2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3503-2411-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/MLSP55844.2023.10285949

PRIVACY-PRESERVING FEDERATED PRIMAL-DUAL LEARNING FOR NON-CONVEX PROBLEMS WITH NON-SMOOTH REGULARIZATION

Yiwei Li[⋆], *Chien-Wei Huang*[⋆], *Shuai Wang*[†], *Chong-Yung Chi*[⋆], *Tony Q. S. Quek*[†]

[⋆]Institute of Communications Engineering, National Tsing Hua University, Hsinchu, Taiwan.
[†]Information Systems Technology and Design,
Singapore University of Technology and Design, Singapore.

Email: lywei0306@foxmail.com, s110064501@gmail.com, shuaiwang@link.cuhk.edu.cn,
cychi@ee.nthu.edu.tw, tonyquek@sutd.edu.sg.

ABSTRACT

Recently, the federated learning (FL) has been a machine learning paradigm for the preservation of data privacy, though high communication cost and privacy protection are still the main concerns of FL. However, in many practical applications, the trained model needs certain nature or characteristics, such as sparseness in classification, otherwise learning performance loss is inevitable. In order to upgrade the learning performance, a suitable non-smooth regularizer (e.g., ℓ_1-norm for the model sparseness) can be added to the loss function (often non-convex) in the considered optimization problem. This paper proposes a novel primal-dual learning algorithm to handle such non-smooth regularization aided non-convex FL problems, that yields much superior learning performance over some state-of-the-art FL algorithms under privacy guarantee by means of differential privacy. Finally, some experimental results are provided to demonstrate the efficacy of the proposed algorithm.

***Index Terms*—** Federated learning, primal-dual method, non-convex and non-smooth optimization, differential privacy.

I. INTRODUCTION

Federated learning (FL) offers the training of machine learning (ML) models in distributed manner, where massively distributed clients repeatedly refine the model parameters under the orchestration of a parameter server (PS) without the need of sharing their private data [1]. The training process of FL faces many challenges due to unreliable and limited network resources [2], [3]. Consequently, the communication between clients and the PS can be very costly and inefficient, certainly resulting in a bottleneck to the applicability of FL [4]. Besides the concern of communication efficiency, FL may still suffer from privacy leakage as the exchanged messages between the clients and the PS can be reversely deduced by professional or experienced adversaries, especially in wireless scenarios [5], [6].

Recent progress and increasing efforts have driven the development of FL algorithms by focusing on the aforementioned issues. Most works tried to improve communication efficiency by allowing partial client participation (PCP) and proper multiple updates of local stochastic gradient descent (SGD) in each communication round [1], [5]. Despite their improved results, few works can effectively handle the following non-convex and non-smooth problem [7]–[9]:

$$\min \big\{ \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x}) + h(\mathbf{x}) \big\}, \tag{1}$$

where $\mathbf{x}$ is the model parameter vector, and N is the number of devices (or clients), f_i is a non-convex loss function, and h is a non-smooth (possibly non-convex) regularizer including the regularization parameter. Problem (1) is pervasive and covers many ML applications, e.g., sparse learning [6], [10]. However, it is quite hard to solve problem (1) due to its non-convexity and non-smoothness in the FL design, especially under uncompromising requirements for both communication efficiency and privacy protection [6], [11].

The primal-dual method (PDM) [8], [9] in distributed optimization has been widely used for exploring frontiers of FL. Recent works applied PDM to FL and demonstrated its good convergence performance over the traditional FL algorithms such as FedAvg [2], [6], [12], thereby having drawn high attention to the communication efficiency improvement. On the other hand, various differential privacy (DP) based mechanisms have been proposed thanks to its well-supported theory and negligible system overhead [5], [6], [13], [14]. Nevertheless, high communication efficiency and privacy guarantee for solving the optimization problem (1) through the combination of primal-dual methods and DP remains to be solved effectively and efficiently.

This paper proposes an algorithm for solving problem (1) in FL system for better communication efficiency and

This work is supported by the Ministry of Science and Technology, Taiwan, under Grants MOST 111-2221-E-007-035-MY2 and 111-2221-E-007-047-MY2-.

stronger privacy guarantee together with superior learning performance. We validate the efficacy of the proposed algorithm by both theoretical analysis and experimental results. So far, quite few effective works have been reported for the development of FL algorithms involving a non-convex loss function and a non-smooth regularizer.

Notation: Let $\mathbb{R}^d$ and $\mathbb{R}^{m\times n}$ denote the set of real $d\times 1$ vectors and $m\times n$ matrices, respectively; $[N]$ denotes the integer set $\{1,\ldots,N\}$; $\{\mathbf{x}_i\}$ denotes the set of $\mathbf{x}_1,\mathbf{x}_2,\ldots$ for all the admissible i. Let $\mathbf{x}^\top$ denote the transpose of vector $\mathbf{x}$; $[\mathbf{X}]_{jk}$ is the (j,k)-th entry of matrix $\mathbf{X}$; $\mathbf{I}_d$ represents the $d\times d$ identity matrix. Let $\|\mathbf{x}\|$, $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_0$ denote the Euclidean norm, ℓ_1-norm and the cardinality of vector $\mathbf{x}$, respectively; the operators $\mathbb{E}[\cdot]$ and $\mathbb{P}[\cdot]$ represent the statistical expectation and the probability function, respectively. $\mathrm{Tr}(\cdot)$ denotes the trace operator; $\nabla f(\mathbf{x})$ denotes the gradient of function $f(\mathbf{x})$ with respect to (w.r.t.) $\mathbf{x}$.

II. Preliminaries and Problem Setup

A. Problem Formulation

We consider a vanilla FL framework consisting of one PS and N clients. Suppose that client $i\in[N]$ holds a private local dataset $\mathcal{D}_i$ with finite size $|\mathcal{D}_i|$, and $f_i(\mathbf{x};\mathcal{D}_i)$ is the given loss function. For simplicity, $f_i(\mathbf{x})$ is used throughout the paper unless a different dataset is used. Then problem (1) under consideration can be expressed as

$$\min\ \Big\{\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}_i)+h(\mathbf{x}_0)\Big\} \tag{2a}$$

$$\text{s.t. } \mathbf{x}_i=\mathbf{x}_0,\quad \forall i\in[N], \tag{2b}$$

where $\mathbf{x}_i\in\mathbb{R}^d$ is the model of client i (local model), and $\mathbf{x}_0\in\mathbb{R}^d$ is the global model. The existing FL algorithms are almost not directly applicable when the loss function of problem (2) is non-convex and non-smooth [2], [6]. This motivated us to develop an effective communication-efficient and privacy-preserving FL algorithm for solving problem (2).

B. Preliminaries of Differential Privacy

For ease of later use, let us briefly review (ϵ,δ)-DP and privacy loss as follows.

Definition 1 *((ϵ,δ)-DP [14]). Suppose that $\mathcal{X}$ is a given dataset, and two neighboring datasets $\mathcal{D},\mathcal{D}'\subset\mathcal{X}$, which differ in only one data sample. A randomized mechanism $\mathcal{M}:\mathcal{X}\to\mathbb{R}^d$ achieves (ϵ,δ)-DP if for any subset of outputs $\mathcal{O}\subseteq Range(\mathcal{M})$:*

$$\mathbb{P}[\mathcal{M}(\mathcal{D})\in\mathcal{O}]\le\exp(\epsilon)\cdot\mathbb{P}\left[\mathcal{M}\left(\mathcal{D}'\right)\in\mathcal{O}\right]+\delta, \tag{3}$$

where $\epsilon>0$ and $\delta\in[0,1)$.

Note that a smaller ϵ means stronger privacy protection (due to interchangeable $\mathcal{D}$ and $\mathcal{D}'$), and δ stands for the probability to break the $(\epsilon,0)$-DP. The (ϵ,δ)-DP can be implemented by properly adding Gaussian noise $\boldsymbol{\xi}\in\mathbb{R}^d$ to protect data privacy [14], that is

$$\mathcal{M}(\mathcal{D})=\boldsymbol{g}(\mathcal{D})+\boldsymbol{\xi},\ \ \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d), \tag{4}$$

where $\boldsymbol{g}(\cdot)$ is a specified query function, and $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)$ is the distribution of a zero-mean Gaussion noise with covariance matrix $\sigma^2\mathbf{I}_d$. The required "noise variance" σ^2 (i.e., variance of every element in $\boldsymbol{\xi}$) for achieving (ϵ,δ)-DP is given by the following lemma.

Lemma 1 *[14, Theorem 3.22] Suppose that the randomized mechanism $\mathcal{M}$ satisfies (ϵ,δ)-DP defined in* (4)*. Then, the minimum required noise variance σ^2 is given by*

$$\sigma^2=\frac{2s^2\ln(1.25/\delta)}{\epsilon^2}, \tag{5}$$

where s, the ℓ_2-norm sensitivity of $\boldsymbol{g}$ in (4)*, is given by*

$$s\triangleq\max_{\mathcal{D},\mathcal{D}'\subset\mathcal{X}}\left\|\boldsymbol{g}(\mathcal{D})-\boldsymbol{g}\left(\mathcal{D}'\right)\right\|, \tag{6}$$

in which $\mathcal{X}$ is the domain of function $\boldsymbol{g}$.

Definition 2 *(Privacy loss [14]). Suppose that a randomized mechanism $\mathcal{M}$ satisfies (ϵ,δ)-DP. Let $\mathcal{D}$ and $\mathcal{D}'$ be two neighboring datasets and $\boldsymbol{o}$ be a possible random vector of $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$. Then, the privacy loss is defined by*

$$\alpha(\boldsymbol{o})=\ln\Big(\frac{\mathbb{P}\big[\mathcal{M}(\mathcal{D})=\boldsymbol{o}\big]}{\mathbb{P}\big[\mathcal{M}(\mathcal{D}')=\boldsymbol{o}\big]}\Big). \tag{7}$$

Note that when $\boldsymbol{o}$ is a continuous random vector, $\mathbb{P}[\cdot]$ stands for its probability density function, and this is exactly the case in our work.

III. Proposed Privacy-preserving Federated Primal-dual Method

The distributed PDM solves problem (2) by sequentially optimizing the augmented Lagrangian (9a), w.r.t. $\mathbf{x},\mathbf{x}_0$ (minimization), and $\boldsymbol{\lambda}$ (maximization). Specifically, for iteration $t=0,1,\ldots,$ they are updated as follows:

$$\mathbf{x}_i^{t+1}=\arg\min_{\mathbf{x}_i}\mathcal{L}_i\big(\mathbf{x}_i,\mathbf{x}_0^t,\boldsymbol{\lambda}_i^t\big),i\in[N], \tag{8a}$$

$$\boldsymbol{\lambda}_i^{t+1}=\boldsymbol{\lambda}_i^t+\rho\big(\mathbf{x}_0^t-\mathbf{x}_i^{t+1}\big),i\in[N], \tag{8b}$$

$$\mathbf{x}_0^{t+1}=\arg\min_{\mathbf{x}_0}\mathcal{L}\big(\mathbf{x}^{t+1},\mathbf{x}_0,\boldsymbol{\lambda}^{t+1}\big), \tag{8c}$$

where

$$\mathcal{L}\big(\mathbf{x},\mathbf{x}_0,\boldsymbol{\lambda}\big)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i\big(\mathbf{x}_i,\mathbf{x}_0,\boldsymbol{\lambda}_i\big)+h(\mathbf{x}_0), \tag{9a}$$

$$\mathcal{L}_i\big(\mathbf{x}_i,\mathbf{x}_0,\boldsymbol{\lambda}_i\big)=f_i\big(\mathbf{x}_i\big)+\boldsymbol{\lambda}_i^\top(\mathbf{x}_0-\mathbf{x}_i)+\frac{\rho}{2}\left\|\mathbf{x}_0-\mathbf{x}_i\right\|^2, \tag{9b}$$

in which $\mathbf{x}=[\mathbf{x}_1^\top,\ldots,\mathbf{x}_N^\top]^\top$, $\boldsymbol{\lambda}=[\boldsymbol{\lambda}_1^\top,\ldots,\boldsymbol{\lambda}_N^\top]^\top$, $\rho>0$ is the penalty parameter which must be large enough such that $\mathcal{L}(\cdot)$ is strongly convex in $\mathbf{x}_i$ [15].

In the FL framework, t corresponds to the number of communication round, and the iterates in (8) may not meet the requirement of communication efficiency and privacy protection. Following the same fashion in FedAvg [2], the

strategy of multiple steps of local SGD in each communication round is considered to improve communication efficiency. In addition, a subset of clients is randomly selected to mitigate the straggler effect in FL [2].

To proceed, we denote $\mathcal{S}_t \subseteq [N]$ with fixed size $|\mathcal{S}_t| = K$ as the set of participated clients at the t-th round. Then, for the t-th round, an inner loop with Q_i^t iterations (where, rather than an preassigned parameter, Q_i^t is determined automatically by the algorithm under consideration), (8a) and (8b) are further replaced by

$$\mathbf{x}_i^{t,r+1} = \mathbf{x}_i^{t,r} - \eta^t \big(\nabla f_i(\mathbf{x}_i^{t,r}; \mathcal{B}_i^{t,r}) - \boldsymbol{\lambda}_i^t + \rho(\mathbf{x}_i^{t,r} - \mathbf{x}_0^t)\big), \quad (10a)$$

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^{t,Q_i^t}, \quad (10b)$$

$$\boldsymbol{\lambda}_i^{t+1} = \boldsymbol{\lambda}_i^t + \rho\left(\mathbf{x}_0^t - \mathbf{x}_i^{t+1}\right), \quad (10c)$$

$$\mathbf{y}_i^{t+1} = \mathbf{x}_i^{t+1} - \boldsymbol{\lambda}_i^{t+1}/\rho, \quad (10d)$$

where η^t denotes the step size, $\mathcal{B}_i^{t,r} \subseteq \mathcal{D}_i$ is a mini-batch dataset with size $|\mathcal{B}_i^{t,r}| = b$. For a preassigned small $\nu > 0$ for controlling the size of the vector $(\mathbf{x}_i^{t,r+1} - \mathbf{x}_i^{t,r})$, the inner loop ends when

$$\left\|\nabla f_i(\mathbf{x}_i^{t,r}; \mathcal{B}_i^{t,r}) - \boldsymbol{\lambda}_i^t + \rho(\mathbf{x}_i^{t,r} - \mathbf{x}_0^t)\right\|^2 \leq \nu, \quad (11)$$

and Q_i^t determined by (11) is the smallest number of iterations spent in the inner loop.

To guarantee (ϵ, δ)-DP for the local model $\mathbf{y}_i^{t+1}$ to be uploaded, the aforementioned artificial Gaussian noise $\boldsymbol{\xi}_i^{t+1}$ is added to $\mathbf{y}_i^{t+1}$, i.e.,

$$\widetilde{\mathbf{y}}_i^{t+1} = \mathbf{y}_i^{t+1} + \boldsymbol{\xi}_i^{t+1} = \mathbf{x}_i^{t+1} - \frac{\boldsymbol{\lambda}_i^{t+1}}{\rho} + \boldsymbol{\xi}_i^{t+1}. \quad (12)$$

Finally, $\mathbf{x}_0^{t+1}$ given by (8c) can alternatively be expressed as

$$\mathbf{x}_0^{t+1} = \mathrm{prox}_{\rho h}\Big(\frac{1}{K}\sum_{i\in\mathcal{S}_t} \widetilde{\mathbf{y}}_i^{t+1}\Big), \quad (13)$$

where $\mathrm{prox}_{\rho h}(\mathbf{u}) \triangleq \arg\min_{\boldsymbol{x}}\{h(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \mathbf{u}\|^2\}$ is the proximal operator [16]. The proof of (13) is given in Appendix A. The proposed FedPDM with DP (FedPDM-DP) is implemented by Algorithm 1.

IV. Privacy and Convergence Analysis

A. Assumptions

Assumption 1 *Each loss function $f_i(\cdot)$ in (2) is L-smooth, i.e., f_i is continuously differentiable and there exists an $L > 0$ such that*

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall i \in [N]. \quad (14)$$

Besides, $\mathcal{L}(\cdot)$ given in (9a) is bounded below, i.e.,

$$\underline{f} \triangleq \inf_{\mathbf{x},\mathbf{x}_0,\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \mathbf{x}_0, \boldsymbol{\lambda}) > -\infty. \quad (15)$$

Assumption 2 *For the mini-batch dataset $\mathcal{B}_i^{t,r}$ with size b at the t-th round, the associated mini-batch gradient satisfies*

$$\mathbb{E}\left[\nabla f_i(\mathbf{x}_i^{t,r}; \mathcal{B}_i^{t,r})\right] = \nabla f_i(\mathbf{x}_i^{t,r}), \quad (16)$$

$$\mathbb{E}\Big[\left\|\nabla f_i(\mathbf{x}_i^{t,r}; \mathcal{B}_i^{t,r}) - \nabla f_i(\mathbf{x}_i^{t,r})\right\|^2\Big] \leq \phi^2, \quad (17)$$

for all $t, r \leq Q_i^t$, where Q_i^t is yielded by Algorithm 1, and the upper bound $\phi^2 \to 0$ as $b \to |\mathcal{D}_i|$ [12].

Algorithm 1: Proposed FedPDM-DP

Input: System parameters b, T, ν, ρ, η^t, K.
Initialize $\mathbf{x}_0^0$, $\{\mathbf{x}_i^{0,0}\}$, $\{\boldsymbol{\lambda}_i^0\}$, and $\mathcal{S}_0 = [N]$.
for $t = 0, 1, \ldots, T-1$ **do**
  Client side:
  for $i \in \mathcal{S}_t$ in parallel **do**
    for $r = 0, 1, \ldots$ **do**
      Sample $\mathcal{B}_i^{t,r}$ from $\mathcal{D}_i$ without replacement.
      Update $\mathbf{x}_i^{t,r+1}$ using (10a).
      if (11) is satisfied **then**
        set $Q_i^t = r + 1$ and go to line 13.
      end if
    end for
    Update $\mathbf{x}_i^{t+1}$ by (10b).
    Update $\boldsymbol{\lambda}_i^{t+1}$ by (10c).
    Update $\mathbf{y}_i^{t+1}$ by (10d).
    Compute $\widetilde{\mathbf{y}}_i^{t+1}$ by (12) and send it to the PS.
  end for
  Server side:
  Update $\mathbf{x}_0^{t+1}$ by (13).
  Update the subset of clients $\mathcal{S}_{t+1} \subseteq [N]$ through randomly sampling without replacement, and then broadcast $\mathbf{x}_0^{t+1}$ to all the clients.
end for

Assumption 3 *The mini-batch gradients $\nabla f_i(\mathbf{x}_i^{t,r}; \mathcal{B}_i^{t,r}), \forall t, r$, are bounded, i.e., $\|\nabla f_i(\mathbf{x}_i^{t,r}; \mathcal{B}_i^{t,r})\| \leq G$.*

B. Privacy Analysis

Theorem 1 *With the noise vector $\boldsymbol{\xi}_i^t \sim \mathcal{N}\big(\mathbf{0}, \sigma_{i,t}^2 \mathbf{I}_d\big)$ added to $\mathbf{y}_i^t$ (cf. (12)), the minimum noise variance $\sigma_{i,t}^2$ for guaranteeing (ϵ, δ)-DP for client i at the t-th communication round is given by*

$$\sigma_{i,t}^2 = \frac{2s_{i,t}^2 \ln\big(1.25/\delta\big)}{\epsilon^2}, \quad (18)$$

where

$$s_{i,t} = \begin{cases} \frac{1-|1-\eta^t\rho|^{Q_i^t}}{1-|1-\eta^t\rho|} 4\eta^t G, & \text{if } \eta^t\rho \neq 2, \\ 4\eta^t Q_i^t G, & \text{otherwise.} \end{cases} \quad (19)$$

Proof: Equation (19) can be derived from the definition of ℓ_2-sensitivity of $\mathbf{y}_i^t$ [14]. The detailed proof of Theorem 1 is omitted due to space limitation. ■

Theorem 1 shows that $\sigma_{i,t}^2$ is larger for larger Q_i^t, though a larger Q_i^t can improve the communication efficiency [6]. Hence, the value of Q_i^t (yielded by Algorithm 1) determines the trade-off between privacy protection and communication efficiency.

The total privacy loss of client i yielded by Algorithm 1, denoted as $\bar{\epsilon}_i^T$, which is the sum of all the privacy losses (cf. (7)) over T communication rounds, can be estimated [14] by multiple methods, e.g., the moments accountant method [17]. By the moments accountant method, we have obtained a

new result on the achievable lower bound of $\bar{\epsilon}_i^T$ given in the following theorem.

Theorem 2 *Let $p_i = K/N$ and $q_i = Q_i^t b/|\mathcal{D}_i|$, that denote the participation fraction of clients, and the fraction of data used by Algorithm 1, respectively. Under the (ϵ, δ)-DP for client i after T communication rounds, an achievable total privacy loss $\bar{\epsilon}_i^T$ is given by*

$$\bar{\epsilon}_i^T = c_0 q_i^2 \epsilon \sqrt{\frac{p_i T}{1-q_i}}, \forall i \in [N], \tag{20}$$

where $c_0 > 0$ is a constant dependent upon δ.

Proof: The proof basically follows that of Theorem 1 reported in [6] for the case of data sampling without replacement. The details are omitted due to space limitation. ■

Theorem 2 shows how the mechanisms of client sampling and data sampling impact on $\bar{\epsilon}_i^T, \forall i$. The larger value of $\bar{\epsilon}_i^T$, the larger value of ϵ by (20), implying weaker privacy protection but better learning performance. This will be justified in our experimental results later.

C. Convergence Analysis

Motivated by [9], the quantity $P(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\})$ used as convergence performance measure is defined by

$$P(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\}) \triangleq \sum_{j=1}^{N} \Big[\big\| \nabla_{\mathbf{x}_j} \mathcal{L}(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\}) \big\|^2 + \big\| \nabla_{\boldsymbol{\lambda}_j} \mathcal{L}(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\}) \big\|^2 \Big] + \big\| \nabla_{\mathbf{x}_0} \mathcal{L}(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\}) \big\|^2. \tag{21}$$

It can be verified that if $P(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\}) \to 0$ as t increases, then a stationary-point solution to (2) can be obtained [12].

Theorem 3 *Suppose that Assumptions 1-3 hold and $2\sqrt{5} - 4 \leq L \leq \rho/4$. Then, the following inequality holds:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[P(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\})\big] \leq \zeta \triangleq \underbrace{\frac{(\mathcal{L}(\{\mathbf{x}_i^0\}, \mathbf{x}_0^0, \{\boldsymbol{\lambda}_i^0\}) - \underline{f})C_0}{T}}_{A_0} + \underbrace{C_1\nu}_{A_1} + \underbrace{C_2\phi^2}_{A_2} + \underbrace{C_3\sigma^2}_{A_3}, \tag{22}$$

where ν, ϕ^2 have been defined in (11) *and Assumption 2, respectively; $\sigma^2 \triangleq \max_{i,t} \sigma_{i,t}^2$; C_0, C_1, C_2, C_3 are constants depending on system parameters ρ and L (cf.* (9b)*,* (14)*).*

Proof: The proof is given in Appendix B. ■

Let us conclude this section with the following two remarks based on Theorem 1 and Theorem 3.

Remark 1 The convergence performance bound in Theorem 3 is dependent upon A_0, A_1, A_2, A_3. Specifically, i) $A_0 \to 0$ as T increases; ii) $A_1 \to 0$ as $\nu \to 0$; iii) A_2 can be made arbitrarily small for b large enough since $\phi^2 \to 0$ as $b \to |\mathcal{D}_i|$ by (17); iv) A_3 can be reduced by letting $\eta^t \rho \to 1$ or increasing ϵ by Theorem 1.

Remark 2 (**Communication complexity**). By Theorem 3, to achieve a ζ-stationary solution under the following parameter setting:

$$T = 4(\mathcal{L}(\{\mathbf{x}_i^0\}, \mathbf{x}_0^0, \{\boldsymbol{\lambda}_i^0\}) - \underline{f})C_0/\zeta, \tag{23a}$$

$$\nu = \zeta/(4C_1), \phi^2 = \zeta/(4C_2), \sigma^2 = \zeta/(4C_3), \tag{23b}$$

where $\sigma^2 \triangleq \max_{i,t} \sigma_{i,t}^2$. It can be inferred that (23b) can be achieved by Remark 1. Finally, we would like to emphasize that the required T is in $\mathcal{O}(1/\zeta)$ by (23a), while to the best of our knowledge, most of state-of-the-art FL algorithms can only attain T in $\mathcal{O}(1/\zeta^2)$ for non-convex problems [18].

V. Experimental Results and Discussions

A. Experiment Setting

We evaluate the performance of the proposed FedPDM-DP by considering the commonly known non-convex and non-smooth logistic regression problem [12] with the loss function

$$F(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{X}) + \gamma \|\mathbf{X}\|_1, \tag{24}$$

where $\|\mathbf{X}\|_1 = \sum_{j=1}^{m} \sum_{k=1}^{n} |[\mathbf{X}]_{jk}|$, $\mathbf{X} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_m^\top]^\top \in \mathbb{R}^{m \times n}$, and $\gamma > 0$ denotes the regularization parameter, and

$$f_i(\mathbf{X}) = \frac{1}{|\mathcal{D}_i|} \sum_{j=1}^{|\mathcal{D}_i|} \Big[\ln\big(1 + \exp(-\mathrm{Tr}(\mathbf{X}\mathbf{X}_{ij}^\top)\big) \Big] + \beta \sum_{j,k} \frac{[\mathbf{X}]_{jk}^2}{1 + [\mathbf{X}]_{jk}^2}, \tag{25}$$

in which $\mathbf{X}_{ij} = \boldsymbol{a}_{ij} \boldsymbol{b}_{ij}^\top$ is sparse, since $\boldsymbol{a}_{ij} \in \mathbb{R}^n$ represents the feature vector and $\boldsymbol{b}_{ij} \in \mathbb{R}^m$ satisfying $\|\boldsymbol{b}_{ij}\|_0 = 1$ denotes the label vector of the j-th sample in $\mathcal{D}_i$. The concavity of f_i increases as β increases. Note that the regularizer $\|\mathbf{X}\|_1$ (convex envelope of $\|\mathbf{X}\|_0$ [15]) is non-smooth for the sparse learned model.

In the testing stage, for the given feature vector, denoted as $\boldsymbol{a}' \in \mathbb{R}^n$, of an unknown class (whose true class number is given by $k' = \arg\max_k \{b'_k, k \in [m]\}$), where b'_k is the k-th element of the true label vector $\boldsymbol{b}' \in \mathbb{R}^m$. We compute the following softmax function [6]

$$z_k = \frac{\exp(\mathbf{x}_k^\top \boldsymbol{a}')}{\sum_{j=1}^{m} \exp(\mathbf{x}_j^\top \boldsymbol{a}')}, \forall k \in [m]. \tag{26}$$

Then the correct decision is made if $\arg\max_k \{z_k\} = k'$. Finally, the overall testing accuracy is obtained as the correct classification rate over the testing dataset.

Datasets: The benchmark dataset MNIST [19] is used, which consists of 60,000 training samples and 10,000 testing samples for which $(m, n) = (10, 785)$. In our experiment, all the training samples are uniformly distributed over $N = 100$ clients.

Parameter setting: The values of system parameters used are $K = 5$, $b = 30$, $\rho = 10$, $\eta^t = 1/\sqrt{1+t}$, $T = 300$, $\delta = 10^{-5}$, $\nu = 10^{-2}$, $\gamma = 10^{-4}$, and $\beta = 10^{-2}$.

Benchmark algorithms: Some state-of-the-art algorithms, including FedAvg-DP [5], SCAFFOLD-DP [20] and

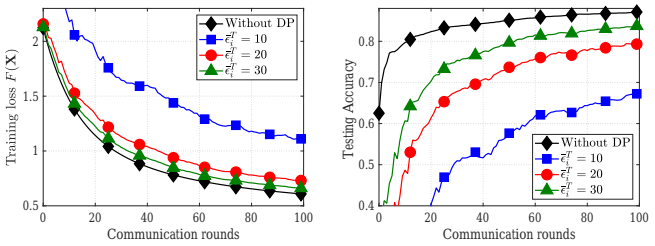

Fig. 1: Performance of the proposed FedPDM-DP for (a) training loss and (b) testing accuracy under different values of $\bar{\epsilon}_i^T \in \{10, 20, 30\}\, \forall i$ and "Without DP" (i.e., $\boldsymbol{\xi}_i^t = \mathbf{0}$).

FedProx-DP [21], are also tested for performance comparison with the proposed FedPDM-DP algorithm.

B. Simulation Results

Impact of total privacy loss $\bar{\epsilon}_i^T$: Figure 1 shows the experimental results of the proposed FedPDM-DP algorithm. One can observe that the training loss (testing accuracy) decreases (increases) with communication rounds; its learning performance is better along with faster convergence rate for larger $\bar{\epsilon}_i^T$ (i.e., weaker privacy protection level).

Performance comparison with benchmark algorithms: Figure 2 shows the experimental results for performance comparison of all the algorithms under test. The performance trends w.r.t. communication rounds are consistent with the observations in Figure 1. Furthermore, the proposed FedPDM-DP significantly outperforms the other benchmark algorithms for both cases ("without DP" case and the case of $\bar{\epsilon}_i^T = 7$). We would like to emphasize that the performance gap between these two cases for the proposed FedPDM-DP is much smaller than those of the other algorithms (cf. Figs. (2a) and (2c) together with (2b) and (2d)), thanks to $\|\mathbf{X}\|_1$ used in the loss function, implying stronger robustness against the added artificial noise for our algorithm.

VI. Conclusions

We have presented a privacy-preserving primal-dual algorithm (i.e., FedPDM-DP) for the FL problem involving a non-convex loss function and a non-smooth regularizer. In addition to analytical results presented for the proposed FedPDM-DP, some experimental results were also provided to demonstrate its superior learning performance over some existing benchmark algorithms and its robustness against the added artificial noise, as well as consistency with all the analytical results.

VII. 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 *Proc. Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

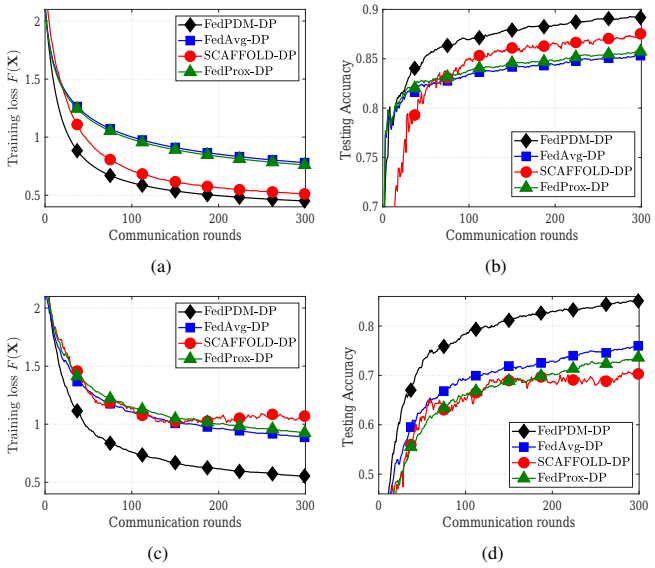

Fig. 2: Performance comparison between the proposed FedPDM-DP and benchmark algorithms, in terms of training loss in (a) and (c), and testing accuracy in (b) and (d); (a) and (b) are for the case of "without DP" case, and (c) and (d) with $\bar{\epsilon}_i^T = 7$.

[2] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. International Conference on Learning Representations (ICLR)*, 2020, pp. 1–26.

[3] Y. Li, S. Wang, C.-Y. Chi, and T. Q. Quek, "Differentially private federated learning in edge networks: The perspective of noise reduction," *IEEE Network*, vol. 36, no. 5, pp. 167–172, 2022.

[4] P. Kairouz, H. B. McMahan, B. Avent, Bellet *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[5] Y. Li, T.-H. Chang, and C.-Y. Chi, "Secure federated averaging algorithm with differential privacy," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6.

[6] Y. Li, S. Wang, T.-H. Chang, and C.-Y. Chi, "Federated stochastic primal-dual learning with differential privacy," *arXiv preprint arXiv:2204.12284*, 2022.

[7] M. Hong and T.-H. Chang, "Stochastic proximal gradient consensus over random networks," *IEEE Trans. Signal Processing*, vol. 65, pp. 2933–2948, 2017.

[8] D. Hajinezhad, M. Hong, T. Zhao, and Z. Wang, "NESTT: A nonconvex primal-dual splitting method for distributed and stochastic optimization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 3207–3215.

[9] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

[10] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," *arXiv preprint arXiv:1802.07876*, 2018.

[11] J. Ding, S. M. Errapotu, H. Zhang, Y. Gong, M. Pan, and Z. Han, "Stochastic ADMM based distributed machine

learning with differential privacy," in *Proc. International Conference on Security and Privacy in Communication Systems*, 2019, pp. 257–277.

[12] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "FedPD: A federated learning framework with adaptivity to non-IID data," *IEEE Trans. Signal Processing*, vol. 69, pp. 6055–6070, 2021.

[13] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Springer Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 2006, pp. 486–503.

[14] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[15] C.-Y. Chi, W.-C. Li, and C.-H. Lin, *Convex Optimization for Signal Processing and Communications: From Fundamentals to Applications*. CRC Press, Boca Raton, FL, Feb. 2017.

[16] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[18] M. Noble, A. Bellet, and A. Dieuleveut, "Differentially private federated learning on heterogeneous data," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2022, pp. 10 110–10 145.

[19] Y. LeCun, C. Cortes, and C. Burges. The MNIST database. [Online]. Available: http://yann.lecun.com/exdb/mnist.

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

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

APPENDIX A
PROOF OF (13)

According to (8c), we have

$$\begin{aligned}
\mathbf{x}_0^{t+1} &= \arg\min_{\mathbf{x}_0} \mathcal{L}\big(\mathbf{x}^{t+1}, \mathbf{x}_0, \boldsymbol{\lambda}^{t+1}\big) \\
&\overset{(a)}{=} \arg\min_{\mathbf{x}_0} h(\mathbf{x}_0) + \sum_{i=1}^{N} \Big({\boldsymbol{\lambda}_i^{t+1}}^{\top}(\mathbf{x}_0 - \mathbf{x}_i^{t+1}) + \frac{\rho}{2}\left\|\mathbf{x}_0 - \mathbf{x}_i^{t+1}\right\|^2 \Big) \\
&\overset{(b)}{=} \arg\min_{\mathbf{x}_0} h(\mathbf{x}_0) + \frac{\rho}{2}\sum_{i=1}^{N} \Big(\big\|\frac{1}{\rho}\boldsymbol{\lambda}_i^{t+1}\big\|^2 + \frac{2}{\rho}{\boldsymbol{\lambda}_i^{t+1}}^{\top}(\mathbf{x}_0^t - \mathbf{x}_i^{t+1}) \\
&\quad + \left\|\mathbf{x}_0 - \mathbf{x}_i^{t+1}\right\|^2 \Big) \\
&\overset{(c)}{=} \arg\min_{\mathbf{x}_0} h(\mathbf{x}_0) + \frac{\rho}{2}\big\|\mathbf{x}_0 - \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^{t+1} - \frac{1}{\rho}\boldsymbol{\lambda}_i^{t+1})\big\|^2 \\
&= \mathrm{prox}_{\rho h}\Big(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^{t+1} - \frac{1}{\rho}\boldsymbol{\lambda}_i^{t+1})\Big), \qquad \text{(A.1)}
\end{aligned}$$

where (a), (b) and (c) hold because $f_i(\mathbf{x}_i)$, $\mathbf{x}_i$, $\boldsymbol{\lambda}_i$, $\forall i$ are constants w.r.t. $\mathbf{x}_0$. ■

APPENDIX B
PROOF OF THEOREM 3

To prove the Theorem 3, we need the following two lemmas and their proofs are omitted here due to the space limitation.

Lemma 2 *Suppose that Assumptions 1-2 hold and* $2\sqrt{5} - 4 \le L \le \rho/4$. *Then,*

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{L}(\{\mathbf{x}_i^{t+1}\}, \mathbf{x}_0^{t+1}, \{\boldsymbol{\lambda}_i^{t+1}\}) - \mathcal{L}(\{\mathbf{x}_i^{t}\}, \mathbf{x}_0^{t}, \{\boldsymbol{\lambda}_i^{t}\})\Big] \\
&\quad - 2\Big[(\frac{8}{\rho} + \frac{L}{4L-2})\nu + \frac{16}{\rho}\phi^2\Big] \\
&\le -\underbrace{\frac{N\rho}{2}}_{B_1}\|\mathbf{x}_0^t - \mathbf{x}_0^{t+1}\|^2 + \sum_{i\in\mathcal{S}_t}\Big[-\underbrace{(\frac{\rho}{2} - \frac{4L^2}{\rho} - L)}_{B_2}\|\mathbf{x}_i^t - \mathbf{x}_i^{t+1}\|^2 \\
&\quad - \underbrace{(\frac{8}{\rho} + \frac{L}{4L-2})}_{B_3}\nu - \underbrace{\frac{16}{\rho}}_{B_4}\phi^2\Big]. \qquad \text{(B.1)}
\end{aligned}$$

Lemma 3 *Suppose that Assumptions 1-2 hold. Then,*

$$\begin{aligned}
&\mathbb{E}\Big[P(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\})\Big] \\
&\le \underbrace{2}_{B_5}\|\mathbf{x}_0^t - \mathbf{x}_0^{t+1}\|^2 + \sum_{i\in\mathcal{S}_t}\Big\{\underbrace{(1 + \frac{4\rho}{N})\frac{16}{\rho^2}}_{B_6}\phi^2 \\
&\quad + \underbrace{\big[(L^2+\rho^2) + \frac{4}{N}(\rho + \frac{4L^2}{\rho}) + \frac{4L^2}{\rho^2}\big]}_{B_7}\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|^2 \\
&\quad + \underbrace{(\frac{L^2}{4L-2} + (1 + \frac{4\rho}{N})\frac{8}{\rho^2})}_{B_8}\nu\Big\} + 8\rho\sigma^2. \qquad \text{(B.2)}
\end{aligned}$$

With $P(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\})$ defined in (21), we have the following inferences:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[P(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\})\big] \\
&\overset{(a)}{\le} \frac{1}{T}\sum_{t=0}^{T-1}\theta_2\Big\{\|\mathbf{x}_0^t - \mathbf{x}_0^{t+1}\|^2 + \sum_{i\in\mathcal{S}_t}\big(\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|^2 + \nu + \phi^2\big)\Big\} \\
&\quad + 8\rho\sigma^2 \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\frac{\theta_1\theta_2}{\theta_1}\Big\{\|\mathbf{x}_0^t - \mathbf{x}_0^{t+1}\|^2 + \sum_{i\in\mathcal{S}_t}\big(\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|^2 + \nu + \phi^2\big)\Big\} \\
&\quad + 8\rho\sigma^2 \\
&\overset{(b)}{\le} \frac{1}{T}\sum_{t=0}^{T-1}\frac{\theta_2}{\theta_1}\Big\{\mathbb{E}\big[\mathcal{L}(\{\mathbf{x}_i^t\}, \mathbf{x}_0^t, \{\boldsymbol{\lambda}_i^t\}) - \mathcal{L}(\{\mathbf{x}_i^{t+1}\}, \mathbf{x}_0^{t+1}, \{\boldsymbol{\lambda}_i^{t+1}\})\big]\Big\} \\
&\quad + C_1\nu + C_2\phi^2 + C_3\sigma^2 \\
&\overset{(c)}{\le} C_0\frac{\big(\mathcal{L}(\{\mathbf{x}_i^0\}, \mathbf{x}_0^0, \{\boldsymbol{\lambda}_i^0\}) - \underline{f}\big)}{T} + C_1\nu + C_2\phi^2 + C_3\sigma^2 \qquad \text{(B.3)}
\end{aligned}$$

which is exactly (22), where (a), (b), (c) hold due to Lemma 3, Lemma 2, and Assumption 1, respectively; $\theta_1 = \min\{B_1, B_2, B_3, B_4\}$, $\theta_2 = \max\{B_5, B_6, B_7, B_8\}$; C_0, C_1, C_2, C_3 and σ^2 were defined in Theorem 3. ■