SignSGD with Federated Defense: Harnessing Adversarial Attacks through Gradient Sign Decoding

Chanho Park [1] **Namyoon Lee** [2]

Abstract

Distributed learning is an effective approach to accelerate model training by using parallel computing power of multiple workers. However, substantial communication delays arise between workers and a parameter server due to the massive costs associated with communicating gradients. SignSGD with majority voting (signSGD-MV) is a simple yet effective optimizer that reduces communication costs through sign quantization, but its convergence rate significantly decreases when adversarial workers arbitrarily manipulate datasets or local gradient updates. In this paper, we consider a distributed learning problem where the workforce comprises a mixture of honest and adversarial workers. In this setting, we show that the convergence rate can remain invariant as long as the number of honest workers providing trustworthy local updates to the parameter server exceeds the number of adversarial workers. The key idea behind this counter-intuitive result is our novel aggregation method, signSGD with federated defense (signSGD-FD). Unlike traditional approaches, signSGD-FD utilizes the gradient information sent by adversarial workers with appropriate weights, obtained through gradient sign decoding. Experimental results demonstrate that signSGD-FD achieves superior convergence rates compared to traditional algorithms in various adversarial attack scenarios.

1. Introduction

Distributed stochastic gradient descent (SGD) stands as a widely adopted technique for tackling large-scale optimization challenges involving data parallelism (Bottou, 2010; Dean et al., 2012). Theoretically, synchronous distributed SGD has the potential to significantly boost the training speed of high-dimensional models in proportion to the number of workers. Nonetheless, the practical implementation of such distributed SGD encounters hurdles, notably the substantial communication costs associated with exchanging gradient information between a central server and individual workers. This communication expense scales linearly with the number of workers. For instance, the cutting-edge large language models (Zhao et al., 2023), using a few billion parameters, necessitate an exchange of tens of Giga bytes of information per iteration for both each worker and the central server. This renders it impractical for distributed training, particularly in the environments with limited communication networks. To resolve this communication bottleneck issue, it becomes imperative to devise communication-efficient distributed learning algorithms that can mitigate communication costs while upholding high learning performance.

In recent years, various techniques have been proposed with the goal of effectively reducing the communication load in distributed learning (Kairouz et al., 2021). The primary approach to cost reduction involves edge devices performing lossy compression on locally computed gradient, which are then transmitted to the parameter server. One simple yet effective algorithm is signSGD with majority voting (signSGD-MV) (Bernstein et al., 2018), in which each worker quantizes the locally computed stochastic gradient with only sign information and sends it to the server. Then, the server aggregates the one-bit gradient information using the MV principle and shares the aggregated one with the workers for performing the model update.

Adversarial attacks aim to interfere with the training models in distributed machine learning systems, introducing security vulnerabilities in their predictive outcomes (Lyu et al., 2020; Baruch et al., 2019; Xie et al., 2020; Blanchard et al., 2017; Alistarh et al., 2018). For instance, malevolent workers seek to compromise the model by introducing inaccurate data or manipulating the model's parameters or gradients. SignSGD-MV has also shown to effective to optimize the model against adversarial attack thanks to the majority vot-

[1]Department of Electrical Engineering, POSTECH, Pohang, South Korea [2]School of Electrical Engineering, Korea University, Seoul, South Korea. Correspondence to: Namyoon Lee <namyoon@korea.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ing principle (Bernstein et al., 2019). Nevertheless, the test accuracy of signSGD-MV deteriorates rapidly as the number of adversarial workers increases.

In this paper, we theoretically and empirically prove that the performance of distributed learning remains unaffected by the number of adversarial workers, as long as the number of honest workers exceeds that of adversarial workers. We summarize our main contributions as follows:

- The key idea showing our counter-intuitive result is a novel distributed learning algorithm called signSGD with *federated defense* (signSGD-FD). Diverging from the traditional majority voting approach, federated defense astutely utilizes gradient information derived not only from honest workers but also from adversarial workers during aggregation. To elucidate this concept, we offer a coding-theoretical interpretation of signSGD-MV. Building upon this novel interpretation, we introduce a progressive weighted majority voting (WMV) method that dynamically adjusts weights throughout iterations. During each iteration, the server estimates the weights of workers by comparing signs between the aggregated gradient through WMV and the local gradient transmitted by the workers. This weight estimation process not only aids in identifying adversarial workers but also leverages the weights to enhance resilience against adversarial attacks.

- We present a unified convergence rate analysis for the signSGD-style algorithm incorporating an arbitrary binary aggregation function. Specifically, when employing signSGD-FD, we demonstrate that the convergence rate remains unaffected in the face of adversarial attacks, as long as the number of adversarial workers is less than that of honest workers. This finding diverges from the previous convergence rate observed with signSGD-MV, where the convergence rate diminishes with an increasing number of adversarial workers.

- We also provide experimental results on MNIST, CIFAR-10 and CIFAR-100 datasets to validate the robustness of signSGD-FD in the presence of malicious attacks. Compared to signSGD-MV and its variants, signSGD-FD can achieve much higher test accuracy in the presence of stochastic sign flip attacks, especially on the $r = 1$ case. Furthermore, we validate the communication efficiency of signSGD-FD by evaluating the communication costs compared to other full-precision attack-robust algorithms.

2. Related Works

In this section, we explain prior works that are closely relevant to our research. Additionally, we provide a brief explanation of the weighted majority voting method, which is a core concept in our proposed *federated defense* aggregation.

Gradient compression: Gradient compression techniques can be categorized into quantization, which compresses the gradient vector into a limited set of codewords, and sparsification, which selectively updates a small number of gradient coordinates to optimize models. Noteworthy quantization methods encompass (Seide et al., 2014; Alistarh et al., 2017; Bernstein et al., 2018; Gandikota et al., 2021; Hönig et al., 2022), while sparsification techniques include (Aji & Heafield, 2017; Wangni et al., 2018; Stich et al., 2018; Rothchild et al., 2020; Li & Hoefler, 2022). To significantly reduce communication costs, some approaches, as exemplified by (Wen et al., 2017; Basu et al., 2019; Sattler et al., 2019; Park & Lee, 2023b; Li & Li, 2023), integrate both quantization and sparsification. Various adaptations of the signSGD-MV algorithm, such as those proposed in (Karimireddy et al., 2019; Jin et al., 2020; Sun et al., 2023; Jin et al., 2024), have been introduced to address additional practical challenges.

Robustness to adversarial attacks: A decentralized learning system is susceptible to malicious attacks, as adversarial attackers can engage in the system. A common attack method in distributed learning is the Byzantine attack (Lamport et al., 1982). To counteract such threats, various defense algorithms modify the aggregation process which traditionally involves computing the average of workers' gradients. These algorithms include coordinate-wise median (Yin et al., 2018), geometric median (Blanchard et al., 2017; Guerraoui et al., 2018), center clipping (Karimireddy et al., 2021), and weighted aggregation (Pillutla et al., 2022). The signSGD-MV algorithm is recognized as an attack-robust solution, as it remains unaffected by attacks on gradient magnitudes. Consequently, several studies (Bernstein et al., 2019; Chen et al., 2020; Sohn et al., 2020; Jin et al., 2020) have focused on enhancing robustness. However, recent developments include new attack methods that can bypass these defense mechanisms (Baruch et al., 2019; Xie et al., 2020), along with the emergence of other backdoor attacks (Bagdasaryan et al., 2020).

Weighted majority voting (WMV): The WMV method has found widespread application across various domains, such as communication systems (Hong et al., 2017; Kim et al., 2019), crowdsourcing (Li & Yu, 2014; Kim et al., 2023a), and ensemble learning (Berend & Kontorovich, 2015; Kim et al., 2023b). Notably, in scenarios involving the transmission of binary information through parallel binary symmetric channels (BSCs), it is established that the WMV decoder, incorporating log-likelihood ratio (LLR) weights, stands as the optimal choice from a maximum likelihood estimation perspective (Jeon et al., 2018). Furthermore, the utility of WMV extends into the realm of federated learning,

where it addresses challenges arising from the heterogeneity of data distribution (Wu & Wang, 2021; Li et al., 2023), as well as mitigating degradation caused by adversarial attacks (Jin et al., 2020). Despite these applications, there remains a notable gap in the literature: specifically, the application of LLR weights on signSGD-based learning algorithms has yet to be explored or studied, to the best of our knowledge.

3. Preliminaries

In this section, we briefly review the classical signSGD-MV algorithm (Bernstein et al., 2018). We also present an adversarial attack mechanism for the signSGD-MV algorithm, in which adversarial attackers flip the stochastic signs of workers' gradients.

3.1. SignSGD-MV

We consider a distributed learning system that consists of one central server and M workers who have their own datasets $\mathcal{D}_m, \forall m \in \mathcal{M}$, where $\mathcal{M}$ is the set of participating workers. These local datasets are the subsets of the global dataset $\mathcal{D}$ with $\mathcal{D} = \bigcup_{m\in\mathcal{M}} \mathcal{D}_m$. Then, the optimization problem of the distributed learning is given by

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) := \arg\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{M} \sum_{m\in\mathcal{M}} f_m(\mathbf{x}), \quad (1)$$

where $f_m(\mathbf{x}) = \mathbb{E}_{\mathbf{d}\sim\mathcal{D}_m}\left[F\left(\mathbf{x};\mathbf{d}\right)\right]$ is a local loss function for the worker $m \in \mathcal{M}$. Since the data samples are distributed among several workers, the worker m trains its model to be suitable only for the own dataset $\mathcal{D}_m$. For given model parameter $\mathbf{x}^t$ at iteration t, the worker m computes the stochastic gradient with batch size B_m as

$$\mathbf{g}_m^t := \frac{1}{B_m} \sum_{\mathbf{d}\in\mathcal{B}_m^t} \nabla f_m\left(\mathbf{x}^t;\mathbf{d}\right) \in \mathbb{R}^N, \quad (2)$$

where $\mathcal{B}_m^t \subset \mathcal{D}_m$. Then, the worker performs one-bit sign quantization for the locally computed stochastic gradient, $\mathsf{sign}\left(\mathbf{g}_m^t\right)$. Subsequently, the sign of the gradient is sent to the server through a band-limited communication network. The server performs the aggregation for $\mathsf{sign}\left(\mathbf{g}_m^t\right)$ using majority voting rule as $\mathsf{sign}\left[\sum_{m\in\mathcal{M}} \mathsf{sign}\left(\mathbf{g}_m^t\right)\right]$ (Bernstein et al., 2018). Then, the server sends the signs of the aggregated gradient to all workers. Lastly, each worker updates the model as follows:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \delta \cdot \mathsf{sign}\left[\sum_{m\in\mathcal{M}} \mathsf{sign}\left(\mathbf{g}_m^t\right)\right], \quad (3)$$

where $\delta \in \mathbb{R}^+$ is a fixed learning rate parameter. This process is repeated until the model converges, and we call this optimization process as *signSGD-MV*.

3.2. Adversarial Attacks

Adversarial attacks in distributed learning give rise to a substantial threat to the accuracy of training models. This paper specifically focuses on the adversarial attack scenario within the framework of signSGD-MV. Our primary emphasis is on the black-box setting, a scenario wherein attackers are unable to access any datasets from honest workers. This particular setting holds practical significance in numerous distributed learning environments. In this setting, we consider two types of adversarial attacks.

Sign inversion attack: Building upon the earlier research conducted by (Bernstein et al., 2019) and (Jin et al., 2020), a potential adversarial attack can be envisioned, wherein the signs of locally computed gradient are inverted. This particular attack scenario is denoted as the sign-inversion attack (SIA).

Stochastic sign flip attack: Inspired by the *Gaussian Byzantine* attack discussed in (Blanchard et al., 2017), an alternative attack mechanism involves the stochastic flipping of the sign information associated with one-bit gradient information. This attack is hereby referred to as the stochastic sign flip attack (SSFA). Specifically, we characterize the stochastic sign flip attack by introducing a sign-flipping probability parameter, denoted as $r \in [0,1]$. The set of compromised workers who have been subjected to adversarial attacks is represented by $\mathcal{L} \subset \mathcal{M}$, where L denotes the cardinality of the set. To elaborate, we define the stochastic sign flip attack method as follows: for each compromised worker $\ell \in \mathcal{L}$, the sign information of the nth coordinate is stochastically flipped with probability r. This can be expressed as follows:

$$\mathsf{sign}\left(\tilde{g}_{\ell,n}^t\right) = \begin{cases} \mathsf{sign}\left(g_{\ell,n}^t\right), & \text{w.p. } 1-r \\ -\mathsf{sign}\left(g_{\ell,n}^t\right), & \text{w.p. } r \end{cases}, \quad (4)$$

where $\tilde{g}_{\ell,n}^t$ is the modified gradient by the adversarial attacks. Notably, the SSFA boils down to the SIA when $r = 1$.

3.3. Coding-Theoretical Interpretation of signSGD-MV

Upper bound of signSGD-MV: Exploring the learning performance of signSGD-MV becomes particularly insightful when we analyze its upper bound. This upper bound is realized when two key conditions are met: i) all workers $m \in \mathcal{M}$ collaboratively utilize the complete local datasets, i.e., $\mathcal{D}_m = \mathcal{D}$, and ii) the gradient computation involves the use of the full-batch size as $B_m = |\mathcal{D}_m|$ for all $m \in \mathcal{M}$. Under these two ideal cases, every worker is empowered to compute the true gradient $\bar{\mathbf{g}}^t$ at each iteration t as

$$\bar{\mathbf{g}}^t = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{d}\in\mathcal{D}} \nabla F\left(\mathbf{x}^t;\mathbf{d}\right). \quad (5)$$

We denote the true sign of nth coordinate as

$$U_n^t = \mathsf{sign}\left(\bar{g}_n^t\right) \tag{6}$$

for $n \in [N]$. Then, the worker updates the model as

$$x_n^{t+1} = x_n^t - \delta \cdot U_n^t. \tag{7}$$

The model update technique utilizing the true gradient exhibits a more rapid convergence rate compared to the conventional signSGD-MV, which relies on one-bit stochastic gradient information. From this ideal case, it is important to decode the true sign of the gradient in every iteration to speed up the convergence rate. We call this true sign U_n^t for all coordinate $n \in [N]$ as *message bits* to be recovered from the aggregation.

Communication model: We now interpret the one-bit stochastic gradient computation process as a communication process. Under the limited data knowledge $\mathcal{D}_m \subset \mathcal{D}$ and batch sizes $B_m \leq |\mathcal{D}_m|$, worker $m \in \mathcal{M}$ computes the sign of the stochastic gradient as

$$Y_{m,n}^t = \mathsf{sign}\left(g_{m,n}^t\right). \tag{8}$$

This stochastic sign is different from the true sign of the gradient U_n^t. We model this sign mismatch effect with a lens through a communication problem. To be specific, all workers send the true sign of the gradient U_n^t through M parallel *binary symmetric channels* (BSCs). Then, the server receives M independent noisy message bits, i.e., $\mathbf{Y}_n^t = \left[Y_{1,n}^t, \cdots, Y_{M,n}^t\right]$. In this modeling, the cross-over probability of the BSC, $p_{m,n}^t$ is defined as

$$p_{m,n}^t = \mathbb{P}\left[Y_{m,n}^t \neq U_n^t\right]. \tag{9}$$

After the server receives all the workers' gradient signs, the server decodes the observations $\mathbf{Y}_n^t$ by using an arbitrary aggregation function $A\left(\mathbf{Y}_n^t\right)$, and we denote the decoded sign as $\hat{U}_n^t$. To mimic the upper bound performance, it is important to design the aggregation function that minimizes the decoding error probability:

$$p_{\mathsf{E},n}^t = \mathbb{P}\left[\hat{U}_n^t \neq U_n^t\right]. \tag{10}$$

Effect of the adversarial attacks in the BSC model: Utilizing our communication model, we offer insights into how adversarial attacks impact the alteration of cross-over probabilities in BSCs. The SIA simply alters the sign of the locally computed gradient among certain workers. Subsequently, when worker $\ell \in \mathcal{L}$ is subjected to the SIA, we can express the effect as an equivalent adjustment in the cross-over probability. Specifically, the probability $p_{\ell,n}^t$ is transformed to $\tilde{p}_{\ell,n}^t = 1 - p_{\ell,n}^t$ for all $n \in [N]$. The SSFA introduces a probability parameter $r \in [0,1]$, and modifies the locally computed gradient according to (4) by changing its sign with this probability. This operation can be conceptualized as incorporating an additional BSC in a cascade fashion. Specifically, let the flipped sign of the stochastic gradient be denoted as $\tilde{Y}_{\ell,n}^t$ for the coordinate $n \in [N]$. Then, the overall cross-over probability of the two consecutively connected BSCs can be derived as:

$$\tilde{p}_{\ell,n}^t = \mathbb{P}\left[\tilde{Y}_{\ell,n}^t \neq U_n^t\right] = p_{\ell,n}^t + r\left(1 - 2p_{\ell,n}^t\right), \tag{11}$$

which can also express that of SIA. Consequently, both SIA and SSFA mechanism can be understood by the changes in the cross-over probabilities of effective BSCs. This interpretation facilitates the establishment of a unified convergence rate applicable to these attack scenarios.

4. SignSGD with Federated Defense

In this section, we put forth a novel distributed learning algorithm called signSGD with *federated defense* (signSGD-FD). SignSGD-FD and conventional signSGD-MV share identical algorithm procedure, differing solely in the gradient aggregation method employed at the server. Consequently, we shall focus on explaining the aggregation technique applied to the gradient sign information $\mathbf{Y}_n^t, \forall n \in [N]$.

4.1. Algorithm

Under the premise that the cross-over probabilities of all workers, $p_{1,n}^t, \cdots, p_{M,n}^t$, are perfectly known at the server, the optimal aggregation method is to perform the maximum likelihood (ML) decoding. To accomplish ML decoding, the server computes the log-likelihood ratio (LLR) as

$$\ln \frac{\mathbb{P}\left[\mathbf{Y}_n^t \middle| U_n^t = +1\right]}{\mathbb{P}\left[\mathbf{Y}_n^t \middle| U_n^t = -1\right]} = \sum_{m \in \mathcal{M}} \ln \frac{1 - p_{m,n}^t}{p_{m,n}^t} Y_{m,n}^t. \tag{12}$$

As a result, the optimal aggregation boils down to the weighted majority voting (WMV) as

$$\hat{U}_n^t = \mathsf{sign}\left(\sum_{m \in \mathcal{M}} w_{m,n}^t Y_{m,n}^t\right), \tag{13}$$

where $w_{m,n}^t = \ln \frac{1-p_{m,n}^t}{p_{m,n}^t}$ is the nth coordinate LLR weight for the worker $m \in \mathcal{M}$. Unfortunately, obtaining the true $p_{m,n}^t$ is an insurmountable task due to the server's inability to access all the data samples from workers. Nevertheless, we can estimate these probabilities from the federated defense mechanism. The key idea behind federated defense lies in leveraging decoding results to estimate the cross-over probabilities $p_{m,n}^t$ over iterations. To elucidate, during the initial phase $t \leq T_{\mathsf{in}}$, the server employs an empirical approach to estimate the probability of computing errors by counting the instances of sign errors across all coordinates

$n \in [N]$, and comparing them with the decoding results as

$$\hat{p}_{m,n}^{t+1} = \frac{\sum_{i=1}^{t} \sum_{n=1}^{N} \mathbf{1}_{\left[Y_{m,n}^{i} \neq \hat{U}_{n}^{i}\right]}}{Nt}, \tag{14}$$

where $\mathbf{1}_{[\cdot]}$ is an indicator function. Here, the decoding results are obtained by the WMV decoding, but with the estimated LLR weights as

$$\hat{U}_{\mathsf{FD},n}^{t} = \mathsf{sign}\left(\sum_{m \in \mathcal{M}} \hat{w}_{m,n}^{t} Y_{m,n}^{t}\right). \tag{15}$$

From the estimated cross-over probabilities in (14), the LLR weight $\hat{w}_{m,n}^{t+1}$ is updated as

$$\hat{w}_{m,n}^{t+1} = \ln \frac{1 - \hat{p}_{m,n}^{t+1}}{\hat{p}_{m,n}^{t+1}}. \tag{16}$$

After the initial phase $t > T_{\mathsf{in}}$, the cross-over probability estimation rule is changed to recursively update $\hat{p}_{m,n}^{t}$ in parallel for each coordinate as

$$\hat{p}_{m,n}^{t+1} = \frac{T_{\mathsf{in}}}{t} \hat{p}_{m,n}^{T_{\mathsf{in}}} + \frac{t - T_{\mathsf{in}}}{t} \frac{\sum_{i=T_{\mathsf{in}}+1}^{t} \mathbf{1}_{\left[Y_{m,n}^{i} \neq \hat{U}_{n}^{i}\right]}}{t - T_{\mathsf{in}}}, \tag{17}$$

where $\hat{p}_{m,n}^{T_{\mathsf{in}}}$ is computed by following (14). Consequently, the model is updated as

$$\mathbf{x}^{t+1} = \mathbf{x}^{t} - \delta \cdot \hat{U}_{\mathsf{FD},n}^{t}. \tag{18}$$

The entire algorithm is summarized in Algorithm 1.

Algorithm 1 signSGD-FD

Input: Initial model $\mathbf{x}^1$, the number of workers M, worker m's batch size B_m, learning rate δ, initial weight $\hat{w}_{m,n}^{1} = 1$, initial phase duration T_{in}, total iteration T

for $t = 1 : T$ **do**

  for each worker $m \in \mathcal{M}$ **do**
    Compute $\mathbf{g}_m^t$ with batch size B_m
    Encode $Y_{m,n}^{t} = \mathsf{sign}\left(g_{m,n}^{t}\right), \forall n \in [N]$
    Send $\mathsf{sign}\left(\mathbf{g}_m^t\right) = \left[Y_{m,1}^{t}, \cdots, Y_{m,N}^{t}\right]$ to **server**
  end for

  for each worker $\ell \in \mathcal{L}$ **attackers do**
    Manipulate $\tilde{Y}_{\ell,n}^{t} = \begin{cases} Y_{\ell,n}^{t}, & \text{w.p. } 1-r \\ -Y_{\ell,n}^{t}, & \text{w.p. } r \end{cases}, \forall n \in [N]$
    Send $\mathsf{sign}\left(\tilde{\mathbf{g}}_\ell^t\right) = \left[\tilde{Y}_{\ell,1}^{t}, \cdots, \tilde{Y}_{\ell,N}^{t}\right]$ to **server**
  end for

  for each coordinate $n = 1 : N$ **server do**
    Decode
    $\hat{U}_n^t = \mathsf{sign}\left(\sum_{m \in \mathcal{M} \setminus \mathcal{L}} \hat{w}_{m,n}^{t} Y_{m,n}^{t} + \sum_{\ell \in \mathcal{L}} \hat{w}_{\ell,n}^{t} \tilde{Y}_{\ell,n}^{t}\right)$
    for $m \in \mathcal{M}$ **do**
      Estimate
      $\hat{p}_{m,n}^{t} = \begin{cases} \frac{\sum_{i=1}^{t} \sum_{n=1}^{N} \mathbf{1}_{\left[Y_{m,n}^{i} \neq \hat{U}_{n}^{i}\right]}}{Nt}, & \text{if } t \leq T_{\mathsf{in}} \\ \frac{\sum_{i=1}^{t} \mathbf{1}_{\left[Y_{m,n}^{i} \neq \hat{U}_{n}^{i}\right]}}{t}, & \text{o.w.} \end{cases}$
      Update $\hat{w}_{m,n}^{t} = \ln \frac{1 - \hat{p}_{m,n}^{t}}{\hat{p}_{m,n}^{t}}$
    end for
    Send $\hat{U}_n^t$ to **all workers** $m \in \mathcal{M}$
  end for

  for each worker $m \in \mathcal{M}$ **do**
    Update $x_n^{t+1} = x_n^t - \delta \cdot \hat{U}_n^t, \forall n \in [N]$
  end for

end for

4.2. Remarks

Universality for adversarial attacks: The proposed signSGD-FD exhibits versatility in addressing a broad spectrum of adversarial attack scenarios. Its adaptability is particularly evident in the nuanced modeling of cross-over probabilities based on distinct adversarial attack mechanisms. In the context of the SIA scenario, the cross-over probability undergoes a transformation from $p_{\ell,n}^{t}$ to $\tilde{p}_{\ell,n}^{t} = 1 - p_{\ell,n}^{t}$ when worker $\ell \in \mathcal{L}$ is targeted. Meanwhile, in the SSFA scenario, which can generalize the SIA method, the cross-over probability is precisely modeled as the expression given by (11). A noteworthy aspect of the proposed FD is its independence from any prior knowledge regarding adversarial attack scenarios. Notably, it eliminates the need to estimate the sign-flipping probability r under the SSFA. The only requisite information is the estimation of effective cross-over probabilities $\hat{p}_{m,n}^{t}$ in (17). Consequently, our signSGD-FD emerges as a versatile solution, applicable across diverse adversarial attack scenarios.

Harnessing compromised workers: Our signSGD-FD algorithm leverages the gradients of compromised workers for aggregation. Specifically, following the cross-over probability estimation process in (14) and (17), the server can pinpoint the compromised workers by identifying workers whose cross-over probabilities exceed 1/2. Typically, during gradient aggregation, the server can eliminate the local gradients from these identified adversarial workers. However, our federated defense mechanism demonstrates that this elimination strategy is notably sub-optimal. To achieve optimal ML decoding performance, it is crucial to utilize the cross-over probabilities of all workers. These probabilities are imperative because the sign of the estimated LLR weights of compromised workers can automatically change if $\hat{p}_{m,n}^{t} > 1/2$. This counter-intuitive result will be verified from the convergence analysis in the subsequent section.

5. Convergence Analysis

In this section, we provide the convergence analysis for signSGD-FD in the presence of stochastic sign flip attacks.

5.1. Assumptions

Before analyzing the convergence guarantee, we first present some assumptions used in the analysis:

Assumption 5.1 (Lower bound)**.** For all $\mathbf{x} \in \mathbb{R}^N$ and a global minimum point $\mathbf{x}^\star$, we have an objective value as

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) = f^\star. \tag{19}$$

Assumption 5.2 (Coordinate-wise smoothness)**.** For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, there exists a vector with non-negative constants $\mathbf{L} = [L_1, \ldots, L_N]$ that satisfies

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \sum_{n=1}^{N} \frac{L_n}{2} (y_n - x_n)^2. \tag{20}$$

Assumption 5.1 is required for the convergence to local minima, and Assumption 5.2 indicates the Lipschitz condition of the objective function. Assumptions 5.1 and 5.2 are commonly used for the convergence analysis of learning algorithms as in (Li et al., 2019; Bernstein et al., 2018), but with a coordinate-wise fashion.

5.2. Convergence Analysis without Attacks

Under these mild assumptions, the convergence rate of signSGD using an arbitrary binary aggregation can be derived as the theorem below.

Theorem 5.3 (Universal convergence rate)**.** *Let $\hat{U}_n^t = A(\mathbf{Y}_n^t) \in \{-1, +1\}$ be a decoded gradient sign for nth coordinate at iteration t. We define the maximum of sign decoding error probability over all $n \in [N]$ and $t \in [T]$ as*

$$P_{\mathsf{E}}^{\mathsf{max}} = \max_{n \in [N], t \in [T]} \mathbb{P}\left[\hat{U}_n^t \neq U_n^t\right]. \tag{21}$$

With a fixed learning parameter $\delta = \sqrt{\frac{2(f^1 - f^\star)}{T \|\mathbf{L}\|_1}}$, the convergence rate of signSGD-type algorithms is given by

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\bar{\mathbf{g}}^t\|_1\right] \leq \frac{1}{1 - 2P_{\mathsf{E}}^{\mathsf{max}}} \sqrt{\frac{2(f^1 - f^\star)\|\mathbf{L}\|_1}{T}}, \tag{22}$$

for $P_{\mathsf{E}}^{\mathsf{max}} < \frac{1}{2}$.

The convergence rate in Theorem 5.3 holds for an arbitrary sign decoding function $A(\mathbf{Y}_n^t)$. From Theorem 5.3, we observe that the convergence rates of signSGD-style algorithms, including signGD, have an order $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$. More importantly, the convergence rate of signSGD-based algorithms improves as decreasing the maximum decoding error probability. From this observation, our focus shifts to establishing an upper bound for the decoding error probability when applying the proposed signSGD-FD algorithm.

Theorem 5.4 (Decoding error bound of signSGD-FD)**.** *For every $n \in [N]$, $m \in \mathcal{M}$, and $t \in [T]$, suppose the ratio between the estimated and true LLR weights are bounded with some constants $\delta_{\mathsf{max}} \in \mathbb{R}^+$ and $\delta_{\mathsf{min}} \in \mathbb{R}^+$ as*

$$1 - \delta_{\mathsf{min}} \leq \frac{\hat{w}_{m,n}^t}{w_{m,n}^t} \leq 1 + \delta_{\mathsf{max}}. \tag{23}$$

Then, the gradient sign decoding error probability when applying the FD aggregation in (15) is upper bounded by

$$P_{\mathsf{E}}^{\mathsf{FD}} \leq \exp\left[-M\left(\frac{1 - \delta_{\mathsf{min}}}{1 + \delta_{\mathsf{max}}}\right)\gamma_{\mathcal{M}}^{\mathsf{WMV}}\right], \tag{24}$$

where $\gamma_{\mathcal{M}}^{\mathsf{WMV}} = \frac{1}{M}\sum_{m \in \mathcal{M}} \frac{1}{2}\left(\frac{1}{2} - p_{m,n}^t\right) \ln \frac{1 - p_{m,n}^t}{p_{m,n}^t}$ is the error exponent of the perfect WMV aggregation for the entire worker set $\mathcal{M}$.

From Theorem 5.4, we observe that the decoding error bound exponentially decreases with the number of workers M. The error exponent $\gamma_{\mathcal{M}}^{\mathsf{WMV}}$ determines how quickly the error probability diminishes as increasing M. In addition, the performance loss of FD aggregation due to weight uncertainty results in an error exponent reduction of $\frac{1 - \delta_{\mathsf{min}}}{1 + \delta_{\mathsf{max}}} \leq 1$. This observation confirms that the accurate $\hat{p}_{m,n}^t$ estimation helps to reduce the decoding error probability.

Theorem 5.5 (Decoding error bound of signSGD-MV)**.** *Suppose the server performs the MV aggregation in (3), i.e., $\hat{U}_{\mathsf{MV},n}^t = \mathsf{sign}\left(\sum_{m \in \mathcal{M}} Y_{m,n}^t\right)$. Then, the decoding error probability is upper bounded by*

$$P_{\mathsf{E}}^{\mathsf{MV}} \leq \exp\left(-M\gamma_{\mathcal{M}}^{\mathsf{MV}}\right), \tag{25}$$

where $\gamma_{\mathcal{M}}^{\mathsf{MV}} = \bar{p}_n^t - \frac{1}{2}\ln\left(2e\bar{p}_n^t\right)$ is the error exponent of the MV aggregation, and $\bar{p}_n^t = \frac{1}{M}\sum_{m \in \mathcal{M}} p_{m,n}^t$ is the average of workers' cross-over probabilities.

The error bound of MV decoder established in Theorem 5.5 follows an exponentially decreasing trend with M, same with the FD aggregation. The error exponent $\gamma_{\mathcal{M}}^{\mathsf{MV}}$ is expressed as the average of workers' computing error probabilities, and this term determines the decoding performance.

5.3. Upper Bounds of Decoding Errors under Attacks

We establish the upper bounds of the decoding error probability of signSGD-FD under the SSFA. When the SSFA is considered, as explained in (11), the effective cross-over probability $\tilde{p}_{\ell,n}^t$ increases, which leads to an increase of decoding error probability. The theorem below elucidates the deterioration of the FD aggregation in the presence of SSFAs.

Theorem 5.6 (Decoding error bound for signSGD-FD under the SSFA)**.** *Suppose the server performs the FD aggregation, and the workers in $\mathcal{L} \subset \mathcal{M}$ with cardinality L are contaminated by the stochastic sign flip attacks with probability r. Then, the decoding error probability is upper bounded by*

$$P_{\mathsf{E}}^{\mathsf{FD}} \leq \exp\left[-(M-L)\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{WMV}}\right], \quad (26)$$

where $\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{WMV}} = \frac{1}{M-L}\left[M\gamma_{\mathcal{M}}^{\mathsf{WMV}} - L\gamma_{\mathcal{L}}^{\mathsf{WMV}} + \sum_{\ell\in\mathcal{L}} g\left(\tilde{p}_{\ell,n}^{t}\right)\right]$ is the modified error exponent of the FD aggregation with a specific function $g(p) = \frac{1}{2}\left(\frac{1}{2}-p\right)\ln\frac{1-p}{p}$.

Compared to Theorem 5.4, Theorem 5.6 shows how the decoding error bound of signSGD-FD changes by the SSFA with probability r. First, the effective number of workers decreases from M to $M-L$, which can increase the decoding error bound. However, $\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{WMV}}$ is greater than $\gamma_{\mathcal{M}}^{\mathsf{WMV}}$, which leads to the decrease of the decoding error bound. To provide a more clear intuition on the two opposite effects, we provide the following corollary for some special cases of the sign flip probability.

Corollary 5.7 (Special case)**.** *Under the stochastic sign flip attacks with the sign flip probability $r \in \{1, \frac{1}{2}\}$, the decoding error probability is upper bounded by*

$$P_{\mathsf{E}}^{\mathsf{WMV}} \leq \begin{cases} \exp\left[-M\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\gamma_{\mathcal{M}}^{\mathsf{WMV}}\right], & \text{if } r=1 \\ \exp\left[-(M-L)\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\gamma_{\mathcal{M}}^{\mathsf{WMV}}\right], & \text{if } r=\frac{1}{2} \end{cases}. \quad (27)$$

Corollary 5.7 clearly shows that the decoding error probability bound does not change under the SSFA with the sign flip probability $r = 1$, i.e., the SIA. This result is counter-intuitive because the adversarial attacks have shown the degradation of the convergence rate. However, our result confirms that the convergence rate remains unchanged if we sagaciously harness the gradient information sent by the adversarial workers with proper weights. It is also noted that when $r = \frac{1}{2}$, the exponent term decreases from M to $M-L$, which slows down the convergence rate. From this result, we also observe that the worst-case attack scenario is to use the sign flip probability of $r = \frac{1}{2}$. This observation aligns with our intuition that the uniformly random gradient sign flipping does not allow the server to exploit the gradient sign information for decoding.

To better appreciate the distinctions, it is informative to juxtapose the decoding error bound of signSGD-MV within the framework of SSFA with r against that of signSGD-FD.

Theorem 5.8 (Decoding error bound of signSGD-MV)**.** *Suppose L workers are under the stochastic sign flip attacks with sign flip probability r. Then, the decoding error probability of the signSGD-MV algorithm is upper bounded by*

$$P_{\mathsf{E}}^{\mathsf{MV}} \leq \exp\left[-(M-2rL)\,\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{MV}}\right], \quad (28)$$

where $\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{MV}} = \frac{1}{M-2rL}\left[M\gamma_{\mathcal{M}}^{\mathsf{MV}} - 2rL\gamma_{\mathcal{L}}^{\mathsf{MV}} + \epsilon_{\mathcal{L}}(r)\right]$ is the modified error exponent of MV aggregation expressed with a sufficiently small parameter $\epsilon_{\mathcal{L}}(r)$.

Theorem 5.8 shows that the decoding error probability worsens as the sign flip probability r increases. Unlike our signSGD-FD method, when $r = 1$, the convergence rate of signSGD-MV is significantly degraded by SSFA because the decoding error bound increases considerably. Nevertheless, signSGD-MV can achieve the identical convergence rate with signSGD-FD under the SSFA when $r = \frac{1}{2}$. As a result, signSGD-FD provides a theoretical guarantee of superior convergence rate than signSGD-MV under the SSFA when $r > \frac{1}{2}$. We also empirically observe from numerical experiments that signSGD-FD provides a faster convergence rate than signSGD-MV under the SSFA, even when $r = \frac{1}{2}$.

6. Experiments

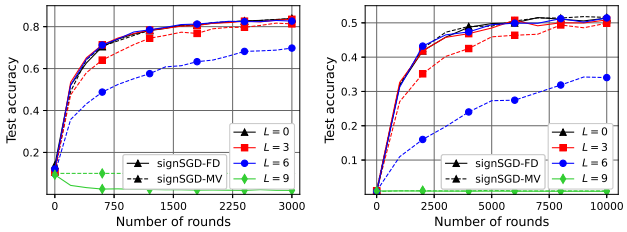
This section presents the experimental results on the image classification to verify the learning performance of signSGD-FD compared to other distributed learning algorithms.

6.1. Settings

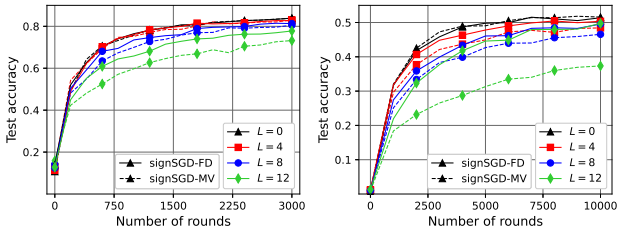
Datasets & Training models. The real-world datasets used for image classification simulation are MNIST (LeCun et al., 1998), CIFAR-10, and CIFAR-100 (Krizhevsky & Hinton, 2009) datasets. For the learning models, we adopt a convolutional neural network (CNN) (LeCun et al., 1998) model for the MNIST dataset, and a ResNet-56 (He et al., 2016) model for the CIFAR-10 and CIFAR-100 datasets. The number of workers M is fixed to 15, and more details about simulation settings are deferred to Appendix A.

Adversarial attacks. To evaluate signSGD-FD in the presence of attacks, we introduce the SIA and SSFA method in Section 3.2. The sign-flipping probability r is set between 0 and 1, especially focusing on the $r = 1$ which is the SIA case and the $r = \frac{1}{2}$ case. Referring to (Karimireddy et al., 2021), we fix the set of workers $\mathcal{L}$ compromised by the attacks during the entire learning process.

Initial phase of signSGD-FD. An additional setting required for signSGD-FD is the aggregation in the initial phase. Appendix B provides test accuracy comparisons according to the initial phase duration T_{in} and the estimation method of cross-over probabilities, and the results in the subsequent sections are based on the settings which achieves the best performance.

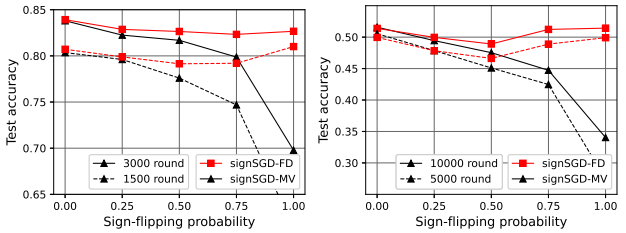


(a) CIFAR-10 with $r = 1$ (b) CIFAR-100 with $r = 1$



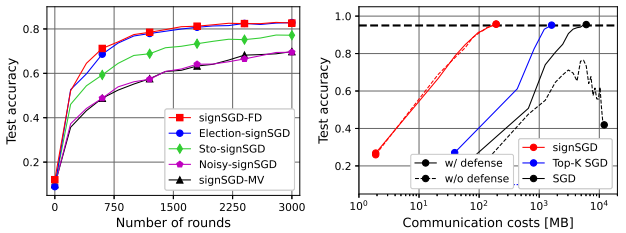
(c) CIFAR-10 with $r = \frac{1}{2}$ (d) CIFAR-100 with $r = \frac{1}{2}$

Figure 1. Test accuracy vs. training rounds varying the number of compromised workers L.



(a) CIFAR-10 dataset (b) CIFAR-100 dataset

Figure 2. Test accuracy vs. training rounds varying sign-flipping probability r.



(a) signSGD-based algorithms (b) Commun. costs comparison

Figure 3. Test accuracy & communication costs comparison among the attack-robust distributed learning algorithms.

6.2. Effects of SIA $(r = 1)$

In Figure 1-(a) and (b), we provide the test accuracy assessment of signSGD-FD on the effect of SIA. The general trend of results is that signSGD-MV deteriorates significantly as the number of compromised workers L increases, while signSGD-FD can achieve almost the same accuracy as in the absence of attacks if $L < \frac{M}{2}$. These results convince us that the LLR weights used in the FD aggregation can provide excellent protection of the training models against the SIA method. Meanwhile, it can be seen that signSGD-FD fails to converge in the $L = 9$ case. This can be considered that the decoding error probability becomes greater than $\frac{1}{2}$ in this case, making it no longer possible to perform the accurate $\hat{p}^t_{m,n}$ estimation. These observations are consistent with the analyses in Section 5 about the impact of $r = 1$ attacks, and even the uncertainty of weights.

6.3. Effects of SSFA $\left(r = \frac{1}{2}\right)$

We also evaluate signSGD-FD in the presence of SSFA with $r = \frac{1}{2}$ in Figure 1-(c) and (d). The test accuracy results show that signSGD-FD still be more robust to the sign flip attacks with $r = \frac{1}{2}$ than signSGD-MV. Nevertheless, signSGD-FD becomes to degrade gradually as the number of compromised workers L increases, which aligns with our analyses in Corollary 5.7. A notable point is that both algorithms can learn the models in $L = 12$ case, unlike the SIA case, since the effective number of workers for both algorithms becomes $M - L > 0$.

6.4. Degradation Trend According to r

The degradation trend according to the sign-flipping probability r is illustrated in Figure 2 with $L = 6$. The results for both datasets have the same trend that the test accuracy loss becomes severer as r increases. On the other hand, signSGD-FD achieves the best accuracy when $r = 0$ or 1, and the worst accuracy for the $r = \frac{1}{2}$ attack, but not that serious. These results have already predicted through the channel capacity analysis on SSFA.

6.5. Algorithms Comparison

We finally compare the robustness against attacks by comparing the test accuracy of signSGD-style algorithms, depicted in Figure 3-(a). Here, we set $L = 6$ with the SIA method for this comparison. The details about the compared algorithms are presented in Appendix A. The comparison results verify that the proposed signSGD-FD algorithm can achieve the highest test accuracy compared to other signSGD-style robust optimizers. Moreover, other algorithms require susceptible parameter settings to obtain high accuracy, but signSGD-FD is not so affected by these settings, which can be observed in Appendix B.

Figure 3-(b) shows each algorithm's communication costs on the MNIST dataset. In this comparison, we adopt *Gaussian Byzantine* attack (Blanchard et al., 2017), and also apply *Multi-Krum* algorithm (Blanchard et al., 2017) on

SGD and Top-K SGD optimizers to be robust against attacks. Appendix A illustrates the details of these algorithms. The results demonstrate that signSGD-type algorithms can significantly reduce communication costs by 30x compared to the Multi-Krum algorithm. We can even observe the robustness of signSGD, showing that signSGD-MV can achieve the same accuracy as signSGD-FD. Top-K SGD-based multi-Krum requires 3x fewer costs than SGD but still needs 10x more costs than our proposed algorithm, which reveals the superiority of signSGD-FD.

7. Conclusion

In this paper, we have demonstrated that the convergence rate of the signSGD-type algorithm remains invariant as long as the number of adversarial workers is lower than that of honest workers. This counterintuitive result is achieved through a novel optimizer, signSGD-FD, which incorporates the concept of federated defense. Federated defense employs learnable weights for weighted majority voting during aggregation. The server dynamically learns these weights based on the reliability estimation of the transmitted local gradient information from the workers. These weights are then used to decode the sign of aggregated local gradients, minimizing sign decoding errors. We have also provided a unified convergence rate analysis framework applicable to various adversarial attack scenarios. Experimental results demonstrate that signSGD-FD outperforms traditional signSGD-MV, showcasing a faster convergence rate, especially in the presence of adversarial attacks.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) Grant through the Ministry of Science and ICT (MSIT), Korea Government, under Grant 2022R1A5A1027646, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT), under Grant RS-2023-00208552.

Impact Statement

This paper presents a work that aims to advance the field of distributed learning under adversarial attacks. Present-day deep neural networks (DNNs) exhibit susceptibility to adversarial attacks, wherein maliciously crafted perturbations applied to input data or model weights can alter or manipulate the classification outcomes. Previously, adversarial attacks have been shown to decrease distributed learning performance. In this paper, however, we show that the distributed learning performance can remain unchanged under some adversarial attacks if we harness the information from adversarial workers in the learning process. This new principle of exploiting adversarial workers can impact the development of many other distributed learning algorithms resilient to adversarial attacks.

References

Aji, A. F. and Heafield, K. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processsing Systems (NeurIPS)*, volume 30, pp. 1709–1720, Long Beach, CA, 2017.

Alistarh, D., Allen-Zhu, Z., and Li, J. Byzantine stochastic gradient descent. In *Advances in Neural Information Processsing Systems (NeurIPS)*, volume 31, Montréal, Canada, 2018.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligenec and Statistics (AISTATS)*, volume 108, pp. 2938–2948, Online, 2020.

Baruch, G., Baruch, M., and Goldberg, Y. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, Vancouver, Canada, 2019.

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processsing Systems (NeurIPS)*, volume 32, pp. 14695–14706, Vancouver, Canada, 2019.

Berend, D. and Kontorovich, A. A finite sample analysis of the naive bayes classifier. *Journal of Machine Learning Research*, 16(1):1519–1545, 2015.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 560–569, Stockholm, Sweden, 2018.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. signSGD with majority vote is communication efficient and fault tolerant. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, 2019.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processsing Systems*, volume 30, Long Beach, CA, 2017.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT)*, pp. 177–186, Paris, France, 2010.

Chen, X., Chen, T., Sun, H., Wu, S. Z., and Hong, M. Distributed training with heterogeneous data: Bridging median-and mean-based algorithms. In *Advances in Neural Information Processsing Systems (NeurIPS)*, volume 33, pp. 21616–21626, Online, 2020.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q., and Ng, A. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pp. 1223–1231, Lake Tahoe, NV, 2012.

Gandikota, V., Kane, D., Maity, R. K., and Mazumdar, A. vqSGD: Vector quantized stochastic gradient descent. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2197–2205, Online, 2021.

Guerraoui, R., Rouault, S., et al. The hidden vulnerability of distributed learning in Byzantium. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3521–3530, Stockholm, Sweden, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, 2016.

Hong, S.-N., Kim, S., and Lee, N. A weighted minimum distance decoding for uplink multiuser MIMO systems with low-resolution ADCs. *IEEE Transactions on Communications*, 66(5):1912–1924, 2017.

Hönig, R., Zhao, Y., and Mullins, R. DAdaQuant: Doubly-adaptive quantization for communication-efficient federated learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 8852–8866, Baltimore, MD, 2022.

Jeon, Y.-S., Lee, N., Hong, S.-N., and Heath, R. W. One-bit sphere decoding for uplink massive MIMO systems with one-bit ADCs. *IEEE Transactions on Wireless Communications*, 17(7):4509–4521, 2018.

Jin, R., Huang, Y., He, X., Dai, H., and Wu, T. Stochastic-sign SGD for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

Jin, R., Liu, Y., Huang, Y., He, X., Wu, T., and Dai, H. Sign-based gradient descent with heterogeneous data: Convergence and Byzantine resilience. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. Early Access.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error feedback fixes signSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3252–3261, Long Beach, CA, 2019.

Karimireddy, S. P., He, L., and Jaggi, M. Learning from history for Byzantine robust optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5311–5319, Online, 2021.

Kearns, M. and Saul, L. Large deviation methods for approximate probabilistic inference. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, Madison, WI, 1998.

Kim, D., Hong, S.-N., and Lee, N. Supervised-learning for multi-hop MU-MIMO communications with one-bit transceivers. *IEEE Journals on Selected Areas in Communications*, 37(11):2559–2572, 2019.

Kim, D., Lee, J., and Chung, H. W. A worker-task specialization model for crowdsourcing: Efficient inference and fundamental limits. *IEEE Transactions on Information Theory*, 2023a. Early Access.

Kim, Y., Shin, J., Cassuto, Y., and Varshney, L. R. Distributed boosting classification over noisy communication channels. *IEEE Journals on Selected Areas in Communications*, 41(1):141–154, 2023b.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Canada, 2009.

Lamport, L., Shostak, R., and Pease, M. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, H. and Yu, B. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*, 2014.

Li, S. and Hoefler, T. Near-optimal sparse allreduce for distributed deep learning. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*, pp. 135–149, Seoul, South Korea, 2022.

Li, X. and Li, P. Analysis of error feedback in federated non-convex optimization with biased compression: Fast convergence and partial participation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 19638–19688, Honolulu, HI, 2023.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Li, Z., Lin, T., Shang, X., and Wu, C. Revisiting weighted aggregation in federated learning with neural networks. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 19767–19788, Honolulu, HI, 2023.

Lyu, L., Yu, H., and Yang, Q. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020.

Park, C. and Lee, N. S^3GD-MV: Sparse-SignSGD with majority vote for communication-efficient distributed learning. In *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pp. 2266–2271, Taipei, Taiwan, 2023a.

Park, C. and Lee, N. Sparse-SignSGD with majority vote for communication-efficient distributed learning. *arXiv preprint arXiv:2302.07475*, 2023b.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

Rothchild, D., Panda, A., Ullah, E., Ivkin, N., Stoica, I., Braverman, V., Gonzalez, J., and Arora, R. FetchSGD: Communication-efficient federated learning with sketching. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 8253–8265, Online, 2020.

Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2019.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Proceedings of the 15th Annual Conference of the International Speech Communications Association (INTERSPEECH)*, pp. 1058–1062, Singapore, 2014.

Sohn, J.-y., Han, D.-J., Choi, B., and Moon, J. Election coding for distributed learning: Protecting signSGD against Byzantine attacks. In *Advances in Neural Information Processsing Systems (NeurIPS)*, volume 33, pp. 14615–14625, Online, 2020.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with memory. In *Advances in Neural Information Processsing Systems (NeurIPS)*, volume 31, pp. 4448–4459, Montréal, Canada, 2018.

Sun, T., Wang, Q., Li, D., and Wang, B. Momentum ensures convergence of signsgd under weaker assumptions. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 33077–33099, Honolulu, HI, 2023.

Wangni, J., Wang, J., Liu, J., and Zhang, T. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processsing Systems (NeurIPS)*, volume 31, pp. 1299–1309, Montréal, Canada, 2018.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. TernGrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processsing Systems (NIPS)*, volume 30, pp. 1–13, Long Beach, CA, 2017.

Wu, H. and Wang, P. Fast-convergent federated learning with adaptive weighting. *IEEE Transactions on Cognitive Communications and Networking*, 7(4):1078–1088, 2021.

Xie, C., Koyejo, O., and Gupta, I. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference (UAI)*, volume 115, pp. 261–270, Online, 2020.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 5650–5659, Stockholm, Sweden, 2018.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. Parallelized stochastic gradient descent. In *Advances in Neural Information Processsing Systems (NIPS)*, volume 23, pp. 2595–2603, Vancouver, Canada, 2010.

A. Details of Experiment Settings

Datasets. The real datasets used for image classification model learning are the MNIST (LeCun et al., 1998), CIFAR-10, and CIFAR-100 (Krizhevsky & Hinton, 2009) datasets. MNIST dataset consists of total 70,000 gray-scale image samples of handwritten digits from 0 to 9 with size 28×28, where 60,000 images are training data and the remaining 10,000 images are test data. CIFAR-10 dataset consists of 60,000 32×32 color images in 10 classes with 50,000 training images and 10,000 test images. CIFAR-100 is the same with CIFAR-10, except for 100 classes. Some common data augmentation methods are used such as random cropping, horizontal random flipping, and random rotation. The data distribution for workers is assumed to be IID that each worker has the same number of image samples for the entire classes of each dataset.

Neural networks. In the training process, we adopt a convolutional neural network (CNN) (LeCun et al., 1998) model for the MNIST dataset, and a ResNet (He et al., 2016) model for the CIFAR-10 and CIFAR-100 datasets. The employed CNN architecture contains $N = 5 \times 10^5$ parameters, consisting of two convolutional layers each with a 5×5 kernel, connected with two fully-connected layers. For the ResNet model, we use ResNet-56 model which has 56 layers with parameters $N = 8.5 \times 10^5$.

Benchmarks. To evaluate the robustness against attacks and the communication efficiency of signSGD-FD, we compare the test accuracy with other conventional DL algorithms. We briefly introduce three signSGD-based DL algorithms and the leveraged parameters, which we adopt in the comparison as below:

- Election-signSGD (Sohn et al., 2020): Election-signSGD uses two-stage majority voting aggregation – multiple polling stations for the first stage, and the final decision aggregation for the second stage. The number of polling stations are the same as the number of workers, and the voting workers for each station is determined by the generator matrix $\mathbf{G}$, which is inspired from coding theory. The matrix $\mathbf{G}$ is generated by the random Bernoulli codes with $r = 2$, the best setting for $M = 15$ in this paper. Notably, the authors assume that the first stage is not affected by any attack.
- Sto-signSGD (Jin et al., 2020): Sto-signSGD leverages stochastic sign quantization depending on the magnitude of each gradient component. To be resilient to the attacks, this algorithm additionally uses the reputation-based weights in the majority voting aggregation, which looks similar with our algorithm. We utilize the best parameter $b = 0.012$ for CIFAR-10 dataset, where b determines the sign quantization probability for the gradient. The main difference from our algorithm is that sto-signSGD assigns the weight of 0 for the suspected workers, while signSGD-FD can assign the negative weights to those workers.
- Noisy signSGD (Chen et al., 2020): Noisy signSGD injects random Gaussian noise to the computed local gradients to mitigate the effect of the adversarial attacks. The original goal of this algorithm is to address the heterogeneity of the data distribution by decreasing the gap between mean and median of each gradient component. However, since it is well known that median-based algorithms are attack resistant, we add this Noisy signSGD algorithm to the comparison. We use the standard deviation parameter $b = 10^{-3}$.

In the communication costs comparison, we use distributed SGD, Top-K SGD optimizers for the baseline. Considering the robustness against attacks, we apply Multi-Krum algorithm to the above baseline algorithms. We elucidate the details for each algorithm as below:

- Distributed SGD (Zinkevich et al., 2010): Distributed SGD leverages 32-bit full-precision information for each gradient component. The aggregation rule is just computing the average of the workers' local gradients.
- Top-K SGD (Stich et al., 2018): Top-K SGD selects only the largest K gradient components in magnitude to update its model. As above, the locally computed gradients are averaged in the aggregation process, and the optimizer updates the model by using 32-bit gradient information. In the comparison, we select only 10% of gradient component, i.e., $\frac{K}{N} = 0.1$.
- Multi-Krum (Blanchard et al., 2017): Multi-Krum is the attack-robust DL algorithm based on SGD optimizers, which trains the model by selecting K workers who do not appear to be affected by the attacks. When there exist L compromised workers in total M workers, the server scores each worker by adding the ℓ-2 distance between the gradient of corresponding worker and the closest $M - L - 2$ gradients of other workers. Then, the server computes the average of gradients by selecting the K lowest-score workers. This can be seen as updating the model with geometric median for workers' gradients.

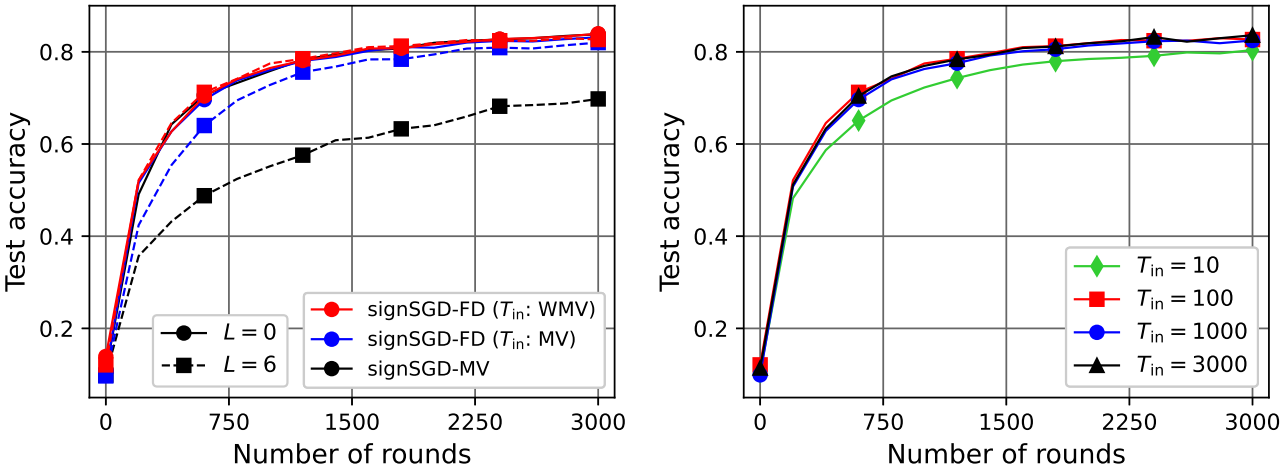
Communication costs. In Section 6.5, we compute the total communication cost for each algorithm required to achieve a certain test accuracy. The communication costs per iteration and worker can be calculated by multiplying the number of the shared gradient components per each worker by the bits required to represent each gradient component. We also consider the costs for the uplink communication (from workers to server) and the downlink communication (from server to workers). Based on this, Table I provides the communication costs for each base algorithm.

Table 1. Total communication costs for each base algorithm.

BASE ALGORITHMS	TOTAL COMMUNICATION COSTS
SGD	$[32N + 32N] \times MT$
TOP-K SGD	$\left[\left(32K + K\log_2\left(\frac{N}{K}\right)\right) + 32N\right] \times MT$
SIGNSGD-MV	$[N + N] \times MT$

Hyper-parameters. For the hyper-parameters that we can tune during the simulations, the number of workers M is fixed to 15 and all workers use the same mini-batch size of $B_m = 64, \forall m \in [M]$. The learning rate of each algorithm is carefully selected by comparing the converged test accuracy, where the value is $\delta = 10^{-3}$ and 10^{-1} for signSGD-based optimizers and SGD-based optimizers, respectively. To stabilize the learning, we do not actively utilize momentum and weight decay.

B. Initial Phase Aggregation Design



(a) Varying initial decoder type

(b) Varying T_{in}

Figure 4. Test accuracy comparison according to the initial phase aggregation of signSGD-FD.

As we observe in Section 5.2, we should design the cross-over probability estimation process delicately. To find a better estimation algorithm, we evaluate the test accuracy of signSGD-FV by changing the learning process in the initial phase, which is described in Figure 4. Here, the number of workers M is 15, and 6 workers are attacked by the SIA method, i.e., $L = 6$. The initial aggregation method of signSGD-FD is to perform the WMV aggregation by considering all gradient coordinates in (14), but the easiest way is to not use the weights as with the MV aggregation. We compare these two aggregation methods through the test accuracy results for CIFAR-10 image classification task. From the results in Figure 4-(a), the original signSGD-FD using WMV aggregation has negligible degradation due to attacks, but the accuracy deterioration begins to emerge as we employ the MV aggregation. This is expected to result in inaccurate weight estimation in the initial phase because the majority voting is greatly affected by the attack of $r = 1$. The effect of the initial phase

duration can be seen in Figure 4-(b), and this shows us that T_{in} does not affect significantly unless the duration is not too short to collect the error samples, such as $T_{\mathsf{in}} = 10$. Therefore, we can summarize that the initial phase aggregation design should be considered carefully, and the proposed signSGD-FD performs well in the presence of attackers. For ease of implementation, we set the initial period T_{in} to 50, 100, 500 for the MNIST, CIFAR-10, and CIFAR-100 datasets, respectively.

C. Proof of Theorem 5.3

Proof. We commence the proof by leveraging Assumption 5.2 to calculate the upper bound of the loss reduction of $f^{t+1} - f^t$. Using the update rule of signSGD-based algorithms in (13), the upper bound can be derived as

$$\begin{aligned} f^{t+1} - f^t &\leq \left\langle \bar{\mathbf{g}}^t, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \sum_{n=1}^{N} \frac{L_n}{2} \left(x_n^{t+1} - x_n^t \right)^2 \\ &= \sum_{n=1}^{N} \left[-\bar{g}_n^t \cdot \delta \hat{U}_n^t + \frac{L_n}{2} \left(-\delta \hat{U}_n^t \right)^2 \right] \\ &= -\delta \sum_{n=1}^{N} \bar{g}_n^t \hat{U}_n^t + \frac{1}{2} \delta^2 \|\mathbf{L}\|_1 \\ &= -\delta \|\bar{\mathbf{g}}^t\|_1 + \frac{\delta^2}{2} \|\mathbf{L}\|_1 + 2\delta \sum_{n=1}^{N} \left| \bar{g}_n^t \right| \mathbf{1}_{\left[U_n^t \neq \hat{U}_n^t \right]}. \end{aligned} \tag{29}$$

By taking expectation according to the randomness of $\hat{U}_n^t$, $f^{t+1} - f^t$ conditioned by $\mathbf{x}^t$ can be upper bounded by

$$\begin{aligned} \mathbb{E}\left[f^{t+1} - f^t \middle| \mathbf{x}^t \right] &\leq -\delta \|\bar{\mathbf{g}}^t\|_1 + \frac{\delta^2}{2} \|\mathbf{L}\|_1 + 2\delta \sum_{n=1}^{N} \left| \bar{g}_n^t \right| \mathbb{P}\left[U_n^t \neq \hat{U}_n^t \right] \\ &\leq -\delta \|\bar{\mathbf{g}}^t\|_1 + \frac{\delta^2}{2} \|\mathbf{L}\|_1 + 2\delta P_{\mathsf{E}}^{\mathsf{max}} \sum_{n=1}^{N} \left| \bar{g}_n^t \right| \\ &= -\delta \left(1 - 2P_{\mathsf{E}}^{\mathsf{max}} \right) \|\bar{\mathbf{g}}^t\|_1 + \frac{\delta^2}{2} \|\mathbf{L}\|_1. \end{aligned} \tag{30}$$

Next, we take the expectation over $\mathbf{x}^t$, and apply a telescoping sum over the iterations, which provides

$$\begin{aligned} f^1 - f^\star &\geq f^1 - \mathbb{E}\left[f^T \right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} f^t - f^{t+1} \right] \\ &\geq \mathbb{E}\left[\sum_{t=1}^{T} \left\{ \delta \left(1 - 2P_{\mathsf{E}}^{\mathsf{max}} \right) \|\bar{\mathbf{g}}^t\|_1 - \frac{\delta^2}{2} \|\mathbf{L}\|_1 \right\} \right], \\ &= \delta \left(1 - 2P_{\mathsf{E}}^{\mathsf{max}} \right) \mathbb{E}\left[\sum_{t=1}^{T} \|\bar{\mathbf{g}}^t\|_1 \right] - \frac{\delta^2 T}{2} \|\mathbf{L}\|_1, \end{aligned} \tag{31}$$

where the last equality holds when δ is fixed according to the training round $t \in [T]$. Consequently, by plugging the learning rate $\delta = \sqrt{\frac{2(f^1 - f^\star)}{T \|\mathbf{L}\|_1}}$ into (31), we obtain

$$\begin{aligned} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \|\bar{\mathbf{g}}^t\|_1 \right] &\leq \frac{1}{1 - 2P_{\mathsf{E}}^{\mathsf{max}}} \left[\frac{1}{\delta T} \left(f^1 - f^\star \right) + \frac{\delta}{2} \|\mathbf{L}\|_1 \right] \\ &= \frac{1}{1 - 2P_{\mathsf{E}}^{\mathsf{max}}} \sqrt{\frac{2 \left(f^1 - f^\star \right) \|\mathbf{L}\|_1}{T}}. \end{aligned} \tag{32}$$

This completes the proof. □

D. Proof of Theorem 5.4

Proof. We define a binary random variable that indicates the decoding error event, i.e., $Z_{m,n}^t = \mathbf{1}_{\left[U_n^t \neq Y_{m,n}^t\right]}$. When employing the FD aggregation with the imperfect LLR weight $\hat{w}_{m,n}^t = \ln \frac{1-\hat{p}_{m,n}^t}{\hat{p}_{m,n}^t}$ with the uncertainty (23), decoding failures arise if the cumulative sum of weights assigned to the incorrectly decoding workers exceeds half of the total weight. Using this fact, the decoding error probability is rewritten as

$$\begin{aligned}
\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{FD},n}^t\right] &= \mathbb{P}\left[\sum_{m=1}^{M} \hat{w}_{m,n}^t Z_{m,n}^t \geq \frac{1}{2}\sum_{m=1}^{M} \hat{w}_{m,n}^t\right] \\
&= \mathbb{P}\left[\sum_{m=1}^{M} \hat{w}_{m,n}^t \left(Z_{m,n}^t - p_{m,n}^t\right) \geq \sum_{m=1}^{M} \hat{w}_{m,n}^t \left(\frac{1}{2} - p_{m,n}^t\right)\right] \\
&= \mathbb{P}\left[\sum_{m=1}^{M} \hat{w}_{m,n}^t \bar{Z}_{m,n}^t \geq \eta\right],
\end{aligned} \tag{33}$$

where $\bar{Z}_{m,n}^t = Z_{m,n}^t - p_{m,n}^t$ and $\eta = \sum_{m=1}^{M} \hat{w}_{m,n}^t \left(\frac{1}{2} - p_{m,n}^t\right)$. Applying Chernoff bound to (33) for $s > 0$ yields an upper bound on the error probability, expressed as:

$$\begin{aligned}
\mathbb{P}\left[\sum_{m=1}^{M} \hat{w}_{m,n}^t \bar{Z}_{m,n}^t \geq \eta\right] &\leq \min_{s>0} e^{-\eta s}\, \mathbb{E}\left[\exp\left(s\sum_{m=1}^{M} \hat{w}_{m,n}^t \bar{Z}_{m,n}^t\right)\right] \\
&= \min_{s>0} e^{-\eta s} \prod_{m=1}^{M} \mathbb{E}\left[e^{s\hat{w}_{m,n}^t \bar{Z}_{m,n}^t}\right].
\end{aligned} \tag{34}$$

Here, we leverage the large deviation bound established in Lemma 1 in (Kearns & Saul, 1998), which is stated as

$$(1-p)e^{-tp} + pe^{t(1-p)} \leq \exp\left(\frac{1-2p}{4\ln\frac{1-p}{p}} t^2\right), \tag{35}$$

for all $p \in [0,1]$ and $|t| < \infty$. Then, we obtain the upper bound of the expectation term in (34) as

$$\begin{aligned}
\mathbb{E}\left[e^{s\hat{w}_{m,n}^t \bar{Z}_{m,n}^t}\right] &= p_{m,n}^t e^{s\hat{w}_{m,n}^t\left(1-p_{m,n}^t\right)} + \left(1-p_{m,n}^t\right) e^{-s\hat{w}_{m,n}^t p_{m,n}^t} \\
&\leq \exp\left[\frac{1-2p_{m,n}^t}{4\ln\frac{1-p_{m,n}^t}{p_{m,n}^t}} \left(\hat{w}_{m,n}^t\right)^2 s^2\right] \\
&= \exp\left[\frac{1}{2}\left(\frac{1}{2} - p_{m,n}^t\right) \frac{\hat{w}_{m,n}^t}{w_{m,n}^t} \hat{w}_{m,n}^t s^2\right],
\end{aligned} \tag{36}$$

where the last equality follows from $w_{m,n}^t = \ln \frac{1-p_{m,n}^t}{p_{m,n}^t}$. From the uncertainty of LLR weight in (23), we can express the upper bound as

$$\mathbb{E}\left[e^{s\hat{w}_{m,n}^t \bar{Z}_{m,n}^t}\right] \leq \exp\left[\frac{1+\delta_{\mathsf{max}}}{2}\left(\frac{1}{2} - p_{m,n}^t\right) \hat{w}_{m,n}^t s^2\right]. \tag{37}$$

Invoking (37) into (34), and also using $\eta = \sum_{m=1}^{M} \hat{w}_{m,n}^t \left(\frac{1}{2} - p_{m,n}^t\right)$, the upper bound of the FD decoding error probability becomes

$$\begin{aligned}
\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{FD},n}^t\right] &\leq \min_{s>0} e^{-\eta s} \exp\left[\frac{\left(1+\delta_{\mathsf{max}}\right) s^2}{2} \sum_{m=1}^{M}\left(\frac{1}{2} - p_{m,n}^t\right) \hat{w}_{m,n}^t\right] \\
&= \min_{s>0} \exp\left[\frac{\left(1+\delta_{\mathsf{max}}\right)\eta}{2} s^2 - \eta s\right] \\
&= \exp\left[-\frac{1}{2\left(1+\delta_{\mathsf{max}}\right)}\eta\right],
\end{aligned} \tag{38}$$

where the last equality follows from the fact that $s = \frac{1}{1+\delta_{\mathsf{max}}}$ is the minimizer of the optimization problem in (38). Now, we derive a lower bound of η in terms of δ_{min} as

$$\begin{aligned}\eta &= \sum_{m=1}^{M} \hat{w}_{m,n}^{t} \left(\frac{1}{2} - p_{m,n}^{t}\right) \\ &= \sum_{m=1}^{M} \left(\frac{1}{2} - p_{m,n}^{t}\right) \frac{\hat{w}_{m,n}^{t}}{w_{m,n}^{t}} \hat{w}_{m,n}^{t} \\ &\geq (1-\delta_{\mathsf{min}}) \sum_{m=1}^{M} \left(\frac{1}{2} - p_{m,n}^{t}\right) w_{m,n}^{t}. \end{aligned} \tag{39}$$

Consequently, by substituting (39) to (38), the upper bound becomes

$$\begin{aligned}\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{FD},n}^t\right] &\leq \exp\left[-M\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right) \cdot \frac{1}{2M} \sum_{m=1}^{M} \left(\frac{1}{2} - p_{m,n}^{t}\right) \ln \frac{1-p_{m,n}^{t}}{p_{m,n}^{t}}\right] \\ &= \exp\left[-M\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right) \gamma_{\mathcal{M}}^{\mathsf{WMV}}\right]. \end{aligned} \tag{40}$$

This concludes the proof.

□

E. Proof of Theorem 5.5

Proof. Similar to the proof in Appendix D, we express the decoding error probability of the MV aggregation in terms of a binary random variable $Z_{m,n}^t = \mathbf{1}_{\left[U_n^t \neq \hat{U}_{\mathsf{MV},n}^t\right]}$ as

$$\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{MV},n}^t\right] = \mathbb{P}\left[\sum_{m=1}^{M} Z_{m,n}^t \geq \frac{M}{2}\right]. \tag{41}$$

By applying Markov's inequality in (41), the MV decoding error probability is upper bounded as

$$\begin{aligned}\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{MV},n}^t\right] &\leq \min_{s>0} e^{-\frac{M}{2}s}\, \mathbb{E}\left[\exp\left(s \sum_{m=1}^{M} Z_{m,n}^t\right)\right] \\ &= \min_{s>0} e^{-\frac{M}{2}s} \prod_{m=1}^{M} \mathbb{E}\left[e^{sZ_{m,n}^t}\right]. \end{aligned} \tag{42}$$

Using the moment generating function of $Z_{m,n}^t$ which follows Bernoulli distribution with $p_{m,n}^t = \mathbb{P}\left[Z_{m,n}^t = 1\right]$, the upper bound of the MV decoding error probability becomes

$$\begin{aligned}\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{MV},n}^t\right] &\leq \min_{s>0} e^{-\frac{M}{2}s} \prod_{m=1}^{M} \left(p_{m,n}^t e^s + 1 - p_{m,n}^t\right) \\ &= \min_{s>0} \exp\left[-\frac{M}{2}s + \sum_{m=1}^{M} \ln\left(p_{m,n}^t\left(e^s - 1\right) + 1\right)\right] \\ &\leq \min_{s>0} \exp\left[-\frac{M}{2}s + \sum_{m=1}^{M} p_{m,n}^t\left(e^s - 1\right)\right], \end{aligned} \tag{43}$$

where the last inequality comes from the property $x \geq \ln(1+x)$ when $x \geq 0$. Let us denote the average of workers' cross-over probabilities as $\bar{p}_n^t = \frac{1}{M}\sum_{m=1}^{M} p_{m,n}^t$. Then, by substituting $s = \ln \frac{1}{2\bar{p}_n^t}$ which minimizes the upper bound in

(43), the upper bound becomes

$$\begin{aligned}\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{MV},n}^t\right] &\leq \exp\left[-\frac{M}{2}\ln\frac{1}{2\bar{p}_n^t} + M\bar{p}_n^t\left(\frac{1}{2\bar{p}_n^t}-1\right)\right] \\ &= \exp\left[-M\left(\bar{p}_n^t - \frac{1}{2}\ln\left(2e\bar{p}_n^t\right)\right)\right] \\ &= \exp\left(-M\gamma_{\mathcal{M}}^{\mathsf{MV}}\right). \end{aligned} \tag{44}$$

This concludes the proof. □

F. Proof of Theorem 5.6

Proof. In order to analyze the deterioration of the FD aggregation caused by the stochastic sign flip attacks, we leverage the result of Theorem 5.4. For ease of expression, we define a function $g(p) = \frac{1}{2}\left(\frac{1}{2}-p\right)\ln\frac{1-p}{p}$ for $p \in [0,1]$. Then, the FD decoding error bound is derived as

$$\begin{aligned}\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{FD},n}^t\right] &\leq \exp\left[-\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\cdot\left(\sum_{m\in\mathcal{M}\setminus\mathcal{L}} g\left(p_{m,n}^t\right) + \sum_{\ell\in\mathcal{L}} g\left(\tilde{p}_{\ell,n}^t\right)\right)\right] \\ &= \exp\left[-\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\cdot\left(\sum_{m\in\mathcal{M}} g\left(p_{m,n}^t\right) - \sum_{\ell\in\mathcal{L}} g\left(p_{\ell,n}^t\right) + \sum_{\ell\in\mathcal{L}} g\left(\tilde{p}_{\ell,n}^t\right)\right)\right] \\ &= \exp\left[-\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\cdot\left(M\gamma_{\mathcal{M}}^{\mathsf{WMV}} - L\gamma_{\mathcal{L}}^{\mathsf{WMV}} + \sum_{\ell\in\mathcal{L}} g\left(\tilde{p}_{\ell,n}^t\right)\right)\right], \end{aligned} \tag{45}$$

where the error exponent of WMV decoder without attacks $\gamma_{\mathcal{M}}^{\mathsf{WMV}} = \frac{1}{M}\sum_{m\in\mathcal{M}} g\left(p_{m,n}^t\right)$ and $\gamma_{\mathcal{L}}^{\mathsf{WMV}} = \frac{1}{L}\sum_{\ell\in\mathcal{L}} g\left(p_{\ell,n}^t\right)$ are exploited. Concentrating on the coefficients of error exponents, the error bound of FD aggregation in the presence of attacks becomes

$$\mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{FD},n}^t\right] \leq \exp\left[-(M-L)\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{WMV}}\right], \tag{46}$$

where $\tilde{\gamma}_{\mathcal{L}}^{\mathsf{WMV}} = \frac{1}{M-L}\left[M\gamma_{\mathcal{M}}^{\mathsf{WMV}} - L\gamma_{\mathcal{L}}^{\mathsf{WMV}} + \sum_{\ell\in\mathcal{L}} g\left(\tilde{p}_{\ell,n}^t\right)\right]$ is the modified error exponent. This concludes the proof. □

G. Proof of Corollary 5.7

Proof. From the result of Theorem 5.6, we modify the $\sum_{\ell\in\mathcal{L}} g\left(\tilde{p}_{\ell,n}^t\right)$ term according to the sign flip probability $r = 1$ and $\frac{1}{2}$. Using the compromised cross-over probability in (11), we can derive the property as

$$\begin{aligned}\frac{1}{2} - \tilde{p}_{\ell,n}^t &= \frac{1}{2} - \left(p_{\ell,n}^t + r\left(1-2p_{\ell,n}^t\right)\right) \\ &= (1-2r)\left(\frac{1}{2} - p_{\ell,n}^t\right) \\ &\triangleq q_{\ell,n}^t. \end{aligned} \tag{47}$$

Then, we can organize the $g\left(\tilde{p}_{\ell,n}^t\right)$ term as

$$\begin{aligned} g\left(\tilde{p}_{\ell,n}^t\right) &= \frac{1}{2}\left(\frac{1}{2} - \tilde{p}_{\ell,n}^t\right)\ln\frac{1-\tilde{p}_{\ell,n}^t}{\tilde{p}_{\ell,n}^t} \\ &= \frac{1}{2}q_{\ell,n}^t \ln\frac{\frac{1}{2}+q_{\ell,n}^t}{\frac{1}{2}-q_{\ell,n}^t}, \end{aligned} \tag{48}$$

where $-\frac{1}{2} < -\left(\frac{1}{2} - p^t_{\ell,n}\right) \leq q^t_{\ell,n} \leq \frac{1}{2} - p^t_{\ell,n} < \frac{1}{2}$. It can be easily checked that $h(x) = x \ln \frac{1/2+x}{1/2-x}$ is an even function, i.e., $h(x) = h(-x)$. Therefore, we can obtain the results that $0 \leq \sum_{\ell\in\mathcal{L}} g\left(\tilde{p}^t_{\ell,n}\right) \leq \sum_{\ell\in\mathcal{L}} g\left(p^t_{\ell,n}\right) = L\gamma^{\mathsf{WMV}}_{\mathcal{L}}$. The lower bound of the inequality can be achieved when $r = \frac{1}{2}$, and the upper bound also can be achieved in the $r = 1$ case. By substituting these results to the error exponent $\tilde{\gamma}^{\mathsf{WMV}}_{\mathcal{L}}$ in Theorem 5.6, the error bound of FD aggregation becomes

$$\mathbb{P}\left[U^t_n \neq \hat{U}^t_{\mathsf{FD},n}\right] \leq \begin{cases} \exp\left[-M\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\gamma^{\mathsf{WMV}}_{\mathcal{M}}\right], & \text{if } r = 1 \\ \exp\left[-\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\left(M\gamma^{\mathsf{WMV}}_{\mathcal{M}} - L\gamma^{\mathsf{WMV}}_{\mathcal{L}}\right)\right], & \text{if } r = \frac{1}{2} \end{cases}. \tag{49}$$

Since our distributed learning system considers that all workers employ the same mini-batch sizes, i.e., $B_m = B_{m'}, \forall m \neq m'$, which results in the identical cross-over probabilities $p^t_{m,n} = p^t_{m',n}$ by referring the proof of Theorem 1 in (Bernstein et al., 2018). Therefore, we can note that the error exponents of worker set $\mathcal{M}$ and $\mathcal{L}$ are the same, i.e., $\gamma^{\mathsf{WMV}}_{\mathcal{M}} = \gamma^{\mathsf{WMV}}_{\mathcal{L}}$. Consequently, we can obtain the final result of Corollary 5.7 as

$$P^{\mathsf{FD}}_{\mathsf{E}} \leq \begin{cases} \exp\left[-M\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\gamma^{\mathsf{WMV}}_{\mathcal{M}}\right], & \text{if } r = 1 \\ \exp\left[-\left(M-L\right)\left(\frac{1-\delta_{\mathsf{min}}}{1+\delta_{\mathsf{max}}}\right)\gamma^{\mathsf{WMV}}_{\mathcal{M}}\right], & \text{if } r = \frac{1}{2} \end{cases}. \tag{50}$$

□

H. Proof of Theorem 5.8

Proof. Aligning with the proof of Theorem 5.5 in Appendix E, we can easily derive the upper bound of MV decoding error probability as

$$\mathbb{P}\left[U^t_n \neq \hat{U}^t_{\mathsf{MV},n}\right] \leq \exp\left[-M\left(\tilde{p}^t_n - \frac{1}{2}\ln\left(2e\tilde{p}^t_n\right)\right)\right], \tag{51}$$

where $\tilde{p}^t_n$ is the average of cross-over probabilities for honest workers $m \in \mathcal{M}\backslash\mathcal{L}$ and compromised workers $\ell \in \mathcal{L}$, i.e.,

$$\begin{aligned} \tilde{p}^t_n &= \frac{1}{M}\left[\sum_{m\in\mathcal{M}\backslash\mathcal{L}} p^t_{m,n} + \sum_{\ell\in\mathcal{L}} \tilde{p}^t_{\ell,n}\right] \\ &= \frac{1}{M}\left[\sum_{m\in\mathcal{M}\backslash\mathcal{L}} p^t_{m,n} + \sum_{\ell\in\mathcal{L}} \left(p^t_{\ell,n} + r\left(1 - 2p^t_{\ell,n}\right)\right)\right] \\ &= \frac{1}{M}\left[\sum_{m\in\mathcal{M}} p^t_{m,n} + r\sum_{\ell\in\mathcal{L}}\left(1 - 2p^t_{\ell,n}\right)\right] \\ &= \frac{1}{M}\sum_{m\in\mathcal{M}} p^t_{m,n} + \frac{rL}{M}\left(1 - \frac{2}{L}\sum_{\ell\in\mathcal{L}} p^t_{\ell,n}\right) \\ &= \bar{p}^t_n + \frac{rL}{M}\left(1 - 2\bar{p}^t_{\mathcal{L},n}\right), \end{aligned} \tag{52}$$

where $\bar{p}^t_{\mathcal{L},n} = \frac{1}{L}\sum_{\ell\in\mathcal{L}} p^t_{\ell,n}$ is the average of true cross-over probabilities for the compromised workers $\ell \in \mathcal{L}$. Using (52), the exponent of error bound in (51) can be expressed as

$$\begin{aligned} M\left[\tilde{p}^t_n - \frac{1}{2}\ln\left(2e\tilde{p}^t_n\right)\right] &= M\bar{p}^t_n + rL\left(1 - 2\bar{p}^t_{\mathcal{L},n}\right) - \frac{M}{2}\ln\left(2e\tilde{p}^t_n\right) \\ &= M\left[\bar{p}^t_n - \frac{1}{2}\ln\left(2e\bar{p}^t_n\right)\right] + rL\left(1 - 2\bar{p}^t_{\mathcal{L},n}\right) - \frac{M}{2}\ln\frac{\tilde{p}^t_n}{\bar{p}^t_n} \\ &= M\gamma^{\mathsf{MV}}_{\mathcal{M}} + rL\left(1 - 2\bar{p}^t_{\mathcal{L},n}\right) - \frac{M}{2}\ln\frac{\tilde{p}^t_n}{\bar{p}^t_n}, \end{aligned} \tag{53}$$

where the error exponent of MV decoder without attack $\gamma_{\mathcal{M}}^{\mathsf{MV}} = \bar{p}_n^t - \frac{1}{2}\ln\left(2e\bar{p}_n^t\right)$ is used. Likewise, we can continue the proof by leveraging the error exponent term $\gamma_{\mathcal{L}}^{\mathsf{MV}} = \bar{p}_{\mathcal{L},n}^t - \frac{1}{2}\ln\left(2e\bar{p}_{\mathcal{L},n}^t\right)$ as

$$\begin{aligned} M\left[\tilde{p}_n^t - \frac{1}{2}\ln\left(2e\tilde{p}_n^t\right)\right] &= M\gamma_{\mathcal{M}}^{\mathsf{MV}} - 2rL\left(\bar{p}_{\mathcal{L},n}^t - \frac{1}{2}\ln\left(2e\bar{p}_{\mathcal{L},n}^t\right)\right) + rL - rL\ln\left(2e\bar{p}_{\mathcal{L},n}^t\right) - \frac{M}{2}\ln\frac{\tilde{p}_n^t}{\bar{p}_n^t} \\ &= M\gamma_{\mathcal{M}}^{\mathsf{MV}} - 2rL\gamma_{\mathcal{L}}^{\mathsf{MV}} - \frac{M}{2}\ln\frac{\tilde{p}_n^t}{\bar{p}_n^t} - rL\ln\frac{\bar{p}_{\mathcal{L},n}^t}{1/2} \\ &= M\gamma_{\mathcal{M}}^{\mathsf{MV}} - 2rL\gamma_{\mathcal{L}}^{\mathsf{MV}} + \epsilon_{\mathcal{L}}(r), \end{aligned} \tag{54}$$

and $\epsilon_{\mathcal{L}}(r) = -\frac{M}{2}\ln\frac{\tilde{p}_n^t}{\bar{p}_n^t} - rL\ln\frac{\bar{p}_{\mathcal{L},n}^t}{1/2}$ is a sufficiently small parameter which contains the ratios of the average cross-over probabilities. Consequently, if we only focus on the coefficients of the error exponents $\gamma_{\mathcal{M}}^{\mathsf{MV}}$ and $\gamma_{\mathcal{L}}^{\mathsf{MV}}$, we can simply express the error bound of the MV decoder affected by the stochastic sign flip attacks with probability r as

$$\begin{aligned} \mathbb{P}\left[U_n^t \neq \hat{U}_{\mathsf{MV},n}^t\right] &\leq \exp\left[-\left(M\gamma_{\mathcal{M}}^{\mathsf{MV}} - 2rL\gamma_{\mathcal{L}}^{\mathsf{MV}} + \epsilon_{\mathcal{L}}(r)\right)\right] \\ &= \exp\left[-\left(M - 2rL\right)\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{MV}}\right], \end{aligned} \tag{55}$$

where $\tilde{\gamma}_{\mathcal{M},\mathcal{L}}^{\mathsf{MV}} = \frac{1}{M-2rL}\left[M\gamma_{\mathcal{M}}^{\mathsf{MV}} - 2rL\gamma_{\mathcal{L}}^{\mathsf{MV}} + \epsilon_{\mathcal{L}}(r)\right]$ is the error exponent of MV decoder in the presence of attacks. This concludes the proof. □