_\ell||^2$ *in* (3) *as Network Lasso does [36]–[38], or using* $\mathrm{tr}(\widehat{W}\Omega\widehat{W}^T)$ *instead of*

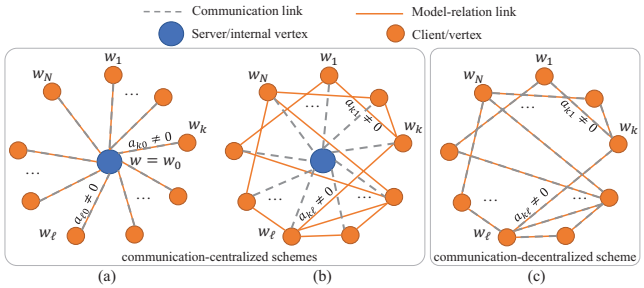


Fig. 1: Illustrations of undirected weighted graphs in FL. (a): Star graph with a server for traditional FL and personalized FL; (b) and (c): Entity graph with and without server for FMTL

(3) *as MOCHA does [16], where* $\widehat{W} := [w_1, \ldots, w_N] \in \mathbb{R}^{d\times N}$. *On the other hand, problem* (1) *is a generalization of the problem in [39] where several algorithms are developed for strongly convex objectives. Problem* (1) *is also similar to the generalized total variation minimization problem [29] which is solved by a primal-dual method for convex objectives. [40] has a convex version of problem* (1) *which is solved by a decentralized algorithm using Alternating Direction Method of Multipliers (ADMM). In* (1)*, we present the FTML problem using the Laplacian regularization matrix* $\mathcal{L}$. *Utilizing the special properties of* $\mathcal{L}$*, we successfully design FMTL algorithms using SGD. Importantly, our algorithms can work (i) in both centralized and decentralized communication schemes, and (ii) with both strongly convex and nonconvex objective functions.*

Assumption 1 (Smoothness)**.** *For each* $k \in \mathcal{N}$, F_k *is* β*-smooth, i.e., for any* $w, w' \in \mathbb{R}^d$,

$$\|\nabla F_k(w) - \nabla F_k(w')\| \leq \beta\|w - w'\|.$$

Assumption 2 (Strong convexity)**.** *For each* $k \in \mathcal{N}$, F_k *is* α*-strongly convex, i.e., for any* $w, w' \in \mathbb{R}^d$,

$$F_k(w) \geq F_k(w') + \langle \nabla F_k(w'), w - w'\rangle + \frac{\alpha}{2}\|w - w'\|^2.$$

Assumption 3 (Bounded variance)**.** *The set of* $\nabla \widetilde{F}_k(w, \zeta_k)$, $k \in \mathcal{N}$ *is unbiased stochastic gradients of* $\nabla F_k(w)$, $k \in \mathcal{N}$, *with total variance bounded by* σ_1^2*, i.e., for any* $W \in \mathbb{R}^{dN}$,

$$\sum_{k=1}^N \mathbb{E}_{\zeta_k} \|\nabla \widetilde{F}_k(w_k, \zeta_k) - \nabla F_k(w_k)\|^2 \leq \sigma_1^2.$$

We note that Assumption 3 is weaker than the assumption of individual bounded variance that is used at each client in FL and personalized FL problems [10], [14], [15]. It should also be noted that (1) shares some similarities to the multi-task learning problem of [41], [42]. However, the latter requires that each $F_k(w_k)$ is twice differential with the Hessian $\nabla^2_{w_k} F_k(w_k)$ uniformly bounded from below and above, which is more restrictive than our assumptions. Moreover, this problem does not take into account the issue of non-IID data distributions among clients, and thus it is not formulated for FL settings.

B. A New View of the FMTL Problem

We first observe that in conventional FL and personalized FL, all clients connect to a server under a communication-centralized scheme shown in Figure 1(a). The relationships among the models of the clients and the server are presented by a star graph. In this graph, a server is considered as a virtually internal vertex 0 with its loss function $F_0 = 0$ and a model w_0. Here, all the models of clients are only related to the server model w_0, i.e., $a_{k0} > 0, \forall k$, but not with each other, i.e., $a_{k\ell} = 0, \forall k, \ell \neq 0$. In this work, we assume that the weights $a_{k\ell}$ are known and focus on the development of FMTL algorithms to solve problem (1). The finding of $a_{k\ell}$ in specific learning applications are referred to [43], [44]. In what follows, we show that the formulated FMTL problem (1) can be used for the conventional FL and some types of personalized FL. For a more general optimization problem of personalized FL, we refer to LSGD-PFL [45].

Relation of FMTL to conventional FL: The objective function of (1) can be seen as a Lagrangian function of the following problem

$$\min_W \sum_{k=1}^N F_k(w_k), \text{ s.t. } w_1 = w_2 = \ldots = w_N, \tag{4}$$

which is equivalent to the conventional FL problem (FedAvg) [1]. Therefore, the solution of the conventional FL problem can be obtained by solving (1).

Relation of FMTL to personalized FL with Moreau envelopes (pFedMe): The problem of pFedMe [14] is formulated as

$$\min_w J(w) = \sum_{k=1}^N \tilde{J}_k(w), \tag{5}$$

where $\tilde{J}_k(w) = \min_{z_k} F_k(z_k) + \frac{\eta}{2}||z_k - w||^2$. We observe that

$$\begin{aligned} J(w) &= \sum_{k=1}^N \min_{z_k} \left(F_k(z_k) + \frac{\eta}{2}\|z_k - w\|^2\right) \\ &= \min_{z_1,\ldots,z_N} \sum_{k=1}^N \left(F_k(z_k) + \frac{\eta}{2}\|z_k - w\|^2\right). \end{aligned}$$

Therefore, (5) is equivalent to the following problem with $z_0 = w$ and $F_0 \equiv 0$:

$$\min_{z_0,z_1,\ldots,z_N} \sum_{k=0}^N F_k(z_k) + \frac{\eta}{2}\sum_{k=0}^N \|z_k - z_0\|^2,$$

which is a special case of (1) with the star graph topology and $a_{k0} = 1, \forall k \in \mathcal{N}$.

Relation of FMTL to meta-learning-based personalized FL (Per-FedAvg): The problem of Per-FedAvg [15] is given by

$$\min_w J(w) = \sum_{k=1}^N F_k(w - \mu\nabla F_k(w)), \tag{6}$$

where $\mu > 0$ and each F_k is assumed to be L_k-Lipschitz continuous. Set $w_k = w - \mu\nabla F_k(w)$ and $\ell_k = \frac{L_k}{2}$, $k \in \mathcal{N}$. Using Lemma 1.2.3 in [46] twice, we have that, for $\mu < \min_k \ell_k$ and for all $z_k \in \mathbb{R}^d$,

$$\begin{aligned} F_k(w_k) &\leq F_k(w) + \langle \nabla F_k(w), w_k - w\rangle + \ell_k\|w_k - w\|^2 \\ &= F_k(w) - (\mu - \ell_k\mu^2)\|\nabla F_k(w)\|^2 \\ &\leq F_k(z_k) + \langle \nabla F_k(w), z_k - w\rangle + \ell_k\|z_k - w\|^2 \\ &\quad - (\mu - \ell_k\mu^2)\|\nabla F_k(w)\|^2 \\ &= F_k(z_k) + a_{k0}\|z_k - w\|^2 \\ &\quad - (\mu - \ell_k\mu^2)\left\|\nabla F_k(w) - \frac{z_k - w}{2(\mu - \ell_k\mu^2)}\right\|^2, \end{aligned}$$

Algorithm 1 FedU

client k's input: local step-size μ
server's input: graph information $\{a_{k\ell}\}$, initial $w_k^{(0)}, \forall k \in \mathcal{N}$, and global step-size $\widetilde{\mu} = \mu R$
for each round $t = 0, \ldots, T-1$ **do**
  server uniformly samples a subset of clients $\mathcal{S}^{(t)}$ of size S and sends $w_k^{(t)}$ to client $k, \forall k \in \mathcal{S}^{(t)}$
  on client $k \in \mathcal{S}^{(t)}$ **in parallel do**
    initialize local model $w_{k,0}^{(t)} \leftarrow w_k^{(t)}$
    for $r = 0, \ldots, R-1$ **do**
      compute mini-batch gradient $\nabla \widetilde{F}_k(w_{k,r}^{(t)})$
      $w_{k,r+1}^{(t)} \leftarrow w_{k,r}^{(t)} - \mu \nabla \widetilde{F}_k(w_{k,r}^{(t)})$
    end for
    send $w_{k,R}^{(t)}$ to the server
  end on client
  on server do
    $w_{k,R}^{(t)} \leftarrow w_k^{(t)}, \forall k \notin \mathcal{S}^{(t)}$
    $w_k^{(t+1)} \leftarrow w_{k,R}^{(t)} - \widetilde{\mu}\,\eta \sum_{\ell \in \mathcal{N}_k} a_{k\ell}(w_{k,R}^{(t)} - w_{\ell,R}^{(t)}), \forall k \in \mathcal{S}^{(t)}$
    $w_k^{(t+1)} \leftarrow w_k^{(t)}, \forall k \notin \mathcal{S}^{(t)}$
  end on server
end for

where $a_{k0} := \ell_k + \frac{1}{4(\mu - \ell_k \mu^2)}$. Hence,

$$F_k(w_k) \leq \min_{z_k} \left(F_k(z_k) + a_{k0} \|z_k - w\|^2 \right),$$

which implies that

$$\begin{aligned} J(w) &\leq \sum_{k=1}^{N} \min_{z_k} \left(F_k(z_k) + a_{k0} \|z_k - w\|^2 \right) \\ &= \min_{z_1, \ldots, z_N} \sum_{k=1}^{N} \left(F_k(z_k) + a_{k0} \|z_k - w\|^2 \right). \end{aligned}$$

Now, (6) can be solved through its following epigraph problem with $z_0 = w$ and $F_0 = 0$:

$$\min_{z_0, z_1, \ldots, z_N} \sum_{k=0}^{N} F_k(z_k) + \frac{\eta}{2} \sum_{k=0}^{N} a_{k0} \|z_k - z_0\|^2,$$

which is also a special case of (1) with the star graph topology and $a_{k0} = 1, \forall k \in \mathcal{N}$.

IV. Federated Multi-Task Learning: Algorithms

A. FedU: Communication-Centralized Algorithm

In this section, we propose an algorithm FedU, which is presented in Algorithm 1, to solve the formulated FL problem (1) under the communication-centralized scheme. Here, we use an entity graph to capture the relationships among the models of clients as shown in Figure 1(b).[1] First, the server uniformly samples a subset of clients $\mathcal{S}^{(t)}$ and sends the latest update of local model w_k to each client $k, \forall k \in \mathcal{S}^{(t)}$. Then, after R local update steps are performed, the server receives the latest local update from the sampled clients to perform model regularization for each local model.

[1]In an entity graph, each vertex is a value of an entity (e.g., a person) and an edge (e.g., friendship) between two entities exists if these entities are perceived to be similar [43].

Note that in the entity graph, the models of clients are only related to other models but not to any server model, as in the star graph of the conventional FL and personalized FL. Therefore, FedU has a key difference compared to the conventional FL algorithms (e.g., FedAvg [1]) and the personalized FL algorithms (e.g., pFedMe [14], and Per-FedAvg [15]). Instead of updating the personalized models only at the clients using a global model from the server, FedU directly updates each local model at both client and server sides without building a global model.

Specifically, as shown in Figure 2, in each communication round, each client $k \in \mathcal{S}^{(t)}$ copies its current local model received from the server: $w_{k,0}^{(t)} = w_k^{(t)}$, and perform R local updates of the form:

$$w_{k,r+1}^{(t)} \leftarrow w_{k,r}^{(t)} - \mu \nabla \widetilde{F}_k(w_{k,r}^{(t)}),$$

where μ is the local step-size. Then server receives $\{w_{k,R}^{(t)}\}$ from sampled clients $k \in \mathcal{S}^{(t)}$, and updates

$$w_{k,R}^{(t)} \leftarrow w_k^{(t)},$$

for any non-sampled client $k \notin \mathcal{S}^{(t)}$. Finally, the server performs its regularization update for any sampled client $k \in \mathcal{S}^{(t)}$ as

$$w_k^{(t+1)} \leftarrow w_{k,R}^{(t)} - \widetilde{\mu}\,\eta \sum_{\ell \in \mathcal{N}_k \cap \mathcal{S}^{(t)}} a_{k\ell}(w_{k,R}^{(t)} - w_{\ell,R}^{(t)}),$$

and for any non-sampled client $k \notin \mathcal{S}^{(t)}$ as

$$w_k^{(t+1)} \leftarrow w_k^{(t)},$$

where $\widetilde{\mu} = \mu R$ is a global step-size. This step finishes one round of communication.

The mechanism of FedU is explained with $N = 2$ example clients as seen in Figure 2. The two clients are the neighbors of each other and share a certain similarity model. Let (w_1^*, w_2^*) be the global solution (true optimum or true opt.) to problem (1), which is presented by orange squares. Denote by $(\widehat{w}_1^*, \widehat{w}_2^*)$ be the local solution (client optimum or client opt.) that obtains the minimum of the local lost function $F_k(w_k)$, which is presented by blue squares. In the case of non i.i.d data, $\widehat{w}_1^*$ and $\widehat{w}_2^*$ are far away from each other, and $(\widehat{w}_1^*, \widehat{w}_2^*)$ is also far away from (w_1^*, w_2^*). At round t, after making $R = 3$ local updates, the updated models $(w_{1,R}^{(t)}, w_{2,R}^{(t)})$ (blue circles) are moved closer to $(\widehat{w}_1^*, \widehat{w}_2^*)$. Then, we make a further step of regularization update in order to move $w_{1,R}^{(t)}$ toward $\widehat{w}_2^*$ and also move $w_{2,R}^{(t)}$ toward $\widehat{w}_1^*$, which finally makes the updated model after round t, i.e., $(w_1^{(t+1)}, w_2^{(t+1)})$, closer to (w_1^*, w_2^*). By doing local and regularization updates in each round, the converged solution of FedU will be (w_1^*, w_2^*).

B. dFedU: Decentralized Version of FedU

We note that the server in FedU needs to known all the graph information $\{a_{k\ell}\}$. This requirement can be achieved by letting all the clients send the information of their neighbors to the server at the beginning of the learning process. However, in a network of massive clients (e.g., thousands), it might be impractical to maintain all the information of the graph

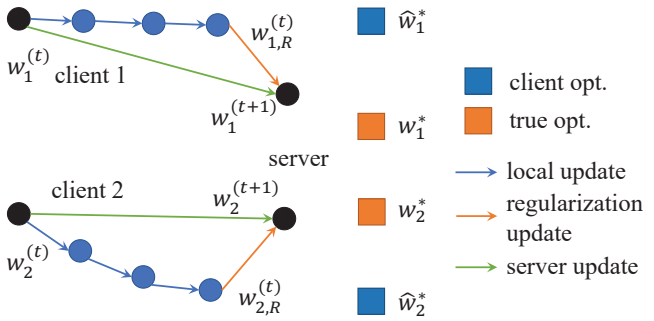


Fig. 2: The update steps of FedU at both client and server sides are illustrated for 2 related tasks (clients) with 3 local steps ($N=2, R=3$) at round t. The local updates $w_{k,r}^{(t)}$ (blue circles) move towards the client optima $\widehat{w}_k^*$ (blue square). The regularization updates (in orange) ensures the server update (in green) moves towards the true optimum $w_k^*, \forall k \in \mathcal{N}$ (orange square).

(e.g., vertices, weighted edge) as well as storage for all model updates at the server. This motivates us to propose dFedU, which is a decentralized version of FedU, and presented in Algorithm 2.

Specifically, in each communication round, each client of an entity graph (as shown in Figure 1(c)) performs R local updates, and sends its updated model to their neighboring clients to perform the model regularization. Here, each client does not need to communicate with the rest of the large number of clients in the whole network. Each client only needs to communicate with its neighbor clients. A client ℓ is a neighbor of client k if and only if it has a communication link (i.e., $a_{k\ell} \neq 0$) and share a certain model similarity with client k (i.e., $a_{k\ell} > 0$). The set of neighboring clients of client k is defined as $\widetilde{\mathcal{N}}_k = \{\ell \mid a_{k\ell} > 0\}$. Note that because there is no server for coordinating the learning, there is no client sampling in dFedU. Compared to the non-FL decentralized scheme [41], [42], dFedU uses R local updates, which are typical in FL algorithm designs.

V. Federated Multi-Task Learning: Convergence Rate

In this section, we present the convergence rate of FedU and dFedU. Let $W^* = [w_1^*, \ldots, w_N^*]$ be the optimal solution to (1).

Lemma 1. *Suppose that Assumption 1 holds and $\eta\rho > 2\beta$, where $\rho := \|\mathcal{L}\|$. Then there exists $\sigma_2 \geq 0$, e.g., $\sigma_2 = \|\nabla F(0)\|\sqrt{\frac{\eta\rho}{\eta\rho - 2\beta}}$ such that, for any $W \in \mathbb{R}^{dN}$,*

$$\sum_{k=1}^N \|\nabla F_k(w_k)\|^2 \leq \sigma_2^2 + \sum_{k=1}^N \|\nabla_{w_k} J(W)\|^2, \quad (7)$$

where $\nabla_{w_k} J(W)$ is the gradient of J with respect to w_k. Consequently, if every F_k is convex, then

$$\sum_{k=1}^N \|\nabla F_k(w_k^*)\|^2 \leq \sigma_2^2. \quad (8)$$

Proof. See Appendix B. □

Algorithm 2 dFedU–Decentralized FedU

client k's input: $\{a_{k\ell}\}$, $\widetilde{\mathcal{N}}_k$, initial $w_k^{(0)}, \forall k \in \mathcal{N}$, local step-size μ, and global step-size $\widetilde{\mu} = \mu R$
for each round $t = 0, \ldots, T-1$ **do**
  on client $k \in \mathcal{N}$ **in parallel do**
    initialize local model $w_{k,0}^{(t)} \leftarrow w_k^{(t)}$
    for $r = 0, \ldots, R-1$ **do**
      compute mini-batch gradient $\nabla \widetilde{F}_k(w_{k,r}^{(t)})$
      $w_{k,r+1}^{(t)} \leftarrow w_{k,r}^{(t)} - \mu \nabla \widetilde{F}_k(w_{k,r}^{(t)})$
    end for
    send $w_{k,R}^{(t)}$ to its neighboring clients in $\widetilde{\mathcal{N}}_k$
  end on client
  on client $k \in \mathcal{N}$ **in parallel do**
    $w_k^{(t+1)} \leftarrow w_{k,R}^{(t)} - \widetilde{\mu}\,\eta \sum_{\ell \in \widetilde{\mathcal{N}}_k} a_{k\ell}(w_{k,R}^{(t)} - w_{\ell,R}^{(t)})$
  end on client
end for

For any given value of ρ, the condition $\eta\rho > 2\beta$ in Lemma 1 can be always achieved by tuning $\eta \in \mathbb{R}$. Therefore, the impact of the relationships among the models of clients (or the graph Laplacian structure encoded by ρ) on the convergence of FedU and dFedU can be controlled by η. One can choose a large η if ρ is small and vice versa to satisfy this condition.

Note that in the conventional FL setting, i.e., $w_k = w, \forall k \in \mathcal{N}$, (7) is rewritten as

$$\frac{1}{N}\sum_{k=1}^N \|\nabla F_k(w)\|^2 \leq \frac{\sigma_2^2}{N} + \gamma^2 \|\nabla_w J(W)\|^2 \text{ with } \gamma = 1,$$

which is exactly the assumptions of $(\sigma_2/\sqrt{N}, \gamma)$-bounded gradient dissimilarity in [10], [22], and the γ-local dissimilarity in [19] with $\sigma_2 = 0$. Here, $\sigma_2 = 0$ and $\gamma = 1$ are for the i.i.d cases, while $\sigma_2 \geq 0$ and $\gamma \geq 1$ for non-IID cases.

From now on, let σ_2 and ρ be defined as in Lemma 1, and $W^{(t)} = [w_1^{(t)}, \ldots, w_N^{(t)}]$ be the collective vector generated by FedU (with client sampling) or dFedU (without client sampling, i.e., $S = N$) at round t. Node that the convergence rate of dFedU is obtained directly from the convergence rate of FedU when $S = N$. In the following theorems, we show that FedU admits linear speedup for strongly convex and sublinear speedup of order $1/2$ for nonconvex objective functions.

Theorem 1 (Convergence in strongly convex cases)**.** *Suppose that Assumptions 1, 2, and 3 hold, and $\eta > \frac{2\beta}{\rho}$. Then there exists $\mu \leq \frac{\widetilde{\mu}_1}{R}$ such that, for any $T \geq \frac{4N}{\widetilde{\mu}_1 \alpha S}$,*

$$\mathbb{E}[J(\widetilde{W}^{(T)}) - J(W^*)] \leq \widetilde{\mathcal{O}}\Bigg(\alpha\Delta^{(0)} e^{-\frac{\widetilde{\mu}_1 \alpha S T}{4N}} + \frac{\sigma_1^2}{(\alpha T)^2 RS} + \frac{\sigma_2^2}{(\alpha T)^2 S} + \frac{\sigma_1^2}{\alpha T RS} + \frac{\sigma_2^2}{\alpha T S}\Bigg), \quad (9)$$

where $\widetilde{\mu}_1 := \min\left\{\frac{1}{q}, \frac{2}{\eta\rho}\right\}$, $q = \frac{128\beta^2\eta\rho}{\alpha^2} + 12(\beta + \eta\rho) + \frac{96\beta^2}{\alpha} + \frac{32p\beta^2}{\alpha\eta\rho}$, $p = 2(\beta + \eta\rho) + \frac{8\eta^2\rho^2}{\alpha} + \frac{64\beta^2}{\alpha} + \frac{12(\beta+\eta\rho)^2}{\eta\rho} + 6\eta\rho + \frac{48\beta^2}{\eta\rho}$, $\Delta^{(0)} := \|W^{(0)} - W^*\|^2$, $\widetilde{W}^{(T)} := \sum_{t=0}^{T-1} \frac{\theta^{(t)} W^{(t)}}{\Theta_T}$, $\Theta_T = \sum_{t=0}^{T-1} \theta^{(t)}$, $\theta^{(t)} = (1 - \mu RS\alpha/(4N))^{-(t+1)}$, *and* $\widetilde{\mathcal{O}}$ *hides*

both constants and polylogarithmic factors. Consequently, the output of FedU *has expected error smaller than* ε *when*

$$T=\widetilde{\mathcal{O}}\left(\frac{1}{\alpha S}+\frac{\sigma_1}{\alpha\sqrt{\varepsilon RS}}+\frac{\sigma_2}{\alpha\sqrt{\varepsilon S}}+\frac{\sigma_1^2}{\alpha RS\varepsilon}+\frac{\sigma_2^2}{\alpha S\varepsilon}\right). \quad (10)$$

Proof. See Appendix D. □

Theorem 2 (Convergence in nonconvex cases)**.** *Suppose that Assumptions 1 and 3 hold, and* $\eta > \frac{2\beta}{\rho}$*. Then there exists* $\mu \leq \frac{\widetilde{\mu}_2}{R}$ *such that, for any* $T > 0$*,*

$$\mathbb{E}\left\|\nabla J(W^{(t^*)})\right\|^2 \leq \mathcal{O}\left(\frac{\Delta_J}{TS}+\frac{\Delta_J^{\frac{2}{3}}M^{\frac{2}{3}}}{T^{\frac{2}{3}}(RS)^{\frac{1}{3}}}+\frac{\Delta_J^{\frac{1}{2}}M^2}{\sqrt{TRS}}\right), \quad (11)$$

where $\widetilde{\mu}_2 := \min\left\{\frac{1}{v}, \frac{2}{\eta\rho}\right\}$*,* $v = 8\big(8\eta\rho+3(\beta+\eta\rho)+12(\beta+\eta\rho)+\frac{8u}{\eta\rho}\big)$*,* $u=\frac{(\beta+\eta\rho)^2}{2}+2\eta^2\rho^2+16\eta\rho\beta^2+\frac{6(\beta+\eta\rho)^3}{\eta\rho}+3\eta\rho(\beta+\eta\rho)+\frac{24(\beta+\eta\rho)\beta^2}{\eta\rho}$*;* $\Delta_J := J(W^{(0)})-J(W^*)$*,* $M^2 = R\sigma_2^2+\sigma_1^2$*, and* t^* *uniformly sampled from* $\{0,\ldots,T-1\}$*. Consequently, the output of* FedU *has expected error smaller than* ε *when*

$$T=\mathcal{O}\left(\frac{1}{S\varepsilon}+\frac{\sigma_1}{\varepsilon^{\frac{3}{2}}\sqrt{RS}}+\frac{\sigma_2}{\varepsilon^{\frac{3}{2}}\sqrt{S}}+\frac{\sigma_1^2}{\varepsilon^2 RS}+\frac{\sigma_2^2}{\varepsilon^2 S}\right). \quad (12)$$

Proof. See Appendix E. □

For illustrative purposes, we compare our rates with those of FL and personalized FL algorithms in i.i.d cases (i.e., $\sigma_2 = 0$ and $\gamma = 1$). The strongly-convex rate of FedU becomes $\frac{\sigma_1^2}{\alpha RS\varepsilon}+\frac{1}{\alpha S}$, which matches the lower-bound for the identical case [47], compared to the latest $\frac{\sigma_1^2}{\alpha RS\varepsilon}+\frac{1}{\alpha}$ by SCAFFOLD [10] and $\frac{\sigma_1^2}{\alpha RS\varepsilon}+\frac{\delta}{\alpha}$ by LSGD-PFL [45] with $\delta \geq 0$. Our rate improvement comes from the advantage of additional information about the structure of the models of clients that is captured by Laplacian regularization. Also, when no variance ($\sigma_1^2 = 0$) and no client sampling, the nonconvex rate of FedU is $\frac{\sigma_2^2}{\varepsilon^2 S}+\frac{\sigma_2}{\varepsilon^{3/2}}+\frac{1}{\varepsilon}$, which is tighter (without γ) than the rate of SCAFFOLD, and less dependent on σ_2 than that of [48].

VI. Experiments

In this section, we evaluate the performance of FedU when the data are heterogeneous and non-i.i.d in both strongly convex and nonconvex settings. We show vital show the advances of FedU with Laplacian regularization in federated multi-task and personalized settings by comparing FedU with cutting-edge learning algorithms including MOCHA, pFedMe, Per-FedAvg, FedProx [19], SCAFFOLD [10], AFL [49], and the vanilla FedAvg. The experimental results show that FedU achieves appreciable performance improvement over others in terms of test accuracy.

A. Experimental Settings

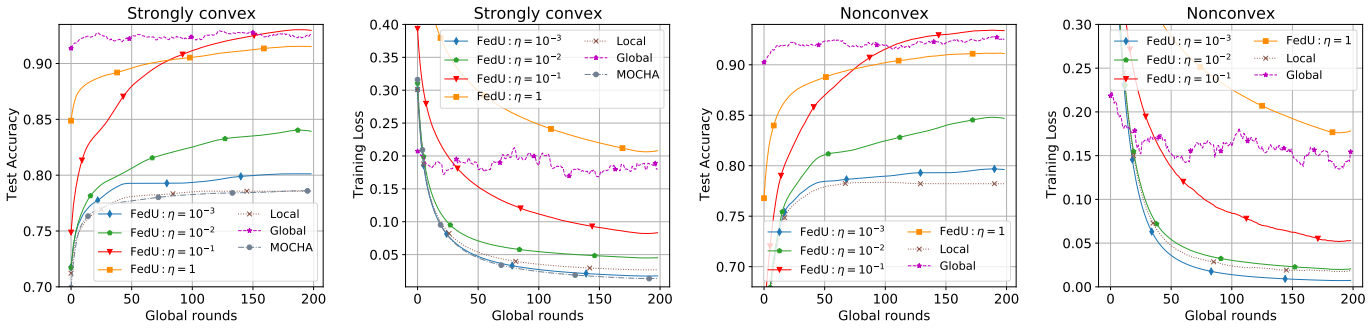
We consider classification problems using real datasets generated in federated settings, including Human Activity Recognition, Vehicle sensor, MNIST, and CIFAR-10.

- **Human Activity Recognition**: The set of data gathered from accelerometers and gyroscopes of cell phones from 30 individuals performing six different activities including lying-down, standing, walking, sitting, walking-upstairs, and walking-downstairs [50]. Each individual is considered as a task (client) classifying 6 different activities.
- **Vehicle Sensor**: Data is collected from a distributed wireless sensor network of 23 sensors including acoustic (microphone), seismic (geophone), and infrared (polarized IR sensor) [51]. It aims to classify types of moving vehicles. We consider each sensor as a separate task (client) performing the binary classification to predict 2 vehicle types: Assault Amphibian Vehicle (AAV) and Dragon Wagon (DW).
- **MNIST**: A handwritten digit dataset [52] includes 10 labels and 70,000 instances. The whole dataset is distributed to $N = 100$ clients. Each client has a different local data size and consists of 2 over 10 labels.
- **CIFAR-10**: An object recognition dataset [53] includes 60,000 colour images belonging to 10 classes. We partition the dataset to $N = 20$ clients and 3 labels per client.

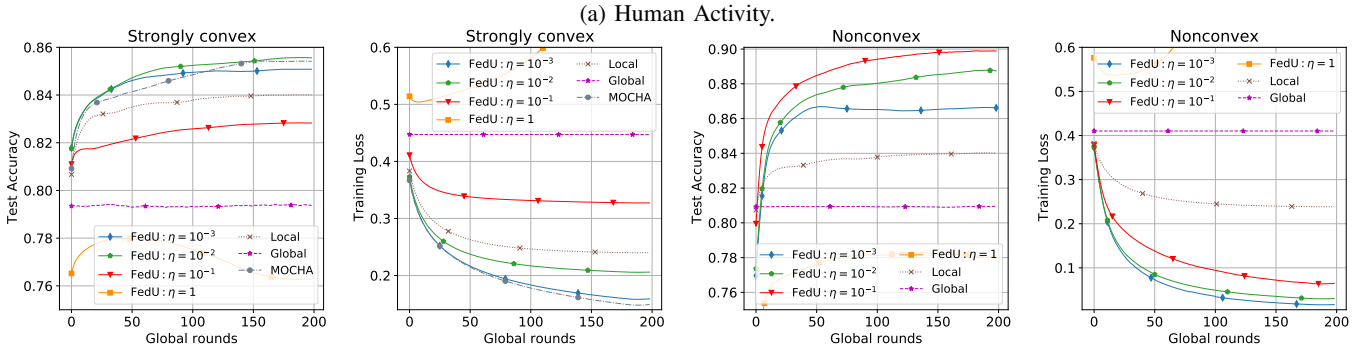
In practical FL networks, some clients have significantly limited data sizes and need collaborative learning with others. For each dataset, we hence down-sample 80% data belonged to a half of the total clients to observe behaviour of all algorithms. We provide all details about datasets and results without down-sampling in the Appendix F. All datasets are split randomly with 75% and 25% for training and testing, respectively.

We use a multinomial logistic regression model (MLR) with cross-entropy loss functions and L_2-regularization term as the strongly convex model for Human Activity Recognition, Vehicle Sensor, and MNIST. For nonconvex setting, we use a simple deep neural network (DNN) with one hidden layer, a ReLU activation function, and a softmax layer at the end of the network for Human Activity and Vehicle Sensor datasets. The size of hidden layer is 100 for Human Activity and 20 for Vehicle Sensor. In the case of MNIST, we use DNN with 2 hidden layers and both layers have the same size of 100. For CIFAR-10, we follow the CNN structure of [1].

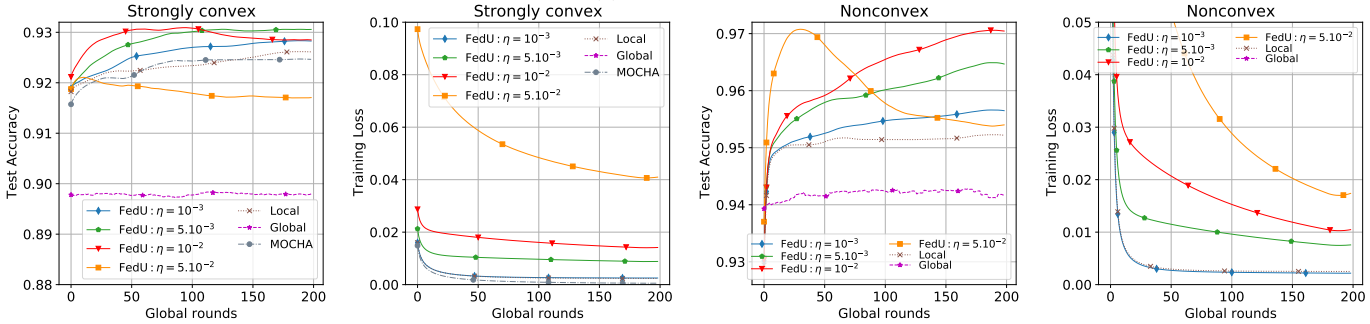
The structural dependence matrix Ω of MOCHA is chosen as $\Omega = (\mathbf{I}_{N\times N} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)^2$ following settings of [16], [24], where $\mathbf{I}_{N\times N}$ is the identity matrix with size $N \times N$ and $\mathbf{1}$ is a vector of all ones size N. Here, Ω is exactly the Laplacian matrix L in problem (1) when all the weights $a_{k\ell} = 1, \forall k, \ell$. As both FedU and dFedU have the same performance when there is no client sampling, in our experiments, we only evaluate the performance of FedU. When comparing FedU with other algorithms, we conduct 5-fold cross-validation to figure out the combination of hyperparameters allowing each algorithm to achieve the highest test accuracy. All experiments are implemented using PyTorch [54] version 1.6. We follow the implementations of [14] for pFedMe, FedAvg, and Per-FedAvg, [24] for MOCHA. All experiments are run on **NVIDIA Tesla T4** GPU. All code and data are published at [2]. The accuracy is reported with mean and standard deviation over 10 runs.



(a) Human Activity.



(b) Vehicle Sensor.



(c) MNIST.

Fig. 3: Performance comparison between MOCHA, local model, global model, and FedU with the various sets of η in both strongly convex and nonconvex settings.

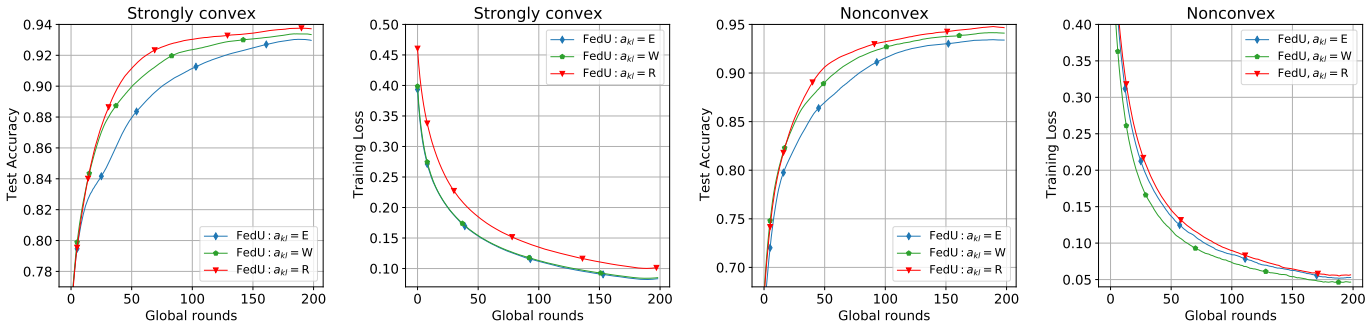
B. Performance of FedU in Federated Multi-Task Learning

We first show benefits of FedU in FMTL setting by comparing FedU with Local model (training one separate model per client), Global model (training one single model on centralized data), and MOCHA, the conventional FMTL algorithm [16]. Note that the performance results of the FMTL algorithm in [29] and MOCHA are reported similar. We evaluate FedU on a wide range of values of $\eta \in \{5.10^{-3}, 10^{-3}, 5.10^{-2}, 10^{-2}, 10^{-1}, 1\}$ and compare with others using their best fine-tuned parameters. In FMTL, each client represents a separate task. All clients have the same weight connection $\{a_{kl}\}$ with others and no client sampling in order to make fair comparisons with Local, Global models, and MOCHA. We also provide details on how to choose the different values of $\{a_{kl}\}$ in supplementary material. We only report the convex setting for MOCHA according to its assumption as stated in Section 3.1 [16].

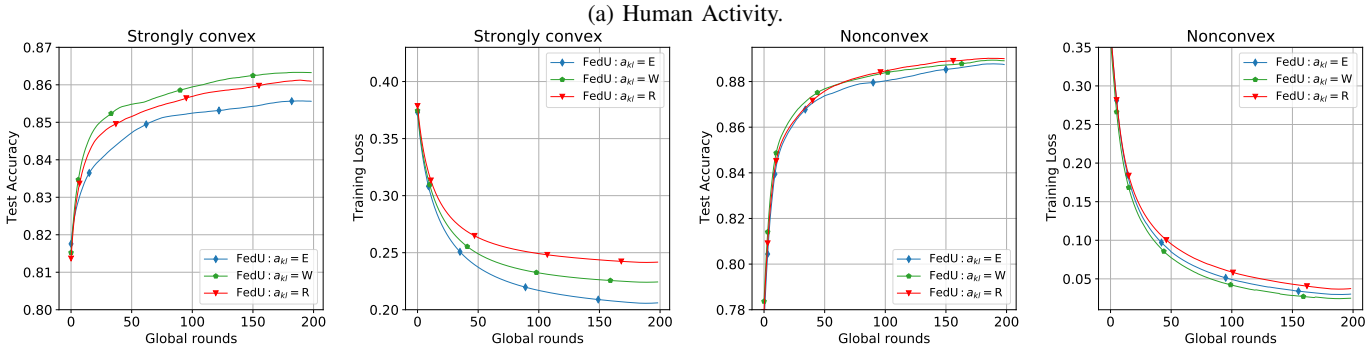
[2]https://github.com/dual-grp/FedU_FMTL

The results in Fig. 3 show that, FedU achieves the highest performance, followed by MOCHA, Local model, and Global model. While the Local model at individual client learns only its own data without any contribution from the model of other clients, the Global model only does a single task that is not well generalized on highly non-i.i.d data. We also recognize that Local model suffers overfitting when the data size at clients is small. By contrast, MOCHA and FedU have the ability to learn models for multiple related tasks simultaneously and capture relationships amongst clients. Especially in the case of FedU, using Laplacian regularization allows utilizing additional information about the structures of clients' models to increase the learning performance, and the contribution from clients having the large data size to those having smaller ones becomes more significant.

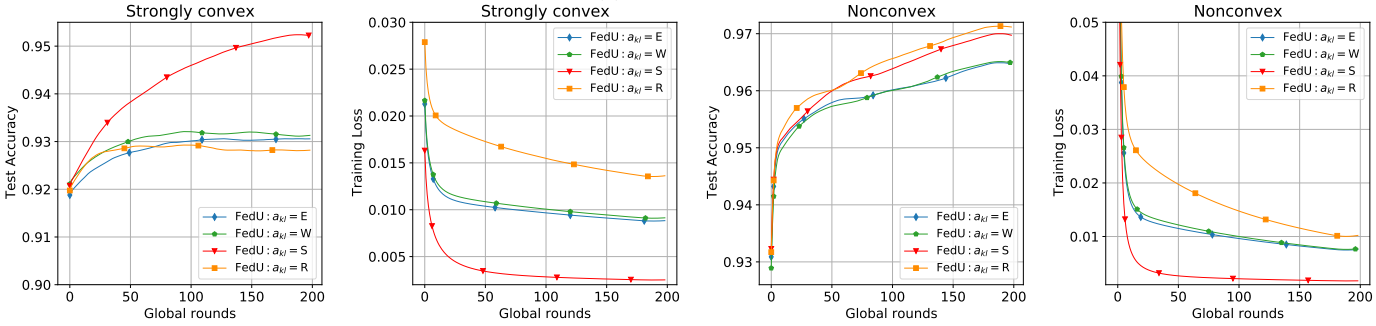
Observing different values of η, we found that the larger η is, the more the coordination from other clients are, then FedU performs better when η is increased. However, when η reaches a certain threshold, it slows down the convergence of



(a) Human Activity.



(b) Vehicle Sensor.



(c) MNIST.

Fig. 4: Effects of graph information $\{a_{kl}\}$ on the convergence of FedU in both convex and nonconvex settings.

FedU, for example, $\eta = 5.10^{-2}$ in Fig. 3. η then should be chosen carefully depends on the dataset.

C. *Effect of the Graph Information* $\{a_{kl}\}$

For the above experiments, we assume that all relationships among a client and its neighbors are equal. However, in practice, the connection weights may have different values and they need to be known in advance. We then evaluate the effect of graph information shown in Fig. 4 by normalizing the values of $\{a_{kl}\}$ in the range of $[0, 1]$ and simulate 4 different scenarios of $\{a_{kl}\}$ as below:

- Random (R): All values of $\{a_{kl}\}$ are generated radomly $\{a_{kl}\} \sim \mathcal{N}(0, 1)$.
- Equal (E): When all clients have the same value for $\{a_{kl}\}$, we can choose any value of $\{a_{kl}\}$ in the range of $[0, 1]$. However, there will be one value of $\eta * \{a_{kl}\}$ allows FedU to achieve the highest accuracy. So, whenever $\{a_{kl}\}$ is large, we can choose a small η, and vice versa. In this experiment, we fix $\{a_{kl}\} = 0.5$ and adjust η accordingly.
- Weighted (W): As there are various clients having significantly small data sizes, we set $\{a_{kl}\} = 0$ on the connection between these clients. We then set $\{a_{kl}\} = 0.5$ on the connection among clients having small data sizes and those having large data sizes, and $\{a_{kl}\} = 1$ for all other connections.
- Similar (S): This scenario is only for MNIST. When distributing data to all clients, each client has 2 labels over 10. Hence, clients may share only one, two similar labels or none of them. We set $\{a_{kl}\} = 0$, $\{a_{kl}\} = 0.5$, and $\{a_{kl}\} = 1$ for the connections among clients having no similar label, one similar label, and two similar labels, respectively.

In most of the cases, the performance of FedU with random $\{a_{kl}\}$ is better than that with equal $\{a_{kl}\}$. When $\{a_{kl}\}$ are weighted, FedU performs better than when all $\{a_{kl}\}$ are equal. Especially for MNIST, when $\{a_{kl}\}$ are weighted based on the similarity of clients, FedU achieves the highest performance compared to other scenarios. Therefore, given knowing the relationship between client's data distribution, for example, in a weather forecasts application, clients in the same geographical

TABLE I: Performance comparison of centralized setting ($R = 5$, $S = 0.1N$, $B = 20$, $T = 200$). There is no convex model for CIFAR-10, we then only report the non-convex case.

Dataset	Algorithm	Test Accuracy Convex	Non Convex
CIFAR-10	FedU		**75.41** $\pm$ 0.29
	pFedMe		74.10 $\pm$ 0.89
	Per-FedAvg		64.70 $\pm$ 1.91
	FedAvg		34.48 $\pm$ 5.34
	FedProx		42.31 $\pm$ 4.21
	SCAFFOLD		45.12 $\pm$ 3.38
	AFL		49.07 $\pm$ 3.35
MNIST	FedU	**96.95** $\pm$ 0.11	97.81 $\pm$ 0.01
	MOCHA	96.18 $\pm$ 0.09	
	pFedMe	93.73 $\pm$ 0.40	**98.64** $\pm$ 0.17
	Per-FedAvg	90.33 $\pm$ 0.84	96.38 $\pm$ 0.40
	FedAvg	87.75 $\pm$ 1.31	91.48 $\pm$ 1.05
	FedProx	88.70 $\pm$ 1.18	91.60 $\pm$ 0.23
	SCAFFOLD	89.45 $\pm$ 0.37	92.15 $\pm$ 0.43
	AFL	89.79 $\pm$ 1.23	92.01 $\pm$ 1.21
Vehicle Sensor	FedU	**88.47** $\pm$ 0.21	**91.79** $\pm$ 0.31
	MOCHA	87.31 $\pm$ 0.23	
	pFedMe	81.38 $\pm$ 0.41	90.62 $\pm$ 0.41
	Per-FedAvg	81.07 $\pm$ 0.71	86.92 $\pm$ 1.3
	FedAvg	79.84 $\pm$ 0.91	84.04 $\pm$ 2.69
	FedProx	82.06 $\pm$ 0.91	87.65 $\pm$ 2.34
	SCAFFOLD	81.97 $\pm$ 0.91	88.48 $\pm$ 0.34
	AFL	82.25 $\pm$ 0.91	87.88 $\pm$ 1.08
Human Activity	FedU	**95.76** $\pm$ 0.46	**95.86** $\pm$ 0.36
	MOCHA	92.33 $\pm$ 0.67	
	pFedMe	95.41 $\pm$ 0.38	95.72 $\pm$ 0.32
	Per-FedAvg	94.78 $\pm$ 0.37	94.80 $\pm$ 0.60
	FedAvg	93.41 $\pm$ 0.95	93.74 $\pm$ 1.01
	FedPro	93.69 $\pm$ 0.84	94.65 $\pm$ 0.72
	SCAFFOLD	93.61 $\pm$ 0.37	94.78 $\pm$ 0.85
	AFL	93.92 $\pm$ 0.34	94.42 $\pm$ 0.34

location may have similar or close weather data, we can set higher values of weight connection for those clients than clients are in different locations to takes advantages of FedU.

D. Comparison with Personalized FL algorithms

Finally, we compare FedU with the conventional FL algorithms FedAvg, FedProx, SCAFFOLD, AFL, MOCHA, and with the state-of-the-art personalized FL algorithms pFedMe and Per-FedAvg. The results are shown in Table. I. We fix the subset of clients $S = 0.1N$ and perform the comparison on all four real datasets. Overall, FedU almost maintains the top performance in all scenarios.

VII. Conclusion

This work has formulated a FMTL problem using Laplacian regularization to capture the relationships among the models of clients. The formulated problem has been proved to be used for traditional FL and personalized FL. We have also proposed both communication-centralized and decentralized algorithms to solve the formulated problem with guaranteed convergence to the optimal solution. Theoretical results show that our algorithms FedU and dFedU achieve the state-of-the art convergence rates. Experimental results with real datasets in both convex and nonconvex objectives demonstrate that the proposed algorithms outperform the conventional MOCHA in FMTL settings, the vanilla FedAvg in FL settings, and pFedMe, and Per-FedAvg in personalized FL settings.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Apr. 2017.

[2] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1, 2021.

[3] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, 2020.

[4] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Trans. Neural Netw. Learn, Syst.*, vol. 32, no. 8, pp. 3710–3722, 2021.

[5] N. Rieke *et al.*, "The future of digital health with federated learning," *NPJ Digital Medicine*, vol. 3, 2020.

[6] J. Xu and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, pp. 1 – 19, 2020.

[7] T. S. Brisimi *et al.*, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[8] J. C. Jiang, B. Kantarci, S. Oktug, and T. Soyata, "Federated learning in smart city sensing: Challenges and opportunities," *Sensors (Basel, Switzerland)*, vol. 20, 2020.

[9] L. Ahmed *et al.*, "Active learning based federated learning for waste and natural disaster image classification," *IEEE Access*, vol. 8, pp. 208 518–208 531, 2020.

[10] S. P. Karimireddy *et al.*, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proceedings of the International Conference on Machine Learning*, vol. 119, 2020.

[11] F. Haddadpour and M. Mahdavi, "On the convergence of local descent methods in federated learning," *arXiv: 1910.14425*, 2019.

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

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

[14] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized federated learning with moreau envelopes," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2020.

[15] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach," in *Advances in Neural Information Processing Systems*, 2020.

[16] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2017.

[17] A. Kumar and H. Daumé, "Learning task grouping and overlap in multi-task learning," in *Proceedings of the International Conference on Machine Learning*, 2012.

[18] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," 2010, p. 733–742.

[19] T. Li *et al.*, "Federated optimization in heterogeneous networks," in *Proceedings of the Machine Learning and Systems 2020*, 2020.

[20] Y. Zhao *et al.*, "Federated Learning with Non-IID Data," *arXiv: 1806.00582*, Jun. 2018.

[21] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," in *Proceedings of International Conference on Learning Representations*, Apr. 2020.

[22] A. Khaled, K. Mishchenko, and P. Richtarik, "Tighter theory for local sgd on identical and heterogeneous data," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 108, 26–28 Aug. 2020.

[23] F. Hanzely and P. Richtárik, "Federated Learning of a Mixture of Global and Local Models," *arXiv:2002.05516*, Feb. 2020.

[24] P. P. Liang *et al.*, "Think Locally, Act Globally: Federated Learning with Local and Global Representations," *arXiv: 2001.01523*, Jun. 2020.

[25] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the International Conference on Machine Learning*, 2017.

[26] Y. Jiang, J. Konečný, K. Rush, and S. Kannan, "Improving Federated Learning Personalization via Model Agnostic Meta Learning," *arXiv: 1909.12488*, Sep. 2019.

[27] A. Nichol, J. Achiam, and J. Schulman, "On First-Order Meta-Learning Algorithms," *arXiv: 1803.02999*, Oct. 2018.

[28] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated Learning with Personalization Layers," *arXiv: 1912.00818*, Dec. 2019.

[29] Y. Sarcheshmehpour, Y. Tian, L. Zhang, and A. Jung, "Networked federated multi-task learning," *arXiv: 2105.12769*, 2021.

[30] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and Robust Federated Learning Through Personalization," in *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021.

[31] J. Shen, X. Zhen, M. Worring, and L. Shao, "Variational Multi-Task Learning with Gumbel-Softmax Priors," in *Proceedings of Advances in Neural Information Processing Systems*, 2021.

[32] R. Li, F. Ma, W. Jiang, and J. Gao, "Online federated multitask learning," in *IEEE International Conference on Big Data*, 2019.

[33] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, p. 243–272, Dec. 2008.

[34] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, p. 1817–1853, Dec. 2005.

[35] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, p. 41–75, Jul. 1997.

[36] A. Jung and Y. SarcheshmehPour, "Local graph clustering with network lasso," *IEEE Signal Processing Letters*, vol. 28, pp. 106–110, 2021.

[37] A. Jung and N. Tran, "Localized linear regression in networked data," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1090–1094, 2019.

[38] D. Hallac, J. Leskovec, and S. Boyd, "Network lasso: Clustering and optimization in large graphs," in *Proceedings of the 21th ACM International Conference on Knowledge Discovery and Data Mining*, 2015.

[39] F. Hanzely, S. Hanzely, S. Horváth, and P. Richtarik, "Lower Bounds and Optimal Algorithms for Personalized Federated Learning," in *Proceedings of Advances in Neural Information Processing Systems*, 2020.

[40] P. Vanhaesebrouck, A. Bellet, and M. Tommasi, "Decentralized Collaborative Learning of Personalized Models over Networks," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Apr 2017, pp. 509–517.

[41] R. Nassif, S. Vlaski, C. Richard, and A. H. Sayed, "Learning over multitask graphs—part i: Stability analysis," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 28–45, 2020.

[42] ——, "Learning over multitask graphs—part II: Performance analysis," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 46–63, 2020.

[43] J. Tuck, S. Barratt, and S. Boyd, "A distributed method for fitting laplacian regularized stratified models," *Journal of Machine Learning Research*, 2021.

[44] J. Tuck and S. Boyd, "Eigen-stratified models," *Optimization and Engineering*, 2021.

[45] F. Hanzely, B. Zhao, and M. Kolar, "Personalized federated learning: A unified framework and universal optimization techniques," in *Proceedings of International Conference on Learning Representations*, 2021.

[46] Y. Nesterov, Ed., *Lectures on Convex Optimization*. Springer International Publishing, 2018, vol. 137.

[47] B. E. Woodworth, J. Wang, A. Smith, B. McMahan, and N. Srebro, "Graph oracle models, lower bounds, and gaps for parallel stochastic optimization," in *Proceedings of the International Conference on Neural Information Processing Systems*, vol. 31, 2018.

[48] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning," vol. 33, no. 01, Jul. 2019.

[49] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic Federated Learning," *arXiv:1902.00146*, Jan. 2019.

[50] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition Using Smartphones," *Computational Intelligence*, p. 6, 2013.

[51] M. F. Duarte and Y. Hen Hu, "Vehicle classification in distributed sensor networks," *Journal of Parallel and Distributed Computing*, vol. 64, no. 7, pp. 826–838, Jul. 2004.

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

[53] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," p. 60, 2009.

[54] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, Vancouver, BC, Canada, 2019.

[55] Y. Arjevani, O. Shamir, and N. Srebro, "A tight convergence analysis for stochastic gradient descent with delayed updates," in *Proceedings of the International Conference on Algorithmic Learning Theory*, vol. 117, Feb. 2020.

[56] S. Stich, "Unified optimal analysis of the (stochastic) gradient method," *arXiv: 1907.04232*, 2019.

[57] A. Kulunchakov and J. Mairal, "Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise," *Journal of Machine Learning Research*, vol. 21, pp. 155:1–155:52, 2020.

[58] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification," *arXiv:1909.06335*, Sep. 2019.

[59] S. J. Reddi *et al.*, "ADAPTIVE FEDERATED OPTIMIZATION," in *International Conference on Learning Representations*, 2021.

[60] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," in *Advances in Neural Information Processing Systems*, 2020.

APPENDIX

A. Technicalities

In this section, we introduce additional definitions and technical lemmas which will be useful for our analysis of FedU.

$$\nabla F(W) := [\nabla_{w_1} F(W)^T, \ldots, \nabla_{w_N} F(W)^T]^T = [\nabla F_1(w_1)^T, \ldots, \nabla F_N(w_N)^T]^T \in \mathbb{R}^{dN} \text{ is the gradient of } F(W) \tag{11}$$

$$\nabla J(W) := [\nabla_{w_1} J(W)^T, \ldots, \nabla_{w_N} J(W)^T]^T \overset{(1)}{=} \nabla F(W) + \eta\, \mathcal{L}\, W \in \mathbb{R}^{dN} \text{ is the gradient of } J(W) \tag{12}$$

$$\zeta = \{\zeta_1, \ldots, \zeta_N\} \text{ is the set of random samples of clients} \tag{13}$$

$$\nabla \widetilde{F}(W, \zeta) := [\nabla \widetilde{F}_1(w_1, \zeta_1)^T, \ldots, \nabla \widetilde{F}_N(w_N, \zeta_N)^T]^T \in \mathbb{R}^{dN} \text{ is the stochastic gradient of } F(W) \tag{14}$$

$$\nabla \widetilde{J}(W, \zeta) := \nabla \widetilde{F}(W, \zeta) + \eta\, \mathcal{L}\, W^{(t)} \in \mathbb{R}^{dN} \text{ is the stochastic gradient of } J(W) \tag{15}$$

$$\widehat{S}^{(t)} = [s_1^{(t)}, \ldots, s_N^{(t)}] \in \mathbb{R}^N \text{ is a client sampling random vector at round } t, \text{ where } s_k^{(t)} = \begin{cases} 1, & \text{if } k \in \mathcal{S}^{(t)} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

$$\widetilde{S}^{(t)} = \operatorname{diag}(\widehat{S}^{(t)}) \otimes I_d \in \mathbb{R}^{dN \times dN} \text{ is a sampling matrix} \tag{17}$$

$$\mathcal{L} = L \otimes I_d \in \mathbb{R}^{dN \times dN} \tag{18}$$

$$C = I - \widetilde{\mu}\, \eta \widetilde{S}\, \mathcal{L} \in \mathbb{R}^{dN \times dN} \text{ is a server-update matrix} \tag{19}$$

$$\tau = \frac{S}{N} \text{ is a client sampling factor.} \tag{20}$$

In what follows, $\|\cdot\|$ represents the 2-norm for matrix and the Euclidean norm for vector.

For a connected graph $\mathcal{G}$, $L = D - A$ is a symmetric positive semi-definite matrix with $\lambda_{\min}(L) = \lambda_1 = 0 < \lambda_2 \leq \cdots \leq \lambda_N = \lambda_{\max}(L) = \rho$, in order. As such, the matrix $\mathcal{L} = L \otimes I_d$ has $\lambda_{\min}(\mathcal{L}) = \lambda_{\min}(L)\lambda_{\min}(I_d) = \lambda_{\min}(L) = 0$ and $\lambda_{\max}(\mathcal{L}) = \lambda_{\max}(L)\lambda_{\max}(I_d) = \rho$. When no client sampling, $\widetilde{S} = I_{dN}$, the matrix $C = I - \widetilde{\mu}\, \eta\, \mathcal{L}$ has $\lambda_{\min}(C) = 1 - \widetilde{\mu}\, \eta \lambda_{\max}(\mathcal{L}) = 1 - \widetilde{\mu}\, \eta\rho$ and $\lambda_{\max}(C) = 1 - \widetilde{\mu}\, \eta \lambda_{\min}(\mathcal{L}) = 1$. Since C is symmetric, we have $\|C\| = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } C\}$. Therefore, C (when no client sampling) is normalized (i.e., $\|C\|^2 = 1$) if and only if

$$\widetilde{\mu}\, \eta\rho \leq 2. \tag{21}$$

Lemma 2 (Sampling matrix's properties)**.** *Let $\widetilde{S}$ be defined as $\widetilde{S}^{(t)}$ in* (17)*. Then*

(a) $\|\widetilde{S}\| = 1$;
(b) $\widetilde{S}^T = \widetilde{S}$;
(c) $\widetilde{S}\widetilde{S} = \widetilde{S}^T\widetilde{S} = \widetilde{S}$;
(d) $\mathbb{E}\, \widetilde{S} = \tau I_{dN}$;
(e) $\mathbb{E}\, \|\widetilde{S}Y\|^2 = \tau\, \mathbb{E}\, \|Y\|^2, \forall Y \in \mathbb{R}^{dN}$

Proof. (a)–(d) follow directly from the definition of $\widetilde{S}$, while (e) from the fact that $\mathbb{E}\, \|\widetilde{S}Y\|^2 = \mathbb{E}\, Y^T \widetilde{S}^T \widetilde{S} Y \overset{(c)}{=} \mathbb{E}\, Y^T \widetilde{S} Y \overset{(d)}{=} \tau\, \mathbb{E}\, Y^T Y = \tau\, \mathbb{E}\, \|Y\|^2$. □

Lemma 3 (Jensen's inequality)**.** *For any vector $X_i \in \mathbb{R}^{dN}, i \in \{1, \ldots, M\}$,*

$$\left\| \sum_{i=1}^{M} X_i \right\|^2 \leq M \sum_{i=1}^{M} \|X_i\|^2. \tag{22}$$

Lemma 4 (Young inequality)**.** *For any vector $X, Y \in \mathbb{R}^{dN}$ and $m > 0$,*

(a) $\langle X, Y \rangle \leq \frac{m}{2}\|X\|^2 + \frac{1}{2m}\|Y\|^2$;
(b) $\|X + Y\|^2 \leq (1 + m)\|X\|^2 + \left(1 + \frac{1}{m}\right)\|Y\|^2$.

Lemma 5 (Smoothness)**.** *Suppose that Assumption 1 holds. Set $\beta_J := \beta + \eta\rho$ with $\rho := \|\mathcal{L}\|$. Then, for any $W, W' \in \mathbb{R}^{dN}$,*

(a) $\|\nabla F(W) - \nabla F(W')\| \leq \beta \|W - W'\|$;
(b) $\|\nabla J(W) - \nabla J(W')\| \leq \beta_J \|W - W'\|$;
(c) $\|\nabla F(W)\|^2 \leq 2\beta^2 \|W - W'\|^2 + 2\|\nabla F(W')\|^2$;
(d) $\|\nabla J(W)\|^2 \leq 2\beta_J^2 \|W - W'\|^2 + 2\|\nabla J(W')\|^2$;
(e) $J(W) - J(W') \leq \langle \nabla J(W'), W - W' \rangle + \frac{\beta_J}{2}\|W - W'\|^2$.

Proof. (a): This directly follows from Assumption 1 and the definition of $\nabla F(W)$ in (11).

(b): Since $\nabla J(W) = \nabla F(W) + \eta\,\mathcal{L}\,W$ and $\rho = \|\,\mathcal{L}\,\|$, the conclusion follows from (a).

(c): Using Lemma 3, we have

$$\|\nabla F(W)\|^2 = \|\nabla F(W) - \nabla F(W') + \nabla F(W')\|^2 \le 2\|\nabla F(W) - \nabla F(W')\|^2 + 2\|\nabla F(W')\|^2, \tag{23}$$

which together with (a) implies (c).

(d): The proof is similar to (c).

(e): This follows from Lemma 1.2.3 in [46]. □

Lemma 6 (Strong convexity)**.** *Suppose that Assumption 2 holds. Then, for any $W, W' \in \mathbb{R}^{dN}$,*

$$J(W) \ge J(W') + \langle \nabla J(W'), W - W' \rangle + \frac{\alpha}{2}\|W - W'\|^2. \tag{24}$$

Proof. It follows from Assumption 2 that F is α-strongly convex with respect to W. Since $J(W) = F(W) + \eta W^T\,\mathcal{L}\,W$ and $\mathcal{L}$ is a positive semi-definite matrix, we derive that J is also α-strongly convex with respect to W, and the conclusion follows. □

Lemma 7 (Smoothness and strong convexity)**.** *Suppose that Assumptions 1 and 2 hold. Let W^* be the optimal solution to* (1)*. Then, for any $W, W', W'' \in \mathbb{R}^{dN}$,*

(a) $\|\nabla J(W)\|^2 \le 2\beta_J[J(W) - J(W^*)]$*;*

(b) $\langle \nabla J(W), W'' - W' \rangle \ge J(W'') - J(W') + \frac{\alpha}{4}\|W' - W''\|^2 - \beta_J\|W'' - W\|^2$*;*

(c) $\|\nabla F(W)\|^2 \le 4\frac{\beta^2}{\alpha}[J(W) - J(W^*)] + 2\sigma_2^2$*, where σ_2 is defined as in Lemma 1.*

Proof. (a) is from Lemmas 5(b), 6, and Theorems 2.1.5 in [46], while (b) is from Lemma 5 in [10].

(c): Applying Lemma 5(c), we have

$$\|\nabla F(W)\|^2 \le 2\beta^2\|W - W^*\|^2 + 2\|\nabla F(W^*)\|^2. \tag{25}$$

It follows from Theorem 2.1.8 in [46] that

$$\|W - W^*\|^2 \le \frac{2}{\alpha}[J(W) - J(W^*)]. \tag{26}$$

By the definition of σ_2 in Lemma 1, $\|\nabla F(W^*)\|^2 \le \sigma_2^2$, which, together with (25) and (26) completes the proof. □

Lemma 8 (Bounded variance)**.** *Suppose that Assumption 3 holds. Then, for any $W \in \mathbb{R}^{dN}$,*

(a) $\mathbb{E}_\zeta\,\|\nabla\widetilde{F}(W,\zeta) - \nabla F(W)\|^2 \le \sigma_1^2$*;*

(b) $\mathbb{E}_\zeta\,\|\nabla\widetilde{J}(W,\zeta) - \nabla J(W)\|^2 \le \sigma_1^2$.

Proof. (a) directly follows from Assumption 3. By the definitions of ∇J and $\nabla\widetilde{J}$ in (12) and (15), (a) implies (b). □

Lemma 9. *Let $\{\widetilde{\mathrm{X}}_1, \ldots, \widetilde{\mathrm{X}}_r, \ldots, \widetilde{\mathrm{X}}_R\}$ be R random variables in $\mathbb{R}^{dN}$ which are not necessarily independent. Suppose each $\widetilde{\mathrm{X}}_r$ has a conditional mean $\mathbb{E}[\widetilde{\mathrm{X}}_r \,|\, \widetilde{\mathrm{X}}_{r-1}, \ldots, \widetilde{\mathrm{X}}_1] = X_r$ (i.e., $\{\widetilde{\mathrm{X}}_r - X_r\}$ form a martingale difference sequence), and a variance $\mathbb{E}\,\|\,\widetilde{\mathrm{X}}_r - X_r\|^2 \le \sigma^2$. Then*

$$\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{S}\,\widetilde{\mathrm{X}}_r\right\|^2 \le \frac{\tau}{R}\sum_{r=0}^{R-1}\mathbb{E}\,\|X_r\|^2 + \frac{\tau\sigma^2}{R}. \tag{27}$$

where $\widetilde{S}$ is defined as $\widetilde{S}^{(t)}$ in (17)*.*

Proof. We see that

$$\mathbb{E}\left\|\sum_{r=0}^{R-1}(\widetilde{\mathrm{X}}_r - X_r)\right\|^2 = \sum_{r=0}^{R-1}\mathbb{E}\,\|\,\widetilde{\mathrm{X}}_r - X_r\|^2 + \sum_{r,i}\mathbb{E}(\widetilde{\mathrm{X}}_r - X_r)^T(\widetilde{\mathrm{X}}_i - X_i) = \sum_{r=0}^{R-1}\mathbb{E}\,\|\,\widetilde{\mathrm{X}}_r - X_r\|^2, \tag{28}$$

where $\sum_{r,i}\mathbb{E}(\widetilde{\mathrm{X}}_r - X_r)^T(\widetilde{\mathrm{X}}_i - X_i) = 0$ because $\{\widetilde{\mathrm{X}}_r - X_r\}$ form a martingale difference sequence. On the other hand,

$$\mathbb{E}\left\|\sum_{r=0}^{R-1}(\widetilde{\mathrm{X}}_r - X_r)\right\|^2 = \mathbb{E}\left\|\sum_{r=0}^{R-1}\widetilde{\mathrm{X}}_r\right\|^2 - \left\|\sum_{r=0}^{R-1}X_r\right\|^2, \tag{29}$$

which, together with (28), implies that

$$\mathbb{E}\left\|\sum_{r=0}^{R-1}\widetilde{\mathrm{X}}_r\right\|^2=\left\|\sum_{r=0}^{R-1}X_r\right\|^2+\sum_{r=0}^{R-1}\mathbb{E}\,\|\,\widetilde{\mathrm{X}}_r-X_r\|^2. \tag{30}$$

Multiplying both sides of (30) by $\frac{1}{R^2}$ and using $\mathbb{E}\,\|\,\widetilde{\mathrm{X}}_r-X_r\|^2\le\sigma^2,\forall r$, we get

$$\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{\mathrm{X}}_r\right\|^2\le\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}X_r\right\|^2+\frac{\sigma^2}{R}. \tag{31}$$

This together with Lemma 2(d) yields

$$\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{S}\,\widetilde{\mathrm{X}}_r\right\|^2\le\tau\,\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{\mathrm{X}}_r\right\|^2\le\tau\,\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}X_r\right\|^2+\frac{\tau\sigma^2}{R}\le\frac{\tau}{R}\sum_{r=0}^{R-1}\mathbb{E}\,\|X_r\|^2+\frac{\tau\sigma^2}{R}, \tag{32}$$

which completes the proof. □

Lemma 10. *Let $\widetilde{\mathrm{X}}$ is a random variable in $\mathbb{R}^{dN}$ with mean $\mathbb{E}\,\widetilde{\mathrm{X}}=X$ and variance $\mathbb{E}\,\|\,\widetilde{\mathrm{X}}-X\|^2\le\sigma^2$. Let $\widetilde{S}$ be defined as $\widetilde{S}^{(t)}$ in* (17)*. Then, for any $\mu\ge0$ and $Y\in\mathbb{R}^{dN}$,*

$$\mathbb{E}\,\|Y-\mu\widetilde{S}\,\widetilde{\mathrm{X}}\,\|^2\le(1-\tau)\,\mathbb{E}\,\|Y\|^2+\tau\,\mathbb{E}\,\|Y-\mu X\|^2+\mu^2\tau\sigma^2. \tag{33}$$

Proof. On one hand, by Lemma 2(c),

$$\begin{aligned}\mathbb{E}\,\|Y-\mu\widetilde{S}\,\widetilde{\mathrm{X}}\,\|^2&=\mathbb{E}(Y-\mu\widetilde{S}\,\widetilde{\mathrm{X}})^T(Y-\mu\widetilde{S}\,\widetilde{\mathrm{X}})\\&=\mathbb{E}(Y^TY-\mu\,\widetilde{\mathrm{X}}^T\,\widetilde{S}Y-\mu Y^T\widetilde{S}\,\widetilde{\mathrm{X}}+\mu^2\,\widetilde{\mathrm{X}}^T\,\widetilde{S}\,\widetilde{\mathrm{X}})\\&=\mathbb{E}(Y^TY-\mu\tau X^TY-\mu\tau Y^TX+\mu^2\tau\,\widetilde{\mathrm{X}}^T\,\widetilde{\mathrm{X}})\end{aligned} \tag{34}$$

On the other hand, $\mathbb{E}\,\widetilde{\mathrm{X}}^T\,\widetilde{\mathrm{X}}\le X^TX+\sigma^2$ since $\mathbb{E}\,\|\,\widetilde{\mathrm{X}}-X\|^2=\mathbb{E}\,\widetilde{\mathrm{X}}^T\,\widetilde{\mathrm{X}}-X^TX\le\sigma^2$. Therefore,

$$\begin{aligned}\mathbb{E}\,\|Y-\mu\widetilde{S}\,\widetilde{\mathrm{X}}\,\|^2&\le\mathbb{E}(Y^TY-\mu\tau X^TY-\mu\tau Y^TX+\mu^2\tau X^TX+\mu^2\tau\sigma^2)\\&=(1-\tau)\,\mathbb{E}(Y^TY)+\tau\,\mathbb{E}(Y^TY-\mu X^TY-\mu Y^T\,\widetilde{\mathrm{X}}+\mu^2X^TX)+\mu^2\tau\sigma^2\\&=(1-\tau)\,\mathbb{E}\,\|Y\|^2+\tau\,\mathbb{E}\,\|Y-\mu X\|^2+\mu^2\tau\sigma^2,\end{aligned} \tag{35}$$

which finishes the proof. □

B. Proof of Lemma 1

It follows from the definition of J and W that (7) can be written as

$$\|\nabla F(W)\|^2\le\sigma_2^2+\|\nabla J(W)\|^2=\sigma_2^2+\|\nabla F(W)+\eta\,\mathcal{L}\,W\|^2, \tag{36}$$

which is equivalent to

$$\mathcal{A}:=-\eta^2\rho^2\|W\|^2-2\eta\langle\nabla F(W),\mathcal{L}\,W\rangle\le\sigma_2^2. \tag{37}$$

By Assumption 1 and the definition of $\nabla F(W)$,

$$\|\nabla F(W)-\nabla F(0)\|\le\beta\|W-0\|=\beta\|W\|, \tag{38}$$

which implies that

$$\|\nabla F(W)\|\le\beta\|W\|+\|\nabla F(0)\|. \tag{39}$$

Combining with Cauchy–Schwartz inequality, we obtain that

$$|\langle\nabla F(W),\mathcal{L}\,W\rangle|\le\|\nabla F(W)\|\|\,\mathcal{L}\,W\|\le\beta\rho\|W\|^2+\psi\rho\|W\|, \tag{40}$$

where $\psi:=\|\nabla F(0)\|$. It follows that

$$-\langle\nabla F(W),\mathcal{L}\,W\rangle\le\beta\rho\|W\|^2+\psi\rho\|W\|, \tag{41}$$

and so

$$\begin{aligned}\mathcal{A} &\leq -\eta^2\rho^2\|W\|^2 + 2\eta(\beta\rho\|W\|^2 + \psi\rho\|W\|) = -\eta\rho(\eta\rho - 2\beta)\|W\|^2 + 2\psi\eta\rho\|W\| \\ &= -\eta\rho(\eta\rho - 2\beta)\left(\|W\|^2 - \frac{\psi}{\eta\rho - 2\beta}\right)^2 + \frac{\psi^2\eta\rho}{\eta\rho - 2\beta} \\ &\leq \frac{\psi^2\eta\rho}{\eta\rho - 2\beta}, \end{aligned} \tag{42}$$

where the last inequality is due to the assumption that $\eta\rho > 2\beta$. Therefore, (37) always holds if $\sigma_2^2 \geq \frac{\psi^2\eta\rho}{\eta\rho - 2\beta}$.

C. Analysis of FedU

For ease of analysis, we rewrite Algorithm 1 as Algorithm 3 with matrix notations. Here, Line 5 of Algorithm 3 represents Lines 9 and 14 of Algorithm 1, while Line 7 of Algorithm 3 represents Lines 14 and 15 of Algorithm 1.

Algorithm 3 FedU with Matrix Notation

1: **server's input**: initial $W^{(0)}$
2: **for** each round $t = 0, \ldots, T-1$ **do**
3: **for** $r = 0, \ldots, R-1$ **do**
4: initialize $W_0^{(t)} \leftarrow W^{(t)}$
5: $W_{r+1}^{(t)} = W_r^{(t)} - \mu\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)})$
6: **end for**
7: $W^{(t+1)} = (I - \widetilde{\mu}\,\eta\widetilde{S}^{(t)}\,\mathcal{L})W_R^{(t)} = C^{(t)}W_R^{(t)}$
8: **end for**

In round t, the local update

$$W_{r+1}^{(t)} = W_r^{(t)} - \mu\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)}) \tag{43}$$

implies that after R local update steps, we have

$$\mu\widetilde{S}^{(t)}\sum_{r=0}^{R-1}\nabla\widetilde{F}(W_r^{(t)}) = \sum_{r=0}^{R-1}(W_r^{(t)} - W_{r+1}^{(t)}) = W_0^{(t)} - W_R^{(t)} = W^{(t)} - W_R^{(t)}. \tag{44}$$

We then rewrite the server update as follows

$$\begin{aligned} W^{(t+1)} &= C^{(t)}W_R^{(t)} \overset{(44)}{=} C^{(t)}\left[W^{(t)} - \mu R\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)})\right] \\ &\overset{(19)}{=} (I - \widetilde{\mu}\,\eta\widetilde{S}^{(t)}\,\mathcal{L})\left[W^{(t)} - \frac{\widetilde{\mu}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)})\right] \\ &= W^{(t)} - \frac{\widetilde{\mu}\,\widetilde{S}^{(t)}}{R}\sum_{r=0}^{R-1}\nabla\widetilde{F}(W_r^{(t)}) - \widetilde{\mu}\,\eta\widetilde{S}^{(t)}\,\mathcal{L}\,W^{(t)} + \frac{\widetilde{\mu}^2\,\eta\widetilde{S}^{(t)}\,\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)}) \\ &= W^{(t)} - \frac{\widetilde{\mu}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\left[\nabla\widetilde{F}(W_r^{(t)}) + \eta\,\mathcal{L}\,W_r^{(t)}\right] + \frac{\widetilde{\mu}\,\eta}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\,\mathcal{L}(W_r^{(t)} - W^{(t)}) + \frac{\widetilde{\mu}^2\,\eta\widetilde{S}^{(t)}\,\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)}) \\ &= W^{(t)} - \widetilde{\mu}\,Z^{(t)}, \end{aligned} \tag{45}$$

where

$$Z^{(t)} = \frac{1}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{J}(W_r^{(t)}) - \frac{\eta}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\,\mathcal{L}(W_r^{(t)} - W^{(t)}) - \frac{\widetilde{\mu}\,\eta\widetilde{S}^{(t)}\,\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)}). \tag{46}$$

Finally, we output $\widetilde{W}^{(T)} = W^{(t)}$ with probability $\frac{\theta^{(t)}}{\sum_{t=0}^{T-1}\theta^{(t)}}$ for some weights $\theta^{(t)}$, and $r \in \{0, \ldots, T-1\}$.

Let $\mathcal{E}^{(t)} := \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\,\|W_r^{(t)} - W^{(t)}\|^2$ be the drift caused by R local update steps at clients, where $\mathbb{E}$ is the expectation taken over all random sources. We now provide some supporting lemmas as follows.

Lemma 11 (Bounded drift)**.** *Suppose that Assumption 3 holds. Then*

$$\mathcal{E}^{(t)} \leq 4\,\widetilde{\mu}^2\,\tau\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \frac{2\,\widetilde{\mu}^2\,\tau\sigma_1^2}{R}. \tag{47}$$

Proof. By Assumption 3, using Lemmas 8(a), 10 and then 4(b), we derive that

$$\begin{aligned}
&\mathbb{E}\,\|W_r^{(t)} - W^{(t)}\|^2 = \mathbb{E}\,\|W_{r-1}^{(t)} - W^{(t)} - \mu\widetilde{S}^{(t)}\nabla\widetilde{F}(W_{r-1}^{(t)})\|^2 \\
&\le (1-\tau)\,\mathbb{E}\,\|W_{r-1}^{(t)} - W^{(t)}\|^2 + \tau\,\mathbb{E}\,\|W_{r-1}^{(t)} - W^{(t)} - \mu\nabla F(W_{r-1}^{(t)})\|^2 + \mu^2\tau\sigma_1^2 \\
&\le (1-\tau)\,\mathbb{E}\,\|W_{r-1}^{(t)} - W^{(t)}\|^2 + \left(1+\frac{1}{R\tau}\right)\tau\,\mathbb{E}\,\|W_{r-1}^{(t)} - W^{(t)}\|^2 + (1+R\tau)\,\mu^2\tau\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \mu^2\tau\sigma_1^2 \\
&\le \left(1+\frac{1}{R}\right)\mathbb{E}\,\|W_{r-1}^{(t)} - W^{(t)}\|^2 + \frac{2\,\widetilde{\mu}^2\,\tau}{R}\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \frac{\widetilde{\mu}^2\,\tau\sigma_1^2}{R^2},
\end{aligned} \tag{48}$$

where the last inequality is due to the fact that $1 + R\tau \le R + R = 2R$ since $R \ge 1$ and $\tau \le 1$. Telescoping (48) yields

$$\mathbb{E}\,\|W_r^{(t)} - W^{(t)}\|^2 \le \left(\frac{2\,\widetilde{\mu}^2\,\tau}{R}\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \frac{\widetilde{\mu}^2\,\tau\sigma_1^2}{R^2}\right)\sum_{r=1}^{R-1}\left(1+\frac{1}{R}\right)^r. \tag{49}$$

Since $\sum_{j=0}^{m-1} x_j = \frac{x^m-1}{x-1}$ and $\left(1+\frac{x}{n}\right)^n \le e^x, \forall x \in \mathbb{R}, n \in \mathbb{N}$, we have $\sum_{r=0}^{R-1}\left(1+\frac{1}{R}\right)^r = \frac{\left(1+\frac{1}{R}\right)^R - 1}{\left(1+\frac{1}{R}\right)-1} \le (e-1)R \le 2R$, and thus

$$\mathbb{E}\,\|W_r^{(t)} - W^{(t)}\|^2 \le 4\,\widetilde{\mu}^2\,\tau\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \frac{2\,\widetilde{\mu}^2\,\tau\sigma_1^2}{R}. \tag{50}$$

Averaging (50) over r, we get the conclusion. □

Lemma 12. *Suppose that Assumptions 1 and 3 hold. Then*

$$\mathbb{E}\,\|Z^{(t)}\|^2 \le \tau(6\beta_J^2 + 3\eta^2\rho^2 + 6\,\widetilde{\mu}^2\,\eta^2\rho^2\beta^2)\,\mathcal{E}^{(t)} + 6\tau\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + 6\tau\,\widetilde{\mu}^2\,\eta^2\rho^2\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \frac{3\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\sigma_1^2}{R}. \tag{51}$$

Proof. Using Lemma 3, we have that

$$\begin{aligned}
\mathbb{E}\,\|Z^{(t)}\|^2 \le& 3\,\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{J}(W_r^{(t)})\right\|^2 + 3\,\mathbb{E}\left\|\frac{\eta\widetilde{S}^{(t)}}{R}\sum_{r=0}^{R-1}\mathcal{L}(W_r^{(t)} - W^{(t)})\right\|^2 \\
&+ 3\,\mathbb{E}\left\|\frac{\widetilde{\mu}\,\eta\widetilde{S}^{(t)}\,\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)})\right\|^2.
\end{aligned} \tag{52}$$

Next, by Lemma 9, Lemma 5(d), and the definition of $\mathcal{E}^{(t)}$,

$$\begin{aligned}
3\,\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{J}(W_r^{(t)})\right\|^2 &\le \frac{3\tau}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla J(W_r^{(t)})\right\|^2 + \frac{3\tau\sigma_1^2}{R} \\
&\le \frac{3\tau}{R}\sum_{r=0}^{R-1}\left(2\beta_J^2\,\mathbb{E}\left\|W_r^{(t)} - W^{(t)}\right\|^2 + 2\,\mathbb{E}\left\|\nabla J(W^{(t)})\right\|^2\right) + \frac{3\tau\sigma_1^2}{R} \\
&= 6\tau\beta_J^2\,\mathcal{E}^{(t)} + 6\tau\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \frac{3\tau\sigma_1^2}{R}.
\end{aligned} \tag{53}$$

It follows from Lemma 2(e) and then Lemma 3 that

$$\begin{aligned}
3\,\mathbb{E}\left\|\frac{\eta\widetilde{S}^{(t)}}{R}\sum_{r=0}^{R-1}\mathcal{L}(W_r^{(t)} - W^{(t)})\right\|^2 &= 3\tau\eta^2\,\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\mathcal{L}(W_r^{(t)} - W^{(t)})\right\|^2 \\
&\le \frac{3\tau\eta^2}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\mathcal{L}(W_r^{(t)} - W^{(t)})\right\|^2 \\
&\le \frac{3\tau\eta^2}{R}\|\,\mathcal{L}\,\|^2\sum_{r=0}^{R-1}\mathbb{E}\left\|(W_r^{(t)} - W^{(t)})\right\|^2 = 3\tau\eta^2\rho^2\,\mathcal{E}^{(t)}.
\end{aligned} \tag{54}$$

Now, using Lemma 2(a) and then proceeding as in (53), we obtain that

$$\begin{aligned}
3\,\mathbb{E}\left\|\frac{\widetilde{\mu}\,\eta\widetilde{S}^{(t)}\,\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)})\right\|^2 &\le 3\,\widetilde{\mu}^2\,\eta^2\rho^2\,\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla\widetilde{F}(W_r^{(t)})\right\|^2 \\
&\le \widetilde{\mu}^2\,\eta^2\rho^2\left(6\tau\beta^2\,\mathcal{E}^{(t)} + 6\tau\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \frac{3\tau\sigma_1^2}{R}\right).
\end{aligned} \tag{55}$$

The proof is completed by combining (52)–(55). □

D. Convergence of FedU and dFedU for Strongly Convex Cases (Proof of Theorem 1)

First, it follows from (45) that

$$\mathbb{E}\,\|W^{(t+1)}-W^*\|^2=\mathbb{E}\,\|W^{(t)}-W^*\|^2-2\,\widetilde{\mu}\,\mathbb{E}\langle Z^{(t)},W^{(t)}-W^*\rangle+\widetilde{\mu}^2\,\mathbb{E}\,\|Z^{(t)}\|^2. \tag{56}$$

Let us now estimate the second term in the right hand side of (56). Using Lemmas 2(b), 7(b), 4(a), and 2(e), we have

$$\begin{aligned}
-2\,\widetilde{\mu}\,\mathbb{E}\langle Z^{(t)},W^{(t)}-W^*\rangle &= \frac{2\,\widetilde{\mu}\,\tau}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\langle\nabla J(W_r^{(t)}),W^*-W^{(t)}\right\rangle+\frac{2\,\widetilde{\mu}\,\tau\eta}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\langle\mathcal{L}(W_r^{(t)}-W^{(t)}),W^{(t)}-W^*\right\rangle\\
&\quad+\frac{2\,\widetilde{\mu}^2\,\eta}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\langle\mathcal{L}\,\widetilde{S}^{(t)}\nabla F(W_r^{(t)}),\widetilde{S}^{(t)}(W^{(t)}-W^*)\right\rangle\\
&\le\frac{2\,\widetilde{\mu}\,\tau}{R}\sum_{r=0}^{R-1}\left(\mathbb{E}[J(W^*)-J(W^{(t)})]-\frac{\alpha}{4}\,\mathbb{E}\,\|W^{(t)}-W^*\|^2+\beta_J\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2\right)\\
&\quad+\frac{2\,\widetilde{\mu}\,\tau\eta}{R}\sum_{r=0}^{R-1}\left(\frac{m}{2}\|\,\mathcal{L}\,\|^2\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2+\frac{1}{2m}\,\mathbb{E}\,\|W^{(t)}-W^*\|^2\right)\\
&\quad+\frac{2\,\widetilde{\mu}^2\,\eta}{R}\sum_{r=0}^{R-1}\left(\frac{n}{2}\|\,\mathcal{L}\,\|^2\,\mathbb{E}\,\|\widetilde{S}^{(t)}\nabla F(W_r^{(t)})\|^2+\frac{1}{2n}\,\mathbb{E}\,\|\widetilde{S}^{(t)}(W^{(t)}-W^*)\|^2\right)\\
&\le\frac{2\,\widetilde{\mu}\,\tau}{R}\sum_{r=0}^{R-1}\left(\mathbb{E}[J(W^*)-J(W^{(t)})]-\frac{\alpha}{4}\,\mathbb{E}\,\|W^{(t)}-W^*\|^2+\beta_J\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2\right)\\
&\quad+\frac{2\,\widetilde{\mu}\,\eta\tau}{R}\sum_{r=0}^{R-1}\left(\frac{m\rho^2}{2}\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2+\frac{1}{2m}\,\mathbb{E}\,\|W^{(t)}-W^*\|^2\right)\\
&\quad+\frac{2\,\widetilde{\mu}^2\,\tau\eta}{R}\sum_{r=0}^{R-1}\left(\frac{n\rho^2}{2}\,\mathbb{E}\,\|\nabla F(W_r^{(t)})\|^2+\frac{1}{2n}\,\mathbb{E}\,\|W^{(t)}-W^*\|^2\right),
\end{aligned}\tag{57}$$

where $m,n>0$ will be chosen later. In addition, Lemmas 5(c) and 7(c) imply that,

$$\begin{aligned}
\mathbb{E}\,\|\nabla F(W_r^{(t)})\|^2&\le 2\beta^2\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2+2\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2\\
&\le 2\beta^2\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2+\frac{8\beta^2}{\alpha}\,\mathbb{E}[J(W^{(t)})-J(W^*)]+4\sigma_2^2.
\end{aligned}\tag{58}$$

By using (57), (58), and the definition of $\mathcal{E}^{(t)}$,

$$\begin{aligned}
T_1\le&-\tau\left(2\,\widetilde{\mu}-\frac{8n\,\widetilde{\mu}^2\,\eta\rho^2\beta^2}{\alpha}\right)\mathbb{E}[J(W^{(t)})-J(W^*)]-\tau\left(\frac{\widetilde{\mu}\,\alpha}{2}-\frac{\widetilde{\mu}\,\eta}{m}-\frac{\widetilde{\mu}^2\,\eta}{n}\right)\mathbb{E}\,\|W^{(t)}-W^*\|^2\\
&+\tau\left(2\,\widetilde{\mu}\,\beta_J+m\,\widetilde{\mu}\,\eta\rho^2+2n\,\widetilde{\mu}^2\,\eta\rho^2\beta^2\right)\mathcal{E}^{(t)}+4n\,\widetilde{\mu}^2\,\tau\eta\rho^2\sigma_2^2.
\end{aligned}\tag{59}$$

Setting $m=\frac{8\eta}{\alpha}$ and $n=\frac{8\,\widetilde{\mu}\,\eta}{\alpha}$, we have

$$\begin{aligned}
T_1\le&-\tau\left(2\,\widetilde{\mu}-\frac{64\,\widetilde{\mu}^3\,\eta^2\rho^2\beta^2}{\alpha^2}\right)\mathbb{E}[J(W^{(t)})-J(W^*)]-\frac{\widetilde{\mu}\,\tau\alpha}{4}\,\mathbb{E}\,\|W^{(t)}-W^*\|^2\\
&+\widetilde{\mu}\,\tau\left(2\beta_J+\frac{8\eta^2\rho^2}{\alpha}+\frac{16\,\widetilde{\mu}^2\,\eta^2\rho^2\beta^2}{\alpha}\right)\mathcal{E}^{(t)}+\frac{32\,\widetilde{\mu}^3\,\tau\eta^2\rho^2\sigma_2^2}{\alpha}.
\end{aligned}\tag{60}$$

Combining this with (56) and Lemma 12, we get

$$\begin{aligned}
&\mathbb{E}\,\|W^{(t+1)}-W^*\|^2\\
&\le \left(1-\frac{\widetilde{\mu}\,\tau\alpha}{4}\right)\mathbb{E}\,\|W^{(t)}-W^*\|^2-\tau\left(2\,\widetilde{\mu}-\frac{64\,\widetilde{\mu}^3\,\eta^2\rho^2\beta^2}{\alpha^2}\right)\mathbb{E}[J(W^{(t)})-J(W^*)]\\
&\quad+\widetilde{\mu}\,\tau\left(2\beta_J+\frac{8\eta^2\rho^2}{\alpha}+\frac{16\,\widetilde{\mu}^2\,\eta^2\rho^2\beta^2}{\alpha}+6\,\widetilde{\mu}\,\beta_J^2+3\,\widetilde{\mu}\,\eta^2\rho^2+6\,\widetilde{\mu}^3\,\eta^2\rho^2\beta^2\right)\mathcal{E}^{(t)}\\
&\quad+6\,\widetilde{\mu}^2\,\tau\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+6\,\widetilde{\mu}^4\,\tau\eta^2\rho^2\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2+\frac{32\,\widetilde{\mu}^3\,\tau\eta^2\rho^2\sigma_2^2}{\alpha}+\frac{3\,\widetilde{\mu}^2\,\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\sigma_1^2}{R}\\
&\le \left(1-\frac{\widetilde{\mu}\,\tau\alpha}{4}\right)\mathbb{E}\,\|W^{(t)}-W^*\|^2-\tau\left(2\,\widetilde{\mu}-\frac{64\,\widetilde{\mu}^3\,\eta^2\rho^2\beta^2}{\alpha^2}\right)\mathbb{E}[J(W^{(t)})-J(W^*)]+\widetilde{\mu}\,\tau p\,\mathcal{E}^{(t)}\\
&\quad+6\,\widetilde{\mu}^2\,\tau\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+6\,\widetilde{\mu}^4\,\tau\eta^2\rho^2\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2+\frac{32\,\widetilde{\mu}^3\,\tau\eta^2\rho^2\sigma_2^2}{\alpha}+\frac{3\,\widetilde{\mu}^2\,\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\sigma_1^2}{R},
\end{aligned}\tag{61}$$

In what follows, we assume that (21) holds. where we use (21) to estimate

$$\begin{aligned}
&2\beta_J+\frac{8\eta^2\rho^2}{\alpha}+\frac{16\,\widetilde{\mu}^2\,\eta^2\rho^2\beta^2}{\alpha}+6\,\widetilde{\mu}\,\beta_J^2+3\,\widetilde{\mu}\,\eta^2\rho^2+6\,\widetilde{\mu}^3\,\eta^2\rho^2\beta^2\\
&\le p=2\beta_J+\frac{8\eta^2\rho^2}{\alpha}+\frac{64\beta^2}{\alpha}+\frac{12\beta_J^2}{\eta\rho}+6\eta\rho+\frac{48\beta^2}{\eta\rho}.
\end{aligned}\tag{62}$$

Using Lemmas 11, 7(a), and 7(c), we have

$$\begin{aligned}
&\mathbb{E}\,\|W^{(t+1)}-W^*\|^2\\
&\le \left(1-\frac{\widetilde{\mu}\,\tau\alpha}{4}\right)\mathbb{E}\,\|W^{(t)}-W^*\|^2-\tau\left(2\,\widetilde{\mu}-\frac{64\,\widetilde{\mu}^3\,\eta^2\rho^2\beta^2}{\alpha^2}\right)\mathbb{E}[J(W^{(t)})-J(W^*)]\\
&\quad+\widetilde{\mu}\,\tau p\left(4\,\widetilde{\mu}^2\,\tau\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2+\frac{2\,\widetilde{\mu}^2\,\tau\sigma_1^2}{R}\right)\\
&\quad+6\,\widetilde{\mu}^2\,\tau\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+6\,\widetilde{\mu}^4\,\tau\eta^2\rho^2\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2+\frac{32\,\widetilde{\mu}^3\,\tau\eta^2\rho^2\sigma_2^2}{\alpha}+\frac{3\,\widetilde{\mu}^2\,\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\sigma_1^2}{R}.
\end{aligned}\tag{63}$$

$$\begin{aligned}
&\le \left(1-\frac{\widetilde{\mu}\,\tau\alpha}{4}\right)\mathbb{E}\,\|W^{(t)}-W^*\|^2-\tau\left(2\,\widetilde{\mu}-\frac{64\,\widetilde{\mu}^3\,\eta^2\rho^2\beta^2}{\alpha^2}\right)\mathbb{E}[J(W^{(t)})-J(W^*)]\\
&\quad+(4p\,\widetilde{\mu}^3\,\tau^2+6\,\widetilde{\mu}^4\,\tau\eta^2\rho^2)\left(4\frac{\beta^2}{\alpha}\,\mathbb{E}[J(W)-J(W^*)]+2\sigma_2^2\right)+\frac{2p\,\widetilde{\mu}^3\,\tau^2\sigma_1^2}{R}\\
&\quad+12\,\widetilde{\mu}^2\,\tau\beta_J\,\mathbb{E}[J(W)-J(W^*)]+\frac{32\,\widetilde{\mu}^3\,\tau\eta^2\rho^2\sigma_2^2}{\alpha}+\frac{3\,\widetilde{\mu}^2\,\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\sigma_1^2}{R}
\end{aligned}\tag{64}$$

$$\begin{aligned}
&\overset{(21)}{\le} \left(1-\frac{\widetilde{\mu}\,\tau\alpha}{4}\right)\mathbb{E}\,\|W^{(t)}-W^*\|^2-\tau\left[2\,\widetilde{\mu}-\widetilde{\mu}^2\underbrace{\left(\frac{128\eta\rho\beta^2}{\alpha^2}+12\beta_J+\frac{96\beta^2}{\alpha}+\frac{32p\beta^2}{\alpha\eta\rho}\right)}_{q}\right]\mathbb{E}[J(W^{(t)})-J(W^*)]\\
&\quad+\widetilde{\mu}^3\,\tau^2\underbrace{\frac{p\left(8R\sigma_2^2+2\sigma_1^2\right)}{R}}_{C_2}+\widetilde{\mu}^2\,\tau\underbrace{\frac{(64\eta\rho R+48)\sigma_2^2+15\sigma_1^2}{\alpha R}}_{C_1}.
\end{aligned}\tag{65}$$

Let $\mu\le\frac{\widetilde{\mu}_1}{R}$. Then $\widetilde{\mu}\le\widetilde{\mu}_1=\min\left\{\frac{1}{q},\frac{2}{\eta\rho}\right\}\le\frac{1}{q}$, which implies that $2\,\widetilde{\mu}-\widetilde{\mu}^2\,q\ge\widetilde{\mu}$, and so

$$\mathbb{E}\,\|W^{(t+1)}-W^*\|^2\le\left(1-\frac{\widetilde{\mu}\,\tau\alpha}{4}\right)\mathbb{E}\,\|W^{(t)}-W^*\|^2-\widetilde{\mu}\,\tau\,\mathbb{E}[J(W^{(t)})-J(W^*)]+\widetilde{\mu}^3\,\tau^2C_2+\widetilde{\mu}^2\,\tau C_1.\tag{66}$$

Recalling that $\Delta^{(t)} = \|W^{(t)} - W^*\|^2$, rearranging the terms, and multiplying both sides of (66) with $\frac{\theta^{(t)}}{\widetilde{\mu}\,\tau\Theta_T}$, where $\Theta_T = \sum_{t=0}^{T-1} \theta^{(t)}$, we obtain that

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{\theta^{(t)}\,\mathbb{E}[J(W^{(t)})]}{\Theta_T} - J(W^*) &\leq \sum_{t=0}^{T-1} \mathbb{E}\left[\left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right) \frac{\theta^{(t)}\Delta^{(t)}}{\widetilde{\mu}\,\tau\Theta_T} - \frac{\theta^{(t)}\Delta^{(t+1)}}{\widetilde{\mu}\,\tau\Theta_T}\right] + \mu^2\tau C_2 + \widetilde{\mu}\,C_1 \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\frac{\theta^{(t-1)}\Delta^{(t)} - \theta^{(t)}\Delta^{(t+1)}}{\widetilde{\mu}\,\tau\Theta_T}\right] + \mu^2\tau C_2 + \widetilde{\mu}\,C_1 \qquad (67) \\
&= \frac{1}{\widetilde{\mu}\,\tau\Theta_T}\Delta^{(0)} - \frac{\theta^{(T-1)}}{\widetilde{\mu}\,\tau\Theta_T}\,\mathbb{E}\,\Delta^{(T)} + \widetilde{\mu}^2\,\tau C_2 + \widetilde{\mu}\,C_1 \\
&\leq \frac{1}{\widetilde{\mu}\,\tau\Theta_T}\Delta^{(0)} + \widetilde{\mu}^2\,\tau C_2 + \widetilde{\mu}\,C_1. \qquad (68)
\end{aligned}$$

Here, (67) follows from the fact that $\left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)\theta^{(t)} = \theta^{(t-1)}$ due to $\theta^{(t)} = \left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{-(t+1)}$. We then have

$$\begin{aligned}
\Theta_T &= \sum_{t=0}^{T-1} \left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{-(t+1)} \\
&= \left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{-T} \sum_{t=0}^{T-1} \left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{t} \\
&= \left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{-T} \frac{1 - \left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{T}}{\frac{\widetilde{\mu}\,\tau\alpha}{4}}.
\end{aligned}$$

Now, let $T \geq \frac{4N}{\widetilde{\mu}_1\,\alpha S}$. Then $\left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{T} \leq \exp\left(-\frac{\widetilde{\mu}\,\tau\alpha T}{4}\right) \leq \exp(-1) \leq \frac{3}{4}$, and thus

$$\Theta_T \geq \left(1 - \frac{\widetilde{\mu}\,\tau\alpha}{4}\right)^{-T} \frac{1}{\widetilde{\mu}\,\tau\alpha} = \frac{\theta^{(T-1)}}{\widetilde{\mu}\,\tau\alpha}, \qquad (69)$$

which yields $\frac{1}{\widetilde{\mu}\,\tau\Theta_T} \leq \frac{\alpha}{\theta^{(T-1)}} \leq \alpha e^{-\frac{\widetilde{\mu}\,\tau\alpha T}{4}}$. Therefore, (68) becomes

$$\sum_{t=0}^{T-1} \frac{\theta^{(t)}\,\mathbb{E}[J(W^{(t)})]}{\Theta_T} - J(W^*) \leq \alpha\Delta^{(0)} e^{-\frac{\widetilde{\mu}\,\tau\alpha T}{4}} + \widetilde{\mu}^2\,\tau C_2 + \widetilde{\mu}\,C_1, \qquad (70)$$

which together with the convexity of J implies that

$$\mathbb{E}\left[J(\widetilde{W}^{(T)}) - J(W^*)\right] = \mathbb{E}\left[J\left(\sum_{t=0}^{T-1} \frac{\theta^{(t)}}{\Theta_T} W^{(t)}\right)\right] - J(W^*) \leq \alpha\Delta^{(0)} e^{-\frac{\widetilde{\mu}\,\tau\alpha T}{4}} + \widetilde{\mu}^2\,\tau C_2 + \widetilde{\mu}\,C_1. \qquad (71)$$

Following the same of approach in [10], [55]–[57], we consider the following cases.

- If $\widetilde{\mu}_1 \geq \widehat{\mu} := \max\left\{\frac{4}{\alpha\tau T}, \frac{4}{\alpha\tau T}\log\left(\frac{\alpha^2\tau\Delta^{(0)}T}{C_1}\right)\right\}$, then we choose $\mu = \widehat{\mu}$ and have

$$\mathbb{E}\left[J(\widetilde{W}^{(T)}) - J(W^*)\right] \leq \widetilde{\mathcal{O}}\left(\frac{C_2}{\alpha^2\tau T^2}\right) + \widetilde{\mathcal{O}}\left(\frac{C_1}{\alpha\tau T}\right). \qquad (72)$$

- If $\frac{4}{\alpha\tau T} \leq \widetilde{\mu}_1 \leq \widehat{\mu}$, then we choose $\mu = \widetilde{\mu}_1$ and have

$$\mathbb{E}\left[J(\widetilde{W}^{(T)}) - J(W^*)\right] \leq \mathcal{O}\left(\alpha\Delta^{(0)} e^{-\frac{\widetilde{\mu}_1\,\alpha\tau T}{4}}\right) + \widetilde{\mathcal{O}}\left(\frac{C_2}{\alpha^2\tau T^2}\right) + \widetilde{\mathcal{O}}\left(\frac{C_1}{\alpha\tau T}\right). \qquad (73)$$

By combining the above two cases,

$$\mathbb{E}\left[J(\widetilde{W}^{(T)}) - J(W^*)\right] \leq \widetilde{\mathcal{O}}\left(\alpha\Delta^{(0)} e^{-\frac{\widetilde{\mu}_1\,\alpha\tau T}{4}} + \frac{R\sigma_2^2 + \sigma_1^2}{(\alpha T)^2 RS} + \frac{R\sigma_2^2 + \sigma_1^2}{\alpha T RS}\right), \qquad (74)$$

which implies (9). The remaining conclusion directly follows from (9).

E. Convergence of FedU and dFedU for Nonconvex Cases (Proof of Theorem 2)

By Lemma 5(e) and (45),

$$\begin{aligned}\mathbb{E}\left[J(W^{(t+1)})-J(W^{(t)})\right] &\le \mathbb{E}\langle\nabla J(W^{(t)}),W^{(t+1)}-W^{(t)}\rangle+\frac{\beta_J}{2}\,\mathbb{E}\,\|W^{(t+1)}-W^{(t)}\|^2\\ &=-\,\mathbb{E}\left\langle\nabla J(W^{(t)}),Z^{(t)}\right\rangle+\frac{\widetilde{\mu}^2\,\beta_J}{2}\,\mathbb{E}\,\|Z^{(t)}\|^2\\ &=\underbrace{-\,\mathbb{E}\left\langle\nabla J(W^{(t)}),\frac{\widetilde{\mu}\,\tau}{R}\sum_{r=0}^{R-1}\nabla J(W_r^{(t)})\right\rangle}_{T_2}+\underbrace{\mathbb{E}\left\langle\nabla J(W^{(t)}),\frac{\widetilde{\mu}\,\eta\tau\,\mathcal{L}}{R}\sum_{r=0}^{R-1}(W_r^{(t)}-W^{(t)})\right\rangle}_{T_3}\\ &\quad+\underbrace{\mathbb{E}\left\langle\nabla J(W^{(t)}),\frac{\widetilde{\mu}^2\,\eta\widetilde{S}^{(t)}\,\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla F(W_r^{(t)})\right\rangle}_{T_4}+\frac{\widetilde{\mu}^2\,\beta_J}{2}\,\mathbb{E}\,\|Z^{(t)}\|^2. \end{aligned}\tag{75}$$

Using the fact that $-xy\le\frac{-2xy+y^2}{2}=\frac{-x^2+(y-x)^2}{2},\forall x,y\in\mathbb{R}$, then Lemma 3 and Lemma 5(b), we have

$$\begin{aligned}T_2&\le-\frac{\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\left\|\frac{1}{R}\sum_{r=0}^{R-1}\nabla J(W_r^{(t)})-\nabla J(W^{(t)})\right\|^2\\ &\le-\frac{\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}\,\tau}{2R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla J(W_r^{(t)})-\nabla J(W^{(t)})\right\|^2\\ &\le-\frac{\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}\,\tau\beta_J^2}{2R}\sum_{r=0}^{R-1}\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2\\ &=-\frac{\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}\,\tau\beta_J^2}{2}\,\mathcal{E}^{(t)}\,.\end{aligned}\tag{76}$$

For the terms T_3, by Lemmas 4(a), and 3,

$$\begin{aligned}T_3&\le\frac{z\,\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}\,\tau}{2z}\,\mathbb{E}\left\|\frac{\eta\,\mathcal{L}}{R}\sum_{r=0}^{R-1}(W_r^{(t)}-W^{(t)})\right\|^2\\ &\le\frac{z\,\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}\,\tau\eta^2}{2zR}\sum_{r=0}^{R-1}\|\,\mathcal{L}\,\|^2\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2\\ &=\frac{z\,\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}\,\tau\eta^2\rho^2}{2z}\,\mathcal{E}^{(t)},\end{aligned}\tag{77}$$

where $z>0$ will be chosen later.

For the terms T_4, using Lemmas 2(b), 4(a), 3, 2(e), 5(d), we derive that

$$\begin{aligned}T_4&=\widetilde{\mu}\,\mathbb{E}\left\langle\widetilde{S}^{(t)}\nabla J(W^{(t)}),\frac{\widetilde{\mu}\,\eta\,\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla F(W_r^{(t)})\right\rangle\\ &\le\frac{s\,\widetilde{\mu}}{2}\,\mathbb{E}\,\|\widetilde{S}^{(t)}\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}^3\,\eta^2}{2s}\,\mathbb{E}\left\|\frac{\mathcal{L}}{R}\sum_{r=0}^{R-1}\widetilde{S}^{(t)}\nabla F(W_r^{(t)})\right\|^2\\ &\le\frac{s\,\widetilde{\mu}}{2}\,\mathbb{E}\,\|\widetilde{S}^{(t)}\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}^3\,\eta^2}{2sR}\sum_{r=0}^{R-1}\|\,\mathcal{L}\,\|^2\,\mathbb{E}\left\|\widetilde{S}^{(t)}\nabla F(W_r^{(t)})\right\|^2\\ &\le\frac{s\,\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}^3\,\tau\eta^2\rho^2}{2sR}\sum_{r=0}^{R-1}\mathbb{E}\,\|\nabla F(W_r^{(t)})\|^2\\ &\le\frac{s\,\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}^3\,\tau\eta^2\rho^2}{2sR}\sum_{r=0}^{R-1}\left(2\beta^2\,\mathbb{E}\,\|W_r^{(t)}-W^{(t)}\|^2+2\|\nabla F(W^{(t)})\|^2\right),\\ &=\frac{s\,\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2+\frac{\widetilde{\mu}^3\,\tau\eta^2\rho^2\beta^2}{s}\,\mathcal{E}^{(t)}+\frac{\widetilde{\mu}^3\,\tau\eta^2\rho^2}{s}\|\nabla F(W^{(t)})\|^2,\end{aligned}\tag{78}$$

where $s>0$ will be chosen later. By the definition of σ_2 in Lemma 1, it holds that $\|\nabla F(W)\|^2 \leq \sigma_2^2 + \|\nabla J(W)\|^2$, which together with (21) and (78) yields

$$\begin{aligned} T_4 &\leq \frac{s\,\widetilde{\mu}\,\tau}{2}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \frac{\widetilde{\mu}^3\,\tau\eta^2\rho^2}{sR}\sum_{r=0}^{R-1}\left(\beta^2\,\mathbb{E}\,\|W_r^{(t)} - W^{(t)}\|^2 + \mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \sigma_2^2\right) \\ &\leq \widetilde{\mu}\,\tau\left(\frac{s}{2} + \frac{2\,\widetilde{\mu}\,\eta\rho}{s}\right)\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \frac{4\,\widetilde{\mu}\,\tau\beta^2}{s}\,\mathcal{E}^{(t)} + \frac{2\,\widetilde{\mu}^2\,\tau\eta\rho\sigma_2^2}{sR}, \end{aligned} \tag{79}$$

Choosing $z = s = \frac{1}{4}$, combining (75)–(78) with Lemma 12, and then using Lemma 11, we obtain that

$$\begin{aligned} \mathbb{E}[J(W^{(t+1)}) - J(W^{(t)})] &\leq -\tau\left(\frac{\widetilde{\mu}}{4} - 3\,\widetilde{\mu}^2\,\beta_J\right)\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \left(4\,\widetilde{\mu}^3\,\tau\eta^2\rho^2 + 3\,\widetilde{\mu}^4\,\tau\eta^2\rho^2\beta_J\right)\|\nabla F(W^{(t)})\|^2 \\ &\quad + \widetilde{\mu}\,\tau\left(\frac{\beta_J^2}{2} + 2\eta^2\rho^2 + 4\,\widetilde{\mu}^2\,\eta^2\rho^2\beta^2 + 3\,\widetilde{\mu}\,\beta_J^3 + \frac{3}{2}\,\widetilde{\mu}\,\eta^2\rho^2\beta_J + 3\,\widetilde{\mu}^3\,\eta^2\rho^2\beta_J\beta^2\right)\mathcal{E}^{(t)} \\ &\quad + \frac{3\,\widetilde{\mu}^2\,\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\beta_J\sigma_1^2}{2R} \\ &\leq -\tau\left(\frac{\widetilde{\mu}}{4} - 3\,\widetilde{\mu}^2\,\beta_J\right)\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \left(4\,\widetilde{\mu}^3\,\tau\eta^2\rho^2 + 3\,\widetilde{\mu}^4\,\tau\eta^2\rho^2\beta_J\right)\|\nabla F(W^{(t)})\|^2 \\ &\quad + \widetilde{\mu}\,\tau u\left(4\,\widetilde{\mu}^2\,\tau\,\mathbb{E}\,\|\nabla F(W^{(t)})\|^2 + \frac{2\,\widetilde{\mu}^2\,\tau\sigma_1^2}{R}\right) + \frac{3\,\widetilde{\mu}^2\,\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\beta_J\sigma_1^2}{2R}, \end{aligned} \tag{80}$$

where we use (21) to estimate

$$\begin{aligned} &\frac{\beta_J^2}{2} + 2\eta^2\rho^2 + 4\,\widetilde{\mu}^2\,\eta^2\rho^2\beta^2 + 3\,\widetilde{\mu}\,\beta_J^3 + \frac{3}{2}\,\widetilde{\mu}\,\eta^2\rho^2\beta_J + 3\,\widetilde{\mu}^3\,\eta^2\rho^2\beta_J\beta^2 \\ &\leq u = \frac{\beta_J^2}{2} + 2\eta^2\rho^2 + 16\beta^2 + \frac{6\beta_J^3}{\eta\rho} + 3\eta\rho\beta_J + \frac{24\beta_J\beta^2}{\eta\rho}. \end{aligned} \tag{81}$$

By the definition of σ_2 in Lemma 1, it holds that $\|\nabla F(W)\|^2 \leq \sigma_2^2 + \|\nabla J(W)\|^2$, which together with (80) and (21) yields

$$\begin{aligned} \mathbb{E}[J(W^{(t+1)}) - J(W^{(t)})] &\leq -\tau\left(\frac{\widetilde{\mu}}{4} - 3\,\widetilde{\mu}^2\,\beta_J\right)\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \left(4\,\widetilde{\mu}^3\,\tau\eta^2\rho^2 + 3\,\widetilde{\mu}^4\,\tau\eta^2\rho^2\beta_J\right)\left(\sigma_2^2 + \mathbb{E}\,\|\nabla J(W)\|^2\right) \\ &\quad + \widetilde{\mu}^3\,\tau^2 u\left[4\left(\sigma_2^2 + \mathbb{E}\,\|\nabla J(W)\|^2\right) + \frac{2\sigma_1^2}{R}\right] + \frac{3\,\widetilde{\mu}^2\,\tau(1+\widetilde{\mu}^2\,\eta^2\rho^2)\beta_J\sigma_1^2}{2R} \\ &\overset{\tau\leq 1}{\leq} -\tau\left(\frac{\widetilde{\mu}}{4} - 8\,\widetilde{\mu}^2\,\tau\eta\rho - 3\,\widetilde{\mu}^2\,\beta_J - 3\,\widetilde{\mu}^2\,\tau\beta_J - \frac{8\,\widetilde{\mu}^2\,u}{\eta\rho}\right)\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 \\ &\quad + \widetilde{\mu}^3\,\tau^2\underbrace{\frac{u\left(4R\sigma_2^2 + 2\sigma_1^2\right)}{R}}_{C_4} + \widetilde{\mu}^2\,\tau\underbrace{\frac{(8\eta\rho + 12\beta_J)R\sigma_2^2 + 27\beta_J\sigma_1^2}{2R}}_{C_3} \end{aligned} \tag{82}$$

Now, let

$$\widetilde{\mu} \leq \widetilde{\mu}_2 = \min\left\{\frac{2}{\eta\rho}, \frac{1}{v}\right\} = \min\left\{\frac{2}{\eta\rho}, \frac{1}{8\left(8\eta\rho + 3\beta_J + 12\beta_J + \frac{8u}{\eta\rho}\right)}\right\}. \tag{83}$$

Then $-\tau\left(\frac{\widetilde{\mu}}{4} - 8\,\widetilde{\mu}^2\,\eta\rho - 3\,\widetilde{\mu}^2\,\beta_J - 12\,\widetilde{\mu}^2\,\beta_J - \frac{8\,\widetilde{\mu}^2\,u}{\eta\rho}\right) \leq -\frac{\widetilde{\mu}\,\tau}{8}$, and so

$$\mathbb{E}\left[J(W^{(t+1)}) - J(W^{(t)})\right] \leq -\frac{\widetilde{\mu}\,\tau}{8}\,\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 + \widetilde{\mu}^3\,\tau^2 C_4 + \widetilde{\mu}^2\,\tau C_3. \tag{84}$$

By re-arranging the terms of (84) and telescoping, we have

$$\begin{aligned} \frac{1}{8T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla J(W^{(t)})\|^2 &\leq \frac{\mathbb{E}\left[J(W^{(0)}) - J(W^{(T)})\right]}{\mu\tau T} + \widetilde{\mu}^2\,\tau C_4 + \widetilde{\mu}\,C_3 \\ &\leq \frac{\mathbb{E}\left[J(W^{(0)}) - J(W^*)\right]}{\mu\tau T} + \widetilde{\mu}^2\,\tau C_4 + \widetilde{\mu}\,C_3. \end{aligned} \tag{85}$$

Let $\Delta_J := J(W^{(0)}) - J(W^*)$. Following the same of approach in [10], [55]–[57], we consider the following cases.

- If $\widetilde{\mu}_2^3 \leq \frac{\Delta_J}{C_4\tau^2 T}$ and $\widetilde{\mu}_2^2 \leq \frac{\Delta_J}{C_3\tau T}$, then we choose $\widetilde{\mu} = \widetilde{\mu}_2$ to get

$$\frac{1}{8T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(W^{(t)})\right\|^2 \leq \frac{\Delta_J}{\widetilde{\mu}_2\,\tau T} + \frac{\Delta_J^{\frac{2}{3}}C_4^{\frac{1}{3}}}{\tau^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\Delta_J^{\frac{1}{2}}C_3^{\frac{1}{2}}}{\tau^{\frac{1}{2}}T^{\frac{1}{2}}}. \tag{86}$$

- If $\widetilde{\mu}_2^3 \geq \frac{\Delta_J}{C_4\tau^2 T}$ or $\widetilde{\mu}_2^2 \geq \frac{\Delta_J}{C_3\tau T}$, then we choose $\widetilde{\mu} = \min\left\{\left(\frac{\Delta_J}{C_4\tau^2 T}\right)^{\frac{1}{3}}, \left(\frac{\Delta_J}{C_5\tau T}\right)^{\frac{1}{2}}\right\}$ to get

$$\frac{1}{8T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(W^{(t)})\right\|^2 \leq \frac{\Delta_J^{\frac{2}{3}}C_4^{\frac{1}{3}}}{\tau^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\Delta_J^{\frac{1}{2}}C_3^{\frac{1}{2}}}{\tau^{\frac{1}{2}}T^{\frac{1}{2}}}. \tag{87}$$

Combining two cases, and with t^* uniformly sampled from $\{0,\ldots,T-1\}$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(W^{(t)})\right\|^2 = \mathbb{E}\left\|\nabla J(W^{(t^*)})\right\|^2 &\leq \mathcal{O}\left(\frac{\Delta_J}{\widetilde{\mu}_2\,\tau T} + \frac{\Delta_J^{\frac{2}{3}}C_4^{\frac{1}{3}}}{\tau^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\Delta_J^{\frac{1}{2}}C_3^{\frac{1}{2}}}{\tau^{\frac{1}{2}}T^{\frac{1}{2}}}\right)\\ &= \mathcal{O}\left(\frac{\Delta_J}{TS} + \frac{\Delta_J^{\frac{2}{3}}M^{\frac{2}{3}}}{T^{\frac{2}{3}}(RS)^{\frac{1}{3}}} + \frac{\Delta_J^{\frac{1}{2}}M^2}{\sqrt{TRS}}\right).\end{aligned} \tag{88}$$

This proves (11), which, in turn, implies (12).

F. Additional Experimental Settings and Results

1) Statistics of All Datasets

We use four real datasets for the experiments including Human Activity Recognition, Vehicle sensor, MNIST, and CIFAR-10. The detailed statistics of all datasets are summarized in Table. II.

TABLE II: Statistics of all datasets using in the experiment.

Dataset	N	Total samples	Num labels / client	Samples / client Mean	Samples / client Std
Human Activity	30	10,299	6	343	35.1
Vehicle Sensor	23	48,303	2	2,100	380.5
MNIST	100	61,866	2	619	343.8
CIFAR-10	20	54,572	3	2729	851.4

2) Learning Tasks at Local Clients

- **Strongly convex setting**: We use a multinomial logistic regression model (MLR) with a cross-entropy loss function and a L_2-regularization term for all strongly convex experiments on Human Activity, Vehicle Sensor, and MNIST datasets. The loss function at each client is defined as follow:

$$F_k(w) = \frac{-1}{D_k}\sum_{j=1}^{D_k}\sum_{c=1}^{C} 1_{\{y_j=c\}}\log\frac{\exp(\langle a_j, w_c\rangle)}{\sum_{i=1}^{C}\exp(\langle a_i, w_i\rangle)} + \frac{\alpha}{2}\sum_{c=1}^{C}\|w_c\|_2^2.$$

- **Nonconvex setting**: We use a simple DNN with one hidden layer, a ReLU activation function, and a softmax layer at the end of the network for Human Activity and Vehicle Sensor datasets. The size of hidden layer is 100 for Human Activity and 20 for Vehicle Sensor. In the case of MNIST, we use DNN with 2 hidden layers and both layers have the same size of 100. For CIFAR-10, we follow the CNN structure of [1].

3) Performance of FedU in Federated Multi-Task Learning without down-sampling data

The result in Table. III shows that FedU still achieves the highest performance, however, the performance gaps between FedU, MOCHA, and Local model are less appreciable. When the local data at a client is large enough, the Local model at one client can learn individually without contributions from others. Therefore, both FedU and MOCHA will show advantages compared to the Local model in federated settings when there are various clients having a small number of data.

4) Comparison with Personalized Federated Learning Algorithms without down-sampling data

Similar to the down-sampling data setting, FedU almost maintains the top performance in all scenarios showing in Table. IV. Only in the nonconvex case on MNIST, pFedMe performs slightly better than FedU.

TABLE III: Performance comparison in mult-task setting without down-sampling data (All tasks participate, $R = 5$, $S = N$, mini-batch size $B = 20$, $\eta = 10^{-2}$, $T = 200$).

Dataset	Algorithm	Test accuracy	
		Convex	Nonconvex
Human Activity	FedU	**99.10** $\pm$ 0.18	**99.21** $\pm$ 0.15
	MOCHA	98.79 $\pm$ 0.04	
	Local	98.29 $\pm$ 0.01	98.34 $\pm$ 0.03
	Global	93.79 $\pm$ 0.27	94.58 $\pm$ 0.16
Vehicle Sensor	FedU	**91.16** $\pm$ 0.02	**95.43** $\pm$ 0.09
	MOCHA	90.94 $\pm$ 0.05	
	Local	88.16 $\pm$ 0.05	92.10 $\pm$ 0.06
	Global	80.21 $\pm$ 0.12	83.00 $\pm$ 0.11
MNIST	FedU	**98.07** $\pm$ 0.02	**98.61** $\pm$ 0.02
	MOCHA	97.99 $\pm$ 0.02	
	Local	97.95 $\pm$ 0.01	97.99 $\pm$ 0.02
	Global	92.04 $\pm$ 0.02	96.19 $\pm$ 0.09

TABLE IV: Performance comparison of centralized setting without down-sampling data ($R = 5$, $S = 0.1N$, $B = 20$, $T = 200$).

Dataset	Algorithm	Test Accurancy	
		Convex	Non Convex
CIFAR-10	FedU		**79.40** $\pm$ 0.25
	pFedMe		78.70 $\pm$ 0.15
	Per-FedAvg		67.61 $\pm$ 0.03
	FedAvg		36.32 $\pm$ 5.57
MNIST	FedU	**97.82** $\pm$ 0.02	98.44 $\pm$ 0.02
	MOCHA	97.80 $\pm$ 0.02	
	pFedMe	95.38 $\pm$ 0.09	**99.04** $\pm$ 0.02
	Per-FedAvg	91.77 $\pm$ 0.23	97.59 $\pm$ 0.30
	FedAvg	90.14 $\pm$ 0.61	90.74 $\pm$ 1.62
Vehicle Sensor	FedU	**89.84** $\pm$ 0.06	**94.18** $\pm$ 0.08
	MOCHA	89.73 $\pm$ 0.89	
	pFedMe	85.87 $\pm$ 0.02	92.23 $\pm$ 0.17
	Per-FedAvg	82.21 $\pm$ 0.01	87.50 $\pm$ 1.21
	FedAvg	81.54 $\pm$ 0.03	85.61 $\pm$ 0.07
Human Activity	FedU	**97.75** $\pm$ 0.21	**97.85** $\pm$ 0.39
	MOCHA	97.69 $\pm$ 0.03	
	pFedMe	97.52 $\pm$ 0.09	97.60 $\pm$ 0.09
	Per-FedAvg	96.04 $\pm$ 0.36	96.21 $\pm$ 0.33
	FedAvg	95.58 $\pm$ 0.05	94.84 $\pm$ 0.07

5) Effect of non-i.i.d levels

To show the effect of different degrees of non-i.i.d on FedU, we did experiments on the MNIST dataset in Fig. 5 for illustration purposes. We use the Dirichlet distribution to generate MNIST non-i.i.d dataset with 100 clients following setting in [58]–[60]. Specifically, client's data is partitioned by using Dirichlet distribution $\text{Dir}_N(\alpha)$, where $N = 100$ is total number of clients and α is concentration parameter. In this setting, $\alpha > 0$ is to control the identicalness among clients. When $\alpha \to \infty$, all clients have identical distributions, corresponding to the i.i.d setting. By contrast, when $\alpha \to 0$, each client holds examples from only one class chosen randomly [58]. In our experiment, we consider three different values for $\alpha \in \{0.01, 0.1, 10.0\}$ to generate populations covering a spectrum of identicalness. In Fig. 5, we consider the network of 100 clients and increase the level of non-i.i.d from left to right. When the level of non-i.i.d increases, personalized algorithms are more stable than traditional federated learning algorithms like FedAvg or FedProx. Importantly, our proposed algorithm FedU performs well in all settings compared to other algorithms and is not much affected by the high degree of non-i.i.d.

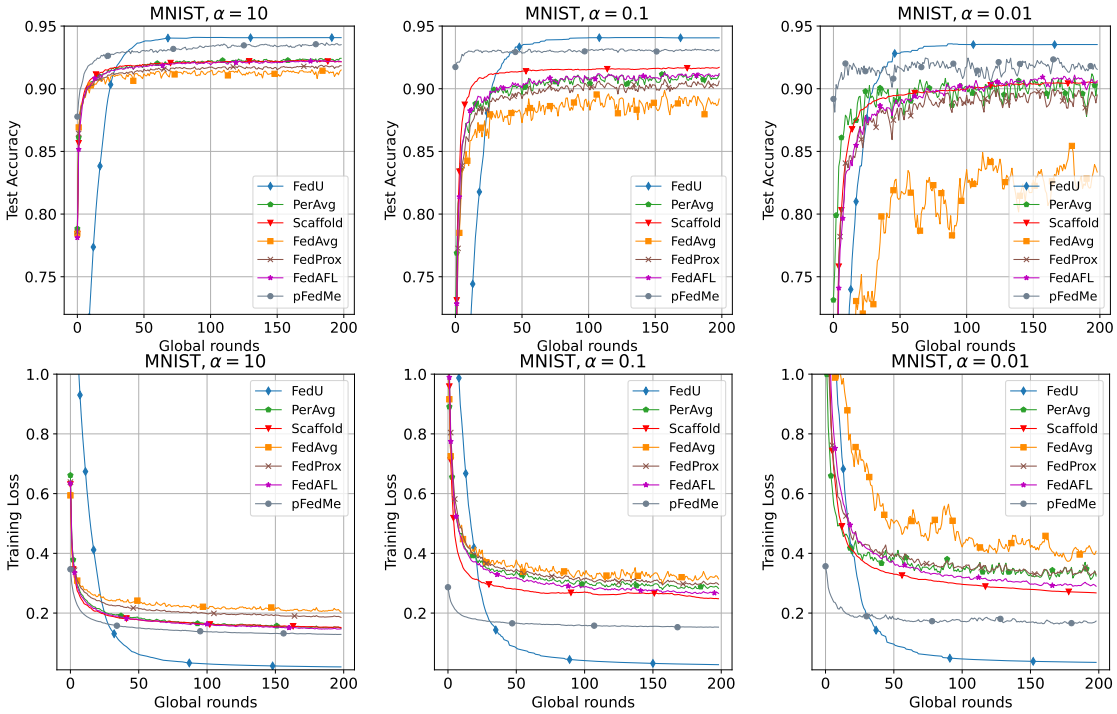


Fig. 5: Effect of different non-i.i.d degrees on Federated Learning algorithms. α is the concentration parameter to control the level of non-i.i.d.