Personalized Federated Learning With Differential Privacy and Convergence Guarantee

Kang Wei, *Member, IEEE*, Jun Li, *Senior Member, IEEE*, Chuan Ma, *Member, IEEE*, Ming Ding, *Senior Member, IEEE*, Wen Chen, *Senior Member, IEEE*, Jun Wu, *Senior Member, IEEE*, Meixia Tao, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*— Personalized federated learning (PFL), as a novel federated learning (FL) paradigm, is capable of generating personalized models for heterogenous clients. Combined with a meta-learning mechanism, PFL can further improve the convergence performance with few-shot training. However, meta-learning based PFL has two stages of gradient descent in each local training round, therefore posing a more serious challenge in information leakage. In this paper, we propose a differential privacy (DP) based PFL (DP-PFL) framework and analyze its convergence performance. Specifically, we first design a privacy budget allocation scheme for inner and outer update stages based on the Rényi DP composition theory. Then, we develop two convergence bounds for the proposed DP-PFL framework under convex and non-convex loss function assumptions, respectively. Our developed convergence bounds reveal that 1) there is an optimal size of the DP-PFL model that can achieve the best convergence performance for a given privacy level, and 2) there is an optimal tradeoff among the number of communication rounds, convergence performance and privacy budget. Evaluations on various real-life datasets demonstrate that our theoretical results are consistent with experimental results. The derived theoretical results can guide the design of various DP-PFL algorithms with configurable tradeoff requirements on the convergence performance and privacy levels.**

***Index Terms*— Federated learning, meta-learning, differential privacy, convergence analysis.**

I. Introduction

IT IS expected that big data-driven artificial intelligence (AI) will soon be applied in many aspects of our daily lives, including health care [1], communications [2], [3], autonomous driving [4], etc. At the same time, the rapid growth of Internet-of-Things (IoT) applications calls for data mining and model learning securely and reliably in distributed systems. In integrating AI in a variety of IoT applications, distributed machine learning (ML) systems are preferred for data processing tasks with edge intelligence [5]. Federated learning (FL) [6], [7], [8], as a recent advance in distributed ML, has been proposed to ensure data being processed locally, thereby protecting clients' privacy.

However, due to the heterogeneity of end devices in IoT environments, e.g., non-independent and identically distributed (non-IID) data, and imbalanced computing capacity, it is not practical to train a generally efficient ML model for all the clients in the conventional FL framework [9], [10], [11]. Therefore, personalized FL (PFL) has been proposed to combat client heterogeneity, with the aim of training a personalized model for each client [12], [13]. To achieve this aim, PFL utilizes a meta-learning mechanism to first obtain a initialized model capable of rapidly adapting to new learning tasks, and then train a personalized model based on this initial model via fine-tuning [14], [15], [16], [17]. To have a better understanding on convergence performance of PFL, the work in [18] derived an upper bound of its loss function, establishing the relationship between the bound and imbalanced data distributions among clients. Furthermore, the work in [17] developed two convergence bounds for PFL based on two measurements of data distribution distances among clients, i.e., Total Variation (TV) and 1-Wasserstein distances.

Similar to conventional FL, the training process of PFL keeps personal data locally, thereby helping protect clients' privacy [19]. However, attackers can still infer private information by analyzing model parameters transmitted by the clients. As one of the well-known privacy-preserving techniques, differential privacy (DP) has been widely applied for protecting clients' sensitive information in conventional FL [20]. The work in [21] derived a convergence bound for meta-leaning with DP. The work in [22] and [23] proposed a DP based PFL (DP-PFL) framework, where personal models are trained via an alternating optimization procedure relying

Manuscript received 4 February 2022; revised 16 December 2022 and 5 May 2023; accepted 6 June 2023. Date of publication 10 July 2023; date of current version 28 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62002170 and Grant 62071296; in part by the Fundamental Research Funds for the Central Universities under Grant 30921013104; in part by the Future Network Grant of Provincial Education Board in Jiangsu; in part by the Youth Foundation Project through the Zhejiang Laboratory under Grant K2023PD0AA01; in part by the National Key Project 2020YFB1807700; in part by Shanghai under Grant 22JC1404000, Grant 20JC1416502, and Grant PKX2021-D02; and in part by the U.S. National Science Foundation under Grant CNS-2128448. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefano Tomasin. *(Corresponding author: Jun Li.)*

Kang Wei is with the School of Electrical and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the Department of Computing, Hong Kong Polytechnic University, Hong Kong 999077, China (e-mail: kang.wei@njust.edu.cn).

Jun Li is with the School of Electrical and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jun.li@njust.edu.cn).

Chuan Ma is with the Zhejiang Laboratory, Hangzhou 311121, China, and also with the Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 211189, China (e-mail: chuan.ma@zhejianglab.edu.cn).

Ming Ding is with Data61, CSIRO, Sydney, NSW 2015, Australia (e-mail: Ming.Ding@data61.csiro.au).

Wen Chen and Meixia Tao are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wenchen@sjtu.edu.cn; mxtao@sjtu.edu.cn).

Jun Wu is with the Shanghai Key Laboratory of Intelligent Information Processing and the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: wujun@fudan.edu.cn).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/TIFS.2023.3293417

on the assumption that loss functions are convex. However, this assumption is not practical since loss functions in PFL are generally non-convex.

In this paper, we propose a novel Rényi DP (RDP) based PFL (RDP-PFL) framework with meta-training and analyze theoretically the tradeoff between privacy guarantee and learning performance. Specifically, we develop two convergence bounds within the proposed framework for convex and non-convex loss function assumptions. The derived expressions for the two bounds reveal that there exist optimal values of the number of communication rounds and model size that can maximize the learning performance for a given privacy level. To the best of the authors' knowledge, this is the first work of its kind that provides a theoretical analysis of the convergence properties of PFL combined with RDP and meta-training.

Our main contributions can be summarized as follows:

- We propose a meta-learning based DP-PFL framework by introducing the concept of RDP to enhance clients' privacy. Specifically, this framework adopts RDP based SGD (RDP-SGD) training in the meta-learning process which provides an initial model that can be rapidly adapted to personalized datasets (e.g., through few-shot learning). Further, to address the possible privacy leakage in the two-stage training procedure in each meta-learning process, we have designed a privacy budget allocation scheme based on RDP composition theory.
- We study the convergence behavior of RDP-PFL and develop two bounds with convex and non-convex loss function assumptions. These two convergence bounds reveal that there exist an optimal model size and an optimal number of communication rounds for maximizing the convergence performance given a fixed privacy level.
- We evaluate the performance of RDP-PFL using a variety of datasets and settings, which demonstrates that our theoretical results are consistent with experimental results. Therefore, our analytical results are helpful for the design of privacy-preserving PFL architectures with different tradeoff requirements on convergence performance and privacy levels.

The remainder of this paper is organized as follows. In Section II, we present the PFL framework and introduce some basics of RDP. Then, we introduce our proposed DP-PFL algorithm in Section III. We derive two convergence bounds for the proposed algorithm in Section IV. Experimental results are described in Section V. Section VI reviews related works on federated meta-learning and distributed learning with DP. Finally, conclusions are drawn in Section VII. The main notation used in this paper is summarized in Tab. I.

II. Preliminaries

In this section, we first present the PFL framework and introduce some basics of RDP.

A. Framework of PFL

Let us consider a general PFL system consisting of U source clients and K target clients, in which each client contains two datasets, i.e., the query dataset and support dataset. In each communication round, each source client trains its local model based on the global one from the central server. Then, all source clients upload their trained models to the server and the server update the global model by aggregating all received local models. The local model update for one epoch contains two steps: the inner update and the outer update. The inner update is based on a one-step gradient descent of the outer update model in the previous epoch, while the outer update model is based on a one-step gradient descent of the inner update model in the current epoch.

TABLE I

Summary of Main Notation

T	The number of communication rounds
τ_0	The number of local training epochs
$\mathcal{M}$	A randomized mechanism for DP
$\mathcal{D}$, $\mathcal{D}'$	Adjacent datasets
ϵ, δ	The parameters related to the original LDP
α	The selective parameter for RDP
$\boldsymbol{I}$	An identity matrix
$\mathcal{D}_i$	The dataset held by the i-th client
$\mathcal{D}_i^{\mathrm{Q}}$	The query dataset held by the i-th client
$\mathcal{D}_i^{\mathrm{S}}$	The support dataset held by the i-th client
$F(\cdot)$	The loss function
$\boldsymbol{w}_i^{t,\tau}$	The model parameter at the t-th communiation round after τ local epochs for the i-th client
$\boldsymbol{\theta}_i^{t,\tau}$	The model parameter with one-step gradient descent based on $\boldsymbol{w}_i^{t,\tau}$ at the t-th communiation round after τ local epochs for the i-th client
$\boldsymbol{n}_i^{\mathrm{Q}}$	The noise vector drawn from $\mathcal{N}(0, C^2\sigma^2)$ for the inner update
$\boldsymbol{n}_i^{\mathrm{S}}$	The noise vector drawn from $\mathcal{N}(0, C^2\sigma^2)$ for the outer update
C	The clipping threshold
S	The number of model parameters (model size)
η	The learning rate for the inner update
β	The learning rate for the outer update

We denote by $\boldsymbol{\theta}_i^{t,\tau}$ and $\boldsymbol{w}_i^{t,\tau}$ the inner and outer update models, respectively, in the τ-th local epoch of the t-th communication round for the i-th source client, $i \in \mathcal{U} \triangleq \{1, 2, \ldots, U\}$. The inner update in the τ-th local epoch of the t-th communication round over query dataset $\mathcal{D}_i^{\mathrm{Q}}$ can be expressed as

$$\boldsymbol{\theta}_i^{t,\tau} = \boldsymbol{w}_i^{t,\tau-1} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau-1}, \mathcal{D}_i^{\mathrm{Q}}), \tag{1}$$

where $F_i(\cdot)$ is the loss function of the i-th client and η is the learning rate. Generally, the loss function $F_i(\cdot)$ is given by the empirical risk and has the same expression for various clients. Then the outer update in the τ-th local epoch of the t-th communication round over support dataset $\mathcal{D}_i^{\mathrm{S}}$ can be expressed as

$$\begin{aligned}\boldsymbol{w}_i^{t,\tau} &= \boldsymbol{w}_i^{t,\tau-1} - \beta \nabla F_i(\boldsymbol{\theta}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}) \\ &= \boldsymbol{w}_i^{t,\tau-1} - \beta(\boldsymbol{I} - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}))\nabla F_i(\boldsymbol{\theta}_i^{t,\tau+1}, \mathcal{D}_i^{\mathrm{S}}),\end{aligned} \tag{2}$$

where β is the learning rate for the outer update.

When all clients complete the local training, they need to upload their local models to the central server. The server performs the global aggregation to obtain a global model parameter $\boldsymbol{w}^{t+1}$, which can be expressed as

$$\boldsymbol{w}^{t+1} = \sum_{i \in \mathcal{U}} p_i \boldsymbol{w}_i^{t,\tau_0}, \tag{3}$$

where τ_0 is the number of local training epochs, $p_i \triangleq |\mathcal{D}_i^{\mathrm{S}}|/|\mathcal{D}^{\mathrm{S}}| \geq 0$ with $\sum_{i \in \mathcal{U}} p_i = 1$, $|\mathcal{D}_i^{\mathrm{S}}|$ represents the size of the support dataset $\mathcal{D}_i^{\mathrm{S}}$, and $|\mathcal{D}^{\mathrm{S}}| = \sum_{i \in \mathcal{U}} |\mathcal{D}_i^{\mathrm{S}}|$ represents the size of all support datasets.

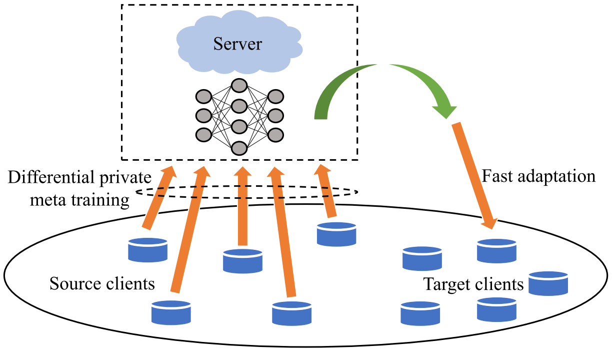


Fig. 1. The diagram of differentially private PFL.

Given a global model $\boldsymbol{w}^t$, the j-th target client can obtain its personalized model $\boldsymbol{\theta}_j$ through one-step gradient descent on $\boldsymbol{w}^t$, $j \in \mathcal{K} \triangleq \{1, 2, \dots, K\}$, which can be expressed as

$$\boldsymbol{\theta}_j = \boldsymbol{w}^t - \eta \nabla F_j(\boldsymbol{w}^t, \mathcal{D}_j), \tag{4}$$

where $\mathcal{D}_j$ is the dataset of the j-th target client.

B. Rényi Differential Privacy

DP with parameters ϵ and δ provides a criterion for privacy protection in distributed data processing systems. Here, $\epsilon > 0$ is the distinguishable bound of all outputs on neighboring datasets $\mathcal{D}, \mathcal{D}'$ in a database, and δ represents the probability of the event that the ratio of the probabilities for two adjacent datasets $\mathcal{D}, \mathcal{D}'$ cannot be bounded by e^{ϵ} after adding a privacy-preserving mechanism.

In this paper, we consider an improved DP definition called RDP, which is strictly stronger than (ϵ, δ)-DP for $\delta > 0$ and allows tighter composition analysis [24]. We formally define RDP as follows.

Definition 1 ((α, ϵ)-RDP [24]): Given a real number $\alpha \in (1, +\infty)$ and privacy budget ϵ, a randomized mechanism $\mathcal{M}$ satisfies (α, ϵ)-RDP for any two adjacent datasets $\mathcal{D}, \mathcal{D}'$, we have

$$D_\alpha[\mathcal{M}(\mathcal{D}) \| \mathcal{M}(\mathcal{D}')] := \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{\mathcal{M}(\mathcal{D})}{\mathcal{M}(\mathcal{D}')}\right)^{\alpha}\right] \le \epsilon, \tag{5}$$

where the expectation is taken over the output of $\mathcal{M}(\mathcal{D})$ and α is a selective parameter. We can note that RDP is a generalization of (ϵ, δ)-DP that adopts Rényi divergence as a distance metric between two distributions. It can be shown that pure (ϵ, δ)-DP is equivalent to (∞, ϵ)-RDP, and, further, that if a model $\mathcal{M}$ satisfies (α, ρ)-RDP, then $\mathcal{M}$ also satisfies $\left(\epsilon + \frac{\log \frac{1}{\delta}}{\alpha - 1}, \delta\right)$-DP for any $\delta \in (0, 1)$. Deep learning models attain RDP guarantees via two alterations to the training process, i.e., the clipping of gradients, and the addition of Gaussian noise to gradients, as known as DP-SGD.

III. Differentially Private Personalized Federated Learning

A. RDP Based PFL

We can note that each client needs to upload its local model $\boldsymbol{w}_i^{t,\tau_0}$, $i \in \mathcal{U}$, to the server. This poses threats on clients' privacy as potential adversaries may reveal sensitive information about individual clients from $\boldsymbol{w}_i^{t,\tau_0}$. Hence, in RDP-PFL, each client performs inner and outer updates by the DP-SGD mechanism. We first use a clipping threshold C to bound the L_2-norm of training gradients in the inner and outer updates. Then we perturb these two clipped gradients by Gaussian noise vectors $\boldsymbol{n}_i^{\mathrm{Q}}$ and $\boldsymbol{n}_i^{\mathrm{S}}$, respectively, where each element in the noise vectors follows the Gaussian distribution $\mathcal{N}(0, C^2\sigma^2)$. Here, σ is the noise standard deviation (SD), which is determined by the privacy budget shown in the next subsection. Specifically, detailed steps to update the inner and outer models for each local epoch are shown as follows.

1) Inner Update: Compute the local gradient and perform the inner update.

$$\nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}) = \frac{\nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}})}{\max\left\{1, \frac{\|\nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}})\|}{C}\right\}}, \tag{6}$$

$$\tilde{\boldsymbol{\theta}}_i^{t,\tau+1} = \boldsymbol{w}_i^{t,\tau} - \eta\left(\nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}) + \boldsymbol{n}_i^{\mathrm{Q}}\right), \tag{7}$$

where $\tilde{\boldsymbol{\theta}}_i^{t,\tau+1}$ represents the inner update mode in RDP-PFL.

2) Outer Update: Compute the local gradient and perform the outer update.

$$\boldsymbol{w}_i^{t,\tau+1} = \boldsymbol{w}_i^{t,\tau} \tag{8}$$

$$- \beta\left(\frac{\boldsymbol{g}_1}{\max\left\{1, \frac{\|\boldsymbol{g}_1\|}{\sqrt{C}}\right\}} \frac{\boldsymbol{g}_2}{\max\left\{1, \frac{\|\boldsymbol{g}_2\|}{\sqrt{C}}\right\}} + \boldsymbol{n}_i^{\mathrm{S}}\right), \tag{9}$$

where

$$\boldsymbol{g}_1 = \boldsymbol{I} - \eta \nabla^2 F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}\right), \tag{10}$$

$$\boldsymbol{g}_2 = \nabla F_i\left(\tilde{\boldsymbol{\theta}}_i^{t,\tau+1}, \mathcal{D}_i^{\mathrm{S}}\right), \tag{11}$$

and $\boldsymbol{I}$ is an identity matrix. Eqs. (10) and (11) are derived by Eq. (2). Note that the source clients can use partial data to train their models in each local training epoch for the inner and outer updates with sampling rates q^{Q} and q^{S}, respectively.

The RDP-PFL process completes after the number of communication rounds reaches a preset number T. We summarize the detailed steps of RDP-PFL in **Algorithm 1**. The Fig. 1 shows the diagram of RDP-PFL. Since the inner and outer models are trained based on query and support datasets, respectively, we need to control the norm of gradients, and then add Gaussian noise to gradients in the inner and outer updates. In addition, we will calculate the accumulation of privacy leakage (privacy budget) for both the inner and outer updates to obtain the SD of the Gaussian noise in the following subsection.

B. Privacy Analysis

It can be noted that privacy budget for multiple access to the training data can be measured by the RDP technique. Based on **Definition 5**, we demonstrate that the privacy budget for $\mathcal{D}_i^{\mathrm{Q}}$ or $\mathcal{D}_i^{\mathrm{S}}$ can be calculated as follows.

Theorem 1: For the meta-training process, the privacy budget ϵ with a given δ for the published local model at each update can be written as

$$\epsilon = \frac{T\tau_0}{\alpha - 1}(\log I^{\mathrm{Q}} + \log I^{\mathrm{S}}) + \frac{\log(\frac{1}{\delta}) + (\alpha - 1)\log(1 - \frac{1}{\alpha}) - \log(\alpha)}{\alpha - 1}, \tag{12}$$

Algorithm 1 Rényi Differential Privacy Based PFL

```
Data: The maximum number of communication rounds $T$, the number of local training epochs
      $\tau_0$, noise SD $\sigma$, clipping threshold $C$, initial model $\boldsymbol{w}^0$;
for $t$: 0 to $T-1$ do
    Server execute:
    Server sends global model $\boldsymbol{w}^t$ to all clients in $\mathcal{U}$;
    Client execute:
    for all $i \in \mathcal{U}$ do
        Update the local model: $\boldsymbol{w}_i^{t,0} = \boldsymbol{w}^t$;
        for $\tau$: 0 to $\tau_0 - 1$ do
            Perform the inner update by Eqs. (6) and (7);
            Perform the outer update by Eqs. (8) and (9);
        Upload the local model $\boldsymbol{w}_i^{t,\tau_0}$ to the server;
    Server execute:
    Aggregate all local models by (3);
    $t \leftarrow t+1$;
Result: $\boldsymbol{w}^T$
```

where

$$I^{\mathrm{Q}} = \int_{-\infty}^{\infty} \mu_0^{\mathrm{Q}}(z)\left((1-q^{\mathrm{Q}}) + \frac{q^{\mathrm{Q}}\mu_1^{\mathrm{Q}}(z)}{\mu_0^{\mathrm{Q}}(z)}\right)^{\alpha}, \tag{13}$$

$$I^{\mathrm{S}} = \int_{-\infty}^{\infty} \mu_0^{\mathrm{S}}(z)\left((1-q^{\mathrm{S}}) + \frac{q^{\mathrm{S}}\mu_1^{\mathrm{S}}(z)}{\mu_0^{\mathrm{S}}(z)}\right)^{\alpha}. \tag{14}$$

In Eq. (13), $\mu_0^{\mathrm{Q}}(z)$ and $\mu_1^{\mathrm{Q}}(z)$ denote the probability density function (PDF) of the Gaussian distribution $\mathcal{N}(0,\sigma)$ and the PDF of a mixture of two Gaussian distributions $q^{\mathrm{Q}}\mathcal{N}(1,\sigma) + (1-q^{\mathrm{Q}})\mathcal{N}(0,\sigma)$, respectively. In Eq. (14), $\mu_0^{\mathrm{S}}(z)$ denotes the $\mathcal{N}(0,\sigma)$ PDF, and $\mu_1^{\mathrm{S}}(z)$ denotes the PDF of a mixture of two Gaussian distributions $q^{\mathrm{S}}\mathcal{N}(1,\sigma) + (1-q^{\mathrm{S}})\mathcal{N}(0,\sigma)$.

Proof: Please see Appendix A. ■

According to Theorem 1, we can note that **Algorithm 1** satisfies the (ϵ,δ)-DP by selecting a proper SD σ. Specifically, we obtain the noise SD based on Eq. (12) via the searching method, in which this noise SD needs to make the accumulation of privacy budget no more than the required one. In RDP-PFL, we adopt the clipping-and-noising technique for inner and outer updates, in which each update will consume part of the privacy budget. Given sampling rates q^{Q} and q^{S}, Theorem 1 presents the privacy budgets allocated to inner and outer updates in RDP-PFL, i.e., $\log I^{\mathrm{Q}}$ and $\log I^{\mathrm{S}}$, respectively.

IV. Convergence Analysis

In this section, we present our theoretical results on the convergence performance of RDP-PFL with convex and non-convex loss function assumptions. Before that, we first mention four customary assumptions required for both convex and non-convex loss function assumptions.

Assumption 1: For any i and $\boldsymbol{w} \in \mathbb{R}^d$, the gradient of $F_i(\boldsymbol{w})$ is bounded by a non-negative constant B, i.e., $\|\nabla F_i(\boldsymbol{w})\| \le B$.

Assumption 2: The loss function $F_i(\boldsymbol{w})$, and its gradient and Hessian are Lipschitz continuous, i.e., for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, existing constants λ, L and ρ,

$$\|F_i(\boldsymbol{w}) - F_i(\boldsymbol{w}')\| \le \lambda\|\boldsymbol{w} - \boldsymbol{w}'\|, \tag{15}$$

$$\|\nabla F_i(\boldsymbol{w}) - \nabla F_i(\boldsymbol{w}')\| \le L\|\boldsymbol{w} - \boldsymbol{w}'\|, \tag{16}$$

$$\|\nabla^2 F_i(\boldsymbol{w}) - \nabla^2 F_i(\boldsymbol{w}')\| \le \rho\|\boldsymbol{w} - \boldsymbol{w}'\|. \tag{17}$$

Assumption 3: The gradient and Hessian of the loss function $F_i(\boldsymbol{w})$ satisfy the following conditions:

$$\|\nabla F_i(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \le \varepsilon_i \; and \; \varepsilon = \sum_{i\in\mathcal{U}} p_i\varepsilon_i, \tag{18}$$

$$\|\nabla^2 F_i(\boldsymbol{w}) - \nabla^2 F(\boldsymbol{w})\| \le \gamma_i, \; and \; \gamma = \sum_{i\in\mathcal{U}} p_i\gamma_i, \tag{19}$$

for any $\boldsymbol{w} \in \mathbb{R}^d$, where $F(\boldsymbol{w}) \triangleq \sum_{i\in\mathcal{U}} p_i F_i(\boldsymbol{w})$, ε_i and γ_i are constants.

Assumption 4: For any client i, a data sample $\boldsymbol{x}$ and $\boldsymbol{w} \in \mathbb{R}^d$, the gradient $\nabla F_i(\boldsymbol{w})$ and Hessian $\nabla^2 F_i(\boldsymbol{w})$ have bounded variances, i.e.,

$$\mathbb{E}\left\{\|\nabla F_i(\boldsymbol{w},\boldsymbol{x}) - \nabla F_i(\boldsymbol{w})\|^2\right\} \le \sigma_G^2, \tag{20}$$

$$\mathbb{E}\left\{\|\nabla^2 F_i(\boldsymbol{w},\boldsymbol{x}) - \nabla^2 F_i(\boldsymbol{w})\|^2\right\} \le \sigma_H^2, \tag{21}$$

where σ_G^2 and σ_H^2 are two variances.

Assumption 1 ensures that the gradient can be bounded by B, where gradient clipping is a popular ingredient of ML. Assumption 2 implies that the local loss function $F_i(\boldsymbol{w})$ as well as its gradient and Hessian are Lipschitz continuous with constants λ, L and ρ, respectively. The conditions in Assumption 3 and Assumption 4 on the bias and variance of stochastic gradients and Hessian are also customary. We can see that Assumptions 1-4 are widely adopted in the theoretical analysis for convergence bounds [16], [17].

To characterize the convergence behavior of RDP-PFL, we first examine the structural properties of the meta-learning objective function $G_i(\boldsymbol{w})$, which is defined as $G_i(\boldsymbol{w}) \triangleq F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w}))$. Based on Assumptions 1-3, we can obtain following lemmas about the meta-learning objective function.

Lemma 1: The gradient of the meta-learning objective function $G_i(\boldsymbol{w})$ is Lipschitz continuous. Moreover, considering the objective function $G_i(\boldsymbol{w})$ and $G_i(\boldsymbol{w}')$, for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, we have

$$\|\nabla G_i(\boldsymbol{w}) - \nabla G_i(\boldsymbol{w}')\| \le L'\|\boldsymbol{w} - \boldsymbol{w}'\|, \tag{22}$$

where $L' = L(1+\eta L)^2 + \eta\rho B$.

Proof: Please see Appendix B. ■

Lemma 1 indicates that the meta-learning objective function $G_i(\boldsymbol{w})$ is obtained by the one-step gradient descent based on $F_i(\boldsymbol{w})$. If we set the leaning rate as $\eta = 0$, we have $L' = L$.

Lemma 2: Considering the local objective function $G_i(\boldsymbol{w})$ and global objective function $G(\boldsymbol{w}) \triangleq \sum_{i\in\mathcal{U}} p_i G_i(\boldsymbol{w})$, for any $\boldsymbol{w} \in \mathbb{R}^d$, we have

$$\|\nabla G_i(\boldsymbol{w}) - \nabla G(\boldsymbol{w})\| \le \varepsilon_i(1+\eta L) + \eta B\gamma_i. \tag{23}$$

Proof: Please see Appendix C. ■

Lemma 2 bounds the bias of the gradient for the meta-learning objective function $G_i(\boldsymbol{w})$ based on Assumptions 1-3.

A. Challenges in Analyzing RDP-PFL

We first briefly highlight the challenges in analyzing the convergence of RDP-PFL based on the intermediate lemmas above.

1) Biased Estimator: We can note that $\nabla F_i(\boldsymbol{w}_i^{t,\tau} - \eta\nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}), \mathcal{D}_i^{\mathrm{S}})$ is not an unbiased estimator of the gradient $\nabla G_i(\boldsymbol{w}_i^{t,\tau})$. In other word, the descent direction utilized in the local training for updating models is a biased estimator. To analyze this bias, we have the following Lemma.

Lemma 3: Suppose that the conditions in Assumptions 1-3 are satisfied. By defining

$$\widetilde{G}_i\left(\boldsymbol{w}_i^{t,\tau}\right) \triangleq F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta\nabla F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}\right), \mathcal{D}_i^{\mathrm{S}}\right), \tag{24}$$

and $\widetilde{\nabla}G_i\left(\boldsymbol{w}_i^{t,\tau}\right)$ *is the gradient of* $\widetilde{G}_i\left(\boldsymbol{w}_i^{t,\tau}\right)$*, we have*

$$\mathbb{E}\{\|\widetilde{\nabla}G_i\left(\boldsymbol{w}_i^{t,\tau}\right) - \nabla G_i\left(\boldsymbol{w}_i^{t,\tau}\right)\|\} \le \frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}. \tag{25}$$

Proof: Please see Appendix D. ■

2) Gradient Clipping: We can notice that the gradient clipping will affect the convergence performance of RDP-PFL. When the clipping threshold C is smaller than the gradient norm, the values of the gradient will be scaled down. The following lemma gives an upper bound to measure the distance between the unclipped gradient and clipped one.

Lemma 4: Considering the unprocessed gradient $\widetilde{\nabla}G_i(\boldsymbol{w}_i^{t,\tau})$ *and the clipped gradient* $\overline{\nabla}G_i(\boldsymbol{w}_i^{t,\tau})$*, we have*

$$\begin{aligned}&\|\overline{\nabla}G_i(\boldsymbol{w}_i^{t,\tau}) - \widetilde{\nabla}G_i(\boldsymbol{w}_i^{t,\tau})\| \\ &\triangleq \Xi(C,S,\sigma) \le (1+\eta L) \\ &\cdot\left(2\eta LB + \eta L\Theta + \frac{\sqrt{C}}{\min\left\{\frac{\sqrt{C}}{B},1\right\}} - \sqrt{C}\min\left\{\frac{\sqrt{C}}{1+\eta L},1\right\}\right),\end{aligned} \tag{26}$$

where

$$\Theta = \begin{cases} 2\sigma C\sqrt{\dfrac{2}{\pi}}\dfrac{S-1}{S-2}\cdot\dfrac{S-3}{S-4}\cdots\dfrac{4}{3}\cdot 2, & \text{if } S \text{ is odd,} \\ 2\sigma C\sqrt{2\pi}\dfrac{S-1}{S-2}\cdot\dfrac{S-3}{S-4}\cdots\dfrac{3}{2}, & \text{if } S \text{ is even,}\end{cases} \tag{27}$$

and S is the number of model parameters.

Proof: See Appendix E. ■

3) Multi-Step Local Update: We can note that multiple local updates before aggregation have a performance loss compared with the centralized meta-learning [16], [25]. Following the same method in [25], we denote $\widehat{\boldsymbol{w}}^{t-1,\tau}$ by the model parameter obtained by the global aggregation result at each local epoch. We denote $\boldsymbol{v}^t$ and $\boldsymbol{v}_i^{t,\tau}$ by the global and local models in RDP-PFL without data sampling and DP mechanism in the training process, respectively. Both $\widehat{\boldsymbol{w}}^{t-1,0}$ and $\boldsymbol{v}^{t-1}$ are synchronized with $\boldsymbol{w}^{t-1}$ at the beginning of the $(t-1)$-th communication round, i.e., $\widehat{\boldsymbol{w}}^{t-1,0} = \boldsymbol{v}^{t-1} = \boldsymbol{w}^{t-1}$. The lemma below gives an upper bound on the difference between $\boldsymbol{v}^t$ and $\widehat{\boldsymbol{w}}^{t-1,\tau_0}$.

Lemma 5: The distance between $\boldsymbol{v}^t$ *and* $\widehat{\boldsymbol{w}}^{t-1,\tau_0}$ *can be bounded as follows*:

$$\begin{aligned}&\mathbb{E}\{\|\boldsymbol{v}^t - \widehat{\boldsymbol{w}}^{t-1,\tau_0}\|\} \\ &\triangleq h(\tau_0) \\ &\le \beta(\varepsilon(1+\eta L) + \eta B\gamma)\left(\frac{(1+\beta L')^{\tau_0}-1}{\beta L'} - \tau_0\right).\end{aligned} \tag{28}$$

Proof: See Appendix F. ■

B. Convex Setting

Now we proceed to establish a convergence bound for RDP-PFL with the strongly convex loss function assumption. We first formally state the strongly convex assumption.

Assumption 5: $F_i(\boldsymbol{w})$ is l-strongly convex, for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, which implies that

$$\langle\nabla F_i(\boldsymbol{w}) - \nabla F_i(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}'\rangle \ge l\|\boldsymbol{w} - \boldsymbol{w}'\|^2. \tag{29}$$

We can note that Assumption 5 is satisfied for many ML models [18], e.g., squared support-vector machine (SVM) and linear regression models.

Based on this strong convexity assumption, we can have the following lemma.

Lemma 6: $G_i(\boldsymbol{w})$ is l'-strongly convex, for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, which implies that

$$\langle\nabla G_i(\boldsymbol{w}) - \nabla G_i(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}'\rangle \ge l'\|\boldsymbol{w} - \boldsymbol{w}'\|^2, \tag{30}$$

where $l' = (1-\eta L)(l-\eta L^2) - \eta\rho B$. Proof: See Appendix G. ■

With the above preparation, we characterize the convergence of RDP-PFL in the following theorem.

Theorem 2: By supposing that **Assumptions 1-5** *are satisfied, we have the following convergence bound for RDP-PFL*:

$$\begin{aligned}&G(\boldsymbol{w}^T) - G(\boldsymbol{w}^\star) \le \zeta^{T\tau_0}(G(\boldsymbol{w}^0) - G(\boldsymbol{w}^\star)) \\ &+ \frac{1-\zeta^{T\tau_0}}{1-\zeta^{\tau_0}}\lambda\Bigg((1+\eta L)\underbrace{h(\tau_0)}_{\text{(a) (caused by multi-step local update)}} \\ &+ \beta\tau_0(1+\eta L)\sum_{i\in\mathcal{U}}p_i\Bigg(\underbrace{\Xi(C,S,\sigma)}_{\text{(b) (caused by DP clipping and DP noise)}} \\ &+ \underbrace{\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}}_{\text{(c) (caused by the number of training data)}}\Bigg) \\ &\cdot\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j\Bigg),\end{aligned} \tag{31}$$

where $\boldsymbol{w}^\star$ is the optimal model and $\zeta = 1 - 2l'\beta + L'l'\beta^2$.

Proof: Please see Appendix H. ■

In Theorem 2, we can select a proper learning rate β to make the parameter ζ smaller than 1. We can observe that the convergence performance of RDP-PFL will be affected by the number of model parameters S, the number of training epochs T, and the training data size $|\mathcal{D}_i^{\mathrm{S}}|$. As can be seen from terms (a)-(c) in (31), both the local data heterogeneity, i.e., ε and γ in $h(\tau_0)$, and noise SD σ, deteriorate the convergence performance. When the DP mechanism is removed in the training process, i.e., $\sigma = 0$ and $C \to +\infty$, it can decrease the convergence bound, i.e., improve the training performance.

Remark 1: Theorem 2 *suggests that there is an optimal value of the number of communication rounds T in terms of convergence performance with a given privacy level ϵ. Specifically, if the number of communication rounds T is larger, the term $\zeta^{T\tau_0}(G(\boldsymbol{w}^0) - G(\boldsymbol{w}^\star))$ in Eq. (31) will be smaller but the terms (a) and (c) in Eq. (31) will be larger due to a larger noise SD caused by DP. By setting $\tau = 1$, $C \to +\infty$ and $\sigma = 0$,* Theorem 2 *recovers the convergence rate of the conventional meta-learning.*

Remark 2: We can also note that a large training data size can enhance the convergence performance, because if $|\mathcal{D}_i^{\mathrm{S}}|$ is larger, the convergence bound in Eq. (31) will be smaller.

*Remark 3: In the conventional meta training, if the number of neurons is larger, the convergence performance will be better over the enough communication rounds. In RDP-PFL, a larger number of model parameters S will result in a larger gradient compression because of the clipping process, which can be seen in **Lemma 4**. Therefore, there exists an optimal number of model parameters for a given clipping threshold C.*

C. Non-Convex Setting

We now present a convergence bound for RDP-PFL with the non-convex loss function assumption.

Theorem 3: If **Assumptions 1-4** *are satisfied, then we have the following convergence bound for RDP-PFL:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{\tau_0-1}\left\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\right\|^2 &\le \frac{G(\boldsymbol{w}^0)-G(\boldsymbol{w}^\star)}{T\beta\left(1-\frac{\beta L'}{2}\right)}\\ &+\frac{\lambda}{\beta\left(1-\frac{\beta L'}{2}\right)}\Bigg((1+\eta L)\underbrace{h(\tau_0)}_{\text{(a) (caused by multi-step local update)}}\\ &+\beta\tau_0\lambda(1+\eta L)\sum_{i\in\mathcal{U}}p_i\Bigg(\underbrace{\Xi(C,S,\sigma)}_{\text{(b) (caused by DP clipping and DP noise)}}\\ &+\underbrace{\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}}_{\text{(c) (caused by the number of training data)}}\Bigg)\\ &\cdot\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j\Bigg).\end{aligned}\tag{32}$$

Proof: Please see Appendix I. ■

Theorem 3 suggests that there is an optimal number of communication rounds. We can notice that the term $\frac{G(\boldsymbol{w}^0)-G(\boldsymbol{w}^\star)}{T\beta(1-\frac{\beta L'}{2})}$ is a decreasing function of T. However, a larger T will lead to a higher SD of the Gaussian noise as shown in Theorem 1, which will damage the training performance according to terms (a) and (c) in Eq. (32). We can also see that, at least in terms of the bounds, the convergence speed of the convex setting, $O(\zeta^{T\tau_0})$, where $\zeta = 1-2l'\beta + L'l'\beta^2 < 1$, is much faster than that in the non-convex setting, $O(\frac{1}{T})$.

V. Experiments

A. Experimental Setup

We examine experimental results for RDP-PFL, on three neural network models, i.e., multi-layer perceptron (MLP), convolutional neural network (CNN) and ResNet-18, and two datasets, i.e., MNIST and CIFAR-10, which are described as follows:

- **Datasets.** MNIST is a dataset of digits consisting of 60, 000 training examples and 10, 000 testing examples formatted as 28×28 size gray scale images [26]. The CIFAR-10 dataset consists of 60, 000 color images in 10 object classes with 6, 000 images included per class [27]. The complete dataset of CIFAR-10 is pre-divided into 50, 000 training images and 10, 000 test images. We use the cross-entropy loss for both MNIST and CIFAR-10.
- **Models.** The MLP model is a simple feed-forward deep neural network with ReLU units and additional softmax layer of 10 classes (corresponding to 10 categories). The CNN model consists of three 3×3 convolution layers (the first with 64 filters, the second with 128 filters, the third with 256 filters, each followed with the 2×2 max pooling layer and ReLu activation function), two fully connected layers (the first with 128 units, the second with 256 units, each followed with the ReLu activation function), and a final softmax output layer. ResNet-18 consists of 17 convolutional layers with the filter size of 3×3, a fully connected layer, and an additional softmax layer [28]. The ResNet-18 model involves 33.16 million parameters, in which the ReLU activation function and batch normalization (BN) are attached to the entire convolutional layers.

We consider that there are 50 clients, in which 60% clients are selected as source clients and the rest of clients are applied for evaluating the fast adaptation performance. In order to illustrate the personalized learning process, we adopt the non-IID data setting as done in [29]. We assign three classes for each client and ensure that the class sets of all clients are different from each other. Specifically, we generate 50 class sets, in which each class set contains three different classes and all class sets are different from each other. We randomly select a specific class set for each client without replacement, and then assign each client 600 data samples, i.e., 200 data samples for each class based on its class set. In addition, for each client, we divide the local training dataset into two datasets, i.e., query and support datasets, in which 50% training data is selected as the query dataset, and the rest of the training data is selected as the support dataset, as described in [15]. For convenience, we set $q^{\mathrm{Q}} = q^{\mathrm{S}}$ in the experiments. For tuning hyper-parameters, we use the grid search method to find the optimal learning rates η and β to conduct our experiments [30]. In particular, we divide the set of clients into a test set (target clients), validation set and training set (source clients) randomly with the ratios 0.2, 0.1, and 0.7, respectively. In addition, we tune the hyper-parameters with the validation set, because the performance of the model generated by the training set is usually evaluated on the validation set. When training with RDP-PFL, we set the learning rates η and β to 0.01 for MNIST, and 0.05 for CIFAR-10, respectively.

B. Evaluation of Privacy Levels

In Fig. 2(a), we choose various privacy levels $\epsilon = 2, 4, 6$ with MLP on the MNIST dataset to show the test accuracy of RDP-PFL. In this experiment, we set $T = 200$ and $\delta = 0.001$ to how the test accuracy changes as the communication round t increases. As shown in Fig. 2(a), values of the test accuracy in RDP-PFL are increasing when we relax the privacy guarantee (increasing ϵ). We also use CNN model on the CIFAR-10 dataset to show how the test accuracy changes as the privacy level varies in Fig. 2(b). In this figure, we choose the values of ϵ as 2, 6 and 10 and can observe the same property as Fig. 2(a). In Fig. 3, we show how the test accuracy of fast adaption changes as the DP parameter δ varies. It can be noted that values of the test accuracy for RDP-PFL are increasing when we choose large values of δ.

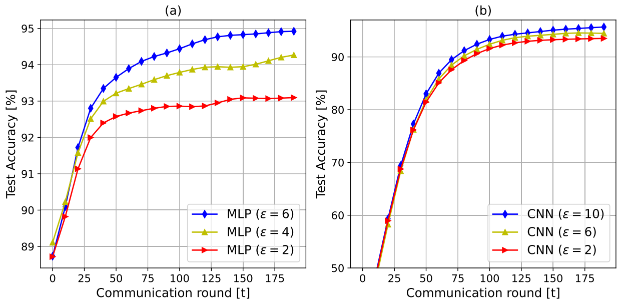


Fig. 2. The test accuracy with different values of ϵ with $\delta = 0.001$ for the whole training process.

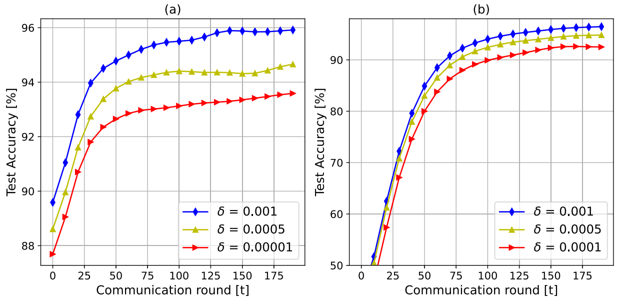


Fig. 3. The test accuracy with different values of δ with $\epsilon = 2$ for the whole training process.

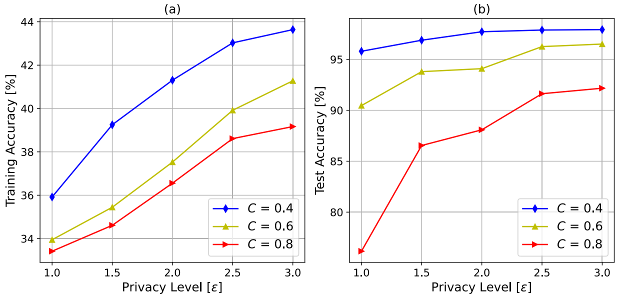


Fig. 4. The training accuracy and test accuracy of RDP-PFL under different values of clipping threshold C and different values of privacy levels ϵ with $\delta = 0.001$.

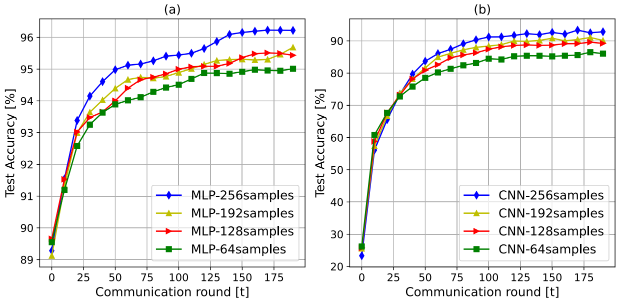


Fig. 5. The test accuracy with different numbers of data samples for the whole training process under $\epsilon = 2$ and $\delta = 0.001$.

We also conduct the proposed RDP-PFL algorithm with ResNet-18 on the CIFAR-10 dataset under different values of the clipping threshold C and different values of the privacy budget ϵ with $\delta = 0.001$. As shown in Fig. 4, both the training accuracy and test accuracy increase as the privacy level ϵ grows. Although the training accuracy is low due to DP, the test accuracy reaches a relatively high level. The reason/intuition is that RDP-PFL aims to learn an initialized model across a set of source clients by a meta-learning approach instead of a model with high training accuracy. Based on the initialized model, the target clients can achieve satisfactory personalized models with fast adaption.

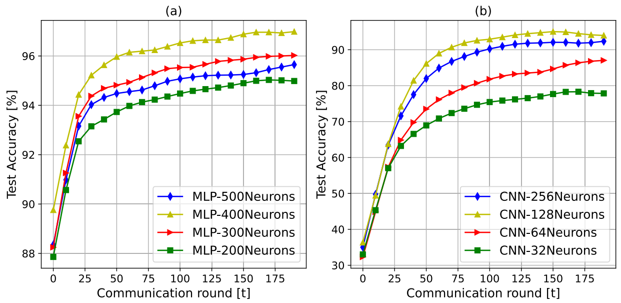


Fig. 6. The test accuracy with different numbers of neurons for the whole training process under $\epsilon = 2$ and $\delta = 0.001$.

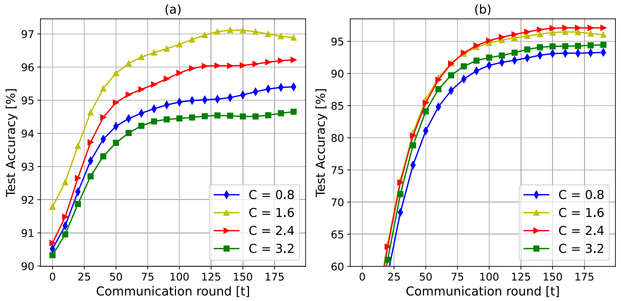


Fig. 7. The test accuracy with different values of clipping thresholds with $\epsilon = 2$ and $\delta = 0.001$.

C. Evaluation of the Number of Data Samples

Figs. 5(a) and (b) show how the test accuracy of fast adaption changes as the number of data samples increases. We conduct the experiments with MLP on MNIST dataset and the CNN model on the CIFAR-10 dataset in Figs. 5(a) and (b), respectively. From Fig. 5(a), we observe that, as the number of data samples increases, the values of the test accuracy increase. This is due to the fact that, as the number of data samples increases, each client in RDP-PFL can use more numbers of data samples for training the meta model. In this way, the aggregated model can finish the fast adaption efficiently. Fig. 5(b) also demonstrates that, when the number of data samples increases, the test accuracy increases quickly.

D. Evaluation of Model Sizes

In Figs. 6(a) and (b), we show how the test accuracy of fast adaption changes as the number of neurons varies. We select various numbers of neurons for the last fully connected layer for MLP and CNN models in this experiment. The fact that the test accuracy remains unchanged demonstrates that the FL algorithm converges. From Figs. 6(a) and (b), we can see that, as the number of neurons increases, the test accuracy of fast adaption of RDP-PFL will increase, and then decrease. This is due to the fact that the the number of neurons is sensitive with the clipping value. If the model size is too large, the convergence performance will be poor.

E. Evaluation of Clipping Thresholds

In Tab. II, we show how the training loss and test loss change as the value of the clipping threshold varies. From Tab. II, we can see that, as the values of the clipping threshold increase, the training loss and test loss of RDP-PFL will first decrease and then increase. We also plot the test accuracy of RDP-PFL with different values of the clipping threshold in Fig. 7. From this figure, we can observe the same property as Tab. II. We can notice that limiting the gradient norm has two

TABLE II
THE TRAINING LOSS AND TEST LOSS WITH DIFFERENT VALUES OF THE CLIPPING THRESHOLD WITH $\epsilon = 2$ AND $\delta = 0.001$

Clipping threshold	MNIST		CIFAR-10	
	Training loss	Test loss	Training loss	Test loss
C = 0.8	0.861	0.219	0.961	0.218
C = 1.6	0.709	0.169	0.770	0.157
C = 2.4	**0.649**	0.156	**0.699**	0.129
C = 3.2	0.657	**0.153**	0.757	**0.114**
C = 4.0	0.712	0.161	0.812	0.166
C = 4.8	1.012	0.211	0.910	0.223

opposing effects. On the one hand, if the clipping threshold C is too small, clipping will destroy the intended gradient direction of parameters. On the other hand, increasing the clipping threshold C forces us to add more privacy-enhancing noise to the parameters.

VI. RELATED WORKS

A. Federated Meta-Learning

The goal of meta-learning is to find an initialized model that current or new clients can easily obtain satisfactory models by performing one-step or a few steps of gradient descent with their own data [14], [15], [16]. The work in [15] first proposed a federated meta-learning framework (FML) with model-agnostic meta-learning (MAML) to train an initialized model capable of rapidly adapting to new learning tasks, instead of a global model in conventional FL. Moreover, extensive experiments on LEAF datasets and a real-world production dataset demonstrated that FML achieves a reduction in required communication cost by $2.82-4.33$ times with a faster convergence rate, and an increase in accuracy by $3.23\% - 14.84\%$ compared with conventional FL [15]. This work [17] evaluated the performance of FML with a convergence bound, in terms of the gradient norm and data distribution distances of clients, for the non-convex loss function assumption. The work in [18] investigated bounds of the training and adaptation performances at the target client in FML in terms of client similarity.

B. Machine Learning With Differential Privacy

Privacy-enhanced ML has attracted intensive attention in recent years [30], [31], [32], [33], as the emergence of ML based open data applications may lead to the leakage of private information. The concept of deep learning with DP was first proposed in [34], which provides an evaluation criterion for privacy guarantees. The work in [35] improved DP-SGD algorithms by dynamically determining the privacy budget and step size for each iteration based on the quality of the gradient of the current training iteration. Further, privacy issues are more critical in distributed ML systems due to the published gradients or model parameters trained locally. The work in [36] provided an approach for analyzing the quality of distributed ML models based on the theoretical analysis of DP-SGD algorithms, which reveals that the quality is related to the privacy level and the size of the datasets. The count sketch algorithm, compressing the local updates using hash functions with bounded errors, was involved to design a communication-efficient and privacy-enhanced FL framework in [37]. The work in [38] developed a privacy-preserving PFL framework for user recommendation models with a hierarchical structure that contains both the public component and private component. Via addressing these components carefully, this privacy-preserving PFL framework in [38] can safeguard the data privacy, where each client uploads the public component directly, while delivering extracted features of the private component.

VII. CONCLUSION

As there is a risk that personal information can be leaked to potential malicious clients in PFL, we have enhanced the privacy protection of data by proposing a novel RDP-PFL framework with meta-training to train an initial shared model under a DP guarantee. Further, we have investigated two convergence bounds for RDP-PFL in terms of the number of communication rounds, model size, and training data size under convex and non-convex loss function assumptions. Finally, we have conducted extensive experiments with three neural networks and two real-world datasets, whose results demonstrate the correctness of our theoretical results. In the proposed RDP-PFL framework, we can see that with an increasing number of communication rounds, the communication cost and privacy budget will increase. Thus, high communication costs and stringent privacy will degrade the training performance. An interesting direction for future work is to apply model compression techniques, such as model pruning or quantization, to enhance the privacy protection and communication efficiency of RDP-PFL.

APPENDIX A
PROOF OF THEOREM 1

Based on the defintion of Rényi DP, we can calculate the Rényi divergence with a positive value α as follows:

$$D_\alpha[\mathcal{M}(\mathcal{D})|\mathcal{M}(\mathcal{D}')] = \frac{1}{\alpha-1}\log\mathbb{E}\left[\left(\frac{\mathcal{M}(\mathcal{D})}{\mathcal{M}(\mathcal{D}')}\right)^{\alpha}\right]. \tag{33}$$

Based on Theorem 20 in [39], we have

$$\epsilon = \epsilon' + \frac{\log(\frac{1}{\delta}) + (\alpha-1)\log(1-\frac{1}{\alpha}) - \log(\alpha)}{\alpha-1}, \tag{34}$$

where ϵ' is the Rényi divergence.

For the inner update, we first use a sampling rate q^{Q} to perform the SGD scheme based on the query dataset and bound the L_2 norm of the gradient vector using the clipping technique with threshold C in Eq. (6). Then we add the Gaussian noise to the clipped gradient vector with SD σ to satisfy DP in Eq. (7). Thus, we can calculate the Rényi divergence for the inner update as follows:

$$D_\alpha[\mu^{\mathrm{Q}}(z)|\mu_0^{\mathrm{Q}}(z)] \tag{35}$$

$$= \frac{1}{\alpha-1}\log\mathbb{E}_{z\sim\mu_0^{\mathrm{Q}}(z)}\left[\left(\frac{(1-q^{\mathrm{Q}})\mu_0^{\mathrm{Q}}(z) + q^{\mathrm{Q}}\mu_1^{Q}(z)}{\mu_0^{\mathrm{Q}}(z)}\right)^{\alpha}\right] \tag{36}$$

$$= \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty}\mu_0^{\mathrm{Q}}(z)\left((1-q^{\mathrm{Q}}) + \frac{q^{Q}\mu_1^{\mathrm{Q}}(z)}{\mu_0^{\mathrm{Q}}(z)}\right)^{\alpha}. \tag{37}$$

For the outer update, we use a sampling rate q^{S} to perform the SGD scheme based on the support dataset and bound the L_2 norm of the gradient vector using the clipping technique with threshold C in Eq. (8). Then we add Gaussian noise to the clipped gradient vector with SD σ to satisfy DP in Eq. (8). Because the term $(I - \eta\nabla^2 F(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}))\nabla F(\boldsymbol{w}_i^{t,\tau} - \eta\nabla F(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}), \mathcal{D}_i^{\mathrm{S}})$ has been bounded by the clipping technique, we can calculate the Rényi divergence for the outer update as follows:

$$D_\alpha[\mu^{\mathrm{S}}(z)|\mu_0^{\mathrm{S}}(z)] \tag{38}$$

$$= \frac{1}{\alpha-1}\log\mathbb{E}_{z\sim\mu_0^{\mathrm{S}}(z)}\left[\left(\frac{(1-q^{\mathrm{S}})\mu_0^{\mathrm{S}}(z) + q^{\mathrm{S}}\mu_1^{\mathrm{S}}(z)}{\mu_0^{\mathrm{S}}(z)}\right)^\alpha\right] \tag{39}$$

$$= \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty}\mu_0^{\mathrm{S}}(z)\left((1-q^{\mathrm{S}}) + \frac{q^{\mathrm{S}}\mu_1^{\mathrm{S}}(z)}{\mu_0^{\mathrm{S}}(z)}\right)^\alpha. \tag{40}$$

Via the composition theorem [39], the total Rényi divergence for the training process can be expressed as

$$\epsilon' = T\tau_0\left(D_\alpha[\mu^{\mathrm{Q}}(z)|\mu_0^{\mathrm{Q}}(z)] + D_\alpha[\mu^{\mathrm{Q}}(z)|\mu_0^{\mathrm{Q}}(z)]\right) \tag{41}$$

$$= \frac{T\tau_0}{\alpha-1}\Bigg[\log\int_{-\infty}^{\infty}\mu_0^{\mathrm{S}}(z)\left((1-q^{\mathrm{S}}) + \frac{q^{\mathrm{S}}\mu_1^{\mathrm{S}}(z)}{\mu_0^{\mathrm{S}}(z)}\right)^\alpha \tag{42}$$

$$+ \log\int_{-\infty}^{\infty}\mu_0^{\mathrm{Q}}(z)\left((1-q^{\mathrm{Q}}) + \frac{q^{\mathrm{Q}}\mu_1^{\mathrm{Q}}(z)}{\mu_0^{\mathrm{Q}}(z)}\right)^\alpha\Bigg] \tag{43}$$

$$= \frac{T\tau_0}{\alpha-1}\log\Bigg[\int_{-\infty}^{\infty}\mu_0^{\mathrm{S}}(z)\left((1-q^{S}) + \frac{q^{\mathrm{S}}\mu_1^{\mathrm{S}}(z)}{\mu_0^{\mathrm{S}}(z)}\right)^\alpha \tag{44}$$

$$\cdot\int_{-\infty}^{\infty}\mu_0^{\mathrm{Q}}(z)\left((1-q^{\mathrm{Q}}) + \frac{q^{\mathrm{Q}}\mu_1^{\mathrm{Q}}(z)}{\mu_0^{\mathrm{Q}}(z)}\right)^\alpha\Bigg] \tag{45}$$

$$= \frac{T\tau_0}{\alpha-1}\log(I^{\mathrm{Q}}I^{\mathrm{S}}), \tag{46}$$

where

$$I^{\mathrm{S}} = \int_{-\infty}^{\infty}\mu_0^{\mathrm{S}}(z)\left((1-q^{S}) + \frac{q^{\mathrm{S}}\mu_1^{\mathrm{S}}(z)}{\mu_0^{\mathrm{S}}(z)}\right)^\alpha, \tag{47}$$

$$I^{\mathrm{Q}} = \int_{-\infty}^{\infty}\mu_0^{\mathrm{Q}}(z)\left((1-q^{\mathrm{Q}}) + \frac{q^{\mathrm{Q}}\mu_1^{\mathrm{Q}}(z)}{\mu_0^{\mathrm{Q}}(z)}\right)^\alpha. \tag{48}$$

Actually, we compute the integrals I^{S} and I^{Q} via the method proposed in Section 3.3 in [40], which has already been adopted in Opacus [41]. In addition, we also use this calculation method based on Opacus in the experiments. This completes the proof. □

Appendix B
Proof of Lemma 1

Via the definition of $G_i(\boldsymbol{w})$, we have

$$\|\nabla G_i(\boldsymbol{w}) - \nabla G_i(\boldsymbol{w}')\| \tag{49}$$

$$= \|\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) - \nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))\| \tag{50}$$

$$= \|(I - \eta\nabla^2 F_i(\boldsymbol{w}))\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{51}$$

$$- (I - \eta\nabla^2 F_i(\boldsymbol{w}'))\nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))\| \tag{52}$$

$$= \|\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{53}$$

$$- \eta\nabla^2 F_i(\boldsymbol{w})\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{54}$$

$$- \nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}')) \tag{55}$$

$$+ \eta\nabla^2 F_i(\boldsymbol{w}')\nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))\| \tag{56}$$

$$= \|[\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) - \nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))] \tag{57}$$

$$\cdot [I - \eta\nabla^2 F_i(\boldsymbol{w}')]\eta\nabla^2 F_i(\boldsymbol{w}')\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{58}$$

$$\cancel{+ -\eta\nabla^2 F_i(\boldsymbol{w}')\nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))} \tag{59}$$

$$- \eta\nabla^2 F_i(\boldsymbol{w})\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{60}$$

$$\cancel{+\eta\nabla^2 F_i(\boldsymbol{w}')\nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))}\| \tag{61}$$

$$= \|[\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) - \nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))] \tag{62}$$

$$\cdot [I - \eta\nabla^2 F_i(\boldsymbol{w}')] - \eta\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{63}$$

$$\cdot [\nabla^2 F_i(\boldsymbol{w}) - \nabla^2 F_i(\boldsymbol{w}')]\|. \tag{64}$$

Using the Lipschitz continuous of $F_i(\boldsymbol{w})$, we can obtain

$$\|\nabla G_i(\boldsymbol{w}) - \nabla G_i(\boldsymbol{w}')\| \tag{65}$$

$$\leq \|\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) - \nabla F_i(\boldsymbol{w}' - \eta\nabla F_i(\boldsymbol{w}'))\| \tag{66}$$

$$\|I - \eta\nabla^2 F_i(\boldsymbol{w}')\| + \eta\|\nabla^2 F_i(\boldsymbol{w}) - \nabla^2 F_i(\boldsymbol{w}')\| \tag{67}$$

$$\|\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w}))\| \tag{68}$$

$$\leq L(1 + \eta L)\|\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w}) - \boldsymbol{w}' + \eta\nabla F_i(\boldsymbol{w}')\| \tag{69}$$

$$+ \eta\rho B\|\boldsymbol{w} - \boldsymbol{w}'\| \tag{70}$$

$$\leq L(1 + \eta L)^2\|\boldsymbol{w} - \boldsymbol{w}'\| + \eta\rho B\|\boldsymbol{w} - \boldsymbol{w}'\| \tag{71}$$

$$\leq L'\|\boldsymbol{w} - \boldsymbol{w}'\|, \tag{72}$$

where $L' = L(1 + \eta L)^2 + \eta\rho B$. This completes the proof. □

Appendix C
Proof of Lemma 2

Via the definitions of $G_i(\boldsymbol{w})$ and $G(\boldsymbol{w})$, we can have

$$\|\nabla G_i(\boldsymbol{w}) - \nabla G(\boldsymbol{w})\| \tag{73}$$

$$= \|\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) - \nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w}))\| \tag{74}$$

$$= \|(I - \eta\nabla^2 F_i(\boldsymbol{w}))\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{75}$$

$$- (I - \eta\nabla^2 F(\boldsymbol{w}))\nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w}))\| \tag{76}$$

$$= \|\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{77}$$

$$- \eta\nabla^2 F_i(\boldsymbol{w})\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{78}$$

$$- \nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w})) \tag{79}$$

$$+ \eta\nabla^2 F(\boldsymbol{w})\nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w}))\| \tag{80}$$

$$= \|[\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) - \nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w}))] \tag{81}$$

$$\cdot [I - \eta\nabla^2 F(\boldsymbol{w})] + \eta\nabla^2 F(\boldsymbol{w})\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{82}$$

$$\cancel{-\eta\nabla^2 F(\boldsymbol{w})\nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w}))} \tag{83}$$

$$- \eta\nabla^2 F_i(\boldsymbol{w})\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) \tag{84}$$

$$\cancel{+\eta\nabla^2 F(\boldsymbol{w})\nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w}))}\| \tag{85}$$

$$= \|[\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w})) - \nabla F(\boldsymbol{w} - \eta\nabla F(\boldsymbol{w}))] \tag{86}$$

$$\cdot [I - \eta\nabla^2 F(\boldsymbol{w})] \tag{87}$$

$$- \eta\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w}))[\nabla^2 F_i(\boldsymbol{w}) - \nabla^2 F(\boldsymbol{w})]\| \tag{88}$$

$$\leq \varepsilon_i\|I - \eta\nabla^2 F(\boldsymbol{w})\| \tag{89}$$

$$+ \eta\|\nabla^2 F_i(\boldsymbol{w}) - \nabla^2 F(\boldsymbol{w})\|\|\nabla F_i(\boldsymbol{w} - \eta\nabla F_i(\boldsymbol{w}))\| \tag{90}$$

$$\leq \varepsilon_i(1 + \eta L) + \eta B\gamma_i. \tag{91}$$

This completes the proof. □

APPENDIX D
PROOF OF LEMMA 3

First, we define $G_i(\boldsymbol{w}) \triangleq F_i(\boldsymbol{w} - \eta \nabla F_i(\boldsymbol{w}))$ and

$$\widetilde{G}_i(\boldsymbol{w}_i^{t,\tau}) \triangleq F_i(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}}), \mathcal{D}_i^{\mathrm{S}}), \tag{92}$$

Further, we define $\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})$ as the gradient of $\widetilde{G}_i(\boldsymbol{w}_i^{t,\tau})$. According to the update rule, we can see the SGD gradient at t-th round can be given by

$$\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) = \left(\boldsymbol{I} - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}})\right) \tag{93}$$

$$\nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}), \mathcal{D}_i^{\mathrm{S}}\right). \tag{94}$$

However, the exact gradient of ∇G_i at $\boldsymbol{w}_i^{t,\tau}$ is given by

$$\nabla G_i\left(\boldsymbol{w}_i^{t,\tau}\right) = \left(\boldsymbol{I} - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau})\right) \tag{95}$$

$$\nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau})\right). \tag{96}$$

We can note that, given ∇G_i and $\boldsymbol{w}_i^{t,\tau}$, the update we used is a biased estimation of $\nabla G_i(\boldsymbol{w}_i^{t,\tau})$. Hence, we have

$$\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) = \left(\boldsymbol{I} - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}) + e_i^H\right) \tag{97}$$

$$\left(\nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau})\right) + e_i^G\right), \tag{98}$$

where

$$e_i^H = \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}) - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}), \tag{99}$$

and

$$e_i^G = \nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}), \mathcal{D}_i^{\mathrm{S}}\right) \tag{100}$$

$$- \nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau})\right). \tag{101}$$

Further, substituting (95) into (97), we can obtain

$$\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})$$

$$= \nabla G_i(\boldsymbol{w}_i^{t,\tau}) + e_i^G\left(\boldsymbol{I} - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau})\right) \tag{102}$$

$$+ e_i^H\left(\nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau})\right)\right) + e_i^G e_i^H. \tag{103}$$

We can calculate the expectation of $\|e_i^H\|^2$ by

$$\mathbb{E}\left\{\|e_i^H\|^2\right\} \tag{104}$$

$$= \eta \mathbb{E}\left\{\left\|\nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}) - \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}})\right\|^2\right\} \tag{105}$$

$$= \eta \mathbb{E}\Bigg\{\Bigg\|\frac{1}{|\mathcal{D}_i^{\mathrm{S}}|}\sum_{\boldsymbol{x} \in \mathcal{D}_i^{\mathrm{S}}}\left(\nabla^2 F_i(\boldsymbol{w}_i^{t,\tau})\right. \tag{106}$$

$$\left.- \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \boldsymbol{x})\right)\Bigg\|^2\Bigg\} \tag{107}$$

$$= \frac{\eta}{|\mathcal{D}_i^{\mathrm{S}}|^2}\mathbb{E}\Bigg\{\sum_{\boldsymbol{x} \in \mathcal{D}_i^{\mathrm{S}}}\|\nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}) \tag{108}$$

$$- \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \boldsymbol{x})\|^2\Bigg\}. \tag{109}$$

Combining Eq. (21), we have

$$\mathbb{E}\left\{\|e_i^H\|^2\right\} \leq \frac{\eta^2 \sigma_H^2}{|\mathcal{D}_i^{\mathrm{S}}|}. \tag{110}$$

Further, we can bound $\mathbb{E}\{\|e_i^H\|\}$ using Jensen's inequality as

$$\mathbb{E}\left\{\|e_i^H\|\right\} \leq \sqrt{\mathbb{E}\left\{\|e_i^H\|^2\right\}} \leq \frac{\eta \sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}. \tag{111}$$

Similarly, we can also bound the expectations of $\|e_i^G\|^2$ and $\|e_i^G\|$ as

$$\mathbb{E}\left\{\|e_i^G\|^2\right\} \leq \frac{\eta^2 L^2 \sigma_G^2}{|\mathcal{D}_i^{\mathrm{S}}|}, \ \mathbb{E}\left\{\|e_i^G\|\right\} \leq \frac{\eta L \sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}. \tag{112}$$

Based on Eq. (102), we can have

$$\mathbb{E}\{\|\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) - \nabla G_i(\boldsymbol{w}_i^{t,\tau})\|\} \tag{113}$$

$$\leq \mathbb{E}\{\|e_i^G\left(I - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau})\right) \tag{114}$$

$$+ e_i^H\left(\nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau})\right)\right) + e_i^G e_i^H\|\} \tag{115}$$

$$\leq \mathbb{E}\Bigg\{\|e_i^G\|\left\|I - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau})\right\| + \|e_i^G\|\|e_i^H\| \tag{116}$$

$$+ \|e_i^H\|\left\|\nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i(\boldsymbol{w}_i^{t,\tau})\right)\right\|\Bigg\}. \tag{117}$$

Substituting Eqs. (111) and (112) into (113), using Assumptions 1-2, we can obtain

$$\mathbb{E}\{\|\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) - \nabla G_i(\boldsymbol{w}_i^{t,\tau})\|\} \tag{118}$$

$$\leq \frac{\eta(1 + \eta L) L \sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta B \sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L \sigma_G \sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}. \tag{119}$$

This completes the proof. □

APPENDIX E
PROOF OF LEMMA 4

According to RDP-PFL, the definitions of $\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})$ and $\overline{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})$ can be expressed as

$$\widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) \triangleq (\boldsymbol{I} - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{Q}})) \tag{120}$$

$$\cdot \nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}\right), \mathcal{D}_i^{\mathrm{S}}\right), \tag{121}$$

$$\overline{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) \triangleq \frac{\boldsymbol{g}_1}{\max\left\{1, \frac{\|\boldsymbol{g}_1\|}{\sqrt{C}}\right\}} \frac{\boldsymbol{g}_2}{\max\left\{1, \frac{\|\boldsymbol{g}_2\|}{\sqrt{C}}\right\}}, \tag{122}$$

where

$$\boldsymbol{g}_1 = I - \eta \nabla^2 F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}), \quad \boldsymbol{g}_2 = \nabla F_i(\widetilde{\boldsymbol{\theta}}_i^{t,\tau+1}, \mathcal{D}_i^{\mathrm{S}}). \tag{123}$$

Based on the definitions of $\widetilde{\nabla} G_i(\boldsymbol{w})$ and $\overline{\nabla} G_i(\boldsymbol{w})$, we can obtain

$$\mathbb{E}\{\|\overline{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) - \widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})\|\} \tag{124}$$

$$= \mathbb{E}\Bigg\{\Bigg\|\frac{\boldsymbol{g}_1}{\max\left\{1, \frac{\|\boldsymbol{g}_1\|}{\sqrt{C}}\right\}} \cdot \frac{\boldsymbol{g}_2}{\max\left\{1, \frac{\|\boldsymbol{g}_2\|}{\sqrt{C}}\right\}} \tag{125}$$

$$- \boldsymbol{g}_1 \cdot \nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}\right), \mathcal{D}_i^{\mathrm{S}}\right)\Bigg\|\Bigg\} \tag{126}$$

$$\leq \mathbb{E}\{\|\boldsymbol{g}_1\|\|C_1 C_2 \cdot \boldsymbol{g}_2 \tag{127}$$

$$- \nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}\right), \mathcal{D}_i^{\mathrm{S}}\right)\|\}, \tag{128}$$

where

$$C_1 = \min\left\{\frac{\sqrt{C}}{\|\boldsymbol{g}_1\|}, 1\right\}, \quad C_2 = \min\left\{\frac{\sqrt{C}}{\|\boldsymbol{g}_2\|}, 1\right\}. \tag{129}$$

Further, we can have

$$\mathbb{E}\{\|\overline{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) - \widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})\|\} \leq \mathbb{E}\big\{\|\boldsymbol{g}_1\|\|C_1 C_2 \cdot \boldsymbol{g}_2 \tag{130}$$

$$- \boldsymbol{g}_2 + \boldsymbol{g}_2 - \nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}\right), \mathcal{D}_i^{\mathrm{S}}\right)\|\big\} \tag{131}$$

$$\leq \mathbb{E}\big\{\|\boldsymbol{g}_1\|\big((1 - C_1 C_2)\|\boldsymbol{g}_2\| \tag{132}$$

$$+ \left\|\boldsymbol{g}_2 - \nabla F_i\left(\boldsymbol{w}_i^{t,\tau} - \eta \nabla F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}\right), \mathcal{D}_i^{\mathrm{S}}\right)\right\|\big). \tag{133}$$

Due to Assumption 2, we can obtain

$$\mathbb{E}\{\|\overline{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) - \widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})\|\} \tag{134}$$

$$\leq \mathbb{E}\big\{\|\boldsymbol{g}_1\|\big((1/C_2 - C_1) C_2 \|\boldsymbol{g}_2\| \tag{135}$$

$$+ L\left\|\widetilde{\boldsymbol{\theta}}_i^{t,\tau} - \boldsymbol{w}_i^{t,\tau} + \eta \nabla F_i\left(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}}\right)\right\|\big). \tag{136}$$

Because $C_2\|\boldsymbol{g}_2\| \leq \sqrt{C}$ and Eq. (7), we have

$$\mathbb{E}\{\|\overline{\nabla} G_i(\boldsymbol{w}_i^{t,\tau}) - \widetilde{\nabla} G_i(\boldsymbol{w}_i^{t,\tau})\|\} \leq \mathbb{E}\bigg\{\|\boldsymbol{g}_1\| \cdot \bigg(\eta L \tag{137}$$

$$\left\|\frac{\nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}})}{\max\left\{1, \frac{\|\nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}})\|}{C}\right\}} - \nabla F_i(\boldsymbol{w}_i^{t,\tau}, \mathcal{D}_i^{\mathrm{S}})\right\| \tag{138}$$

$$+ \eta L \|\boldsymbol{n}_i^{\mathrm{Q}}\| + \frac{\sqrt{C}}{C_2} - \sqrt{C} C_1\bigg)\bigg\} \tag{139}$$

$$\leq (1 + \eta L)(2\eta L B + \eta L \mathbb{E}\{\|\boldsymbol{n}_i^{\mathrm{Q}}\|\} + \sqrt{C}/C_2 - \sqrt{C} C_1) \tag{140}$$

$$\leq (1 + \eta L)\left(2\eta L B + \eta L \Theta + \frac{\sqrt{C}}{\min\left\{\frac{\sqrt{C}}{B}, 1\right\}}\right. \tag{141}$$

$$\left. - \sqrt{C} \min\left\{\frac{\sqrt{C}}{1 + \eta L}, 1\right\}\right), \tag{142}$$

where Θ is an upper bound on of $\mathbb{E}\{\|\boldsymbol{n}_i^{\mathrm{Q}}\|\}$ and can be derived as follows. We can note that the elements in $\boldsymbol{n}_i^{\mathrm{Q}}$ and $\boldsymbol{n}_i^{\mathrm{S}}$ are drawn from the same Gaussian distribution $\mathcal{N}(0, C^2\sigma^2)$. We assume that a noise vector $\boldsymbol{n}$ is generated using the same method as $\boldsymbol{n}_i^{\mathrm{Q}}$ and $\boldsymbol{n}_i^{\mathrm{S}}$, and then derive an upper bound on $\mathbb{E}\{\|\boldsymbol{n}\|\}$ as follows:

$$\mathbb{E}\{\|\boldsymbol{n}\|\}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^S \int_{-\infty}^{+\infty} \mathrm{d}n_1 \tag{143}$$

$$\cdots \int_{-\infty}^{+\infty} \sqrt{n_1^2 + \cdots + n_S^2}\, e^{-\frac{n_1^2 + \cdots + n_S^2}{2}} \mathrm{d}n_S \tag{144}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^S \int_{-\infty}^{+\infty} \mathrm{d}n_1 \cdots \int_{-\infty}^{+\infty} \mathrm{d}n_{S-2} \int_0^{+\infty} \int_{-\pi}^{\pi} \tag{145}$$

$$r\sqrt{n_1^2 + \cdots + n_{S-2}^2 + r^2}\, e^{-\frac{n_1^2 + \cdots + n_{S-2}^2 + r^2}{2}} \mathrm{d}r\mathrm{d}\theta \tag{146}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^S \int_{-\infty}^{+\infty} \mathrm{d}n_1 \cdots \int_{-\infty}^{+\infty} \mathrm{d}n_{S-2} \int_0^{+\infty} \tag{147}$$

$$2\pi r\sqrt{n_1^2 + \cdots + n_{S-2}^2 + r^2}\, e^{-\frac{n_1^2 + \cdots + n_{S-2}^2 + r^2}{2}} \mathrm{d}r \tag{148}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^S \int_{-\infty}^{+\infty} \mathrm{d}n_1 \cdots \int_{-\infty}^{+\infty} \mathrm{d}n_{S-3} \int_0^{+\infty} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \tag{149}$$

$$2\pi r^2 \cos\theta \sqrt{n_1^2 + \cdots + n_{S-3}^2 + r^2}\, e^{-\frac{n_1^2 + \cdots + n_{S-3}^2 + r^2}{2}} \mathrm{d}r\mathrm{d}\theta, \tag{150}$$

where n_s represents the s-th element in $\boldsymbol{n}$, $s \in \{1, \ldots, S\}$, and S is the size of $\boldsymbol{n}$. Recursively, we have

$$\mathbb{E}\{\|\boldsymbol{n}\|\} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^S \int_{-\frac{\pi}{2}}^{+\infty} r^S e^{-\frac{r^2}{2}} \mathrm{d}r \tag{151}$$

$$2\pi \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos\theta \mathrm{d}\theta \cdots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{S-2}\theta \mathrm{d}\theta. \tag{152}$$

The first integral can be calculated by

$$\int_0^{+\infty} r^S e^{-\frac{r^2}{2}} \mathrm{d}r = -\sigma^2 r^{S-1} e^{-\frac{r^2}{2}} \mid_0^{-\infty} \tag{153}$$

$$+ \int_0^{+\infty} (S-1)\sigma^2 r^{S-2} e^{-\frac{r^2}{2}} \mathrm{d}r \tag{154}$$

$$= \begin{cases} (S-1)(S-3)\cdots 2\sigma^{S+1}, & \text{if } S \text{ is odd}, \\ \sqrt{2\pi}(S-1)(S-3)\cdots 3\sigma^{S+1}, & \text{if } S \text{ is even}. \end{cases} \tag{155}$$

Because

$$\int_0^{\frac{\pi}{2}} \cos^S\theta \mathrm{d}\theta = \begin{cases} \frac{S-1}{S} \cdot \frac{S-3}{S-2} \cdots \frac{4}{5} \cdot \frac{2}{3}, & \text{if } S \text{ is odd}, \\ \frac{S-1}{S} \cdot \frac{S-3}{S-2} \cdots \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{\pi}{2}, & \text{if } S \text{ is even}, \end{cases} \tag{156}$$

and

$$\int_0^{\frac{\pi}{2}} \cos^S\theta \mathrm{d}\theta \int_0^{\frac{\pi}{2}} \cos^{S-1}\theta \mathrm{d}\theta = \frac{\pi}{2S}, \tag{157}$$

we can have

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos\theta \mathrm{d}\theta \cdots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{S-2}\theta \mathrm{d}\theta \tag{158}$$

$$= 2^{S-2} \int_0^{\frac{\pi}{2}} \cos\theta \mathrm{d}\theta \cdots \int_0^{\frac{\pi}{2}} \cos^{S-2}\theta \mathrm{d}\theta \tag{159}$$

$$= \begin{cases} 2^{S-2} \cdot \frac{\pi}{2(S-2)} \cdot \frac{\pi}{2(S-4)} \cdots \frac{\pi}{2 \cdot 3} \cdot 2, & \text{if } S \text{ is odd}, \\ 2^{S-2} \cdot \frac{\pi}{2(S-2)} \cdot \frac{\pi}{2(S-4)} \cdots \frac{\pi}{2 \cdot 2} \cdot 2, & \text{if } S \text{ is even}. \end{cases} \tag{160}$$

Based on (153) and (158), we can obtain

$$\mathbb{E}\{\|\boldsymbol{n}\|\} \triangleq \Theta \tag{161}$$

$$= \begin{cases} 2\sqrt{\frac{2}{\pi}}\sigma \frac{S-1}{S-2} \cdot \frac{S-3}{S-4} \cdots \frac{4}{3} \cdot 2, & \text{if } S \text{ is odd}, \\ 2\sqrt{2\pi}\sigma \frac{S-1}{S-2} \cdot \frac{S-3}{S-4} \cdots \frac{3}{2}, & \text{if } S \text{ is even}. \end{cases} \tag{162}$$

This completes the proof. □

Appendix F
Proof of Lemma 5

Due to the update rule of $\widehat{\boldsymbol{w}}^{t-1,\tau}$, i.e.,

$$\widehat{\boldsymbol{w}}^{t-1,\tau} = \widehat{\boldsymbol{w}}^{t-1,\tau-1} - \beta \sum_{i \in \mathcal{U}} p_i \nabla G_i(\widehat{\boldsymbol{w}}^{t-1,\tau-1}) \tag{163}$$

$$= \widehat{\boldsymbol{w}}^{t-1,\tau-1} - \beta \nabla G(\widehat{\boldsymbol{w}}^{t-1,\tau-1}), \tag{164}$$

we can have

$$\left\| \boldsymbol{v}^t - \widehat{\boldsymbol{w}}^{t-1,\tau_0} \right\| \leq \left\| \sum_{i \in \mathcal{U}} p_i (\boldsymbol{v}_i^{t-1,\tau_0} - \widehat{\boldsymbol{w}}^{t-1,\tau_0}) \right\| \tag{165}$$

$$\leq \left\| \boldsymbol{v}^{t-1,\tau_0-1} - \beta \sum_{i \in \mathcal{U}} p_i \nabla G_i(\boldsymbol{v}_i^{t-1,\tau_0-1}) \right. \tag{166}$$

$$\left. - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} + \beta \nabla G(\widehat{\boldsymbol{w}}^{t-1,\tau_0-1}) \right\| \tag{167}$$

$$\leq \left\| \boldsymbol{v}^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| \tag{168}$$

$$+ \beta \sum_{i \in \mathcal{U}} p_i \left\| \nabla G_i(\boldsymbol{w}_i^{t-1,\tau_0-1}) - \nabla G(\widehat{\boldsymbol{v}}^{t-1,\tau_0-1}) \right\| \tag{169}$$

$$\leq \left\| \boldsymbol{v}^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| \tag{170}$$

$$+ \beta L' \sum_{i \in \mathcal{U}} p_i \left\| \boldsymbol{v}_i^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\|, \tag{171}$$

where $\boldsymbol{v}^t$ and $\boldsymbol{v}_i^{t,\tau}$ are the global and local models in RDP-PFL without data sampling and DP mechanism in the training process, respectively. By invoking **Lemma 3** in [25], we have

$$\left\| \boldsymbol{v}_i^{t-1,\tau_0} - \widehat{\boldsymbol{w}}^{t-1,\tau_0} \right\| \tag{172}$$

$$= \left\| \boldsymbol{v}_i^{t-1,\tau_0-1} - \beta \nabla G_i(\boldsymbol{w}_i^{t-1,\tau_0-1}) \right. \tag{173}$$

$$\left. - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} + \beta \nabla G(\widehat{\boldsymbol{w}}^{t-1,\tau_0-1}) \right\| \tag{174}$$

$$\leq \left\| \boldsymbol{v}_i^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| + \beta \|\boldsymbol{n}_i^{\mathrm{S}}\| \tag{175}$$

$$+ \beta \left\| \nabla G_i(\boldsymbol{v}_i^{t-1,\tau_0-1}) - \nabla G(\widehat{\boldsymbol{w}}^{t-1,\tau_0-1}) \right\| \tag{176}$$

$$\leq \left\| \boldsymbol{v}_i^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| \tag{177}$$

$$+ \beta \left\| \nabla G_i(\boldsymbol{v}_i^{t-1,\tau_0-1}) - \nabla G_i(\widehat{\boldsymbol{v}}^{t-1,\tau_0-1}) \right\| \tag{178}$$

$$+ \beta \left\| \nabla G_i(\widehat{\boldsymbol{w}}^{t-1,\tau_0-1}) - \nabla G(\widehat{\boldsymbol{w}}^{t-1,\tau_0-1}) \right\|. \tag{179}$$

Due to **Lemma 1** and **Lemma 2**, we can obtain

$$\left\| \boldsymbol{v}_i^{t-1,\tau_0} - \widehat{\boldsymbol{w}}^{t-1,\tau_0} \right\| \tag{180}$$

$$\leq (1 + \beta L') \left\| \boldsymbol{v}_i^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| \tag{181}$$

$$+ \beta \left\| \nabla G_i(\widehat{\boldsymbol{w}}^{t-1,\tau_0-1}) - \nabla G(\widehat{\boldsymbol{w}}^{t-1,\tau_0-1}) \right\| \tag{182}$$

$$\leq (1 + \beta L') \left\| \boldsymbol{v}_i^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| \tag{183}$$

$$+ \beta(\varepsilon_i(1 + \eta L) + \eta B \gamma_i). \tag{184}$$

Recursively, we have

$$\mathbb{E}\left\{ \left\| \boldsymbol{v}_i^{t-1,\tau_0} - \widehat{\boldsymbol{w}}^{t-1,\tau_0} \right\| \right\}$$

$$\leq \mathbb{E}\left\{ \left\| \boldsymbol{v}_i^{t-1,0} - \widehat{\boldsymbol{w}}^{t-1,0} \right\| \right\} \tag{185}$$

$$+ \beta \left(\varepsilon_i (1 + \eta L) + \eta B \gamma_i \right) \sum_{j=0}^{\tau_0 - 1} (1 + \beta L')^j. \tag{186}$$

Due to $\boldsymbol{v}_i^{t-1,0} = \widehat{\boldsymbol{w}}^{t-1,0}$ and Eq. (161), we have

$$\mathbb{E}\left\{ \left\| \boldsymbol{w}_i^{t-1,\tau_0} - \widehat{\boldsymbol{w}}^{t-1,\tau_0} \right\| \right\} \tag{187}$$

$$\leq \frac{\varepsilon_i (1 + \eta L) + \eta B \gamma_i + \Theta}{L'} ((1 + \beta L')^{\tau_0} - 1). \tag{188}$$

Substituting (187) into (165), we have

$$\mathbb{E}\left\{ \left\| \boldsymbol{v}^t - \widehat{\boldsymbol{w}}^{t-1,\tau_0} \right\| \right\} \tag{189}$$

$$\leq \mathbb{E}\left\{ \left\| \boldsymbol{v}^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| \right\} \tag{190}$$

$$+ \beta L' \left((1 + \beta L')^{\tau_0 - 1} - 1 \right) \tag{191}$$

$$\cdot \sum_{i \in \mathcal{U}} \frac{p_i (\varepsilon_i (1 + \eta L) + \eta B \gamma_i)}{L'} \tag{192}$$

$$\leq \mathbb{E}\left\{ \left\| \boldsymbol{v}^{t-1,\tau_0-1} - \widehat{\boldsymbol{w}}^{t-1,\tau_0-1} \right\| \right\} \tag{193}$$

$$+ \beta(\varepsilon(1 + \eta L) + \eta B \gamma) \left((1 + \beta L')^{\tau_0 - 1} - 1 \right). \tag{194}$$

Recursively, we have

$$\mathbb{E}\left\{ \left\| \boldsymbol{v}^t - \widehat{\boldsymbol{w}}^{t-1,\tau_0} \right\| \right\} \triangleq h(\tau_0) \tag{195}$$

$$\leq \mathbb{E}\left\{ \left\| \boldsymbol{v}^{t-1,0} - \widehat{\boldsymbol{w}}^{t-1,0} \right\| \right\} \tag{196}$$

$$+ \beta(\varepsilon(1 + \eta L) + \eta B \gamma) \sum_{j=1}^{\tau_0} \left((1 + \beta L')^{j-1} - 1 \right) \tag{197}$$

$$\leq \beta(\varepsilon(1 + \eta L) + \eta B \gamma) \left(\frac{(1 + \beta L')^{\tau_0} - 1}{\beta L'} - \tau_0 \right). \tag{198}$$

This completes the proof. □

Appendix G
Proof of Lemma 6

According to the definition of $G_i(\boldsymbol{w})$, we have

$$[\nabla G_i(\boldsymbol{w}) - \nabla G_i(\boldsymbol{w}')]^\top (\boldsymbol{w} - \boldsymbol{w}') \tag{199}$$

$$\geq -\eta \rho B \|\boldsymbol{w} - \boldsymbol{w}'\|^2 + (1 - \eta L)(\boldsymbol{w} - \boldsymbol{w}')^\top \tag{200}$$

$$[\nabla F_i(\boldsymbol{w} - \eta \nabla F_i(\boldsymbol{w})) - \nabla F_i(\boldsymbol{w}' - \eta \nabla F_i(\boldsymbol{w}'))] \tag{201}$$

$$= -\eta \rho B \|\boldsymbol{w} - \boldsymbol{w}'\|^2 + (1 - \eta L)(\boldsymbol{w} - \boldsymbol{w}')^\top \tag{202}$$

$$[\nabla F_i(\boldsymbol{w} - \eta \nabla F_i(\boldsymbol{w})) - \nabla F_i(\boldsymbol{w} - \eta \nabla F_i(\boldsymbol{w}')) \tag{203}$$

$$+ \nabla F_i(\boldsymbol{w} - \eta \nabla F_i(\boldsymbol{w}')) - \nabla F_i(\boldsymbol{w}' - \eta \nabla F_i(\boldsymbol{w}'))]. \tag{204}$$

Because $F_i(\boldsymbol{w})$ is l-strongly convex, we can obtain

$$[\nabla G_i(\boldsymbol{w}) - \nabla G_i(\boldsymbol{w}')]^\top (\boldsymbol{w} - \boldsymbol{w}') \tag{205}$$

$$\geq -\eta \rho B \|\boldsymbol{w} - \boldsymbol{w}'\|^2 + (1 - \eta L)(l - \eta L^2) \|\boldsymbol{w} - \boldsymbol{w}'\|^2 \tag{206}$$

$$\geq l' \|\boldsymbol{w} - \boldsymbol{w}'\|^2, \tag{207}$$

where

$$l' = (1 - \eta L)(l - \eta L^2) - \eta \rho B. \tag{208}$$

This completes the proof. □

APPENDIX H
PROOF OF THEOREM 2

First, we can derive a convergence bound on the loss function of the centralized learning. Because $G(\cdot)$ is L'-Lipschitz smooth, we have

$$G(\widehat{\boldsymbol{w}}^{t,\tau+1}) - G(\widehat{\boldsymbol{w}}^{t,\tau}) \tag{209}$$

$$\leq \nabla G(\widehat{\boldsymbol{w}}^{t,\tau})^\top (\widehat{\boldsymbol{w}}^{t,\tau+1} - \widehat{\boldsymbol{w}}^{t,\tau}) + \frac{L'}{2}\|\widehat{\boldsymbol{w}}^{t,\tau+1} - \widehat{\boldsymbol{w}}^{t,\tau}\|^2 \tag{210}$$

$$= -\beta \nabla G(\widehat{\boldsymbol{w}}^{t,\tau})^\top \nabla G(\widehat{\boldsymbol{w}}^{t,\tau}) + \frac{L'\beta^2}{2}\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\|^2 \tag{211}$$

$$\leq -\beta\left(1 - \frac{L'\beta}{2}\right)\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\|^2. \tag{212}$$

Moreover, $G(\boldsymbol{w})$ is l'-strongly convex, and it follows that

$$G(\widehat{\boldsymbol{w}}^{t,\tau+1}) - G(\boldsymbol{w}^\star) \leq \frac{1}{2l'}\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau+1})\|^2. \tag{213}$$

Substituting (213) into (209), we obtain

$$G(\widehat{\boldsymbol{w}}^{t,\tau+1}) - G(\boldsymbol{w}^\star) \leq G(\widehat{\boldsymbol{w}}^{t,\tau}) - G(\boldsymbol{w}^\star) \tag{214}$$

$$-\beta\left(1 - \frac{L'\beta}{2}\right)\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\|^2 \tag{215}$$

$$\leq \left(1 + 2l'\beta - L'l'\beta^2\right)(G(\widehat{\boldsymbol{w}}^{t,\tau}) - G(\boldsymbol{w}^\star)). \tag{216}$$

Recursively, we have

$$G(\widehat{\boldsymbol{w}}^{t,\tau_0}) - G(\boldsymbol{w}^\star) \tag{217}$$

$$\leq \left(1 - 2l'\beta + L'l'\beta^2\right)^{\tau_0}(G(\widehat{\boldsymbol{w}}^{t,0}) - G(\boldsymbol{w}^\star)). \tag{218}$$

Due to $\widehat{\boldsymbol{w}}^{t,0} = \boldsymbol{w}^t$, we obtain

$$G(\widehat{\boldsymbol{w}}^{t,\tau_0}) - G(\boldsymbol{w}^\star) \tag{219}$$

$$\leq \left(1 - 2l'\beta + L'l'\beta^2\right)^{\tau_0}(G(\widehat{\boldsymbol{w}}^{t-1,\tau_0}) - G(\boldsymbol{w}^\star)) \tag{220}$$

$$+ \left(1 - 2l'\beta + L'l'\beta^2\right)^{\tau_0}(G(\boldsymbol{w}^t) - G(\widehat{\boldsymbol{w}}^{t-1,\tau_0})). \tag{221}$$

Then we want to bound $\|G(\boldsymbol{w}^t) - G(\widehat{\boldsymbol{w}}^{t-1,\tau_0})\|$ as follows:

$$\|G(\boldsymbol{w}^t) - G(\widehat{\boldsymbol{w}}^{t-1,\tau_0})\| \tag{222}$$

$$\leq \|G(\boldsymbol{w}^t) - G(\boldsymbol{v}^t)\| + \|G(\boldsymbol{v}^t) - G(\widehat{\boldsymbol{w}}^{t-1,\tau_0})\|, \tag{223}$$

where $\boldsymbol{v}^t$ is the model without data sampling and DP mechanism in the training process. We first bound $\|G(\boldsymbol{w}^t) - G(\widetilde{\boldsymbol{w}}^t)\|$ as

$$\|G(\boldsymbol{w}^t) - G(\boldsymbol{v}^t)\|$$

$$\leq \lambda(1 + \eta L)\|\boldsymbol{w}^t - \boldsymbol{v}^t\| \tag{224}$$

$$= \lambda(1 + \eta L)\left\|\sum_{i\in\mathcal{U}} p_i(\boldsymbol{w}_i^{t-1,\tau_0} - \boldsymbol{v}_i^{t-1,\tau_0})\right\| \tag{225}$$

$$\overset{(a)}{\leq} \lambda(1 + \eta L)\sum_{i\in\mathcal{U}} p_i \underbrace{\|\boldsymbol{w}_i^{t-1,\tau_0} - \boldsymbol{v}_i^{t-1,\tau_0}\|}_{H_1}, \tag{226}$$

where (a) is due to Jensen's inequation. Then we bound H_1 as

$$H_1 \leq \left\|\sum_{\tau=0}^{\tau_0-1}(\boldsymbol{w}_i^{t-1,\tau} - \beta\overline{\nabla}G_i(\boldsymbol{w}_i^{t-1,\tau})\right. \tag{227}$$

$$\left. - \boldsymbol{v}_i^{t-1,\tau} + \beta\nabla G_i(\boldsymbol{v}_i^{t-1,\tau}))\right\| \tag{228}$$

$$\overset{(b)}{\leq} \beta\sum_{\tau=0}^{\tau_0-1}\underbrace{\|\overline{\nabla}G_i(\boldsymbol{w}_i^{t-1,\tau}) - \nabla G_i(\boldsymbol{v}_i^{t-1,\tau})\|}_{H_2} \tag{229}$$

$$+ \sum_{\tau=0}^{\tau_0-1}\|\boldsymbol{w}_i^{t-1,\tau} - \boldsymbol{v}_i^{t-1,\tau}\|, \tag{230}$$

where (b) is due to triangle inequality. We can bound H_2 as

$$H_2 \overset{(c)}{\leq} \|\overline{\nabla}G_i(\boldsymbol{w}_i^{t-1,\tau}) - \widetilde{\nabla}G_i(\boldsymbol{w}_i^{t-1,\tau})\| \tag{231}$$

$$+ \|\widetilde{\nabla}G_i(\boldsymbol{w}_i^{t-1,\tau}) - \nabla G_i(\boldsymbol{w}_i^{t-1,\tau})\| \tag{232}$$

$$+ \|\nabla G_i(\boldsymbol{w}_i^{t-1,\tau}) - \nabla G_i(\boldsymbol{v}_i^{t-1,\tau})\| \tag{233}$$

$$\overset{(d)}{\leq} \frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|} \tag{234}$$

$$+ \Xi(C, S, \sigma) + \|\nabla G_i(\boldsymbol{w}_i^{t-1,\tau}) - \nabla G_i(\boldsymbol{v}_i^{t-1,\tau})\| \tag{235}$$

$$\overset{(e)}{\leq} \frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|} \tag{236}$$

$$+ \Xi(C, S, \sigma) + L'\|\boldsymbol{w}_i^{t-1,\tau} - \boldsymbol{v}_i^{t-1,\tau}\|, \tag{237}$$

where (c) is due to triangle inequality, (d) is from **Lemma 3** and **Lemma 4**, and (e) is due to the L'-Lipschitz smoothness. Substituting (231) into (227), we obtain

$$H_1 \leq (1 + \beta L')\sum_{\tau=0}^{\tau_0-1}\|\boldsymbol{w}_i^{t-1,\tau} - \boldsymbol{v}_i^{t-1,\tau}\| \tag{238}$$

$$+ \beta\tau_0\left(\Xi(C, S, \sigma) + \frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}\right. \tag{239}$$

$$\left. + \frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\right). \tag{240}$$

Recursively, we obtain

$$H_1 \leq \beta\tau_0\left(\Xi(C, S, \sigma) + \frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}\right. \tag{241}$$

$$\left. + \frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\right)\sum_{j=0}^{\tau_0-1}(\tau_0 - j)(1 + \beta L')^j. \tag{242}$$

Substituting Eq. (241) into Eq. (224), we obtain

$$\|G(\boldsymbol{w}^t) - G(\boldsymbol{v}^t)\| \tag{243}$$

$$\leq \beta\tau_0\lambda(1 + \eta L)\sum_{i\in\mathcal{U}} p_i\left(\Xi(C, S, \sigma) + \frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}\right. \tag{244}$$

$$\left. + \frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} + \frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\right)\sum_{j=0}^{\tau_0-1}(\tau_0 - j)(1 + \beta L')^j. \tag{245}$$

Substituting Eqs. (243) and (195) into Eq. (224), we obtain

$$\|G(\boldsymbol{w}^t) - G(\widehat{\boldsymbol{w}}^{t-1,\tau_0})\| \tag{246}$$

$$\le \beta\tau_0\lambda(1+\eta L)\sum_{i\in\mathcal{U}} p_i\Bigg(\Xi(C,S,\sigma)+\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} \tag{247}$$

$$+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\Bigg)\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j \tag{248}$$

$$+\lambda(1+\eta L)h(\tau_0). \tag{249}$$

Substituting (246) into (217), we have

$$G(\widehat{\boldsymbol{w}}^{T-1,\tau_0})-G\left(\boldsymbol{w}^\star\right) \tag{250}$$

$$\le \left(1-2l'\beta+L'l'\beta^2\right)^{\tau_0}\left(G(\widehat{\boldsymbol{w}}^{T-2,\tau_0})-G(\boldsymbol{w}^\star)\right) \tag{251}$$

$$+\left(1-2l'\beta+L'l'\beta^2\right)^{\tau_0}\lambda\Bigg((1+\eta L)h(\tau_0) \tag{252}$$

$$+\beta\tau_0(1+\eta L)\sum_{i\in\mathcal{U}} p_i\Bigg(\Xi(C,S,\sigma)+\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} \tag{253}$$

$$+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\Bigg)\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j\Bigg). \tag{254}$$

Recursively, we obtain

$$G(\widehat{\boldsymbol{w}}^{T-1,\tau_0})-G\left(\boldsymbol{w}^\star\right)\le\zeta^{T\tau_0}(G(\boldsymbol{w}^0)-G(\boldsymbol{w}^\star)) \tag{255}$$

$$+\sum_{k=1}^{T-1}\zeta^{k\tau_0}\lambda\Bigg((1+\eta L)h(\tau_0) \tag{256}$$

$$+\beta\tau_0(1+\eta L)\sum_{i\in\mathcal{U}} p_i\Bigg(\Xi(C,S,\sigma)+\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} \tag{257}$$

$$+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\Bigg)\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j\Bigg), \tag{258}$$

where $\zeta=1-2l'\beta+L'l'\beta^2$. Further, Due to $G(\boldsymbol{w}^T)-G(\boldsymbol{w}^\star)=G(\widehat{\boldsymbol{w}}^{T-1,\tau_0})-G(\boldsymbol{w}^T)+G(\boldsymbol{w}^T)-G(\boldsymbol{w}^\star)$ and Eq. (246), we have

$$G(\boldsymbol{w}^T)-G(\boldsymbol{w}^\star)$$

$$\le\zeta^{T\tau_0}(G(\boldsymbol{w}^0)-G(\boldsymbol{w}^\star)) \tag{259}$$

$$+\frac{1-\zeta^{T\tau_0}}{1-\zeta^{\tau_0}}\lambda\Bigg((1+\eta L)h(\tau_0) \tag{260}$$

$$+\beta\tau_0(1+\eta L)\sum_{i\in\mathcal{U}} p_i\Bigg(\Xi(C,S,\sigma)+\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} \tag{261}$$

$$+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\Bigg)\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j\Bigg). \tag{262}$$

This completes the proof. □

Appendix I
Proof of Theorem 3

Based on Eq. (209), we obtain

$$G(\widehat{\boldsymbol{w}}^{t,\tau_0})-G(\widehat{\boldsymbol{w}}^{t,0}) \tag{263}$$

$$\le-\beta\left(1-\frac{L'\beta}{2}\right)\sum_{\tau=0}^{\tau_0-1}\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\|^2. \tag{264}$$

Due to $\widehat{\boldsymbol{w}}^{t,0}=\boldsymbol{w}^t$ and Eq. (246), we have

$$G(\boldsymbol{w}^{t+1})-G(\boldsymbol{w}^t) \tag{265}$$

$$\le-\beta\left(1-\frac{L'\beta}{2}\right)\sum_{\tau=0}^{\tau_0-1}\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\|^2 \tag{266}$$

$$+\beta\tau_0\lambda(1+\eta L)\sum_{i\in\mathcal{U}} p_i\Bigg(\Xi(C,S,\sigma)+\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} \tag{267}$$

$$+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\Bigg)\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j \tag{268}$$

$$+\lambda(1+\eta L)h(\tau_0). \tag{269}$$

Now summing above equation over $t=0,1,\ldots,T-1$ and rearranging the terms yield that

$$\beta\left(1-\frac{L'\beta}{2}\right)\sum_{t=0}^{T-1}\sum_{\tau=0}^{\tau_0-1}\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\|^2$$

$$\le G(\boldsymbol{w}^0)-G(\boldsymbol{w}^T) \tag{270}$$

$$+T\beta\tau_0\lambda(1+\eta L)\sum_{i\in\mathcal{U}} p_i\Bigg(\Xi(C,S,\sigma)+\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} \tag{271}$$

$$+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\Bigg)\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j \tag{272}$$

$$+T\lambda(1+\eta L)h(\tau_0). \tag{273}$$

Because $G(\boldsymbol{w}^0)-G(\boldsymbol{w}^T)\le G(\boldsymbol{w}^0)-G(\boldsymbol{w}^\star)$, we can conclude that

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{\tau_0-1}\|\nabla G(\widehat{\boldsymbol{w}}^{t,\tau})\|^2\le\frac{G(\boldsymbol{w}^0)-G(\boldsymbol{w}^\star)}{T\beta\left(1-\frac{\beta L'}{2}\right)} \tag{274}$$

$$+\frac{\lambda}{\beta\left(1-\frac{\beta L'}{2}\right)}\Bigg((1+\eta L)h(\tau_0) \tag{275}$$

$$+\beta\tau_0\lambda(1+\eta L)\sum_{i\in\mathcal{U}} p_i\Bigg(\Xi(C,S,\sigma)+\frac{\eta(1+\eta L)L\sigma_G}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}} \tag{276}$$

$$+\frac{\eta B\sigma_H}{\sqrt{|\mathcal{D}_i^{\mathrm{S}}|}}+\frac{\eta^2 L\sigma_G\sigma_H}{|\mathcal{D}_i^{\mathrm{S}}|}\Bigg)\sum_{j=0}^{\tau_0-1}(\tau_0-j)(1+\beta L')^j\Bigg). \tag{277}$$

This completes the proof. □

References

[1] Y. Deng, F. Bao, Q. Dai, L. F. Wu, and S. J. Altschuler, "Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning," *Nature Methods*, vol. 16, no. 4, pp. 311–314, Apr. 2019.

[2] X. Deng et al., "Blockchain assisted federated learning over wireless channels: Dynamic resource allocation and client scheduling," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 3537–3553, May 2023.

[3] P. Wu, J. Li, L. Shi, M. Ding, K. Cai, and F. Yang, "Dynamic content update for wireless edge caching via deep reinforcement learning," *IEEE Commun. Lett.*, vol. 23, no. 10, pp. 1773–1777, Oct. 2019.

[4] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[5] D. C. Nguyen et al., "Enabling AI in future wireless networks: A data life cycle perspective," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 553–595, 1st Quart., 2021.

[6] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, Jan. 2020.

[7] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 72–80, Apr. 2020.

[8] X. Deng et al., "Low-latency federated learning with DNN partition in distributed industrial IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 3, pp. 755–775, Mar. 2023.

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

[10] K. Wei et al., "Low-latency federated learning over wireless channels with differential privacy," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 290–307, Jan. 2022.

[11] M. Mendieta, T. Yang, P. Wang, M. Lee, Z. Ding, and C. Chen, "Local learning matters: Rethinking data heterogeneity in federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8387–8396.

[12] J. Zhang, S. Guo, X. Ma, H. Wang, W. Xu, and F. Wu, "Parameterized knowledge transfer for personalized federated learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 10092–10104.

[13] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 2089–2099.

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

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

[16] K. Ji, J. Yang, and Y. Liang, "Theoretical convergence of multi-step model-agnostic meta-learning," *J. Mach. Learn. Res.*, vol. 23, pp. 1–41, Jan. 2022.

[17] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. NeurIPS Conf.*, vol. 33, Dec. 2020, pp. 3557–3568.

[18] S. Lin, G. Yang, and J. Zhang, "A collaborative learning framework via federated meta-learning," in *Proc. IEEE 40th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Singapore, Nov. 2020, pp. 289–299.

[19] V. Kulkarni, M. Kulkarni, and A. Pant, "Survey of personalization techniques for federated learning," in *Proc. 4th World Conf. Smart Trends Syst., Secur. Sustainability*, Jul. 2020, pp. 794–797.

[20] K. Wei et al., "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, 2020.

[21] J. Li, M. Khodak, S. Caldas, and A. Talwalkar, "Differentially private meta-learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2020.

[22] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong, "Privacy-preserving personalized federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[23] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong, "Personalized federated learning with differential privacy," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9530–9539, Oct. 2020.

[24] I. Mironov, "Rényi differential privacy," in *Proc. IEEE Comput. Secur. Found. Symp. (CSF)*, Santa Barbara, CA, USA, Aug. 2017, pp. 263–275.

[25] S. Wang et al., "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

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

[27] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009. [Online]. Available: http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] W. Zhuang, Y. Wen, and S. Zhang, "Divergence-aware federated self-supervised learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2022.

[30] R. Hu, Y. Gong, and Y. Guo, "Federated learning with sparsification-amplified privacy and adaptive optimization," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1463–1469.

[31] C. Ma et al., "On safeguarding privacy and security in the framework of federated learning," *IEEE Netw.*, vol. 34, no. 4, pp. 242–248, Jul. 2020.

[32] Z. Luo, D. J. Wu, E. Adeli, and L. Fei-Fei, "Scalable differential privacy with sparse network finetuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5057–5066.

[33] K. Wei et al., "User-level privacy-preserving federated learning: Analysis and performance optimization," *IEEE Trans. Mobile Comput.*, vol. 21, no. 9, pp. 3388–3401, Sep. 2022.

[34] A. Martin et al., "Deep learning with differential privacy," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, Vienna, Austria, Oct. 2016, pp. 308–318.

[35] J. Lee and D. Kifer, "Concentrated differentially private gradient descent with adaptive per-iteration privacy budget," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1656–1665.

[36] N. Wu, F. Farokhi, D. Smith, and M. A. Kaafar, "The value of collaboration in convex machine learning with differential privacy," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Francisco, CA, USA, May 2020, pp. 304–317.

[37] T. Li, Z. Liu, V. Sekar, and V. Smith, "Privacy for free: Communication-efficient learning with differential privacy using sketches," 2019, *arXiv:1911.00972*.

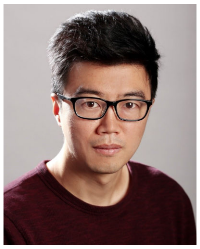
[38] J. Wu et al., "Hierarchical personalized federated learning for user modeling," in *Proc. Web Conf.*, Apr. 2021, pp. 957–968.

[39] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato, "Hypothesis testing interpretations and Renyi differential privacy," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Aug. 2020, pp. 2496–2506.

[40] I. Mironov, K. Talwar, and L. Zhang, "Rényi differential privacy of the sampled Gaussian mechanism," 2019, *arXiv:1908.10530*.

[41] A. Yousefpour et al., "Opacus: User-friendly differential privacy library in PyTorch," in *Proc. NeurIPS Workshop Privacy Mach. Learn.*, Dec. 2021.

Kang Wei (Member, IEEE) received the B.S. degree in information engineering from Xidian University, Xi'an, China, in 2014, and the Ph.D. degree from the Nanjing University of Science and Technology. He is currently a Post-Doctoral Fellow with The Hong Kong Polytechnic University. He mainly focuses on privacy protection and optimization techniques for edge intelligence, including federated learning, differential privacy, and network resource allocation.



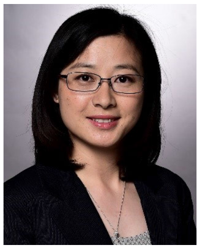
Jun Li (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009. From January 2009 to June 2009, he worked with the Department of Research and Innovation, Alcatel Lucent Shanghai Bell, as a Research Scientist. From June 2009 to April 2012, he was a Post-Doctoral Fellow with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia. From April 2012 to June 2015, he was a Research Fellow with the School of Electrical Engineering, The University of Sydney, Australia. Since June 2015, he has been a Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He was a Visiting Professor with Princeton University from 2018 to 2019. His research interests include network information theory, game theory, distributed intelligence, multiple agent reinforcement learning and their applications in ultra-dense wireless networks, mobile edge computing, network privacy and security, and the Industrial Internet of Things. He has coauthored more than 200 papers in IEEE journals and conferences and holds one U.S. patent and more than ten Chinese patents in these areas. He is serving as an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATION and a TPC member for several flagship IEEE conferences.

Chuan Ma (Member, IEEE) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013, and the Ph.D. degree from The University of Sydney, Australia, in 2018. From 2018 to 2022, he worked as a Lecturer with the Nanjing University of Science and Technology. He is currently a Principal Investigator with the Zhejiang Laboratory. He has published more than 40 journals and conference papers, including a best paper in WCNC 2018 and the Best Paper Award of IEEE Signal Processing Society in 2022. His research interests include stochastic geometry, wireless caching networks, and distributed machine learning, with a focus on the big data analysis and privacy-preserving.

Ming Ding (Senior Member, IEEE) received the B.S. and M.S. degrees (Hons.) in electronics engineering and the Doctor of Philosophy (Ph.D.) degree in signal and information processing from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2004, 2007, and 2011, respectively. From April 2007 to September 2014, he worked with the Sharp Laboratories of China, Shanghai, as a Researcher/Senior Researcher/Principal Researcher. Currently, he is a Principal Research Scientist with Data61, CSIRO, Sydney, NSW, Australia. He has authored more than 200 papers in IEEE journals and conferences, all in recognized venues, and around 20 3GPP standardization contributions and two books, i.e., *Multi-Point Cooperative Communication Systems: Theory and Applications* (Springer, 2013) and *Fundamentals of Ultra-Dense Wireless Networks* (Cambridge University Press, 2022). Also, he holds 21 U.S. patents and has co-invented another more than 100 patents on 4G/5G technologies. His research interests include information technology, data privacy and security, and machine learning and AI. Currently, he is an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He has served as a guest editor/co-chair/co-tutor/TPC member for multiple IEEE top-tier journals/conferences. He received several awards for his research work and professional services, including the prestigious IEEE Signal Processing Society Best Paper Award in 2022.

Wen Chen (Senior Member, IEEE) is currently a tenured Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is also the Director of the Broadband Access Network Laboratory. He has published more than 130 papers in IEEE journals and more than 120 papers in IEEE conferences, with citations of more than 9000 in Google Scholar. His research interests include multiple access, wireless AI, and meta-surface communications. He is a fellow of the Chinese Institute of Electronics and a Distinguished Lecturer of IEEE Communications Society and IEEE Vehicular Technology Society. He is also the Shanghai Chapter Chair of IEEE Vehicular Technology Society and an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE ACCESS, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.

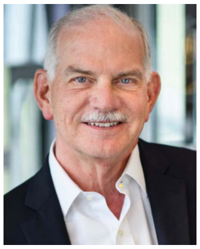
Jun Wu (Senior Member, IEEE) received the B.S. degree in information engineering and the M.S. degree in communication and electronic system from Xidian University in 1993 and 1996, respectively, and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications in 1999. He was a Professor with the Department of Computer Science and Technology, Tongji University. He was also a Principal Scientist with Broadcom before he joined Tongji University. He is currently a Full Professor with the School of Computer Science, Fudan University. His research interests include wireless networks, machine learning, and signal processing.



Meixia Tao (Fellow, IEEE) received the B.S. degree in electronic engineering from Fudan University, Shanghai, China, in 1999, and the Ph.D. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology in 2003.

She is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. Her current research interests include wireless edge learning, coded caching, reconfigurable intelligence surfaces, and semantic communications.

Dr. Tao received the 2019 IEEE Marconi Prize Paper Award, the 2013 IEEE Heinrich Hertz Award for Best Communications Letters, the IEEE/CIC International Conference on Communications in China (ICCC) 2015 Best Paper Award, and the International Conference on Wireless Communications and Signal Processing (WCSP) 2012 and 2022 Best Paper Award. She also received the 2009 IEEE ComSoc Asia–Pacific Outstanding Young Researcher Awards. She is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY and an Editor-at-Large of the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY. She served as a member of the Executive Editorial Committee for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2015 to 2019. She was also an Editorial Board Member of several other journals as an Editor or a Guest Editor, including the IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. She also served as the TPC Co-Chair for IEEE ICC 2023.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science and their applications in wireless networks, and energy systems and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is also a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.