mathematics

MDPI

Article

Convergence Analysis for Differentially Private Federated Averaging in Heterogeneous Settings †

Yiwei Li [1], Shuai Wang [2] and Qilong Wu [1,*]

[1] Fujian Key Laboratory of Communication Network and Information Processing, Xiamen University of Technology, Xiamen 361024, China; lywei0306@gapp.nthu.edu.tw
[2] National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, Chengdu 611731, China; shuaiwang@uestc.edu.cn
* Correspondence: wql@xmut.edu.cn
† This article is an expanded version of a paper entitled "Secure federated averaging algorithm with differential privacy", which was presented at IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Espoo, Finland, 17–20 September 2020.

Abstract: Federated learning (FL) has emerged as a prominent approach for distributed machine learning, enabling collaborative model training while preserving data privacy. However, the presence of non-i.i.d. data and the need for robust privacy protection introduce significant challenges in theoretically analyzing the performance of FL algorithms. In this paper, we present novel theoretical analysis on typical differentially private federated averaging (DP-FedAvg) by judiciously considering the impact of non-i.i.d. data on convergence and privacy guarantees. Our contributions are threefold: (i) We introduce a theoretical framework for analyzing the convergence of DP-FedAvg algorithm by considering different client sampling and data sampling strategies, privacy amplification and non-i.i.d. data. (ii) We explore the privacy–utility tradeoff and demonstrate how client strategies interact with differential privacy to affect learning performance. (iii) We provide extensive experimental validation using real-world datasets to verify our theoretical findings.

Keywords: federated learning; convergence analysis; privacy analysis; data heterogeneity

MSC: 68W15



Received: 15 January 2025
Revised: 26 January 2025
Accepted: 27 January 2025
Published: 2 February 2025

Citation: Li, Y.; Wang, S.; Wu, Q. Convergence Analysis for Differentially Private Federated Averaging in Heterogeneous Settings. *Mathematics* **2025**, *13*, 497. https://doi.org/10.3390/math13030497

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid growth of the Internet of Things (IoT) in recent years, the number of intelligent devices has surged globally. Many of these devices are equipped with various sensors and increasingly powerful hardware, enabling them to collect and process data at unprecedented scales [1–3]. However, this growth has been accompanied by rising concerns about data privacy. Federated learning (FL) has emerged as a key solution in this context, which enables collaborative machine learning (ML) by allowing a parameter server (PS) to coordinate the training of a global model across distributed clients, thus eliminating the need to centralize sensitive data [4,5]. However, FL systems may involve millions of clients, and clients collect their own data based on their local environments and usage patterns; the size and distribution of local datasets vary significantly across clients [6]. This variation leads to data heterogeneity, as data generated by different clients are typically non-independent and identically distributed (non-i.i.d.) [7]. Data heterogeneity is also reflected in the diverse dynamics exhibited by different devices and their communication network.

Federated Averaging (FedAvg), proposed by [5], is one of the most popular FL algorithms. In FedAvg, a small subset of clients is randomly selected to perform local model

updates at each communication round, which are then aggregated by the PS to update the global model. Due to its simplicity and effectiveness, FedAvg serves as the foundation for most subsequent FL algorithms. Despite its widespread adoption, the convergence behavior of FedAvg under data heterogeneity remains poorly understood. Theoretical results, such as those in [8,9], suggest that data heterogeneity significantly impacts FedAvg, requiring many more communication rounds to converge as client gradients diverge. These findings are consistent with observations on synthetic or artificially partitioned datasets [10,11]. However, in many real-world FL tasks, FedAvg often performs remarkably well [5], contrary to these theoretical predictions. Furthermore, advanced methods designed to address data heterogeneity, such as SCAFFOLD [12], demonstrate significant advantages on synthetic heterogeneous datasets but exhibit similar performance to FedAvg on realistic datasets. This inconsistency between theory and real-world performance raises questions about the practical relevance of data heterogeneity and the adequacy of existing theoretical models for real-world FL systems.

On the other hand, FedAvg still faces significant challenges, particularly in ensuring robust privacy protections. Data privacy in FL is inherently at risk during the exchange of model parameters between clients and the PS, as adversaries may attempt to infer sensitive information from these exchanges [13]. Differential privacy (DP) has emerged as a key tool to address this risk by introducing calibrated noise into the learning process [14], offering mathematically rigorous privacy guarantees. However, the added DP noise can negatively affect model accuracy and algorithm convergence, requiring a careful balance between privacy and utility [15,16].

Currently, there are numerous works focused on analyzing the impact of data heterogeneity in FedAvg [17,18]. However, to the best of our knowledge, none of the existing works have provided a complete theoretical analysis on FedAvg for handling data heterogeneity together with the consideration of client and data sampling strategies as well as privacy amplification analysis, hence motivating us to develop an advanced theoretical framework for a DP-based FedAvg algorithm (DP-FedAvg) over non-i.i.d. data.

1.1. Related Work

This section provides a focused review of the literature to contextualize our work within the broader scope of FL research:

- FL in the presence of data heterogeneity: A substantial body of research has tackled the challenges posed by non-i.i.d. data in FL. Studies such as [6,19] indicate that the convergence rate of FL models in non-i.i.d. settings is significantly lower than in i.i.d. scenarios. These works also provide theoretical analyses suggesting that effective client selection schemes can improve convergence rates in non-i.i.d. scenarios. Many existing FL frameworks [5,11,20] rely on random client selection strategies [19]. Additionally, some works [21,22] consider device heterogeneity by taking into account factors like wireless channel capacity and computational power. However, a unified theoretical analysis that examines both non-i.i.d. data effects and client sampling strategies in a single framework remains unexplored.
- FL with privacy considerations: Privacy protection has become a critical focus in the FL system. The works [23,24] have explored DP as a robust mechanism to protect data privacy in FL, demonstrating its effectiveness in achieving privacy guarantees without incurring excessive computational overhead. Numerous privacy protection techniques based on DP algorithm in FL have been investigated, including DP-SGD [25], DP-FedAvg [1,26] and the DP-based primal-dual method (DP-PDM) [27], where additive noise is applied to local gradients/models. Notably, the impact of this additive noise on learning performance can be quantified through convergence analysis, and total

privacy loss can be readily tracked during the training process. However, few of these studies consider privacy amplification techniques [28,29] to reduce the adverse effect of DP noise, which could potentially improve the privacy–utility tradeoff.

To the best of our knowledge, none of the existing works have fully explored the interplay between DP noise and non-i.i.d. data on the convergence of the DP-FedAvg algorithm. Although various attempts have been made to integrate DP into FL algorithms, such as DP-SGD [1,26], DP-Prox [30] and DP-SCAFFOLD [31], the privacy–utility tradeoffs achieved by these methods remain suboptimal in non-i.i.d. settings, especially, they did not include the privacy amplification analyses. This poor privacy–utility tradeoff can be attributed to both data heterogeneity and the added DP noise, as existing FL works do not comprehensively integrate DP noise, privacy amplification, data heterogeneity, client selection and data sampling strategies within one work. This work addresses these gaps, providing robust privacy protections through DP while offering strong theoretical guarantees.

1.2. Contributions

The main contributions are summarized as follows:

- **Theoretical framework:** We propose a novel theoretical framework to analyze the convergence of DP-FedAvg under various client sampling strategies, with a particular focus on the impact of privacy protection and non-i.i.d. data. This framework enables us to quantify how client sampling methods and data heterogeneity influence the algorithm's convergence rate and privacy guarantees.
- **Privacy–utility tradeoff:** We explore the tradeoff between privacy protection and learning performance, taking into account the effects of privacy amplification. Our theoretical results demonstrate how various system parameters influence this tradeoff and highlights that the privacy protection level and gradient clipping play a crucial role in determining the privacy–utility tradeoff in practical FL systems.
- **Empirical validation:** We empirically validate our theoretical findings through extensive experiments on real-world datasets, demonstrating the behavior of DP-FedAvg under non-i.i.d. data conditions.

Synopsis: For ease of the ensuing presentation, all the mathematical notations used are listed in Table 1. Section 2 introduces the preliminaries of FL, DP and the existing DP-FedAvg algorithm. Section 3 provides the convergence analysis and privacy analysis for the DP-FedAvg algorithm. Experimental results are presented in Section 4, and Section 5 concludes the paper.

Table 1. Notations.

Notation	Definition
$\|\cdot\|$	Euclidean norm.
$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$	The inner product operator.
ϵ, δ	The DP parameter.
$\triangle$	The ℓ_2-norm sensitivity.
N	The total number of clients.
T	The total number of iterations.
Q	The number of local SGD updates.
$\|\mathcal{D}\|$	The size of set $\mathcal{D}$.
$\boldsymbol{w}^\top$	Transpose of vector $\boldsymbol{w}$.

Table 1. *Cont.*

Notation	Definition
$\boldsymbol{w}_t$	The global model at t-th iteration.
$\boldsymbol{w}_t^k$	The local model for k-th client at t-th iteration.
$F_k(\boldsymbol{w})$	The objective function for the client k.
$\nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right)$	The mini-batch gradient using the data $\mathcal{B}_t^k$ at client k.
Γ	The degree of non-i.i.d. data.
$\mathcal{G}$	The gradient clipping level.
n, n_k	The number of data samples for all client and the client k.
b	The size of mini-batch data.
d	The dimension of model parameters.
ϕ_k	The variance of stochastic gradients for the k-th client.
L	L-smooth parameter.
μ	μ-strong convex parameter.
$\mathbb{E}[\cdot]$, $\mathbb{P}[\cdot]$	The statistical expectation and the probability function, respectively.

2. Preliminaries

2.1. Federated Learning

We consider a vanilla FL system that consists of one parameter server (PS) and N clients in which each client has its local dataset $\mathcal{D}_k := \left\{\left(\boldsymbol{x}_{k,j}, \boldsymbol{y}_{k,j}\right), \forall k \in [N], j \in [n_k]\right\}$, where n_k and $n = \sum_{k=1}^N n_k$ denote the number of k-th client's training samples and its number of training samples, respectively. $\boldsymbol{x}_{k,j}$ denotes the training data sample and $\boldsymbol{y}_{k,j}$ is the corresponding label, while $[N]$ denotes the set $\{1, 2, \cdots, N\}$. The FL algorithm is to solve the following distributed optimization problem:

$$\min_{\boldsymbol{w}} \left\{F(\boldsymbol{w}) \triangleq \sum_{k=1}^N p_k F_k(\boldsymbol{w})\right\} \tag{1}$$

where $\boldsymbol{w}$ is the global model and $p_k = n_k/n$ is the weight of the k-th client and $\sum_{k=1}^N p_k = 1$; the F_k is the local objective function for client k, which is defined by

$$F_k(\boldsymbol{w}) = \frac{1}{n_k} \sum_{j=1}^{n_k} \ell\left(\boldsymbol{w}; \mathbf{x}_{k,j}, \mathbf{y}_{k,j}\right), \tag{2}$$

where $\ell(\cdot)$ is a user-specified loss function. Problem (1) can be solved using the well-known FedAvg algorithm [5]. Specifically, FedAvg involves T communication rounds, and each round contains the following steps:

1. First, at the beginning of round $t \in \{0, \ldots, T-1\}$, the PS selects a subset of clients $\mathcal{S}_t \subseteq [N]$ to participate and sends them the latest global model $\boldsymbol{w}_t$.
2. Each client $i \in \mathcal{S}_t$ initializes its local model $\boldsymbol{w}_{t+1}^k$ to be the global model $\boldsymbol{w}_t$ and then performs Q iterations of stochastic gradient descent (SGD) on its local dataset.

$$\boldsymbol{w}_{t+1}^k = \boldsymbol{w}_t, \tag{3}$$

$$\boldsymbol{w}_{t+j}^k = \boldsymbol{w}_{t+j-1}^k - \eta_{t+j-1}\left(\nabla F_k\left(\boldsymbol{w}_{t+j-1}^k, \mathcal{B}_{t+j-1}^k\right)\right), j \in [Q], k \in \mathcal{S}_t, \tag{4}$$

where η_{t+j-1} represents the learning rate, and $\mathcal{B}_{t+j-1}^k \subseteq \mathcal{D}^k$ is the mini-batch data with $|\mathcal{B}_{t+j-1}^k| = b$; $\nabla F_k\left(\boldsymbol{w}_{t+j-1}^k, \mathcal{B}_{t+j-1}^k\right)$ denotes the stochastic gradient.

3. Each client $i \in \mathcal{S}_t$ uploads the final local model update $\boldsymbol{w}_{t+Q}^k$ to the PS.
4. The PS aggregates the local model updates from all participating clients and improves the global model as

$$\boldsymbol{w}_{t+Q} = \sum_{k \in \mathcal{S}_t} p_k \boldsymbol{w}_{t+Q}^k. \tag{5}$$

The same procedure repeats for the next round until the algorithm converges.

2.2. Differential Privacy

Although FedAvg avoids the direct information leakage by keeping the data locally, the intermediate updates exchanged during the collaboration process such as $\boldsymbol{w}_{t+Q}^k$ and $\boldsymbol{w}_{t+Q}$ could still leak private information about the local data as demonstrated in recent advanced attacks such as model inversion attacks and membership attacks [32]. Figure 1 depicts the FL system model under attacks from an adversary. The adversary in this context can be either an "honest-but-curious" aggregation PS or clients within the system. The aggregation PS is assumed to honestly follow the designed training protocol but may be curious about clients' private data and attempt to infer it from the shared messages. To provide strong privacy protection for local data, we adopt (ϵ, δ)-DP, which is defined as follows.

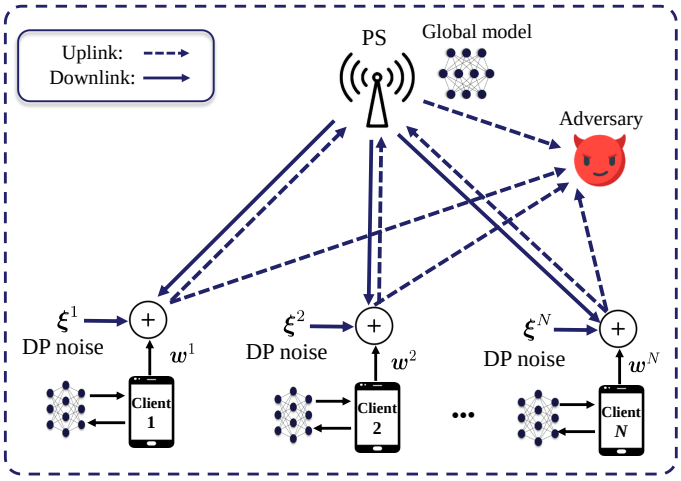


Figure 1. A typical FL system.

Definition 1 ((ϵ, δ)-DP [14]). *A randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if for any two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in only one record, and for any subsets of outputs $S \subseteq$ range$(\mathcal{M})$,*

$$\mathbb{P}\left[\mathcal{M}(\mathcal{D}) \in S\right] \leq \exp(\epsilon) \cdot \mathbb{P}\left[\mathcal{M}\left(\mathcal{D}'\right) \in S\right] + \delta, \tag{6}$$

where $\epsilon > 0$ accounts for privacy protection level, and $\delta \geq 0$ is the probability threshold to break $(\epsilon, 0)$-DP.

For a query function $g : \mathcal{X} \to \mathbb{R}$, the (ϵ, δ)-DP mechanism can be implemented by adding artificial Gaussian noise to function g as follows [14]:

$$\mathcal{M}(\mathcal{D}) \triangleq g(\mathcal{D}) + \mathcal{N}\left(0, \Delta^2 \cdot \sigma^2\right), \tag{7}$$

where Δ is the sensitivity of function g, defined by

$$\Delta = \max_{\mathcal{D},\mathcal{D}'} \|g(\mathcal{D}) - g(\mathcal{D}')\|, \tag{8}$$

in which $\mathcal{D}, \mathcal{D}' \subseteq \mathcal{X}$ are neighboring datasets. σ^2 denotes the minimal required "noise scale" for achieving (ϵ, δ)-DP given by the following lemma in [14].

Lemma 1. *Suppose that the query function g accesses the dataset $\mathcal{D}$ via randomized mechanism $\mathcal{M}$. Then, $\mathcal{M}$ satisfies (ϵ, δ)-DP if*

$$\sigma^2 \geq \frac{2\Delta^2 \ln(1.25/\delta)}{\epsilon^2}. \tag{9}$$

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

$$c(\boldsymbol{o}) = \ln\Big(\frac{\mathbb{P}\big[\mathcal{M}(\mathcal{D}) = \boldsymbol{o}\big]}{\mathbb{P}\big[\mathcal{M}(\mathcal{D}') = \boldsymbol{o}\big]}\Big). \tag{10}$$

It has been known that, running on a randomly generated subset of a dataset, the DP mechanism can yield stronger privacy protection than running on the entire dataset. This fact implies that the DP noise required for achieving a predefined DP level can be reduced when partial data are randomly selected at each local update. For the ease of later analysis, we restate the privacy amplification theorem as follows [7].

Theorem 1. *Suppose that a mechanism $\mathcal{M}$ is (ϵ, δ)-DP over a given dataset $\mathcal{D}$ with size $|\mathcal{D}| = n$. Consider the subsampling mechanism that outputs a random sample uniformly over all subsets $\mathcal{D}_s \subseteq \mathcal{D}$ with size $|\mathcal{D}_s| = b$. Then, when $\epsilon \leq 1$, executing $\mathcal{M}$ mechanism on the subset $\mathcal{D}_s$ guarantees (ϵ', δ')-DP, where ϵ' and δ' are given by*

$$\epsilon' = \min\big(2q\epsilon, \epsilon\big), \tag{11}$$

$$\delta' = q\delta, \tag{12}$$

where $q = b/n$ is the data sampling ratio when data are sampled without replacement.

According to Theorem 1, the privacy would be amplified when $q \leq 1/2$. Note that the privacy amplification is pervasively adopted in the existing literature on FL [1,7,23] since only a small portion of data are currently being used in local SGD.

2.3. The Typical DP-FedAvg Algorithm

In this subsection, we present the typical DP-FedAvg, which is summarized in Algorithm 1. The main steps follow the FedAvg framework outlined in Section 2.1. The key distinctions between DP-FedAvg and FedAvg are twofold: (i) Gaussian noise is added to local gradients to ensure robust privacy protection and (ii) gradient clipping is applied to the locally updated model. Specifically, to protect the privacy of the local model $\boldsymbol{w}_t^k$, we apply DP by adding Gaussian noise $\boldsymbol{\xi}_{t+1}^k \sim \mathcal{N}(0, (\sigma_{t+1}^k)^2 \mathbf{I}_d)$ to the local gradient as follows:

$$\boldsymbol{w}_{t+1}^k = \boldsymbol{w}_t^k - \eta_t\Big(\tilde{\boldsymbol{g}}_t^k + \boldsymbol{\xi}_{t+1}^k\Big), \tag{13}$$

where $\tilde{\boldsymbol{g}}_t^k$ is the clipped local gradient, defined by

$$\tilde{\boldsymbol{g}}_t^k = \boldsymbol{g}_t^k / \max\Big(1, \frac{\|\boldsymbol{g}_t^k\|}{\mathcal{G}}\Big), \tag{14}$$

in which g_t^k is the local stochastic gradient, given as follows:

$$g_t^k = \nabla F_k\left(w_t^k, \mathcal{B}_t^k\right). \tag{15}$$

It is important to note that determining the ℓ_2-sensitivity requires bounded local gradients. Therefore, gradient clipping is applied to the locally updated model. Specifically, if $\|g_{t+1}^k\| \geq \mathcal{G}$, the gradient is scaled down to have a norm of $\mathcal{G}$; otherwise, it remains unchanged. Gradient clipping is a standard operation in SGD with DP, particularly for determining sensitivity in FL systems.

Algorithm 1 Proposed DP-FedAvg algorithm

Input: Initial values of $w_0^1, \ldots, w_0^N$ at the clients, and initial value of w_0 at the server.
for $t = 0, 1, \ldots, T-1$ **do**
  Client side:
  for $k \in \mathcal{S}_t$ in parallel **do**
    Calculate local gradients: g_t^k via (15).
    Perform gradient clipping: $\tilde{g}_t^k$ via (14).
    if $t+1 \neq zQ, z = 1, 2, \ldots$ **then**
      Update $w_{t+1}^k = w_t^k - \eta_t g_t^k$
    else if $t+1 = zQ, z = 1, 2, \ldots$ **then**
      Update w_{t+1}^k via (13).
    end if
    Server side:
    Update $w_{t+1} = \sum_{k=1}^K p_k w_{t+1}^k$.
    Update the subset of clients $\mathcal{S}_{t+1} \subseteq [N]$ through uniformly sampling, and broadcast $\mathbf{x}_0^t$ to all clients.
  end for
end for

3. Theoretical Analysis for DP-FedAvg

In this section, we give the formal privacy guarantee and rigorous convergence analysis of DP-FedAvg. Before stating our results, we make the following assumptions.

3.1. Assumptions

Assumption 1. *F_1, F_2,..., F_N are all L-smooth: for all w and v, $F_k(v) \leq F_k(w) + (v - w)^\top \nabla F_k(w) + \frac{L}{2}\|v - w\|^2, \forall k \in [N]$.*

Assumption 2. *F_1, F_2,..., F_N are all μ-strong convex: for all w and v, $F_k(v) \geq F_k(w) + (v - w)^\top \nabla F_k(w) + \frac{\mu}{2}\|v - w\|^2, \forall k \in [N]$.*

Assumption 3. *Denote $\mathcal{B}_t^k$ as the mini-batch data from the dataset in client k with size $|\mathcal{B}_t^k| = b$; then, the variance of stochastic gradients in each client is bounded, that is, $\mathbb{E}\left[\left\|\nabla F_k(w_t^k, \mathcal{B}_t^k) - \nabla F_k(w_t^k)\right\|^2\right] \leq \phi_k^2/b, \forall k \in [N]$.*

Assumption 4. *The local gradient is bounded, i.e., $\left\|\nabla F_k(w_t^k, \mathcal{B}_t^k)\right\| \leq \mathcal{G}$.*

Quantifying the degree of non-i.i.d. data. Inspired by [6], we quantify data heterogeneity through the degree of non-i.i.d. data, defined as

$$\Gamma = F^\star - \sum_{k=1}^K p_k F_k^\star, \tag{16}$$

where $F^\star$ and $F_k^\star$ represent the minimum values of F and F_k, respectively. A larger value of Γ indicates a higher degree of data heterogeneity. Specifically, when the data are i.i.d., Γ approaches zero.

3.2. Privacy Analysis

We firstly estimate the sensitivity on client perspective. The ℓ_2-norm sensitivity for each local uploaded model is given in following Lemma.

Lemma 2. *Supposed that the Assumptions 1–4 hold and* $\mathrm{mod}(t+1, Q) = 0$. *The* ℓ_2*-norm sensitivity is given by*

$$\triangle_{t+1} = \max_{\mathcal{D}_k, \mathcal{D}_k'} \left\| \boldsymbol{w}_{t+1,\mathcal{D}_k}^k - \boldsymbol{w}_{t+1,\mathcal{D}_k'}^k \right\| = 4Q\mathcal{G}\eta_{t+1}, \tag{17}$$

where $\boldsymbol{w}_{t+1,\mathcal{D}_k}^k$ *and* $\boldsymbol{w}_{t+1,\mathcal{D}_k'}^k$ *denote the local model parameters updated from two neighboring datasets* $\mathcal{D}_k$ *and* $\mathcal{D}_k'$*, respectively.*

Proof. See Appendix A. □

Lemma 3. *According to the privacy amplification theorem in Theorem 1, the required privacy protection level to guarantee* (ϵ, δ)*-DP is given by*

$$\epsilon' = \frac{\epsilon}{2q}, \tag{18}$$

$$\delta' = \frac{\delta}{q}, \tag{19}$$

where q *is the data sampling ratio.*

By Lemma 2 and Lemma 3, the total DP noise scale for t-th iteration is

$$\sigma_t^2 = \sum_{k=1}^{K} p_k (\sigma_t^k)^2 = \frac{128\eta_t^2 Q^4 b^2 \mathcal{G}^2}{n\epsilon^2} \sum_{k=1}^{K} \frac{\log(1.25Qb/n_k\delta)}{n_k}. \tag{20}$$

One can observe from (20) that the increasing Q promptly increases the noise scale, which is consistent the result in Lemma 2. According to (20), the DP noise scale is primarily determined by the privacy protection level ϵ, the gradient clipping threshold $\mathcal{G}$ and other algorithm-specific parameters, such as Q, b and η_t. Thus, in practical FL systems, achieving a balance between privacy protection and learning performance relies on effectively controlling the key parameters ϵ and $\mathcal{G}$.

Then, the accumulated noise over the entire training process with total T/Q communication rounds can be obtained by

$$\sigma_{total}^2 = \sum_{t=1}^{T/Q} \sum_{k=1}^{K} p_k (\sigma_t^k)^2 = \frac{128TQ^3 b^2 \mathcal{G}^2}{n\epsilon^2} \sum_{t=1}^{T/Q} \eta_t^2 \sum_{k=1}^{K} \frac{\log(1.25Qb/n_k\delta)}{n_k}. \tag{21}$$

By considering privacy amplification, the total privacy loss after T communication rounds has been explored in [27]. For the sake of clarity and ease of later analysis, we restate it here as follows.

Theorem 2. *Suppose that any client* $i, \forall i \in [N]$ *in Algorithm 1 is uniformly sampled with probability* p_i *and each communication round guarantees* (ϵ, δ)*-DP. The total privacy loss for client* i *after* T *communication rounds satisfies*

$$\bar{\epsilon}_i = c_0 q_i \epsilon \sqrt{\frac{p_i T}{1-q_i}}, \forall i \in [N], \tag{22}$$

where $c_0 > 0$ *is a constant and* q_i *is given by*

$$q_i = \begin{cases} \frac{Qb}{|\mathcal{D}_i|}, & \text{data sampling WOR,} \\ 1-(1-\frac{1}{|\mathcal{D}_i|})^{Qb}, & \text{data sampling WR,} \end{cases} \tag{23}$$

where WOR and WR denote data sampling without and with replacement, respectively.

Remark 1. *By Theorem 2, the total privacy loss is significantly impacted by data sampling strategies. Specifically, the total privacy loss for data sampling with replacement is smaller than that for data sampling without replacement due to* $1-(1-1/|\mathcal{D}_i|)^{Qb} \leq Qb/|\mathcal{D}_i|$.

3.3. Convergence Analysis for DP-FedAvg

In this section, we analyze the convergence performance of the DP-FedAvg algorithm. Let T denote the total number of iterations and $\boldsymbol{w}^\star$ the optimal solution of $F(\boldsymbol{w})$. The convergence performance is measured by $\mathbb{E}[F(\boldsymbol{w}_T) - F(\boldsymbol{w}^\star)]$, where the expectation is taken over the algorithm's randomness. This metric represents the gap between the achieved objective value and the optimal objective value after T iterations. Notably, when $\mathbb{E}[F(\boldsymbol{w}_T) - F(\boldsymbol{w}^\star)] = 0$, the algorithm converges to the optimal solution.

Theorem 3. *For the full-participation case, assume that Assumption 1, Assumption 2 and Assumption 3 hold. Let* $\kappa = L/\mu$, $\alpha = \max\{8\kappa, Q\}$, *and the learning rate* $\eta_t = 2/(\mu(\alpha+t))$. *Then, the following inequality holds:*

$$\mathbb{E}\left[F(\bar{\boldsymbol{w}}_t) - F^\star\right] \leq \frac{2\kappa}{t+\alpha}\left(\frac{A+B}{\mu} + 2L\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|^2\right), \tag{24}$$

where

$$A = \sum_{k=1}^{K} p_k^2 \phi_k^2 / b + 8(Q-1)^2\mathcal{G}^2 + 6L\Gamma, \tag{25}$$

$$B = \frac{8dQ^4 b^2 \mathcal{G}^2}{nL^2\epsilon^2} \log(n) \log(1.25^2 Q^2 b^2 / n\delta^2) \left(2(Q-1)^2 + 1\right). \tag{26}$$

Proof. See Appendix B. □

Remark 2. *In Theorem 3, since* $\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|^2 \leq \frac{1}{\mu^2}\left(\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}^\star)\right)^2 \leq \frac{4\mathcal{G}^2}{\mu^2}$, *the dominating term in (24) is* $\mathcal{O}\left(\frac{A+B}{\mu T}\right)$. *Therefore, the DP-FedAvg algorithm achieves a convergence rate in the order of*

$$\mathcal{O}\left(\frac{\sum_{k=1}^{K} p_k^2 \phi_k^2 / b + Q^2\mathcal{G}^2 + L\Gamma}{\mu T} + \frac{dQ^4 b^2 \mathcal{G}^2 \log(n) \log(Q^2 b^2 / n\delta^2)}{\mu T n L^2 \epsilon^2}\right). \tag{27}$$

Therefore, an increase in Q can significantly increase the noise scale, thereby deteriorating the convergence performance.

Next, we present the convergence analysis by considering client sampling. Inspired by [6], we make the following assumption.

Assumption 5. *Assume the total number of clients is* N *and the size of the selected client set is* $|\mathcal{S}_t| = K$. *If we uniformly sample the participating clients with replacement with sampling probabilities* $p_1, \cdots, p_N$, *then the aggregation step on the server side performs* $\boldsymbol{w}_t = \frac{1}{K}\sum_{k\in\mathcal{S}_t} \boldsymbol{w}_t^k$.

Otherwise, if we uniformly sample the participated clients without replacement from $[N]$*, then the aggregation step on the PS performs* $\boldsymbol{w}_t = \frac{N}{K}\sum_{k\in\mathcal{S}_t} p_k \boldsymbol{w}_t^k$.

Theorem 4. *Suppose that Assumptions 1, 2, 3 and 5 hold. Let* κ, α, η_t, A *and* B *be defined in Theorem 3. Then, the following inequality holds:*

$$\mathbb{E}\left[F(\bar{\boldsymbol{w}}_t) - F^\star\right] \leq \frac{2\kappa}{t+\alpha}\left(\frac{A+B+H}{\mu} + 2L\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|^2\right), \tag{28}$$

if sampling the clients with replacement:

$$H = \frac{4}{K}Q^2\mathcal{G}^2, \tag{29}$$

and if sampling the clients without replacement:

$$H = \frac{N-K}{N-1}\frac{4}{K}Q^2\mathcal{G}^2. \tag{30}$$

Proof. See Appendix D. □

Theorem 4 provides an upper bound of the average total local SGDs over t iterations. A smaller value indicates a higher convergence rate and a lower cost function value in (28). Based on Theorems 3 and 4, we have following remarks.

Remark 3. *(The impact of* Q *on communication efficiency) By denoting* T_ε *as the number of required iterations for DP-FedAvg to achieve an* ε *accuracy. Based on the result in (27), the number of required communication rounds* T_ε/Q *is*

$$\frac{T_\varepsilon}{Q} \propto \mathcal{O}\left(\frac{db^2\mathcal{G}^2\log(n)\log(Q^2b^2/n\delta^2)}{\mu T n L^2\epsilon^2}Q^3 + \mathcal{G}^2Q + \frac{\sum_{k=1}^K p_k^2\phi_k^2/b + L\Gamma}{Q}\right). \tag{31}$$

From (31), we observe that the communication round T_ε/Q *is the function of* Q*, and* T_ε/Q *first decreases and then increases, i.e., there exists an optimal* Q *such that the communication cost is optimal. Furthermore, if* Q *is too large, the* $\boldsymbol{w}_k$ *will promptly achieve local optimal, making FedAvg behave like one-shot average [33]. However, the one-shot averaging does not perform well in the non-i.i.d. data case as demonstrated in [6].*

Remark 4. *(The impact of* Q *on learning performance) The DP noise scale increases significantly with larger values of* Q*, which inevitably degrades learning performance. However, a larger* Q *can also reduce the number of communication rounds, as stated in Remark 3. Therefore, there exists an optimal value of* Q *that balances learning performance and communication efficiency.*

Remark 5. *(The impact of gradient clipping bound level* $\mathcal{G}$*) The convergence performance is influenced by gradient clipping level* $\mathcal{G}$*. However, limiting the gradient norm has the following shortcomings:* (i) *Gradient clipping destroys the unbiasedness of the gradient estimate.* (ii) *If* $\mathcal{G}$ *is too small, the average of the clipped gradients may point in a direction significantly different from the true gradient, potentially leading to poor convergence.* (iii) *If* $\mathcal{G}$ *is too large, the gradients will carry more DP noise, as* $\mathcal{G}$ *determines the sensitivity.*

4. Numerical Experiments

4.1. Experimental Setting

In this section, we evaluate the performance of DP-FedAvg under a non-i.i.d. data setting. Denote the $\ell(\boldsymbol{w}; \mathbf{x}_i)$ as the prediction model with parameter $\boldsymbol{w} = (\mathbf{W}, \mathbf{b})$, then the form for logistic regression and softmax regression are $\ell(\boldsymbol{w}; \mathbf{x}_i, \mathbf{y}_i) = \text{logisitc}(\mathbf{W}\mathbf{x} + \mathbf{b})$ and $\ell(\boldsymbol{w}; \mathbf{x}_i, \mathbf{y}_i) = \text{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})$, respectively. The loss function for both problems is given by

$$F_k(\boldsymbol{w}) = \frac{1}{n_k} \sum_{j=1}^{n_k} \text{CrossEntropy}\left(\ell(\boldsymbol{w}; \mathbf{x}_j), \mathbf{y}_j\right). \tag{32}$$

Note that this is a convex and smooth problem. The testing accuracy (learning accuracy over the testing data) is evaluated as the learning performance in the experiments.

Datasets: Two benchmark datasets, Adult [34] and MNIST [35], were considered for performance evaluation. The MNIST dataset consists of 60,000 training samples and 10,000 testing samples, while the Adult dataset consists of 32,561 training samples and 16,281 testing samples. To simulate the FL system, we distributed all training samples across $N = 100$ clients in a **non-i.i.d.** manner, ensuring that each client possessed partial labels for the data:

- For the MNIST dataset, following the heterogeneous data partition method in [6], each client was allocated data samples of only four different labels with $|\mathcal{D}_i| = 600$. This led to a high degree of non-i.i.d. datasets among clients.
- For the Adult dataset, following [1], all the training samples were uniformly distributed among a total $N = 100$ clients such that each client only contained data from one class with $|\mathcal{D}_i| = 325$.

Parameter setting: For all experiments, we maintained the following parameter settings: $K = 10$, $b = 10$ and $\delta = 10^{-5}$. The choice of the learning rate η^t depended on the dataset under consideration.

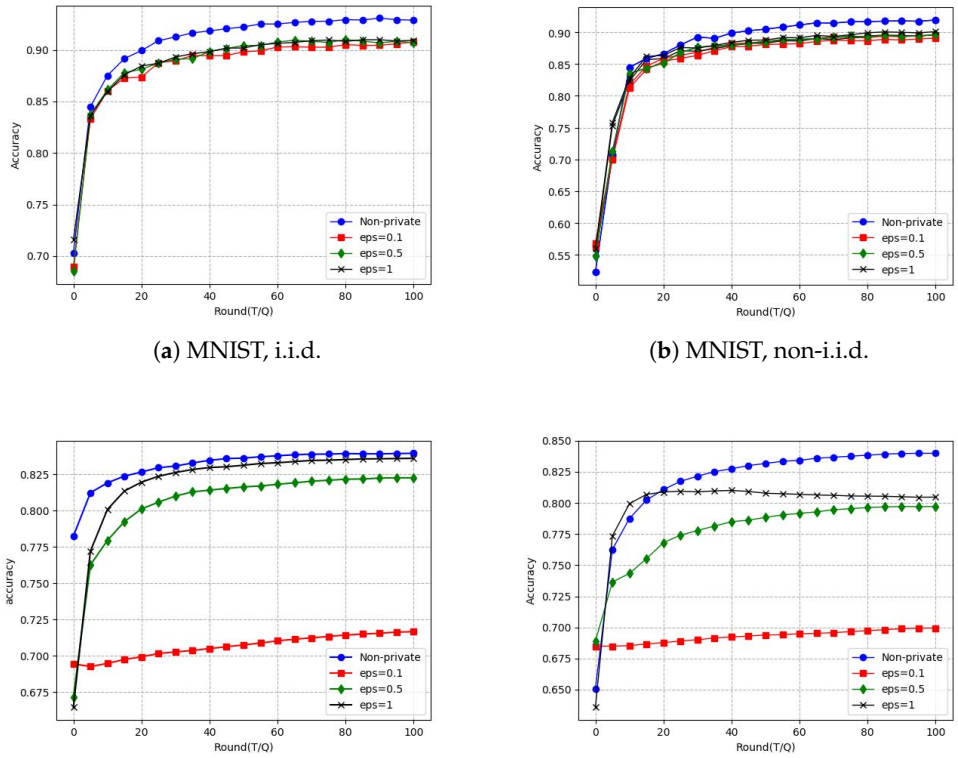
For the MNIST dataset, the local learning rate was set to $\eta^t = 0.04/\sqrt{1+t}$, while for the Adult dataset, $\eta^t = 0.01/\sqrt{1+t}$. Following the rule of gradient clipping in [25], we selected the value of $\mathcal{G}$ by taking the median of the norms of the unclipped gradients over the training process.

4.2. Impact of the Privacy Protection Level ϵ

To validate the learning performance at various privacy protection levels, the parameters were set as $K = 10$, $Q = 5$ and $\delta = 10^{-5}$, and the testing accuracy was computed with respect to communication rounds (T/Q). We compared the accuracy with various values of ϵ under both i.i.d. and non-i.i.d. scenarios. As illustrated in Figure 2, both the MNIST and Adult datasets provided strong privacy protection without a significant decrease in testing accuracy. Moreover, we find that MNIST has a higher tolerance to DP noise than the Adult dataset. Specifically, the difference in testing accuracy between the non-private and private models for both i.i.d. and non-i.i.d. datasets is much smaller than the corresponding drop in accuracy for the Adult dataset. The reason for the MNIST dataset's robustness to DP noise compared with the Adult dataset may have been due to the following reasons:

- Impact of non-i.i.d. data: The presence of non-i.i.d. data across clients in FL amplifies the impact of DP noise. The Adult dataset, which contains categorical and numerical features, is more sensitive to such heterogeneity. Conversely, the MNIST dataset, consisting of grayscale images of handwritten digits with simpler patterns, is inherently more resilient to non-i.i.d. data settings.

- Inherent data structure: The MNIST dataset benefits from the spatial coherence of pixel intensities, where patterns such as the shapes of digits remain discernible when DP noise is introduced. This property of image data enables models to identify meaningful features despite perturbations. On the other hand, the Adult dataset, being tabular with categorical and numerical features, lacks such spatial structure. DP noise in tabular data can obscure critical relationships between features and labels, making it harder for the FL model to extract meaningful patterns under noisy conditions.



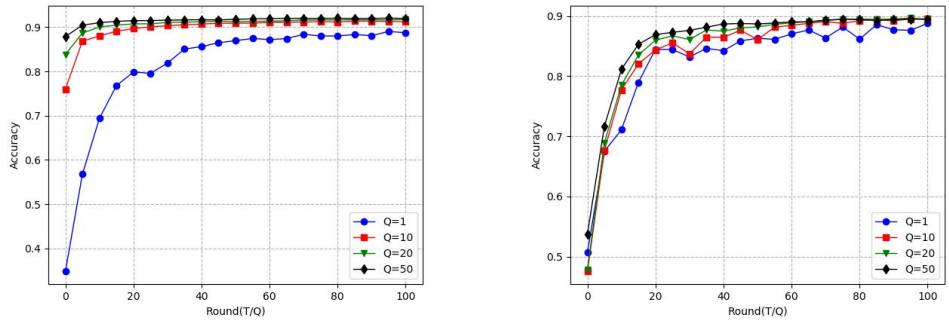
(**a**) MNIST, i.i.d. (**b**) MNIST, non-i.i.d.

(**c**) Adult, i.i.d. (**d**) Adult, non-i.i.d.

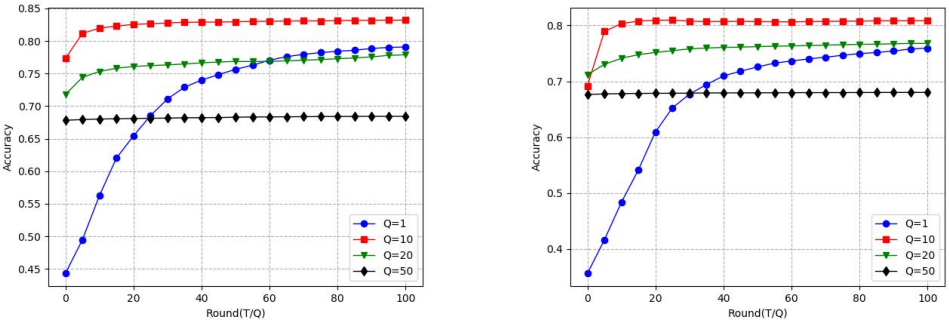
Figure 2. The testing accuracy versus communication rounds of the DP-FedAvg algorithm on various privacy protection levels under Adult and MNIST datasets.

4.3. Impact of Local Epoch Length Q

In this experiment, we compared the learning performance under the parameter settings $\epsilon = 0.5$, $K = 10$ and $\delta = 10^{-5}$ for various values of Q. As shown in Figure 3, we observed that both excessively large and excessively small values of Q reduced the convergence rate on the Adult dataset. However, for the MNIST dataset, a large Q had a negligible effect on convergence. This may have been due to two reasons: (i) The optimal value of Q for MNIST is significantly larger than for the Adult dataset, and (ii) MNIST exhibits a higher tolerance to noise compared to the Adult dataset. Although a large Q alleviates the communication burden, it substantially increases the noise scale, thereby degrading learning performance as discussed in Remark 4. Additionally, our simulation results indicate that the optimal value of Q is sensitive to both the parameter settings and the training dataset. Furthermore, for both datasets, the convergence performance on i.i.d. data is better than on non-i.i.d. data under the same parameter settings.



(**a**) MNIST, i.i.d. (**b**) MNIST, non-i.i.d.



(**c**) Adult, i.i.d. (**d**) Adult, non-i.i.d.

Figure 3. The testing accuracy versus communication rounds of the DP-FedAvg algorithm with various Q under Adult and MNIST datasets.

5. Conclusions and Future Works

In this paper, we conducted a thorough theoretical analysis of DP-FedAvg under different client and data sampling strategies, with a particular focus on the impact of non-i.i.d. data on the convergence and privacy protection of FL algorithms. Our work fills a significant gap in the literature by providing a unified framework that integrates data heterogeneity with DP. Through our analysis, we demonstrated that the DP-FedAvg algorithm maintains robustness toward data heterogeneity and DP noise, achieving a sub-linear convergence rate for convex FL problems. Additionally, our convergence and privacy analyses offer valuable insight and perspectives into how various system parameters in the FL framework influence learning performance, which can be used as guidelines for the practical FL algorithm design. Our theoretical findings are supported by extensive numerical experiments on real-world datasets, which are consistent with our theoretical findings.

While our theoretical analysis provides useful insights for DP-FedAvg under the data heterogeneity scenario, several promising directions remain for future research. First, our work is based on the typical star network framework, consisting of one server and multiple clients. A promising direction would be to explore hierarchical FL systems, where multiple servers collaborate to train a global model. This framework could involve more clients and larger datasets, potentially improving learning performance. Second, another promising avenue for future work involves examining the tradeoffs between privacy, utility and computational overhead when implementing advanced methods in resource-constrained FL systems, such as those in mobile edge computing environments. Finally, while the DP mechanism offers strong privacy guarantees, more sophisticated approaches could be employed to further enhance privacy protection. Techniques such as homomorphic encryption could be integrated to strengthen the security of DP-FedAvg. Addressing these

challenges and exploring innovative theoretical analyses will be an exciting and valuable direction for future research.

Author Contributions: Methodology, Y.L. and S.W.; Formal analysis, Y.L., S.W. and Q.W.; Investigation, Y.L. and Q.W.; Writing—original draft preparation, Y.L.; Writing—review and editing, Y.L., S.W. and Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

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

Appendix A. Proof of the Lemma 2

Proof. Since there is only one different record between $\mathcal{D}_k, \mathcal{D}'_k$, and each client performs Q steps of SGD per communication round, the local update scheme is $\boldsymbol{w}^k_{t+1,\mathcal{D}_k} = \boldsymbol{w}^k_{t+1-Q,\mathcal{D}_k} - \sum_{\tau=0}^{Q-1} \eta_{t+1-Q+\tau} \boldsymbol{g}^k_{t+1-Q+\tau,\mathcal{D}_k}$. Thus, we have

$$\begin{aligned}
\left\| \boldsymbol{w}^k_{t+1,\mathcal{D}_k} - \boldsymbol{w}^k_{t+1,\mathcal{D}'_k} \right\|^2 &= \left\| \sum_{\tau=0}^{Q-1} \eta_{t+1-Q+\tau} \boldsymbol{g}^k_{t+1-Q+\tau,\mathcal{D}_k}) - \sum_{\tau=0}^{Q-1} \eta_{t+1-Q+\tau} \boldsymbol{g}^k_{t+1-Q+\tau,\mathcal{D}'_k}) \right\|^2 \\
&\overset{(a)}{\leq} 4\eta^2_{t+1} \left\| \sum_{\tau=0}^{Q-1} \left(\boldsymbol{g}^k_{t+1-Q+\tau,\mathcal{D}_k} - \boldsymbol{g}^k_{t+1-Q+\tau,\mathcal{D}'_k} \right) \right\|^2 \\
&\leq 4Q^2\eta^2_{t+1} \sum_{\tau=0}^{Q-1} \left\| \boldsymbol{g}^k_{t+1-Q+\tau,\mathcal{D}_k} - \boldsymbol{g}^k_{t+1-Q+\tau,\mathcal{D}'_k} \right\|^2 \\
&\overset{(b)}{\leq} 16Q^2\mathcal{G}^2\eta^2_{t+1},
\end{aligned} \tag{A1}$$

where (a) holds due to $\eta_t \leq 2\eta_{t+1}$; (b) follows from Assumption 4. □

Appendix B. Proof of Theorem 3

Proof. Inspired by the perturbed iterate framework in [6], we define $\mathcal{I}_Q = \{zQ | z = 1, 2, \cdots\}$ as the set of global synchronization steps. When $t+1 \notin \mathcal{I}_Q$, it is the time that clients run the local model update; otherwise, it is the time step to communication. Then, the update rule of DP-FedAvg can be described as

$$\boldsymbol{v}^k_{t+1} = \begin{cases} \boldsymbol{w}^k_t - \eta_t \nabla F_k\left(\boldsymbol{w}^k_t, \mathcal{B}^k_t\right) & \text{if } t+1 \notin \mathcal{I}_Q \\ \boldsymbol{w}^k_t - \eta_t \left(\nabla F_k\left(\boldsymbol{w}^k_t, \mathcal{B}^k_t\right) + \boldsymbol{\xi}^k_t\right) & \text{if } t+1 \in \mathcal{I}_Q \end{cases} \tag{A2}$$

$$\boldsymbol{w}^k_{t+1} = \begin{cases} \boldsymbol{v}^k_{t+1} & \text{if } t+1 \notin \mathcal{I}_Q \\ \sum_{k=1}^K p_k \boldsymbol{v}^k_{t+1} & \text{if } t+1 \in \mathcal{I}_Q \end{cases} \tag{A3}$$

where $\boldsymbol{\xi}^k_t$ is the Gaussian noise added to gradients, $\boldsymbol{v}^k_{t+1}$ denotes one-step SGD update from $\boldsymbol{w}^k_t$ and $\boldsymbol{w}^k_{t+1}$ can be interpreted as the result after communication steps.

Motivated by [6], we define two virtual sequences, $\bar{\boldsymbol{w}}_t = \sum_{k=1}^K p_k \boldsymbol{w}^k_t$ and $\bar{\boldsymbol{v}}_t = \sum_{k=1}^K p_k \boldsymbol{v}^k_t$. For simplicity of analysis, when $t+1 \in \mathcal{I}_Q$, define

$$\bar{\boldsymbol{g}}_t = \sum_{k=1}^K p_k \left(\nabla F_k(\boldsymbol{w}^k_t) + \boldsymbol{\xi}^k_t \right), \tag{A4}$$

$$\boldsymbol{g}_t = \sum_{k=1}^K p_k \left(\nabla F_k\left(\boldsymbol{w}^k_t, \mathcal{B}^k_t\right) + \boldsymbol{\xi}^k_t \right). \tag{A5}$$

When $t+1 \notin \mathcal{I}_Q$, define

$$\boldsymbol{g}_t = \sum_{k=1}^{K} p_k \left(\nabla F_k \left(\boldsymbol{w}_t^k, \mathcal{B}_t^k \right) \right), \tag{A6}$$

$$\bar{\boldsymbol{g}}_t = \sum_{k=1}^{K} p_k \left(\nabla F_k \left(\boldsymbol{w}_t^k \right) \right). \tag{A7}$$

Note that, for both cases, we always have $\bar{\boldsymbol{v}}_{t+1} = \bar{\boldsymbol{w}}_t - \eta_t \boldsymbol{g}_t$ and $\mathbb{E}[\boldsymbol{g}_t] = \bar{\boldsymbol{g}}_t$. □

Appendix B.1. Key Lemmas

To clearly convey the proof of Theorem 3, we need the following key lemmas.

Lemma A1. *(one-step SGD) If Assumption 1 and Assumption 2 hold and $\eta_t \leq \frac{1}{4L}$. Then, when $t+1 \notin \mathcal{I}_Q$,*

$$\begin{aligned} \mathbb{E}\left[\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^\star\|^2 \right] \leq & (1 - \mu\eta_t) \mathbb{E}\left[\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2 \right] + \eta_t^2 \mathbb{E}\left[\|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2 \right] \\ & + 2\mathbb{E}\left[\sum_{k=1}^{K} p_k \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2 \right] + 6L\eta_t^2 \Gamma, \end{aligned} \tag{A8}$$

when $t+1 \in \mathcal{I}_Q$,

$$\begin{aligned} \mathbb{E}\left[\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^\star\|^2 \right] \leq & (1 - \mu\eta_t) \mathbb{E}\left[\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2 \right] + \eta_t^2 \mathbb{E}\left[\|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2 \right] \\ & + 2\mathbb{E}\left[\sum_{k=1}^{K} p_k \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2 \right] + 6L\eta_t^2 \Gamma + 2d\eta_t^2 \mathbb{E}\left[\sum_{k=1}^{K} p_k \|\boldsymbol{\xi}_t^k\|^2 \right]. \end{aligned} \tag{A9}$$

Lemma A2. *Suppose that Assumption 3 holds. Then,*

$$\mathbb{E}\left[\left\| \boldsymbol{g}_t - \bar{\boldsymbol{g}}_t \right\|^2 \right] \leq \sum_{k=1}^{K} p_k^2 \phi_k^2 / b. \tag{A10}$$

Lemma A3. *Assume η_t is non-increasing and $\eta_t \leq 2\eta_{t+Q}$, then, when $t+1 \in \mathcal{I}_Q$,*

$$\mathbb{E}\left[\sum_{k=1}^{K} p_k \left\| \bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k \right\|^2 \right] \leq 4(Q-1)^2 \eta_t^2 \left(\mathcal{G}^2 + d \sum_{k=1}^{K} p_k \sigma_{t,k}^2 \right). \tag{A11}$$

When $t+1 \notin \mathcal{I}_Q$,

$$\mathbb{E}\left[\sum_{k=1}^{K} p_k \left\| \bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k \right\|^2 \right] \leq 4(Q-1)^2 \mathcal{G}^2 \eta_t^2. \tag{A12}$$

Lemma A4. *Assume η_t is non-increasing and $\eta_t \leq \frac{1}{4L}$, then the total noise in each communication round is given by*

$$\mathbb{E}\left[\sum_{k=1}^{K} p_k \left\| \boldsymbol{\xi}_t^k \right\|^2 \right] \leq \frac{4Q^4 b^2 \mathcal{G}^2}{nL^2 \epsilon^2} \log(n) \log(1.25^2 Q^2 b^2 / n\delta^2). \tag{A13}$$

Appendix B.2. Completing the Proof of Theorem 3

Proof. Let $\triangle_{t+1} = \mathbb{E}\Big[\|\bar{\boldsymbol{w}}_{t+1} - \boldsymbol{w}^\star\|^2\Big]$. Based on the results of Lemma A1–A4, when $t+1 \notin \mathcal{I}_Q$, it follows that

$$\mathbb{E}\Big[\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^\star\|^2\Big] \leq (1-\mu\eta_t)\mathbb{E}\Big[\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2\Big] + A\eta_t^2, \tag{A14}$$

where

$$A = \sum_{k=1}^{K} p_k^2\phi_k^2/b + 8(Q-1)^2\mathcal{G}^2 + 6L\Gamma. \tag{A15}$$

When $t+1 \in \mathcal{I}_Q$,

$$\mathbb{E}\Big[\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^\star\|^2\Big] \leq (1-\mu\eta_t)\mathbb{E}\Big[\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2\Big] + (A+B)\eta_t^2, \tag{A16}$$

where

$$B = \frac{8dQ^4b^2\mathcal{G}^2\log(n)\log(1.25^2Q^2b^2/n\delta^2)}{nL^2\epsilon^2}\big(2(Q-1)^2+1\big). \tag{A17}$$

Then, (A16) becomes

$$\triangle_{t+1} \leq (1-\mu\eta_t)\,\triangle_t + \eta_t^2(A+B). \tag{A18}$$

For a diminishing stepsize, $\eta_t = \frac{\beta}{t+\alpha}$ for some $\beta > \frac{1}{\mu}$ and $\alpha \geq 0$ such that $\eta_0 \leq \min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+Q}$, this also implies that $\alpha \geq Q$:

$$\begin{aligned}
\triangle_{t+1} &\leq (1-\mu\eta_t)\,\triangle_t + \eta_t^2(A+B)\\
&= \Big(1-\frac{\mu\beta}{t+\alpha}\Big)\frac{\nu}{t+\alpha} + \frac{\beta^2}{(t+\alpha)^2}(A+B)\\
&\leq \Big(1-\frac{\mu\beta}{t+\alpha}\Big)\frac{\nu}{t+\alpha} + \frac{\beta^2(A+B)}{(t+\alpha)^2}\\
&\leq \frac{t+\alpha-1}{(t+\alpha)^2}\nu + \left(\frac{\beta^2(A+B)}{(t+\alpha)^2} - \frac{(\mu\beta-1)\nu}{(t+\alpha)^2}\right)\\
&\leq \frac{t+\alpha-1}{(t+\alpha)^2-1}\nu\\
&= \frac{\nu}{t+\alpha+1},
\end{aligned} \tag{A19}$$

where $\nu \geq \beta^2(A+B)/(\mu\beta-1)$. By the L-smoothness of $F(\cdot)$, we have

$$\mathbb{E}\big[F(\bar{\boldsymbol{w}}_t) - F^\star\big] \leq \frac{L}{2}\mathbb{E}\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2 = \frac{L}{2}\triangle_t \leq \frac{L}{2}\frac{\nu}{t+\alpha}. \tag{A20}$$

Note that, from (A20), we have $\nu \geq \alpha\triangle_0$. Thus, we obtain $\nu = \max\Big\{\frac{\beta^2(A+B)}{\mu\beta-1}, \alpha\triangle_0\Big\}$. If we choose $\beta = 2/\mu$ and then $\eta_t = \frac{2}{\mu}\frac{1}{\alpha+t}$, where $\alpha \geq \max\{8\kappa, Q\}$, $\kappa = \frac{L}{\mu}$. The range of α due to $\eta_0 = \frac{2}{\mu}\frac{1}{\alpha} \leq \frac{1}{4L}$ implies that $\alpha \geq 8\kappa$. Then, we have

$$\mathbb{E}\big[F(\bar{\boldsymbol{w}}_t) - F^\star\big] \leq \frac{2\kappa}{t+\alpha}\Big(\frac{A+B}{\mu} + 2L\,\triangle_0\Big), \tag{A21}$$

where if $\frac{\beta^2(A+B)}{\mu\beta-1} \geq \alpha\triangle_0$, then

$$\mathbb{E}\left[F(\bar{\boldsymbol{w}}_t) - F^\star\right] \leq \frac{L}{2}\frac{1}{t+\alpha}\frac{\beta^2 A}{\mu\beta-1} = \frac{2\kappa}{t+\alpha}\frac{A+B}{\mu}; \tag{A22}$$

otherwise, if $\frac{\beta^2(A+B)}{\mu\beta-1} < \alpha\triangle_0$, and in this case, $\alpha = 8\kappa$, then

$$\mathbb{E}\left[F(\bar{\boldsymbol{w}}_t) - F^\star\right] \leq \frac{L}{2}\frac{1}{t+\alpha}\alpha\triangle_0 = \frac{4\kappa L\triangle_0}{t+\alpha}. \tag{A23}$$

By combining (A21), (A22) and (A23), we obtain the result. □

Appendix C. Proofs of Key Lemmas in Theorem 3

Appendix C.1. Proof of Lemma A1

Proof. Since we have $\bar{\boldsymbol{v}}_{t+1} = \bar{\boldsymbol{w}}_t - \eta_t \boldsymbol{g}_t$, then

$$\begin{aligned}\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^\star\|^2 =& \|\bar{\boldsymbol{w}}_t - \eta_t\boldsymbol{g}_t - \boldsymbol{w}^\star\|^2 = \|\bar{\boldsymbol{w}}_t - \eta_t\boldsymbol{g}_t - \boldsymbol{w}^\star - \eta_t\bar{\boldsymbol{g}}_t + \eta_t\bar{\boldsymbol{g}}_t\|^2 \\ =& \underbrace{\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star - \eta_t\bar{\boldsymbol{g}}_t\|^2}_{\triangleq\mathcal{A}_1} + 2\eta_t\langle\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star - \eta_t\bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t\rangle \\ &+ \eta_t^2\underbrace{\|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2}_{\triangleq\mathcal{A}_2}.\end{aligned} \tag{A24}$$

Then, $\mathbb{E}\left[\langle\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star - \eta_t\bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t\rangle\right] = 0$ due to $\mathbb{E}[\boldsymbol{g}_t] = \bar{\boldsymbol{g}}_t$. Then, for the $\mathcal{A}_1$ term, we have

$$\begin{aligned}\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star - \eta_t\bar{\boldsymbol{g}}_t\|^2 =& \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2 + \eta_t^2\|\bar{\boldsymbol{g}}_t\|^2 - 2\eta_t\langle\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star, \bar{\boldsymbol{g}}_t\rangle \\ \overset{(a)}{\leq}& \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2 + \eta_t^2\sum_{k=1}^K p_k\|\nabla F_k(\boldsymbol{w}_t^k) + \boldsymbol{\xi}_t^k\|^2 \\ &- 2\eta_t\sum_{k=1}^K p_k\left\langle\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k + \boldsymbol{w}_t^k - \boldsymbol{w}^\star, \nabla F_k(\boldsymbol{w}_t^k) + \boldsymbol{\xi}_t^k\right\rangle \\ =& \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2 - 2\eta_t\sum_{k=1}^K p_k\left\langle\boldsymbol{w}_t^k - \boldsymbol{w}^\star, \nabla F_k(\boldsymbol{w}_t^k) + \boldsymbol{\xi}_t^k\right\rangle \\ &- 2\eta_t\sum_{k=1}^K p_k\left\langle\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k, \nabla F_k(\boldsymbol{w}_t^k) + \boldsymbol{\xi}_t^k\right\rangle \\ &+ \eta_t^2\sum_{k=1}^K p_k\|\nabla F_k(\boldsymbol{w}_t^k) + \boldsymbol{\xi}_t^k\|^2,\end{aligned} \tag{A25}$$

where (a) follows from Jensen's inequality of function $\|\cdot\|^2$. Then, by the L-smoothness of $F_k(\cdot)$, we have $\left\|\nabla F_k(\boldsymbol{w}_t^k)\right\|^2 \leq 2L\left(F_k(\boldsymbol{w}_t^k) - F_k^\star\right)$. Then,

$$\begin{aligned}\eta_t^2\sum_{k=1}^K p_k\|\Pi_{\mathcal{G}}\left(\nabla F_k(\boldsymbol{w}_t^k)\right)+\boldsymbol{\xi}_t^k\|^2 \leq&\eta_t^2\sum_{k=1}^K p_k\|\nabla F_k(\boldsymbol{w}_t^k)+\boldsymbol{\xi}_t^k\|^2\\ =&\eta_t^2\sum_{k=1}^K p_k\left(\|\nabla F_k(\boldsymbol{w}_t^k)\|^2+2\nabla F_k(\boldsymbol{w}_t^k)^\top\boldsymbol{\xi}_t^k+\|\boldsymbol{\xi}_t^k\|^2\right)\\ \leq&2L\eta_t^2\sum_{k=1}^K p_k\left(F_k(\boldsymbol{w}_t^k)-F_k^\star\right)+\eta_t^2\sum_{k=1}^K p_k\|\boldsymbol{\xi}_t^k\|^2\\ &+2\eta_t^2\sum_{k=1}^K p_k\left(\nabla F_k(\boldsymbol{w}_t^k)^\top\boldsymbol{\xi}_t^k\right),\end{aligned}\tag{A26}$$

where $\Pi_{\mathcal{G}}(\cdot)$ denotes the gradient clipping operation. By μ-strong convex of $F_k(\cdot)$, we have $\nabla F_k(\boldsymbol{w}_t^k)^\top(\boldsymbol{w}_t^k-\boldsymbol{w}^\star)\geq\left(F_k(\boldsymbol{w}_t^k)-F_k(\boldsymbol{w}^\star)\right)+\frac{\mu}{2}\left\|\boldsymbol{w}_t^k-\boldsymbol{w}^\star\right\|^2$, then

$$\begin{aligned}-\left\langle\boldsymbol{w}_t^k-\boldsymbol{w}^\star,\nabla F_k(\boldsymbol{w}_t^k)+\mathcal{G}\cdot\boldsymbol{\xi}_t^k\right\rangle=&-\left\langle\boldsymbol{w}_t^k-\boldsymbol{w}^\star,\nabla F_k(\boldsymbol{w}_t^k)\right\rangle-\mathcal{G}\left\langle\boldsymbol{w}_t^k-\boldsymbol{w}^\star,\boldsymbol{\xi}_t^k\right\rangle\\ \leq&-\left(F_k(\boldsymbol{w}_t^k)-F_k(\boldsymbol{w}^\star)\right)-\frac{\mu}{2}\left\|\boldsymbol{w}_t^k-\boldsymbol{w}^\star\right\|^2\\ &-\mathcal{G}\left\langle\boldsymbol{w}_t^k-\boldsymbol{w}^\star,\boldsymbol{\xi}_t^k\right\rangle.\end{aligned}\tag{A27}$$

By the AM-GM inequality, i.e., $2\langle\boldsymbol{x},\boldsymbol{y}\rangle\leq\zeta\|\boldsymbol{x}\|^2+\zeta^{-1}\|\boldsymbol{y}\|^2$, for any $\zeta>0$, we have,

$$\begin{aligned}-2\left\langle\bar{\boldsymbol{w}}_t-\boldsymbol{w}_t^k,\nabla F_k(\boldsymbol{w}_t^k)+\boldsymbol{\xi}_t^k\right\rangle\leq&\frac{1}{\eta_t}\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_t^k\|^2+\eta_t\|\nabla F_k(\boldsymbol{w}_t^k)+\boldsymbol{\xi}_t^k\|^2\\ \leq&\frac{1}{\eta_t}\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_t^k\|^2+\eta_t\|\nabla F_k(\boldsymbol{w}_t^k)\|^2\\ &+2\eta_t\|\nabla F_k(\boldsymbol{w}_t^k)^\top\boldsymbol{\xi}_t^k\|+\eta_t\|\boldsymbol{\xi}_t^k\|^2\\ \leq&\frac{1}{\eta_t}\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_t^k\|^2+2L\eta_t\left(F_k(\boldsymbol{w}_t^k)-F_k^\star\right)\\ &+2\eta_t\|\nabla F_k(\boldsymbol{w}_t^k)^\top\boldsymbol{\xi}_t^k\|+\eta_t\|\boldsymbol{\xi}_t^k\|^2.\end{aligned}\tag{A28}$$

By applying these estimates, we have

$$\begin{aligned}\|\bar{\boldsymbol{w}}_t-\eta_t\bar{\boldsymbol{g}}_t-\boldsymbol{w}^\star\|^2\leq&\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}^\star\|^2+4L\eta_t^2\sum_{k=1}^K p_k\left(F_k(\boldsymbol{w}_t^k)-F_k^\star\right)\\ &-2\eta_t\sum_{k=1}^K p_k\left(F_k(\boldsymbol{w}_t^k)-F_k(\boldsymbol{w}^\star)\right)+4\eta_t^2\sum_{k=1}^K p_k\left(\nabla F_k(\boldsymbol{w}_t^k)^\top\boldsymbol{\xi}_t^k\right)\\ &+2\eta_t^2\sum_{k=1}^K p_k\|\boldsymbol{\xi}_t^k\|^2+\sum_{k=1}^K p_k\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_t^k\|^2-\mu\eta_t\sum_{k=1}^K p_k\left\|\boldsymbol{w}_t^k-\boldsymbol{w}^\star\right\|^2\\ &-2\eta_t\sum_{k=1}^K p_k(\boldsymbol{w}_t^k-\boldsymbol{w}^\star)^\top\boldsymbol{\xi}_t^k\\ \overset{(a)}{\leq}&(1-\mu\eta_t)\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}^\star\|^2+4L\eta_t^2\sum_{k=1}^K p_k\left(F_k(\boldsymbol{w}_t^k)-F_k^\star\right)\\ &-2\eta_t\sum_{k=1}^K p_k\left(F_k(\boldsymbol{w}_t^k)-F_k(\boldsymbol{w}^\star)\right)+2\eta_t^2\sum_{k=1}^K p_k\|\boldsymbol{\xi}_t^k\|^2\\ &+\sum_{k=1}^K p_k\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_t^k\|^2,\end{aligned}\tag{A29}$$

where in (a), we apply the fact that $\boldsymbol{w}_t^k$ is independent of the $\boldsymbol{\xi}_t^k$ and thus, $\mathbb{E}[\boldsymbol{\xi}_t^k] = 0$. Furthermore, we also use the inequality $-\sum_{k=1}^K p_k \left\| \boldsymbol{w}_t^k - \boldsymbol{w}^\star \right\|^2 \leq -\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2$, which can be proved by

$$\begin{aligned}
\sum_{k=1}^K p_k \left\| \boldsymbol{w}_t^k - \boldsymbol{w}^\star \right\|^2 &\geq \sum_{k=1}^K p_k \left(\|\boldsymbol{w}_t^k\|^2 \right) - 2\|\bar{\boldsymbol{w}}_t\| \|\boldsymbol{w}^\star\| + \|\boldsymbol{w}^\star\|^2 \\
&\geq \sum_{k=1}^K \left(\|p_k \boldsymbol{w}_t^k\|^2 \right) - 2\|\bar{\boldsymbol{w}}_t\| \|\boldsymbol{w}^\star\| + \|\boldsymbol{w}^\star\|^2 \\
&= \left\| \bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star \right\|^2. \qquad \text{(A30)}
\end{aligned}$$

Then, by letting $\Psi(t) \triangleq 4L\eta_t^2 \sum_{k=1}^K p_k \left(F_k(\boldsymbol{w}_t^k) - F_k^\star \right) - 2\eta_t \sum_{k=1}^K p_k \left(F_k(\boldsymbol{w}_t^k) - F_k(\boldsymbol{w}^\star) \right)$, we have

$$\begin{aligned}
\Psi(t) &= 4L\eta_t^2 \sum_{k=1}^K p_k \Big(F_k(\boldsymbol{w}_t^k) - F_k^\star \Big) - 2\eta_t \sum_{k=1}^K p_k \Big(F_k(\boldsymbol{w}_t^k) - F_k(\boldsymbol{w}^\star) \Big) \\
&= \Big(4L\eta_t^2 - 2\eta_t \Big) \sum_{k=1}^K p_k \Big(F_k(\boldsymbol{w}_t^k) - F_k(\boldsymbol{w}^\star) \Big) + 4L\eta_t^2 \sum_{k=1}^K p_k \left(F_k(\boldsymbol{w}^\star) - F_k^\star \right) \\
&= \Big(4L\eta_t^2 - 2\eta_t \Big) \sum_{k=1}^K p_k \Big(F_k(\boldsymbol{w}_t^k) - F^\star \Big) + 4L\eta_t^2 \Big(F^\star - \sum_{k=1}^K p_k F_k^\star \Big) \\
&= \Big(4L\eta_t^2 - 2\eta_t \Big) \sum_{k=1}^K p_k \Big(F_k(\boldsymbol{w}_t^k) - F^\star \Big) + 4L\eta_t^2 \Gamma. \qquad \text{(A31)}
\end{aligned}$$

Since $\eta_t \leq \frac{1}{4L}$, which implies that $4L\eta_t^2 - 2\eta_t \leq 0$, then, $\sum_{k=1}^K p_k \left(F_k(\boldsymbol{w}_t^k) - F^\star \right)$ can be bounded by

$$\begin{aligned}
\sum_{k=1}^K p_k \Big(F_k(\boldsymbol{w}_t^k) - F^\star \Big) &= \sum_{k=1}^K p_k \Big(F_k(\boldsymbol{w}_t^k) - F_k(\bar{\boldsymbol{w}}_t) \Big) + \sum_{k=1}^K p_k (F_k(\bar{\boldsymbol{w}}_t) - F^\star) \\
&\overset{(a)}{\geq} \sum_{k=1}^K p_k \Big\langle \nabla F_k(\bar{\boldsymbol{w}}_t), \bar{\boldsymbol{w}}_t^k - \overline{\boldsymbol{w}}_t \Big\rangle + (F(\bar{\boldsymbol{w}}) - F^\star) \\
&\overset{(b)}{\geq} -\frac{1}{2} \sum_{k=1}^K p_k \Big(\eta_t \|\nabla F_k(\bar{\boldsymbol{w}}_t)\|^2 + \frac{1}{\eta_t} \Big\| \boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t \Big\|^2 \Big) + (F(\bar{\boldsymbol{w}}) - F^\star) \\
&\overset{(c)}{\geq} -L\eta_t \sum_{k=1}^K p_k (F_k(\bar{\boldsymbol{w}}_t) - F_k^\star) - \frac{1}{2\eta_t} \sum_{k=1}^K p_k \Big\| \boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t \Big\|^2 \\
&\quad + \left(F(\bar{\boldsymbol{w}}) - F^\star \right), \qquad \text{(A32)}
\end{aligned}$$

where (a) follows from the convexity of $F_k(\cdot)$; (b) holds due to the AM-GM inequality; in (c), we invoke the L-smoothness of $F_k(\cdot)$. Thus, by plugging (A32) into (A31), we have

$$\begin{aligned}
\Psi(t) \leq & L\eta_t \Big(2\eta_t - 4L\eta_t^2 \Big) \sum_{k=1}^K p_k (F_k(\bar{\boldsymbol{w}}_t) - F_k^\star) + (1 - 2L\eta_t) \sum_{k=1}^K p_k \Big\| \boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t \Big\|^2 \\
& - \Big(2\eta_t - 4L\eta_t^2 \Big) (F(\bar{\boldsymbol{w}}) - F^\star) + 4L\eta_t^2 \Gamma \\
\leq & (L\eta_t - 1) \Big(2\eta_t - 4L\eta_t^2 \Big) (F(\bar{\boldsymbol{w}}) - F^\star) + (1 - 2L\eta_t) \sum_{k=1}^K p_k \Big\| \boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t \Big\|^2 \\
& + \Big(4L\eta_t^2 + L\eta_t \Big(2\eta_t - 4L\eta_t^2 \Big) \Big) \Gamma \\
\overset{(a)}{\leq} & \sum_{k=1}^K p_k \Big\| \boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t \Big\|^2 + 6L\eta_t^2 \Gamma, \qquad \text{(A33)}
\end{aligned}$$

where (a) follows from $L\eta_t - 1 < 0$ and $2\eta_t - 4L\eta_t^2 \leq 2\eta_t$. Then, combining the results of (A29), (A31)–(A33), we have

$$\mathbb{E}\Big[\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^\star\|^2\Big] \leq (1-\mu\eta_t)\mathbb{E}\Big[\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2\Big] + 2\mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2\Big] + 2p\eta_t^2 \mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\boldsymbol{\xi}_t^k\|^2\Big] + 6L\Gamma\eta_t^2. \tag{A34}$$

When $t+1 \notin \mathcal{I}_Q$, we have

$$\boldsymbol{g}_t = \sum_{k=1}^{K} p_k \nabla F_k\Big(\boldsymbol{w}_t^k, \mathcal{B}_t^k\Big), \tag{A35a}$$

$$\bar{\boldsymbol{g}}_t = \sum_{k=1}^{K} p_k \nabla F_k\big(\boldsymbol{w}_t^k\big). \tag{A35b}$$

By following the same procedure, we obtain the similar result as follows:

$$\mathbb{E}\Big[\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^\star\|^2\Big] \leq (1-\mu\eta_t)\mathbb{E}\Big[\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|^2\Big] + 2\mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2\Big] + 6L\Gamma\eta_t^2. \tag{A36}$$

Thus, we complete the proof. □

Appendix C.2. Proof of Lemma A2

Proof. According to Assumption 3, the variance of stochastic gradients in each client is bounded by ϕ_k^2/b. Then, the $\mathcal{A}_2$ term in (A24) can be bounded by

$$\begin{aligned}
\mathbb{E}\Big[\|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2\Big] &= \mathbb{E}\Big[\Big\|\sum_{k=1}^{K} p_k\Big(\Pi_{\mathcal{G}} \nabla F_k\Big(\boldsymbol{w}_t^k, \mathcal{B}_t^k\Big) + \boldsymbol{\xi}_t^k\Big) - \sum_{k=1}^{K} p_k\Big(\Pi_{\mathcal{G}} \nabla F_k\big(\boldsymbol{w}_t^k\big) + \boldsymbol{\xi}_t^k\Big)\Big\|^2\Big] \\
&= \sum_{k=1}^{K} p_k^2 \mathbb{E}\Big[\|\nabla F_k\Big(\boldsymbol{w}_t^k, \mathcal{B}_t^k\Big) - \nabla F_k\big(\boldsymbol{w}_t^k\big)\|^2\Big] \\
&\leq \sum_{k=1}^{K} p_k^2 \phi_k^2 / b.
\end{aligned} \tag{A37}$$

Thus, the proof is completed. □

Appendix C.3. Proof of Lemma A3

Proof. For any $t \in [0, Q-1]$, there exists $t_0 \leq t$ such that $t - t_0 \leq Q-1$ and $\bar{\boldsymbol{w}}_{t_0} = \boldsymbol{w}_{t_0}^k$ for $k \in [0, N]$. Assume $\eta_{t_0} \leq 2\eta_t$ and η_t is non-increasing, i.e., $\eta_t \leq \eta_{t_0}$. When $t+1 \in \mathcal{I}_Q$, we have

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2\Big] &= \mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\Big(\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{t_0}\Big) - (\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t_0})\|^2\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{t_0}\|^2\Big] - \mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{t_0}\|^2\Big] \\
&\overset{(a)}{\leq} \mathbb{E}\Big[\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{t_0}\|^2\Big] \\
&\overset{(b)}{\leq} \sum_{k=1}^{K} p_k \mathbb{E}\Big[\sum_{\tau=t_0}^{t} (Q-1)\eta_\tau^2 \|\nabla F_k\Big(\boldsymbol{w}_\tau^k, \mathcal{B}_\tau^k\Big) + z_\tau^k\|^2\Big] \\
&\leq \sum_{k=1}^{K} p_k \sum_{\tau=t_0}^{t} (Q-1)\eta_{t_0}^2 \Big(\mathcal{G}^2 + d\sigma_{t,k}^2\Big) \\
&\leq 4(Q-1)^2 \eta_t^2 \Big(\mathcal{G}^2 + d\sum_{k=1}^{K} p_k \sigma_{t,k}^2\Big),
\end{aligned} \tag{A38}$$

where (a) follows from the convexity of $\|\cdot\|^2$, and (b) holds since $\mathbb{E}\|\boldsymbol{X}-\mathbb{E}\boldsymbol{X}\|^2=\mathbb{E}\|\boldsymbol{X}\|^2-\|\mathbb{E}[\boldsymbol{X}]\|^2$. Similarly, when $t+1\notin\mathcal{I}_Q$, we can derive a similar result:

$$\mathbb{E}\Big[\sum_{k=1}^{K}p_k\big\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_t^k\big\|^2\Big]\leq 4(Q-1)^2\mathcal{G}^2\eta_t^2. \tag{A39}$$

Thus, the proof is completed. □

Appendix C.4. Proof of Lemma A4

Proof. When $t+1\in\mathcal{I}_Q$, the DP noise in each round is bounded by

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k=1}^{K}p_k\|\boldsymbol{\xi}_t^k\|^2\Big]&=\sum_{k=1}^{K}p_k(\sigma_t^k)^2\\
&=\frac{128\eta_t^2Q^4b^2\mathcal{G}^2}{n\epsilon^2}\sum_{k=1}^{K}\frac{\log(1.25Qb/n_k\delta)}{n_k}\\
&\overset{(a)}{\leq}\frac{128\eta_t^2Q^4b^2\mathcal{G}^2}{n\epsilon^2}\int_{k=1}^{n}\frac{\log(1.25Qb/k\delta)}{k}dk\\
&=-\frac{128\eta_t^2Q^4b^2\mathcal{G}^2}{n\epsilon^2}\int_{k=1}^{n}\log(k\delta/1.25Qb)d\log(k\delta/1.25Qb)\\
&=-\frac{64\eta_t^2Q^4b^2\mathcal{G}^2}{n\epsilon^2}\Big(\log^2(n\delta/1.25Qb)-\log^2(\delta/1.25Qb)\Big)\\
&=\frac{64\eta_t^2Q^4b^2\mathcal{G}^2}{n\epsilon^2}\log(n)\log(1.25^2Q^2b^2/n\delta^2)\\
&\overset{(b)}{\leq}\frac{4Q^4b^2\mathcal{G}^2}{nL^2\epsilon^2}\log(n)\log(1.25^2Q^2b^2/n\delta^2),
\end{aligned}\tag{A40}$$

where (a) follows $\sum_{k=1}^{K}\log(1.25Qb/n_k\delta)/n_k\leq\int_{k=1}^{n}\big(\log(1.25Qb/k\delta)/k\big)dk$, and (b) holds due to $\eta_t\leq 1/4L$. □

Appendix D. Proof of Theorem 4

As defined in Appendix A, let $\mathcal{I}_Q=\{zQ|z=1,2,\cdots\}$ be the set of global synchronization steps. Note that, for each communication round, the PS randomly activates a subset of clients $\mathcal{S}_t$ according to some sampling schemes, and then FedAvg only performs updates on selected clients. Thus, the $\mathcal{S}_t$ varies in each communication round. Additionally, we define two virtual sequences $\bar{\boldsymbol{w}}_t=\sum_{k=1}^{K}p_k\boldsymbol{w}_t^k$ and $\bar{\boldsymbol{v}}_t=\sum_{k=1}^{K}p_k\boldsymbol{v}_t^k$; when $t+1\in\mathcal{I}_Q$, we define

$$\bar{\boldsymbol{g}}_t=\sum_{k=1}^{K}p_k\Big(\nabla F_k(\boldsymbol{w}_t^k)+\boldsymbol{\xi}_t^k\Big), \tag{A41}$$

$$\boldsymbol{g}_t=\sum_{k=1}^{K}p_k\Big(\nabla F_k(\boldsymbol{w}_t^k,\mathcal{B}_t^k)+\boldsymbol{\xi}_t^k\Big). \tag{A42}$$

When $t+1\notin\mathcal{I}_Q$, we define

$$\boldsymbol{g}_t=\sum_{k=1}^{K}p_k\nabla F_k(\boldsymbol{w}_t^k,\mathcal{B}_t^k), \tag{A43}$$

$$\bar{\boldsymbol{g}}_t=\sum_{k=1}^{K}p_k\nabla F_k(\boldsymbol{w}_t^k). \tag{A44}$$

Thus, we always have $\bar{\boldsymbol{v}}_{t+1} = \bar{\boldsymbol{w}}_t - \eta_t \boldsymbol{g}_t$ and $\mathbb{E}[\boldsymbol{g}_t] = \bar{\boldsymbol{g}}_t$. Then, the update scheme of DP-FedAvg with PCP being taken into consideration is

$$\boldsymbol{v}_{t+1}^k = \begin{cases} \boldsymbol{w}_t^k - \eta_t \nabla F_k(\boldsymbol{w}_t^k, \mathcal{B}_t^k), & \text{if } t+1 \notin \mathcal{I}_Q, \\ \boldsymbol{w}_t^k - \eta_t \left(\nabla F_k(\boldsymbol{w}_t^k, \mathcal{B}_t^k) + \boldsymbol{\xi}_t^k \right), & \text{if } t+1 \in \mathcal{I}_Q, \end{cases} \tag{A45}$$

$$\boldsymbol{w}_{t+1}^k = \begin{cases} \boldsymbol{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_Q, \\ \text{samples } \mathcal{S}_{t+1} \text{ and average } \left\{ \boldsymbol{v}_{t+1}^k \right\}_{k \in \mathcal{S}_{t+1}} & \text{if } t+1 \in \mathcal{I}_Q, \end{cases} \tag{A46}$$

where $\boldsymbol{v}_{t+1}^k$ denotes a one-step SGD update from $\boldsymbol{w}_t^k$ and $\boldsymbol{w}_{t+1}^k$ can be interpreted as the result after the communication steps.

Appendix D.1. Key Lemmas

Motivated by [6], when $t+1 \notin \mathcal{I}_Q$, we still have $\bar{\boldsymbol{w}}_{t+1} = \bar{\boldsymbol{v}}_{t+1}$. However, when $t+1 \in \mathcal{I}_Q$, we aim to establish this relation in the sense of expectation, i.e., $\mathbb{E}_{\mathcal{S}_{t+1}}\left[\bar{\boldsymbol{w}}_{t+1}\right] = \bar{\boldsymbol{v}}_{t+1}$. To achieve this, the sampling strategy on the PS must be unbiased. We identify two sampling and averaging schemes that satisfy this requirement and provide convergence guarantees.

(I) If the PS performs the sampling with replacement to establish $\mathcal{S}_{t+1}$, then each element in $\mathcal{S}_{t+1}$ may occur multiple times, with each client being sampled with equal probability. In this case, the PS averages the parameters by $\boldsymbol{w}_{t+1}^k = \frac{1}{K} \sum_{k \in \mathcal{S}_{t+1}} \boldsymbol{v}_{t+1}^k$.

(II) If the PS performs the sampling without replacement to establish $\mathcal{S}_{t+1}$, then each element in $\mathcal{S}_{t+1}$ can appear only once. In this case, the PS averages the parameters by $\boldsymbol{w}_{t+1}^k = \frac{N}{K} \sum_{k \in \mathcal{S}_{t+1}} p_k \boldsymbol{v}_{t+1}^k$.

Due to the fact that $\mathcal{S}_{t+1}$ varies in each communication round, we assume that the DP-FedAvg algorithm activates all devices at the beginning of each communication round. The updated parameters depend only on the sampled clients, and the next-round parameters are broadcasted to all clients. This approach simplifies our theoretical analysis. As expected, $\boldsymbol{w}_{t+1}^k$ is equal to the weighted average of the parameters across all devices after the SGD updates.

Lemma A5. *If $t+1 \in \mathcal{I}_Q$, for both client sampling with/without replacement schemes, we have*

$$\mathbb{E}_{\mathcal{S}_{t+1}}\left[\bar{\boldsymbol{w}}_{t+1}\right] = \bar{\boldsymbol{v}}_{t+1}. \tag{A47}$$

Proof. Please refer to the proof of Lemma 4 in [6]. □

Lemma A6. *For $t+1 \in \mathcal{I}_Q$, assume η_t is non-increasing and $\eta_t \leq 2\eta_{t+Q}$; then, for client sampling with the replacement scheme,*

$$\mathbb{E}_{\mathcal{S}_{t+1}}\left[\left\| \bar{\boldsymbol{w}}_{t+1} - \bar{\boldsymbol{w}}_{t+1} \right\|^2 \right] = \frac{4}{K} \eta_t^2 Q^2 \mathcal{G}^2, \tag{A48}$$

and for client sampling without the replacement scheme,

$$\mathbb{E}_{\mathcal{S}_{t+1}}\left[\left\| \bar{\boldsymbol{w}}_{t+1} - \bar{\boldsymbol{w}}_{t+1} \right\|^2 \right] = \frac{N-K}{N-1} \frac{4}{K} \eta_t^2 Q^2 \mathcal{G}^2. \tag{A49}$$

Proof. Please refer to the proof of Lemma 5 in [6]. □

Appendix D.2. Completing the Proof of Theorem 4

Proof.

$$\begin{aligned}\|\bar{\boldsymbol{w}}_{t+1}-\boldsymbol{w}^\star\|^2 &= \|\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1}+\bar{\boldsymbol{v}}_{t+1}-\boldsymbol{w}^\star\|^2 \\ &= \|\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1}\|^2+\|\bar{\boldsymbol{v}}_{t+1}-\boldsymbol{w}^\star\|^2+2\underbrace{\langle \bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1},\bar{\boldsymbol{v}}_{t+1}-\boldsymbol{w}^\star\rangle}_{\triangleq\mathcal{A}_3}, \end{aligned} \tag{A50}$$

where the $\mathcal{A}_3$ term will be 0 due to Lemma A5. When $t+1\notin\mathcal{I}_Q$, the first term will vanish due to $\bar{\boldsymbol{w}}_{t+1}=\bar{\boldsymbol{v}}_{t+1}$. Then, we have

$$\mathbb{E}\Big[\big\|\bar{\boldsymbol{w}}_{t+1}-\boldsymbol{w}^\star\big\|^2\Big]\le(1-\mu\eta_t)\mathbb{E}\Big[\big\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}^\star\big\|^2\Big]+A\eta_t^2. \tag{A51}$$

When $t+1\in\mathcal{I}_Q$, we use Lemma A6 to bound the second term in the second equality:

$$\begin{aligned}\mathbb{E}\Big[\big\|\bar{\boldsymbol{w}}_{t+1}-\boldsymbol{w}^\star\big\|^2\Big] &= \mathbb{E}\Big[\big\|\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1}\big\|^2\Big]+\mathbb{E}\Big[\big\|\bar{\boldsymbol{v}}_{t+1}-\boldsymbol{w}^\star\big\|^2\Big] \\ &\le(1-\mu\eta_t)\mathbb{E}\Big[\big\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}^\star\big\|^2\Big]+(A+B+H)\eta_t^2, \end{aligned} \tag{A52}$$

where H is defined in Theorem 4. The main difference between (A16) and (A52) is the additional term H. Thus, we can follow the same progress and argument to prove the theorem. To be specific, we choose $\nu=\max\Big\{\frac{\beta^2(A+B+H)}{\mu\beta-1},\alpha\triangle_0\Big\}$, $\beta=\frac{2}{\mu}$ and $\eta_t=\frac{2}{\mu}\frac{1}{\alpha+t}$, where $\alpha\ge\max\{8\kappa,Q\}$, $\kappa=\frac{L}{\mu}$. By the L-smoothness of $F(\cdot)$, we have

$$\mathbb{E}\big[F(\bar{\boldsymbol{w}}_t)-F^\star\big]\le\frac{L}{2}\mathbb{E}\Big[\big\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}^\star\big\|^2\Big]\le\frac{L}{2}\frac{\nu}{t+\alpha}, \tag{A53}$$

then,

$$\mathbb{E}\big[F(\bar{\boldsymbol{w}}_t)-F^\star\big]\le\frac{2\kappa}{t+\alpha}\Big(\frac{A+B+H}{\mu}+2L\big\|\boldsymbol{w}_1-\boldsymbol{w}^\star\big\|^2\Big). \tag{A54}$$

Thus, we complete the proof. □

References

1. Li, Y.; Chang, T.H.; Chi, C.Y. Secure Federated Averaging Algorithm with Differential Privacy. In Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Espoo, Finland, 17–20 September 2020; pp. 1–6.
2. Sattler, F.; Wiedemann, S.; Müller, K.R.; Samek, W. Robust and communication-efficient federated learning from non-iid data. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 3400–3413. [CrossRef] [PubMed]
3. Wang, S.; Chang, T.-H. Federated Matrix Factorization: Algorithm Design and Application to Data Clustering. *IEEE Trans. Signal Process.* **2022**, *70*, 1625–1640. [CrossRef]
4. Wang, S.; Xu, Y.; Yuan, Y.; Quek, T.Q.S. Toward Fast Personalized Semi-Supervised Federated Learning in Edge Networks: Algorithm Design and Theoretical Guarantee. *IEEE Trans. Wireless Commun.* **2024**, *23*, 1170–1183. [CrossRef]
5. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
6. Li, X.; Huang, K.; Yang, W.; Wang, S.; Zhang, Z. On the Convergence of FedAvg on non-IID Data. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020; pp. 1–26.
7. Li, Y.; Wang, S.; Chi, C.Y.; Quek, T.Q.S. Differentially Private Federated Clustering Over Non-IID Data. *IEEE Internet Things J.* **2024**, *11*, 6705–6721. [CrossRef]
8. Glasgow, M.R.; Yuan, H.; Ma, T. Sharp bounds for federated averaging (local SGD) and continuous perspective. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Virtual, 28–30 March 2022; pp. 9050–9090.
9. Woodworth, B.E.; Patel, K.K.; Srebro, N. Minibatch vs local SGD for heterogeneous distributed learning. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020; pp. 6281–6292.

10. Hsu, T.M.H.; Qi, H.; Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv* **2019**, arXiv:1909.06335.
11. Li, T.; Sahu, A.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.J.; Stich, S.U.; Suresh, A.T. Scaffold: Stochastic controlled averaging for on-device federated learning. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020; pp. 5132–5143.
13. Geiping, J.; Bauermeister, H.; Dröge, H.; Moeller, M. Inverting gradients-how easy is it to break privacy in federated learning? In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Virtual, 6–12 December 2020; pp. 16937–16947.
14. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends® Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
15. Li, Y.; Huang, C.W.; Wang, S.; Chi, C.Y.; Quek, T.Q. Privacy-Preserving Federated Primal-Dual Learning for Non-Convex and Non-Smooth Problems With Model Sparsification. *IEEE Internet Things J.* **2024**, *11*, 25853–25866. [CrossRef]
16. Wang, X.; Wang, S.; Li, Y.; Fan, F.; Li, S.; Lin, X. Differentially Private and Heterogeneity-Robust Federated Learning with Theoretical Guarantee. *IEEE Trans. Artif. Intell.* **2024**, *5*, 6369–6384. [CrossRef]
17. Li, Z.; He, Y.; Yu, H.; Kang, J.; Li, X.; Xu, Z.; Niyato, D. Data heterogeneity-robust federated learning via group client selection in industrial IoT. *IEEE Internet Things J.* **2022**, *9*, 17844–17857. [CrossRef]
18. Wang, S.; Xu, Y.; Wang, Z.; Chang, T.-H.; Quek, T.Q.S.; Sun, D. Beyond ADMM: A Unified Client-Variance-Reduced Adaptive Federated Learning Framework. In Proceedings of the AAAI, Washington, DC, USA, 7–14 February 2023; pp. 10175–10183.
19. Wu, H.; Wang, P. Fast-convergent federated learning with adaptive weighting. *IEEE Trans. Cogn. Commun. Netw.* **2021**, *7*, 1078–1088. [CrossRef]
20. Wu, H.; Tang, X.; Zhang, Y.J.A.; Gao, L. Incentive Mechanism for Federated Learning With Random Client Selection. *IEEE Trans. Netw. Sci. Eng.* **2024**, *11*, 1922–1933. [CrossRef]
21. Saha, R.; Misra, S.; Chakraborty, A.; Chatterjee, C.; Deb, P.K. Data-centric client selection for federated learning over distributed edge networks. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *34*, 675–686. [CrossRef]
22. Pang, J.; Yu, J.; Zhou, R.; Lui, J.C. An incentive auction for heterogeneous client selection in federated learning. *IEEE Trans. Mob. Comput.* **2022**, *22*, 5733–5750. [CrossRef]
23. 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]
24. Li, Y.; Wang, S.; Chi, C.Y.; Quek, T.Q. Differentially Private Federated Learning in Edge Networks: The Perspective of Noise Reduction. *IEEE Netw.* **2022**, *36*, 167–172. [CrossRef]
25. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318.
26. McMahan, H.B.; Ramage, D.; Talwar, K.; Zhang, L. Learning Differentially Private Recurrent Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
27. 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]
28. Erlingsson, Ú.; Feldman, V.; Mironov, I.; Raghunathan, A.; Talwar, K.; Thakurta, A. Amplification by shuffling: From local to central differential privacy via anonymity. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, San Diego, CA, USA, 6–9 January 2019; pp. 2468–2479.
29. Balle, B.; Barthe, G.; Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 3–8 December 2018; pp. 6277–6287.
30. Wei, K.; Li, J.; Ding, M.; Ma, C.; Yang, H.H.; Farokhi, F.; Jin, S.; Quek, T.Q.S.; Poor, H.V. Federated Learning With Differential Privacy: Algorithms and Performance Analysis. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3454–3469. [CrossRef]
31. Noble, M.; Bellet, A.; Dieuleveut, A. Differentially Private Federated Learning on Heterogeneous Data. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Virtual, 28–30 March 2022; pp. 10110–10145.
32. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the IEEE symposium on security and privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 3–18.
33. Zhang, Y.; Duchi, J.C.; Wainwright, M.J. Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.* **2013**, *14*, 3321–3363.

34. Blake, C.L.; Merz, C.J. UCI Repository of Machine Learning Databases. 1996. Available online: https://archive.ics.uci.edu/dataset/2/adult (accessed on 30 April 1996).
35. LeCun, Y.; Cortes, C.; Burges, C. The MNIST Database. 1998. Available online: http://yann.lecun.com/exdb/mnist (accessed on 1 November 1998).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.