Article

Privacy-Preserving Distributed Learning via Newton Algorithm

Zilong Cao, Xiao Guo and Hai Zhang *

School of Mathematics, Northwest University, Xi'an 710127, China; nwu_czl@stumail.nwu.edu.cn (Z.C.); xiaoguo@nwu.edu.cn (X.G.)
* Correspondence: zhanghai@nwu.edu.cn

Abstract: Federated learning (FL) is a prominent distributed learning framework. The main barriers of FL include communication cost and privacy breaches. In this work, we propose a novel privacy-preserving second-order-based FL method, called GDP-LocalNewton. To improve the communication efficiency, we use Newton’s method to iterate and allow local computations before aggregation. To ensure strong privacy guarantee, we make use of the notion of differential privacy (DP) to add Gaussian noise in each iteration. Using advanced tools of Gaussian differential privacy (GDP), we prove that the proposed algorithm satisfies the strong notion of GDP. We also establish the convergence of our algorithm. It turns out that the convergence error comes from the local computation and Gaussian noise for DP. We conduct experiments to show the merits of the proposed algorithm.

Keywords: federated learning; differential privacy; second-order method

MSC: 68P27; 68W15

1. Introduction

Federated learning (FL) is a popular distributed learning paradigm that enables a large amount of clients to collaboratively train a global model without sharing their individual data [1]. The most popular algorithm is called federated averaging (FedAvg). In each round, the server broadcasts the current model to all the clients. The clients then run multiple steps of stochastic gradient descent (SGD) in a distributed fashion. After that, the server updates the global model by aggregating results from local clients. The convergence rate of FedAvg has been studied extensively in recent years [2–4].

One desiderata of a real FL paradigm is the communication efficiency. The communication cost of FL mainly comes from the latency cost, that is, the fixed cost of sending messages, which is proportional to the communication rounds, regardless of the size of the message. In pursuit of this, many first-order gradient-based methods have been developed, such as local stochastic gradient descent (Local SGD) [1,5–7] and mini-batch SGD [8–10]. These algorithms reduce the communication cost by performing local computations at the client devices before aggregation. It works when the clients are mobile resources which have reasonable computational power but may suffer from the communication latency [11]. However, for some serverless systems, say, the cloud-based systems, the latency cost is severe. The communication failures are more frequent. Hence, the communication rounds should be further reduced. Recent years have seen a few work on using second-order-based methods [11–15] to improve the convergence of first-order-based methods. Wang et al. [12] proposed the GIANT algorithm, a new distributed approximation of Newton’s method, and showed an improved convergence rate over the distributed first-order competitors. Dünner et al. [13] developed a method which approximates the global Hessian matrix by using the block Hessian matrices from users, thereby reducing the computational burden of the global Hessian matrix. Gupta et al. [11] proposed and analyzed a second-order method with local computations, called LocalNewton. Due to its second-order nature and local computations, LocalNewton is superior in terms of the communication cost.
The other desideratum of a real FL paradigm is the privacy preservation. The data from local clients often contain sensitive information of individuals. To some extent, FL could preserve the individuals’ privacy because the original data never leave the clients. However, there exist adversarial attacks, which can result in privacy leakage by using the information during the communication; see [16] for example. Hence, the vanilla FL paradigm does not have a rigorous privacy guarantee. Differential privacy (DP; Dwork [17]) is a standard and well-adopted framework which provides strong guarantee of an individual’s privacy by ensuring that no individual has a significant influence on the algorithm’s output. The worst-case influence is indeed termed the privacy budget. A differentially private (DP) algorithm is often achieved by randomizing the algorithm’s output. Besides the vanilla \((e, \delta)\)-DP, several DP notions are developed, say, the Rényi differential privacy (RDP) [18] and the Gaussian differential privacy (GDP) [19]. RDP and GDP enjoy an elegant composition property, that is, the theoretical bound of the privacy leakage of repeated queries can be tight and lossless, which would bring benefits when accounting for the privacy leakage in iterative-based algorithms. Using the notion of DP, there have been several studies on developing a privacy-preserving FL algorithm; see McMahan et al. [2], Geyer et al. [3], Triastcyn and Faltins [4], Noble et al. [20], Wei et al. [21], Girgis et al. [22], Cheu et al. [23], Rastogi and Nath [24], Huang et al. [25], among others. Most of these algorithms are first-order based.

With the two desiderata in mind, we develop a novel privacy-preserving second-order-based method called \textit{GDP-LocalNewton} within the FL framework. Building upon Newton’s algorithm, each device performs Newton’s iterations locally. To avoid privacy leakage, we take advantage of the notion of GDP [19] to add Gaussian noise to the updates in each iteration. In particular, we propose a novel linear search method to determine the step size in the Newton update. After several local steps, the devices send parameter updates to the central server. The server then aggregates these updates and broadcasts them to synchronize all local parameters. This process is iterated until convergence or specific conditions are attained. It is worth mentioning that we assume a specific adversary termed as the curious onlooker, who can eavesdrop on the communication between the server and the local machines. The server is a typical example of a curious onlooker.

We rigorously show that under proper parameters set-up, the proposed algorithm satisfies GDP. We also analyze the convergence bound of \textit{GDP-LocalNewton}. It turns out that the convergence error comprises two components. One comes from the DP noises, and the other comes from the local computations. We also conduct experiments to show the merits of \textit{GDP-LocalNewton}.

2. GDP-LocalNewton

In this section, we first provide the problem formulation and the framework of GDP. Then, we develop the \textit{GDP-LocalNewton} algorithm. After that, we provide rigorous privacy analysis and convergence analysis of the proposed algorithm.

2.1. Problem Formulation

2.1.1. Some Notations and Symbols

We represent vectors (e.g., \(\mathbf{g}\)) and matrices (e.g., \(\mathbf{H}\)) as bold lowercase and uppercase letters, respectively. \(\|\mathbf{g}\|\) denotes its \(\ell_2\) norm, and the spectral norm of matrix \(\mathbf{H}\) is denoted by \(\|\mathbf{H}\|\). \(\mathbf{I}\) denotes an identity matrix, and the set \(\{1, 2, \ldots, n\}\) is denoted by \([n]\). To distinguish the index of workers and iterations, we denote the worker index as the superscript (e.g., \(\mathbf{g}^k\)) and denote the iteration counter as the subscript (e.g., \(\mathbf{g}_t\)).

Our paper contains a large number of parameters. For ease of reading and distinction, we list them in Table 1.
2.1.2. Fundamental Problem

We consider the empirical risk minimization problems of the following form:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ f(\mathbf{w}) \equiv \frac{1}{n} \sum_{j=1}^{n} f_j(\mathbf{w}) \right\},$$

(1)

where $f_j(\cdot) : \mathbb{R}^d \to \mathbb{R}$, for all $j \in \left[ n \right] = 1, 2, \ldots, n$, represents the loss of the $j$-th observation, given an underlying parameter $\mathbf{w} \in \mathbb{R}^d$. Generally, we call the $f(\cdot)$ the global loss function. In the field of machine learning, such problems are general, e.g., logistic and linear regression, support vector machines, neural networks and graphical models. Taking the logistic regression as an example, we have

$$f_j(\mathbf{w}) = \ell_j(\mathbf{w}^T \mathbf{x}_j) = \log(1 + \exp^{-y_j \mathbf{w}^T \mathbf{x}_j}) + \frac{\gamma}{2} \| \mathbf{w} \|_2^2,$$

where $\ell_j(\cdot)$ is the loss function for sample $j \in \left[ n \right]$ and $\gamma$ is an approximately chosen regularization parameter. $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ is the sample matrix containing $n$ data, and $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ is the corresponding label vector. Hence, $(\mathbf{X}, \mathbf{y})$ defines the training dataset, which is composed as $(\mathbf{x}_j, y_j)$, $j = 1, 2, \ldots, n$.

At the $t$-th iteration, the gradient and the Hessian are denoted by

$$\mathbf{g}_t = \nabla f(\mathbf{w}_t) = \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(\mathbf{w}_t),$$

and

$$\mathbf{H}_t = \nabla^2 f(\mathbf{w}_t) = \frac{1}{n} \sum_{j=1}^{n} \nabla^2 f_j(\mathbf{w}_t),$$

respectively, where $\mathbf{w}_t$ is the updated parameter at the $t$-th iteration. Specially, in this paper, we make some assumptions about loss functions. Global loss function $f(\cdot)$ is smooth and strongly convex. $f_i(\cdot)$ is twice differentiable, $\nabla^2 f_i(\mathbf{w}) \preceq B \mathbf{I}$ and $\| \nabla f_i(\cdot) \|_2 \leq \Gamma$, where $B \in \mathbb{R}$ and $\Gamma \in \mathbb{R}$ are some fixed constants.

2.1.3. Data Distribution

Assume that the FL system has $K$ workers in total. $\mathcal{S}_k$, for all $k \in [K] = \{1, 2, \ldots, K\}$, represents a subset chosen uniformly from $\left[ n \right] = \{1, 2, \ldots, n\}$ without replacement randomly, and the workers can be identified by the subsets, uniquely. The number of each
workers’ sample is equal to \( s = |S_k| \) for any \( k \in [K] \). Due to the sampling without replacement assumption, we have \( S_1 \cup S_2 \cup \ldots \cup S_k = [n] \) and \( S_i \cap S_j = \emptyset \) for all \( i, j \in [K] \). Also, \( K = n/s \) represents the number of workers.

2.2. Gaussian Differential Privacy

In this subsection, we review the definition of Gaussian differential privacy (GDP) [26]. Then, we give some theorems and tools about GDP [26], which are the base of the proof that our algorithm satisfies \( \mu - GDP \). Lemmas 1 and 2 show that every user’s output satisfies \( \mu_k - GDP \). Lemma 3 guarantees DP when the entire dataset is divided as disjointed parts.

**Definition 1** (f-DP). Let \( h : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p \) be a randomized function, and \( \mathcal{D} = \{x_1, \ldots, x_i, \ldots, x_n\} \) be a dataset. If the output of \( h \) for any \( a \)-level test between the hypotheses of the formulation \( H_0 : x_i = t \) vs. \( H_1 : x_i = s \), which has power function \( \beta(\alpha) \leq 1 - f(\alpha) \), where \( f \) is a convex, continuous, non-increasing function and for all \( \alpha \in [0, 1] \) \( f(\alpha) \leq 1 - \alpha \), then we say \( h \) satisfies \( f \)-DP.

**Definition 2** (Gaussian differential privacy). Define \( \Phi(\cdot) \) as the standard Gaussian cumulative distribution. If \( h \) is \( f \)-DP and

\[
f(\alpha) \geq \Phi(\Phi^{-1}(1 - \alpha) - \mu), \alpha \in [0, 1],
\]

then \( h \) is \( \mu \)-Gaussian differentially private (\( \mu \)-GDP).

**Definition 3** (Global sensitivity). Let \( g : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p \) be a deterministic function. The global sensitivity of \( g \) is the (possibly infinite) number

\[
 GS_g = \sup_{x_{1:n}, x'_{1:n} \in \mathbb{R}^{n \times m}} \{ \|g(x'_{1:n}) - g(x_{1:n})\|_2 : d_H(x_{1:n}, x'_{1:n}) = 1 \}, \tag{2}
\]

where \( x_{1:n} \) means a dataset \( \{x_1, x_2, \ldots, x_n\} \), and \( x'_{1:n} \) is different from \( x_{1:n} \) at some datum.

**Lemma 1** (Gaussian Mechanism of GDP [26]). Let \( g : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p \) be a function with finite global sensitivity \( GS_g \). Let \( Z \) be a standard normal \( p \)-dimensional random vector. For all \( \mu > 0 \) and \( x \in \mathbb{R}^{n \times m} \), the random function \( h(x) = g(x) + \frac{GS_g}{\mu}Z \) is \( \mu \)-GDP.

**Lemma 2** (Matrix Gaussian mechanism [26]). Consider a data matrix \( A \in \mathbb{R}^{n \times m} \) such that each row vector \( a_i \) satisfies \( \|a_i\| \leq 1 \). Further define the function \( h(A) = \frac{1}{\mu}A^TA \). Let \( W \) be a symmetric random matrix whose upper-triangular elements and the diagonal are i.i.d. \( \frac{1}{\mu}N(0, 1) \). Then the random function \( h(A) = h(A) + W \) satisfies \( \mu \)-GDP.

**Lemma 3** (Parallel Composition [27]). Suppose we have a set of privacy mechanisms \( \mathcal{M} = \{M_1, \ldots, M_m\} \). If each \( M_i \) provides a \( \mu_i \)-GDP guarantee on a disjointed subset of the entire dataset, \( \mathcal{M} \) will provide \( (\max\{\mu_1, \ldots, \mu_m\}) \)-GDP.

**Lemma 4** (Composition of GDP[19]). The \( n \)-fold composition of \( \mu_i - GDP \) mechanisms is \( \sqrt{\mu_1^2 + \ldots + \mu_n^2} - GDP \).

2.3. GDP-LocalNewton Algorithm

In this subsection, we propose the GDP-LocalNewton algorithm for privacy-preserving distributed learning; see Algorithm 1.
Algorithm 1 GDP-LocalNewton

1: **Input:** Initial iteration $w_0 \in \mathbb{R}^d$; Max linear search parameter $\alpha^*$; Linear search parameter $0 < \beta \leq 1/2$; Privacy parameter $\mu$; Loss function parameter $B$ and $\Gamma$; Iteration parameter $T$; $I_T \subseteq \{1, 2, \ldots, T\}$ is the set of communication nodes.

2: **for** $k = 1$ to $K$ in parallel **do**

3: **Initialization:** $w_0^k = w_0$

4: **for** $t = 0$ to $T - 1$ **do**

5: if $t \in I_t$ then

6: $\tilde{w}_t = \frac{1}{k} \sum_{k=1}^{k} w_t^k$

7: $w_t^k = \tilde{w}_t$

8: end if

9: $\hat{g}_t^k = \hat{g}^k + \frac{2B\sqrt{T}}{\mu s} Z_t^k$

10: $\hat{H}_t^k = \hat{H}_t^k + \frac{2B\sqrt{T}}{\mu s} U_t^k$

11: $\hat{p}_t^k = \hat{H}_t^k - \hat{g}_t^k$

12: Find step-size $\alpha_t^k$ using line search (Equation (10))

13: Update model: $w_{t+1}^k = w_t^k - \alpha_t^k \hat{p}_t^k$

14: end for

15: end for

At the $k$-th worker in the $t$-th iteration, we define the **local loss function** (at the local iteration $w_t^k$) as

$$f^k(w_t^k) = \frac{1}{\bar{s}} \sum_{j \in S_k} f_j(w_t^k).$$

(3)

The $k$-th worker’s target is to minimize the local loss in Equation (3). We give the corresponding local gradient $g_t^k$ and local Hessian $H_t^k$ at the $k$-th worker in the $t$-th iteration as the following:

$$g_t^k = g^k(w_t^k) = \nabla f^k(w_t^k) = \frac{1}{\bar{s}} \sum_{j \in S_k} \nabla f_j(w_t^k),$$

$$H_t^k = H^k(w_t^k) = \nabla^2 f^k(w_t^k) = \frac{1}{\bar{s}} \sum_{j \in S_k} \nabla^2 f_j(w_t^k).$$

The updates of the local workers’ parameter are given by

$$w_{t+1}^k = \begin{cases} w_t^k - \alpha_t^k (\hat{H}_t^k)^{-1} \hat{g}_t^k, & \text{if } t \notin I_t \\ \tilde{w}_t - \alpha_t^k \hat{H}_t^k (\tilde{w}_t) - \hat{g}_t^k (\tilde{w}_t), & \text{if } t \in I_t \end{cases}$$

(4)

where $\tilde{w}_t = \frac{1}{k} \sum_{k=1}^{k} w_t^k$ and

$$\hat{g}_t^k = g_t^k + \frac{2B\sqrt{T}}{\mu s} Z_t^k,$$

(5)

$$\hat{g}^k(\tilde{w}_t) = g^k(\tilde{w}_t) + \frac{2B\sqrt{T}}{\mu s} Z_t^k,$$

(6)

and

$$\hat{H}_t^k = H_t^k + \frac{2B\sqrt{T}}{\mu s} U_t^k,$$

(7)

$$\hat{H}^k(\tilde{w}_t) = H^k(\tilde{w}_t) + \frac{2B\sqrt{T}}{\mu s} U_t^k,$$

(8)

then,

$$\hat{p}_t^k = (\hat{H}_t^k)^{-1} \hat{g}_t^k,$$

(9)

where $Z_t^k$ is a sequence of i.i.d. standard $d$-dimensional Gaussian random vectors and $U_t^k$ is a sequence of i.i.d. symmetric random matrices, whose upper triangular elements and the
diagonals are i.i.d. standard normal randoms. \( \mu \) is the privacy parameter, \( B \) is the norm bound of the Hessian matrix and \( \Gamma \) is the norm bound of the gradient. \( \mathcal{I}_T \) is the set of communication nodes, where every user communicates their own updated parameters to the server, \( T \) represents the set of iterations, and \( L \) means the number of local computations. We have \( |T| = \lfloor |\mathcal{I}_T| \times L \rfloor \), e.g., \( T = \{1, 2, \ldots, L, \ldots, 2L, \ldots, \lfloor |\mathcal{I}_T| \times L \rfloor \} \). In particular, if \( L = 1 \), this means no local computation.

**Remark 1.** In practice, \( \hat{H}^k(w_i) \) may not be positive definite. So we need to truncate the eigenvalues as \( \max \{ \lambda_j, \epsilon \} \), where \( \epsilon > 0 \) is equal to the regularization parameter. The matrix is also differentially private as a result of the post-processing property of GDP ([19], Proposition 2.8).

Here, we propose a novel step selection, coping with the negative influence of noises, and it plays an important role in the convergence of GDP-LocalNewton.

**Step-size selection:** Let each worker locally choose a step size according to the following rule:

\[
a^*_k = \max_{\alpha \in \alpha^*} \text{ such that } f^k(w^k - \alpha \hat{p}^k) \leq f^k(w_k^k) - \alpha \beta(\hat{p}^k)^T g^k_k + \alpha \gamma (\hat{p}^k)^T H^k_k \tilde{N}^k_k,
\]

for some \( \beta \in (0, 1/2] \) and \( \gamma = 1 - \beta \) so that

\[
\alpha^* \leq \min \left\{ \frac{2(1-\beta)\kappa(1-\epsilon)}{M(1+\epsilon)}, \frac{2\alpha^2}{3M[(M-\kappa/4)L]} \right\} < 1,
\]

where \( \tilde{N}^k_k = (\hat{H}^k_k)^{-1} g^k_k - (H^k_k)^{-1} g^k_k \), \( M \) and \( \kappa \) are the parameters of the functional smooth and strong convex, proposed at the next subsection.

**Remark 2.** In the subsequent theoretical section, we can observe the advantages and significance of the linear search in the algorithm’s convergence analysis. In part of the experiment, we will compare our strategy with the fixed step size and decaying step size. The results show that our strategy is optimal. The decaying step size refers to compressing the step size of the previous iteration at each iteration, which means \( \alpha_i + 1 = \alpha^{i+1}_k \alpha_i \), where \( \alpha \) is the decaying rate.

### 2.4.1. Theoretical Results

#### 2.4.1. Assumptions on the Loss Functions

We need the following assumptions on the loss functions.

**Assumption 1.** \( f_i(\cdot) \), for all \( i \in [n] \), is twice differentiable.

**Assumption 2.** \( f(\cdot) \) is \( \kappa \)-strongly convex, that is, \( \nabla^2 f(w) \succ \kappa I \).

**Assumption 3.** \( f(\cdot) \) is \( M \)-smooth, that is, \( \nabla^2 f(w) \preceq MI \).

**Assumption 4.** For \( i \in [n] \), \( \|\nabla^2 f_i(\cdot)\|_2 \) and \( \|\nabla f_i(\cdot)\|_2 \) are upper bounded such that \( \nabla^2 f_i(w) \preceq BI \) and \( \|\nabla f_i(\cdot)\|_2 \leq \Gamma \), where \( B \in \mathbb{R} \) and \( \Gamma \in \mathbb{R} \) are some fixed constants.

#### 2.4.2. Privacy

Theorem 1 shows that GDP-LocalNewton satisfies \( \mu - \text{GDP} \).

**Theorem 1.** Assuming Assumptions 1–4 hold, the output of the \( k \)-worker (4) satisfies \( \frac{\mu}{\sqrt{T}} - \text{GDP} \). Furthermore, after \( T \) iterations, the whole algorithm satisfies \( \mu - \text{GDP} \).

**Proof.** Firstly, we compute the global sensitivity of \( s g^k \) and \( s H^k \) as the following:

\[
GS_{sg^k} = \|s g^k(x'_{1:n}) - s g^k(x_{1:n})\| = \|\nabla f_i(\cdot) - \nabla f_j(\cdot)\| \leq 2\Gamma,
\]

(11)
we know that the output of the $k$-worker (4) satisfies $\frac{\mu}{\sqrt{s}} - \text{GDP}$. Finally, through Lemmas 3 and 4, the whole algorithm satisfies $\mu - \text{GDP}$. The proof is completed. \qed

2.4.3. Convergence Analysis

The following three lemmas show that we can separate the Gaussian noises added in the algorithm and bound them so as to guarantee the convergence of $\text{GDP-LocalNewton}$.

**Lemma 5** ([26]). Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy $\sigma^2$. For any $\alpha > 0$, with probability at least $1 - \alpha$,

$$\|X\|_2 \leq 4\sigma \sqrt{d} + 2\sigma \sqrt{2 \log(1/\alpha)}.$$

**Lemma 6** ([26]). Let $W$ be a symmetric $d \times d$ random matrix whose upper triangular elements, including the diagonal, are i.i.d. $N(0, 1)$. For any $\alpha > 0$, with probability at least $1 - \alpha$,

$$\|W\|_2 \leq \sqrt{2d \log(2d/\alpha)}$$

**Lemma 7**. Assumptions 1–4 hold. Fix the total iteration $T$ and the GDP parameter $\mu$, let $s \geq \frac{4B\sqrt{T^2 \log(4d/\delta_0)}}{k(1-\epsilon)^3\mu}$, and use Lemma 1 and Lemma 2, then we can rewrite the noisy term in (4) as

$$(\hat{H}^k)_{-1}g^k = (H^k)_{-1}g^k + \hat{N}_k^k.$$

Here, with probability at least $1 - \delta_0$, we have

$$\|\hat{N}_k^k\| \leq C\hat{M}_{\text{privacy}},$$

where $\hat{M}_{\text{privacy}} = \frac{\Gamma B \sqrt{Td \log(d/\delta_0)}}{\mu \epsilon \sigma (1-\epsilon)^3}$ and $C'$ is a constant.

**Proof.** See Appendix B.1. \qed

**Lemma 8**. Let the function $f(\cdot)$ satisfy Assumptions 1–3, and suppose that step size $\alpha^k$ satisfies the line-search condition in (10). Also, let $0 < \epsilon < 1/2$, $0 < \delta, \delta_0 < 1$, $\mu$ and $T$ be fixed constants. Moreover, let the sample size $s \geq \frac{4B\sqrt{T^2 \log(4d/\delta_0)}}{k(1-\epsilon)^3\mu}$. Then, defined in Equation (4), at the $k$-th worker, we have

$$f^k(w^k_{t+1}) - f^k(w^k_t) \leq -\psi\|g^k_t\|^2 + \alpha^k(1 - 2\beta)(p^k_t)^T H^k_t \hat{N}_k^k + \alpha^k(1 - \beta)M(1 + \epsilon)\|\hat{N}_k^k\|^2,$$

with probability of at least $1 - \delta$.

If $s \geq \frac{4B\sqrt{T^2 \log(4d/\delta_0)}}{k(1-\epsilon)^3\mu}$, $\beta = 1/2$ and using Lemma 7, we obtain

$$f^k(w^k_{t+1}) - f^k(w^k_t) \leq -\psi\|g^k_t\|^2 + \frac{\alpha^k}{2} M(1 + \epsilon)(C\hat{M}_{\text{privacy}})^2,$$

with at least probability $1 - \delta - \delta_0$, where $\psi = \frac{\alpha^k}{M(1-\epsilon)}$. Note that $w^k_{t+1} = w^k_t - \alpha^k\hat{p}^k_t$, which means the workers’ local gradient $\hat{N}_k^k$, $\hat{M}_{\text{privacy}}$ and $C'$ are from Lemma 7.

**Proof.** See Appendix B.1. \qed

**Remark 3.** In Lemma 8, we can find that the latter two terms come from the noise by DP. The second term can be eliminated by setting the $\beta = 1/2$ and it is the merit of our new linear selection...
that can reduce the random negative effectiveness. As a result, Lemma 8 shows that the novel linear selection can improve the convergence of the algorithm. The small $\|N_t^k\|^2$ leads to the small third term.

**Theorem 2.** /L = 1 case/ Suppose that Assumptions 1–4 hold, and the step-size $\alpha_t^k$ satisfies the line-search condition (10). Also, let $T, \mu, 0 < \delta, \xi_0 < 1, 0 < \epsilon, \epsilon_1 < 1/2$ be fixed constants and let $\beta = 1/2, \Gamma = \max_{1 \leq i \leq n} \|\nabla f_i(\cdot)\|$. Moreover, assume that the sample size for each worker satisfies $s \geq \frac{4B^2T}{\epsilon^2} \log \frac{2d}{\delta}$, where the samples are chosen without replacement. When the GDP-LocalNewton is in the $L = 1$ case, we obtain the following:

1. If $s \geq \max \left\{ \frac{4B^2T \sqrt{2d \log(4p/\xi_0)}}{\epsilon^2(1-\epsilon)^2}, \frac{r^2}{\epsilon^2} \log(d/\delta) \right\}$ for $G = \min_k \|g^k(\bar{w}_t)\|$, we obtain, with probability at least $1 - K(6\delta + \xi_0)$,

$$f(\bar{w}_{t+1}) - f(\bar{w}^*) \leq \rho_1(f(\bar{w}_t) - f(\bar{w}^*)) + \bar{M}_\phi,$$

(14)

2. If $s \geq \frac{2B^2T \sqrt{2d \log(4p/\xi_0)}}{\epsilon^2(1-\epsilon)^2}$, we obtain, with probability at least $1 - K(6\delta + \xi_0)$,

$$f(\bar{w}_{t+1}) - f(\bar{w}^*) \leq \rho_2(f(\bar{w}_t) - f(\bar{w}^*)) + \bar{M}_\phi + \frac{\eta T}{\epsilon^2}.$$

(15)

Note that $\rho_i = 1 - 2\epsilon\xi_i$, for $i = \{1, 2\}$, $\xi_1 = \left( \frac{\epsilon_1}{\epsilon} - \frac{\epsilon_1}{\epsilon} - \frac{M-\epsilon_1(1-\epsilon)}{\epsilon^2(1-\epsilon)^2} \right)$, $C_2 = \left( \frac{\epsilon_1}{\epsilon} - \frac{\epsilon_1}{\epsilon} - \frac{M-\epsilon_1(1-\epsilon)}{\epsilon^2(1-\epsilon)^2} \right)$, $\psi = \frac{\epsilon^2}{\epsilon^2(1-\epsilon)^2}$, $\eta = \left( 1 + \frac{2\log(\frac{1}{\epsilon})}{\epsilon^2} \right) \frac{\sqrt{\Gamma}}{\sqrt{2}}$. $\bar{M}_\phi = \epsilon_1 (C^* \bar{M}_{\text{privacy}}(\Gamma) + \frac{1}{2} M + \epsilon)$ and $\bar{M}_\phi = \eta (C^* \bar{M}_{\text{privacy}}(\Gamma) + \frac{1}{2} M + \epsilon)$ are constants. $C^* \bar{M}_{\text{privacy}}(\Gamma)$ is from Lemma 7.

**Proof.** See Appendix C.1 and C.2. □

**Remark 4.** Compared to LocalNewton when $L = 1$ ([11]), the protection of privacy leads to an extra error term $\bar{M}_\phi$. The error term shows the trade-off between the protection of privacy and the usefulness of the model. If we want to decrease the error term $\bar{M}_\phi$, it needs more data in every worker.

In this theorem, we choose $\beta = 1/2$ so that the second random term in the Lemma 8 is eliminated. That is because the less random terms commonly make the algorithm more stable.

**Theorem 3.** /L > 1 case/ Suppose Assumptions 1–4 hold and step size $\alpha_t^k$ satisfies the line search condition (10). Also, let $T, \mu, 0 < \delta, \xi_0 < 1, 0 < \epsilon < 1/2$ be fixed constants and let $\Gamma = \max_{1 \leq i \leq n} \|\nabla f_i(\cdot)\|$. Moreover, assume that the sample size for each worker satisfies $s \geq \max \left\{ \frac{4B^2T}{\epsilon^2} \log \frac{2d}{\delta}, \frac{2B^2T \sqrt{2d \log(4p/\xi_0)}}{\epsilon^2(1-\epsilon)^2} \right\}$. Then, the GDP LocalNewton updates, $\bar{w}_{t+1}$, within the case of $L > 1$, satisfy

$$f(\bar{w}_{t+1}) - f(\bar{w}_t) \leq -\frac{C}{R} \sum_{k=1}^{K} \sum_{r=0}^{L} \|b^k_r\|^2 + \frac{\eta T}{\epsilon^2} \bar{M}_\phi + L \bar{M}_\phi,$$

(16)

with probability at least $1 - KL(6\delta + \xi_0)$. Here, $C = \psi - \frac{L(M-\epsilon_1(1-\epsilon))\|\alpha^*\|^2}{\epsilon^2(1-\epsilon)^2}, \psi = \frac{\epsilon^2}{\epsilon^2(1-\epsilon)^2}$, and $\eta = \left( 1 + \frac{2\log(\frac{1}{\epsilon})}{\epsilon^2} \right) \frac{\sqrt{\Gamma}}{\sqrt{2}}$. $\bar{M}_\phi = \eta (C^* \bar{M}_{\text{privacy}}(\Gamma) + \frac{1}{2} M + \epsilon)$ is a constant and $C^* \bar{M}_{\text{privacy}}(\Gamma)$ is from Lemma 7.

**Proof.** See Appendix C.3. □
3. Empirical Evaluation

In this section, we evaluate the numerical performance of GDP-LocalNewton under \( L = 1 \) and \( L > 1 \), respectively. In addition, we design experiments to explore the performance of different step size strategies, such as the linear search step size, fixed step size, and decaying step size. In particular, we conduct both the simulation and the real data experiments.

The simulated datasets are generated according to the logistic model. Supposing \( d = 10, n = 50,000 \), each entry of the model parameter vector \( w = (w_1, \ldots, w_{10}) \) is generated i.i.d. from \( U(-0.5, 0.5) \). For each sample, each entry of the predictor vector is generated i.i.d. from the standard Gaussian. The labels are then generated from the logistic regression. We also generate 10,000 samples for testing.

The real datasets we use are summarized in Table 2, which are publicly available in LIBSVM [28].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training Samples (n)</th>
<th>Features (d)</th>
<th>Testing Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>a9a</td>
<td>32,000</td>
<td>123</td>
<td>16,000</td>
</tr>
<tr>
<td>Covtype</td>
<td>500,000</td>
<td>54</td>
<td>80,000</td>
</tr>
<tr>
<td>ijcnn1</td>
<td>49,000</td>
<td>22</td>
<td>91,000</td>
</tr>
</tbody>
</table>

3.1. Without Local Computation (\( L = 1 \))

We fix the total clients \( K \) to be 50. Hence, the number of samples per client \( s \) is \( n/50 \). We fix the total communication rounds \( T \) as 10. The regularization parameter is 0.001. We set the privacy parameter \( \mu = 1, 5 \) and 10, respectively. We compare our method with GDP-GD algorithm with \( \mu = 1 \) and LocalNewton with \( \mu = \infty \). The details of GDP-GD can be found in Appendix D. Note that we set the max step size \( \alpha^* \) of GDP-LocalNewton to be the fixed step size of GDP-GD. This is because, due to noise accumulation, the step sizes of algorithms perturbed by GDP should be much smaller than those without DP perturbation in order to ensure convergence.

Figures 1 and 2 display the training loss and test accuracy of each method on one simulated dataset and three real datasets, where the training loss is the empirical average of the loss function evaluated at the training data points, and the test accuracy is the proportion of test samples that are correctly classified. As Figures 1 and 2 show, GDP-LocalNewton with different \( \mu \) converges much faster than GDP-GD with \( \mu = 1 \) in all datasets. It means that our privacy-preserving second-order algorithm is much better than the privacy-preserving first-order algorithm GDP-GD. We can also see that as the privacy parameter \( \mu \) increases, the performance of GDP-LocalNewton increases, especially when the communication round is large.

3.2. With Local Computation (\( L > 1 \))

We use the a9a and the ijcnn1 datasets to show how different values of \( L \) influence the performance of GDP-LocalNewton. We set the GDP parameter \( \mu = 2 \) and \( \mu = 4 \) for the a9a dataset and set \( \mu = 1 \) and \( \mu = 2 \) for the ijcnn1 dataset, the local computation parameter \( L = 1 \) and \( L = 3 \), and regularization \( \gamma = 0.5 \). The communication round is 8. The other parameter set-ups are the same as those of the former experiments.

As Figure 3 shows, GDP-LocalNewton converges in all datasets. When the values of \( \mu \) are equal, \( L = 3 \) can speed up the convergence of the algorithm. That means the noises do not influence the speed of convergence too much. At the 7-th communication of the algorithm, the error from the privacy protection is bigger than the error of the local computation. That means that the privacy protection influences the usefulness of GDP-LocalNewton more than the local computation.
Figure 1. The training loss of each method on four datasets with respect to the $L = 1$ case.

Figure 2. Cont.
3.3. Different Step Size Strategies

In this section, we explore how different step size strategies, including linear-search step size ($\alpha^* = 0.03$), fixed step size (fix $\alpha_k = 0.03$), and decaying step size (the initial step size $\alpha = 0.03$), impact $GDP$-$LocalNewton$ under different levels of privacy protection (different $\mu$ values). In addition, we set different decaying rates to show its impact. Under $L = 1$, we use the simulation and the a9a data. Due to the varying characteristics of the datasets, we made different parameter settings for some of them. In the generated data,
we set the $\mu = 1$ and 5 and a better decaying rate $\sigma$ as 0.5. In the a9a data, we set the $\mu = 5$ and 10 and a better decaying rate $\sigma$ as 0.9.

As Figures 4 and 5 show, the linear-search step size performs well and stably in every case. This validates the effectiveness of our strategy both theoretically and experimentally. The reason for the similarity between the fixed step size and the search step size is that the noise causes some of the search steps to be equal to the maximum step size set, making them equivalent to the fixed step size. For the decaying step size, the experiments show that the decaying rate is very important for the convergence. An appropriate decaying rate can help the algorithm converge. In the later stages of the algorithm, decaying to a smaller step size helps resist the negative impact of noise on the algorithm and further promotes convergence, such as Figures 4a and 5a. An excessive decaying rate leads to algorithm non-convergence, like the 0.1 decaying rate in every figure. However, in practice, adjusting the decaying rate will increase the operational costs and raise the risk of privacy leakage. Therefore, the linear search strategy is more effective, efficient, and stable.

Figure 4. The training loss of GDP-LocalNewton on the generated data with respect to different strategies of step size. (a,b) Correspond to $\mu = 1$ and $\mu = 5$, respectively. An appropriate decaying rate is the half decay ($\sigma = 0.5$).

Figure 5. The training loss of GDP-LocalNewton on the a9a data with respect to different strategies of step size. (a,b) Correspond to $\mu = 5$ and $\mu = 10$, respectively. An appropriate decaying rate is the nine-tenth decay ($\sigma = 0.9$).
4. Conclusions

In this paper, we proposed a novel algorithm called GDP-LocalNewton for privacy-preserving and communication-efficient distributed learning. To improve the communication efficiency, we developed the algorithm based on Newton’s method and traded more local computations between communications. To handle possible privacy leakage via the curious onlooker, we adopted the notion of GDP [19] by adding Gaussian noise to the updates of each local machine. In particular, we developed a step searching strategy to determine the step size in the noisy Newton update. We validated the effectiveness, efficiency, and stability of our strategy through experiments. We theoretically studied the convergence of GDP-LocalNewton, which turns out to have two error terms, one corresponding to the privacy protection and the other corresponding to the local computation. The experiments corroborated the theoretical findings.

There are some interesting problems that deserve further study. First, the DP framework would lead to strong privacy protection; however, a DP algorithm often leads to much accuracy loss compared with its non-DP counterpart. Therefore, it is desirable to consider other weaker notions of privacy [29]. Second, how to handle the heterogeneity among local datasets algorithmically and theoretically in the second-order-based framework deserves further study. Finally, it is worth tackling other issues or challenges that FL faces within the second-order-based framework.

Author Contributions: Conceptualization, H.Z.; methodology, Z.C., X.G. and H.Z.; experiment, Z.C.; theory, Z.C. and X.G.; writing, Z.C., X.G. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by: 1. National Natural Science Foundation of China-Guangdong Joint Fund (U1811461). 2. Natural Science Foundation of Shaanxi Province (No. 2021JQ-429).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Appendix A. Some Auxiliary Lemmas

Here, we prove the auxiliary lemmas that are used in the main proofs of the paper.

Lemma A1 ([11]). Let \( f(\cdot) \) satisfy Assumptions 1 – 4, and \( 0 < \epsilon < 1/2 \) and \( 0 < \delta < 1 \) be fixed constants. Then, if \( s > 4B \frac{1}{\kappa^2} \log \frac{2d}{\delta} \), the local Hessian matrix at the \( k \)-th worker satisfies

\[
(1 - \epsilon)\kappa \leq \nabla^2 f^k(w) = H^k(w) \leq (1 + \epsilon)M, \tag{A1}
\]

for all \( w \in \mathbb{R}^d \) and \( k \in [K] \) with probability at least \( 1 - \delta \).

Lemma A2 (McDiarmid’s inequality). Let \( X = (X_1, X_2, \ldots, X_m) \) be \( m \)-independent random variables taking values from the set \( A \), and assume that \( f : A^m \to \mathbb{R} \) satisfies the following condition (bounded difference):

\[
\sup_{x_1, \ldots, x_m, \hat{x}_i} |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, \hat{x}_i, \ldots, x_m)| \leq c_i, \tag{A2}
\]

for all \( i \in 1, \ldots, m \). Then for any \( \epsilon > 0 \), we have

\[
P[f(X_1, \ldots, X_m) - \mathbb{E}[f(X_1, \ldots, X_m)] \geq \epsilon] \leq \exp \left( -\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2} \right). \tag{A3}
\]
Lemma A3 ([11]). Let \( S \in \mathbb{R}^{n \times s} \) be any uniform sampling matrix, then for any matrix \( B = [b_1, \ldots, b_n] \in \mathbb{R}^{d \times n} \) with probability \( 1 - \delta \), for any \( \delta > 0 \), we have
\[
\left\| \frac{1}{n} \text{BSS}^T 1 - \frac{1}{n} B 1 \right\| \leq \left( 1 + \sqrt{\frac{2 \log \left( \frac{1}{\delta} \right)}{\frac{1}{n}}} \right) \sqrt{\frac{1}{2}} \max \| b_i \|,  
\]
(A4)
where the vector \( B 1 \) is the sum of column of the matrix \( B \) and \( \text{BSS}^T 1 \) is the sum of uniformly sampled and scaled column of the matrix \( B \), where the scaling factor is \( \frac{1}{\sqrt{n}} \) with \( p = \frac{1}{n} \). If \((i_1, \ldots, i_s)\) is the set of sampled indices, the \( \text{BSS}^T 1 = \sum_{i \in (i_1, \ldots, i_s)} \frac{1}{\sqrt{n}} b_i \).

From Lemma A1, we can obtain a key corollary easily, which is taken to bound \( |(p_i^k)^T g(\tilde{w}_i) - (p_i^k)^T g^k(\tilde{w}_i)| \). And this bound is crucial for our main theorems.

Corollary A1. Let \( g^k(\tilde{w}_i) \) be the gradient of the loss function in the \( k \)-th worker, and \( g(\tilde{w}_i) \) be the gradient of the global loss function, where \( \tilde{w}_i \) is the \( t \)-th communication (\( L \geq 1 \)). Then, we obtain the following:

1. Provided that \( \| g_i(\tilde{w}_i) \| \leq \Gamma \), where \( \Gamma = \max_{i \in [n]} \| \nabla f_i(\cdot) \| \), then
\[
\| g^k(\tilde{w}_i) - g(\tilde{w}_i) \| \leq (1 + \sqrt{\frac{2 \log \left( \frac{1}{\delta} \right)}{\frac{1}{n}}} ) \sqrt{\frac{1}{2}} \Gamma.  
\]
(A5)
Writing, \( \eta = (1 + \sqrt{\frac{2 \log \left( \frac{1}{\delta} \right)}{\frac{1}{n}}} ) \sqrt{\frac{1}{2}} \Gamma \), we succinctly have
\[
\langle \nu, g^k(\tilde{w}_i) - g(\tilde{w}_i) \rangle \leq \| \nu \| \| g^k(\tilde{w}_i) - g(\tilde{w}_i) \| \leq \eta \| \nu \|  
\]
with probability at least \( 1 - \delta \), where \( \eta = O(1/\sqrt{s}) \) is small.

2. Provided that \( \| g_i(\tilde{w}_i) \| \leq \Gamma \), where \( \Gamma = \max_{i \in [n]} \| \nabla f_i(\cdot) \| \) and \( \| g^k(\tilde{w}_i) \| \geq G \), then using vector Bernstein inequality with \( t = \epsilon_1 \| g^k \| \), we obtain
\[
P \left( \| g^k(\tilde{w}_i) - g(\tilde{w}_i) \| \geq \epsilon_1 \| g^k(\tilde{w}_i) \| \right) \leq d \exp \left( -s \epsilon_1^2 \frac{G^2}{2s} + 1/4 \right) \leq d \exp \left( -s \frac{\epsilon_1^2 G^2}{32s} + 1/4 \right).  
\]
(A7)
So, as long as
\[
G^2 = \Omega \left( \frac{s^2 \log (d/\delta)}{\epsilon_1^2} \right)  
\]
or,
\[
s \gtrsim \left( \frac{r^2}{\epsilon_1^2 G^2} \log (d/\delta) \right),  
\]
we have
\[
\langle \nu, g^k(\tilde{w}_i) - g(\tilde{w}_i) \rangle \leq \| \nu \| \| g^k(\tilde{w}_i) - g(\tilde{w}_i) \| \leq \epsilon_1 \| \nu \| \| g^k(\tilde{w}_i) \|  
\]
with probability at least \( 1 - \delta \).

Appendix B. The Proofs of Some Lemmas in The Context

Appendix B.1. The Proof of Lemma 7

**Proof.** Let \( \tilde{U}_i^k = \frac{28 \sqrt{tt}}{p \delta} U_i^k \) and \( \tilde{Z}_i^k = \frac{2 \sqrt{tt}}{p \delta} Z_i^k \), and note that the Neumann series formula leads to the identity
\[
\left\{ H_i^k + \frac{28 \sqrt{tt}}{p \delta} U_i^k \right\}^{-1} = \left\{ H_i^k \right\}^{-1} \sum_{j=0}^{\infty} \left[ -\tilde{U}_i^k \left\{ H_i^k \right\}^{-1} \right]^j,  
\]
where we use the fact which satisfies the new line search condition

$$k \in C$$

Thus, we obtain

$$\lambda_k = (H_k)^{-1}g_k + \tilde{N}_k,$$

where

$$\tilde{N}_k = (H_k)^{-1} \left( \tilde{Z}_k + \sum_{j=1}^{\infty} \left( - \tilde{U}_j (H_k)^{-1} \right)^j \{ g_k + \tilde{Z}_j \} \right).$$

It remains to bound $\| \tilde{N}_k \|_2$ with high probability in order to complete the proof. Note that with probability at least $1 - \xi_0$, we have $\| Z_k \|_2 \leq 4d + 2\sqrt{2 \log(2/\xi_0)}$ and $\| \tilde{U}_k \|_2 \leq \sqrt{2d \log(4d/\xi_0)},$ $\| \tilde{N}_k \| = \| (H_k)^{-1} \| \| Z_k \| + \sum_{j=1}^{\infty} \left( (H_k)^{-1} \right)^j \| \{ g_k + \tilde{Z}_j \} \|$

$$\leq \frac{1}{\kappa(1-\epsilon)} \| Z_k \| + \frac{\| \tilde{U}_k \|}{\kappa(1-\epsilon)} \left( \| g_k \| + \| Z_k \| \right)$$

$$\leq \frac{\Gamma \sqrt{T(4d+2\sqrt{2\log(2/\xi_0)})}}{\kappa(1-\epsilon)} + \frac{2B \sqrt{TT} \kappa(1-\epsilon)}{\mu} \left( \Gamma + \frac{2B \sqrt{T(4d+2\sqrt{2\log(2/\xi_0)})}}{\mu} \right)$$

$$\leq C_0 \frac{\Gamma \sqrt{T(4\sqrt{d+2\log(2/\xi_0)})} + \mu B \sqrt{T \log(2/\xi_0)}}{\mu \kappa(1-\epsilon)^2} = C' \hat{M}_{\text{privacy}},$$

where $C_0$ and $C'$ are constants and $\hat{M}_{\text{privacy}} = \frac{\Gamma B \sqrt{T \log(2/\xi_0)}}{\mu \kappa(1-\epsilon)^2}$. Note that in order to make the second and third inequalities right, we need $\| \tilde{U}_k \| \leq \frac{\kappa(1-\epsilon)}{2}$. So let $s \geq \frac{4B \sqrt{T(2\log(4d/\xi_0))}}{\kappa(1-\epsilon)}$, $\| \tilde{U}_k \| \leq \frac{\kappa(1-\epsilon)}{2}$ hold with probability at least $1 - \xi_0/2$. □

**Appendix B.2. The Proof of Lemma 8**

**Proof.** Because the $f^k(\cdot)$ is $M(1+\epsilon)$-smooth, we have

$$f^k(w_k^0 - a\hat{p}_k^0) - f^k(w_k^0) \leq (-a\hat{p}_k^0)^T g_k^0 (w_k^0) + \frac{M(1+\epsilon)}{2} a^2 \| \hat{p}_k^0 \|^2.$$  

(A10)

The above inequality is satisfied for all $a \in \mathbb{R}$. We know that $a_k^0$, the local step-size at worker $k$, satisfies the line search constraint in Equation (10). Thus, there exists $a_k^* \in (0, 1)$, which satisfies the new line search condition

$$-a_k^0 \hat{p}_k^k T g_k^0 + \frac{M(1+\epsilon)}{2} a_k^0 \| \hat{p}_k^0 \|^2 \leq -a_k^0 \beta_k \hat{p}_k^k T g_k^0 + a_k^0 \gamma_k \hat{p}_k^k T g_k^0 + \tilde{N}_k^0.$$  

(A11)

Thus, a must satisfy

$$\frac{M(1+\epsilon)}{2} a \| \hat{p}_k^0 \|^2 \leq (1 - \beta_k) \hat{p}_k^k T g_k^0 + \gamma_k \hat{p}_k^k T g_k^0 - \tilde{N}_k^0$$

$$= (1 - \beta_k) \hat{p}_k^k T g_k^0 - \tilde{N}_k^0 + \gamma_k \hat{p}_k^k T g_k^0 - \tilde{N}_k^0$$

(A12)

where we use the fact $\gamma_k = 1 - \beta_k$ and $g_k^0 = H_k^0 (\hat{p}_k^k - \tilde{N}_k^0)$. Note that due to the $\kappa(1-\epsilon)$-strong convexity of $f^k(\cdot)$, the local line-search constraint must satisfy

$$a^* \leq \frac{2(1-\beta)\kappa(1-\epsilon)}{M(1+\epsilon)}.$$  

(A13)

Hence, if we choose $a^* \leq \frac{2(1-\beta)\kappa(1-\epsilon)}{M(1+\epsilon)}$ or $a^* \leq \frac{\kappa(1-\epsilon)}{M}$ for $\epsilon < 1/2$, we can guarantee that the computation steps sufficiently satisfying the line-search condition from Equation (10) are satisfied with $a_k^0 = a^*$. Hence, making use of the new line search condition, we obtain
\[ f^k(w^k_i - a_i^k p^k_i) - f^k(w^k_i) \leq -a^* \beta (p^k_i)^T g^k_i + a^* \gamma (p^k_i)^T H^k_i N^k_i \]
\[ = -a^* \beta (p^k_i + \bar{N}^k_i)^T g^k_i + a^* \gamma (p^k_i)^T H^k_i N^k_i \]
\[ = -a^* \beta (p^k_i)^T g^k_i - a^* \beta (\bar{N}^k_i)^T g^k_i + a^* \gamma (p^k_i)^T H^k_i N^k_i \]
\[ = -a^* \beta (g^k_i)^T (H^k_i)^{-1} g^k_i - a^* \beta (\bar{N}^k_i)^T H^k_i p^k_i + a^* \gamma (p^k_i)^T H^k_i N^k_i \]
\[ \leq -a^* \beta \frac{M(1+\epsilon)}{M(1+\epsilon)} \|g^k_i\|^2 + a^* (\gamma p^k_i - \beta p^k_i)^T H^k_i N^k_i \]
\[ = -a^* \beta \frac{M(1+\epsilon)}{M(1+\epsilon)} \|g^k_i\|^2 + a^* (\gamma p^k_i + \gamma \bar{N}^k_i - \beta p^k_i)^T H^k_i N^k_i \]
\[ = -a^* \beta \frac{M(1+\epsilon)}{M(1+\epsilon)} \|g^k_i\|^2 + a^* (\gamma - \beta) (p^k_i)^T H^k_i N^k_i + a^* \gamma (N^k_i)^T H^k_i N^k_i \]
\[ = -a^* \beta \frac{M(1+\epsilon)}{M(1+\epsilon)} \|g^k_i\|^2 + a^* (1 - \beta) (p^k_i)^T H^k_i N^k_i + a^* (1 - \beta) M(1 + \epsilon) \|N^k_i\|^2, \]

w.p. 1 - \delta. Because we have \( H^k_i \leq M(1+\epsilon)I \) and 1 - \beta > 0, we immediately obtain
\[ f^k(w^k_i - a_i^k p^k_i) - f^k(w^k_i) \leq -\psi \|g^k_i\|^2 + a^* (1 - \beta) (p^k_i)^T H^k_i N^k_i + a^* (1 - \beta) M(1 + \epsilon) \|N^k_i\|^2, \] (A14)
where \( \psi = \frac{a^* \beta}{M(1+\epsilon)}. \)

**Appendix C. The Proofs of Main Theorems**

**Appendix C.1. The Proof of Claim (14)**

**Proof.** We define some key symbols
\[ w^k_{i+1} = w_i - a_i^k p^k_i, \quad \text{and} \quad w_{i+1} = w_i - \frac{1}{K} \sum_{k=1}^{K} a_i^k p^k_i. \]

Invoking the M-smoothness of the function \( f(\cdot) \), we have
\[ f(w_i) - f(w_{i+1}) \geq -\frac{M}{2} \|w_i - w_{i+1}\|^2 + \langle g(w_i), w_i - w_{i+1} \rangle \]
\[ \geq -\frac{M}{2\epsilon^2} \| \sum_{k=1}^{K} a_i^k p^k_i \|^2 + \langle g(w_i), \frac{1}{K} \sum_{k=1}^{K} a_i^k p^k_i \rangle \]
\[ \geq \frac{1}{2} \sum_{k=1}^{K} (a_i^k)^2 \| p^k_i \|^2 + \langle g(w_i), \frac{1}{K} \sum_{k=1}^{K} a_i^k p^k_i \rangle \]
\[ = \frac{1}{K} \sum_{k=1}^{K} (a_i^k)^2 \langle p^k_i, g(w_i) - \frac{\epsilon}{K} (a_i^k)^2 \| p^k_i \|^2 \rangle, \] (A15)
where (i) is from
\[ \left\| \sum_{k=1}^{H} a^k \right\|^2 \leq \frac{1}{H} \sum_{k=1}^{H} \| a^k \|^2. \] (A16)

We make use of \( \|g^k_i\| \geq G \) and **Corollary A1**, combining with Lemma 8, then we have
\[ a_i^k (p^k_i)^T g(w_i) \geq f^k(w_i) - f^k(w_{i+1}) + \frac{(1-\epsilon)}{2} (a_i^k)^2 \| p^k_i \|^2 - \epsilon_1 \| p^k_i \| \| g^k_i \| \]
\[ \geq \psi \| g^k_i \|^2 - a_i^k (1 - 2\beta) (p^k_i)^T H^k_i N^k_i - a_i^k (1 - \beta) M(1 + \epsilon) \| N^k_i \|^2 \]
\[ + \frac{k-\epsilon}{2} (a_i^k)^2 \| p^k_i \|^2 - \epsilon_1 \| p^k_i \| \| g^k_i \| \]
\[ \geq \psi \| g^k_i \|^2 - a_i^k (1 - 2\beta) (p^k_i)^T H^k_i N^k_i - a_i^k (1 - \beta) M(1 + \epsilon) \| N^k_i \|^2 \]
\[ + \frac{k-\epsilon}{2} (a_i^k)^2 \| p^k_i \|^2 - \epsilon_1 \| p^k_i \| \| g^k_i \| - \epsilon_1 \| N^k_i \| \| g^k_i \| \]
\[ \geq \left( \psi - \frac{\epsilon_1}{k} (a_i^k)^2 \| p^k_i \|^2 - a_i^k (1 - 2\beta) (p^k_i)^T H^k_i N^k_i - \epsilon_1 \| N^k_i \| \| g^k_i \| \right) \]
\[ - a_i^k (1 - \beta) M(1 + \epsilon) \| N^k_i \|^2 + \frac{\epsilon}{2} (a_i^k)^2 \| p^k_i \|^2, \] (A17)
with probability at least $1 - 4\delta$, where the second inequality uses the fact $\|\hat{p}_k^T\| \leq \|p_k^T\| + \|\hat{N}_k^T\|$. And we let
\[
\varphi(\hat{N}_k^T) = \alpha_k^2 (1 - 2\beta)(p_k^T)^TH_k^T\hat{N}_k^T + \epsilon_1 \|\hat{N}_k^T\|\|g_k^T\| + \beta(1 - \beta)M(1 + \epsilon)\|\hat{N}_k^T\|^2. \tag{A18}
\]

So we have
\[
\alpha_k^2(p_k^T)^Tg(\hat{w}_i) - \frac{M}{2} (\alpha_k^2)^2\|p_k^T\|^2 \geq \left( \psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|g_k^T\|^2 + \frac{1}{\kappa(1-\epsilon)} (\alpha_k^2)^2\|p_k^T\|^2 - \frac{M}{2} (\alpha_k^2)^2\|p_k^T\|^2 - \varphi(\hat{N}_k^T)
\]
\[
\geq \left( \psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|g_k^T\|^2 - \frac{M-\kappa(1-\epsilon)}{2\kappa(1-\epsilon)} (\alpha_k^2)^2\|p_k^T\|^2 - \varphi(\hat{N}_k^T)
\]
\[
\geq \left( \psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|g_k^T\|^2 - \frac{M-\kappa(1-\epsilon)}{\kappa^2(1-\epsilon)^2} (\alpha_k^2)^2\|p_k^T\|^2 - \varphi(\hat{N}_k^T)
\]
\[
\geq \left( \psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|g_k^T\|^2 - [M-\kappa(1-\epsilon)](\alpha^*)^2\|\hat{N}_k^T\|^2 - \varphi(\hat{N}_k^T),
\]
with probability at least $1 - 4\delta$, where the third inequality is due to $\|\hat{p}_k^T\|^2 \leq 2\|p_k^T\|^2 + 2\|\hat{N}_k^T\|^2$. Rearranging the terms about $\|\hat{N}_k^T\|$, we have
\[
\hat{\varphi}(\hat{N}_k^T; \beta) = \alpha_k^2 (1 - 2\beta)(p_k^T)^T\hat{H}_k^T\hat{N}_k^T + \epsilon_1 \|\hat{N}_k^T\|\|g_k^T\| + \{ (1 - \beta)M(1 + \epsilon) + [M-\kappa(1-\epsilon)]\alpha^* \}(\alpha^*) \|\hat{N}_k^T\|^2. \tag{A20}
\]

Hence, we obtain
\[
\alpha_k^2(p_k^T)^Tg(\hat{w}_i) - \frac{M}{2} (\alpha_k^2)^2\|p_k^T\|^2 \geq \left( \psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} - \frac{M-\kappa(1-\epsilon)}{\kappa^2(1-\epsilon)^2} (\alpha^*)^2 \right)\|g_k^T\|^2 - \varphi(\hat{N}_k^T; \beta), \tag{A21}
\]
with probability at least $1 - 6\delta$, and from the inequality, we obtain
\[
f(\hat{w}_i) - f(\hat{w}_{i+1}) \geq \frac{1}{K} \sum_{k=1}^{K} \left( \psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} - \frac{M-\kappa(1-\epsilon)}{\kappa^2(1-\epsilon)^2} (\alpha^*)^2 \right)\|g_k^T\|^2 - \frac{1}{K} \sum_{k=1}^{K} \hat{\varphi}(\hat{N}_k^T; \beta), \tag{A22}
\]
with probability at least $1 - 6K\delta$.

If we choose $\beta = 1/2$, we will eliminate the $(p_k^T)^T\hat{H}_k^T\hat{N}_k^T$ and obtain a more concise result as following
\[
f(\hat{w}_i) - f(\hat{w}_{i+1}) \geq \frac{1}{K} \sum_{k=1}^{K} C_1 \|g_k^T\|^2 - \frac{1}{K} \sum_{k=1}^{K} \hat{\varphi}(\hat{N}_k^T; 1/2), \tag{A23}
\]
where $C_1 = \left( \psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} - \frac{M-\kappa(1-\epsilon)}{\kappa^2(1-\epsilon)^2} (\alpha^*)^2 \right)$ and $\hat{\varphi}(\hat{N}_k^T; 1/2) = \epsilon_1 \|\hat{N}_k^T\|\|g_k^T\| + \{ \frac{1}{2}M(1 + \epsilon) + [M-\kappa(1-\epsilon)]\alpha^* \}(\alpha^*) \|\hat{N}_k^T\|^2$.

Making use of inequality (A16), we obtain
\[
\frac{1}{K} \sum_{k=1}^{K} C_1 \|g_k^T\|^2 \geq \frac{C_1}{2} \|g(\hat{w}_i)\|^2, \tag{A24}
\]
then, combining inequality (A23), we obtain
\[
f(\hat{w}_i) - f(\hat{w}_{i+1}) \geq C_1 \|g(\hat{w}_i)\|^2 - \frac{1}{K} \sum_{k=1}^{K} \hat{\varphi}(\hat{N}_k^T; 1/2). \tag{A25}
\]

Invoking the $\kappa$-strong convexity of the function $f$
\[
f(\hat{w}_i) - f(\hat{w}^*) \leq \frac{\kappa}{2} \|g(\hat{w}_i)\|^2, \tag{A26}
\]
and combining the inequality (A25) and (A26), we obtain
\[
f(\hat{w}_{i+1}) - f(\hat{w}^*) \leq (1 - 2\kappa C_1)(f(\hat{w}_i) - f(\hat{w}^*)) + \frac{1}{K} \sum_{k=1}^{K} \hat{\varphi}(\hat{N}_k^T; 1/2), \tag{A27}
\]
with probability at least $1 - 6K\delta$. If we use Lemma 7, we have
\[ \phi(\hat{N}_k^k; 1/2) \leq c_1 \hat{M}_{\text{privacy}} \Gamma + \{ \frac{1}{2} M(1 + \epsilon) + [M - \kappa(1 - \epsilon)] \alpha^* \} \hat{M}_{\text{privacy}}^2 = \hat{M}_\phi, \]

with probability at least 1 - \( \xi_0 \). Hence, we have

\[ f(\hat{w}_{t+1}) - f(\hat{w}^*) \leq (1 - 2\kappa C_1)(f(\hat{w}_t) - f(\hat{w}^*)) + \hat{M}_\phi, \quad (A28) \]

with probability at least 1 - \( K(6\delta + \xi_0) \). \( \square \)

**Appendix C.2. The Proof of Claim (15)**

**Proof.** From the uniform subsampling property and the case 1 of Corollary A1, we obtain

\[ |(\hat{p}_t^k)^T g(\hat{w}_t) - (\hat{p}_t^k)^T g^*(\hat{w}_t)| \leq \eta \| \hat{p}_t^k \| \text{ w.p. } 1 - \delta. \quad (A29) \]

Thus,

\[ (\hat{p}_t^k)^T g(\hat{w}_t) \geq (\hat{p}_t^k)^T g^*(\hat{w}_t) - \eta \| \hat{p}_t^k \| \text{ w.p. } 1 - \delta. \quad (A30) \]

Now, since the function \( f^k \) is \( \kappa(1 - \epsilon) \) strongly convex with probability \( 1 - \delta \), we have the following bound with probability of at least \( 1 - \delta \):

\[ a_k^k(\hat{p}_t^k)^T g^*(\hat{w}_t) \geq (f^k(\hat{w}_t) - f^k(\hat{w}_{t+1})) + \frac{\kappa(1-\epsilon)}{2} \| \hat{p}_t^k \|^2. \quad (A31) \]

Combining the equations, and similar to the previous analysis, we have

\[ a_k^k(\hat{p}_t^k)^T g(\hat{w}_t) \geq (f^k(\hat{w}_t) - f^k(\hat{w}_{t+1})) + \frac{\kappa(1-\epsilon)}{2} \| \hat{p}_t^k \|^2 - \eta \| \hat{p}_t^k \| \]

\[ \geq \left( \psi - \frac{\| \hat{M}_{\text{privacy}} \| \| \hat{N}_k^k \|}{\kappa(1-\epsilon)} \right) \| g^* \|^2 - \phi(\hat{N}_k^k; \beta) - \frac{\eta f}{\kappa(1-\epsilon)}, \quad (A32) \]

with probability at least \( 1 - \delta \), where \( \phi(\hat{N}_k^k; \beta) = a_k^k(1 - 2\beta)(\hat{p}_t^k)^T H_k^k \hat{N}_k^k + \eta \| \hat{N}_k^k \| + \{(1 - \beta)M(1 + \epsilon) + [M - \kappa(1 - \epsilon)] \alpha^* \} \| \hat{N}_k^k \|^2 \). So take Equation (A32) into Equation (A15) and obtain

\[ f(\hat{w}_t) - f(\hat{w}_{t+1}) \geq \frac{1}{K} \sum_{k=1}^K C_2 \| g_t^k \|^2 - \frac{1}{K} \sum_{k=1}^K \phi(\hat{N}_k^k; \beta) - \frac{\eta f}{\kappa(1-\epsilon)}, \quad (A33) \]

with probability at least \( 1 - 6K\delta \).

Then, let \( \beta = 1/2 \) and making a similar analysis like inequality (A25) and (A26), we obtain

\[ f(\hat{w}_{t+1}) - f(\hat{w}^*) \leq (1 - 2\kappa C_2)(f(\hat{w}_t) - f(\hat{w}^*)) + \frac{1}{K} \sum_{k=1}^K \phi(\hat{N}_k^k; 1/2) + \frac{\eta f}{\kappa(1-\epsilon)}, \quad (A34) \]

with probability at least \( 1 - 6K\delta \), where \( C_2 = \left( \psi - \frac{\| \hat{M}_{\text{privacy}} \| \| \hat{N}_k^k \|}{\kappa(1-\epsilon)} \right) \| g^* \|^2 \) and \( \phi(\hat{N}_k^k; 1/2) = \eta \| \hat{N}_k^k \| + \{(1 - \beta)M(1 + \epsilon) + [M - \kappa(1 - \epsilon)] \alpha^* \} \| \hat{N}_k^k \|^2 \). If we use Lemma 7, we have

\[ \phi(\hat{N}_k^k; 1/2) \leq \eta \hat{M}_{\text{privacy}} + \{ \frac{1}{2} M(1 + \epsilon) + [M - \kappa(1 - \epsilon)] \alpha^* \} \| \hat{M}_{\text{privacy}} \|^2 = \hat{M}_{\phi}, \]

with probability at least \( 1 - \xi_0 \). Then, we obtain the claim

\[ f(\hat{w}_{t+1}) - f(\hat{w}^*) \leq (1 - 2\kappa C_2)(f(\hat{w}_t) - f(\hat{w}^*)) + \hat{M}_\phi + \frac{\eta f}{\kappa(1-\epsilon)}, \quad (A35) \]

with probability at least \( 1 - K(6\delta + \xi_0) \). \( \square \)

**Appendix C.3. The Proof of Theorem 3**

**Proof.** We define some symbols as the following:

\[ \hat{w}_{t_0+1} = \hat{w}_{t_0} - \sum_{t=t_0}^L \hat{p}_t, \quad (A36) \]
where $\hat{\mathbf{p}}_\tau = \frac{1}{L} \sum_{k=1}^{K} a_k^k \hat{\mathbf{p}}_\tau^k$ is the average descent direction and $\hat{\mathbf{p}}_\tau^k = (\hat{\mathbf{H}}_\tau^k)^{-1} \hat{\mathbf{g}}_\tau$ is the local disturbed descent direction at the $k$-th worker at the iteration $\tau$. $L$ is the local computing times.

Invoking the $M$ smoothness of the function $f$, we have

\[
f(\hat{\mathbf{w}}_{0}) - f(\hat{\mathbf{w}}_{0+1}) \geq -\frac{M}{\kappa} \left\| \sum_{\tau=L_0}^{L} \hat{\mathbf{p}}_{\tau} \right\|^2 + \left\langle \mathbf{g}(\hat{\mathbf{w}}_{0}), \sum_{\tau=L_0}^{L} \hat{\mathbf{p}}_{\tau} \right\rangle
\]

\[
= -\frac{M}{\kappa} \left\| \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} a_k^k \hat{\mathbf{p}}_{\tau} \right\|^2 + \frac{1}{\kappa} \sum_{k=1}^{K} \left\{ \mathbf{g}(\hat{\mathbf{w}}_{0}), a_k^k \hat{\mathbf{p}}_{\tau} \right\}
\]

\[
\geq -\frac{M}{\kappa^2} \sum_{k=1}^{K} \left\| a_k^k \hat{\mathbf{p}}_{\tau} \right\|^2 + \frac{1}{\kappa} \sum_{k=1}^{K} \left\{ \mathbf{g}(\hat{\mathbf{w}}_{0}), a_k^k \hat{\mathbf{p}}_{\tau} \right\},
\]

(A37)

we make use of inequality (A16) in the last inequality.

Invoking the $\kappa(1 - \epsilon)$ strong convexity of function $f^k(\cdot)$, we have

\[
f^k(\mathbf{w}_{0}^k) - f^k(\mathbf{w}_{0+1}^k) \leq \frac{-\kappa(1 - \epsilon)}{2} \left\| \sum_{\tau=L_0}^{L} a_k^k \hat{\mathbf{p}}_{\tau} \right\|^2 + \left\langle \mathbf{g}_{0}^k, \sum_{\tau=L_0}^{L} a_k^k \hat{\mathbf{p}}_{\tau} \right\rangle
\]

(A38)

with probability $1 - \delta$.

At every communication round, we have

\[
\frac{1}{\kappa} \sum_{k=1}^{K} \left( f^k(\mathbf{w}_{0}^k) - f^k(\mathbf{w}_{0+1}^k) \right) \leq \frac{-\kappa(1 - \epsilon)}{2\kappa} \sum_{k=1}^{K} \left\| \sum_{\tau=L_0}^{L} a_k^k \hat{\mathbf{p}}_{\tau} \right\|^2 + \frac{1}{\kappa} \sum_{k=1}^{K} \left\{ \mathbf{g}_{0}^k, a_k^k \hat{\mathbf{p}}_{\tau} \right\}.
\]

(A39)

Then, combining the inequality (A37) and (A39) and using the inequality (A30), we eliminate the terms $\frac{1}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \left\{ \mathbf{g}(\hat{\mathbf{w}}_{0}), a_k^k \hat{\mathbf{p}}_{\tau} \right\}$ and $\frac{1}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \left\{ \mathbf{g}_{0}^k, a_k^k \hat{\mathbf{p}}_{\tau} \right\}$. So, we have

\[
f(\hat{\mathbf{w}}_{0}) - f(\hat{\mathbf{w}}_{0+1}) \geq \frac{1}{\kappa} \sum_{k=1}^{K} \left( f^k(\mathbf{w}_{0}^k) - f^k(\mathbf{w}_{0+1}^k) \right) - \frac{M - \kappa(1 - \epsilon)}{2\kappa} \sum_{k=1}^{K} \left\| a_k^k \hat{\mathbf{p}}_{\tau} \right\|^2 - \frac{1}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \eta a_k^k \left\| \hat{\mathbf{p}}_\tau \right\|.
\]

(A40)

Making use of inequality (A16), we obtain

\[
f(\hat{\mathbf{w}}_{0}) - f(\hat{\mathbf{w}}_{0+1}) \geq \frac{1}{\kappa} \sum_{k=1}^{K} \left( f^k(\mathbf{w}_{0}^k) - f^k(\mathbf{w}_{0+1}^k) \right) - \frac{L(M - \kappa(1 - \epsilon))}{2\kappa} \sum_{k=1}^{K} \left( \sum_{\tau=L_0}^{L} a_k^k \right) \left\| \hat{\mathbf{p}}_{\tau} \right\|^2
\]

\[
- \frac{1}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \eta a_k^k \left\| \hat{\mathbf{p}}_\tau \right\|.
\]

(A41)

Then, from Lemma 8, we obtain, with probability at least $1 - KL\delta$,

\[
f(\hat{\mathbf{w}}_{0}) - f(\hat{\mathbf{w}}_{0+1}) \geq \frac{\kappa}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \left\| \hat{\mathbf{p}}_{\tau} \right\|^2 - \frac{1}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} a^k M(1 + \epsilon) \left\| \tilde{N}_{\tau}^k \right\|^2
\]

\[
- \frac{L(M - \kappa(1 - \epsilon))}{2\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \left( a^k \right)^2 \left( \left\| \hat{\mathbf{p}}_{\tau} \right\|^2 + \left\| \tilde{N}_{\tau}^k \right\|^2 \right) - \frac{1}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \eta \left( \left\| \hat{\mathbf{p}}_{\tau} \right\| + \left\| \tilde{N}_{\tau}^k \right\| \right)
\]

\[
\geq \frac{\kappa}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \left\| \hat{\mathbf{p}}_{\tau} \right\|^2 - \frac{\eta LF}{\kappa(1 - \epsilon)} - \frac{1}{\kappa} \sum_{k=1}^{K} \sum_{\tau=L_0}^{L} \varphi(\tilde{N}_{\tau}^k; 1/2),
\]

(A42)

where $C = \psi - \frac{L(M - \kappa(1 - \epsilon))}{2\kappa(1 - \epsilon)^2}$ and $\varphi(\tilde{N}_{\tau}^k; 1/2) = \eta \left\| \tilde{N}_{\tau}^k \right\| + \left\{ \frac{1}{2} M(1 + \epsilon) + [M - \kappa(1 - \epsilon)] a^* \right\} \left\| \tilde{N}_{\tau}^k \right\|^2$. 

If using Lemma 7, we have

\[
\varphi(\tilde{N}^k_{\tau}; 1/2) \leq \eta \tilde{M}_{\text{privacy}} + \{ \frac{1}{2} M(1 + \epsilon) + [M - \kappa(1 - \epsilon)] a^* \} \alpha^* \tilde{M}^2_{\text{privacy}} = \tilde{M}_\varphi,
\]

with probability at least \(1 - \xi_0\). Then we obtain Theorem 3

\[
f(\bar{w}_t) - f(\bar{w}_{t+1}) \geq C \sum_{k=1}^{K} \sum_{\tau=t_0}^{\tau} \|g^k_{\tau}\|^2 - \frac{\eta L \Gamma}{\kappa(1 - \epsilon)} - L \tilde{M}_\varphi,
\]

(A43)

with probability at least \(1 - KL(6\delta + \xi_0)\).

\[\square\]

Appendix D. The FL Gradient Descent Method with \(\mu\)-GDP

Appendix D.1. Algorithm

Algorithm A1 GDP-GD

1. **Input:** Initial iteration \(\bar{w}_0 \in \mathbb{R}^d\); Privacy parameter \(\mu\); Loss function parameter \(\gamma\); Iteration parameter \(T\); Step size \(\alpha\).
2. **for** \(t = 0\) to \(T\) **do**
3. **Initialization:** \(w^k_t = \bar{w}_t\)
4. **for** \(k = 1\) to \(K\) in parallel **do**
5. \(\tilde{g}^k_t = \epsilon^k_{\tau} + \frac{\epsilon T}{\mu s} z^k_{\tau}^i\)
6. **Send to the server:** \(w^k_{t+1} = w^k_t - \alpha \tilde{g}^k_t\)
7. **end for**
8. **The server updates the parameter:** \(\bar{w}_{t+1} = \sum_{k=1}^{K} w^k_{t+1}\)
9. **end for**

Appendix D.2. Theory

**Lemma A4** ([26]). After \(T\) communication rounds, the output of any user satisfies \(\mu\)-GDP.

**Theorem A1.** After \(T\) communication rounds, the GDP-GD algorithm satisfies \(\mu\)-GDP.

**Proof.** The result is from Lemmas 3 and A4. \(\square\)

**References**


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.