Stochastic-Sign SGD for Federated Learning with Theoretical Guarantees

Richeng Jin[*] **Yufan Huang**[†] **Xiaofan He**[‡] **Tianfu Wu**[§] **Huaiyu Dai**[¶]

Abstract

Federated learning (FL) has emerged as a prominent distributed learning paradigm. FL entails some pressing needs for developing novel parameter estimation approaches with theoretical guarantees of convergence, which are also communication efficient, differentially private and Byzantine resilient in the heterogeneous data distribution settings. Quantization-based SGD solvers have been widely adopted in FL and the recently proposed signSGD with majority vote shows a promising direction. However, no existing methods enjoy all the aforementioned properties. In this paper, we propose an intuitively-simple yet theoretically-sound method based on signSGD to bridge the gap. We present Stochastic-Sign SGD which utilizes novel stochastic-sign based gradient compressors enabling the aforementioned properties in a unified framework. We also present an error-feedback variant of the proposed Stochastic-Sign SGD which further improves the learning performance in FL. We test the proposed method with extensive experiments using deep neural networks on the MNIST dataset. The experimental results corroborate the effectiveness of the proposed method.

1 Introduction

Recently, Federated Learning (FL) has become a prominent distributed learning paradigm since it allows training on a large amount of decentralized data residing on devices like mobile phones [1]. However, FL imposes several critical challenges. First of all, the communication capability of the mobile devices can be a significant bottleneck. Furthermore, the training data on a given worker is typically based on its usage of the mobile devices, which results in heterogeneous data distribution. In addition, the local data usually contains some sensitive information of a particular mobile device user. Therefore, there is an pressing need to develop a privacy-preserving distributed learning algorithm. Finally, similar to many distributed learning methods, FL may suffer from malicious participants. As is shown in [2], even a single Byzantine worker, which may transmit arbitrary information, can severely disrupt the convergence of distributed gradient descent algorithms. However, to the best of our knowledge, no existing methods can cope with all the aforementioned challenges. To alleviate the communication burden of the workers, there have been various gradient quantization methods [3–7] in the literature, among which the recently proposed signSGD with majority vote [8] is of particular interest due to its robustness and communication efficiency.[1] In signSGD, during each communication round, only the signs of the gradients and aggregation results are exchanged between the workers and the server, which leads to around $32\times$ less communication than full-precision

[*]North Carolina State University, Email: rjin2@ncsu.edu.

[†]North Carolina State University, Email: yhuang20@ncsu.edu.

[‡]Wuhan University, Email: xiaofanhe@whu.edu.cn.

[§]North Carolina State University, Email: tianfu_wu@ncsu.edu.

[¶]Corresponding author; North Carolina State University, Email: hdai@ncsu.edu.

[1]Note that all the algorithms considered in this work use the idea of majority vote. Therefore, we ignore the term "with majority vote" in the following discussions for the ease of presentation.

Preprint. Under review.

distributed SGD. Nonetheless, it has been shown in [9] that SIGNSGD fails to converge when the data on different workers are heterogeneous (i.e., drawn from different distributions), which is one of the most important features in FL.

In this work, inspired by the idea of adding carefully designed noise before taking the sign operation in [9], we present Stochastic-Sign SGD, which is a class of stochastic-sign based SGD algorithms. In particular, we first propose a stochastic compressor $sto\text{-}sign$, which extends SIGNSGD to its stochastic version Sto-SIGNSGD. In this scheme, instead of directly transmitting the signs of gradients, the workers adopt a two-level stochastic quantization and transmit the signs of the quantized results. We note that different from the existing 1-bit stochastic quantization schemes (e.g., QSGD [3], cpSGD [7]), the proposed algorithm also uses the majority vote rule in gradient aggregation, which allows the server-to-worker communication to be 1-bit compressed and ensures robustness as well. Then, to further resolve the privacy concerns, a differentially private stochastic compressor $dp\text{-}sign$ is proposed, which can accommodate the requirement of (ϵ, δ)-local differential privacy [10]. The corresponding algorithm is termed as DP-SIGNSGD. We then prove that when the number of workers is large enough, both of the proposed algorithms converge with a rate of $O(\frac{\sqrt{d}}{\sqrt{T}})$ under heterogeneous data distribution, where d is the dimension of the hypothesis vector and T is the total number of training iterations. In addition, assuming that there are M normal (benign) workers, it is shown that the Byzantine resilience of the proposed algorithms is upper bounded by $|\sum_{m=1}^{M}(\boldsymbol{g}_m^{(t)})_i|/b_i, \forall i$, where $(\boldsymbol{g}_m^{(t)})_i$ is the i-th entry of worker m's gradient at iteration t and $b_i \geq \max_m(\boldsymbol{g}_m^{(t)})_i$ is some design parameter. Particularly, b_i depends on the data heterogeneity (through $\max_m(\boldsymbol{g}_m^{(t)})_i$). As a special case, the proposed algorithms can tolerate $M-1$ Byzantine workers when the normal workers can access the same dataset (i.e., $(\boldsymbol{g}_m^{(t)})_i = (\boldsymbol{g}_j^{(t)})_i, \forall 1 \leq j, m \leq M$), which recovers the result of SIGNSGD.

We also extend the proposed algorithm to its error-feedback variant, termed as Error-Feedback Stochastic-Sign SGD. In this scheme, the server keeps track of the error induced by the majority vote operation and compensates for the error in the next communication round. Both the convergence and the Byzantine resilience are established. Extensive simulations are performed to demonstrate the effectiveness of all the proposed algorithms.

2 Related Works

Gradient Quantization: To accommodate the need of communication efficiency in distributed learning, various gradient compression methods have been proposed. Most of the existing works focus on unbiased methods [11, 12]. QSGD [3], TernGrad [4] and ATOMO [13] propose to use stochastic quantization schemes, based on which a differentially private variant is proposed in [7]. Due to the unbiased nature of such quantization methods, the convergence of the corresponding algorithms can be established.

The idea of sharing the signs of gradients in SGD can be traced back to 1-bit SGD [14]. [15] and [5,8] show theoretical and empirical evidence that sign based gradient schemes can converge well despite the biased approximation nature in the homogeneous data distribution scenario. In the heterogeneous data distribution case, [9] shows that the convergence of SIGNSGD is not guaranteed and proposes to add carefully designed noise to ensure a convergence rate of $O(d^{\frac{3}{4}}/T^{\frac{1}{4}})$. However, their analysis assumes second order differentiability of the noise probability density function and cannot be applied to some commonly used noise distributions (e.g., uniform and Laplace distributions). In addition, their analysis requires that the variance of the noise goes to infinity as the number of iterations grows, which may be unrealistic in practice.

Error-Compensated SGD: Instead of directly using the biased approximation of the gradients, [14] corrects the quantization error by adding error feedback in subsequent updates and observes almost no accuracy loss empirically. [6] proposes the error-compensated quantized SGD in quadratic optimization and proves its convergence for unbiased stochastic quantization. [16] proves the convergence of the proposed error compensated algorithm for strongly-convex loss functions and [17] proves the convergence of sparsified gradient methods with error compensation for both convex and non-convex loss functions. In addition, [18] proposes EF-SIGNSGD, which combines the error compensation methods and SIGNSGD; however, only the single worker scenario is considered. [19] further extends it to the multi-worker scenario and the convergence is established. However, it is required in these two works that the compressing error cannot be larger than the magnitude of the original vector, which is not the case for some biased compressors like SIGNSGD. [20] considers more general compressors and proves the convergence under the assumption that the compressors have bounded magnitude of error. However, to the best our knowledge, none of the existing works consider the Byzantine resilience of the error-compensated methods.

Byzantine Tolerant SGD in Heterogenous Environment: There have been significant research interests in developing SGD based Byzantine tolerant algorithms, most of which consider homogeneous data distribution, e.g., Krum [21], ByzantineSGD [22], and the median based algorithms [23]. [8] shows that SIGNSGD can tolerate up to half "blind" Byzantine workers who determine how to manipulate their gradients before observing the gradients.

To accommodate the need for robust FL, some Byzantine tolerant algorithms that can deal with heterogeneous data distribution have been developed. [24] proposes to incorporate a regularized term with the objective function. However, it requires strong convexity and can only converge to the neighborhood of the optimal solution. [25] uses trimmed mean to aggregate the shared parameters. Nonetheless, it can only tolerate a small (unknown) number of Byzantine workers. In addition, both [24] and [25] assume model aggregation, i.e., both the workers and the parameter server share their models with others in full precision, which may incur significant communication cost.

Our Contributions. This paper makes three main contributions to the field of FL as follows.

1. We propose the framework of Stochastic-Sign SGD, which utilizes the stochastic-sign based gradient compressors to overcome the convergence issue of SIGNSGD in the presence of heterogeneous data distribution. In particular, two novel stochastic compressors, $sto\text{-}sign$ and $dp\text{-}sign$, are proposed, which extend SIGNSGD to Sto-SIGNSGD and DP-SIGNSGD, respectively. DP-SIGNSGD is shown to improve the privacy and the accuracy simultaneously, without sacrificing any communication efficiency. We further improve the learning performance of the proposed algorithm by incorporating the error-feedback method.
2. We prove that all the algorithms converge with a rate of $O(\frac{\sqrt{d}}{\sqrt{T}})$ in the heterogeneous data distribution scenario, which matches SIGNSGD in the homogeneous data distribution scenario.
3. We theoretically show the Byzantine resilience of the proposed algorithms, which depends on the heterogeneity of the local datasets of the workers.

3 Problem Formulation

In this paper, we consider a typical federated optimization problem with M normal workers as in [1]. Formally, the goal is to minimize the finite-sum objective of the form

$$\min_{w\in\mathbb{R}^d} F(w) \quad \text{where} \quad F(w) \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} f_m(w). \tag{1}$$

For a machine learning problem, we have a sample space $I = X \times Y$, where X is a space of feature vectors and Y is a label space. Given the hypothesis space $\mathcal{W} \subseteq \mathbb{R}^d$, we define a loss function $l : \mathcal{W} \times I \to \mathbb{R}$ which measures the loss of prediction on the data point $(x, y) \in I$ made with the hypothesis vector $w \in \mathcal{W}$. In such a case, $f_m(w)$ is a local function defined by the local dataset of worker m and the hypothesis w. More specifically,

$$f_m(w) = \frac{1}{|D_m|}\sum_{(x_n,y_n)\in D_m} l(w; (x_n, y_n)), \tag{2}$$

where $|D_m|$ is the size of worker m's local dataset D_m.

In many FL applications, the local datasets of the workers are heterogeneously distributed. In this case, we have $D_m \neq D_j$ and therefore $\nabla f_m(w) \neq \nabla f_j(w), \forall m \neq j$.

We consider a parameter server paradigm. At each communication round t, each worker m forms a batch of training samples, based on which it computes and transmits the stochastic gradient $\boldsymbol{g}_m^{(t)}$ as an estimate to the true gradient $\nabla f_m(w_m^{(t)})$. When the worker m evaluates the gradient over its whole local dataset, we have $\boldsymbol{g}_m^{(t)} = \nabla f_m(w_m^{(t)})$. After receiving the gradients from the workers, the server performs aggregation and sends the aggregated gradient back to the workers. Finally, the workers update their local model weights using the aggregated gradient. In this sense, the classic stochastic gradient descent (SGD) algorithm [26] performs iterations of the form

$$w_m^{(t+1)} = w_m^{(t)} - \frac{\eta}{M}\sum_{m=1}^{M} \boldsymbol{g}_m^{(t)}. \tag{3}$$

In this case, since all the workers adopt the same update rule using the aggregated gradient, $w_m^{(t)}$'s are the same for all the workers. Therefore, in the following discussions, we omit the worker index m for the ease of presentation. To accommodate the requirement of communication efficiency in FL, we adopt the popular idea of gradient quantization and assume that each worker m quantizes the gradient with a stochastic 1-bit compressor $q(\cdot)$ and sends $q(\boldsymbol{g}_m^{(t)})$ instead of its actual local gradient $\boldsymbol{g}_m^{(t)}$. Combined with the idea of majority vote in [5], the corresponding algorithm is presented in Algorithm 1.

Algorithm 1 Stochastic-Sign SGD with majority vote

Input: learning rate η, current hypothesis vector $w^{(t)}$, M workers each with an independent gradient $\boldsymbol{g}_m^{(t)}$, the 1-bit compressor $q(\cdot)$.
on server:
pull $q(\boldsymbol{g}_m^{(t)})$ **from** worker m.
push $\tilde{\boldsymbol{g}}^{(t)} = sign\big(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})\big)$ **to** all the workers.
on each worker:
update $w^{(t+1)} = w^{(t)} - \eta\tilde{\boldsymbol{g}}^{(t)}$.

Intuitively, the performance of Algorithm 1 is limited by the probability of wrong aggregation, which is given by

$$sign\bigg(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^{(t)})\bigg). \tag{4}$$

In SIGNSGD, $q(\boldsymbol{g}_m^{(t)}) = sign(\boldsymbol{g}_m^{(t)})$ and (4) holds when $\nabla f_m(w^{(t)}) \neq \nabla f_j(w^{(t)}), \forall m \neq j$ with a high probability, which prevents its convergence. In this work, we propose two compressors $sto\text{-}sign$ and $dp\text{-}sign$, which guarantee that (4) holds with a probability that is strictly smaller than 0.5 and therefore the convergence of Algorithm 1 follows. Moreover, $dp\text{-}sign$ is differentially private, i.e., given the quantized gradient $q(\boldsymbol{g}_m^{(t)})$, the adversary cannot distinguish the local dataset of worker m from its neighboring datasets that differ in only one data point with a high probability. The detailed definition of differential privacy can be found in Section 1 of the supplementary document.

In addition to the M normal workers, it is assumed that there exist B Byzantine attackers, and its set is denoted as $\mathcal{B}$. Instead of using $sto\text{-}sign$ and $dp\text{-}sign$, the Byzantine attackers can use an arbitrary compressor denoted by $byzantine\text{-}sign$. In this work, we consider the scenario that the Byzantine attackers have access to the average gradients of all the M normal workers (i.e., $\boldsymbol{g}_j^{(t)} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{g}_m^{(t)}$, $\forall j \in \mathcal{B}$) and follow the same procedure as the normal workers. Therefore, we assume that the Byzantine attacker j shares the opposite signs of the true gradients, i.e., $byzantine\text{-}sign(\boldsymbol{g}_j^{(t)}) = -sign(\boldsymbol{g}_j^{(t)})$.

In order to facilitate the convergence analysis, the following commonly adopted assumptions are made.

Assumption 1. *(Lower bound). For all w and some constant F^*, we have objective value $F(w) \geq F^*$.*

Assumption 2. *(Smoothness). $\forall w_1, w_2$, we require for some non-negative constant L*

$$F(w_1) \leq F(w_2) + < \nabla F(w_2), w_1 - w_2 > + \frac{L}{2}||w_1 - w_2||_2^2, \tag{5}$$

where $< \cdot, \cdot >$ is the standard inner product.

Assumption 3. *(Variance bound). For any worker m, the stochastic gradient oracle gives an independent unbiased estimate g_m that has coordinate bounded variance:*

$$\mathbb{E}[g_m] = \nabla f_m(w), \mathbb{E}[((g_m)_i - \nabla f_m(w)_i)^2] \leq \sigma_i^2, \tag{6}$$

for a vector of non-negative constants $\bar{\sigma} = [\sigma_1, \cdots, \sigma_d]$; $(g_m)_i$ and $\nabla f_m(w)_i$ are the i-th coordinate of the stochastic and the true gradient, respectively.

Assumption 4. *The total number of workers is odd.*

We note that Assumptions 1, 2 and 3 are standard for non-convex optimization and Assumption 4 is just to ensure that there is always a winner in the majority vote [9], which can be easily relaxed.

Experimental Settings. To facilitate empirical discussions on our proposed algorithms in the remaining sections, we first introduce our experimental settings here. We implement our proposed method with a two-layer fully connected neural network on the standard MNIST dataset [27]. We consider a scenario of $M = 31$ normal workers. To simulate the heterogeneous data distribution scenario, each worker only stores exclusive data for one out of the ten categories, unless otherwise noted. We use a constant learning rate and tune the parameter from the set $\{1, 0.1, 0.01, 0.005, 0.001, 0.0001\}$. We compare the proposed algorithms with two baselines: SIGNSGD [5] and FedAvg [28]. More details about the implementation can be found in Section 5 of the supplementary document.

4 Algorithms and Convergence Analysis

In this section, we propose two compressors $sto\text{-}sign$ and $dp\text{-}sign$ for the Stochastic-Sign SGD framework, which can deal with the heterogeneous data distribution scenario. The basic ideas of the two compressors are given as follows.

- $sto\text{-}sign$: instead of directly sharing the signs of the gradients, $sto\text{-}sign$ first performs a two-level stochastic quantization and then transmits the signs of the quantized results.
- $dp\text{-}sign$: it is a differentially private version of $sto\text{-}sign$. The probability of each coordinate of the gradients mapping to $\{-1, 1\}$ is designed to accommodate the local differential privacy requirements.

In this section, we first consider the scenario in which all the workers are benign. The Byzantine resilience of $sto\text{-}sign$ and $dp\text{-}sign$ will be discussed in Section 5. In addition, we assume that each worker evaluates the gradients over its whole local dataset for simplicity (i.e., $\boldsymbol{g}_m^{(t)} = \nabla f_m(w^{(t)}), \forall 1 \leq m \leq M$). Particularly, in federated learning, the workers usually compute $\nabla f_m(w^{(t)})$ due to the small size of the local dataset. The discussion about stochastic gradients is presented in Section 6. The proofs of the theoretical results are provided in Section 2 of the supplementary document.

4.1 The Stochastic Compressor $sto\text{-}sign$

Formally, the compressor $sto\text{-}sign$ is defined as follows.

Definition 1. *For any given gradient $\boldsymbol{g}_m^{(t)}$, the compressor sto-sign outputs $sto\text{-}sign(\boldsymbol{g}_m^{(t)}, \boldsymbol{b})$, where $\boldsymbol{b}$ is a vector of design parameters. The i-th entry of $sto\text{-}sign(\boldsymbol{g}_m^{(t)}, \boldsymbol{b})$ is given by*

$$sto\text{-}sign(\boldsymbol{g}_m^{(t)}, \boldsymbol{b})_i = \begin{cases} 1, \text{ with probability } \frac{b_i + (\boldsymbol{g}_m^{(t)})_i}{2b_i}, \\ -1, \text{with probability } \frac{b_i - (\boldsymbol{g}_m^{(t)})_i}{2b_i}, \end{cases} \tag{7}$$

where $(\boldsymbol{g}_m^{(t)})_i$ and $b_i \geq max_m|(\boldsymbol{g}_m^{(t)})_i|$ are the i-th entry of $\boldsymbol{g}_m^{(t)}$ and $\boldsymbol{b}$, respectively.

Recall that the performance of Algorithm 1 largely depends on the probability of wrong aggregation (c.f. (4)). When $q(\boldsymbol{g}_m^{(t)}) = sign(\boldsymbol{g}_m^{(t)})$, whether (4) holds or not is determined by the gradients $\boldsymbol{g}_m^{(t)}$'s, which are unknown. As a result, the convergence of SIGNSGD is not guaranteed. The key idea of $sto\text{-}sign$ is to introduce the stochasticity such that the probability of wrong aggregation can be theoretically bounded for an arbitrary realization of $\boldsymbol{g}_m^{(t)}$'s.

In the following discussion, we term Algorithm 1 with $q(\boldsymbol{g}_m^{(t)}) = sto\text{-}sign(\boldsymbol{g}_m^{(t)}, \boldsymbol{b})$ as Sto-SIGNSGD. For the ease of presentation, we consider the scalar case and obtain the following results for the compressor $sto\text{-}sign$. They can be readily generalized to the vector case by applying the results independently on each coordinate.

Theorem 1. *Let $u_1, u_2, \cdots, u_M$ be M known and fixed real numbers and consider random variables $\hat{u}_m = sto\text{-}sign(u_m, b)$, $1 \le m \le M$. Then we have*

$$P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M}\hat{u}_m\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}u_m\bigg)\bigg) < \big[\big(1-x\big)e^{x}\big]^{\frac{M}{2}}, \tag{8}$$

where $x = \frac{|\sum_{m=1}^{M}u_m|}{bM}$.

Here we provide some intuition about the proof. Given the majority vote rule, the aggregation result is wrong if more than half of the workers share the wrong signs. In addition, based on (7), we can obtain the probability of each worker sharing 1 or -1. Therefore, the number of workers that share the wrong signs can be modeled as a Poisson binomial variable, denoted as Z. The key difficulty is that the correct sign $sign(\frac{1}{M}\sum_{m=1}^{M}u_m)$ is unknown. However, thanks to the special structure of (7), the mean of the number of workers sharing either -1 or 1 depends on $\frac{1}{M}\sum_{m=1}^{M}u_m$ rather than on each individual u_m. That being said, we can always obtain the expectation of Z as a function of $\frac{1}{M}\sum_{m=1}^{M}u_m$. As a result, we can invoke the Markov inequality and obtain (8) after some algebra.

Remark 1. ***(selection of b)*** *Some discussions on the choice of the vector **b** in (7) are in order. We take the i-th entry of **b** as an example. In the FL application, the i-th entry of the gradient $\boldsymbol{g}_m^{(t)}$ corresponds to u_m in Theorem 1. On the one hand, according to the definition of sto-sign, $b_i \ge max_m|(\boldsymbol{g}_m^{(t)})_i|$. On the other hand, it can be shown that $\big(1-x\big)e^{x}$ in (8) is a decreasing function of x (and therefore an increasing function of b_i) when $x < 1$. Therefore, to minimize the probability of wrong aggregation, it is optimal to select $b_i = max_m|(\boldsymbol{g}_m^{(t)})_i|$. In addition, since the true gradients change during the training process, the optimal b_i varies across the iterations too. In the implementation of sto-sign, for a fixed vector **b**, it is possible that $b_i < max_m|(\boldsymbol{g}_m^{(t)})_i|$ for some coordinates. In such cases, the probabilities defined in (7) may fall out of the range $[0, 1]$. We round them to 1 if they are positive and 0 otherwise. However, in practice, since $max_m|(\boldsymbol{g}_m^{(t)})_i|$ is unknown, the selection of an appropriate **b** is an interesting problem deserving further investigation.*

Theorem 2. *Given the same $\{u_m\}_{m=1}^{M}$ and $\{\hat{u}_m\}_{m=1}^{M}$ as those in Theorem 1, for a sufficiently large b, we have $P\big(sign\big(\frac{1}{M}\sum_{m=1}^{M}\hat{u}_m\big) \neq sign\big(\frac{1}{M}\sum_{m=1}^{M}u_m\big)\big) < \frac{1}{2}$.*

Theorem 3. *Suppose Assumptions 1, 2 and 4 are satisfied, and set the learning rate $\eta = \frac{1}{\sqrt{Td}}$. Then for any M, by running Sto-SIGNSGD for T iterations, we have*

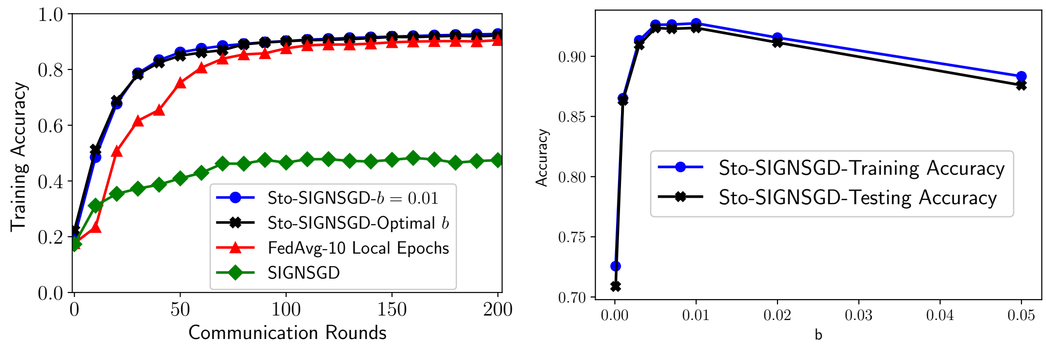
$$\frac{1}{T}\sum_{t=1}^{T}c||\nabla F(w^{(t)})||_1 \le \frac{(F(w^{(0)}) - F^*)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2bd\Delta(M), \tag{9}$$

where c is some positive constant, and $\Delta(M)$ is the solution to $\big[\big(1-x\big)e^{x}\big]^{\frac{M}{2}} = \frac{1}{2}$.

Given the results in Theorem 1, the proof of Theorem 3 follows the well known strategy of relating the norm of the gradient to the expected improvement of the global objective in a single iteration. Then accumulating the improvement over the iterations yields the convergence rate of the algorithm.

Remark 2. *Similar to SIGNSGD, the convergence rate of Sto-SIGNSGD depends on the L_1-norm of the gradient. A detailed discussion on this feature can be found in [5]. Note that compared to the convergence rate of SIGNSGD, there is a positive coefficient $c < 1$. This can be understood as the cost of dealing with the heterogeneous data distribution.*

Remark 3. *It can be verified that $\Delta(M)$ is a decreasing function of M and $\lim_{M\to\infty}\Delta(M) = 0$. In addition, if we select $b \propto \frac{1}{\sqrt{dT}}$, the right hand side of (9) is upper bounded by $O(\frac{\sqrt{d}}{\sqrt{T}})$. Moreover, Theorem 3 holds for any b and the last term in (9) captures the gap induced by the scenarios where the probability of wrong aggregation is larger than $\frac{1}{2}$. According to Theorem 2, the probability of wrong aggregation is always smaller than $\frac{1}{2}$ when b is sufficiently large and therefore, the last term in (9) can be eliminated in such a case. However, setting a larger b increases the probability of wrong aggregation and decreases the positive constant c, which negatively impact the convergence.*



https://www.overleaf.com/project

Figure 1: The left figure compares the training accuracy of Sto-SIGNSGD with SIGNSGD and FedAvg [28]. The right figure shows the training and the testing accuracy of Sto-SIGNSGD for different $\boldsymbol{b} = b \cdot \mathbf{1}$. The results are averaged over 5 repeats. For FedAvg, we tune the number of local epochs from the set $\{1, 10, 20, 30\}$ and present the best results.

Experimental results. We perform experiments to examine the learning performance of Sto-SIGNSGD for different selection of $\boldsymbol{b}$. Throughout our experiments, in the fixed $\boldsymbol{b}$ scenarios, we set $\boldsymbol{b} = b \cdot \mathbf{1}$ for some positive constant b. For "Optimal $\boldsymbol{b}$", we set $b_i = max_m|(\boldsymbol{g}_m^{(t)})_i|, \forall i$. The results are shown in Figure 1. It can be observed that Sto-SIGNSGD outperforms SIGNSGD and FedAvg, and the performance of "$b = 0.01$" is almost the same as "Optimal $\boldsymbol{b}$". That beind said, compared to FedAvg, Sto-SIGNSGD achieves better performance while requires less communication overhead per communication round. In addition, it can be observed that for fixed $\boldsymbol{b}$, b should be large enough to optimize the performance. Then when b keeps increasing, both the training accuracy and the testing accuracy decrease, which corroborates our analysis above.

4.2 The Differentially Private Compressor dp-$sign$

In this subsection, we present the differentially private version of sto-$sign$. Formally, the compressor dp-$sign$ is defined as follows.

Definition 2. *For any given gradient $\boldsymbol{g}_m^{(t)}$, the compressor dp-sign outputs dp-sign$(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)$. The i-th entry of dp-sign$(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)$ is given by*

$$dp\text{-}sign(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)_i = \begin{cases} 1, & \text{with probability } \Phi\big(\frac{(\boldsymbol{g}_m^{(t)})_i}{\sigma}\big) \\ -1, & \text{with probability } 1 - \Phi\big(\frac{(\boldsymbol{g}_m^{(t)})_i}{\sigma}\big) \end{cases} \tag{10}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the normalized Gaussian distribution; $\sigma = \frac{\Delta_2}{\epsilon}\sqrt{2\ln(\frac{1.25}{\delta})}$, where ϵ and δ are the differential privacy parameters and Δ_2 is the sensitivity measure.[2]

Theorem 4. *The proposed compressor dp-sign$(\cdot, \epsilon, \delta)$ is (ϵ, δ)-differentially private for any $\epsilon, \delta \in (0, 1)$.*

Remark 4. *Note that throughout this paper, we assume $\delta > 0$. For the $\delta = 0$ scenario, the Laplace mechanism [10] can be used by replacing the cumulative distribution function of the normalized Gaussian distribution in (10) with that of the Laplace distribution. The corresponding discussion is provided in the supplementary document.*

We term Algorithm 1 with $q(\boldsymbol{g}_m^{(t)}) = dp\text{-}sign(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)$ as DP-SIGNSGD. Similar to sto-$sign$, we consider the scalar case and obtain the following result for dp-$sign(\cdot, \epsilon, \delta)$.

[2]Please refer to Section 1 of the supplementary document for detailed information about the differential privacy parameters (ϵ,δ) and the sensitivity measure Δ_2.

Theorem 5. *Let $u_1, u_2, \cdots, u_M$ be M known and fixed real numbers. Further define random variables $\hat{u}_i = dp\text{-}sign(u_i, \epsilon, \delta), \forall 1 \leq i \leq M$. Then there always exist a constant σ_0 such that when $\sigma \geq \sigma_0$, $P(sign(\frac{1}{M}\sum_{m=1}^{M}\hat{u}_i) \neq sign(\frac{1}{M}\sum_{m=1}^{M}u_i)) < \big[(1-x)e^x\big]^{\frac{M}{2}}$, where $x = \frac{|\sum_{m=1}^{M}u_m|}{\sigma M}$.*

Given Theorem 5, the convergence of DP-SIGNSGD can be obtained by following a similar analysis to that of Theorem 3.

5 Byzantine Resilience

In this section, the Byzantine resilience of the proposed algorithms is investigated. We note that the convergence of Sto-SIGNSGD and DP-SIGNSGD is limited by the probability of wrong aggregation (i.e., more than half of the workers share the wrong signs). Let Z_i denote the number of normal workers that share (quantized) gradients with different signs from the true gradient $\nabla F(w^{(t)})$ on the i-th coordinate (i.e., $q(\boldsymbol{g}_m^{(t)})_i \neq sign(\nabla F(w^{(t)})_i)$). Then, Z_i is a Poisson binomial variable. In order to tolerate k_i Byzantine workers on the i-th coordinate of the gradient, we need to have $P(Z_i \geq \frac{M-k_i}{2}) < \frac{1}{2}$, where M is the number of benign workers. Therefore, we can prove the following theorem.

Theorem 6. *There exists a positive constant s_0 such that when $s > s_0$, Sto-SIGNSGD and DP-SIGNSGD can at least tolerate k_i Byzantine attackers on the i-th coordinate of the gradient at t-th iteration and k_i satisfies*

$$k_i < \frac{|\sum_{m=1}^{M}\nabla f_m(w^{(t)})_i|}{s}, \quad \big[(1-x)e^x\big]^{\frac{M-k_i}{2}} < \frac{1}{2}, \tag{11}$$

where $x = \frac{|\sum_{m=1}^{M}\nabla f_m(w^{(t)})_i| - sk_i}{(M-k_i)s}$, $s = \sigma$ for DP-SIGNSGD and $s = b_i \geq max_m|(\boldsymbol{g}_m^{(t)})_i|$ for Sto-SIGNSGD.

Overall, the number of Byzantine workers that the algorithms can tolerate is given by $min_{1\leq i\leq d}k_i$.

In this case, $\big[(1-x)e^x\big]^{\frac{M-k_i}{2}}$ measures the probability of wrong aggregation after taking the Byzantine workers into consideration. As we know, $(1-x)e^x$ is decreasing function of x (and therefore an increasing function of s). So the second equation of (11) indicates that the Byzantine tolerance decreases as s increases, which conforms to the observation from the first equation of (11).

According to (11), when $sto\text{-}sign$ is used, we can set $b_i = max_m|\nabla f_m(w^{(t)})_i|$. In this case, $k_i < \frac{|\sum_{m=1}^{M}\nabla f_m(w^{(t)})_i|}{max_m|\nabla f_m(w^{(t)})_i|}$, which means that the Byzantine resilience depends on the heterogeneity of the local datasets. When the workers can access the same dataset, i.e., $\nabla f_m(w^{(t)})_i = \nabla f_n(w^{(t)})_i, \forall m, n$, Theorem 6 gives $x = 1$ and $k_i < M$. Therefore, it can tolerate $M-1$ Byzantine workers.

Remark 5. *Our analysis of the convergence and the Byzantine resilience is based on each individual coordinate of the gradients, which corresponds to the generalized Byzantine attacks and the dimensional Byzantine resilience [29]. Furthermore, it also indicates that the parameter σ in $dp\text{-}sign$ can be different across coordinates and iterations, which allows one to select suitable parameters for different coordinates and iterations to improve the privacy performance of the algorithm. A similar idea has been explored in [30] without considering quantization.*

Experimental results. Fig. 2 shows the performance of Sto-SIGNSGD for different selection of $\boldsymbol{b} = b \cdot \mathbf{1}$ and different number of Byzantine workers B. It can be seen that when $b = 0.001$, it is not large enough to optimize the performance according to our results in Section 4. Setting $b = 0.003$ and $b = 0.01$ achieves almost the same performance as "Optimal $\boldsymbol{b}$" when there is no Byzantine worker ($B = 0$). However, as the number of Byzantine workers increases, both the training and the testing accuracy of "$b = 0.01$" drop much faster than those of "$b = 0.003$", which conforms to our analysis above that a lager b results in worse Byzantine resilience.

To examine the impact of data heterogeneity, we vary the number of labels of each worker's local training dataset. Table 1 shows the testing accuracy of Sto-SIGNSGD with optimal $\boldsymbol{b}$. It can be observed that the Byzantine resilience of Sto-SIGNSGD increases as the number of labels increases. Up to now, we examine the performance of Sto-SIGNSGD, the results for DP-SIGNSGD are deferred to Section 7.

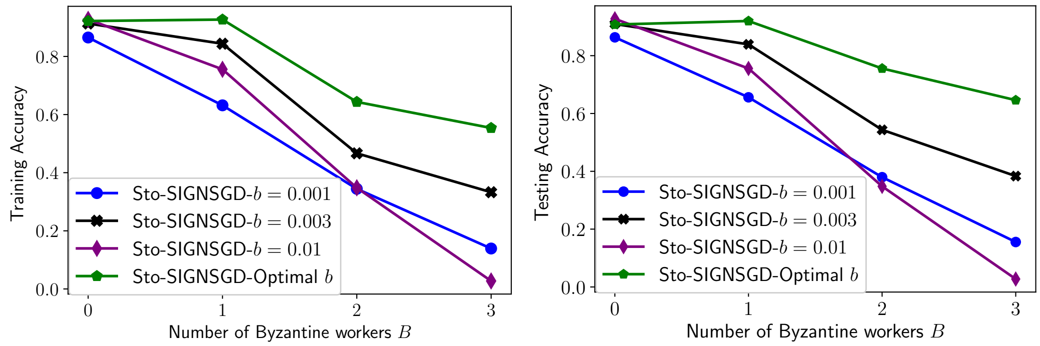


Figure 2: The training and the testing accuracy of Sto-SIGNSGD for different number of Byzantine workers and different $\boldsymbol{b}$.

Table 1: Testing Accuracy of Sto-SIGNSGD

B	1 LABEL	2 LABELS	4 LABELS
0	90.78%	91.03%	92.38%
1	91.98%	90.78%	92.66%
2	75.58%	90.77%	93.30%
3	64.61%	77.88%	93.22%
4	11.62%	53.99%	88.22%

6 Extending to SGD

Up until this point in the paper, the discussions are based on the assumption that each worker can evaluate its local true gradient $\nabla f_m(w^{(t)})$ for the ease of presentation. In the SGD scenario, we have to further account for the sampling noise. Particularly, the following theorem for Sto-SIGNSGD can be proved. The corresponding result for DP-SIGNSGD can be obtained following a similar strategy.

Theorem 7. *Suppose Assumptions 1-4 are satisfied, and set the learning rate $\eta = \frac{1}{\sqrt{Td}}$. Then, when $\boldsymbol{b} = b \cdot \mathbf{1}$ and b is sufficiently large, Sto-SIGNSGD converges to the (local) optimum with a rate of $O(\frac{1}{\sqrt{T}})$ if either of the following two conditions is satisfied.*

- $P\big(sign(\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i) \neq sign(\nabla F(w^t)_i\big) < 0.5, \forall 1 \leq i \leq d.$
- *The mini-batch size of stochastic gradient at each iteration is at least T.*

Remark 6. *Note that the first condition is not hard to satisfy. One sufficient condition is that the sampling noise of each worker is symmetric with zero mean. This assumption is also used in [8], which shows that the sampling noise is approximately not only symmetric, but also unimodal.*

Remark 7. *We note that by replacing the compressor sign in SIGNSGD with sto-sign or dp-sign, we can obtain the improved rate (a factor of $\frac{1}{\sqrt{M}}$ in the variance term) without assuming unimodal and symmetric stochastic gradient sampling noise as in [5]. More details can be found in Section 2 of the supplementary document.*

Remark 8. *We note that the above discussion assumes that b is sufficiently large, which guarantees that the probability of wrong aggregation is less than 0.5. For an arbitrary b that satisfies the condition in the definition of sto-sign, we believe that it is possible to prove that the algorithm converges to the neighborhood of the (local) optimum. In particular, similar to the proof of Theorem 3, there will be an additional term $\sum_{i=1}^{d}|\nabla F(w^t)_i|\mathbb{1}_{|\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i|\leq b\Delta(M)}$. It is possible to upper bound this additional term given the fact that $\mathbb{E}[\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i] = \nabla F(w^t)_i$, despite that more efforts are required to make the analysis rigorous.*

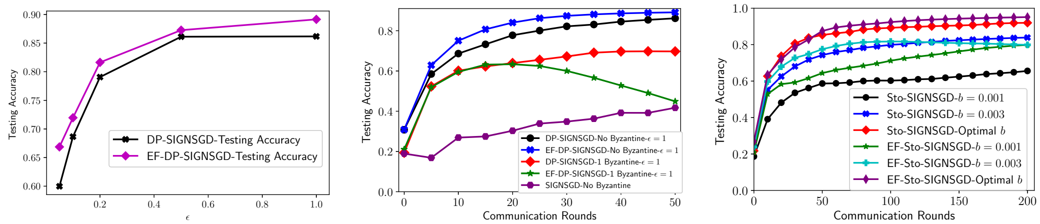


Figure 3: The first figure shows the performance of DP-SIGNSGD and EF-DP-SIGNSGD for different ϵ when $\delta = 10^{-5}$, without Byzantine attackers. The ϵ's measure the per epoch privacy guarantee of the algorithms. The second figure compares EF-DP-SIGNSGD with DP-SIGNSGD when $\epsilon = 1$. The last figure compares Sto-SIGNSGD with EF-Sto-SIGNSGD in the presence of 1 Byzantine attacker.

7 Extending to Error-feedback Variant

To further improved the performance of Algorithm 1, we incorporate the error-feedback technique and propose its error-feedback variant, which is presented in Algorithm 2.

Algorithm 2 Error-Feedback Stochastic-Sign SGD with majority vote

---

Input: learning rate η, current hypothesis vector $w^{(t)}$, current residual error vector $\tilde{\boldsymbol{e}}^{(t)}$, M workers each with an independent gradient $\boldsymbol{g}_m^{(t)} = \nabla f_m(w^{(t)})$, the 1-bit compressor $q(\cdot)$.

on server:

pull $q(\boldsymbol{g}_m^{(t)})$ **from** worker m.

push $\tilde{\boldsymbol{g}}^{(t)} = sign\big(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)}\big)$ **to** all the workers,

update residual error:

$$\tilde{\boldsymbol{e}}^{(t+1)} = \frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)} - \frac{1}{M}\tilde{\boldsymbol{g}}^{(t)}. \tag{12}$$

on each worker:

update $w^{(t+1)} = w^{(t)} - \eta\tilde{\boldsymbol{g}}^{(t)}$.

Remark 9. *Note that in Algorithm 2, only the server adopts the error-feedback method. When dp-$sign$ is used, implementing error-feedback on the worker's side may increase the privacy leakage. Accounting for the additional privacy leakage caused by error-feedback is left as future work.*

Remark 10. *In (12), by adding the coefficient $\frac{1}{M}$ to $\tilde{\boldsymbol{g}}^{(t)}$, the server keeps the magnitude information about the aggregation results and enables more effective error-feedback performance. More discussion about the parameter $\frac{1}{M}$ is provided in the supplementary document.*

Both sto-$sign$ and dp-$sign$ can be used in Algorithm 2 and the corresponding algorithms are termed as EF-Sto-SIGNSGD and EF-DP-SIGNSGD, respectively. In the following, we show the convergence and Byzantine resilience of Algorithm 2 when dp-$sign$ is used. The results can be easily adapted for sto-$sign$. Particularly, the following theorems can be proved.

Theorem 8. *When Assumptions 1, 2 and 4 are satisfied, there exists a σ_0 such that when $\sigma \geq \sigma_0$, by running Algorithm 2 with $\eta = \frac{1}{M\sqrt{Td}}$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{||\nabla F(w^{(t)})||_2^2}{\sigma} \leq \frac{(F(w_0) - F^*)\sqrt{d}}{\sqrt{T}} + \frac{(1 + L + L^2\beta)\sqrt{d}}{\sqrt{T}}, \tag{13}$$

where β is some positive constant.

The proof of Theorem 8 follows the strategy of taking $y^{(t)} = w^{(t)} - \eta\tilde{e}^{(t)}$ such that $y^{(t)}$ is updated in the same way as $w^{(t)}$ in the non error-feedback scenario. A key technical challenge is to bound the norm of the residual error $||\tilde{e}^{(t)}||_2^2$. Utilizing the fact that the output of the compressor $q(\cdot) \in \{-1, 1\}$, we upper bound it by first proving that in this case, the server's compressor is an α-approximate compressor [18] for some $\alpha < 1$.

Besides the fact that error-feedback is only used on the server's side, another difference between Algorithm 2 and those in [18, 19] is that it does not require the workers to share the magnitude information about the gradients. On the one hand, it saves communication overhead. On the other hand, it keeps the resilience against the re-scaling attacks. By following a similar strategy to the proofs of Theorem 8 and considering the impact of Byzantine attackers, we obtain the Byzantine resilience of Algorithm 2 as follows.

Theorem 9. *At each iteration t, there exists a constant σ_0 such that when $\sigma > \sigma_0$, Algorithm 2 can at least tolerate $k_i = |\sum_{m=1}^{M} \nabla f_m(w^{(t)})_i|/\sigma$ Byzantine attackers on the i-th coordinate of the gradient. Overall, the number of Byzantine workers that Algorithm 2 can tolerate is given by $min_{1 \leq i \leq d} k_i$.*

Experimental results. For DP-SIGNSGD and EF-DP-SIGNSGD, we follow the idea of gradient clipping in [31] to bound the sensitivity Δ_2. After computing the gradient for each individual training sample in the local dataset, each worker clips it in its L_2 norm for a clipping threshold C to ensure that $\Delta_2 \leq C$. We set $C = 4$ in the experiments and the results are shown in Fig. 3. It can be observed from the first two figures that when there is no Byzantine attackers, EF-DP-SIGNSGD outperforms DP-SIGNSGD for all the examined ϵ's, which demonstrates its effectiveness. In addition, both DP-SIGNSGD and EF-DP-SIGNSGD outperform SIGNSGD, while providing privacy guarantees.

Another observation is that the error-feedback variants do not necessarily perform better. For instance, in the second figure of Fig. 3, when there is one Byzantine attacker and $\epsilon = 1$, the testing accuracy of EF-DP-SIGNSGD is worse than that of DP-SIGNSGD. In the beginning of the training process, k_i's in Theorem 9 are large enough such that the algorithm can tolerate the Byzantine attacker. As the gradients decrease, the probability of wrong aggregation increases. In this case, the error-feedback mechanism may carry the wrong aggregations to the future iterations and have a negative impact on the learning process. Similar results are obtained for Sto-SIGNSGD when $b = 0.003$ in the last figure of Fig. 3. In the meantime, for $b = 0.001$ and "Optimal $\boldsymbol{b}$", the error-feedback variant can tolerate the Byzantine attacker and therefore provide better performance.

8 Conclusion

We propose a Stochastic-Sign SGD framework which utilizes two novel gradient compressors and can deal with heterogeneous data distribution. The proposed algorithms are proved to converge in the heterogeneous data distribution scenario with the same rate as SIGNSGD in the homogeneous data distribution case. In particular, the proposed differentially private compressor $dp\text{-}sign$ improves the privacy and the accuracy simultaneously without sacrificing any communication efficiency. Then, we further improve the learning performance of the proposed method by incorporating the error-feedback scheme. In addition, the Byzantine resilience of the proposed algorithms is shown analytically. It is expected that the proposed algorithms can find wide applications in the design of communication efficient, differentially private and Byzantine resilient FL algorithms.

Supplementary Material

The supplementary material is organized as follows. In Section 1, we formally provide the definition of local differential privacy [10]. In Section 2, we provide the proofs of the theoretical results presented in the main document. Discussions about the extended differentially private compressor dp-$sign$ when $\delta = 0$ are provided in Section 3. Discussions about the server's compressor $\frac{1}{M}sign(\cdot)$ in Algorithm 2 are provided in Section 4. The details about the implementation of our experiments and some additional experimental results are presented in Section 5.

1 Definition of Local Differential Privacy

In this work, we study the privacy guarantee of the proposed algorithms from the lens of local differential privacy [10], which provides a strong notion of individual privacy in data analysis. The definition of local differential privacy is formally given as follows.

Definition 3. *Given a set of local datasets $\mathcal{D}$ provided with a notion of neighboring local datasets $\mathcal{N}_{\mathcal{D}} \subset \mathcal{D} \times \mathcal{D}$ that differ in only one data point. For a query function $f : \mathcal{D} \rightarrow \mathcal{X}$, a mechanism $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{O}$ to release the answer of the query is defined to be (ϵ, δ)-locally differentially private if for any measurable subset $\mathcal{S} \subseteq \mathcal{O}$ and two neighboring local datasets $(D_1, D_2) \in \mathcal{N}_{\mathcal{D}}$,*

$$P(\mathcal{M}(f(D_1)) \in \mathcal{S}) \leq e^{\epsilon} P(\mathcal{M}(f(D_2)) \in \mathcal{S}) + \delta. \tag{14}$$

A key quantity in characterizing local differential privacy for many mechanisms is the sensitivity of the query f in a given norm l_r, which is defined as

$$\Delta_r = \max_{(D_1, D_2) \in \mathcal{N}_{\mathcal{D}}} ||f(D_1) - f(D_2)||_r. \tag{15}$$

For more details about the concept of differential privacy, the reader is referred to [10] for a survey.

2 Proofs

2.1 Proof of Theorem 1

Theorem 1. *Let $u_1, u_2, \cdots, u_M$ be M known and fixed real numbers and consider random variables $\hat{u}_m = sto\text{-}sign(u_m, b)$, $1 \leq m \leq M$. Then we have*

$$P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M}\hat{u}_m\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}u_m\bigg)\bigg) < \bigg[\bigg(1 - \frac{1}{x}\bigg)e^{\frac{1}{x}}\bigg]^{\frac{M}{2}}, \tag{16}$$

where $x = \frac{bM}{|\sum_{m=1}^{M} u_m|}$.

We first provide some intuition about the proof. Given the majority vote rule, the aggregation result is wrong if more than half of the workers share the wrong signs. In addition, based on the definition of sto-$sign$, we can obtain the probability of each worker sharing 1 or -1. Therefore, the number of workers that share the wrong signs can be modeled as a Poisson binomial variable, denoted as Z. The key difficulty is that the correct sign $sign(\frac{1}{M}\sum_{m=1}^{M} u_m)$ is unknown. However, it can be shown that the mean of the number of workers sharing either -1 or 1 depends on $\frac{1}{M}\sum_{m=1}^{M} u_m$ rather than on each individual u_m. That being said, we can always obtain the expectation of Z as a function of $\frac{1}{M}\sum_{m=1}^{M} u_m$. As a result, we can invoke the Markov inequality and obtain (16) after some algebra.

Proof. Without loss of generality, assume $u_1 \leq u_2 \leq \cdots \leq u_K < 0 \leq u_{K+1} \leq \cdots \leq u_M$ and $\frac{1}{M}\sum_{m=1}^{M} u_m < 0$. Note that similar analysis can be done when $\frac{1}{M}\sum_{i=m}^{M} u_m > 0$. Further define a series of random variables $\{X_m\}_{m=1}^{M}$ given by

$$X_m = \begin{cases} 1, & \text{if } \hat{u}_m \neq sign\Big(\frac{1}{M}\sum_{m=1}^{M} u_m\Big), \\ 0, & \text{if } \hat{u}_m = sign\Big(\frac{1}{M}\sum_{m=1}^{M} u_m\Big). \end{cases} \tag{17}$$

In particular, X_m can be considered as the outcome of one Bernoulli trial with successful probability $P(X_m = 1)$. Let $Z = \sum_{m=1}^{M} X_m$ and we have

$$P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M}\hat{u}_m\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}u_m\bigg)\bigg) = P\bigg(Z \geq \frac{M}{2}\bigg). \tag{18}$$

Note that according to the definition of $sto\text{-}sign$, b is large enough such that $b > \max_m |u_m|$. The probability of $X_m = 1$ is given by

$$P(X_m = 1) = \frac{b + u_m}{2b}. \tag{19}$$

Then, Z follows the Poisson binomial distribution with mean and variance given by

$$\begin{aligned} \mu &= \sum_{m=1}^{M} P(X_m = 1) = \frac{M}{2} + \frac{\sum_{m=1}^{M} u_m}{2b}, \\ \sigma^2 &= \sum_{m=1}^{M} \frac{(b - u_m)(b + u_m)}{4b^2}. \end{aligned} \tag{20}$$

For any variable $a > 0$, we have

$$\mathbb{E}\big[e^{aZ}\big] = \mathbb{E}\big[e^{a\sum_{m=1}^{M} X_m}\big] = \mathbb{E}\bigg[\prod_{m=1}^{M} e^{aX_m}\bigg] = \prod_{m=1}^{M} \mathbb{E}\big[e^{aX_m}\big], \tag{21}$$

where the last equality is due to the independence among X_m's. In addition,

$$\mathbb{E}\big[e^{aX_m}\big] = P(X_m = 1)e^a + P(X_m = 0) = 1 + P(X_m = 1)(e^a - 1) \leq e^{P(X_m=1)(e^a-1)}, \tag{22}$$

where the last inequality is due to the inequality $1 + y \leq e^y$.

Combining (21) and (22), we have

$$\mathbb{E}\big[e^{aZ}\big] = \prod_{m=1}^{M} \mathbb{E}\big[e^{aX_m}\big] \leq \prod_{m=1}^{M} e^{P(X_m=1)(e^a-1)} \leq e^{(e^a-1)\mu}. \tag{23}$$

Therefore,

$$P\bigg(Z \geq \frac{M}{2}\bigg) = P\big(e^{aZ} \geq e^{\frac{Ma}{2}}\big) \leq \frac{\mathbb{E}\big[e^{aZ}\big]}{e^{\frac{Ma}{2}}} \leq \frac{e^{(e^a-1)\mu}}{e^{\frac{Ma}{2}}}, \tag{24}$$

where we invoke the Markov's inequality.

Since $\frac{\sum_{m=1}^{M} u_m}{2b} < 0$ by our assumption, it can be verified that $\frac{M}{2\mu} > 1$. Let $a = \ln(\frac{M}{2\mu}) > 0$, we have

$$P\bigg(Z \geq \frac{M}{2}\bigg) \leq \frac{e^{\left(e^{\ln(\frac{M}{2\mu})}-1\right)\mu}}{e^{\frac{M\ln(\frac{M}{2\mu})}{2}}} = \frac{e^{\frac{M}{2}-\mu}}{(\frac{M}{2\mu})^{\frac{M}{2}}} = \frac{e^{-\frac{\sum_{m=1}^{M} u_m}{2b}}}{(\frac{M}{M+\frac{1}{b}\sum_{m=1}^{M} u_m})^{\frac{M}{2}}}. \tag{25}$$

Let $x = \frac{|\sum_{m=1}^{M} u_m|}{bM}$ and it can be verified that $x < 1$ since $b > max_m|u_m|$. Then (25) can be reduced to

$$P\bigg(Z \geq \frac{M}{2}\bigg) \leq \big[(1-x)e^x\big]^{\frac{M}{2}}. \tag{26}$$

□

2.2 Proof of Theorem 2

Theorem 2. *Given the same $\{u_m\}_{m=1}^{M}$ and $\{\hat{u}_m\}_{m=1}^{M}$ as those in Theorem 1, for a sufficiently large b, we have $P\big(sign\big(\frac{1}{M}\sum_{m=1}^{M}\hat{u}_m\big) \neq sign\big(\frac{1}{M}\sum_{m=1}^{M}u_m\big)\big) < \frac{1}{2}$.*

Proof. Without loss of generality, assume $u_1 \leq u_2 \leq \cdots \leq u_K < 0 \leq u_{K+1} \leq \cdots \leq u_M$. According to the definition of $sto\text{-}sign$, we have

$$\hat{u}_m = sto\text{-}sign(u_m, b) = \begin{cases} 1, & \text{with probability } \frac{b+u_m}{2b}, \\ -1, & \text{with probability } \frac{b-u_m}{2b}, \end{cases} \tag{27}$$

Further define a series of random variables $\{\hat{X}_m\}_{m=1}^M$ given by

$$\hat{X}_m = \begin{cases} 1, & \text{if } \hat{u}_m = 1, \\ 0, & \text{if } \hat{u}_m = -1. \end{cases} \tag{28}$$

In particular, $\hat{X}_m$ can be considered as the outcome of one Bernoulli trial with successful probability $P(\hat{X}_m = 1)$. Let $\hat{Z} = \sum_{m=1}^M \hat{X}_m$, then

$$P\left(sign\left(\frac{1}{M}\sum_{m=1}^M \hat{u}_m\right) = 1\right) = P\left(\hat{Z} \geq \frac{M}{2}\right) = \sum_{H=\frac{M+1}{2}}^M P(\hat{Z} = H). \tag{29}$$

In addition,

$$P(\hat{Z} = H) = \frac{\sum_{A \in F_H} \prod_{i \in A}(b+u_i) \prod_{j \in A^c}(b-u_j)}{(2b)^M} = \frac{a_{M,H}b^M + a_{M-1,H}b^{M-1} + \cdots + a_{0,H}b^0}{(2b)^M}, \tag{30}$$

in which F_H is the set of all subsets of H integers that can be selected from $\{1, 2, 3, ..., M\}$; $a_{m,H}, \forall 0 \leq m \leq M$ is some constant. It can be easily verified that $a_{M,H} = \binom{M}{H}$.

When b is sufficiently large, $P(\hat{Z} = H)$ is dominated by the first two terms in (30). In particular, $\forall m$, we have

$$\begin{aligned} \sum_{A \in F_H} \prod_{i \in A}(b+u_i) \prod_{j \in A^c}(b-u_j) = (b+u_m) \sum_{A \in F_H, m \in A} \prod_{i \in A/\{m\}}(b+u_i) \prod_{j \in A^c}(b-u_j) \\ + (b-u_m) \sum_{A \in F_H, m \notin A} \prod_{i \in A}(b+u_i) \prod_{j \in A^c/\{m\}}(b-u_j). \end{aligned} \tag{31}$$

As a result, when $\frac{M+1}{2} \leq H \leq M-1$, the u_m related term in $a_{M-1,H}$ is given by

$$\left[\binom{M-1}{H-1} - \binom{M-1}{H}\right] u_m. \tag{32}$$

When $H = M$, the u_m related term in $a_{M-1,H}$ is given by

$$\left[\binom{M-1}{H-1}\right] u_m. \tag{33}$$

By summing over m, we have

$$a_{M-1,H} = \left[\binom{M-1}{H-1} - \binom{M-1}{H}\right] \sum_{m=1}^M u_m, \tag{34}$$

and

$$a_{M-1,H} = \left[\binom{M-1}{H-1}\right] \sum_{m=1}^M u_m, \tag{35}$$

when $\frac{M+1}{2} \leq H \leq M-1$ and $H = M$, respectively.

By summing over H, we have

$$\sum_{H=\frac{M+1}{2}}^M a_{M,H} = \sum_{H=\frac{M+1}{2}}^M \binom{M}{H} = 2^{M-1}, \tag{36}$$

$$\sum_{H=\frac{M+1}{2}}^M a_{M-1,H} = \binom{M-1}{\frac{M-1}{2}} \sum_{m=1}^M u_m \tag{37}$$

As a result,

$$P\left(\hat{Z} \geq \frac{M}{2}\right) = \sum_{H=\frac{M+1}{2}}^{M} P(\hat{Z} = H) = \frac{2^{M-1}b^M + \binom{M-1}{\frac{M-1}{2}} \sum_{m=1}^{M} u_m b^{M-1}}{(2b)^M} + O\left(\frac{1}{b^2}\right)$$
$$= \frac{1}{2} + \frac{\binom{M-1}{\frac{M-1}{2}}}{2^M b} \sum_{m=1}^{M} u_m + O\left(\frac{1}{b^2}\right). \tag{38}$$

Therefore, if the second term dominates the third term (i.e., b is sufficiently large), $P\big(\hat{Z} \geq \frac{M}{2}\big) > \frac{1}{2}$ when $\sum_{m=1}^{M} u_m > 0$; $P\big(\hat{Z} \geq \frac{M}{2}\big) < \frac{1}{2}$ when $\sum_{m=1}^{M} u_m < 0$. That being said, the probability of wrong aggregation is always smaller than 1/2. $\square$

2.3 Proof of Theorem 3

Theorem 3. *Suppose Assumptions 1, 2 and 4 are satisfied, and set the learning rate $\eta = \frac{1}{\sqrt{Td}}$. Then for any M, by running Sto-SIGNSGD for T iterations, we have*

$$\frac{1}{T}\sum_{t=1}^{T} c||\nabla F(w^{(t)})||_1 \leq \frac{(F(w^{(0)}) - F^*)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2bd\Delta(M), \tag{39}$$

where c is some positive constant, and $\Delta(M)$ is the solution to $\big[(1-x)e^x\big]^{\frac{M}{2}} = \frac{1}{2}$.

The proof of Theorem 3 follows the well known strategy of relating the norm of the gradient to the expected improvement of the global objective in a single iteration. Then accumulating the improvement over the iterations yields the convergence rate of the algorithm.

Proof. According to Assumption 2, we have

$$\begin{aligned}
&F(w^{(t+1)}) - F(w^{(t)}) \\
&\leq < \nabla F(w^{(t)}), w^{(t+1)} - w^{(t)} > + \frac{L}{2}||w^{(t+1)} - w^{(t)}||^2 \\
&= -\eta < \nabla F(w^{(t)}), sign\bigg(\frac{1}{M}\sum_{m=1}^{M} sto\text{-}sign(\boldsymbol{g}_m^{(t)})\bigg) > + \frac{L}{2}\bigg|\bigg|\eta sign\bigg(\frac{1}{M}\sum_{m=1}^{M} sto\text{-}sign(\boldsymbol{g}_m^{(t)})\bigg)\bigg|\bigg|^2 \\
&= -\eta < \nabla F(w^{(t)}), sign\bigg(\frac{1}{M}\sum_{m=1}^{M} sto\text{-}sign(\boldsymbol{g}_m^{(t)})\bigg) > + \frac{L\eta^2 d}{2} \\
&= -\eta||\nabla F(w^{(t)})||_1 + \frac{L\eta^2 d}{2} + 2\eta\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{sign(\frac{1}{M}\sum_{m=1}^{M} sto\text{-}sign(\boldsymbol{g}_m^{(t)})_i) \neq sign(\nabla F(w^{(t)})_i)},
\end{aligned} \tag{40}$$

where $\nabla F(w^{(t)})_i$ is the i-th entry of the vector $\nabla F(w^{(t)})$ and η is the learning rate. Taking expectation on both sides yeilds

$$\begin{aligned}
&\mathbb{E}[F(w^{(t+1)}) - F(w^{(t)})] \leq -\eta||\nabla F(w^{(t)})||_1 + \frac{L\eta^2 d}{2} \\
&+ 2\eta\sum_{i=1}^{d}|\nabla F(w^{(t)})_i| P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} sto\text{-}sign(\boldsymbol{g}_m^{(t)})_i\bigg) \neq sign(\nabla F(w^{(t)})_i)\bigg).
\end{aligned} \tag{41}$$

Let $\Delta(M)$ denote the solution to $\big[(1-x)e^x\big]^{\frac{M}{2}} = \frac{1}{2}$. Since $\big[(1-x)e^x\big]$ is a decreasing function of x for $x < 1$, it can be verified that $\big[(1-x)e^x\big]^{\frac{M}{2}} < \frac{1}{2}$ when $x > \Delta(M)$ and $\big[(1-x)e^x\big]^{\frac{M}{2}} \geq \frac{1}{2}$

otherwise. According to Theorem 1, we have two possible scenarios as follows.

$$P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} sto\text{-}sign(\boldsymbol{g}_m^{(t)})_i\bigg) \neq sign(\nabla F(w^{(t)})_i)\bigg) \begin{cases} \leq \frac{1}{2} - \frac{c_1}{2}, \text{if } \frac{|\nabla F(w^{(t)})_i|}{b} > \Delta(M), \\ \leq \frac{1}{2} + \frac{c_2}{2}, \text{if } \frac{|\nabla F(w^{(t)})_i|}{b} \leq \Delta(M), \end{cases} \tag{42}$$

in which and $c_1, c_2 < 1$ are some positive constants.

Plugging (42) into (41), we can obtain

$$\begin{aligned}
&\mathbb{E}[F(w^{(t+1)}) - F(w^{(t)})] \\
&\leq -\eta||\nabla F(w^{(t)})||_1 + \frac{L\eta^2 d}{2} \\
&+ \eta\bigg[(1-c_1)\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{\frac{|\nabla F(w^{(t)})_i|}{b} > \Delta(M)} + (1-c_2)\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{\frac{|\nabla F(w^{(t)})_i|}{b} \leq \Delta(M)}\bigg] \\
&+ 2\eta c_2\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{\frac{|\nabla F(w^{(t)})_i|}{b} \leq \Delta(M)} \\
&\leq -\eta||\nabla F(w^{(t)})||_1 + \frac{L\eta^2 d}{2} + \eta(1-c)||\nabla F(w^{(t)})||_1 + 2\eta c_2\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{\frac{|\nabla F(w^{(t)})_i|}{b} \leq \Delta(M)} \\
&= -\eta c||\nabla F(w^{(t)})||_1 + \frac{L\eta^2 d}{2} + 2\eta c_2\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{\frac{|\nabla F(w^{(t)})_i|}{b} \leq \Delta(M)},
\end{aligned} \tag{43}$$

where $c = min\{c_1, c_2\}$. Adjusting the above inequality and averaging both sides over $t = 1, 2, \cdots, T$, we can obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\eta c||\nabla F(w^{(t)})||_1 &\leq \frac{\mathbb{E}[F(w^{(0)}) - F(w^{(T+1)})]}{T} + \frac{L\eta^2 d}{2} \\
&+ \frac{2\eta c_2}{T}\sum_{t=1}^{T}\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{\frac{|\nabla F(w^{(t)})_i|}{b} \leq \Delta(M)} \\
&\leq \frac{\mathbb{E}[F(w^{(0)}) - F(w^{(T+1)})]}{T} + \frac{L\eta^2 d}{2} + 2\eta db\Delta(M).
\end{aligned} \tag{44}$$

Letting $\eta = \frac{1}{\sqrt{dT}}$ and dividing both sides by η gives

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}c||\nabla F(w^{(t)})||_1 &\leq \frac{\mathbb{E}[F(w^{(0)}) - F(w^{(T+1)})]\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2db\Delta(M) \\
&\leq \frac{(F(w^{(0)}) - F^*)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2db\Delta(M),
\end{aligned} \tag{45}$$

which completes the proof. □

2.4 Proof of Theorem 4

Theorem 4. *The proposed compressor $dp\text{-}sign(\cdot, \epsilon, \delta)$ is (ϵ, δ)-differentially private for any $\epsilon, \delta \in (0, 1)$.*

Proof. We start from the one-dimension scenario and consider any a, b that satisfy $||a - b||_2 \leq \Delta_2$. Without loss of generality, assume that $dp\text{-}sign(a, \epsilon, \delta) = dp\text{-}sign(b, \epsilon, \delta) = 1$. Then we have

$$\begin{aligned}
P(dp\text{-}sign(a, \epsilon, \delta) = 1) &= \Phi\bigg(\frac{a}{\sigma}\bigg) = \int_{-\infty}^{a}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}dx, \\
P(dp\text{-}sign(b, \epsilon, \delta) = 1) &= \Phi\bigg(\frac{b}{\sigma}\bigg) = \int_{-\infty}^{b}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}dx.
\end{aligned} \tag{46}$$

Furthermore,

$$\frac{P(dp\text{-}sign(a,\epsilon,\delta)=1)}{P(dp\text{-}sign(b,\epsilon,\delta)=1)}=\frac{\int_{-\infty}^{a}e^{-\frac{x^2}{2\sigma^2}}dx}{\int_{-\infty}^{b}e^{-\frac{x^2}{2\sigma^2}}dx}=\frac{\int_{0}^{\infty}e^{-\frac{(x-a)^2}{2\sigma^2}}dx}{\int_{0}^{\infty}e^{-\frac{(x-b)^2}{2\sigma^2}}dx}. \tag{47}$$

According to Theorem A.1 in [10], given the parameters ϵ,δ and σ, it can be verified that $e^{-\epsilon}\leq\left|\frac{P(dp\text{-}sign(a,\epsilon,\delta)=1)}{P(dp\text{-}sign(b,\epsilon,\delta)=1)}\right|\leq e^{\epsilon}$ with probability at least $1-\delta$.

For the multi-dimension scenario, consider any vector $\boldsymbol{a}$ and $\boldsymbol{b}$ such that $||\boldsymbol{a}-\boldsymbol{b}||_2\leq\Delta_2$ and $\boldsymbol{v}\in\{-1,1\}^d$, we have

$$\frac{P(dp\text{-}sign(\boldsymbol{a},\epsilon,\delta)=\boldsymbol{v})}{P(dp\text{-}sign(\boldsymbol{b},\epsilon,\delta)=\boldsymbol{v})}=\frac{\int_D e^{-\frac{||\boldsymbol{x}-\boldsymbol{a}||_2^2}{2\sigma^2}}d\boldsymbol{x}}{\int_D e^{-\frac{||\boldsymbol{x}-\boldsymbol{b}||_2^2}{2\sigma^2}}d\boldsymbol{x}}, \tag{48}$$

where D is some integral area depending on $\boldsymbol{v}$. Similarly, it can be shown that $e^{-\epsilon}\leq\left|\frac{P(dp\text{-}sign(\boldsymbol{a},\epsilon,\delta)=\boldsymbol{v})}{P(dp\text{-}sign(\boldsymbol{b},\epsilon,\delta)=\boldsymbol{v})}\right|\leq e^{\epsilon}$ with probability at least $1-\delta$. $\square$

2.5 Proof of Theorem 5

Theorem 5. *Let $u_1,u_2,\cdots,u_M$ be M known and fixed real numbers. Further define random variables $\hat{u}_i=dp\text{-}sign(u_i,\epsilon,\delta),\forall 1\leq i\leq M$. Then there always exist a constant σ_0 such that when $\sigma\geq\sigma_0$, $P(sign(\frac{1}{M}\sum_{m=1}^M\hat{u}_i)\neq sign(\frac{1}{M}\sum_{m=1}^M u_i))<\big[(1-x)e^x\big]^{\frac{M}{2}}$, where $x=\frac{|\sum_{m=1}^M u_m|}{\sigma M}$.*

The proof of Theorem 5 follows a similar strategy to that of Theorem 1. The difficulty we need to overcome is that unlike $sto\text{-}sign$, the expectation of the number of workers that share the wrong signs is not a function of $\frac{1}{M}\sum_{m=1}^M u_i$ due to the nonlinearity introduced by $\Phi(\cdot)$. However, when σ is large enough, we show that it can be upper bounded as a function of $\frac{1}{M}\sum_{m=1}^M u_i$.

Proof. Without loss of generality, assume $u_1\leq u_2\leq\cdots\leq u_K<0\leq u_{K+1}\leq\cdots\leq u_M$ and $\frac{1}{M}\sum_{i=1}^M u_i<0$. Note that similar analysis can be done when $\frac{1}{M}\sum_{i=1}^M u_i>0$. Further define a series of random variables $\{X_i\}_{i=1}^M$ given by

$$X_i=\begin{cases}1, & \text{if } \hat{u}_i\neq sign\big(\frac{1}{M}\sum_{i=1}^M u_i\big),\\ 0, & \text{if } \hat{u}_i=sign\big(\frac{1}{M}\sum_{i=1}^M u_i\big).\end{cases} \tag{49}$$

In particular, X_i can be considered as the outcome of one Bernoulli trial with successful probability $P(X_i=1)$. Let $Z=\sum_{i=1}^M X_i$ and we have

$$P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^M\hat{u}_i\bigg)\neq sign\bigg(\frac{1}{M}\sum_{m=1}^M u_i\bigg)\bigg)=P\bigg(Z\geq\frac{M}{2}\bigg). \tag{50}$$

In addition,

$$P(X_m=1)=\Phi\bigg(\frac{u_m}{\sigma}\bigg). \tag{51}$$

Then, Z follows the Poisson binomial distribution with mean and variance given by

$$\begin{aligned}\mu&=\sum_{m=1}^M P(X_m=1)=\sum_{m=1}^M\Phi\bigg(\frac{u_m}{\sigma}\bigg),\\ \sigma^2&=\sum_{m=1}^M\Phi\bigg(\frac{u_m}{\sigma}\bigg)\bigg(1-\Phi\bigg(\frac{u_m}{\sigma}\bigg)\bigg).\end{aligned} \tag{52}$$

Let n denote a zero-mean Gaussian noise with variance σ, according to the assumption that $u_1 \leq u_2 \leq \cdots \leq u_K < 0 \leq u_{K+1} \leq \cdots \leq u_M$, we have

$$\begin{aligned} \Phi\bigg(\frac{u_m}{\sigma}\bigg) &= \frac{1}{2} - P(u_m < n < 0), \;\; \forall 1 \leq m \leq K, \\ \Phi\bigg(\frac{u_m}{\sigma}\bigg) &= \frac{1}{2} + P(0 < n < u_m), \;\; \forall K+1 \leq m \leq M. \end{aligned} \tag{53}$$

Therefore,

$$\mu = \sum_{m=1}^{M} \Phi\bigg(\frac{u_m}{\sigma}\bigg) = \frac{M}{2} - \bigg[\sum_{m=1}^{K} P(u_m < n < 0) - \sum_{m=K+1}^{M} P(0 < n < u_m)\bigg]. \tag{54}$$

Note that for any Gaussian noise, $P(a_1 < n < 0) + P(a_2 < n < 0) \geq P(a_1 + a_2 < n < 0)$ for any $a_1 < 0, a_2 < 0$. Therefore, we consider the worst case scenario such that $\sum_{m=1}^{K} P(u_m < n < 0) - \sum_{m=K+1}^{M} P(0 < n < u_m)$ is minimized, i.e., $K = 1$. In this case,

$$\begin{aligned} &\sum_{m=1}^{K} P(u_m < n < 0) - \sum_{m=K+1}^{M} P(0 < n < u_m) \\ &= P\bigg(u_1 < n \leq -\sum_{m=2}^{M} u_m\bigg) + P\bigg(-\sum_{m=2}^{M} u_m < n < 0\bigg) - \sum_{m=2}^{M} P(0 < n < u_m) \\ &> \bigg|\sum_{m=1}^{M} u_m\bigg|\bigg[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{u_1^2}{2\sigma^2}}\bigg] + P\bigg(-\sum_{m=2}^{M} u_m < n < 0\bigg) - \sum_{m=2}^{M} P(0 < n < u_m) \\ &> \bigg|\sum_{m=1}^{M} u_m\bigg|\bigg[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{u_1^2}{2\sigma^2}}\bigg] - \bigg|\sum_{m=2}^{M} u_m\bigg| \frac{1}{\sqrt{2\pi}\sigma}\bigg[1 - e^{-\frac{(\sum_{m=2}^{M} u_m)^2}{2\sigma^2}}\bigg] \\ &= \frac{1}{\sqrt{2\pi}\sigma}\bigg[\bigg|\sum_{m=1}^{M} u_m\bigg| e^{-\frac{u_1^2}{2\sigma^2}} + \bigg|\sum_{m=2}^{M} u_m\bigg|\bigg[e^{-\frac{(\sum_{m=2}^{M} u_m)^2}{2\sigma^2}} - 1\bigg]\bigg], \end{aligned} \tag{55}$$

where the first inequality is due to $f(a) > f(u_1)$ for $a \in (u_1, \sum_{m=1}^{M} u_m]$ and the second inequality is due to $f(a) < \frac{1}{\sqrt{2\pi}\sigma}$ for any $a > 0$, where $f(\cdot)$ is the probability density function of the normal distribution.

In particular, as $\sigma \to \infty$, $|\sum_{m=1}^{M} u_m| e^{-\frac{u_1^2}{2\sigma^2}}$ increases and converges to $|\sum_{m=1}^{M} u_m|$ while $|\sum_{m=2}^{M} u_m| \bigg[e^{-\frac{(\sum_{m=2}^{M} u_m)^2}{2\sigma^2}} - 1\bigg]$ increases and converges to 0. Therefore, we have

$$\frac{1}{\sqrt{2\pi}\sigma}\bigg[\bigg|\sum_{m=1}^{M} u_m\bigg| e^{-\frac{u_1^2}{2\sigma^2}} + \bigg|\sum_{m=2}^{M} u_m\bigg|\bigg[e^{-\frac{(\sum_{m=2}^{M} u_m)^2}{2\sigma^2}} - 1\bigg]\bigg] \xrightarrow{\sigma \to \infty} -\frac{\sum_{m=1}^{M} u_m}{\sqrt{2\pi}\sigma}. \tag{56}$$

As a result, there exists a σ_0 such that when $\sigma \geq \sigma_0$, we have

$$\mu = \sum_{m=1}^{M} \Phi\bigg(\frac{u_m}{\sigma}\bigg) \leq \frac{M}{2} + \frac{\sum_{m=1}^{M} u_m}{2\sigma}. \tag{57}$$

Following the same analysis as that in the proof of Theorem 1, we can show that

$$P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} \hat{u}_i\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M} u_i\bigg)\bigg) < \big[(1-x)e^{x}\big]^{\frac{M}{2}}, \tag{58}$$

where $x = \frac{|\sum_{m=1}^{M} u_m|}{\sigma M}$. $\square$

2.6 Proof of Theorem 6

Theorem 6. *There exists a positive constant s_0 such that when $s > s_0$, Sto-SIGNSGD and DP-SIGNSGD can at least tolerate k_i Byzantine attackers on the i-th coordinate of the gradient and k_i satisfies*

$$k_i < \frac{|\sum_{m=1}^{M} \nabla f_m(w^{(t)})_i|}{s}, \quad \big[(1-x)e^x\big]^{\frac{M-k_i}{2}} < \frac{1}{2}, \tag{59}$$

where $x = \frac{|\sum_{m=1}^{M} \nabla f_m(w^{(t)})_i| - sk_i}{(M-k_i)s}$, $s = \sigma$ for DP-SIGNSGD and $s = b_i \geq max_m|(\boldsymbol{g}_m^{(t)})_i|$ for Sto-SIGNSGD.

Overall, the number of Byzantine workers that the algorithms can tolerate is given by $min_{1 \leq i \leq d} k_i$.

We first provide some intuition about the proof. It has been shown in the proof of Theorem 3 that the convergence of Sto-SIGNSGD and DP-SIGNSGD is guaranteed if the probability of more than half of the workers sharing wrong signs is less than 0.5. On the i-th coordinate of the gradient, if there are k_i Byzantine workers that always share the wrong signs, then at most $\frac{M-k_i}{2}$ normal workers can share wrong signs such that the aggregated result is still correct.

Proof. By replacing M with $M - k_i$ in (24), we can obtain

$$P\bigg(Z \geq \frac{M-k_i}{2}\bigg) \leq \frac{e^{(e^a-1)\mu}}{e^{\frac{(M-k_i)a}{2}}}, \tag{60}$$

where $\mu = \mathbb{E}[Z]$. It is shown in the proof of Theorem 1 and Theorem 4 that there exists a positive constant s_0 such that when $s > s_0$, $\mu \leq \frac{M}{2} - \frac{|\sum_{m=1}^{M} \nabla f_m(w^{(t)})_i|}{2s}$. Let $a = \ln(\frac{M-k_i}{2\mu})$, we have

$$P\bigg(Z \geq \frac{M-k_i}{2}\bigg) \leq \frac{e^{\frac{M-k_i}{2}-\mu}}{(\frac{M-k_i}{2\mu})^{\frac{M-k_i}{2}}} = \Bigg(\frac{e^{\frac{M-2\mu-k_i}{M-k_i}}}{(\frac{M-k_i}{2\mu})}\Bigg)^{\frac{M-k_i}{2}} \leq \big[(1-x)e^x\big]^{\frac{M-k_i}{2}}, \tag{61}$$

where $x = \frac{\frac{|\sum_{m=1}^{M} \nabla f_m(w^{(t)})_i|}{s} - k_i}{M-k_i}$. In addition, the above inequality requires $\ln(\frac{M-k_i}{2\mu}) > 0$ and therefore $k_i < \frac{|\sum_{m=1}^{M} \nabla f_m(w^{(t)})_i|}{s}$. $\square$

2.7 Proof of Theorem 7

Theorem 7. *Suppose Assumptions 1-4 are satisfied, and set the learning rate $\eta = \frac{1}{\sqrt{Td}}$. Then, when $\boldsymbol{b} = b \cdot \mathbf{1}$ and b is sufficiently large, Sto-SIGNSGD converges to the (local) optimum with a rate of $O(\frac{1}{\sqrt{T}})$ if either of the following two conditions is satisfied.*

- $P\big(sign(\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i) \neq sign(\nabla F(w^t)_i\big) < 0.5, \forall 1 \leq i \leq d.$
- *The mini-batch size of stochastic gradient at each iteration is at least T.*

Proof. Note that in the proof of Theorem 3, we obtain

$$\begin{aligned} &\mathbb{E}[F(w^{t+1}) - F(w^t)] \leq -\eta||\nabla F(w^t)||_1 + \frac{L\eta^2 d}{2} \\ &+ 2\eta\sum_{i=1}^{d}|\nabla F(w^t)_i|P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^t)_i\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(w^t)_i\bigg)\bigg), \end{aligned} \tag{62}$$

where $q(\boldsymbol{g}_m^t) = sto\text{-}sign(\boldsymbol{g}_m^t)$. For the ease of notation, let

$$p_{i,1} = P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^t)_i\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i\bigg)\bigg),$$
$$p_{i,2} = P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i\bigg)\bigg) < \frac{1}{2}, \tag{63}$$
$$p_{i} = P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^t)_i\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(w^t)_i\bigg)\bigg).$$

Then

$$p_i = p_{i,1}(1-p_{i,2}) + p_{i,2}(1-p_{i,1}) = p_{i,1} + p_{i,2} - 2p_{i,1}p_{i,2}. \tag{64}$$

We first prove the convergence under the first condition. According to Theorem 2, we have $p_{i,1} < \frac{1}{2}$. In this case, it can be verified that p_i is an increasing function of both $p_{i,1}$ and $p_{i,2}$ and therefore $p_i < \frac{1}{2} + \frac{1}{2} - \frac{1}{2} = \frac{1}{2}$. Following similar analysis to that in the proof of Theorem 3, it can be shown that Sto-SIGNSGD converges to the (local) optimum with a rate of $O(\frac{\sqrt{d}}{\sqrt{T}})$.

Then, we prove the convergence under the second condition. According to (64), it is obvious that $p_i \leq p_{i,1} + p_{i,2}$. Therefore, we have

$$\sum_{i=1}^{d}|\nabla F(w^t)_i|p_i \leq \sum_{i=1}^{d}|\nabla F(w^t)_i|p_{i,1} + \sum_{i=1}^{d}|\nabla F(w^t)_i|p_{i,2}. \tag{65}$$

In particular,

$$\begin{aligned}
p_{i,2} &= P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i\bigg)\bigg) \\
&\leq P\bigg(\bigg|\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i - \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i\bigg| \geq \bigg|\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i\bigg|\bigg) \\
&\leq \frac{\mathbb{E}[|\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i - \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i|]}{|\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i|} \\
&\leq \frac{\sqrt{\mathbb{E}[(\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i - \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{g}_m^t)_i)^2]}}{|\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(w^t)_i|} \\
&\leq \frac{\sigma_i}{\sqrt{MT}|\nabla F(w^t)_i|}.
\end{aligned} \tag{66}$$

As a result, the second term in (65) is bounded by $O(\frac{||\bar{\sigma}||_1}{\sqrt{MT}})$. Following the same analysis as that in the proof of Theorem 3, it can be shown that Sto-SIGNSGD converges with a rate of $O(\frac{1}{\sqrt{T}})$. □

2.8 Proof of Theorem 8

Theorem 8. *When Assumptions 1, 2 and 4 are satisfied, there exists a σ_0 such that when $\sigma \geq \sigma_0$, by running Algorithm 2 with $\eta = \frac{1}{M\sqrt{Td}}$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{||\nabla F(w^{(t)})||_2^2}{\sigma} \leq \frac{(F(w^{(0)}) - F^*)\sqrt{d}}{\sqrt{T}} + \frac{(1+L+L^2\beta)\sqrt{d}}{\sqrt{T}}, \tag{67}$$

where β is some positive constant.

The proof of Theorem 8 follows the strategy of taking $y^{(t)} = w^{(t)} - \tilde{\eta}\tilde{\boldsymbol{e}}^{(t)}$ such that $y^{(t)}$ is updated in the same way as $w^{(t)}$ in the non error-feedback scenario. A key technical challenge is to bound the norm of the residual error $||\tilde{\boldsymbol{e}}^{(t)}||_2^2$. Utilizing the fact that the output of the compressor $q(\cdot) \in \{-1, 1\}$, we upper bound it by first proving that in this case, the server's compressor is an α-approximate compressor [18] for some $\alpha < 1$. Therefore, before proving Theorem 8, we first prove the following lemmas.

Lemma 1. *Let $y^{(t)} = w^{(t)} - \eta M\tilde{\boldsymbol{e}}^{(t)}$, we have*

$$y^{(t+1)} = y^{(t)} - \eta \sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta). \tag{68}$$

Proof.

$$\begin{aligned}
y^{(t+1)} &= w^{(t+1)} - \eta M\tilde{\boldsymbol{e}}^{(t+1)} \\
&= w^{(t)} - \eta\tilde{\boldsymbol{g}}^{(t)} - \eta M\tilde{\boldsymbol{e}}^{(t+1)} \\
&= w^{(t)} - \eta\bigg(\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + M\tilde{\boldsymbol{e}}^{(t)} - M\tilde{\boldsymbol{e}}^{(t+1)}\bigg) - \eta M\tilde{\boldsymbol{e}}^{(t+1)} \\
&= w^{(t)} - \eta\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) - \eta M\tilde{\boldsymbol{e}}^{(t)} \\
&= y^{(t)} - \eta\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta).
\end{aligned} \tag{69}$$

□

Lemma 2. *There exists a positive constant $\beta > 0$ such that $\mathbb{E}[||\tilde{\boldsymbol{e}}^{(t)}||_2^2] \leq \beta d, \forall t$.*

Proof. We first prove that for the 1-bit compressor $q(\boldsymbol{g}_m^{(t)})$, there exists some constant α such that the following inequality always holds.

$$\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)} - \frac{1}{M}sign\bigg(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)}\bigg)\bigg|\bigg|_2^2 \leq (1-\alpha)\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)}\bigg|\bigg|_2^2, \tag{70}$$

where $\alpha < 1$ is some positive constant.

For the ease of presentation, we let $\boldsymbol{r}_i^{(t)}$ denote the i-th entry of $\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)}$. Then, we can rewrite the left hand side of (70) as follows,

$$\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)} - \frac{1}{M}sign\bigg(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)}\bigg)\bigg|\bigg|_2^2 = \sum_{i=1}^{d}\bigg(\boldsymbol{r}_i^{(t)} - \frac{1}{M}sign(\boldsymbol{r}_i^{(t)})\bigg)^2. \tag{71}$$

In particular, we have

$$\bigg(\boldsymbol{r}_i^{(t)} - \frac{1}{M}sign(\boldsymbol{r}_i^{(t)})\bigg)^2 = \bigg((\boldsymbol{r}_i^{(t)})^2 + \frac{1}{M^2} - \frac{2|\boldsymbol{r}_i^{(t)}|}{M}\bigg) = \bigg[1 - \frac{1}{M(\boldsymbol{r}_i^{(t)})^2}\bigg(2|\boldsymbol{r}_i^{(t)}| - \frac{1}{M}\bigg)\bigg](\boldsymbol{r}_i^{(t)})^2. \tag{72}$$

If $2|\boldsymbol{r}_i^{(t)}| - \frac{1}{M} > 0, \forall i$, then there exist a positive constant α such that

$$\sum_{i=1}^{d}\bigg(\boldsymbol{r}_i^{(t)} - \frac{1}{M}sign(\boldsymbol{r}_i^{(t)})\bigg)^2 \leq \sum_{i=1}^{d}(1-\alpha)(\boldsymbol{r}_i^{(t)})^2 = (1-\alpha)\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)}\bigg|\bigg|_2^2. \tag{73}$$

In order to prove that $2|\boldsymbol{r}_i^{(t)}| - \frac{1}{M} > 0, \forall i$, we first show that $M(\tilde{\boldsymbol{e}}^{(t)})_i$ is an even number for any t by induction. In particular, according to Assumption 4 and $(\tilde{\boldsymbol{e}}^{(0)})_i = 0$, $M\boldsymbol{r}_i^{(0)} = \sum_{m=1}^{M} q(\boldsymbol{g}_m^{(0)})_i$ is an odd number. Therefore, $M(\tilde{\boldsymbol{e}}^{(1)})_i = \sum_{m=1}^{M} q(\boldsymbol{g}_m^{(0)})_i - sign(\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})_i)$ is an even number. In addition,

$$M(\tilde{\boldsymbol{e}}^{(t+1)})_i = \sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})_i + M(\tilde{\boldsymbol{e}}^{(t)})_i - sign\bigg(\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})_i + M(\tilde{\boldsymbol{e}}^{(t)})_i\bigg). \tag{74}$$

Given that $M(\tilde{\boldsymbol{e}}^{(t)})_i$ is even, we can show that $M(\tilde{\boldsymbol{e}}^{(t+1)})_i$ is even as well. Therefore, $M\boldsymbol{r}_i^{(t)} = \sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})_i + M(\tilde{\boldsymbol{e}}^{(t)})_i$ is odd and $2|\boldsymbol{r}_i^{(t)}| \geq \frac{2}{M} > \frac{1}{M}, \forall t, i$.
Given (70), we can show that

$$\begin{aligned}
\mathbb{E}||\tilde{\boldsymbol{e}}^{(t+1)}||_2^2 &\leq (1-\alpha)\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)}) + \tilde{\boldsymbol{e}}^{(t)}\bigg|\bigg|_2^2 \\
&\leq (1-\alpha)(1+\rho)\mathbb{E}||\tilde{\boldsymbol{e}}^{(t)}||_2^2 + (1-\alpha)\bigg(1+\frac{1}{\rho}\bigg)\mathbb{E}\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})\bigg|\bigg|_2^2 \\
&\leq \sum_{j=0}^{t}[(1-\alpha)(1+\rho)]^{t-j}(1-\alpha)\bigg(1+\frac{1}{\rho}\bigg)\mathbb{E}\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})\bigg|\bigg|_2^2 \\
&\leq \frac{(1-\alpha)\Big(1+\frac{1}{\rho}\Big)d}{1-(1-\alpha)(1+\rho)},
\end{aligned} \tag{75}$$

where we invoke Young's inequality recurrently and ρ can be any positive constant. Therefore, there exists some constant $\beta > 0$ such that $\mathbb{E}[||\tilde{\boldsymbol{e}}^{(t)}||_2^2] \leq \beta d, \forall t$. □

Now, we are ready to prove Theorem 8.

Proof. Let $y^{(t)} = w^{(t)} - \eta M \tilde{\boldsymbol{e}}^{(t)}$, and $\tilde{\eta} = M\eta$, according to Lemma 1, we have

$$\begin{aligned}
&\mathbb{E}[F(y^{(t+1)}) - F(y^{(t)})] \\
&\leq -\tilde{\eta}\mathbb{E}\bigg[< \nabla F(y^{(t)}), \frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta) > \bigg] + \frac{L}{2}\mathbb{E}\bigg[\bigg|\bigg|\tilde{\eta}\frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta)\bigg|\bigg|_2^2\bigg] \\
&= \tilde{\eta}\mathbb{E}\bigg[< \nabla F(w^{(t)}) - \nabla F(y^{(t)}), \frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta) > \bigg] \\
&+ \frac{L\tilde{\eta}^2}{2}\mathbb{E}\bigg[\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta)\bigg|\bigg|_2^2\bigg] \\
&- \tilde{\eta}\mathbb{E}\bigg[< \nabla F(w^{(t)}), \frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta) > \bigg].
\end{aligned} \tag{76}$$

We first bound the first term, in particular, we have

$$\begin{aligned}
&< \nabla F(w^{(t)}) - \nabla F(y^{(t)}), \frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta) > \\
&\leq \frac{\tilde{\eta}}{2}||\frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta)||_2^2 + \frac{1}{2\tilde{\eta}}||\nabla F(w^{(t)}) - \nabla F(y^{(t)})||_2^2 \\
&\leq \frac{\tilde{\eta}}{2}||\frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta)||_2^2 + \frac{L^2}{2\tilde{\eta}}||y^{(t)} - w^{(t)}||_2^2 \\
&= \frac{\tilde{\eta}}{2}||\frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta)||_2^2 + \frac{L^2\tilde{\eta}}{2}||\tilde{\boldsymbol{e}}^{(t)}||_2^2 \\
&\leq \frac{\tilde{\eta}}{2}||\frac{1}{M}\sum_{m=1}^{M} \textit{dp-sign}(\boldsymbol{g}_m^{(t)};\epsilon,\delta)||_2^2 + \frac{L^2\tilde{\eta}\beta d}{2},
\end{aligned} \tag{77}$$

where the second inequality is due to the L-smoothness of F.

Then, we can bound the last term as follows.

$$
\begin{aligned}
&-\mathbb{E}\Big[<\nabla F(w^{(t)}), \frac{1}{M}\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)};\epsilon,\delta)>\Big] \\
&= -\mathbb{E}\Big[\sum_{i=1}^{d}\nabla F(w^{(t)})_i \frac{1}{M}\sum_{m=1}^{M} dp\text{-}sign((\boldsymbol{g}_m^{(t)})_i;\epsilon,\delta)\Big] \\
&= -\sum_{i=1}^{d}\nabla F(w^{(t)})_i \frac{1}{M}\sum_{m=1}^{M}\bigg(2\Phi\Big(\frac{\nabla f_m(w^{(t)})_i}{\sigma}\Big)-1\bigg) \\
&\leq -\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\frac{|\nabla F(w^{(t)})_i|}{\sigma} \\
&= -\frac{||\nabla F(w^{(t)})||_2^2}{\sigma},
\end{aligned}
\tag{78}
$$

where the inequality is due to (57) in the proof of Theorem 4.

Plugging (77) and (78) into (76) yields

$$
\begin{aligned}
&\mathbb{E}[F(y^{(t+1)})-F(y^{(t)})] \\
&\leq \frac{\tilde{\eta}^2+L\tilde{\eta}^2}{2}\mathbb{E}\bigg[\bigg|\bigg|\frac{1}{M}\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)};\epsilon,\delta)\bigg|\bigg|_2^2\bigg] + \frac{L^2\tilde{\eta}^2\beta d}{2} - \frac{\tilde{\eta}||\nabla F(w^{(t)})||_2^2}{\sigma} \\
&\leq \frac{(\tilde{\eta}^2+L\tilde{\eta}^2+L^2\tilde{\eta}^2\beta)d}{2} - \frac{\tilde{\eta}||\nabla F(w^{(t)})||_2^2}{\sigma}.
\end{aligned}
\tag{79}
$$

Rewriting (79) and taking average over $t=0,1,2,\cdots,T-1$ on both sides yields

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{||\nabla F(w^{(t)})||_2^2}{\sigma} \\
&\leq \sum_{t=0}^{T-1}\frac{\mathbb{E}[F(y^{(t)})-F(y^{(t+1)})]}{\tilde{\eta}T} + \frac{(\tilde{\eta}+L\tilde{\eta}+L^2\tilde{\eta}\beta)d}{2}.
\end{aligned}
\tag{80}
$$

Taking $\tilde{\eta}=\frac{1}{\sqrt{Td}}$ and $w^{(0)}=y^{(0)}$ yields

$$
\frac{1}{T}\sum_{t=0}^{T-1}\frac{||\nabla F(w^{(t)})||_2^2}{\sigma} \leq \frac{(F(w^{(0)})-F^*)\sqrt{d}}{\sqrt{T}} + \frac{(1+L+L^2\beta)\sqrt{d}}{\sqrt{T}}. \tag{81}
$$

$\square$

2.9 Proof of Theorem 9

Theorem 9. *At each iteration t, there exists a constant σ_0 such that when $\sigma>\sigma_0$, Algorithm 2 can at least tolerate $k_i=|\sum_{m=1}^{M}\nabla f_m(w^{(t)})_i|/\sigma$ Byzantine attackers on the i-th coordinate of the gradient. Overall, the number of Byzantine workers that Algorithm 2 can tolerate is given by $min_{1\leq i\leq d}k_i$.*

By following a similar strategy to the proof of Theorem 8 and taking the impact of Byzantine attackers into consideration, the convergence of Algorithm 2 in the presence of Byzantine attackers can be established.

Proof. Without loss of generality, assume that the first M workers are normal and the last B are Byzantine. Following a similar procedure to the proof of Theorem 8, we can show that

$$
\begin{aligned}
&\mathbb{E}[F(y^{(t+1)}) - F(y^{(t)})] \\
&\leq -\tilde{\eta}\mathbb{E}\bigg[< \nabla F(y^{(t)}), \frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg] > \bigg] \\
&+ \frac{L}{2}\mathbb{E}\bigg[\bigg|\bigg|\tilde{\eta}\frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg]\bigg|\bigg|_2^2\bigg] \\
&= \tilde{\eta}\mathbb{E}\bigg[< \nabla F(w^{(t)}) - \nabla F(y^{(t)}), \frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg] > \bigg] \\
&+ \frac{L\tilde{\eta}^2}{2}\mathbb{E}\bigg[\bigg|\bigg|\frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg]\bigg|\bigg|_2^2\bigg] \\
&- \tilde{\eta}\mathbb{E}\bigg[< \nabla F(w^{(t)}), \frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg] > \bigg].
\end{aligned} \tag{82}
$$

For the first term, we have

$$
\begin{aligned}
&< \nabla F(w^{(t)}) - \nabla F(y^{(t)}), \frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg] > \\
&\leq \frac{\tilde{\eta}}{2}\bigg|\bigg|\frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg]\bigg|\bigg|_2^2 \\
&+ \frac{1}{2\tilde{\eta}}||\nabla F(w^{(t)}) - \nabla F(y^{(t)})||^2 \\
&\leq \frac{\tilde{\eta}d}{2} + \frac{L^2}{2\tilde{\eta}}||y^{(t)} - w^{(t)}||^2 \\
&= \frac{\tilde{\eta}d}{2} + \frac{L^2\tilde{\eta}}{2}||\tilde{\boldsymbol{e}}^{(t)}||^2 \\
&\leq \frac{\tilde{\eta}d}{2} + \frac{L^2\tilde{\eta}\beta d}{2}.
\end{aligned} \tag{83}
$$

For the third term, if $B < \frac{|\sum_{m=1}^{M}(\boldsymbol{g}_m^{(t)})_i|}{\sigma}$, we have

$$
\begin{aligned}
&- \mathbb{E}\bigg[< \nabla F(w^{(t)}), \frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign(\boldsymbol{g}_m^{(t)}; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign(\boldsymbol{g}_j^{(t)})\bigg] > \bigg] \\
&= -\mathbb{E}\bigg[\sum_{i=1}^{d} \nabla F(w^{(t)})_i \frac{1}{M+B}\bigg[\sum_{m=1}^{M} dp\text{-}sign((\boldsymbol{g}_m^{(t)})_i; \epsilon, \delta) + \sum_{j=1}^{B} byzantine\text{-}sign((\boldsymbol{g}_j^{(t)})_i)\bigg] > \bigg] \\
&\leq -\sum_{i=1}^{d} |\nabla F(w^{(t)})_i| \frac{1}{M+B}\bigg[\frac{|\sum_{m=1}^{M}(\boldsymbol{g}_m^{(t)})_i|}{\sigma} - B\bigg] \\
&\leq -c||\nabla F(w^{(t)})||_1,
\end{aligned} \tag{84}
$$

where c is some positive constant.

Following the same analysis as that in the proof of Theorem 8, the convergence of Algorithm 2 can be established. □

3 Discussions about dp-$sign$ with $\delta = 0$

In this section, we present the differentially private compressor dp-$sign$ with $\delta = 0$.

Definition 4. *For any given gradient $\boldsymbol{g}_m^t$, the compressor* dp-$sign$ *outputs* dp-$sign(\boldsymbol{g}_m^t, \epsilon, 0)$. *In particular, the i-th entry of* dp-$sign(\boldsymbol{g}_m^t, \epsilon, 0)$ *is given by*

$$dp\text{-}sign(\boldsymbol{g}_m^t, \epsilon, 0)_i = \begin{cases} 1, & \textit{with probability } \frac{1}{2} + \frac{1}{2} sign((\boldsymbol{g}_m^t)_i)\big(1 - e^{-\frac{|(\boldsymbol{g}_m^t)_i|}{\lambda}}\big), \\ -1, & \textit{with probability } \frac{1}{2} - \frac{1}{2} sign((\boldsymbol{g}_m^t)_i)\big(1 - e^{-\frac{|(\boldsymbol{g}_m^t)_i|}{\lambda}}\big), \end{cases} \tag{85}$$

where $\lambda = \frac{\Delta_1}{\epsilon}$ and Δ_1 is the sensitivity measures defined in (15).

Theorem 10. *The proposed compressor* dp-$sign(\cdot, \epsilon, 0)$ *is $(\epsilon, 0)$-differentially private.*

Proof. Consider any vector $\boldsymbol{a}$ and $\boldsymbol{b}$ such that $||\boldsymbol{a} - \boldsymbol{b}||_1 \leq \Delta_1$ and $\boldsymbol{v} \in \{-1, 1\}^d$, we have

$$\frac{P(dp\text{-}sign(\boldsymbol{a}, \epsilon, 0) = \boldsymbol{v})}{P(dp\text{-}sign(\boldsymbol{b}, \epsilon, 0) = \boldsymbol{v})} = \frac{\int_D e^{-\frac{||\boldsymbol{x}-\boldsymbol{a}||}{\lambda}} d\boldsymbol{x}}{\int_D e^{-\frac{||\boldsymbol{x}-\boldsymbol{b}||}{\lambda}} d\boldsymbol{x}}, \tag{86}$$

where D is some integral area depending on $\boldsymbol{v}$. It can be verified that $e^{-\epsilon} \leq |\frac{e^{-\frac{||\boldsymbol{x}-\boldsymbol{a}||}{\lambda}}}{e^{-\frac{||\boldsymbol{x}-\boldsymbol{b}||}{\lambda}}}| \leq e^{\epsilon}$ always holds, which indicates that $e^{-\epsilon} \leq |\frac{P(dp\text{-}sign(\boldsymbol{a}, \epsilon, 0) = \boldsymbol{v})}{P(dp\text{-}sign(\boldsymbol{b}, \epsilon, 0) = \boldsymbol{v})}| \leq e^{\epsilon}$. □

Theorem 11. *Let $u_1, u_2, \cdots, u_M$ be M known and fixed real numbers. Further define random variables $\hat{u}_i =$ dp-$sign(u_i, \epsilon, \delta), \forall 1 \leq i \leq M$. Then there always exist a constant σ_0 such that when $\sigma \geq \sigma_0$, $P(sign(\frac{1}{M}\sum_{m=1}^M \hat{u}_i) \neq sign(\frac{1}{M}\sum_{m=1}^M u_i)) < \big[(1-x)e^x\big]^{\frac{M}{2}}$, where $x = \frac{|\sum_{m=1}^M u_m|}{\gamma\lambda M}$ and γ is some positive constant.*

Proof. Without loss of generality, assume $u_1 \leq u_2 \leq \cdots \leq u_K < 0 \leq u_{K+1} \leq \cdots \leq u_M$ and $\frac{1}{M}\sum_{i=1}^M u_i < 0$. Note that similar analysis can be done when $\frac{1}{M}\sum_{i=1}^M u_i > 0$. Further define a series of random variables $\{X_i\}_{i=1}^M$ given by

$$X_i = \begin{cases} 1, & \text{if } \hat{u}_i \neq sign\big(\frac{1}{M}\sum_{i=1}^M u_i\big), \\ 0, & \text{if } \hat{u}_i = sign\big(\frac{1}{M}\sum_{i=1}^M u_i\big). \end{cases} \tag{87}$$

In particular, X_i can be considered as the outcome of one Bernoulli trial with successful probability $P(X_i = 1)$. Let $Z = \sum_{i=1}^M X_i$ and we have

$$P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^M \hat{u}_i\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^M u_i\bigg)\bigg) = P\bigg(Z \geq \frac{M}{2}\bigg). \tag{88}$$

Z follows the Poisson binomial distribution with mean and variance given by

$$\begin{aligned} \mu &= \sum_{m=1}^M P(X_m = 1) = \frac{M}{2} - \bigg[\sum_{m=1}^K P(u_m < n < 0) - \sum_{i=K+1}^M P(0 < n < u_m)\bigg], \\ \sigma^2 &= \sum_{m=1}^M P(n > -u_m)(1 - P(n > -u_m)), \end{aligned} \tag{89}$$

where $n \sim Laplace(0, \lambda)$. Similar to the analysis for dp-$sign$ with $\delta > 0$, we can show that

$$
\begin{aligned}
&\sum_{m=1}^{K} P(u_m < n < 0) - \sum_{m=K+1}^{M} P(0 < n < u_m) \\
&= P\bigg(u_1 < n \leq -\sum_{m=2}^{M} u_m\bigg) + P\bigg(-\sum_{m=2}^{M} u_m < n < 0\bigg) - \sum_{m=2}^{M} P(0 < n < u_m) \\
&> \bigg|\sum_{m=1}^{M} u_m\bigg|\bigg[\frac{1}{2\lambda} e^{-\frac{|u_1|}{\lambda}}\bigg] + P\bigg(-\sum_{m=2}^{M} u_m < n < 0\bigg) - \sum_{m=2}^{M} P(0 < n < u_m) \\
&> \bigg|\sum_{m=1}^{M} u_m\bigg|\bigg[\frac{1}{2\lambda} e^{-\frac{|u_1|}{\lambda}}\bigg] - \bigg|\sum_{m=2}^{M} u_m\bigg|\frac{1}{2\lambda}\bigg[1 - e^{-\frac{|\sum_{m=2}^{M} u_m|}{\lambda}}\bigg] \\
&= \frac{1}{2\lambda}\bigg[\bigg|\sum_{m=1}^{M} u_m\bigg| e^{-\frac{|u_1|}{\lambda}} + \bigg|\sum_{m=2}^{M} u_m\bigg|\bigg[e^{-\frac{|\sum_{m=2}^{M} u_m|}{\lambda}} - 1\bigg]\bigg].
\end{aligned}
\tag{90}
$$

As a result, there exists a λ_0 such that when $\lambda \geq \lambda_0$, we have

$$
\mu = \sum_{m=1}^{M} P(X_m = 1) \leq \frac{M}{2} + \frac{\sum_{m=1}^{M} u_m}{2\lambda\gamma}, \tag{91}
$$

where γ is some constant larger than 1.

Following the same analysis as that in the proof of Theorem 1, we can show that there exists a positive constant M_0 such that when $M \geq M_0$

$$
P\bigg(sign\bigg(\frac{1}{M}\sum_{m=1}^{M} \hat{u}_m\bigg) \neq sign\bigg(\frac{1}{M}\sum_{m=1}^{M} u_m\bigg)\bigg) = P\bigg(Z \geq \frac{M}{2}\bigg) < \big[(1-x)e^{x}\big]^{\frac{M}{2}}, \tag{92}
$$

where $x = \frac{|\sum_{m=1}^{M} u_m|}{\gamma\lambda M}$ and γ is some positive constant. $\square$

4 Discussions about the server's compressor $\frac{1}{M} sign(\cdot)$ in Algorithm 2

Recall that the update rule of the residual error is given by

$$
\tilde{\boldsymbol{e}}^{t+1} = \frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^t) + \tilde{\boldsymbol{e}}^t - \frac{a}{M} sign\bigg(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^t) + \tilde{\boldsymbol{e}}^t\bigg), \tag{93}
$$

where $a = 1$ in the proposed Algorithm.

Theorem 12. *In Algorithm 2, when the total number of workers is odd and $a = \frac{1}{2k+1}$ for any non-negative integer k, the server's compressor $\frac{a}{M} sign(\cdot)$ is an α-approximate compressor for some $\alpha > 0$.*

Proof. The goal is to show that $\frac{a}{M} sign(\cdot)$ is an α-approximate compressor, i.e., for $\boldsymbol{r}^t = \frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^t) + \tilde{\boldsymbol{e}}^t$,

$$
\big|\big|\boldsymbol{r}^t - \frac{a}{M} sign(\boldsymbol{r}^t)\big|\big|_2^2 = \sum_{i=1}^{d}\bigg(\boldsymbol{r}_i^t - \frac{a}{M} sign(\boldsymbol{r}_i^t)\bigg)^2 \leq (1-\alpha)||\boldsymbol{r}^t||_2^2 = (1-\alpha)\sum_{i=1}^{d}(\boldsymbol{r}_i^t)^2, \tag{94}
$$

where $\boldsymbol{r}_i^t$ is the i-th entry of $\boldsymbol{r}^t$. In addition,

$$
\bigg(\boldsymbol{r}_i^t - \frac{a}{M} sign(\boldsymbol{r}_i^t)\bigg)^2 = \bigg((\boldsymbol{r}_i^t)^2 - \frac{2a|\boldsymbol{r}_i^t|}{M} + \frac{a^2}{M^2}\bigg) = \bigg(1 - \frac{a}{M(\boldsymbol{r}_i^t)^2}\bigg(2|\boldsymbol{r}_i^t| - \frac{a}{M}\bigg)\bigg)(\boldsymbol{r}_i^t)^2. \tag{95}
$$

It can be seen that a sufficient condition for (94) with some $\alpha > 0$ is given by

$$
2|\boldsymbol{r}_i^t| - \frac{a}{M} > 0, \forall 1 \leq i \leq d. \tag{96}
$$

Given that the total number of workers is odd, it is obvious that $(2k+1)M\boldsymbol{r}_i^0$ is an odd number. As a result, $(2k+1)M(\tilde{\boldsymbol{e}}^{(1)})_i = (2k+1)M\boldsymbol{r}_i^0 - sign(\boldsymbol{r}_i^0)$ is an even number. Similar to the proof of Lemma 2, it can be shown by induction that $(2k+1)M\boldsymbol{r}_i^t$ is odd for any t. Therefore,

$$2|\boldsymbol{r}_i^t| \geq \frac{2}{(2k+1)M} > \frac{a}{M}. \tag{97}$$

□

For $a \neq \frac{1}{2k+1}$, we consider four cases as follows.

Case 1: $a \geq 2$**.** In this case, since $(\tilde{\boldsymbol{e}}^0)_i = 0$, then if $\frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^0)_i = \frac{1}{M}$, (96) is not satisfied.

Case 2: $1 < a < 2$**.** In this case, we consider a sequence of $\{\frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^t)_i\}_{t=0}^n$, given by

$$\frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^t)_i = \frac{1}{M}, \forall 0 \leq t \leq n. \tag{98}$$

Then

$$(\tilde{\boldsymbol{e}}^{(1)})_i = \frac{1-a}{M},\ \ \boldsymbol{r}_i^{(1)} = \frac{2-a}{M}. \tag{99}$$

Suppose that $2|\boldsymbol{r}_i^t| > \frac{a}{M}, \forall t$. Now we show that given $(\tilde{\boldsymbol{e}}^t)_i = \frac{t-ta}{M}, \boldsymbol{r}_i^t = \frac{t+1-ta}{M}$, and $a < \frac{t+1}{t}$, we have $(\tilde{\boldsymbol{e}}^{t+1})_i = \frac{t+1-t+1a}{M}, \boldsymbol{r}_i^{t+1} = \frac{t+2-t+1a}{M}$ and $a < \frac{t+2}{t+1}$. To satisfy $2|\boldsymbol{r}_i^t| > \frac{a}{M}$, we have $a < \frac{2t+2}{2t+1} < \frac{t+2}{t+1}$. In addition, according to (93), $(\tilde{\boldsymbol{e}}^{t+1})_i = \boldsymbol{r}_i^t - \frac{a}{M}sign(\boldsymbol{r}_i^t) = \frac{t+1-t+1a}{M}$. Then, $\boldsymbol{r}_i^{t+1} = \frac{t+2-t+1a}{M}$. As a result, by induction, we can show that $a < \frac{n+1}{n}$. As n increases, a decreases and approaches 1.

Case 3: $\frac{1}{2} < a < 1$**.** Again, we consider the sequence of $\{\frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^t)_i\}_{t=0}^n$ given by (98). Then similarly, it can be shown that

$$(\tilde{\boldsymbol{e}}^t)_i = \frac{t-ta}{M}, \forall 1 \leq t \leq n+1. \tag{100}$$

Let $\frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^{(n+1)})_i = -\frac{1}{M}$, we have

$$\boldsymbol{r}_i^{(n+1)} = \frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^{(n+1)})_i + (\tilde{\boldsymbol{e}}^{(n+1)})_i = \frac{n-(n+1)a}{M}. \tag{101}$$

Taking $2|\boldsymbol{r}_i^{(n+1)}| > \frac{a}{M}$ yields

$$a \begin{cases} > \frac{2n}{2n+1} \geq \frac{n+1}{n+2}, & \text{if } a > \frac{n}{n+1}, \\ < \frac{2n}{2n+3}, & \text{if } a < \frac{n}{n+1}. \end{cases} \tag{102}$$

When $n=1$, $\frac{n}{n+1} = \frac{1}{2}$. According to (102), $a > \frac{n}{n+1}$ is a sufficient condition of $a > \frac{n+1}{n+2}$. Therefore, as n increases, a increases and approaches 1.

Case 4: $\frac{1}{2k+2} < a \leq \frac{1}{2k}$**, for any positive integer** k**.** Again, we consider the sequence of $\{\frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^t)_i\}_{t=0}^n$ given by (98). Similarly,

$$(\tilde{\boldsymbol{e}}^t)_i = \frac{t-ta}{M}, \forall 1 \leq t \leq n+1. \tag{103}$$

Let $\frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^{(n+1)})_i = -\frac{s}{M}$, then we have

$$\boldsymbol{r}_i^{(n+1)} = \frac{1}{M}\sum_{m=1}^M q(\boldsymbol{g}_m^{(n+1)})_i + (\tilde{\boldsymbol{e}}^{(n+1)})_i = \frac{(n+1-s)-(n+1)a}{M}. \tag{104}$$

First of all, let $n = s = 2k-1$, we have $\boldsymbol{r}_i^{(n+1)} = \frac{1-2ka}{M}$. Therefore, if $a = \frac{1}{2k}$, it is possible that $2|\boldsymbol{r}_i^{(n+1)}| = 0 \leq \frac{a}{M}$.

When $a < \frac{n+1-s}{n+1}$, $2|\boldsymbol{r}_i^{(n+1)}| > \frac{a}{M}$ yields

$$a < \frac{2(n+1-s)}{2n+3}. \tag{105}$$

Let $n = (2k+1) \times 2^y - 2$ and $s = 2k \times 2^y - 1$, where y is some non-negative integer, we have $\frac{n+1-s}{n+1} = \frac{2^y}{(2k+1)\times 2^y - 1}$. According to (105), $a < \frac{2^{y+1}}{(2k+1)\times 2^{y+1}-1}$. When $y = 0$, $\frac{n+1-s}{n+1} = \frac{2^y}{(2k+1)\times 2^y - 1} = \frac{1}{2k}$. As y (and therefore n and s) increases, a decreases and approaches $\frac{1}{2k+1}$.

When $a > \frac{n+1-s}{n+1}$, $2|\boldsymbol{r}_i^{(n+1)}| > \frac{a}{M}$ yields

$$a > \frac{2(n+1-s)}{2n+1}. \tag{106}$$

Let $n = (2k+1) \times 2^y$ and $s = 2k \times 2^y + 1$, where y is some non-negative integer, we have $\frac{n+1-s}{n+1} = \frac{2^y}{(2k+1)\times 2^y + 1}$. According to (106), $a > \frac{2^{y+1}}{(2k+1)\times 2^{y+1}+1}$. When $y = 0$, $\frac{n+1-s}{n+1} = \frac{2^y}{(2k+1)\times 2^y + 1} = \frac{1}{2k+2}$. As y (and therefore n and s) increases, a increases and approaches $\frac{1}{2k+1}$.

Remark 11. *By assuming that the total number of workers is odd, it is guaranteed that there is always a winner in the majority vote. When the number of workers is even, it is possible that $\boldsymbol{r}_i^t = 0$ and therefore (96) does not hold. This issue can be addressed if the server ignores the communication rounds (e.g., does not transmit anything) during which there is no winner in the majority vote.*

5 Details of the Implementation

Our experiments are mainly implemented using Python 3.7.4 with packages tensorflow 2.0 and numpy 1.16.5. One Intel i7-9700 CPU with 32 GB of memory and one NVIDIA GeForce RTX 2070 SUPER GPU are used in the experiments.

5.1 Dataset and Pre-processing

We perform experiments on the standard MNIST dataset for handwritten digit recognition consisting of 60,000 training samples and 10,000 testing samples.[3] Each sample is a 28×28 size gray-level image. We normalize the data by dividing it with the max RGB value (i.e., 255.0).

5.2 Dataset Assignment

In our experiments, we consider 31 normal workers and measure the data heterogeneity by the number of labels of data that each worker stores. We first partition the training dataset according to the labels. For each worker, we randomly generate a set of size n which indicates the labels of training data that should be assigned to this worker. Then, a subset of training data from the corresponding labels is randomly sampled and assigned to the worker without replacement. The size of the subset depends on n and the size of the training data for each label. More specifically, we set the size of the subset as $\lfloor 60000/(31n) \rfloor$ in the beginning. When there are not enough training data for a label, we reduce the size of the subset accordingly. We consider the scenarios that all the workers have the same n. For the results in the third figure in Fig. 2, we set $n = 1, 2, 4$ for "1 LABEL", "2 LABELS", "4 LABELS", respectively. For the rest of the results, we set $n = 1$.

5.3 Neural Network Setting

We implement a two-layer fully connected neural network with softmax of classes with cross-entropy loss. The hidden layer has 128 hidden ReLU units.

5.4 Learning Rate Tuning

We use a constant learning rate η and tune the parameters from the set $\{1, 0.1, 0.01, 0.005, 0.001, 0.0001\}$. In particular, for Sto-SIGNSGD and EF-Sto-SIGNSGD, we set $\eta = 0.005$; for DP-SIGNSGD and EF-DP-SIGNSGD, we set $\eta = 0.01$. For FedAvg [28], we tune the learning rate from the set $\{2, 1.5, 1, 0.1, 0.01, 0.005, 0.001, 0.0001\}$, the number of local epochs from the set $\{1, 10, 20, 30\}$ and present the best result.

[3]Avaiable at `http://yann.lecun.com/exdb/mnist/`

5.5 Additional Experimental Results

We report some additional experimental results for Sto-SIGNSGD and EF-Sto-SIGNSGD in Fig. 4 and Fig. 5. In particular, when there are no Byzantine attackers, we also compare our proposed algorithms with FedAvg [28] which is not Byzantine resilient. For FedAvg, we tune the learning rate from the set $\{2, 1.5, 1, 0.1, 0.01, 0.005, 0.001, 0.0001\}$, the number of local epochs from the set $\{1, 10, 20, 30\}$ and present the best result.

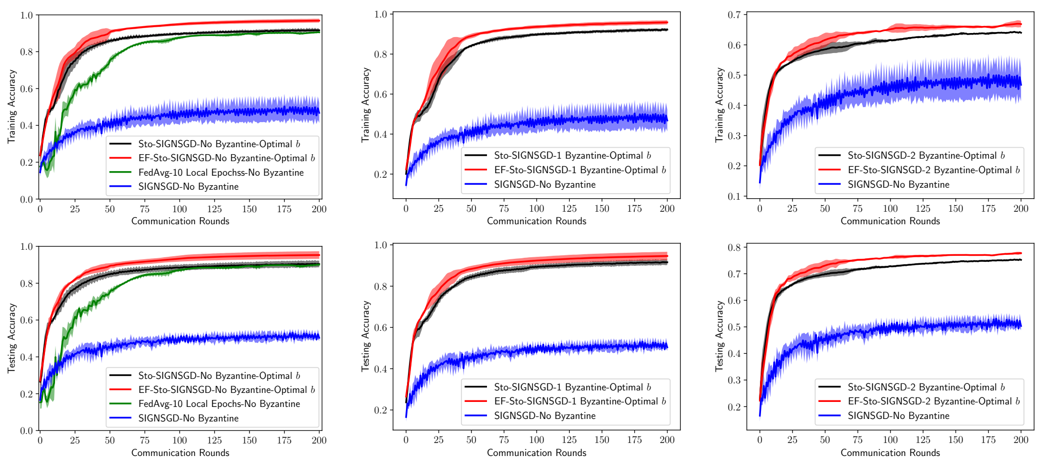


Figure 4: Experimental results showing the training and testing accuracy of Sto-SIGNSGD and EF-Sto-SIGNSGD for Optimal $\boldsymbol{b}$ and different number of Byzantine attackers. The solid curves represent the mean value and the shaded region spans one standard deviation obtained over five repetitions. EF-Sto-SIGNSGD outperforms Sto-SIGNSGD, SIGNSGD in all the examined scenarios. In addition, it can be observed that the proposed algorithms perform better than FedAvg while requires less communication overhead per round.

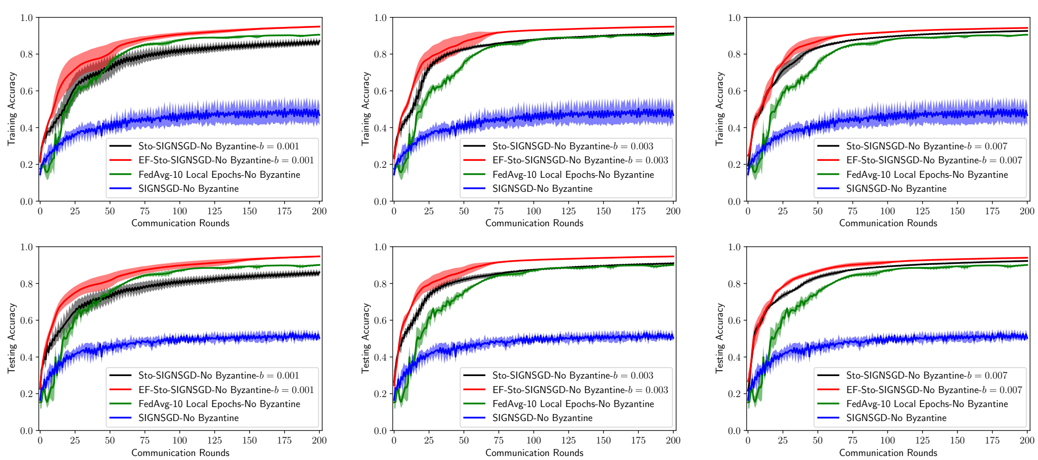


Figure 5: Experimental results showing the training and testing accuracy of Sto-SIGNSGD and EF-Sto-SIGNSGD for different $\boldsymbol{b} = b \cdot \mathbf{1}$. The solid curves represent the mean value and the shaded region spans one standard deviation obtained over five repetitions. EF-Sto-SIGNSGD outperforms Sto-SIGNSGD, SIGNSGD and FedAvg when there is no Byzantine attacker.

We report some additional experimental results for DP-SIGNSGD and EF-DP-SIGNSGD in Fig. 6.

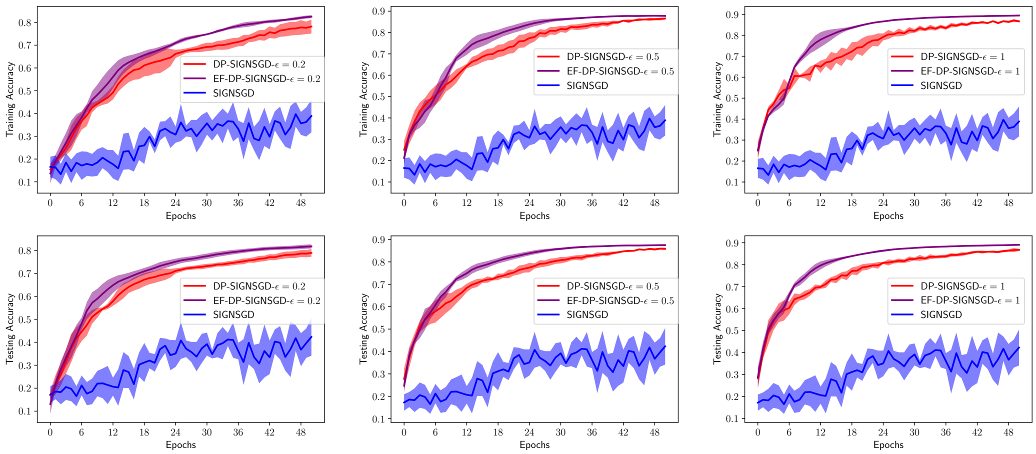


Figure 6: Experimental results showing the training and testing accuracy of DP-SIGNSGD and EF-DP-SIGNSGD for different ϵ when there is no Byzantine attacker. The solid curves represent the mean value and the shaded region spans one standard deviation obtained over five repetitions. The ϵ's measure the per epoch privacy guarantee of the algorithms. EF-DP-SIGNSGD outperforms DP-SIGNSGD and SIGNSGD.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[2] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, p. 44, 2017.

[3] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1709–1720.

[4] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," in *Advances in neural information processing systems*, 2017, pp. 1509–1519.

[5] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*, 2018, pp. 560–569.

[6] J. Wu, W. Huang, J. Huang, and T. Zhang, "Error compensated quantized SGD and its applications to large-scale distributed optimization," in *International Conference on Machine Learning*, 2018, pp. 5325–5333.

[7] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," in *Advances in Neural Information Processing Systems*, 2018, pp. 7564–7575.

[8] J. Bernstein, J. Zhao, K. Azizzadenesheli, and A. Anandkumar, "signSGD with majority vote is communication efficient and byzantine fault tolerant," in *In Seventh International Conference on Learning Representations (ICLR)*, 2019.

[9] X. Chen, T. Chen, H. Sun, Z. S. Wu, and M. Hong, "Distributed training with heterogeneous data: Bridging median and mean based algorithms," *arXiv preprint arXiv:1906.01736*, 2019.

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

[11] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Advances in Neural Information Processing Systems*, 2018, pp. 7652–7662.

[12] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," in *Advances in Neural Information Processing Systems*, 2018, pp. 2525–2536.

[13] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright, "ATOMO: Communication-efficient learning via atomic sparsification," in *Advances in Neural Information Processing Systems*, 2018, pp. 9850–9861.

[14] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[15] D. Carlson, Y. P. Hsieh, E. Collins, L. Carin, and V. Cevher, "Stochastic spectral descent for discrete graphical models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 296–311, 2015.

[16] S. U. Stich, J. B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Advances in Neural Information Processing Systems*, 2018, pp. 4447–4458.

[17] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Advances in Neural Information Processing Systems*, 2018, pp. 5973–5983.

[18] S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi, "Error feedback fixes signSGD and other gradient compression schemes," in *International Conference on Machine Learning*, 2019, pp. 3252–3261.

[19] S. Zheng, Z. Huang, and J. Kwok, "Communication-efficient distributed blockwise momentum SGD with error-feedback," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 446–11 456.

[20] H. Tang, X. Lian, T. Zhang, and J. Liu, "DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," *Proceedings of the 36th International Conference on Machine Learning-Volume 97*, pp. 6155–6165, 2019.

[21] P. Blanchard, R. Guerraoui, J. Stainer *et al.*, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 119–129.

[22] D. Alistarh, Z. Allen-Zhu, and J. Li, "Byzantine stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2018, pp. 4613–4623.

[23] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, 2018, pp. 5650–5659.

[24] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1544–1551.

[25] C. Xie, S. Koyejo, and I. Gupta, "SLSGD: Secure and efficient distributed on-device machine learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019.

[26] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[28] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[29] C. Xie, O. Koyejo, and I. Gupta, "Generalized byzantine-tolerant SGD," *arXiv preprint arXiv:1802.10116*, 2018.

[30] L. Xiang, J. Yang, and B. Li, "Differentially-private deep learning from an optimization perspective," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 559–567.

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