mathematics

MDPI

Article

Privacy-Enhanced Federated Learning for Non-IID Data

Qingjie Tan, Shuhui Wu * and Yuanhong Tao

School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China; q_taner@163.com (Q.T.); taoyuanhong12@zust.edu.cn (Y.T.)
* Correspondence: s.wu@zust.edu.cn or swuhzh@163.com

Abstract: Federated learning (FL) allows the collaborative training of a collective model by a vast number of decentralized clients while ensuring that these clients' data remain private and are not shared. In practical situations, the training data utilized in FL often exhibit non-IID characteristics, hence diminishing the efficacy of FL. Our study presents a novel privacy-preserving FL algorithm, HW-DPFL, which leverages data label distribution similarity as a basis for its design. Our proposed approach achieves this objective without incurring any additional overhead communication. In this study, we provide evidence to support the assertion that our approach improves the privacy guarantee and convergence of FL both theoretically and empirically.

Keywords: federated learning; Hellinger distance; differential privacy; non-IID data

MSC: 68CS; 94ICC

1. Introduction

Federated learning (FL) enables the collaborative training of a shared model by multiple decentralized clients, eliminating the demand for direct data exchange between clients. The utilization of this paradigm guarantees the localization of the training data and the protection of clients' privacy [1]. Consequently, FL have gained significant traction in addressing numerous practical challenges across diverse fields, including the medical arena [2]. Nevertheless, the training data disseminated among numerous participating clients typically exhibit non-IID characteristics [3–5], which is a vital issue in the field of FL, as highlighted in reference [1]. The impact of label distributions in clients' training data on the overall performance of classification tasks has been observed [6]. Non-IID data significantly impact FL from two distinct perspectives: two primary factors contribute to the divergence of local models. One is that the data distributions vary significantly among various clients, and another is that the local data are imbalanced. The non-IID distributed data might result in a phenomenon known as "weight divergence" during a model's training process. Furthermore, this could have a detrimental effect on the efficiency of the global model [7,8].

One approach to address the aforementioned issue is mitigating the impact of data category imbalance on FL model by employing data augmentation techniques in situations where there exists a substantial disparity in data categories across various client datasets [9]. Nevertheless, the predominant obstacle in practical implementation is the inefficient utilization of communication resources resulting from the disproportionate allocation and dissemination of client data. The FedAvg algorithm, introduced by McMahan et al. [10], is widely acknowledged in this context. The system efficiently combines the model updates from several clients by utilizing a weighted averaging technique on the parameters. The system efficiently combines the model updates from several clients by employing a weighted averaging technique on the model parameters. This study examines the variation in client data while making the assumption that the global data follow the IID assumption. Furthermore, it is observed that there has been limited progress in enhancing the performance of

check for updates

Citation: Tan, Q.; Wu, S.; Tao, Y. Privacy-Enhanced Federated Learning for Non-IID Data. *Mathematics* **2023**, *11*, 4123. https://doi.org/10.3390/math11194123

Academic Editors: Zhaoquan Gu and Jianxin Li

Received: 16 August 2023
Revised: 15 September 2023
Accepted: 23 September 2023
Published: 29 September 2023

the algorithm. Based on this premise, researchers initiated investigations into enhanced FL methodologies, including FedProx [11], SCAFFOLD [12], and MOON [13], with the aim of enhancing the performance of FedAvg in non-IID data scenarios through refining local training procedures. Nevertheless, in this particular situation, it is worth noting that the enhancements achieved by FedProx may have been more gratifying, whereas the SCAFFOLD and MOON approaches imposed a considerable amount of supplementary communication overhead. The enhanced FL algorithm for non-IID data aims to optimize the aggregation weight in order to enhance the performance of FedAvg. The enhancement of aggregation weight places greater emphasis on the computation of similarity between the local and the global model [14,15]. But this approach has significant storage and time costs and does not effectively merge variations in client data distribution. Hence, the primary focus of our research paper is centered in the non-IID data scenario. Our objective is to assess the similarity of the distribution of clients' data and subsequently modify the aggregate weight based on these findings. The objective is to mitigate the communication bottleneck.

While FL can provide a certain level of privacy protection by sharing model parameters like gradients, it is vital to consider the non-IID scenario. In such cases, attackers can deduce model parameter information, hence posing a potential danger of privacy breach [16]. To augment the privacy protection capacity of the model during transmission, current approaches integrate FL with additional privacy protection technologies. These technologies include Differential Privacy (DP) [17], Homomorphic Encryption [18], and Secure Multi-Party Computation [19]. DP, in particular, possesses both a rigorous mathematical foundation and the ability to quantify the level of data privacy protection through the concept of privacy budget. Currently, DP has emerged as a highly effective method for safeguarding data privacy in the context of FL. Two primary application approaches of DP are commonly utilized: centralized DP and localized DP. The conventional approach to distributed processing involves the centralization of data processing and storage on the central server. However, this strategy is susceptible to single-point failures and potential breaches of privacy [20]. The concept of localized data processing entails the distribution of data processing and protection tasks over multiple local devices, hence enhancing privacy safeguards. In [21], the authors offer a privacy protection strategy called FedProx at the client level. However, they do not provide sufficient evidence to establish that the scheme fully satisfies the notion of DP. The study in [22] offers a theoretical demonstration; however, it fails to consider the trade-off between privacy parameters and model utility. Hence, for the non-IID scenario, the pressing issue at hand pertains to the reduction of communication costs while simultaneously guaranteeing the privacy of FL. Hence, this study presents a privacy-preserving FL technique that leverages the similarity in distribution of client data.

The primary contributions of our study are as follows:

(1) To address the issue of suboptimal FL algorithm models resulting from non-IID data, we have put out a proposed scheme, which involves utilizing the Hellinger distance to quantify the disparity between the local data distributions of clients and the ideal balanced distribution. By doing so, we aim to alleviate the divergence in the model;
(2) To address the issue of excessive communication usage in FL while dealing with non-IID data, we propose an aggregation technique that incorporates similarity weighting. This method leverages the similarity results obtained from analyzing the data distribution of each client, allowing for fast transfer of local model information to the Parameter Server (PS);
(3) To address the privacy disclosure issue in FL, we employ DP as a solution. During the training process, Gaussian noise is incorporated into the client's output in order to enhance privacy and security measures.

The remainder of our paper is organized as follows. Following this introduction, the relevant preliminary is presented in Section 2 and the proposed system model for the privacy-preserving FL algorithm the context of non-IID is presented in Section 3. The findings pertaining to privacy theory are presented in Section 4, whereas the results

concerning convergence theory can be found in Section 5. In Section 6, the discussion is given. Finally, the concluding remarks are presented in Section 7.

2. Preliminary

This section primarily outlines the fundamental framework of FL, elucidates the notion of DP mechanism, and examines the influence of non-IID data on model optimization.

2.1. Federated Learning

FL refers to a collaborative training procedure that involves the interaction between local clients and PS [19]. Supposing a standard FL system with N local clients and a PS, each client $k \in \{1, 2, \ldots, N\}$ has its private training dataset D_k, and the dataset size is n_k; here, $D_k = \left\{ \left(u_i^{(k)}, v_i^{(k)} \right) \right\}_{i=1}^{n_k}$, $u_i^{(k)}$ represents data point i of client k, and $v_i^{(k)}$ indicates the label of the data point i of client k. The client communicates with the PS to train the global model cooperatively without transmitting the original data. Therefore, the optimization problem of FL could be described as

$$\min_w F(w) \triangleq \sum_{k=1}^{N} p_k F_k(w), \tag{1}$$

where $F(w)$ denotes a global objective function, $w \in R^d$ stands for a model parameter vector, $p_k = n_k / \sum_{k=1}^{N} n_k$ refers to aggregate weights, and $F_k(w)$ denotes the local target function of client k. Specifically, we assume that the n_k training data of the client k is $D_k = \left\{ \left(u_1^{(k)}, v_1^{(k)} \right), \left(u_2^{(k)}, v_2^{(k)} \right), \ldots \left(u_{n_k}^{(k)}, v_{n_k}^{(k)} \right) \right\}$; then, the local objective function $F_k(w)$ can be defined as

$$F_k(w) \triangleq \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(w; u_i^{(k)}, v_i^{(k)}), \tag{2}$$

where $\ell(\cdot)$ refers to the loss function specified by the client. Cross-entropy is often used as a loss function in image recognition tasks.

The FedAvg algorithm [21] is a commonly employed approach in federated optimization. The PS computes the average of the local model parameters submitted by individual clients and thereafter distributes the aggregated outcomes to each client. In the conventional FedAvg algorithm, the initial step involves the client retrieving the latest global model parameters from the PS and subsequently initializing the local model. In the scenario where clients are chosen at random for training, the selected clients engage in training the local data update model. This is achieved by individually executing E rounds of random gradient descent steps and afterwards reporting the results to the PS. Ultimately, the PS obtains the model that has been modified locally and proceeds to aggregate it through an averaging process.

2.2. Differential Privacy

Dwork et al. proposed DP [23] to solve the privacy protection problem in databases. As a proven privacy protection technology, DP can ensure that the impact of a single sample on the whole is always lower than a certain threshold when outputting information, which makes it impossible for attackers to analyze the situation of a single sample from the change in output.

Definition 1. ((ε, δ)-DP [23]): *Consider any two neighboring datasets D and D', which differ in only one data sample. A randomized mechanism $M : \mathrm{D} \to \mathbb{R}$ with domain D and range $\mathbb{R}$ guarantees (ε, δ)-DP ((ε, δ)-DP), and, for any subsets of outputs $S \subset \mathbb{R}$, it holds that*

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D\prime) \in S] + \delta. \tag{3}$$

This ensures that the output of (ε, δ)-DP is indistinguishable, regardless of differences in a single record. The parameter $\varepsilon > 0$ is known as the privacy budget; smaller ε indicates stronger privacy protection level, and $\delta \in [0, 1]$ represents the probability to break the ε-DP.

Typically, ε-DP and (ε, δ)-DP assurance can be achieved through the Laplace Mechanism and Gaussian Mechanism, but this paper focuses on guaranteeing (ε, δ)-DP by adding random noise that conforms to the Gaussian distribution $N(0, \sigma^2 I^d)$ to the output function. To meet the requirements of DP, this mechanism controls the noise variance within a certain range to ensure that it meets the following conditions:

$$\sigma^2 \geq \frac{2(\Delta f)^2 \log(1.25/\delta)}{\varepsilon^2}, \tag{4}$$

Here, the notation $\Delta f \triangleq \max_{D,D'} \|f(D) - f(D')\|_2$ stands for the L_2-norm sensitivity of the function.

Sampling processes are frequently employed in machine learning algorithms. The privacy amplification characteristic, as suggested by differential privacy (DP) [24], demonstrates that the DP mechanism, when applied to a randomly chosen subset of the dataset, provides superior privacy safeguards compared to when it is applied to the entire dataset.

Lemma 1. (*Privacy amplification by subsampling [25]): If M is (ε, δ)-DP, then $M \circ Subsampling$ obeys (ε', δ')-DP, with $\varepsilon' = \log(1 + \gamma(e^{\varepsilon} - 1))$ and $\delta' = \gamma\delta$.*

The privacy amplification theorem demonstrates that, by sub-sampling the client, it is possible to effectively decrease the noise variance needed to attain the desired level of privacy protection, as specified by DP. In a broader sense, the lemma suggests that it is vital to exploit the randomness in sub-sampling because, if M is (ε, δ)-DP, then a sub-sampled mechanism with probability $\gamma < 1$ obeys $(O(\gamma\varepsilon), \gamma\delta)$-DP for a sufficiently small ε.

2.3. Impact of Non-IID Data

In every global iteration of FL, each client aims to reduce their loss function based on its local data. The existence of non-IID attributes in the local dataset can result in significant discrepancies between the local and the global model. In certain instances, it has been shown that the gradient of local models may exhibit a contrasting direction compared to that of the global model, leading to a phenomenon known as drift inside the local model [12,26]. Put differently, the revised local model exhibits a bias towards the local optimum and deviates from the global optimum state. Assume that the parameters of these local models are uploaded to the PS for the purpose of aggregation. The precision of the global model will be impacted, and there will also be a significant utilization of network capacity, resulting in a decrease in communication efficiency.

Figure 1 illustrates the FedAvg problem in both IID and non-IID scenarios. In IID scenarios, it can be observed that the global optimal value exhibits a strong proximity to the local ideal value. In other words, the global average model converges towards the global optimum. In non-IID scenarios, the discrepancy between the global optimal value and the local ideal value results in a considerable distance between the averaged global model and the global optimal state. Hence, it is imperative to investigate the methodologies for developing a proficient FL in non-IID scenarios.

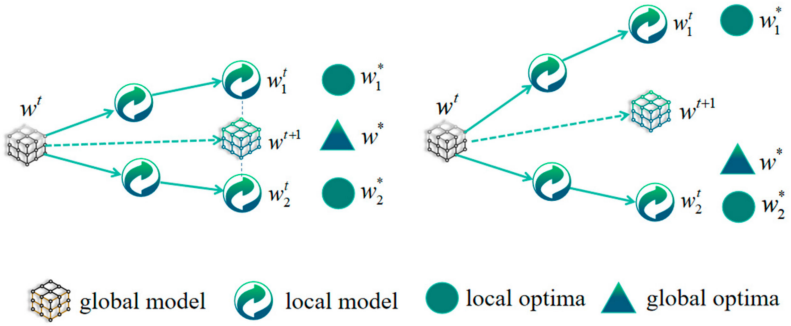


Figure 1. The FedAvg problem in IID and non-IID data.

3. System Model

In this section, we introduce the privacy-preserving FL algorithm (HW-DPFL), which is designed on the basis of the concept of probability distribution similarity of data labels. Subsequently, the method's specific process is described.

Firstly, it is vital to note that, in the FedAvg algorithm, the PS is responsible for aggregating and averaging the local model parameters. Thus, the effectiveness of FedAvg is greatly influenced by the weighting method employed. Typically, the weight assigned to each local dataset is determined by calculating the ratio of that dataset to the entire dataset. Nevertheless, in non-IID cases, this approach can have an impact on the rate of convergence and potentially compromise privacy. Hence, it is imperative to choose a more suitable approach for determining the weight. To address the problems at hand, this section presents a privacy-preserving FL approach called HW-DPFL, which leverages the similarity of probability distributions of data labels. The flow of the algorithm is depicted in Figure 2. During the process of model aggregation, the algorithm computes the Hellinger distance of the label distribution for each client's dataset. It then extracts the local model information from this calculation and aggregates it using an updated weighting approach. The proposed approach mitigates the challenges associated with training non-IID data and enhances the efficiency of model training.

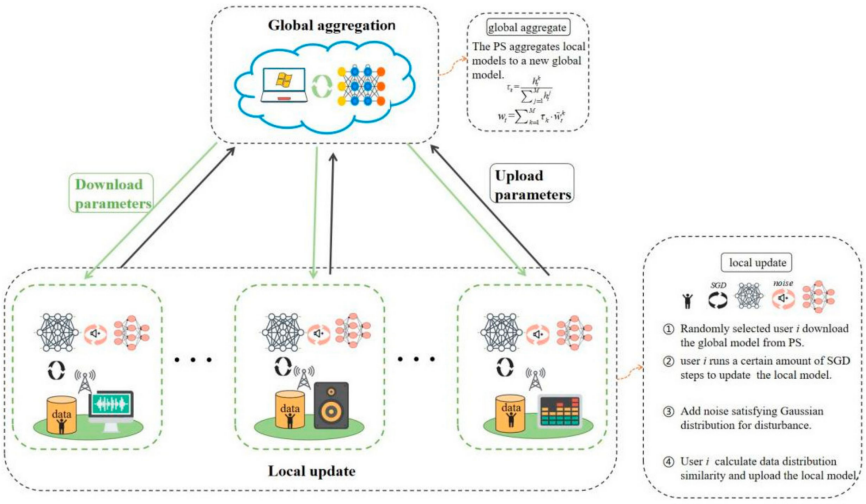


Figure 2. Schematic diagram of the HW-DPFL algorithm process.

In each iteration t, the label distribution of the client k dataset can be represented by the label vector G^k:

$$G^k = [n_{k,1}, n_{k,2}, \ldots, n_{k,C}] \tag{5}$$

where C denotes the total number of label types, $n_{k,C}$ indicates the number of C- type labels possessed by the client k.

The Hellinger distance is computed based on the label distribution G^k of the client k local dataset and the standard balanced data label distribution S:

$$h_t = \mathrm{H}(G,S) = \frac{1}{\sqrt{2}}\left\|\sqrt{G}-\sqrt{S}\right\|_2 = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}\left(\sqrt{G_i}-\sqrt{S_i}\right)^2} \tag{6}$$

Hellinger distance is a metric employed in the field of probability and statistics to quantify the degree of similarity between two probability distributions [27]. In the context of non-IID data, the Hellinger distance could be employed as a metric to assess the similarity between two classes, hence enabling algorithmic enhancements. Hence, the measure of similarity between each client's local dataset and the designated standard balanced dataset can be determined by computing the Hellinger distance.

The parameters for updating the weight of the model vary depending on the number of iterations:

$$\mathrm{w}_t = \sum_{k=1}^{M}\tau_k \cdot w_t^k, \tag{7}$$

$$\tau_k = \frac{h_t^k}{\sum_{j=1}^{M} h_t^j}. \tag{8}$$

Furthermore, given the PS's inclination towards honesty and curiosity while adhering to the FL protocol, it demonstrates a greater interest in the client's data information. Simultaneously, the system is susceptible to additional external attacks during the transmission of model parameters. To address this issue, we propose the incorporation of noise that adheres to a Gaussian distribution, thereby ensuring DP. Algorithm 1 provides a concise representation of the privacy-preserving FL method suggested in this research, which is founded on the concept of data label distribution similarity (HW-DPFL).

Algorithm 1: HW-DPFL

Input: K denotes the number of terminals; B denotes the local batch size; E denotes the local training times of the terminal model; F denotes the proportion of clients participating in training; η denotes learning rate; S denotes standard balanced data label distribution.
Output: model parameter
The PS does
Initialize global model parameters
for each round t = 1, 2, ⋯, do
$M \leftarrow \max(C \cdot K, 1)$ // Determine the number of clients for this round of communication
$S_t \leftarrow (random\,set\,of\,M\,client)$ // Randomly select M clients to participate in training
for each client $k \in S_t$ in parallel, do
$w_t^k \leftarrow ClientUpdate(k, w_{t-1})$
$h_t \leftarrow GetWeight(k)$
$w_t \leftarrow HW - DPFL\left(\widetilde{w}_t^k\right)$
def GetWeight(k): // Get the aggregation weight of user k
$G \leftarrow$(Get local dataset label distribution)
$h_t \leftarrow W(G,S)$ // Calculate the distribution similarity of user data labels
return h_t
def HW-DPFL(w,h): // Weighted aggregation
$w_t \leftarrow \frac{\sum_{k=1}^{K}\widetilde{w}_t^k \cdot h_t^k}{\sum_{k=1}^{K} h_t^k}$
return w_t
def ClientUpdate(k,w): // Model update
$B \leftarrow$(Batch Local Datasets)
for each local epoch from 1 to E do
for batch $b \in B$ do // Train each batch of data
$\widetilde{w}_t^k \leftarrow (w_t^k - \eta\nabla F_t^k(w;b)) + Z_t^k, Z_t^k \sim N(0, \sigma_{t,k}^2 I^d)$
return to server

During the process of model iteration, the method introduces noise to the model parameter information, thereby perturbing the data in a manner that significantly hinders the attacker's ability to extract meaningful information from it. The determination of noise parameters and privacy budget in DP is contingent upon the specific requirements for privacy protection. The combination theorem enables clients to effectively compute the privacy loss incurred throughout each iteration of the training process. In order to enhance clarity, the tth round, on the basis of the HW-DPFL algorithm, might be denoted as follows:

$$\text{Client } k: \begin{array}{l} w_t^k = ClientUpdate(k, w_{t-1}) \\ \widetilde{w}_t^k = w_t^k + Z_t^k, Z_t^k \sim N(0, \sigma^2 I^d) \end{array} \tag{9}$$

$$\text{Server}: \ \mathrm{w}_t = \sum_{k=1}^{M} \tau_k \cdot \widetilde{w}_t^k, \tag{10}$$

where w_t is the global model parameter of round t, w_t^k denotes the local model parameter of client k in round t, $ClientUpdate(k, w_{t-1})$ means the local random gradient descent process of client k, and d is the dimension of model parameters.

4. Privacy Analysis

In this section, we focus on the analysis of the privacy guarantees offered by the HW-DPFL algorithm. We begin by analyzing the sensitivity of the local parameter update function in relation to the L_2-norm. Following this, we proceed to assess the level of privacy guarantee in each subsequent iteration. Finally, we calculate the total privacy budget after the conclusion of all T iterations.

4.1. L_2-Norm Sensitivity

To achieve DP, we incorporate the Gaussian technique with L_2-norm sensitivity by introducing noise. Thus, we elucidate the sensitivity towards the local parameter updating function.

Assumption 1. *Suppose ζ_t^k is a uniform random sampling from the local data of client k in the iteration t. The squared norm of gradients for all clients is uniformly bounded, so $\left|\left|\nabla F_k(w_t^{k,e}; \zeta_t^{k,e})\right|\right|_2 \leq G^2$ for $k = 1, \ldots, N$, $e = 1, \ldots, E$, and $t = 1, \ldots, T$.*

Paper [21] *has successfully used Assumption 1 for DP-based research proof, as evidenced by the application of a gradient clipping methodology* [28].

Lemma 2. *If Assumption 1 holds, then the L_2-norm sensitivity of the local update parameters for user k in the iteration t is*

$$\Delta f_t^k \triangleq \max_{D, D\prime} ||w_t^k(D_k) - w_t^k(D_k')||_2 = 2\eta EG \tag{11}$$

The proof of Lemma 2 is shown in Appendix A.

4.2. Privacy Guarantee in Round T

Subsequently, a sub-sampling privacy amplification lemma is employed to mitigate the noise variance, ensuring that each client adheres to the noise variance constraint in every iteration.

Theorem 1. *Without replacement sampling in mini-batches, given that the noise level $\sigma_{t,k}^2$ and the added noise Z_t^k are obtained from sampling from a Gaussian distribution $N(0, \sigma_{t,k}^2 I^d)$, then we have*

$$\sigma_{t,k}^2 \geq \frac{32\gamma^2\eta^2E^2G^2\log(1.25\gamma/\delta)}{\varepsilon^2}, \tag{12}$$

where the sampling probability is $\gamma = Eb/n_k$.

Proof of Theorem 1. According to the privacy amplification by sub-sampling, the Gaussian noise level in fact can describe $(\log(1+\gamma(e^{\varepsilon}-1)),\gamma\delta)$-$DP$. Since

$$\log(1+\gamma(e^{\varepsilon}-1)) \leq \gamma(e^{\varepsilon}-1) \leq 2\gamma\varepsilon, \tag{13}$$

we can then obtain that the Gaussian noise level achieves at least $(2\gamma\varepsilon,\gamma\delta)$-$DP$. Specifically, in the iteration t, in order to satisfy the (ε,δ)-DP guarantee of client k, the Gaussian noise level can be decreased to

$$\begin{aligned}\sigma_{t,k}^2 &\geq \frac{8\left(\Delta f_t^k\right)^2\gamma^2\log(1.25\gamma/\delta)}{\varepsilon^2}\\ &= \frac{32\gamma^2\eta^2E^2G^2\log(1.25\gamma/\delta)}{\varepsilon^2}\end{aligned} \tag{14}$$

The proof is finished; the text continues here. □

4.3. The Total Privacy Loss

In this paper, we employ the moment accountant approach to quantify the cumulative privacy loss across T rounds. Our proposed methodology offers a more stringent constraint for quantifying the overall extent of privacy compromise compared to prior research efforts.

Theorem 2. *Assume that the noise Z_t^k obeys a Gaussian distribution $N(0,\sigma_{t,k}^2I^d)$; then, the HW-DPFL algorithm guarantees $(\hat{\varepsilon},\delta)$-DP. We have*

$$\hat{\varepsilon} = \varepsilon\left(\frac{T\log(1/\delta)}{2\log(1.25\gamma/\delta)}\right)^{\frac{1}{2}}. \tag{15}$$

Proof of Theorem 2. According to [28], we define the log of the moment-generating function evaluated at e for client k in iteration t as

$$\alpha_t^k(e) = \log\left(E_{\widetilde{w}_t^k}\left[\left(\frac{\Pr[\widetilde{w}_t^k|D_k]}{\Pr[\widetilde{w}_t^k|D_k']}\right)^e\right]\right). \tag{16}$$

Suppose that u_0 and u_1 stand for the probability density function of $N(0,\sigma_{t,k}^2I^d)$ and $N\left(\Delta f_t^k,\sigma_{t,k}^2I^d\right)$, respectively. Let u denotes the mixture of two Gaussian distributions as $u=(1-\gamma)u_0+\gamma u_1$. Therefore, we have

$$\alpha_t^k(e) = log(max(E_1,E_2)) \tag{17}$$

where

$$\begin{aligned}\mathrm{E}_1 &= E_{z\sim u_0}\left[\left(\frac{u_0(z)}{u(z)}\right)^e\right],\\ \mathrm{E}_2 &= E_{z\sim u}\left[\left(\frac{u(z)}{u_0(z)}\right)^e\right].\end{aligned} \tag{18}$$

According to composability for moment accountant method and Lemma 3 in [27], we have

$$\alpha_t^k(e) \leq \frac{T\gamma^2\left(\Delta f_t^k\right)^2e^2}{\sigma_{t,k}^2} = \frac{T\varepsilon^2e^2}{8\log(1.25\gamma/\delta)}. \tag{19}$$

Next, following Theorem 2.2 in [28], the HW-DPFL algorithm satisfies $(\hat{\varepsilon},\hat{\delta})$-$DP$. Here,

$$\begin{aligned}\hat{\delta} &= \min_{e\in Z^+}\exp(\alpha_k(e)-e\hat{\varepsilon})\\ &= \min_{e\in Z^+}\exp\left(\frac{T\varepsilon^2e^2}{8\log(1.25\gamma/\delta)}-e\hat{\varepsilon}\right)\end{aligned} \tag{20}$$

Since the above formula is a quadratic function of e, we assume that $\theta(x) = T\varepsilon^2 e^2/8log(1.25\gamma/\delta) - \hat{\varepsilon}e, e = 1, \ldots E$. Then,

$$\hat{\delta} < \exp(\theta(e^* + 1)), \tag{21}$$

where e^* is the minimum point of the function $\theta(x)$.

To make the HW-DPFL algorithm satisfy $(\hat{\varepsilon}, \hat{\delta})$-$DP$, let

$$\theta(e^* + 1) = \frac{T\varepsilon^2}{8\log(1.25\gamma/\delta)} - \frac{2\log(1.25\gamma/\delta)\hat{\varepsilon}^2}{T\varepsilon^2} \leq \log(\delta). \tag{22}$$

Thus, we have

$$\log(1/\delta) \leq -\frac{T\varepsilon^2}{8\log(1.25\gamma/\delta)} + \frac{2\log(1.25\gamma/\delta)\hat{\varepsilon}^2}{T\varepsilon^2} \leq \frac{2\log(1.25\gamma/\delta)\hat{\varepsilon}^2}{T\varepsilon^2}. \tag{23}$$

and

$$\hat{\varepsilon} \geq \varepsilon\left(\frac{T\log(1/\delta)}{2\log(1.25\gamma/\delta)}\right)^{\frac{1}{2}}. \tag{24}$$

The proof is finished, the text continues here. □

The coexistence of b and E adds to the acceleration of the convergence of Stochastic Gradient Descent (SGD) [29]. Moreover, as stated in Theorem 1, in cases when both b and E exhibit substantial magnitudes, it becomes imperative to provide a higher level of noise in order to guarantee differential privacy. However, this increased noise may potentially hinder the convergence of the algorithm. This suggests that there is a trade-off between the speed at which the algorithm reaches convergence and the degree of privacy protection. The aforementioned trade-off is subjected to further analysis in the future section.

5. Convergence Analysis

This section primarily focuses on the analysis of the convergence of the HW-DPFL algorithm described herein. Let us commence by establishing certain assumptions.

Assumption 2. *For all $k \in [N]$, each F_k is L-smooth, i.e., for all x and y, $F_k(x) \leq F_k(y) + (x-y)^T\nabla F_k(y) + L\|x-y\|^2/2$.*

Assumption 3. *For all $k \in [N]$, each F_k is μ strong convex, i.e., for all x and y, $F_k(x) \geq F_k(y) + (x-y)^T\nabla F_k(y) + \mu\|x-y\|^2/2$.*

Assumption 4. *For all $k \in [N]$, the stochastic gradients for each client satisfy $E[\nabla F_k(w_t^k; \zeta_t^k, b)] = \nabla F_k(w_t^k)$ and $E[\left\|\nabla F_k(w_t^k; \zeta_t^k, b) - \nabla F_k(w_t^k)\right\|^2] \leq \rho_k^2$.*

Let F^* and F_k^* denote the optimal values of total objective functions and objective function of the client k, respectively. According to [30], we assume that the degree of data heterogeneity can be expressed as $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$. It can be observed that, when the client data are IID, $\Gamma = 0$. The more heterogeneous the data, the greater the value of $|\Gamma|$.

Suppose w_t^k is the model parameter of the client k in the round t, and E is the total number of local epochs. The command set $\Omega_E = \{nE|n = 1, 2, \ldots\}$ represents the times of the client communicates with the PS. Considering that a subset of clients is randomly selected to participate the training according to the sampling scheme, at this time, if $t+1 \in \Omega_E$, it means that the PS aggregates the local models' parameters to obtain the global model and sends the latest model parameters to each client, if $t+1 \notin \Omega_E$, the client updates the local model parameters with its local data. Because clients participating in the training have to perform multiple rounds of iterations locally, we use an intermediate

variable v_{t+1}^k to represent the results of the one-step SGD, and the updated results can be expressed as

$$v_{t+1}^k = w_t^k - \eta \nabla F_k(w_t^k, \zeta_t^k) \tag{25}$$

$$w_{t+1}^k = \begin{cases} v_{t+1}^k, & t+1 \notin \Omega_E \\ \sum_{k=1}^{M} \tau_k v_{t+1}^k, & t+1 \in \Omega_E \end{cases}. \tag{26}$$

In order to enhance the comprehensibility of the proof, we shall introduce the subsequent lemma.

Lemma 3. *(Results for each round t) In iteration t, suppose that Assumptions 1 to 4 hold. Then,*

$$E\left[||\hat{v}_{t+1} - w^*||_2^2\right] \leq (1 - \mu\eta) E\left[||\hat{w}_t - w^*||_2^2\right] + \Psi + T\Lambda, \tag{27}$$

where

$$\Psi = 2(E-1)^2 \eta^2 G^2 + 2(M+2)\eta^2 L\Gamma + \eta^2 \sum_{k=1}^{M} \tau_k^2 \rho_k^2 \tag{28}$$

$$\Lambda = \mathrm{d} \sum_{k=1}^{M} \tau_k^2 \sigma_{t,k}^2 = \mathrm{d} \sum_{k=1}^{M} \tau_k^2 \frac{32\gamma^2 \eta^2 E^2 G^2 \log(1.25\gamma/\delta)}{\varepsilon^2}, \tag{29}$$

where w^* stands for the global optimal solution.

The proof of Lemma 3 is shown in Appendix B.

Theorem 3. *Suppose Assumptions 1 to 4 hold; then, the convergence rate of the HW-DPFL algorithm satisfies*

$$\begin{aligned} E\left[F(\hat{w}^T)\right] - F(w^*) &\leq \tfrac{L}{2} E\left[\|\hat{w}_T - w^*\|_2^2\right] \\ &\leq \tfrac{L(1-\mu\eta)^T}{2} E\left[\|\hat{w}_0 - w^*\|_2^2\right] + L\mu\eta\left(\tfrac{\Psi + T\Lambda}{2} + \tfrac{N-M}{N-1}\tfrac{2}{M}\eta^2 E^2 G^2\right) \end{aligned} \tag{30}$$

Proof of Theorem 3. If $t+1 \notin \Omega_E$, it can be observed that $\hat{w}_{t+1} = \hat{v}_{t+1}$. And if $t+1 \in \Omega_E$, the two are not equal. Assuming that there is no communication loss among the selected clients in each round, we hope that the model parameters obtained after sub-sampling and average aggregation are unbiased; thus in the HW-DPFL algorithm, when $t+1 \in \Omega_E$, we have

$$E_{S_t}[\hat{w}_{t+1}] = \hat{v}_{t+1}. \tag{31}$$

Here, it is used to express the expectation of the set S_t of randomly selected partial clients.

Lemma 4. (Bounding the variance of $\{\hat{w}_t\}$ [29]). *If PS samples S_t uniformly without replacement, then the variance of $\{\hat{w}_t\}$ is bounded by*

$$E_{S_t}\left[\|\hat{w}_{t+1} - \hat{v}_{t+1}\|_2^2\right] \leq \frac{N-M}{N-1}\frac{4}{M}\eta^2 E^2 G^2. \tag{32}$$

Note that

$$\begin{aligned} E\left[\|\hat{w}_{t+1} - w^*\|_2^2\right] &= E\left[\|\hat{w}_{t+1} - \hat{v}_{t+1} + \hat{v}_{t+1} - w^*\|_2^2\right] \\ &= \underbrace{E\left[\|\hat{w}_{t+1} - \hat{v}_{t+1}\|_2^2\right]}_{R_1} + \underbrace{E\left[\|\hat{v}_{t+1} - w^*\|_2^2\right]}_{R_2} + \underbrace{2E[\langle \hat{w}_{t+1} - \hat{v}_{t+1}, \hat{v}_{t+1} - w^* \rangle]}_{R_3}. \end{aligned} \tag{33}$$

Then, we use $E_{S_t}[\hat{w}_{t+1}] = \hat{v}_{t+1}$, and the term $R_3 = 0$.

Case 1. If $t+1 \notin \Omega_E$, then $R_1 = 0$ because $\hat{w}_{t+1} = \hat{v}_{t+1}$. According to Lemma 3, we have

$$E\left[\|\hat{w}_{t+1} - w^*\|_2^2\right] = E\left[\|\hat{v}_{t+1} - w^*\|_2^2\right] \leq (1-\mu\eta)E\left[\|\hat{w}_t - w^*\|_2^2\right] + \Psi + T\Lambda. \tag{34}$$

Case 2. If $t+1 \in \Omega_E$, according to Lemmas 3 and 4, it follows that

$$\begin{aligned} E\left[\|\hat{w}_{t+1} - w^*\|_2^2\right] &= E\left[\|\hat{v}_{t+1} - w^*\|_2^2\right] + E\left[\|\hat{w}_{t+1} - \hat{v}_{t+1}\|_2^2\right] \\ &\leq (1-\mu\eta)E\left[\|\hat{w}_t - w^*\|_2^2\right] + \Psi + T\Lambda + \tfrac{N-M}{N-1}\tfrac{4}{M}\eta^2E^2G^2. \end{aligned} \tag{35}$$

Unrolling the recursion, we can obtain

$$\begin{aligned} E\left[\|\hat{w}_T - w^*\|_2^2\right] &\leq (1-\mu\eta)^T E\left[\|\hat{w}_0 - w^*\|_2^2\right] + \sum_{t=1}^{T-1}(1-\mu\eta)^t\left(\Psi + T\Lambda + \tfrac{N-M}{N-1}\tfrac{4}{M}\eta^2E^2G^2\right) \\ &\leq (1-\mu\eta)^T E\left[\|\hat{w}_0 - w^*\|_2^2\right] + \mu\eta\left(\Psi + T\Lambda + \tfrac{N-M}{N-1}\tfrac{4}{M}\eta^2E^2G^2\right). \end{aligned} \tag{36}$$

Since $F_k(.)$ is L-smooth, we have

$$\begin{aligned} E\left[F(\hat{w}^T)\right] - F(w^*) &\leq \tfrac{L}{2}E\left[\|\hat{w}_T - w^*\|_2^2\right] \\ &\leq \tfrac{L(1-\mu\eta)^T}{2}E\left[\|\hat{w}_0 - w^*\|_2^2\right] + L\mu\eta\left(\tfrac{\Psi+T\Lambda}{2} + \tfrac{N-M}{N-1}\tfrac{2}{M}\eta^2E^2G^2\right) \end{aligned} \tag{37}$$

The proof is finished; the text continues here. □

By Theorem 3, the convergence upper bound of the HW-DPFL algorithm is affected by several factors, namely, the number of transmission rounds T, the mini-batch size b, the noise level σ, and the number of local update steps E. It is important to recognize that an increase in E has the potential to enhance the algorithm's convergence rate. The potential for enhancing convergence rates exists when the mini batch size b is increasing at the local level. Nevertheless, the algorithm's convergence rate may be impeded by the significant magnitudes of E and b. Increasing the degree of noise σ has the potential to improve the effectiveness of privacy measures. However, this may lead to a decrease in the rate of convergence.

6. Experiment

In this section, we assess the efficacy of the HW-DPFL. The experiments primarily employ Convolutional Neural Networks for the purpose of classifying the MNIST dataset.

MNIST dataset: The dataset was publicly provided by the National Institute of Standards and Technology. It is a binary image dataset, which consists of 70,000 grayscale images that have been manually scribbled. Each image is associated with a numerical designation ranging from 0 to 9. The resolution of the image is fixed at 28 × 28 pixels. For the MNIST dataset, a resolution of 28 × 28 has been considered a relatively low resolution, which has been widely accepted and effectively applied in practice. Some image examples from the MNIST dataset are in Figure 3.

A total of 60,000 images were designated as the training dataset, while the remaining 10,000 images were allocated for testing the model. During the model training process, it is necessary to specify the overall number of clients and ensure an equitable distribution of 60,000 images among them. This allocation ensures that each client receives an equal share of 600 photographs. The proportion of customers whose data are dependent and identically distributed is established at 0.8.

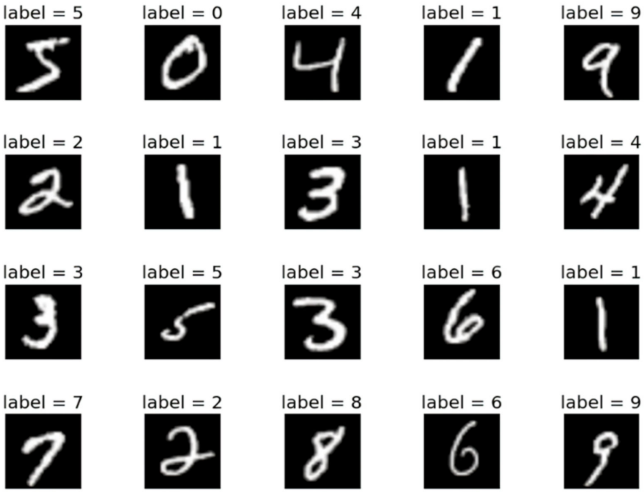


Figure 3. Size-normalized examples from the MNIST database.

Parameter setting: We set $\delta = 10^{-5}$ and the maximal local gradient norm to 1. It should be noted that the loss function is defined as cross-entropy and represents a highly convex optimization issue:

(1) Impact of local mini-batch size b: The impact of varying local mini-batch sizes on the training loss of the HW-DPFL algorithm is depicted in Figure 4. We set the values of local mini-batch size $b = \{10, 15, 20, 50\}$. Based on the findings from the experimental results, it is not difficult to see that an optimal state is present in two distinct contexts. In the IID case, it is observed that an increase b results in accelerated convergence and greater reduction in training loss. Nevertheless, the outcome is detrimental when the magnitude is above a specific threshold. The decrease in training loss is more pronounced when handling non-IID data, and the disparity in convergence between distinct b values is more noticeable;

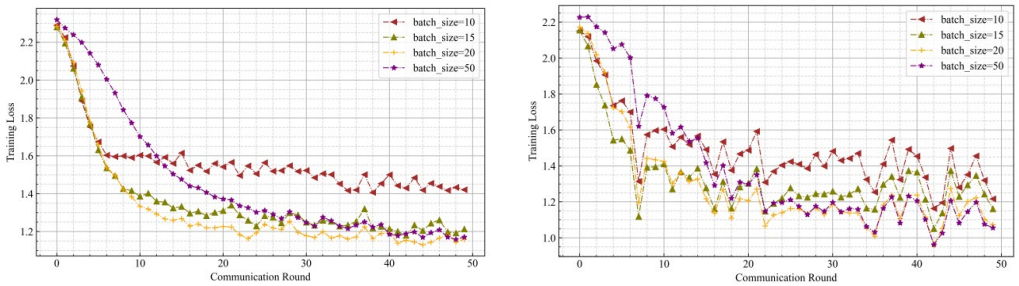
(2) Impact of the number of local update steps E: We also analyze the performance of the HW-DPFL algorithm with different local update steps. In the experiment, we set the number of local update steps as $E = \{1, 10, 20, 50\}$. The outcomes are depicted in Figure 5. For a fixed $\varepsilon = 1$, there is an optimal E value which makes the HW-DPFL perform the best in two scenarios. Moreover, increasing the value of E can result in expedited algorithm convergence. Nevertheless, the rate of convergence decelerates significantly when E is too large. In addition, for non-IID, an excessively large E results in a higher degree of variability in the training loss. The presence of larger E can result in more significant variations in weights among clients, hence impeding the convergence of the HW-DPFL;

(3) Impact of the noise level σ**:** The experience results of the HW-DPFL with different noise levels σ are presented in Figure 6. We set the noise level $\sigma = \{0.2, 0.5, 1\}$. The results indicate a steady decrease in training loss as the noise level increases. This can be attributed to the detrimental impact of high noise levels on the model's convergence performance, leading to a substantial increase in training loss. In both IID and non-IID cases, the training loss of the HW-DPFL exhibits an initial steep decline. In non-IID scenarios, the training loss experiences a greater reduction. Furthermore, the HW-DPFL has the potential to enhance the resilience of the training model in the face of DP injection noise;

The above experiments examines the impact of various factors on the efficacy of the HW-DPFL algorithm. It is evident that the HW-DPFL algorithm demonstrates superior performance across several data features. When the data follow the IID assumption, correctly raising the local mini-batch size b and the number of local update steps E enhances the convergence speed and decreases training losses. However, surpassing a particular threshold would result in the reverse effect. When dealing with non-IID data, it has been

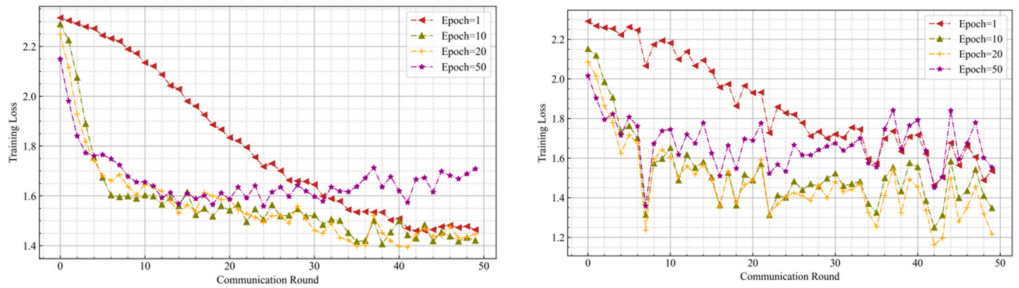
seen that increasing the value of b can effectively decrease training losses. However, it should be noted that, as the value of E increases, there is a corresponding increase in the variability of training losses. Furthermore, it is vital to consider the trade-off between utility and privacy when dealing with both IID and non-IID scenarios. The excessive noise level significantly impacts the convergence performance of the model. In circumstances where the data are non-IID distributed, the HW-DPFL algorithm exhibits reduced training losses and demonstrates enhanced capacity for improving the robustness of the model. Hence, the performance of the HW-DPFL algorithm can be enhanced through the adjustment of parameters such as the local mini-batch size b, the number of local update steps E, and the high noise level σ. In relation to the b value, it is imperative to select a suitable magnitude that aligns with the IID characteristics of the data. Regarding the E value, it is crucial to maintain control within a moderate range to prevent fluctuations and mitigate the adverse impact on the rate of convergence that may arise from an excessively large size. As for the σ value, it is essential to strike a balance between utility and privacy considerations, thereby opting for an appropriate level of noise that guarantees privacy while ensuring the desired level of utility. Implementing these modifications will enhance the training efficacy of the HW-DPFL algorithm and bolster the resilience of the model;

(4) Algorithm performance comparison: In IID and non-IID scenarios, HW-DPFL exhibits a greater level of accuracy compared to both the DP-FedAvg [8] and DP-FL [19]. Simultaneously, HW-DPFL demonstrates comparable accuracy to the DP-FL algorithm in the non-IID case in Table 1, thereby confirming the practicality and efficacy of the HW-DPFL algorithm in non-IID data scenarios.



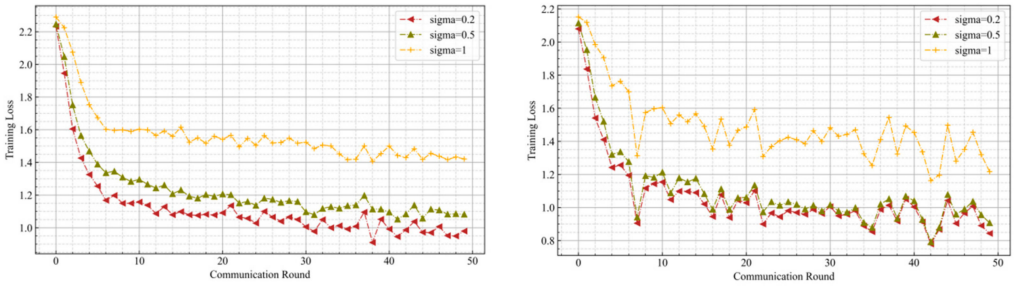
(**a**) IID data (**b**) Non-IID data

Figure 4. Comparing the impact of local mini-batch size on training loss in different data scenarios: (**a**) IID data. (**b**) Non-IID data.



(**a**) IID data (**b**) Non-IID data

Figure 5. Comparing the impact of the number of local update steps on training loss in different data scenarios: (**a**) IID data. (**b**) Non-IID data.



(**a**) IID data (**b**) Non-IID data

Figure 6. Comparing the impact of the noise level on training loss in different data scenarios: (**a**) IID data. (**b**) Non-IID data.

Table 1. Comparison accuracy of DP-FedAvg, DP-FL, and HW-DPFL.

	Clients	**Acc on**	**Acc on**	**Acc on**
		DP-FedAvg	**DP-FL**	**HW-DPFL**
IID data	100	96.41%	94.20%	96.67%
Non-IID data	100	90.03%	93.90%	95.21%

7. Discussion

This section examines three key aspects: data dissemination, privacy protection, and training time. In this study, we evaluate the efficacy of three techniques in the context of non-IID data, using heterogeneous and homogeneous models.

The primary focus of HW-DPFL lies in training non-IID data inside both homogeneous and heterogeneous models while emphasizing the implementation of robust privacy protection measures. The process of fine-tuning primarily takes place during the training stage and relies on weight aggregation. The Hellinger distance metric is also utilized to quantify the similarity between two probability distributions [26]. The performance of a system is influenced by the configuration of its models and the distribution of its data, both at the local level throughout numerous iterations and at the global level during aggregations. The substantial variance in updates leads to a departure of the global model from the genuine optimization outcomes.

Models are commonly perceived as entities that serve as repositories for storing knowledge derived from diverse datasets. The complexity of a model is influenced by various factors, including its structural design, dimensions, the distribution of data, and the size of the dataset. The augmentation of hidden units or parameters results in generalization mistakes. When various techniques are utilized to train the model under identical conditions, such as measuring the model complexity using CNN, it is noted that the accuracy of the model surpasses that of a shallow model. However, it is worth noting that the training time is extended.

During the preparation of this paper, it has come to our attention that a study conducted by [3] in the IEEE Internet of Things Journal in January 2023 explores an issue closely related to our research. It also investigates the application of FL to non-IID datasets, using DP techniques, yielding promising outcomes. However, it fails to address the content of our study adequately. Four distinct points of divergence exist between our paper and the work mentioned above: (1) The optimization of the gradient in FL was enhanced by [3] by utilizing historical gradient information. In contrast, our approach focuses on optimizing the gradient by adjusting the server-side aggregation strategy of parameters; (2) The reference [3] employs the K-means algorithm to cluster the label distribution of user data, aiming to address the issue of non-IID. In our work, we utilize the Hamming distance as a metric to quantify the difference between the IID and non-IID distribution; (3) Regarding the DP mechanism, [3] employs Laplace noise and a simple combination theorem to calculate privacy loss. In contrast, we introduce Gaussian noise and utilize the

moment accountant method to calculate privacy loss; (4) The reference [3] solely presents empirical experiments to demonstrate their results, while we provide theoretical proof of privacy and convergence in our work.

The HW-DPFL is configured with three distinct levels of noise, which are afterward linked to DP privacy protection. The hyper-parameters indicate the privacy protection level for both data and models.

8. Conclusions

This paper has studied an FL framework toward non-IID data, and a novel approach called HW-DPFL is proposed based on weighted aggregation of data distribution, which aims to improve FL's efficiency and protect the FL's privacy in non-IID data scenarios. Based on Hellinger distance, the algorithm quantifies the distribution balance degree of the clients' local privacy data labels to readjust the weight information of FL aggregation on PS so that the algorithm can converge faster while ensuring that the client information is fully trained. To effectively deal with the problem of information leakage, we add Gaussian noise to the shared parameters before uploading the parameters to PS. The algorithm can obtain local differential privacy with adjustable noise in FL architectures. Theoretical guarantees on the privacy protection capabilities and convergence of HW-DPFL were derived. The HW-DPFL algorithm was subsequently assessed using the MNIST dataset. The experimental findings exhibited the enhancement of HW-DPFL about non-IID data across several dimensions. The findings also suggest that HW-DPFL demonstrates potential usefulness and robust convergence in the face of non-IID data. Moreover, DP is incorporated into the upgraded FL framework to ensure the scheme's privacy.

Further research can be conducted to explore additional examinations of the theorems and the efficacy of HW-DPFL in future endeavors. Additionally, it is important to address various non-IID settings, such as feature-based non-IID scenarios. The potential strengths of DP-shuffle can be enhanced through the manipulation of various levels of noise. Furthermore, due to its nature as a local sample federated scheme, HW-DPFL has the potential for seamless integration into many upcoming federated learning frameworks as a fundamental operational component.

Author Contributions: Conceptualization, Q.T. and S.W.; methodology, Q.T. and S.W.; software, Q.T.; validation, Q.T., S.W. and Y.T.; formal analysis, Q.T.; investigation, Q.T.; data curation, Q.T.; writing—original draft preparation, Q.T.; writing—review and editing, Q.T., S.W. and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Zhejiang University of Science and Technology Postgraduate Research and Innovation Fund, grant number 2022yjskc24.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Lemma 2. During each iteration t, each client will initialize their local model with w_{t-1} and perform E steps of local SGD to obtain w_t^k, starting from w_{t-E},

$$w_t^k = w_{t-1} - \sum_{e=1}^{E} \eta g_t^{k,e}, \tag{A1}$$

where $g_t^{k,e}$ is the local gradient vector on the basis of the local datasets. Thus,

$$\begin{aligned} &||w_t^k(D_k) - w_t^k(D_k')||_2 \\ &= ||\sum_{e=1}^{E} \eta g_t^{k,e}(D_k) - \sum_{e=1}^{E} g_t^{k,e}(D_k')||_2 \\ &= \eta||\sum_{e=1}^{E} \eta g_t^{k,e}(D_k) - g_t^{k,e}(D_k')||_2 \\ &= \eta \sum_{e=1}^{E} ||g_t^{k,e}(D_k) - g_t^{k,e}(D_k')||_2 \\ &\leq 2\eta EG \end{aligned} \tag{A2}$$

where the last inequality is obtained from Assumption 1.

The proof is finished. □

Appendix B

Proof of Lemma 3. First, let $\hat{v}_t = \sum_{k=1}^{M} \tau_k v_t^k$, $\hat{w}_t = \sum_{k=1}^{M} \tau_k w_t^k$, $\hat{g}_t = \sum_{k=1}^{M} \tau_k \cdot \nabla F_k(w_t^k)$, and $g_t = \sum_{k=1}^{M} \tau_k \cdot \nabla F_k(w_t^k; \zeta_t^k, b)$. Then, $\hat{g}_t = E[g_t]$. Notice that $\hat{v}_{t+1} = \hat{w}_t - \eta g_t + Z_t$; then, we have

$$\begin{aligned} &E\left[\|\hat{v}_{t+1} - w^*\|_2^2\right] = E\left[\|\hat{w}_t - \eta g_t + Z_t - w^*\|_2^2\right] \\ &= E\left[\|\hat{w}_t - w^* - \eta g_t + Z_t\|_2^2\right] \\ &= \underbrace{E\left[\|\hat{w}_t - w^* - \eta g_t\|_2^2\right]}_{A_1} + \underbrace{E\left[\|Z_t\|_2^2\right]}_{A_2} \underbrace{-2\eta\langle \hat{w}_t - w^* - \eta\hat{g}_t, Z_t\rangle}_{A_3} \end{aligned} \tag{A3}$$

Because the added noise obeys the Gaussian distribution $N(0, \sigma_{t,k}^2 I^d)$, we have $A_3 = 0$. Next, we consider the bounding A_1. Note that

$$\begin{aligned} &E\left[\|\hat{w}_t - w^* - \eta g_t\|_2^2\right] = E\left[\|\hat{w}_t - w^* - \eta g_t + \eta\hat{g}_t - \eta\hat{g}_t\|_2^2\right] \\ &= \underbrace{E\left[\|\hat{w}_t - w^* - \eta\hat{g}_t\|_2^2\right]}_{B_1} + \underbrace{\eta^2 E\left[\|\hat{g}_t - g_t\|_2^2\right]}_{B_2} + \underbrace{2\eta E[\langle \hat{w}_t - w^* - \eta\hat{g}_t, \hat{g}_t - g_t\rangle]}_{B_3} \end{aligned} \tag{A4}$$

Then, $B_3 = 0$, according to $\hat{g}_t = E[g_t]$. Next, we prove that the term B_1 is bounded. We have

$$B_1 = E\left[\|\hat{w}_t - w^*\|_2^2\right] + \underbrace{\eta^2 E\left[\|\hat{g}_t\|_2^2\right]}_{C_1} \underbrace{-2\eta E[\langle \hat{w}_t - w^*, \hat{g}_t\rangle]}_{C_2}. \tag{A5}$$

Since $F_k(.)$ is L-smooth, it follows that

$$\left\|\nabla_k(w_t^k)\right\|_2^2 \leq 2L\left(F_k(w_t^k) - F_k^*\right). \tag{A6}$$

According to the convexity of $\|.\|_2^2$, it follows that

$$\left\|\nabla_k(w_t^k)\right\|_2^2 \leq 2L\left(F_k(w_t^k) - F_k^*\right). \tag{A7}$$

If any non-negative constants τ_k satisfy $\sum_{k=1}^{M} \tau_k = 1$, then

$$\begin{aligned} &\left\|\sum_{k=1}^{M} \tau_k(w_t^k)\right\|_2^2 \leq \sum_{k=1}^{M} \tau_k \left\|w_t^k\right\|_2^2, \\ &\left\|\sum_{k=1}^{M} w_t^k\right\|_2^2 \leq M \sum_{k=1}^{M} \left\|w_t^k\right\|_2^2 \end{aligned} \tag{A8}$$

Applying that the above formulas, we have

$$\begin{aligned} C_1 &= \eta^2 E\left[\|\hat{g}_t\|_2^2\right] \\ &= \eta^2 E\left[\left\|\sum_{k=1}^{M}\tau_k \nabla F_k(w_t^k)\right\|_2^2\right] \\ &\le \eta^2 M E\left[\sum_{k=1}^{M}\tau_k\left\|\nabla F_k(w_t^k)\right\|_2^2\right] \\ &\le 2ML\eta^2 E\left[\sum_{k=1}^{M}\tau_k\left(F_k(w_t^k)-F_k^*\right)\right]. \end{aligned} \tag{A9}$$

And we bound the term C_2 as follows:

$$\begin{aligned} C_2 &= -2\eta E[\langle \hat{w}_t - w^*, \hat{g}_t\rangle] \\ &= -2\eta E\left[\sum_{k=1}^{M}\tau_k\langle \hat{w}_t - w^*, \nabla F_k(w_t^k)\rangle\right] \\ &= -2\eta E\left[\sum_{k=1}^{M}\tau_k\langle \hat{w}_t - w_t^k, \nabla F_k(w_t^k)\rangle + \sum_{k=1}^{M}\tau_k\langle w_t^k - w^*, \nabla F_k(w_t^k)\rangle\right]. \end{aligned} \tag{A10}$$

Since the μ-strong convexity of $F_k(.)$ is true, it follows that

$$-\langle w_t^k - w^*, \nabla F_k(w_t^k)\rangle \le -\left(F_k(w_t^k) - F_k(w^*)\right) - \frac{\mu}{2}\left\|w_t^k - w^*\right\|_2^2. \tag{A11}$$

According to AM–GM inequality and Cauchy–Schwarz inequality, we have

$$-2\langle \hat{w}_t - w_t^k, \nabla F_k(w_t^k)\rangle, \nabla F_k(w_t^k)\rangle \le \frac{1}{\eta}\left\|\hat{w}_t - w_t^k\right\|_2^2 + \eta\left\|\nabla F_k(w_t^k)\right\|_2^2. \tag{A12}$$

Then,

$$\begin{aligned} C_2 &= -2\eta E\left[\sum_{k=1}^{M}\tau_k\langle \hat{w}_t - w_t^k, \nabla F_k(w_t^k)\rangle\right] - 2\eta E\left[\sum_{k=1}^{M}\tau_k\langle w_t^k - w^*, \nabla F_k(w_t^k)\rangle\right] \\ &\le \eta\sum_{k=1}^{M}\tau_k E\left(\frac{1}{\eta}\left\|\hat{w}_t - w_t^k\right\|_2^2 + \eta\left\|\nabla F_k(w_t^k)\right\|_2^2\right) - 2\eta\sum_{k=1}^{M}\tau_k E\left(\frac{\mu}{2}\left\|w_t^k - w^*\right\|_2^2 + F_k(w_t^k) - F_k(w^*)\right) \end{aligned} \tag{A13}$$

Combining (A5), (A9), and (A13), we have

$$\begin{aligned} B_1 &\le E\left[\|\hat{w}_t - w^*\|_2^2\right] + 2\mathrm{ML}\eta^2\sum_{k=1}^{M}\tau_k E\left[F_k(w_t^k) - F_k^*\right] \\ &+\eta\sum_{k=1}^{M}\tau_k E\left(\frac{1}{\eta}\left\|\hat{w}_t - w_t^k\right\|_2^2 + \eta\left\|\nabla F_k(w_t^k)\right\|_2^2\right) - 2\eta\sum_{k=1}^{M}\tau_k E\left(\frac{\mu}{2}\left\|w_t^k - w^*\right\|_2^2 + F_k(w_t^k) - F_k(w^*)\right) \\ &\le E\left[\|\hat{w}_t - w^*\|_2^2\right] - \mu\eta\sum_{k=1}^{M}\tau_k E\left[\left\|w_t^k - w^*\right\|_2^2\right] + \sum_{k=1}^{M}\tau_k E\left[\left\|\hat{w}_t - w_t^k\right\|_2^2\right] \\ &+2L\eta^2(M+1)\sum_{k=1}^{M}\tau_k E\left[F_k(w_t^k) - F_k^*\right] - 2\eta\sum_{k=1}^{M}\tau_k E\left[F_k(w_t^k) - F_k(w^*)\right] \\ &\le (1-\mu\eta)E\left[\|\hat{w}_t - w^*\|_2^2\right] + \sum_{k=1}^{M}\tau_k E\left[\left\|\hat{w}_t - w_t^k\right\|_2^2\right] \\ &\underbrace{+2L\eta^2(M+1)\sum_{k=1}^{M}\tau_k E\left[F_k(w_t^k) - F_k^*\right] - 2\eta\sum_{k=1}^{M}\tau_k E\left[F_k(w_t^k) - F_k(w^*)\right]}_{D} \end{aligned} \tag{A14}$$

where the third inequality comes from $\hat{w}_t = \sum_{k=1}^{M}\tau_t^k w_t^k$, and, from $\left\|\sum_{k=1}^{M}\tau_k(w_t^k)\right\|_2^2 \le \sum_{k=1}^{M}\tau_k\left\|w_t^k\right\|_2^2$, it follows that $\sum_{k=1}^{M}E\left[\left\|w_t^k - w^*\right\|_2^2\right] \ge E\left[\|\hat{w}_t - w^*\|_2^2\right]$.

We next aim to bound D. Let $\varphi = 2\eta(1-\eta L(M+1))$. Note that $\eta < 1/(L(M+1))$. We have $0 < \varphi < 2\eta$. Then,

$$\begin{aligned} D &= 2L\eta^2(M+1)\sum_{k=1}^{M}\tau_k E\Big[F_k(w_t^k) - F_k^*\Big] - 2\eta\sum_{k=1}^{M}\tau_k E\Big[F_k(w_t^k) - F_k(w^*)\Big] \\ &= \varphi\sum_{k=1}^{M}\tau_k E\Big[F_k(w^*) - F_k(w_t^k)\Big] + (2\eta-\varphi)\sum_{k=1}^{M}\tau_k E\big[F_k(w^*) - F_k^*\big] \\ &= \underbrace{-\varphi\sum_{k=1}^{M}\tau_k E\Big[F_k(w_t^k) - F_k(w^*)\Big]}_{J} + 2\eta^2 L(M+1)\Gamma, \end{aligned} \tag{A15}$$

where $\Gamma = F(\mathrm{w}^*) - \sum_{k=1}^{M}\tau_k F_k^*$.

For the term J, according to the convexity of $F_k(.)$ and AM–GM inequality, we find

$$\begin{aligned} J &= -\varphi\sum_{k=1}^{M}\tau_k E\Big[F_k(w_t^k) - F_k(w^*)\Big] \\ &= -\varphi\sum_{k=1}^{M}\tau_k E\Big[F_k(w_t^k) - F_k(\hat{\mathrm{w}}_t)\Big] - \varphi[F(\hat{\mathrm{w}}_t) - F(w^*)] \\ &\le -\varphi\sum_{k=1}^{M}\tau_k E\langle\nabla F_k(\hat{\mathrm{w}}_t), w_t^k - \mathrm{w}_{t-1}\rangle - \varphi[F(\hat{\mathrm{w}}_t) - F(w^*)] \\ &\le \varphi\sum_{k=1}^{M}\tau_k E\Big[\eta L\big(F_k(\hat{\mathrm{w}}_t) - F_k^*\big) + \tfrac{1}{2\eta}\Big\|w_t^k - \hat{\mathrm{w}}_t\Big\|_2^2\Big] - \varphi[F(\hat{\mathrm{w}}_t) - F(w^*)] \end{aligned} \tag{A16}$$

Thus, we can obtain

$$\begin{aligned} D &\le \varphi\sum_{k=1}^{M}\tau_k E\Big[\eta L\big(F_k(\hat{\mathrm{w}}_t) - F_k^*\big) + \tfrac{1}{2\eta}\Big\|w_t^k - \hat{\mathrm{w}}_t\Big\|_2^2\Big] - \varphi[F(\hat{\mathrm{w}}_t) - F(w^*)] + 2\eta^2 L(M+1)\Gamma \\ &= \varphi\sum_{k=1}^{M}\tau_k E\Big[\eta L(F_k(\hat{\mathrm{w}}_t) - F(w^*)) + \eta L\big(F(w^*) - F_k^*\big) + \tfrac{1}{2\eta}\Big\|w_t^k - \hat{\mathrm{w}}_t\Big\|_2^2\Big] \\ &-\varphi[F(\hat{\mathrm{w}}_t) - F(w^*)] + 2\eta^2 L(M+1)\Gamma \\ &= \varphi\eta L\sum_{k=1}^{M}\tau_k E[(F_k(\hat{\mathrm{w}}_t) - F(w^*))] - \varphi[F(\hat{\mathrm{w}}_t) - F(w^*)] + \tfrac{\varphi}{2\eta}\sum_{k=1}^{M}\tau_k E\Big[\Big\|w_t^k - \hat{\mathrm{w}}_t\Big\|_2^2\Big] \\ &+\eta L\Gamma(2\eta(M+1)+\varphi) \\ &= \varphi(\eta L - 1)[(F(\hat{\mathrm{w}}_t) - F(w^*))] + \tfrac{\varphi}{2\eta}\sum_{k=1}^{M}\tau_k E\Big[\Big\|w_t^k - \hat{\mathrm{w}}_t\Big\|_2^2\Big] + \eta L\Gamma(2\eta(M+1)+\varphi) \\ &\le \sum_{k=1}^{M}\tau_k E\Big[\Big\|w_t^k - \hat{\mathrm{w}}_t\Big\|_2^2\Big] + 2(M+2)\eta^2 L\Gamma \end{aligned} \tag{A17}$$

where the last inequality results from $\eta L - 1 < 0, 0 < \varphi < 2\eta$, and $\sum_{k=1}^{M} E[(F_k(\hat{\mathrm{w}}_t) - F(w^*))] = F(\hat{\mathrm{w}}_t) - F(w^*) \ge 0$.

Recalling the expression of B_1, we have

$$B_1 \le (1-\mu\eta)E\Big[\|\hat{w}_t - w^*\|_2^2\Big] + \underbrace{2\sum_{k=1}^{M}\tau_k E\Big[\Big\|\hat{w}_t - w_t^k\Big\|_2^2\Big]}_{Q} + 2(M+2)\eta^2 L\Gamma \tag{A18}$$

For the term Q, we analyze that HW-DPFL requires communication every E steps; then, we can bound the divergence of $\left\{w_t^k\right\}$:

$$\begin{aligned} Q &= \sum_{k=1}^{M} \tau_k E\left[\left\|\hat{w}_t - w_t^k\right\|_2^2\right] \\ &= \sum_{k=1}^{M} \tau_k E\left[\left\|(w_t^k - \hat{w}_{t_0}) - (\hat{w}_t - \hat{w}_{t_0})\right\|_2^2\right] \\ &\leq \sum_{k=1}^{M} \tau_k E\left[\left\|(w_t^k - \hat{w}_{t_0})\right\|_2^2\right] \\ &\leq \sum_{k=1}^{M} \tau_k \sum_{t=t_0}^{t-1} (E-1)\eta^2 \left\|\nabla F_k(w_t^k; \zeta_t^k, b)\right\|_2^2 \\ &\leq (E-1)^2 \eta^2 G^2 \end{aligned} \tag{A19}$$

To sum up, we can have

$$B_1 \leq (1-\mu\eta) E\left[\|\hat{w}_t - w^*\|_2^2\right] + 2(E-1)^2\eta^2 G^2 + 2(M+2)\eta^2 L\Gamma, \tag{A20}$$

where, in the first inequality, we use $E\|X - EX\|_2^2 \leq E\|X\|_2^2$ and $X = w_t^k - \hat{w}_{t_0}$, with probability τ_k. For any $t \geq 0$, there is a $t_0 < t < t_0 + E$ such that $t - t_0 \leq E - 1$ and $w_{t_0}^k = \hat{w}_{t_0}$. Note that we use Jensen inequality in the second inequality; it follows that

$$\left\|w_t^k - \hat{w}_{t_0}\right\|_2^2 = \left\|\sum_{t=t_0}^{t-1} \eta \nabla F_k(w_t^k; \zeta_t^k, b)\right\|_2^2 \leq (t-t_0) \sum_{t=t_0}^{t-1} \eta^2 \left\|\nabla F_k(w_t^k; \zeta_t^k, b)\right\|_2^2. \tag{A21}$$

We next focus on bounding the term B_2:

$$\begin{aligned} B_2 &= \eta^2 E\left[\|\hat{g}_t - g_t\|_2^2\right] \\ &= \eta^2 E\left[\left\|\sum_{k=1}^{M} \tau_k \cdot \left(\nabla F_k(w_t^k; \zeta_t^k, b) - \nabla F_k(w_t^k)\right)\right\|_2^2\right] \\ &\leq \eta^2 \sum_{k=1}^{M} \tau_k^2 \cdot E\left[\left\|\left(\nabla F_k(w_t^k; \zeta_t^k, b) - \nabla F_k(w_t^k)\right)\right\|_2^2\right] \\ &\leq \eta^2 \sum_{k=1}^{M} \tau_k^2 \rho_k^2 \end{aligned} \tag{A22}$$

Here in the last inequality, we use the variance of the stochastic gradients for each client, satisfying $E[\left\|\nabla F_k(w_t^k; \zeta_t^k, b) - \nabla F_k(w_t^k)\right\|^2] \leq \rho_k^2$.

Combining, (A20) and (A22), we have

$$A_1 \leq (1-\mu\eta) E\left[\|\hat{w}_t - w^*\|_2^2\right] + 2(E-1)^2\eta^2 G^2 + 2(M+2)\eta^2 L\Gamma + \eta^2 \sum_{k=1}^{M} \tau_k^2 \rho_k^2 \tag{A23}$$

where $\Psi = 2(E-1)^2\eta^2 G^2 + 2(M+2)\eta^2 L\Gamma + \eta^2 \sum_{k=1}^{M} \tau_k^2 \rho_k^2$.

Since the HW-DPFL algorithm guarantees $(\hat{\varepsilon}, \delta) - DP$, the Gaussian noise level can be represented as

$$\sigma_{t,k}^2 = \frac{32\gamma^2\eta^2 E^2 G^2 \log(1.25\gamma/\delta)}{\varepsilon^2} \tag{A24}$$

For the term A_2,we can have

$$A_2 = E\left[\|Z_t\|_2^2\right] = \mathrm{d} \sum_{k=1}^{M} \tau_k^2 \sigma_{t,k}^2 < T\Lambda \tag{A25}$$

where $\Lambda = \mathrm{d}\sum_{k=1}^{M}\tau_k^2\sigma_{t,k}^2 = \mathrm{d}\sum_{k=1}^{M}\tau_k^2\frac{32\gamma^2\eta^2E^2G^2\log(1.25\gamma/\delta)}{\varepsilon^2}$.

Combining (A23) and (A25), results for each iteration t can be expressed as

$$E\left[\|\hat{v}_{t+1} - w^*\|_2^2\right] \leq (1-\mu\eta)E\left[\|\hat{w}_t - w^*\|_2^2\right] + \Psi + T\Lambda. \tag{A26}$$

The proof is finished. □

References

1. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
2. Cazzato, G.; Massaro, A.; Colagrande, A.; Lettini, T.; Cicco, S.; Parente, P.; Nacchiero, E.; Lospalluti, L.; Cascardi, E.; Giudice, G.; et al. Dermatopathology of Malignant Melanoma in the Era of Artificial Intelligence: A Single Institutional Experience. *Diagnostics* **2022**, *12*, 1972. [CrossRef] [PubMed]
3. Li, Q.; Wen, Z.; Wu, Z.; Hu, S.; Wang, N.; Li, Y.; Liu, X.; He, B. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 3347–3366. [CrossRef]
4. You, X.; Liu, X.; Jiang, N.; Cai, J.; Ying, Z. Reschedule Gradients: Temporal Non-IID Resilient Federated Learning. *IEEE Internet Things J.* **2023**, *10*, 747–762. [CrossRef]
5. Ma, X.; Zhu, J.; Lin, Z.; Chen, S.; Qin, Y. A state-of-the-art survey on solving non-IID data in Federated Learning. *Future Generation Comput Syst.* **2022**, *135*, 244–258. [CrossRef]
6. Bassily, R.; Smith, A.; Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, Philadelphia, PA, USA, 18–21 October 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 464–473.
7. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated learning with non-IID data. *arXiv* **2018**, arXiv:1806.00582. [CrossRef]
8. Wang, H.; Yurochkin, M.; Sun, Y.; Dimitris Papailiopoulos, D.; Khazaeni, Y. Federated learning with matched averaging. *arXiv* **2020**, arXiv:2002.06440.
9. Duan, M.; Liu, D.; Chen, X.; Tan, Y.; Ren, J.; Qiao, L.; Liang, L. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications. In Proceedings of the 2019 IEEE 37th International Conference on Computer Design (ICCD), Abu Dhabi, United Arab Emirates, 17–20 November 2019; pp. 246–254.
10. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Aguera, B. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
11. Li, T.; Sahu, K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
12. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. Scaffold: Stochastic controlled averaging for federated learning. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5132–5143.
13. Li, Q.; He, B.; Song, D. Model-contrastive federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10713–10722.
14. Wu, H.; Wang, P. Fast-convergent federated learning with adaptive weighting. *IEEE Trans. Cogn. Commun. Netw.* **2021**, *7*, 1078–1088. [CrossRef]
15. Sattler, F.; Müller, K.R.; Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 3710–3722. [CrossRef]
16. Geiping, J.; Bauermeister, H.; Dröge, H.; Michael Moeller, M. Inverting gradients how easy is it to break privacy in federated learning? In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual, 6–12 December 2020; pp. 16937–16947.
17. Liu, D.; Simeone, O. Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 170–185. [CrossRef]
18. Ma, J.; Naas, A.; Sigg, S.; Lyu, X. Privacy-preserving federated learning based on multi-key homomorphic encryption. *Int. J. Intell. Syst.* **2022**, *37*, 5880–5893. [CrossRef]
19. Byrd, D.; Polychroniadou, A. Differentially private secure multi-party computation for federated learning in financial applications. In Proceedings of the First ACM International Conference on AI in Finance, New York, NY, USA, 15–16 October 2020; pp. 1–9.
20. Zhang, Y.; Huang, K.; Yang, J. Federated learning with privacy protection: A survey. *J. Syst. Eng. Electron.* **2021**, *32*, 797–809.
21. Geyer, C.; Klein, T.; Nabi, M. Differentially private federated learning: A client level perspective. *arXiv* **2017**, arXiv:1712.07557.
22. Shen, X.; Liu, Y.; Zhang, Z. Performance-enhanced federated learning with differential privacy for internet of things. *IEEE Internet Things J.* **2022**, *9*, 24079–24094. [CrossRef]
23. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]

24. Huang, Z.; Hu, R.; Guo, Y.; Chan-Tin, E.; Gong, Y. DP-ADMM: ADMM-based distributed learning with differential privacy. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 1002–1012. [CrossRef]
25. Balle, B.; Barthe, G.; Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *arXiv* **2018**. [CrossRef]
26. Li, Q.; Diao, Y.; Chen, Q.; He, B. Federated learning on non-iid data silos: An experimental study. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 9–12 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 965–978.
27. Wu, J.; Karunamuni, J. Profile Hellinger distance estimation. *Statistics* **2015**, *49*, 711–740. [CrossRef]
28. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318.
29. Stich, U.; Cordonnier, B.; Jaggi, M. Sparsified SGD with memory. *arXiv* **2018**. [CrossRef]
30. Li, X.; Huang, K.; Yang, W.; Wang, S.; Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv* **2019**, arXiv:1907.02189.