Momentum Ensures Convergence of SIGNSGD under Weaker Assumptions

Tao Sun[*] [1] **Qingsong Wang**[*] [2] **Dongsheng Li**[1] **Bao Wang**[2]

Abstract

Sign Stochastic Gradient Descent (SIGNSGD) is a communication-efficient stochastic algorithm that only uses the sign information of the stochastic gradient to update the model's weights. However, the existing convergence theory of SIGNSGD either requires increasing batch sizes during training or assumes the gradient noise is symmetric and unimodal. Error feedback has been used to guarantee the convergence of SIGNSGD under weaker assumptions at the cost of communication overhead. This paper revisits the convergence of SIGNSGD and proves that momentum can remedy SIGNSGD under weaker assumptions than previous techniques; in particular, our convergence theory does not require the assumption of bounded stochastic gradient or increased batch size. Our results resonate with echoes of previous empirical results where, unlike SIGNSGD, SIGNSGD with momentum maintains good performance even with small batch sizes. Another new result is that SIGNSGD with momentum can achieve an improved convergence rate when the objective function is second-order smooth. We further extend our theory to SIGNSGD with major vote and federated learning.

1. Introduction

This paper considers the following optimization problem arising from machine learning:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \mathbb{E}_{\xi\sim\mathcal{D}} f(\boldsymbol{x};\xi), \tag{1}$$

where $\mathcal{D}$ is the data distribution. Bernstein et al. (2018a) propose SIGNSGD, which solves the optimization problem in (1) via the following iterations:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \gamma \text{Sign}(\nabla f(\boldsymbol{x}^t;\xi^t)), \quad \text{for } t \geq 0, \tag{2}$$

where ξ^t is the data sampled i.i.d. from the distribution $\mathcal{D}$ in the t-th iteration, and $\gamma > 0$ is the learning rate (or step size). In each iteration, SIGNSGD only uses the sign information of the stochastic gradient, which reduces communication costs in each iteration. However, using the sign operator leads to biased expectations and prevents convergence unless the batch size is increased or the gradient noise is symmetric and unimodal (Bernstein et al., 2018a;b).

The error feedback technique has been developed to address this issue and eliminate the need for large sampling costs (Karimireddy et al., 2019). The error feedback technique is a variant of the momentum method, which has been widely used for both theoretical and empirical improvements in optimization (Karimireddy et al., 2019; Lin et al., 2018; Stich et al., 2018). The basic momentum scheme for the SIGNSGD, i.e., SIGNSGD with Simple Momentum (SIGNSGD-SIM), is shown in Algorithm 1. However, the current convergence guarantees for SIGNSGD-SIM, as presented in (Bernstein et al., 2018a;b)[1], are established based on the same assumptions as those for SIGNSGD.

Algorithm 1 SIGNSGD with SImple Momentum (SIGNSGD-SIM)

```
Require: parameters $\gamma > 0, 0 \leq \theta < 1$
  Initialization: $\boldsymbol{x}^1 = \boldsymbol{0}, \boldsymbol{m}^0 = \boldsymbol{0}$
  for $t = 1, 2, \ldots$
    step 1: Sample $\xi^t \sim \mathcal{D}$ i.i.d and update
            $\boldsymbol{m}^t = \theta\boldsymbol{m}^{t-1} + (1-\theta)\nabla f(\boldsymbol{x}^t;\xi^t)$
    step 2: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \gamma\text{Sign}(\boldsymbol{m}^t)$
  end for
```

Empirical performance of SIGNSGD and SIGNSGD-SIM under different batch sizes. To investigate if the simple momentum scheme can fix the convergence of SIGNSGD, we conduct experiments on training ResNet110 from (He et al., 2016) for CIFAR-100 (Krizhevsky et al., 2009) classification using SIGNSGD and SIGNSGD-SIM with various batch sizes. The momentum parameter is set to 0.9 for SIGNSGD-SIM, and the other hyperparameters are the same as SIGNSGD, as described in more detail in Section 4.1. The experimental results are shown in Figure 1, where both SIGNSGD and SIGNSGD-SIM achieve a testing accuracy of around 71%. As the batch size decreases,

[*]Equal contribution [1]College of Computer, National University of Defense Technology, Hunan, China. [2]University of Utah. Correspondence to: Dongsheng Li < dsli@nudt.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] SIGNSGD-SIM is denoted as SIGNUM in their paper

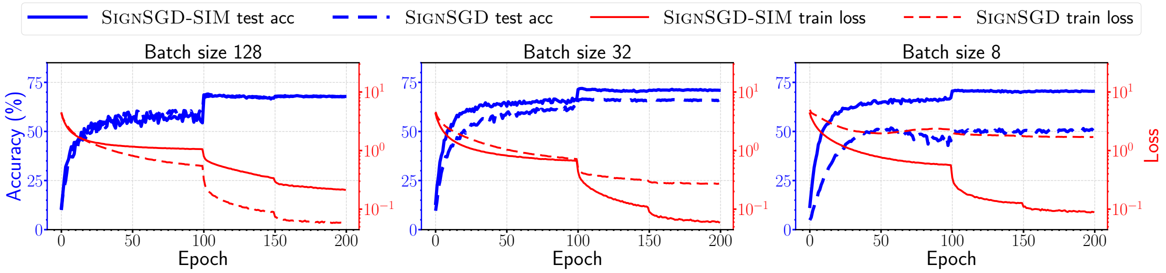


Figure 1. Experimental results of train losses, train accuracies, and test accuracies for training **ResNet110** on **CIFAR-100** with different batch sizes. As the batch size reduces, both training loss and test accuracy of SIGNSGD deteriorate substantially. In contrast, the performance of SIGNSGD-SIM is nearly preserved, even when using a very small batch size.

SIGNSGD-SIM maintains similar training curves while the performance of SIGNSGD drops significantly. The above experimental results align with previous empirical findings in (Karimireddy et al., 2019; Bernstein et al., 2018b), which show that SIGNSGD-SIM often outperforms SIGNSGD, especially when the batch size is small.

Indeed, in Theorem 1 in Section 2, we present an improved convergence guarantee for SIGNSGD-SIM that does not require the assumption of bounded stochastic gradient, large batch size, or symmetric unimodal gradient noise. The assumptions in our theorem are even weaker than those used in the error feedback scheme. Additionally, we show that SIGNSGD-SIM has a faster convergence rate when the objective function is second-order smooth. We also extend these theoretical findings to other variants of SIGNSGD, including SIGNSGD with major vote and federated learning.

1.1. Additional related works

SIGNSGD and its variants: Gradient quantization is a technique that aims to reduce the amount of information that needs to be transmitted when training machine learning models in a parallel setting (Alistarh et al., 2017). One specific type of gradient quantization is sign-based quantization, which involves transmitting only the sign of the gradient rather than the full gradient. While the reduction in information can lead to biased estimators, SIGNSGD is empirically shown to perform well even when using just 1-bit of information (Seide et al., 2014; Strom, 2015; Li et al., 2014) and has been used for distributed learning since the early days of deep learning (Li et al., 2014). Meanwhile, Balles & Hennig (2018) show that the sign-based method has a deep connection with the well-known Adam method (Kingma & Ba, 2015; Zhang et al.).

Thus, there has been a line of research to understand sign-based methods. For example, Bernstein et al. (2018a) provide a theoretical analysis of iteration error bound on the SIGNSGD. Bernstein et al. (2018b) propose the Major Vote (MV-) SIGNSGD, an extension of SIGNSGD for distributed setting that only requires transmitting the sign information. Motivated by black-box adversarial attacks in robust deep learning, Liu et al. (2019) propose the Zeroth Order (ZO-) SIGNSGD, which gets rid of employing the stochastic gradient directly. Al-Dujaili & O'Reilly (2020) present a new black-box adversarial attack algorithm by exploiting a sign-based gradient estimation approach.

However, most of the papers about SIGNSGD mentioned above require an increasing batch size to guarantee their convergence. Some researchers have recently proposed modifications of SIGNSGD that reduce sampling costs. For example, in (Karimireddy et al., 2019), the authors introduce the error feedback technique to remove the large sampling assumption and propose SIGNSGD-EF. The paper (Safaryan & Richtarik, 2021) modifies the scheme of SIGNSGD by comparing the global objective function values in each iteration, resulting in the algorithm SIGNSGD-CF. By considering a coordinate Lipschitz-like property, Crawshaw et al. (2022) propose a robust general SIGNSGD algorithm with stepsize adjusted by historical gradients.

Applications of SIGNSGD: By using SIGNSGD, in (Sohn et al., 2020), SIGNSGD is used to develop a coding method that minimizes the worker-master communication load to guarantee Byzantine-robustness for distributed learning. In (Jin et al., 2020), the sign-based method is applied to the federated training tasks and shown to have provable convergence guarantees.

1.2. Why do we need a simpler scheme?

The need for large batch sizes to achieve convergence can be costly in terms of both sampling and computation. This requirement can make it challenging to use SIGNSGD and its variants in specific settings. Although the error feedback and function value comparison methods are proposed to reduce the sampling cost of SIGNSGD, these methods also have limitations for applying to the distributed systems, such as the need to transmit additional information or the requirement for extra computation. In particular, the error feedback (SIGNSGD-EF) requires transmitting the magnitudes of the gradients, which accounts for 32-bit per layer in addition

References	BSG-F	FBS-F	IR	AS	FLE	EA-F
SIGNSGD (Bernstein et al., 2018a)	√	×	×	×	×	√
MV-SIGNSGD (Bernstein et al., 2018b)	√	×	×	√	×	√
ZO-SIGNSGD (Liu et al., 2019)	×	×	×	×	×	√
SIGNSGD-EF (Karimireddy et al., 2019)	×	√	×	×	×	√
SIGNSGD-CF (Safaryan & Richtarik, 2021)	√	√	×	×	×	×
Federated MV-SIGNSGD (Jin et al., 2020)	×	×	×	√	√	×
SIGNSGD-SIM (**This paper**)	√	√	√	×	×	√
MV-STO-SIGNSGD-SIM (**This paper**)	×	√	√	√	×	√
Federated MV-STO-SIGNSGD-SIM (**This paper**)	×	√	√	√	√	√

Table 1. Comparisons with previous closely related works on different assumptions for convergence. "BSG-F " stands for "Bounded Stochastic Gradient Free" (i.e, we do not need to assume $\mathbb{E}\|\nabla f(\boldsymbol{x};\xi)\|^2 \leq \hat{\sigma}^2$ for some $\hat{\sigma} > 0$.), "FBS" stands for "Fixed Batch Size", "IR" is short for " Improved Rates", "AS" means the "transmitted information are All Signs", "FLE" is "Federated Learning Extension", and "EA-F" is short for "Extra Assumptions Free".

to the signs, hurting communication efficiency. The function value comparison method requires obtaining the global training function value, which needs extra computational costs and is even untractable in the distributed setting. Compared with the vanilla SIGNSGD, the simple momentum scheme in Algorithm 1 only recruits an extra addition of vectors, which costs only a few computations and minimal memory overhead. SIGNSGD-SIM can be easily extended to the distributed schemes where communication efficiency is achieved by only transmitting 1-bit sign information.

1.3. Contributions

In this paper, we prove that the SIGNSGD with simple momentum converges under weaker assumptions. Different convergence results are established for other variants, of which all needed assumptions are weakened. We elaborate on our contributions below.

- We prove $\mathcal{O}(\frac{1}{T^{1/4}})$ convergence of SIGNSGD with simple momentum without the need for increasing batch sizes. When comparing with the error feedback SIGNSGD, the studied algorithm only requires the assumption of bounded variance rather than the bounded gradient.
- When the objective function f is second-order smooth, we prove that the SIGNSGD-SIM enjoys a faster convergence rate as $\mathcal{O}(\frac{1}{T^{2/7}})$ if we modify the momentum update slightly.
- We develop two communication-efficient variants for the distributed setting cases; the MV-SIGNSGD with simple momentum (MV-STO-SIGNSGD-SIM), and its federated scheme, in which all transmitted information just signs. We prove the convergence of these algorithms under both first- and second-order smooth assumptions.

We list the comparisons with related works on different assumptions for convergence in Table 1.

Notation. Throughout this paper, we use boldface letters to denote vectors, e.g., $\boldsymbol{x} \in \mathbb{R}^d$. The j-th coordinate of a vector $\boldsymbol{x}$ is denoted by x_j. We use $\mathrm{Diag}(\gamma_1, \gamma_2, \ldots, \gamma_d)$ to denote the diagonal matrix whose diagonal entries are $\gamma_1, \gamma_2, \ldots, \gamma_d$. Given a function $f(\boldsymbol{x};\xi)$, the gradient $\nabla f(\boldsymbol{x};\xi)$ is taken with respect to variable $\boldsymbol{x}$. The Hessian matrix of the function f is denoted by $\nabla^2 f$. We denote $\mathbb{E}[\cdot]$ as the expectation with respect to the underlying probability space. We use $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_{\text{op}}$ to denote the L_2-, L_1-norm, and spectral norm, respectively. We denote the minimum value of the function f as $\min f$. Given non-negative sequences $(a_t, b_t)_{t\geq 0}$, we use $a_t = \mathcal{O}(b_t)$ if $a_t \leq Cb_t$ with some constant $C > 0$ and we write $a_t = \Theta(b_t)$ if $a_t = \mathcal{O}(b_t)$ and $b_t = \mathcal{O}(a_t)$.

2. SIGNSGD with Simple Momentum

2.1. Assumptions

We collect several common and necessary assumptions in this subsection.

Assumption 1 *Function $f(\cdot)$ is differentiable with the Lipschitz gradient, i.e.,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \tag{3}$$

Assumption 1 can be replaced by Lipschitz smoothness on $f(\boldsymbol{x};\xi)$, i.e., $\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{y};\xi)\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \xi \sim \mathcal{D}$. Nevertheless, such an assumption is stronger than Assumption 1 because $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \|\mathbb{E}\nabla f(\boldsymbol{x};\xi) - \mathbb{E}\nabla f(\boldsymbol{y};\xi)\| \leq \mathbb{E}\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{y};\xi)\|$.

Assumption 2 *The stochastic sample $\xi \sim \mathcal{D}$ is i.i.d. and $\mathbb{E}\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x})\|^2 \leq \sigma^2$ for any $\boldsymbol{x} \in \mathbb{R}^d$.*

Assumption 1 is the Lipschitz smoothness and Assumption 2 indicates the uniform bounded variance of the stochastic gradient. These two assumptions are standard in the analysis of stochastic optimization algorithms (Bottou et al., 2018).

2.2. Convergence under weaker assumptions

Theorem 1 *Let $(\boldsymbol{x}^t)_{t\geq 0}$ be generated by the SIGNSGD-SIM, and Assumptions 1 and 2 hold. For integer $T \geq 2$, if*

$\gamma = \frac{1}{LT^{3/4}}$ *and* $\theta = 1 - \frac{1}{\sqrt{T}}$*, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{f(\boldsymbol{x}^0) - \min f}{LT^{1/4}} + \frac{2d}{T^{1/4}} + \frac{2\sqrt{d}\sigma}{T^{1/4}} + \frac{2\sqrt{d}\|\boldsymbol{\epsilon}^0\|}{T^{1/2}} + \frac{d}{2T^{3/4}}.$$

We present an explanation on the specific use of the learning rate and momentum: Note that the optimal convergence rate for nonconvex first-order stochastic methods is $\mathcal{O}\Big(\frac{1}{T^{1/4}}\Big)$ (Arjevani et al., 2023). To ensure this rate in our proof and considering the term $\frac{1}{\gamma T}$, we need to choose $\gamma = \Theta\Big(\frac{1}{T^{3/4}}\Big)$. Combining the last two terms, $\frac{\gamma L}{1-\theta}$ and $\sqrt{1-\theta}\sigma$, yielding $1-\theta = \Theta\Big(\frac{1}{T^{1/2}}\Big)$, which leads to the specific choices of learning rate and momentum parameter in Theorem 1.

We want to clarify that while Bernstein et al. (2018a) present a convergence result for signSGD with momentum (Theorem 3), their algorithm requires "warmup," meaning that the first few steps use plain signSGD instead of signSGD with momentum. The number of "warmups" needed depends on the momentum parameter θ in such a way that prevents a possible momentum scheduling. Additionally, their convergence rate is $\mathcal{O}(\frac{1}{T^{1/4}} + \frac{\sigma}{(1-\theta)\sqrt{B}})$, where B is the batchsize, and T denotes the number of iterations, and σ^2 is the variation of the stochastic noise. Consequently, B must grow as $\Theta(T^{1/2})$ to achieve this convergence rate, significantly increasing the sampling costs. In comparison, Theorem 1 has an advantage over Bernstein et al. (2018a) by demonstrating that the same convergence rate $\mathcal{O}(\frac{1}{T^{1/4}})$ can be achieved without increasing batching size.

Theorem 1 indicates that a simple momentum is sufficient to guarantee the convergence of signSGD. The convergence rate is $\mathcal{O}\Big(\frac{f(\boldsymbol{x}^0)-\min f + \sqrt{d}\sigma}{T^{1/4}}\Big)$, which also achieves the same convergence rate as signSGD with error feedback (Karimireddy et al., 2019). An interesting finding is that compared with the signSGD with error feedback, Theorem 1 uses a weaker assumption on the stochastic gradient: we just need the boundedness of the variance, i.e., $\mathbb{E}\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x})\|^2 \leq \sigma^2$; while in (Karimireddy et al., 2019), the authors use a stronger assumption, i.e., $\mathbb{E}\|\nabla f(\boldsymbol{x};\xi)\|^2 \leq \hat{\sigma}^2$ for some $\hat{\sigma} > 0$.

In SGD, the convergence rate can be as fast as

$$\min_{1\leq i\leq T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\| = \mathcal{O}\Big(\frac{\sqrt{f(\boldsymbol{x}^0) - \min f + \sigma^2}}{T^{1/4}}\Big);$$

see (Ghadimi & Lan, 2013) for details. Notice that $\|\boldsymbol{x}\|_1 \leq \sqrt{d}\|\boldsymbol{x}\|$ for any $\boldsymbol{x} \in \mathbb{R}^d$, thus for SGD, we have

$$\min_{1\leq i\leq T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 = \mathcal{O}\Big(\frac{\sqrt{d(f(\boldsymbol{x}^0) - \min f) + d\sigma^2}}{T^{1/4}}\Big),$$

Therefore, we have i) If $\sigma^2 \gg f(\boldsymbol{x}^0) - \min f$, signSGD-SIM can be as fast as SGD. ii) If $\sigma^2 \ll f(\boldsymbol{x}^0) - \min f$, the rate of signSGD-SIM is dominated by the term $\mathcal{O}\Big(\frac{f(\boldsymbol{x}^0)-\min f + \sqrt{d(f(\boldsymbol{x}^0)-\min f)}}{T^{1/4}}\Big)$, and the rate of SGD is dominated by $\mathcal{O}\Big(\frac{\sqrt{d(f(\boldsymbol{x}^0)-\min f)}}{T^{1/4}}\Big)$. In this case, signSGD-SIM is worse than SGD if $f(\boldsymbol{x}^0) - \min f \gg d$. While when $f(\boldsymbol{x}^0) - \min f \ll d$, both algorithms perform similarly. We stress that signSGD-SIM only uses the signs for the update, signSGD-SIM consumes significantly fewer gradient communication costs than the SGD if they need the same iterations to output the desired solution.

In signSGD-SIM, the constant learning rate can be replaced with the coordinate version, i.e., **step 2** in Algorithm 1 can be replaced by the following iteration $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \boldsymbol{A}\cdot \text{Sign}(\boldsymbol{m}^t)$ with $\boldsymbol{A} := \text{Diag}(\gamma_1, \gamma_2, \ldots, \gamma_d)$ and $\gamma_j > 0$ $(j = 1, 2, \ldots, d)$. Using Lemma 1 in the supplementary material, the above variant also enjoys the same convergence rate when $\gamma_j = \Theta(\frac{1}{LT^{3/4}})$ as $j = 1, 2, \ldots, d$.

2.3. Smoothness improves convergence rates

When the objective function is second-order smooth, we can obtain an improved convergence rate for signSGD with momentum compared to that in Section 2.2. Motivated by (Arnold et al., 2019; Cutkosky & Mehta, 2020), we consider the scheme as follows

$$\begin{aligned}
\boldsymbol{y}^t &= \boldsymbol{x}^t + \frac{\theta}{1-\theta}(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}),\\
\boldsymbol{m}^t &= \theta\boldsymbol{m}^{t-1} + (1-\theta)\nabla f(\boldsymbol{y}^t;\xi^t),\\
\boldsymbol{x}^{t+1} &= \boldsymbol{x}^t - \gamma\text{Sign}(\boldsymbol{m}^t).
\end{aligned} \tag{4}$$

Compared with Algorithm 1, (4) uses an extra momentum before calculating the stochastic gradient. The second-order smoothness is used to characterize the Hessian matrix of the objective function, and our theory also relies on the following assumption.

Assumption 3 *The Hessian matrix of function f satisfies $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\|_{\text{op}} \leq \rho\|\boldsymbol{x} - \boldsymbol{y}\|$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.*

With the second-order smoothness assumption, we are prepared to present the improved convergence of scheme (4). In particular, we have the following theorem.

Theorem 2 *Let $(\boldsymbol{x}^t)_{t\geq 0}$ be generated by scheme (4), and Assumptions 1, 2 and 3 hold. For integer $T \geq 2$, if $\theta = 1 - \frac{1}{T^{4/7}}$ and $\gamma = \frac{1}{\max\{\sqrt{\rho}, L\}T^{5/7}}$, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{\max\{\sqrt{\rho}, L\}(f(\boldsymbol{x}^0) - \min f)}{T^{2/7}} + \frac{2\sqrt{d}\sigma + d^{3/2}}{T^{2/7}} + \frac{1}{2T^{5/7}} + \frac{2\sqrt{d}\|\nabla f(\boldsymbol{x}^0)\|}{T^{3/7}}.$$

The convergence rate, in this case, is

$$\mathcal{O}\Big(\frac{f(\boldsymbol{x}^0) - \min f + \sqrt{d}\sigma + d^{3/2}}{T^{2/7}}\Big),$$

which is faster than the speed with only Lipschitz smoothness when d is not very large. The previous sign-based methods mentioned in the introduction do not consider the second-order case. As far as we are aware, our paper is the first to investigate the improved convergence rate of the sign-based method.

3. The Major Vote signSGD with Momentum

This section considers the sign-based methods in distributed settings. In particular, we study two algorithms: the first one is signSGD with major vote and momentum, and the second one further extends the algorithm to the federated learning setting, resulting in Federated MV-signSGD-SIM.

3.1. Major vote

Now we consider the following distributed training task

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}),\ f_i(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} f_i(\boldsymbol{x};\xi), \quad (5)$$

where $\mathcal{D}_i$ denotes the data distribution of the i-th client and $f_i(\boldsymbol{x};\xi)$ is the loss function associated with the training data ξ. The centralized distributed system assumes that there is a Parameter Server (PS) connects n workers, and $\mathcal{D}_i$ is stored in the i-th worker. In Major Vote (MV-) signSGD, each worker stores the parameter with the same initialization. In the t-th iteration, the workers calculate local stochastic gradients and send their signs to the PS. After collecting all sign information, the PS then sends the sign of the average to all workers. All information transmitted in the networks is signed, leading to remarkable communication efficiency.

Unlike the distributed SGD, the PS only processes the information received from the workers but does not update the parameter in MV-signSGD. The parameter update is performed in all workers. Because the workers get the same feedback from the PS, the parameters in all workers are identical, given the same initialization.

We cannot directly extend Algorithm 1 to the major vote scheme presented in (Bernstein et al., 2018b), i.e., replacing Sign$[\cdot]$ with Sign$[\sum_{i=1}^n \text{Sign}(\cdot)]$ in Algorithm 1 because the two-layer sign operator breaks some critical properties. Alternatively, we use the stochastic method to sign the gradient information introduced by Jin et al. (2020). We denote the stochastic sign method employed major vote signSGD in (Jin et al., 2020) as MV-sto-signSGD. Before establishing our theoretical convergence of MV-sto-signSGD, we present another necessary assumption below.

Assumption 4 *For any $\boldsymbol{x}\in\mathbb{R}^d$, and $\xi\sim\mathcal{D}$, the stochastic gradient satisfies $\|\nabla f(\boldsymbol{x};\xi)\|^2 \leq R^2$ with $R > 0$.*

In many neural network training tasks, the stochastic gradient is usually bounded due to the activation functions. When Assumption 4 holds, we have $\|\nabla f(\boldsymbol{x})\|^2 = \|\mathbb{E}\nabla f(\boldsymbol{x};\xi)\|^2 \leq R^2$ and $\mathbb{E}\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x})\|^2 \leq 2\mathbb{E}\|\nabla f(\boldsymbol{x};\xi)\|^2 + 2\mathbb{E}\|\nabla f(\boldsymbol{x})\|^2 \leq 4R^2$. Thus, Assumption 4 implies Assumption 2. This assumption is also used to analyze the convergence of distributed signSGD in (Jin et al., 2020).

Given a d-dimensional vector $\boldsymbol{v}$ that satisfies $\|\boldsymbol{v}\| \leq R$, we denote a stochastic sign operator as

$$[\mathcal{S}_R(\boldsymbol{v})]_i = \begin{cases} -\text{Sign}(\boldsymbol{v}_i), \text{ with probability } \frac{1}{2} - \frac{|\boldsymbol{v}_i|}{2R}, \\ \ \ \text{Sign}(\boldsymbol{v}_i), \text{ with probability } \frac{1}{2} + \frac{|\boldsymbol{v}_i|}{2R}. \end{cases} \quad (6)$$

It is easy to see that $\mathcal{S}_R(\boldsymbol{v})$ only contains signs but follows $\mathbb{E}(\mathcal{S}_R(\boldsymbol{v})) = \frac{\boldsymbol{v}}{R}$. The unbiased expectation yields favorable properties in the proof.

In MV-signSGD with momentum for this distributed optimization, each worker receives the same information from the parameter server for updating and thus enjoys the same parameter if we use the same initialization for all workers. To this end, we can use $\boldsymbol{y}^t$ and $\boldsymbol{x}^t$ to denote the parameters in all workers. We formulate MV-sto-signSGD-SIM in Algorithm 2.

Algorithm 2 Major Vote signSGD with stochastic SImple Momentum (MV-sto-signSGD-SIM)

```
Require: parameters $\gamma > 0$, $0 \leq \theta < 1$, $\alpha \geq 0$ , $R > 0$
  Initialization: $\boldsymbol{x}^0 = \boldsymbol{y}^0 = \boldsymbol{0}$, $\boldsymbol{m}^0 = \boldsymbol{0}$
  for $t = 1, 2, \ldots$
    step 1: All workers update $\boldsymbol{y}^t = \boldsymbol{x}^t + \alpha(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$
    step 2: Worker $i$ samples $\xi^t(i) \sim \mathcal{D}_i$ i.i.d and update
        $\boldsymbol{m}^t(i) = \theta\boldsymbol{m}^{t-1}(i) + (1-\theta)\nabla f_i(\boldsymbol{y}^t;\xi^t(i))$
    step 3: PS pulls $[\mathcal{S}_R(\boldsymbol{m}^t(i))]_{1\leq i\leq n}$ and update
        $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \gamma\text{Sign}\Big[\sum_{i=1}^n \mathcal{S}_R(\boldsymbol{m}^t(i))\Big]$
  end for
```

Indeed, we consider two types of "momentum": one is Algorithm 1 (single momentum), and the other one is the scheme in equation (4) (double momentum). In Algorithm 2, we employ the parameter α to combine single momentum and double momentum for major vote variants, where when $\alpha = 0$, Algorithm 2 becomes single momentum while $\alpha > 0$ means the double one. Therefore, the case when $\alpha > 0$ is used to prove the nonconvex acceleration, but we still presented the results for $\alpha = 0$.

We are prepared to present the convergence of MV-sto-signSGD with momentum. In particular, we have

Theorem 3 *Let $(\boldsymbol{x}^t)_{t\geq 0}$ be generated by MV-sto-signSGD-SIM, Assumptions 1, 2, and 4 hold for $\xi^t(i)$ and*

$\nabla f_i(\cdot;\cdot)$. *For integer* $T \geq 2$*, if* $\alpha = 0$, $\gamma = \frac{1}{LT^{3/4}}$*, and* $\theta = 1 - \frac{1}{\sqrt{T}}$*, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{f(\boldsymbol{x}^1) - \min f}{LT^{1/4}} + \frac{2d\gamma}{T^{1/4}} + \frac{2\sqrt{d}\sigma}{\sqrt{n}T^{1/4}} + \frac{d}{2T^{3/4}} + \frac{2\sqrt{d}\|\nabla f(\boldsymbol{x}^0)\|}{T^{1/2}} + \frac{2dR}{\sqrt{n}}.$$

Furthermore, if Assumptions 1, 2, 3, and 4 hold for $\xi^t(i)$ *and* $\nabla f_i(\cdot;\cdot)$*, and* $\alpha = \frac{\theta}{1-\theta}$, $1-\theta = \frac{1}{T^{4/7}}$ *and* $\gamma = \frac{1}{\max\{\sqrt{\rho},L\}T^{5/7}}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{\max\{\sqrt{\rho},L\}[f(\boldsymbol{x}^0) - \min f]}{T^{2/7}} + \frac{d^{3/2}}{T^{2/7}} + \frac{2\sqrt{d}\sigma}{\sqrt{n}T^{2/7}} + \frac{d}{2T^{5/7}} + \frac{2\sqrt{d}\|\nabla f(\boldsymbol{x}^0)\|}{T^{3/7}} + \frac{2dR}{\sqrt{n}}.$$

The sign operator used in our algorithm is slightly different from that (Bernstein et al., 2018a;b) and identical to the one used in (Jin et al., 2020). Compared with the MV-SIGNSGD (Bernstein et al., 2018a;b) and MV-STO-SIGNSGD (Jin et al., 2020), our algorithm achieves the same rate without increasing the batch size in each iteration when $\alpha = 0$. While when $\alpha = \frac{\theta}{1-\theta}$, the algorithm enjoys faster convergence with second-order smooth properties. Although the error-feedback MV-STO-SIGNSGD has been proposed in (Jin et al., 2020), it is required to send non-sign compressed information, while in our algorithm, all information transmitted is signed. Theorem 3 shows that the error bound can be improved if the number of workers n increases, which means that the algorithm is friendly for large-scale distributed training tasks.

Regarding the majority vote algorithm, we have proved a better rate under weaker assumptions than Bernstein et al. (2018a). They assume that the gradient noise is symmetric and unimodal, which is removed for our momentum majority vote scheme. The convergence rate in Bernstein et al. (2018a) is $\mathcal{O}\left(\frac{1}{T^{1/4}} + \frac{\sigma}{\sqrt{Bn}}\right)$, where n is the number of nodes. Hence, to reach ϵ error for $\min_{1\leq t\leq T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1$, one needs to set $T = \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$ and $B = \Theta\left(\frac{1}{n\epsilon^2}\right)$ for majority vote signSGD. While our algorithm only needs $T = \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$, also saving significant sampling costs.

3.2. Towards a federated formulation

We consider the variant of the major vote for federated learning. Specifically, each node performs local SGD several times (we set it to be K) before sending the signs to the PS; that is, in each iteration, each worker performs the following iteration

$$\boldsymbol{z}^{t,k+1}(i) = \boldsymbol{z}^{t,k}(i) - \eta\nabla f_i(\boldsymbol{z}^{t,k}(i);\xi^{t,k}(i))$$

with $\boldsymbol{z}^{t,0}(i) = \boldsymbol{y}^t$. The federated version performs a similar scheme as Algorithm 2 and modifies the momentum updating as follows

$$\boldsymbol{m}^t(i) = \theta\boldsymbol{m}^{t-1}(i) + (1-\theta)\nabla f_i(\boldsymbol{y}^t(i);\xi^t(i))$$

with $\boldsymbol{y}^t(i) = \boldsymbol{z}^{t,K}(i)$. We formulate MV-STO-SIGNSGD with momentum in Algorithm 3. The convergence results of the federated MV-STO-SIGNSGD with momentum is presented in Theorem 4.

Algorithm 3 Federated Major Vote stochastic SIGNSGD with SImple Momentum (Federated MV-STO-SIGNSGD-SIM)

```
Require: parameters $\gamma > 0$, $\alpha \geq 0$, $0 \leq \theta < 1$, $R > 0$, $K > 0$
  Initialization: $\boldsymbol{x}^0 = \boldsymbol{y}^0 = \boldsymbol{0}$, $\boldsymbol{m}^0 = \boldsymbol{0}$
  for $t = 1, 2, \ldots$
    step 1: All workers update $\boldsymbol{y}^t = \boldsymbol{x}^t + \alpha(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$ and set $\boldsymbol{z}^{t,0}(i) = \boldsymbol{y}^t$ for $i \in \{1, 2, \ldots, n\}$
    step 2: Worker $i$
      for $k = 1, 2, \ldots, K$
        samples $\xi^{t,k}(i) \sim \mathcal{D}_i$ i.i.d and updates
          $\boldsymbol{z}^{t,k+1}(i) = \boldsymbol{z}^{t,k}(i) - \eta\nabla f_i(\boldsymbol{z}^{t,k}(i);\xi^{t,k}(i))$
      end for outputs $\boldsymbol{y}^t(i) = \boldsymbol{z}^{t,K}(i)$
      samples $\xi^t(i) \sim \mathcal{D}_i$ i.i.d and update
          $\boldsymbol{m}^t(i) = \theta\boldsymbol{m}^{t-1}(i) + (1-\theta)\nabla f_i(\boldsymbol{y}^t(i);\xi^t(i))$
    step 3: PS pulls $[\mathcal{S}_R(\boldsymbol{m}^t(i))]_{1\leq i\leq n}$ and update
          $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \gamma\mathrm{Sign}\left[\sum_{i=1}^n \mathcal{S}_R(\boldsymbol{m}^t(i))\right]$
  end for
```

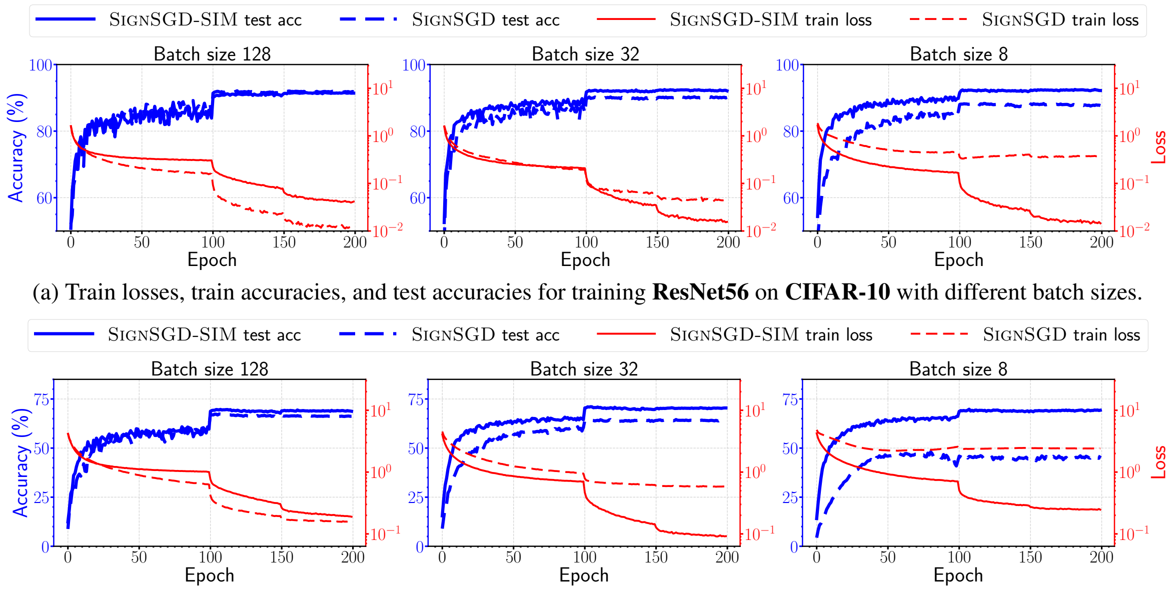
Theorem 4 *Let* $(\boldsymbol{x}^t)_{t\geq 0}$ *be generated by federated MV-SIGNSGD with momentum, Assumptions 1, 2 and 4 hold for* $\xi^t(i)$ *and* $\nabla f_i(\cdot;\cdot)$*. For integer* $T \geq 2$*, if* $\alpha = 0$, $\gamma = \frac{1}{LT^{3/4}}$, $\theta = 1 - \frac{1}{\sqrt{T}}$, $\eta = \frac{1}{4LK^2}$*, and* $K = T^{1/4}$*, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{f(\boldsymbol{x}^1) - \min f}{LT^{1/4}} + \frac{2d\gamma}{RT^{1/4}} + \frac{d}{2T^{3/4}} + \frac{2\sqrt{d}\sigma}{\sqrt{n}T^{1/4}} + \frac{2\sqrt{d}\|\nabla f(\boldsymbol{x}^0)\|}{T^{1/2}} + \frac{2dR}{\sqrt{n}} + \frac{\sqrt{2\sigma^2 + R^2}}{T^{1/4}}.$$

Furthermore, if Assumptions 1, 2, 3 and 4 hold, and $\alpha = \frac{\theta}{1-\theta}$*, if* $1-\theta = \frac{1}{T^{4/7}}$, $\gamma = \frac{1}{\max\{\sqrt{\rho},L\}T^{5/7}}$*, and* $\eta = \frac{1}{4LK^2}$*, and* $K = T^{1/4}$*, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{\max\{\sqrt{\rho},L\}[f(\boldsymbol{x}^0) - \min f]}{T^{2/7}} + \frac{d^{3/2}}{T^{2/7}} + \frac{2\sqrt{d}\sigma}{\sqrt{n}T^{2/7}} + \frac{\max\{\sqrt{\rho},L\}Ld}{2T^{5/7}} + \frac{2\sqrt{d}\|\nabla f(\boldsymbol{x}^0)\|}{T^{3/7}} + \frac{2dR}{\sqrt{n}} + \frac{\sqrt{2\sigma^2 + R^2}}{T^{1/4}}.$$

The results show that the federated MV-STO-SIGNSGD with momentum can be as fast as MV-STO-SIGNSGD with



(a) Train losses, train accuracies, and test accuracies for training **ResNet56** on **CIFAR-10** with different batch sizes.

(b) Train losses, train accuracies, and test accuracies for training **ResNet56** on **CIFAR-100** with different batch sizes.

Figure 2. Performance comparison between SIGNSGD-SIM and SIGNSGD on training ResNet56: in sharp contrast, SIGNSGD-SIM maintains its performance with smaller batch sizes while the performance of SIGNSGD drops significantly.

momentum. As the number of workers n increases, the algorithm can achieve better convergence rates. Compared with the federated algorithm in (Jin et al., 2020), our method does not require a large batch size in each iteration.

4. Numerical Results

We provide numerical verification of our convergence result in Theorems 1, 3, and 4. The results confirm that i) compared with SIGNSGD (MV-STO-SIGNSGD), the added momentum in SIGNSGD-SIM and (MV-STO-SIGNSGD-SIM and its federated variant) maintains good performance with small batch sizes and ii) the training of MV-STO-SIGNSGD-SIM and its federated variant can be benefited from more workers (see Appendix G.2).

4.1. Experimental evaluation of SIGNSGD-SIM

Models and dataset. We train various ResNet models from (He et al., 2016) on CIFAR-10/CIFAR-100 (Krizhevsky et al., 2009) with the optimization algorithms SIGNSGD and SIGNSGD-SIM. The CIFAR-10 and CIFAR-100 are popular image classification datasets that consist of small, 32x32 pixel colored images. CIFAR-10 contains 60,000 images, divided into 10 classes with 6,000 images per class. CIFAR-100 contains 60,000 images but is divided into 100 classes with 600 images per class. Both datasets are split into a training set of 50,000 images and a test set of 10,000 images. Our code is based on open-source libraries[2].

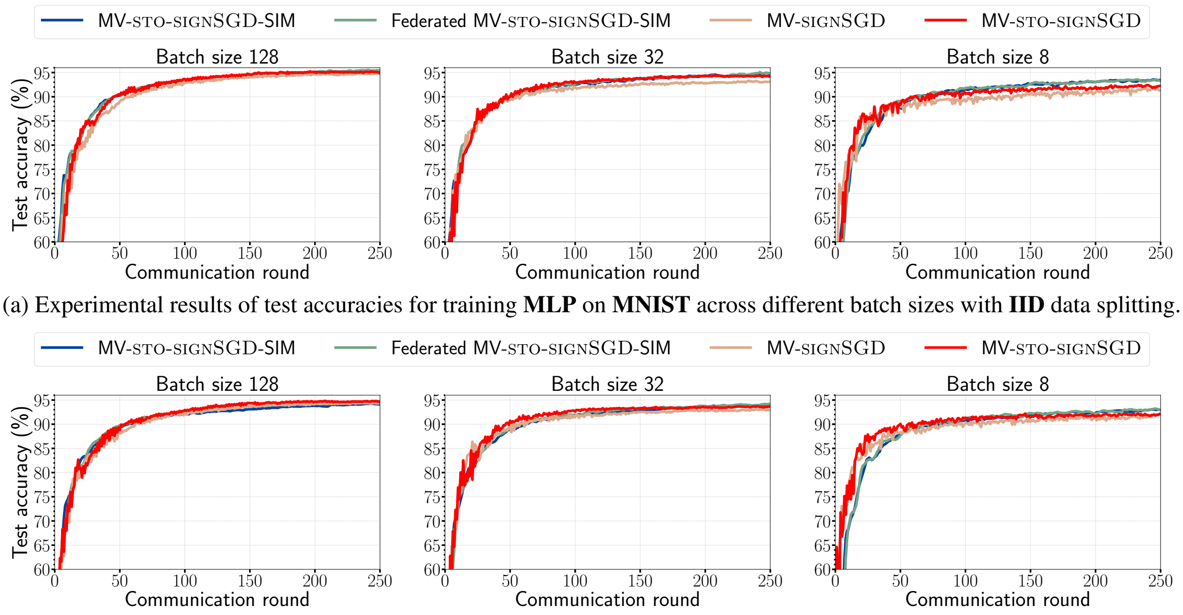
The result of training ResNet110 on CIFAR-100 is already presented in Figure 1 in the introduction where SIGNSGD-SIM significantly outperforms SIGNSGD when the batch size is small. Indeed, similar phenomena occur when training ResNet20/32/56 on CIFAR-10/CIFAR-100. We present the result of ResNet56 (see Table 2 and Figure 2) in this section and leave the results of ResNet20/32 in Appendix G.1 for the sake of space.

Algorithms	Batch Size		
	128	32	8
	CIFAR-10		
SGD-M	92.37%	93.89%	93.80%
signSGD	92.12%	90.32%	88.32%
signSGD-SIM	91.60%	92.48%	92.51%
	CIFAR-100		
SGD-M	68.22%	71.71%	72.16%
signSGD	67.36%	65.54%	48.09%
signSGD-SIM	69.70%	71.07%	69.79%

Table 2. The testing accuracies of training **ResNet56** on **CIFAR-10/CIFAR-100** with different batch sizes using SIGNSGD, and SIGNSGD-SIM. We also include the results with (uncompressed) gradient descent with momentum (SGD-M) as a reference.

Training parameters. All of the algorithms in the study are

[2] github.com/akamaster/pytorch_resnet_cifar10, github.com/epfml/error-feedback-SGD



(a) Experimental results of test accuracies for training **MLP** on **MNIST** across different batch sizes with **IID** data splitting.

(b) Experimental results of test accuracies for training **MLP** on **MNIST** across different batch sizes with **Non-IID** data splitting.

Figure 3. In both **IID** and **Non-IID** cases, MV-STO-SIGNSGD-SIM and Federated MV-STO-SIGNSGD-SIM maintains better performance for smaller batch sizes.

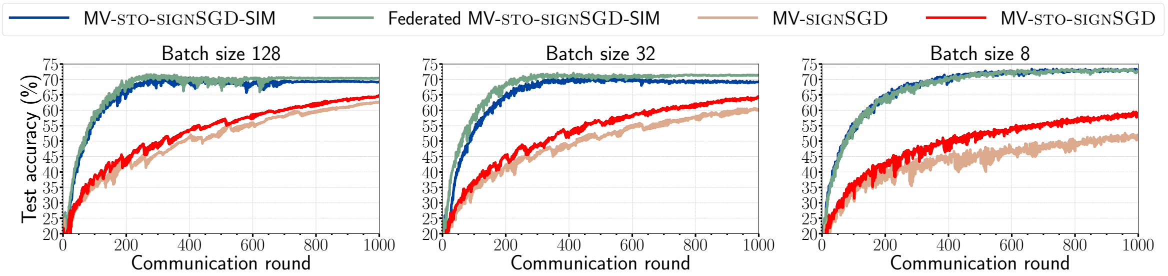


Figure 4. Experimental results of test accuracies for training **CNN** on **CIFAR-10** across different batch sizes. Both MV-STO-SIGNSGD-SIM and Federated MV-STO-SIGNSGD-SIM obtain excellent performance across batch sizes. In sharp contrast, both MV-STO-SIGNSGD and MV-SIGNSGD train much slower with deteriorating accuracies for smaller batch sizes.

run for a total of 200 epochs. The learning rate is decimated twice during this time, first at 100 epochs and again at 150 epochs. The initial learning rate for a batch size of 128 is 1×10^{-3}. For smaller batch sizes, the learning rate was proportionally reduced, as suggested in (Goyal et al., 2017) and adopted in a previous study (Karimireddy et al., 2019). The momentum parameter of SIGNSGD-SIM is set to 0.9, and the weight decay for both algorithms is set to 1×10^{-4}.

Results. The averaged results from 5 independent runs are reported in Table 2 and Figure 2. As shown in the results, SIGNSGD experiences convergence issues when the batch size is small, while SIGNSGD-SIM consistently performs well regardless of the batch size. These results align with previous findings across various architectures and datasets ((Karimireddy et al., 2019; Bernstein et al., 2018b)) that SIGNSGD-SIM is not as sensitive to batch size as SIGNSGD, supporting Theorem 1.

4.2. Experimental evaluation of MV-STO-SIGNSGD-SIM and its federated variant

In this section, we run experiments comparing MV-STO-SIGNSGD-SIM (Algorithm 2) and its federated variant Federated MV-STO-SIGNSGD-SIM (Algorithm 3) with MV-SIGNSGD (Bernstein et al., 2018a;b) and MV-STO-SIGNSGD (Jin et al., 2020) on training Multi-Layer Perceptron (MLP) on MNIST (LeCun et al., 1998) and Convolutional neural network (CNN) on CIFAR-10 dataset.

Multi-layer perceptron and MNIST. We train an MLP with 2-hidden layers with 64 units each using ReLU activations on the MNIST dataset. MNIST is a dataset of 28×28 grayscale images of digits from 0 to 9, containing $60,000$ training samples and $10,000$ testing samples. We investigate two methods for dividing the MNIST data among clients: **IID**, where the data is shuffled and evenly distributed into 50 clients, and **Non-IID**, where the clients receive data according to a Dirichlet distribution with concentration parameter $\alpha = 1$ (Hsu et al., 2019).

Convolutional neural network and CIFAR-10. We train a CNN model on CIFAR-10. The model has two 5×5 convolutional layers with max-pooling, two fully-connected layers with 384 and 192 units, respectively, and a final softmax output layer. The $50,000$ training images are shuffled and evenly distributed to 100 clients in the **IID** fashion.

Training parameters. All the optimization algorithms in the study run for a total of 250 (1000) commutation rounds on MNIST (CIFAR-10) experiments with 50 (100) workers. We tune the learning rate from the set $\{0.1, 0.01, 0.001, 0.0001\}$ for each algorithm in both experiments. The optimal learning rates for MV-STO-SIGNSGD and MV-SIGNSGD are 0.01 for MNIST and 0.001 for CIFAR-10. The optimal learning rate for MV-STO-SIGNSGD-SIM and Federated MV-STO-SIGNSGD-SIM is 0.001 for all experiments. We adopt the constant learning rate in the MNIST experiment and a decaying learning rate at the rate of 0.999 per commutation round for the CIFAR-10 experiment. For MV-STO-SIGNSGD, MV-STO-SIGNSGD-SIM and Federated MV-STO-SIGNSGD-SIM, we fix the parameter R to be 0.001. In the algorithms SIGNSGD-SIM and Federated MV-STO-SIGNSGD-SIM, the momentum parameter θ is set to 0.9, and the parameter α is set to 0. The Federated MV-STO-SIGNSGD-SIM performs 5 local iterations on each worker, i.e. $K = 5$.

Results. The averaged results from 5 independent runs are reported in Figure 3 and Figure 4. Both MV-STO-SIGNSGD-SIM and its federated variant maintain good performance with smaller batch sizes validating Theorem 4.

5. Concluding Remarks

In this paper, we first provide a theoretical interpretation of why momentum benefits the convergence of SIGNSGD; in particular, we show that SIGNSGD with momentum guarantees convergence with smaller batch size than the vanilla SIGNSGD, echoing existing numerical evidence. We further extend our theory to SIGNSGD in distributed settings, including SIGNSGD with major vote and SIGNSGD in a federated learning scenario. We verify our theory with various numerical experiments in different settings.

Acknowledgements

Dongsheng Li and Tao Sun are supported in part by the National Science Foundation of China (62025208), Hunan Provincial Natural Science Foundation of China (2022JJ10065), and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References

Al-Dujaili, A. and O'Reilly, U.-M. Sign bits are all you need for black-box attacks. In *International Conference on Learning Representations*, 2020.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199 (1-2):165–214, 2023.

Arnold, S., Manzagol, P. A., Babanezhad, R., Mitliagkas, I., and Roux, N. L. Reducing the variance in online optimization by transporting past gradients. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

Balles, L. and Hennig, P. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 10–15 Jul 2018.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signSGD: Compressed optimisation for non-convex problems. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 10–15 Jul 2018a.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. signsgd with majority vote is communication efficient and fault tolerant. *arXiv*, 2018b.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Crawshaw, M., Liu, M., Orabona, F., Zhang, W., and Zhuang, Z. Robustness to unbounded smoothness of generalized signsgd. *arXiv preprint arXiv:2208.11195*, 2022.

Cutkosky, A. and Mehta, H. Momentum improves normalized SGD. In *International Conference on Machine Learning*. PMLR, 2020.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM*

Journal on Optimization, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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*, pp. 770–778, 2016.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

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

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error feedback fixes SignSGD and other gradient compression schemes. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, 09–15 Jun 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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

Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., and Su, B.-Y. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation*, OSDI'14, pp. 583–598, USA, 2014. USENIX Association. ISBN 9781931971164.

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, B. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

Liu, S., Chen, P.-Y., Chen, X., and Hong, M. signSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

Safaryan, M. and Richtarik, P. Stochastic Sign descent methods: New algorithms and better theory. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9224–9234. PMLR, 18–24 Jul 2021.

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 *Conference of the International Speech Communication Association*, 2014.

Sohn, J.-y., Han, D.-J., Choi, B., and Moon, J. Election coding for distributed learning: Protecting signsgd against byzantine attacks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14615–14625. Curran Associates, Inc., 2020.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified sgd with memory. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Strom, N. Scalable distributed dnn training using commodity gpu cloud computing. In *Interspeech 2015*, 2015.

Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z.-Q. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*.

Supplementary materials for

Momentum Ensures Convergence of SIGNSGD *under Weaker Assumptions*

A. Technical Lemmas

Lemma 1 *Let $\boldsymbol{x}^\dagger, \boldsymbol{m} \in \mathbb{R}^d$ be arbitrary vectors, and $\boldsymbol{A} = \mathrm{Diag}(a_1, a_2, \ldots, a_d) \in \mathbb{R}^d$ with $a_i > 0$, $i \in \{1, 2, \ldots, d\}$. We denote*

$$\boldsymbol{x}^\ddagger = \boldsymbol{x}^\dagger - \gamma \boldsymbol{A} \cdot \mathrm{Sign}(\boldsymbol{m}), \tag{7}$$

and $\boldsymbol{\epsilon} := \boldsymbol{m} - \nabla f(\boldsymbol{x}^\dagger)$. If Assumption 1 holds, we have

$$f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) \leq -\gamma \|\boldsymbol{A}\nabla f(\boldsymbol{x}^\dagger)\|_1 + 2\gamma\|\boldsymbol{A}\|_F\|\boldsymbol{\epsilon}\| + \frac{L\gamma^2\|\boldsymbol{A}\|_F^2}{2}.$$

Specifically, if $\boldsymbol{A}$ is the identity matrix,

$$f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) \leq -\gamma \|\nabla f(\boldsymbol{x}^\dagger)\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}\| + \frac{L\gamma^2 d}{2}.$$

Lemma 2 *Let $\boldsymbol{x}^\dagger$ be an arbitrary vector, and $\boldsymbol{m}(i) \in \mathbb{R}^d$ be a vector associated with node i such that $\|\boldsymbol{m}(i)\| \leq R$. If we denote*

$$\boldsymbol{x}^\ddagger = \boldsymbol{x}^\dagger - \gamma \cdot \mathrm{Sign}(\sum_{i=1}^n \mathcal{S}_R(\boldsymbol{m}(i))),$$

and $\boldsymbol{\epsilon} := \frac{\sum_i^n \boldsymbol{m}(i)}{nR} - \frac{\nabla f(\boldsymbol{x}^\dagger)}{R}$, it then holds

$$\mathbb{E}f(\boldsymbol{x}^\ddagger) - \mathbb{E}f(\boldsymbol{x}^\dagger) \leq -\gamma\mathbb{E}\|\nabla f(\boldsymbol{x}^\dagger)\|_1 + 2\sqrt{d}\gamma R\mathbb{E}\|\boldsymbol{\epsilon}\| + \frac{2d\gamma R}{\sqrt{n}} + \frac{L\gamma^2 d}{2}.$$

Lemma 3 *Let Assumptions 1, 2 and 4 hold for $\xi^t(i)$ and $\nabla f_i(\cdot;\cdot)$. Assume node i performs local SGD as*

$$\boldsymbol{z}^{t,k+1}(i) = \boldsymbol{z}^{t,k}(i) - \eta\nabla f_i(\boldsymbol{z}^{t,k}(i); \xi^{t,k}(i))$$

with $\boldsymbol{z}^{t,0}(i) = \boldsymbol{y}^t$. Since $0 < \eta \leq \frac{1}{4LK}$, it holds

$$\mathbb{E}\|\boldsymbol{y}^t(i) - \boldsymbol{y}^t\| \leq \eta\sqrt{8K\sigma^2 + 16K^2\sigma^2 + 16K^2R^2}.$$

B. Proof of Theorem 1

We use the notations $\boldsymbol{g}^t := \nabla f(\boldsymbol{x}^t; \xi^t)$, and $\boldsymbol{\epsilon}^t := \boldsymbol{m}^t - \nabla f(\boldsymbol{x}^t)$, and $\boldsymbol{\delta}^t := \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t)$. Therefore,

$$\begin{aligned} \boldsymbol{m}^t &= \theta\boldsymbol{m}^{t-1} + (1-\theta)\boldsymbol{g}^t = \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^{t-1})) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t)) \\ \Rightarrow \boldsymbol{\epsilon}^t &= \boldsymbol{m}^t - \nabla f(\boldsymbol{x}^t) = \theta\boldsymbol{\epsilon}^{t-1} + \theta\underbrace{(\nabla f(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^t))}_{:=\boldsymbol{s}^t} + (1-\theta)\boldsymbol{\delta}^t. \end{aligned}$$

That is,

$$\boldsymbol{\epsilon}^t = \theta\boldsymbol{\epsilon}^{t-1} + \boldsymbol{s}^t + (1-\theta)\boldsymbol{\delta}^t.$$

Notice that, the smoothness of the gradient implies

$$\|\boldsymbol{s}^t\| = \|\nabla f(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^t)\| \leq L\|\boldsymbol{x}^{t-1} - \boldsymbol{x}^t\| \leq L\sqrt{d}\gamma. \tag{8}$$

Using Mathematical Induction, we have

$$\boldsymbol{\epsilon}^t = \theta^t\boldsymbol{\epsilon}^0 + \sum_{i=1}^t \theta^{t-i}\boldsymbol{s}^i + (1-\theta)\sum_{i=1}^t \theta^{t-i}\boldsymbol{\delta}^i. \tag{9}$$

By taking the norms of both sides of inequality (9) and use the bound of $\boldsymbol{s}^i$ in inequality (8), we obtain

$$\|\boldsymbol{\epsilon}^t\| \le L\sqrt{d}\gamma\sum_{i=1}^{t}\theta^{t-i} + (1-\theta)\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\| + \theta^t\|\boldsymbol{\epsilon}^0\|.$$

By further taking the expectations of both sides of the above inequality, there is

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \le \frac{L\sqrt{d}\gamma}{1-\theta} + (1-\theta)\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\| + \theta^t\|\boldsymbol{\epsilon}^0\|. \tag{10}$$

Notice that the random variables $(\boldsymbol{\delta}^i)_{1\le i\le t}$ are independent. We have

$$\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\| \le \sqrt{\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\|^2} = \sqrt{\mathbb{E}\sum_{i=1}^{t}\theta^{2t-2i}\|\boldsymbol{\delta}^i\|^2} \le \frac{\sigma}{\sqrt{1-\theta^2}},$$

by using Cauchy's inequality and the fact that $\mathbb{E}\|\boldsymbol{\delta}^i\|^2 \le \sigma^2$. Finally, we obtain the following inequality by plugging the above result into inequality (10):

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \le \frac{L\sqrt{d}\gamma\theta}{1-\theta} + \frac{\sqrt{1-\theta}}{\sqrt{1+\theta}}\sigma + \theta^t\|\boldsymbol{\epsilon}^0\| \le \frac{L\sqrt{d}\gamma}{1-\theta} + \sqrt{1-\theta}\sigma + \theta^t\|\boldsymbol{\epsilon}^0\|.$$

Using Lemma 1 with $\boldsymbol{x}^\dagger \to \boldsymbol{x}^t$ and $\boldsymbol{m} \to \boldsymbol{m}^t$, we obtain

$$f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^t) \le -\gamma\|\nabla f(\boldsymbol{x}^t)\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}^t\| + \frac{L\gamma^2 d}{2}.$$

Summing the recursion from $t = 1$ to T and taking expectations,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \le \frac{f(\boldsymbol{x}^1) - \min f}{\gamma T} + \frac{2Ld\gamma}{1-\theta} + 2\sqrt{d}\sqrt{1-\theta}\sigma + \frac{L\gamma d}{2} + 2\sqrt{d}\sum_{t=1}^{T}\theta^t\|\boldsymbol{\epsilon}^0\|/T.$$

As $1-\theta = \frac{1}{\sqrt{T}}$ and $\gamma = \frac{1}{LT^{3/4}}$, we then have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \le \frac{f(\boldsymbol{x}^1) - \min f}{LT^{1/4}} + \frac{2d}{T^{1/4}} + \frac{2\sqrt{d}\sigma}{T^{1/4}} + \frac{2\sqrt{d}\|\boldsymbol{\epsilon}^0\|}{T^{1/2}} + \frac{d}{2T^{3/4}}.$$

Notice that $\boldsymbol{x}^1 = \boldsymbol{x}^0$ due to the initialization $\boldsymbol{m}^0 = \boldsymbol{0}$, we then proved the result.

C. Proof of Theorem 2

We adopt the following notations: $\boldsymbol{g}^t := \nabla f(\boldsymbol{y}^t;\xi^t)$, and $\boldsymbol{\epsilon}^t := \boldsymbol{m}^t - \nabla f(\boldsymbol{x}^t)$, and $\boldsymbol{\delta}^t := \boldsymbol{g}^t - \nabla f(\boldsymbol{y}^t)$, and $\mathbf{H}(\boldsymbol{y},\boldsymbol{x}) := \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - [\nabla^2 f(\boldsymbol{x})](\boldsymbol{y}-\boldsymbol{x})$. Recall $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\|_{\text{op}} \le \rho\|\boldsymbol{x}-\boldsymbol{y}\|$ for $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$ from Assumption 3, it is then easy to see

$$\begin{aligned}\|\mathbf{H}(\boldsymbol{y},\boldsymbol{x})\| &= \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - [\nabla^2 f(\boldsymbol{x})](\boldsymbol{y}-\boldsymbol{x})\|\\ &= \|\int_{s=0}^{1}[\nabla^2 f(\boldsymbol{x}+(\boldsymbol{y}-\boldsymbol{x})s) - \nabla^2 f(\boldsymbol{x})](\boldsymbol{x}-\boldsymbol{y})ds\|\\ &\le \int_{s=0}^{1}\|\nabla^2 f(\boldsymbol{x}+(\boldsymbol{y}-\boldsymbol{x})s) - \nabla^2 f(\boldsymbol{x})\|_{\text{op}}\|\boldsymbol{x}-\boldsymbol{y}\|ds\\ &\le \rho\int_{s=0}^{1}(1-s)\|\boldsymbol{x}-\boldsymbol{y}\|^2 ds = \frac{\rho}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2.\end{aligned}$$

Through direct computations, we have

$$\begin{aligned}
\boldsymbol{m}^t &= \theta \boldsymbol{m}^{t-1} + (1-\theta)\boldsymbol{g}^t = \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^{t-1})) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{y}^t))\\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^t) + [\nabla^2 f(\boldsymbol{x}^t)](\boldsymbol{x}^{t-1} - \boldsymbol{x}^t) + \mathbf{H}(\boldsymbol{x}^{t-1}, \boldsymbol{x}^t))\\
&\qquad + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t) + [\nabla^2 f(\boldsymbol{x}^t)] \underbrace{(\boldsymbol{y}^t - \boldsymbol{x}^t)}_{=\frac{\theta}{1-\theta}(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})} + \mathbf{H}(\boldsymbol{y}^t, \boldsymbol{x}^t))\\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^t) + \mathbf{H}(\boldsymbol{x}^{t-1}, \boldsymbol{x}^t)) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t) + \mathbf{H}(\boldsymbol{y}^t, \boldsymbol{x}^t))\\
\Rightarrow \boldsymbol{\epsilon}^t &= \boldsymbol{m}^t - \nabla f(\boldsymbol{x}^t) = \theta\boldsymbol{\epsilon}^{t-1} + \theta\mathbf{H}(\boldsymbol{x}^{t-1}, \boldsymbol{x}^t) + (1-\theta)\mathbf{H}(\boldsymbol{y}^t, \boldsymbol{x}^t) + (1-\theta)\boldsymbol{\delta}^t.
\end{aligned}$$

Using Mathematical Induction, we have

$$\boldsymbol{\epsilon}^t = \theta^t \boldsymbol{\epsilon}^0 + \theta \sum_{i=1}^t \theta^{t-i}\mathbf{H}(\boldsymbol{x}^{i-1}, \boldsymbol{x}^i) + (1-\theta)\sum_{i=1}^t \theta^{t-i}\mathbf{H}(\boldsymbol{y}^i, \boldsymbol{x}^i) + (1-\theta)\sum_{i=1}^t \theta^{t-i}\boldsymbol{\delta}^i.$$

Taking the norms and expectations of both sides of the above equation, we obtain

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\epsilon}^t\| &\le \theta\sum_{i=1}^t \theta^{t-i}\mathbb{E}\|\mathbf{H}(\boldsymbol{x}^{i-1}, \boldsymbol{x}^i)\| + (1-\theta)\sum_{i=1}^t \theta^{t-i}\mathbb{E}\|\mathbf{H}(\boldsymbol{y}^i, \boldsymbol{x}^i)\| + \mathbb{E}\left\|(1-\theta)\sum_{i=1}^t \theta^{t-i}\boldsymbol{\delta}^i\right\| + \theta^t\|\boldsymbol{\epsilon}^0\|\\
&\le \frac{\rho}{2}\theta\sum_{i=1}^t \theta^{t-i}\mathbb{E}\|\boldsymbol{x}^{i-1} - \boldsymbol{x}^i\|^2 + \frac{\rho}{2}(1-\theta)\sum_{i=1}^t \theta^{t-i}\mathbb{E}\|\boldsymbol{y}^i - \boldsymbol{x}^i\|^2 + \mathbb{E}\left\|(1-\theta)\sum_{i=1}^t \theta^{t-i}\boldsymbol{\delta}^i\right\| + \theta^t\|\boldsymbol{\epsilon}^0\|\\
&\le \frac{\rho}{2}\frac{\theta}{1-\theta}\sum_{i=1}^t \theta^{t-i}\mathbb{E}\|\boldsymbol{x}^{i-1} - \boldsymbol{x}^i\|^2 + \mathbb{E}\left\|(1-\theta)\sum_{i=1}^t \theta^{t-i}\boldsymbol{\delta}^i\right\| + \theta^t\|\boldsymbol{\epsilon}^0\|,
\end{aligned}$$

where we used $\boldsymbol{y}^i - \boldsymbol{x}^i = \frac{\theta}{1-\theta}(\boldsymbol{x}^i - \boldsymbol{x}^{i-1})$. With the scheme of the algorithm, we have

$$\mathbb{E}\|\boldsymbol{x}^{i-1} - \boldsymbol{x}^i\|^2 = \mathbb{E}\|\gamma \text{Sign}(\boldsymbol{m}^{i-1})\|^2 \le \gamma^2 d.$$

On the other hand, by independence between the random variables $(\boldsymbol{\delta}^i)_{i\ge 0}$, there is

$$\mathbb{E}\left\|(1-\theta)\sum_{i=1}^t \theta^{t-i}\boldsymbol{\delta}^i\right\| \le \sqrt{\mathbb{E}\left\|(1-\theta)\sum_{i=1}^t \theta^{t-i}\boldsymbol{\delta}^i\right\|^2} = (1-\theta)\sqrt{\sum_{i=1}^t \theta^{2t-2i}\mathbb{E}\|\boldsymbol{\delta}^i\|^2} \le \frac{\sqrt{1-\theta}}{\sqrt{1+\theta}}\sigma \le \sqrt{1-\theta}\sigma.$$

Hence, we obtain the following bound for $\mathbb{E}\|\boldsymbol{\epsilon}^t\|$ by combining the computations above:

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \le \frac{d\rho}{2}\frac{\theta}{(1-\theta)^2}\gamma^2 + \sqrt{1-\theta}\sigma + \theta^t\|\boldsymbol{\epsilon}^0\|.$$

Using Lemma 1 with $\boldsymbol{x}^\dagger \to \boldsymbol{x}^t$ and $\boldsymbol{m} \to \boldsymbol{m}^t$ in the algorithm, we obtain

$$f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^t) \le -\gamma\|\nabla f(\boldsymbol{x}^t)\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}^t\| + \frac{L\gamma^2 d}{2}.$$

By summing the above inequalities with t ranging from 1 to T and then taking the the expectation, we have

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \le \frac{f(\boldsymbol{x}^1) - \min f}{\gamma T} + d^{3/2}\rho\frac{\theta}{(1-\theta)^2}\gamma^2 + 2\sqrt{d}\sqrt{1-\theta}\sigma + \frac{L\gamma d}{2} + 2\sqrt{d}\sum_{t=1}^T \theta^t\|\boldsymbol{\epsilon}^0\|/T.$$

We then conclude the proof of this theorem by noting that $1-\theta = \frac{1}{T^{4/7}}$, $\gamma = \frac{1}{\max\{\sqrt{\rho}, L\}T^{5/7}}$ and $\boldsymbol{x}^0 = \boldsymbol{x}^1$.

D. Proof of Theorem 3

I.) $\alpha = 0$: In this case, $\boldsymbol{y}^t = \boldsymbol{x}^t$. We use the notations $\boldsymbol{g}^t := \frac{\sum_{i=1}^n \nabla f_i(\boldsymbol{x}^t;\xi^t(i))}{nR}$, and $\boldsymbol{\epsilon}^t := \frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} - \nabla f(\boldsymbol{x}^t)/R$, and $\boldsymbol{\delta}^t := \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t)/R$. With the scheme of the algorithm,

$$\begin{aligned}\frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} &= \theta\frac{\sum_i^n \boldsymbol{m}^{t-1}(i)}{nR} + (1-\theta)\boldsymbol{g}^t \\ &= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^{t-1})/R) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t)/R).\end{aligned}$$

Hence, we are led to

$$\boldsymbol{\epsilon}^t = \frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} - \nabla f(\boldsymbol{x}^t)/R = \theta\boldsymbol{\epsilon}^{t-1} + \boldsymbol{s}^t + (1-\theta)\boldsymbol{\delta}^t,$$

where $\boldsymbol{s}^t := \theta(\nabla f(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^t))/R$. With the scheme of the algorithm,

$$\|\boldsymbol{s}^t\| = \frac{\theta}{R}\|\nabla f(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^t)\| \leq \frac{L}{R}\|\boldsymbol{x}^{t-1} - \boldsymbol{x}^t\| \leq \frac{L\sqrt{d}\gamma}{R}.$$

Because $(\xi^t(i))_{1\leq i\leq n, t\geq 1}$ are independent with each other, $\mathbb{E}\langle\boldsymbol{\delta}^i, \boldsymbol{\delta}^j\rangle = 0$ when $i \neq j$; and

$$\mathbb{E}\|\boldsymbol{\delta}^i\|^2 = \frac{\mathbb{E}\|\sum_{i=1}^n \nabla f_i(\boldsymbol{x}^t;\xi^t(i)) - \sum_{i=1}^n \nabla f_i(\boldsymbol{x}^t)\|^2}{n^2R^2} = \frac{\sigma^2}{nR^2}.$$

Similar to the proof of Theorem 3, it follows

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \frac{L\sqrt{d}\gamma}{R(1-\theta)} + \frac{\sqrt{1-\theta}\sigma}{R\sqrt{n}} + \theta^t\|\boldsymbol{\epsilon}^0\|.$$

Using Lemma 2 with $\boldsymbol{x}^\dagger \to \boldsymbol{x}^t$ and $\boldsymbol{m}(i) \to \boldsymbol{m}^t(i)$, we get

$$\mathbb{E}f(\boldsymbol{x}^{t+1}) - \mathbb{E}f(\boldsymbol{x}^t) \leq -\gamma\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 + 2\sqrt{d}\gamma R\mathbb{E}\|\boldsymbol{\epsilon}^t\| + \frac{2d\gamma R}{\sqrt{n}} + \frac{L\gamma^2 d}{2}. \tag{11}$$

The sum of the recursion from $t = 1$ to T yields

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{f(\boldsymbol{x}^1) - \min f}{\gamma T} + \frac{2Ld\gamma}{1-\theta} + \frac{2\sqrt{d}\sqrt{1-\theta}\sigma}{\sqrt{n}} + \frac{L\gamma d}{2} + \frac{2dR}{\sqrt{n}} + 2\sqrt{d}R\sum_{t=1}^T \theta^t\|\boldsymbol{\epsilon}^0\|/T.$$

Note $\boldsymbol{x}^0 = \boldsymbol{x}^1$, and we then proved the result. Let $\gamma = \frac{1}{LT^{3/4}}$, and $\theta = 1 - \frac{1}{\sqrt{T}}$, we have

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{f(\boldsymbol{x}^1) - \min f}{LT^{1/4}} + \frac{2d\gamma}{T^{1/4}} + \frac{2\sqrt{d}\sigma}{\sqrt{n}T^{1/4}} + \frac{d}{2T^{3/4}} + \frac{2\sqrt{d}\|\nabla f(\boldsymbol{x}^0)\|}{T^{1/2}} + \frac{2dR}{\sqrt{n}}.$$

II.) $\alpha = \frac{\theta}{1-\theta}$. In this case, by denoting $\boldsymbol{g}^t := \frac{\sum_{i=1}^n \nabla f_i(\boldsymbol{y}^t;\xi^t(i))}{nR}$, and $\boldsymbol{\epsilon}^t := \frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} - \nabla f(\boldsymbol{x}^t)/R$, and $\boldsymbol{\delta}^t := \boldsymbol{g}^t - \nabla f(\boldsymbol{y}^t)/R$, and $\mathbf{H}(\mathbf{x}, \boldsymbol{y}) := \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) - [\nabla^2 f(\boldsymbol{x})](\boldsymbol{x} - \boldsymbol{y})$, it follows

$$\begin{aligned}\frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} &= \theta\frac{\sum_i^n \boldsymbol{m}^{t-1}(i)}{nR} + (1-\theta)\boldsymbol{g}^t = \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^{t-1})/R) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{y}^t)/R) \\ &= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^t)/R + [\nabla^2 f(\boldsymbol{x}^t)](\boldsymbol{x}^{t-1} - \boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{x}^{t-1}, \boldsymbol{x}^t)/R) \\ &\qquad + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t)/R + [\nabla^2 f(\boldsymbol{x}^t)](\boldsymbol{y}^t - \boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{y}^t, \boldsymbol{x}^t)/R) \\ &= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{x}^{t-1}, \boldsymbol{x}^t)/R) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{y}^t, \boldsymbol{x}^t)/R) \\ &\Rightarrow \boldsymbol{\epsilon}^t = \frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} - \nabla f(\boldsymbol{x}^t) = \theta\boldsymbol{\epsilon}^{t-1} + \theta\mathbf{H}(\boldsymbol{x}^{t-1}, \boldsymbol{x}^t)/R + (1-\theta)\mathbf{H}(\boldsymbol{y}^t, \boldsymbol{x}^t)/R + (1-\theta)\boldsymbol{\delta}^t.\end{aligned}$$

With Mathematical Induction, we have

$$\boldsymbol{\epsilon}^t = \theta\sum_{i=1}^{t}\theta^{t-i}\mathbf{H}(\boldsymbol{x}^{i-1},\boldsymbol{x}^i)/R + (1-\theta)\sum_{i=1}^{t}\theta^{t-i}\mathbf{H}(\boldsymbol{y}^i,\boldsymbol{x}^i)/R + (1-\theta)\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i + \theta^t\|\boldsymbol{\epsilon}^0\|.$$

Similar to the proof of Theorem 2, we obtain the following bound

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \le \frac{d\rho}{2}\frac{\theta}{(1-\theta)^2 R}\gamma^2 + \frac{\sqrt{1-\theta}\sigma}{R\sqrt{n}} + \theta^t\|\boldsymbol{\epsilon}^0\|.$$

The sum of the recursion from $t = 1$ to T yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \le \frac{f(\boldsymbol{x}^0) - \min f}{\gamma T} + d^{3/2}\rho\frac{\theta}{(1-\theta)^2}\gamma^2 + \frac{2\sqrt{d}\sqrt{1-\theta}\sigma}{\sqrt{n}} + \frac{L\gamma d}{2} + 2\sqrt{d}R\sum_{t=1}^{T}\theta^t\|\boldsymbol{\epsilon}^0\|/T + \frac{2dR}{\sqrt{n}}.$$

When $1-\theta = \frac{1}{T^{4/7}}$ and $\gamma = \frac{1}{\max\{\sqrt{\rho},L\}T^{5/7}}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \le \frac{\max\{\sqrt{\rho},L\}[f(\boldsymbol{x}^0) - \min f]}{T^{2/7}} + \frac{d^{3/2}}{T^{2/7}} + \frac{2\sqrt{d}\sigma}{\sqrt{n}T^{2/7}} + \frac{\max\{\sqrt{\rho},L\}Ld}{2T^{5/7}} + \frac{2\sqrt{d}R\|\boldsymbol{\epsilon}^0\|}{T^{3/7}} + \frac{2dR}{\sqrt{n}}.$$

E. Proof of Theorem 4

I.) $\alpha = 0$: In this case, $\boldsymbol{x}^t = \boldsymbol{y}^t$. Denote $\hat{\boldsymbol{g}}^t := \frac{\sum_{i=1}^n \nabla f_i(\boldsymbol{x}^t(i);\xi^t(i))}{nR}$, and follow the notation in [I, the proof of Theorem 3]. In the federated SIGNSGD, it follows

$$\begin{aligned}
\frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} &= \theta\frac{\sum_i^n \boldsymbol{m}^{t-1}(i)}{nR} + (1-\theta)\hat{\boldsymbol{g}}^t \\
&= \theta\frac{\sum_i^n \boldsymbol{m}^{t-1}(i)}{nR} + (1-\theta)\boldsymbol{g}^t + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t] \\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^{t-1})/R) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t)/R) + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t].
\end{aligned}$$

Thus, we derive

$$\boldsymbol{\epsilon}^t = \frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} - \nabla f(\boldsymbol{x}^t)/R = \theta\boldsymbol{\epsilon}^{t-1} + \boldsymbol{s}^t + (1-\theta)\boldsymbol{\delta}^t + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t],$$

where $\boldsymbol{s}^t := \theta(\nabla f(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^t))/R$. We have

$$\boldsymbol{\epsilon}^t = \theta^t\boldsymbol{\epsilon}^0 + \sum_{i=1}^{t}\theta^{t-i}\boldsymbol{s}^i + (1-\theta)\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i + (1-\theta)\sum_{i=1}^{t}\theta^{t-i}\left[\hat{\boldsymbol{g}}^i - \boldsymbol{g}^i\right].$$

With the definition of $\hat{\boldsymbol{g}}^i$, note that $\boldsymbol{x}^t = \boldsymbol{y}^t$, with Lemma 3,

$$\mathbb{E}\|\hat{\boldsymbol{g}}^i - \boldsymbol{g}^i\| \le \frac{L\sum_{i=1}^n \mathbb{E}\|\boldsymbol{x}^t(i) - \boldsymbol{x}^t\|}{nR} \le \frac{\eta L\sqrt{8K\sigma^2 + 16K^2\sigma^2 + 16K^2R^2}}{R}.$$

Taking the expectations and norms of both sides,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\epsilon}^t\| &\le \frac{L\sqrt{d}\gamma}{R(1-\theta)} + (1-\theta)\mathbb{E}\left\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\right\| + (1-\theta)\mathbb{E}\left\|\sum_{i=1}^{t}\theta^{t-i}\left[\hat{\boldsymbol{g}}^i - \boldsymbol{g}^i\right]\right\| + \theta^t\|\boldsymbol{\epsilon}^0\| \\
&\le \frac{L\sqrt{d}\gamma}{R(1-\theta)} + \frac{\sqrt{1-\theta}\sigma}{R\sqrt{n}} + \frac{\eta L\sqrt{8K\sigma^2 + 16K^2\sigma^2 + 16K^2R^2}}{R} + \theta^t\|\boldsymbol{\epsilon}^0\| \\
&\le \frac{L\sqrt{d}\gamma}{R(1-\theta)} + \frac{\sqrt{1-\theta}\sigma}{R\sqrt{n}} + \frac{\eta L\sqrt{24K^2\sigma^2 + 16K^2R^2}}{R} + \theta^t\|\boldsymbol{\epsilon}^0\|.
\end{aligned}$$

Note that (11) still holds,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{f(\boldsymbol{x}^0)-\min f}{\gamma T} + \frac{2Ld\gamma}{(1-\theta)} + \frac{2\sqrt{d}\sqrt{1-\theta}\sigma}{\sqrt{n}} + \frac{L\gamma d}{2} + \frac{2dR}{\sqrt{n}}$$
$$+ \sqrt{24\sigma^2+16R^2}\eta LK + \frac{2\sqrt{d}\|\nabla f(\boldsymbol{x}^0)\|}{T^{1/2}}.$$

II.) $\alpha = \frac{\theta}{1-\theta}$: Denote $\hat{\boldsymbol{g}}^t := \frac{\sum_{i=1}^n \nabla f_i(\boldsymbol{y}^t(i);\xi^t(i))}{nR}$, and follow the notation in [II, the proof of Theorem 3],

$$\begin{aligned}
\frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} &= \theta\frac{\sum_i^n \boldsymbol{m}^{t-1}(i)}{nR} + (1-\theta)\hat{\boldsymbol{g}}^t = \theta\frac{\sum_i^n \boldsymbol{m}^{t-1}(i)}{nR} + (1-\theta)\boldsymbol{g}^t + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t] \\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^{t-1})/R) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{y}^t)/R) + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t] \\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^t)/R + [\nabla^2 f(\boldsymbol{x}^t)](\boldsymbol{x}^{t-1}-\boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{x}^{t-1},\boldsymbol{x}^t)/R) + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t] \\
&\quad + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t)/R + [\nabla^2 f(\boldsymbol{x}^t)](\boldsymbol{y}^t - \boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{y}^t,\boldsymbol{x}^t)/R) + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t] \\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{x}^{t-1},\boldsymbol{x}^t)/R) + (1-\theta)(\boldsymbol{\delta}^t + \nabla f(\boldsymbol{x}^t)/R + \mathbf{H}(\boldsymbol{y}^t,\boldsymbol{x}^t)/R) + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t]
\end{aligned}$$

Thus, we have

$$\boldsymbol{\epsilon}^t = \frac{\sum_i^n \boldsymbol{m}^t(i)}{nR} - \nabla f(\boldsymbol{x}^t) = \theta\boldsymbol{\epsilon}^{t-1} + \theta\mathbf{H}(\boldsymbol{x}^{t-1},\boldsymbol{x}^t)/R + (1-\theta)\mathbf{H}(\boldsymbol{y}^t,\boldsymbol{x}^t)/R + (1-\theta)\boldsymbol{\delta}^t + (1-\theta)[\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t].$$

Noticing $\boldsymbol{\epsilon}^0 = \mathbf{0}$, and $\hat{\boldsymbol{g}}^0 = \boldsymbol{g}^0$, we have

$$\boldsymbol{\epsilon}^t = \theta\sum_{i=1}^{t}\theta^{t-i}\mathbf{H}(\boldsymbol{x}^{i-1},\boldsymbol{x}^i)/R + (1-\theta)\sum_{i=1}^{t}\theta^{t-i}\mathbf{H}(\boldsymbol{y}^i,\boldsymbol{x}^i)/R + (1-\theta)\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i + (1-\theta)\sum_{i=1}^{t}\theta^{t-i}\Big[\hat{\boldsymbol{g}}^i - \boldsymbol{g}^i\Big] + \theta^t\boldsymbol{\epsilon}^0.$$

Taking the expectations and norms of both sides gives us

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\epsilon}^t\| &\leq \frac{d\rho}{2}\frac{\theta}{(1-\theta)^2 R}\gamma^2 + (1-\theta)\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\| + (1-\theta)\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\Big[\hat{\boldsymbol{g}}^i - \boldsymbol{g}^i\Big]\Big\| + \theta^t\|\boldsymbol{\epsilon}^0\| \\
&\leq \frac{d\rho}{2}\frac{\theta}{(1-\theta)^2 R}\gamma^2 + \frac{\sqrt{1-\theta}\sigma}{R\sqrt{n}} + \frac{\eta L\sqrt{8K\sigma^2 + 16K^2\sigma^2 + 16K^2R^2}}{R} + \theta^t\|\boldsymbol{\epsilon}^0\|,
\end{aligned}$$

and thus

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{f(\boldsymbol{x}^0)-\min f}{\gamma T} + d^{3/2}\rho\frac{\theta}{(1-\theta)^2}\gamma^2 + \frac{2\sqrt{d}\sqrt{1-\theta}\sigma}{\sqrt{n}} + \frac{L\gamma d}{2}$$
$$+ \frac{2dR}{\sqrt{n}} + \eta L\sqrt{24K^2\sigma^2 + 16K^2R^2} + 2\sqrt{d}R\sum_{t=1}^{T}\theta^t\|\boldsymbol{\epsilon}^0\|/T.$$

When $1-\theta = \frac{1}{T^{4/7}}$, $\gamma = \frac{1}{\max\{\sqrt{\rho},L\}T^{5/7}}$, and $\eta = \frac{1}{4LK^2}$, and $K = T^{1/4}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^t)\|_1 \leq \frac{\max\{\sqrt{\rho},L\}[f(\boldsymbol{x}^0)-\min f]}{T^{2/7}} + \frac{d^{3/2}}{T^{2/7}} + \frac{2\sqrt{d}\sigma}{\sqrt{n}T^{2/7}}$$
$$+ \frac{\max\{\sqrt{\rho},L\}Ld}{2T^{5/7}} + \frac{2\sqrt{d}R\|\boldsymbol{\epsilon}^0\|}{T^{3/7}} + \frac{2dR}{\sqrt{n}} + \frac{\sqrt{2\sigma^2+R^2}}{T^{1/4}}.$$

F. Proofs of Technical Lemmas

F.1. Proofs of Lemma 1

The proof of this lemma is motivated by (Cutkosky & Mehta, 2020). The smoothness of the gradient gives us

$$\begin{aligned}
f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) &\le \langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{x}^\ddagger - \boldsymbol{x}^\dagger \rangle + \frac{L}{2}\|\boldsymbol{x}^\ddagger - \boldsymbol{x}^\dagger\|^2 \\
&\le -\gamma \langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{A} \cdot \text{Sign}(\boldsymbol{m}) \rangle + \frac{L\gamma^2 \sum_{i=1}^d a_i^2}{2} \\
&= -\gamma \langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{A} \cdot \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] \rangle + \gamma \langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{A} \cdot \text{Sign}(\boldsymbol{m}) - \boldsymbol{A} \cdot \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] \rangle + \frac{L\gamma^2 \sum_{i=1}^d a_i^2}{2} \\
&= -\gamma \|\boldsymbol{A} \nabla f(\boldsymbol{x}^\dagger)\|_1 + \gamma \langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{A} \cdot \text{Sign}(\boldsymbol{m}) - \boldsymbol{A} \cdot \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] \rangle + \frac{L\gamma^2 \sum_{i=1}^d a_i^2}{2}.
\end{aligned}$$

Notice that

$$\langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{A} \cdot \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] - \boldsymbol{A} \cdot \text{Sign}(\boldsymbol{m}) \rangle = \sum_{i=1}^d a_i [\nabla f(\boldsymbol{x}^\dagger)]_i \cdot [\text{Sign}([\nabla f(\boldsymbol{x}^\dagger)]_i) - \text{Sign}([\boldsymbol{m}]_i)].$$

1) If $\text{Sign}([\nabla f(\boldsymbol{x}^\dagger)]_i) = \text{Sign}([\boldsymbol{m}]_i)$, $a_i [\nabla f(\boldsymbol{x}^\dagger)]_i \cdot [\text{Sign}([\nabla f(\boldsymbol{x}^\dagger)]_i) - \text{Sign}([\boldsymbol{m}]_i)] = 0$.

2) Otherwise, $[\nabla f(\boldsymbol{x}^\dagger)]_i \cdot [\boldsymbol{m}]_i \le 0$ which means

$$|\boldsymbol{\epsilon}_i| = |[\nabla f(\boldsymbol{x}^\dagger)]_i - [\boldsymbol{m}]_i| \ge |[\nabla f(\boldsymbol{x}^\dagger)]_i|.$$

In this case, it then holds

$$a_i [\nabla f(\boldsymbol{x}^\dagger)]_i \cdot [\text{Sign}([\nabla f(\boldsymbol{x}^\dagger)]_i) - \text{Sign}([\boldsymbol{m}]_i)] \le 2a_i |[\nabla f(\boldsymbol{x}^\dagger)]_i| \le 2a_i |\boldsymbol{\epsilon}_i|. \tag{12}$$

In summary, for any $i \in [d]$, the inequality (12) holds. Thus, we have

$$\langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{A} \cdot \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] - \boldsymbol{A} \cdot \text{Sign}(\boldsymbol{m}) \rangle \le 2 \sum_{i=1}^d a_i |\boldsymbol{\epsilon}_i| \le 2\|\boldsymbol{A}\|_F \|\boldsymbol{\epsilon}\|$$

and

$$f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) \le -\gamma \|\boldsymbol{A} \nabla f(\boldsymbol{x}^\dagger)\|_1 + 2\gamma \|\boldsymbol{A}\|_F \|\boldsymbol{\epsilon}\| + \frac{L\gamma^2 \|\boldsymbol{A}\|_F^2}{2}.$$

When $\boldsymbol{A}$ is the identity matrix, $\|\boldsymbol{A}\|_F = \sqrt{d}$, which directly gives

$$f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) \le -\gamma \|\nabla f(\boldsymbol{x}^\dagger)\|_1 + 2\sqrt{d}\gamma \|\boldsymbol{\epsilon}\| + \frac{L\gamma^2 d}{2}.$$

F.2. Proof of Lemma 2

Using a shorthand notation $\boldsymbol{m} := \sum_{i=1}^n \mathcal{S}_R(\boldsymbol{m}(i))/n$, it holds

$$\boldsymbol{x}^\ddagger = \boldsymbol{x}^\dagger - \text{Sign}(\boldsymbol{m}).$$

Using the Lipschitz property of the gradient, we have

$$\begin{aligned}
f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) &\le \langle \nabla f(\boldsymbol{x}^\dagger), \boldsymbol{x}^\ddagger - \boldsymbol{x}^\dagger \rangle + \frac{L}{2}\|\boldsymbol{x}^\ddagger - \boldsymbol{x}^\dagger\|^2 \\
&\le -\gamma \langle \nabla f(\boldsymbol{x}^\dagger), \text{Sign}(\boldsymbol{m}) \rangle + \frac{Ld\gamma^2}{2} \\
&= -\gamma \langle \nabla f(\boldsymbol{x}^\dagger), \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] \rangle + \gamma \langle \nabla f(\boldsymbol{x}^\dagger), \text{Sign}(\boldsymbol{m}) - \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] \rangle + \frac{Ld\gamma^2}{2} \\
&= -\gamma \|\nabla f(\boldsymbol{x}^\dagger)\|_1 + \gamma \langle \nabla f(\boldsymbol{x}^\dagger), \text{Sign}(\boldsymbol{m}) - \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] \rangle + \frac{Ld\gamma^2}{2}.
\end{aligned}$$

If $[\nabla f(\boldsymbol{x}^\dagger)/R]_i \cdot [\boldsymbol{m}]_i > 0$, i.e., $\text{Sign}([\nabla f(\boldsymbol{x}^\dagger)]_i) = \text{Sign}([\boldsymbol{m}]_i)$, then we get

$$[\nabla f(\boldsymbol{x}^\dagger)]_i \cdot [\text{Sign}([\nabla f(\boldsymbol{x}^\dagger)]_i) - \text{Sign}([\boldsymbol{m}]_i)] = 0.$$

If $[\nabla f(\boldsymbol{x}^\dagger)/R]_i \cdot [\boldsymbol{m}]_i \leq 0$,

$$|[\nabla f(\boldsymbol{x}^\dagger)]_i/R - [\boldsymbol{m}]_i| \geq |[\nabla f(\boldsymbol{x}^\dagger)]_i/R|$$

due to $1/R > 0$. In this case, for each coordinate i, it then holds

$$[\nabla f(\boldsymbol{x}^\dagger)]_i \cdot [\text{Sign}([\nabla f(\boldsymbol{x}^\dagger)]_i) - \text{Sign}([\boldsymbol{m}]_i)] \leq 2|[\nabla f(\boldsymbol{x}^\dagger)]_i| = 2R|[\nabla f(\boldsymbol{x}^\dagger)]_i/R| \leq 2R|\nabla f(\boldsymbol{x}^\dagger)]_i/R - [\boldsymbol{m}]_i|. \tag{13}$$

Thus, we have the following estimate:

$$\langle \nabla f(\boldsymbol{x}^\dagger), \text{Sign}[\nabla f(\boldsymbol{x}^\dagger)] - \text{Sign}(\boldsymbol{m})\rangle \leq 2R\sum_{i=1}^{d} |[\nabla f(\boldsymbol{x}^\dagger)]_i/R - [\boldsymbol{m}]_i| \leq 2\sqrt{d}R\|\nabla f(\boldsymbol{x}^\dagger)/R - \boldsymbol{m}\|.$$

Then, we get the following inequality

$$f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) \leq -\gamma\|\nabla f(\boldsymbol{x}^\dagger)\|_1 + 2\sqrt{d}\gamma R\|\nabla f(\boldsymbol{x}^\dagger)/R - \boldsymbol{m}\| + \frac{L\gamma^2 d}{2}. \tag{14}$$

Recall the definition of $\boldsymbol{m}$, inequality (14) is equivalent to the following expression:

$$f(\boldsymbol{x}^\ddagger) - f(\boldsymbol{x}^\dagger) \leq -\gamma\|\nabla f(\boldsymbol{x}^\dagger)\|_1 + 2\sqrt{d}\gamma R\left\|\sum_{i=1}^{n}\mathcal{S}_R(\boldsymbol{m}(i))/n - \nabla f(\boldsymbol{x}^\dagger)/R\right\| + \frac{L\gamma^2 d}{2}.$$

For any $\boldsymbol{v} \in \mathbb{R}^d$ such that $\|\boldsymbol{v}\| \leq R$, with direct computations, there is

$$\mathbb{E}\|\mathcal{S}_R(\boldsymbol{v}) - \boldsymbol{v}/R\|^2 \leq d - \frac{\|\boldsymbol{v}\|^2}{R^2} \leq d.$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\left(\left\|\sum_{i=1}^{n}\mathcal{S}_R(\boldsymbol{m}(i))/n - \nabla f(\boldsymbol{x}^\dagger)/R\right\| \,\middle|\, \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right) \\
&\leq \mathbb{E}\left(\left\|\sum_{i=1}^{n}\mathcal{S}_R(\boldsymbol{m}(i))/n - \sum_{i=1}^{n}\boldsymbol{m}(i)/(nR)\right\| \,\middle|\, \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right) + \mathbb{E}\left(\left\|\sum_{i=1}^{n}\boldsymbol{m}(i)/(nR) - \nabla f(\boldsymbol{x}^\dagger)/R\right\| \,\middle|\, \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right) \\
&\leq \sqrt{\mathbb{E}\left(\left\|\sum_{i=1}^{n}\mathcal{S}_R(\boldsymbol{m}(i))/n - \sum_{i=1}^{n}\boldsymbol{m}(i)/(nR)\right\|^2 \,\middle|\, \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right)} + \mathbb{E}\left(\|\boldsymbol{\epsilon}\| \,\middle|\, \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right) \\
&= \sqrt{\sum_{i=1}^{n}\frac{1}{n^2}\mathbb{E}\left(\|\mathcal{S}_R(\boldsymbol{m}(i)) - \boldsymbol{m}(i)/R\|^2 \middle| \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right)} + \mathbb{E}\left(\|\boldsymbol{\epsilon}\| \,\middle|\, \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right) \\
&\leq \sqrt{\frac{d}{n}} + \mathbb{E}\left(\|\boldsymbol{\epsilon}\| \,\middle|\, \{\boldsymbol{m}(i)\}_{1\leq i\leq n}\right).
\end{aligned}$$

Taking full expectations, we then obtain the desired results.

F.3. Proof of Lemma 3

Note that for any $k \in \{0, 1, \ldots, K-1\}$, in node i,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{z}^{t,k+1}(i) - \boldsymbol{y}^t\|^2 &= \mathbb{E}\|\boldsymbol{z}^{t,k}(i) - \eta\nabla f(\boldsymbol{z}^{t,k}(i); \xi^{t,k}(i)) - \boldsymbol{y}^t\|^2 \\
&\leq \mathbb{E}\Big\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{y}^t - \eta\Big(\nabla f(\boldsymbol{z}^{t,k}(i); \xi^{t,k}(i)) - \nabla f_i(\boldsymbol{z}^{t,k}(i)) \\
&\qquad + \nabla f_i(\boldsymbol{z}^{t,k}(i)) - \nabla f_i(\boldsymbol{y}^t) + \nabla f_i(\boldsymbol{y}^t)\Big)\Big\|^2,
\end{aligned}$$

By using the Cauchy's inequality

$$\mathbb{E}\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\frac{1}{\psi})\mathbb{E}\|\mathbf{a}\|^2 + (1+\psi)\mathbb{E}\|\mathbf{b}\|^2$$

with $\boldsymbol{a} = \boldsymbol{z}^{t,k}(i) - \boldsymbol{x}^t - \eta(\nabla f(\boldsymbol{z}^{t,k}(i);\xi^{t,k}(i)) - \nabla f_i(\boldsymbol{z}^{t,k}(i))$, $\mathbf{b} = \eta(\nabla f_i(\boldsymbol{z}^{t,k}(i)) - \nabla f_i(\boldsymbol{x}^t) + \nabla f_i(\boldsymbol{x}^t)$ and $\psi = 2K-1$. Denote that $\Re := (1+\frac{1}{2K-1})\mathbb{E}\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{x}^t - \eta(\nabla f(\boldsymbol{z}^{t,k}(i);\xi^{t,k}(i)) - \nabla f_i(\boldsymbol{z}^{t,k}(i))\|$ and $\Im := 2K\eta^2\mathbb{E}\|(\nabla f_i(\boldsymbol{z}^{t,k}(i)) - \nabla f_i(\boldsymbol{y}^t) + \nabla f_i(\boldsymbol{y}^t)\|^2$. The unbiased expectation property of $\nabla f(\boldsymbol{z}^{t,k}(i);\xi^{t,k}(i))$ gives us

$$\begin{aligned}\Re &= (1+\frac{1}{2K-1})\Big(\mathbb{E}\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{y}^t\|^2 + \eta^2\mathbb{E}\|\nabla f(\boldsymbol{z}^{t,k}(i);\xi^{t,k}(i)) - \nabla f_i(\boldsymbol{z}^{t,k}(i))\|^2\Big)\\ &\leq (1+\frac{1}{2K-1})\Big(\mathbb{E}\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{y}^t\|^2 + \eta^2\sigma^2\Big)\end{aligned}$$

On the other hand, we have the following bound

$$\begin{aligned}\Im &\leq 4K\eta^2\mathbb{E}\|\nabla f_i(\boldsymbol{z}^{t,k}(i)) - \nabla f_i(\boldsymbol{y}^t)\|^2 + 4K\eta^2\mathbb{E}\|\nabla f_i(\boldsymbol{y}^t)\|^2\\ &\leq 4L^2K\eta^2\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{y}^t\|^2 + 4K\eta^2R^2.\end{aligned}$$

When $0 < \eta \leq \frac{1}{4LK}$,

$$1 + \frac{1}{2K-1} + 4L^2K\eta^2 \leq 1 + \frac{1}{K-1},$$

and we can obtain

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{z}^{t,k+1}(i) - \boldsymbol{y}^t\|^2\\ &\leq (1+\frac{1}{2K-1} + 4L^2K\eta^2)\mathbb{E}\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{y}^t\|^2 + 2\eta^2\sigma^2 + 4K\eta^2\sigma^2 + 4K\eta^2R^2\\ &\leq (1+\frac{1}{K-1})\mathbb{E}\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{y}^t\|^2 + 2\eta^2\sigma^2 + 4K\eta^2\sigma^2 + 4K\eta^2R^2.\end{aligned}$$

The recursion from $j = 0$ to k yields

$$\begin{aligned}\mathbb{E}\|\boldsymbol{z}^{t,k}(i) - \boldsymbol{y}^t\|^2 &\leq \sum_{j=0}^{K-1}(1+\frac{1}{K-1})^j\Big[2\eta^2\sigma^2 + 4K\eta^2\sigma^2 + 4K\eta^2R^2\Big]\\ &\leq (K-1)\Big[(1+\frac{1}{K-1})^K - 1\Big] \times \Big[2\eta^2\sigma^2 + 4K\eta^2\sigma^2 + 4K\eta^2R^2\Big]\\ &\leq 8K\eta^2\sigma^2 + 16K^2\eta^2\sigma^2 + 16K^2\eta^2R^2,\end{aligned}$$

where we used the inequality $(1+\frac{1}{K-1})^K \leq 5$ holds for any $K \geq 1$. Noticing the fact $\boldsymbol{y}^t(i) = \boldsymbol{z}^{t,K}(i)$, we then prove

$$\mathbb{E}\|\boldsymbol{y}^t(i) - \boldsymbol{y}^t\|^2 \leq 8K\eta^2\sigma^2 + 16K^2\eta^2\sigma^2 + 16K^2\eta^2R^2$$

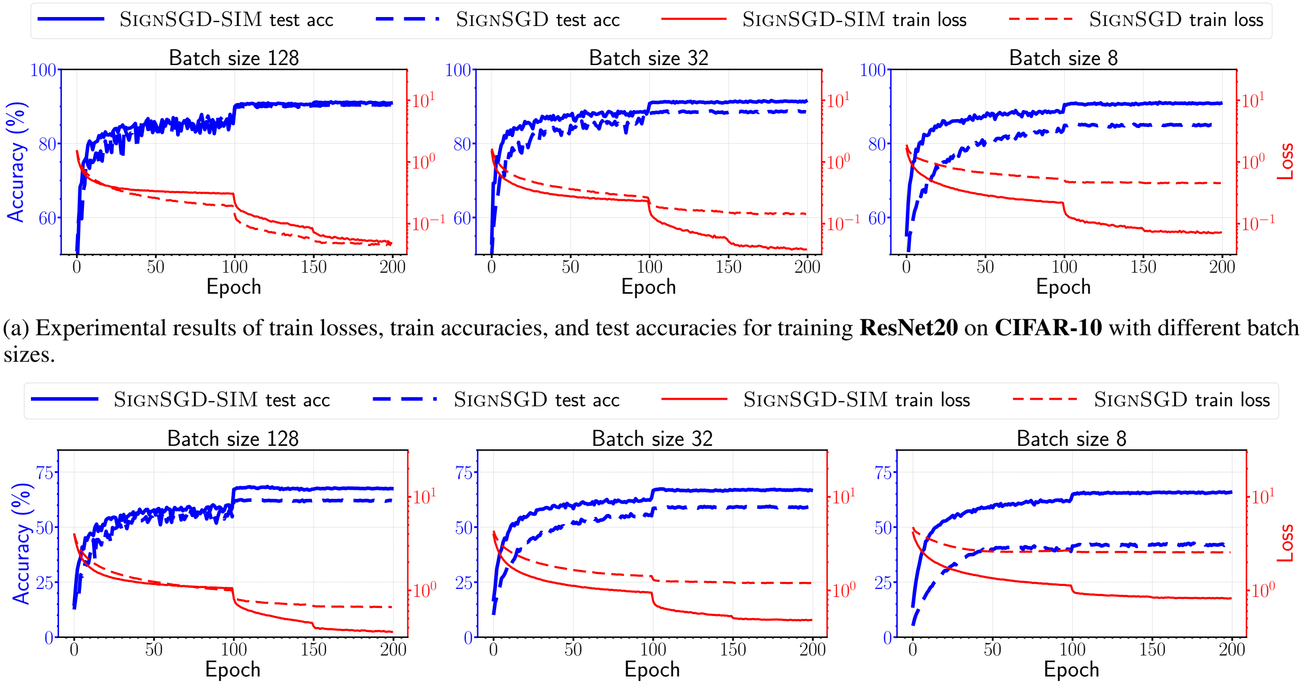
for Assumption 4. With Cauchy's inequality,

$$\mathbb{E}\|\boldsymbol{y}^t(i) - \boldsymbol{y}^t\| \leq \sqrt{\mathbb{E}\|\boldsymbol{y}^t(i) - \boldsymbol{y}^t\|^2} \leq \eta\sqrt{8K\sigma^2 + 16K^2\sigma^2 + 16K^2R^2}.$$

G. Additional experiments

G.1. Additional experiments on CIFAR-10/CIFAR-100

In this section, we provide additional experiments of training ResNet20/32 on CIFAR-10/CIFAR-100 using SIGN-SGD and SIGN-SGD-SIM. The results in Figure 5, Figure 6, and Table 3 align with the result in Section 4.1 that SIGN-SGD-SIM does not require a large batch size to guarantee convergence.



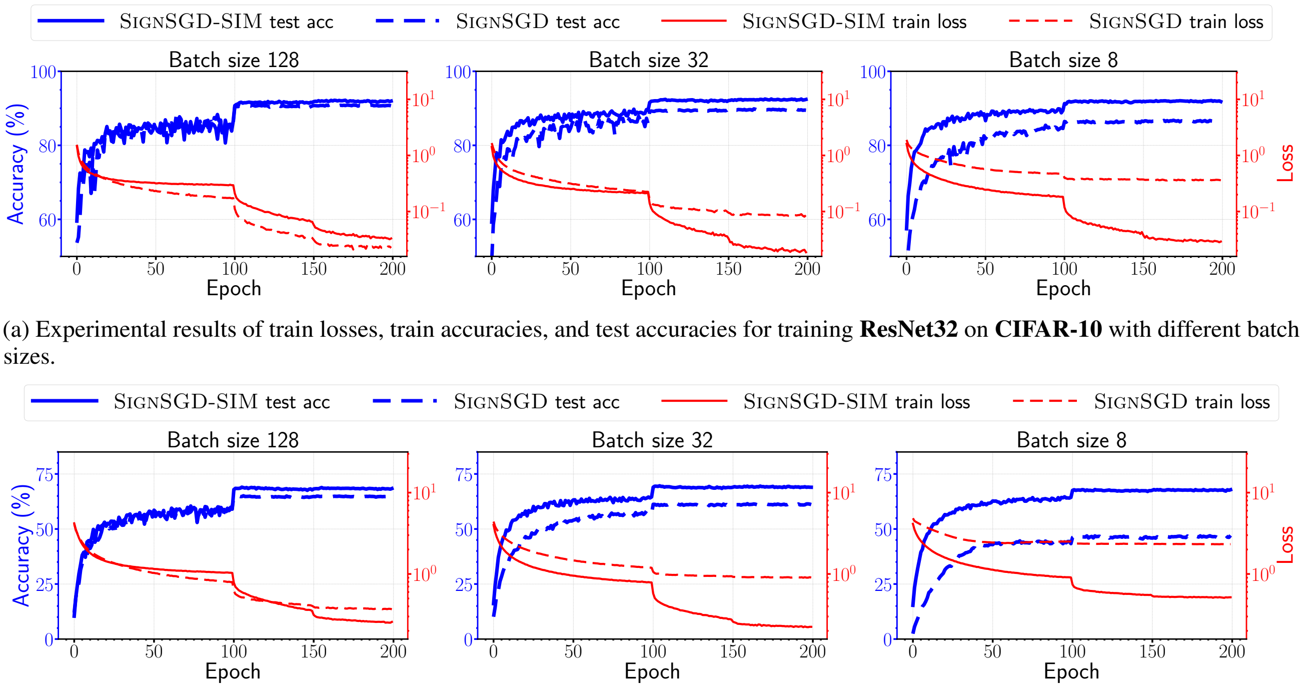
(a) Experimental results of train losses, train accuracies, and test accuracies for training **ResNet20** on **CIFAR-10** with different batch sizes.

(b) Experimental results of train losses, train accuracies, and test accuracies for training **ResNet20** on **CIFAR-100** with different batch sizes.

Figure 5. Performance comparison between SIGNSGD-SIM and SIGNSGD on training ResNet20: in sharp contrast, SIGNSGD-SIM maintains its performance with smaller batch sizes while the performance of SIGNSGD drops significantly.

Batch size	ResNet-20, CIFAR-10		ResNet-20, CIFAR-100		ResNet-32, CIFAR-10		ResNet-32, CIFAR-100	
	SIGNSGD	SIGNSGD-SIM	SIGNSGD	SIGNSGD-SIM	SIGNSGD	SIGNSGD-SIM	SIGNSGD	SIGNSGD-SIM
128	90.53%	91.26%	62.43%	68.45%	90.92%	92.19%	65.12%	69.01%
32	88.88%	91.60%	59.50%	67.36%	89.86%	92.53%	61.53%	69.63%
8	85.23%	91.15%	42.72%	65.99%	86.78%	92.12%	47.11%	67.94%

Table 3. The testing accuracies of training **ResNet20** and **ResNet32** on **CIFAR-10/CIFAR-100** with different batch sizes using SIGNSGD and SIGNSGD-SIM.



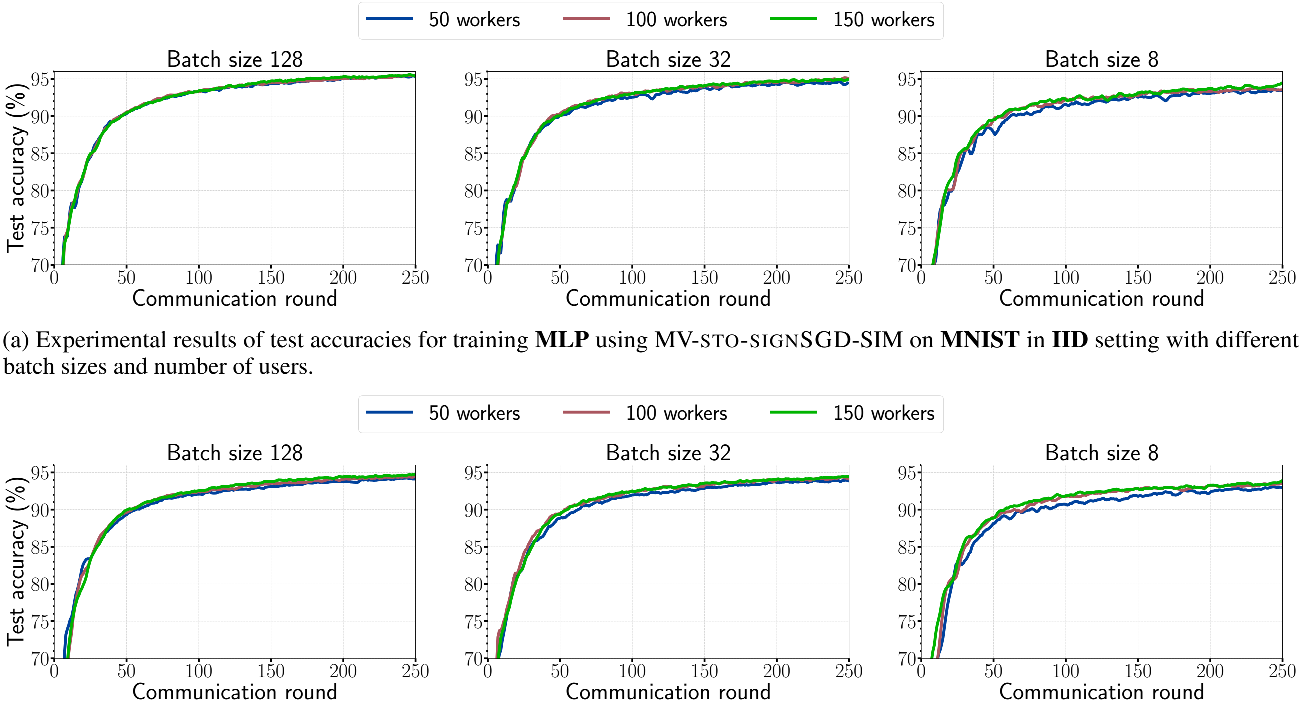
(a) Experimental results of train losses, train accuracies, and test accuracies for training **ResNet32** on **CIFAR-10** with different batch sizes.

(b) Experimental results of train losses, train accuracies, and test accuracies for training **ResNet32** on **CIFAR-100** with different batch sizes.

Figure 6. Performance comparison between SIGNSGD-SIM and SIGNSGD on training ResNet32: in sharp contrast, SIGNSGD-SIM maintains its performance with smaller batch sizes while the performance of SIGNSGD drops significantly.

G.2. Additional experiments on MNIST with different number of workers

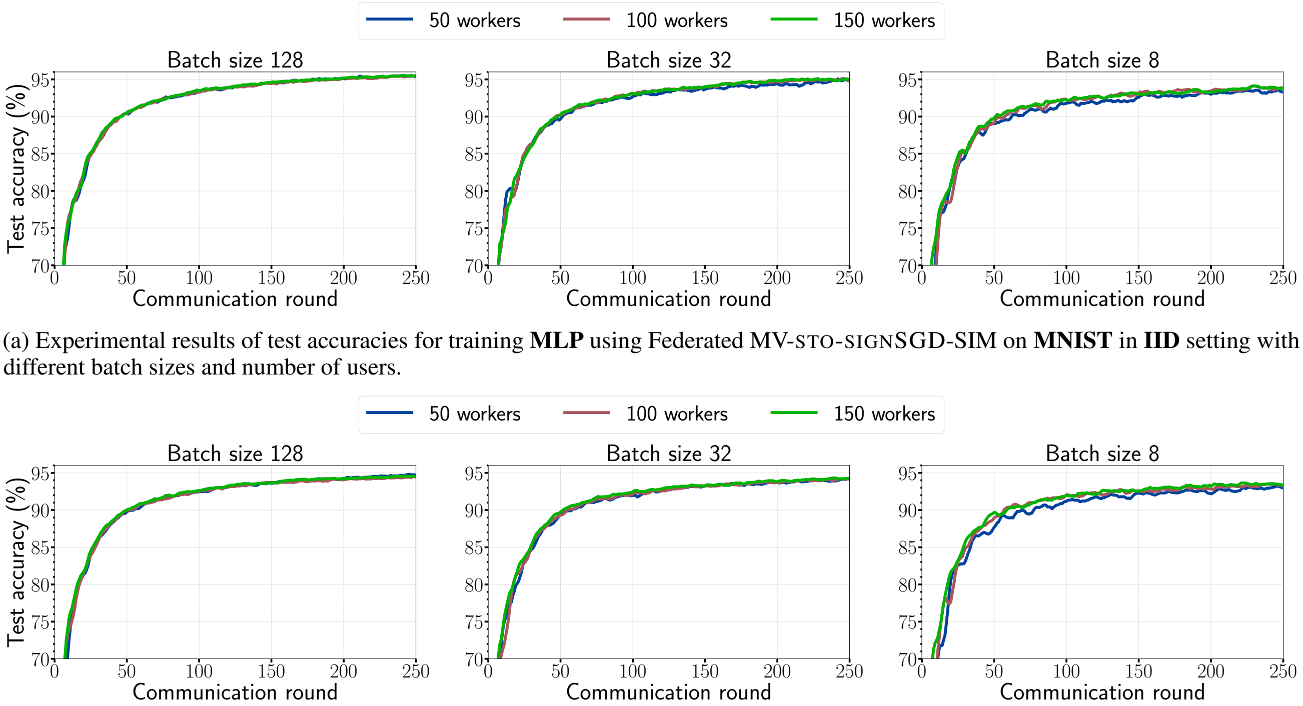
In this section, we examine how the number of workers affects the performance of MV-STO-SIGNSGD-SIM and Federated MV-STO-SIGNSGD-SIM. We follow the setting of training MLP on MNIST in Section 4.2 and consider the cases of 50, 100, and 150 workers. The results for MV-STO-SIGNSGD-SIM and Federated MV-STO-SIGNSGD-SIM are presented in Figure 7 and Figure 8 respectively. In both cases, the algorithms benefit from more workers, especially when the batch is small.



(a) Experimental results of test accuracies for training **MLP** using MV-STO-SIGNSGD-SIM on **MNIST** in **IID** setting with different batch sizes and number of users.

(b) Experimental results of test accuracies for training **MLP** using MV-STO-SIGNSGD-SIM on **MNIST** in **Non-IID** setting with different batch sizes and number of users.

Figure 7. The experimental results in both **Non-IID** and **Non-IID** settings demonstrate that incorporating more users in the training procedure improves the speed and stability of MV-STO-SIGNSGD-SIM, particularly when the batch size is small.



(a) Experimental results of test accuracies for training **MLP** using Federated MV-STO-SIGNSGD-SIM on **MNIST** in **IID** setting with different batch sizes and number of users.

(b) Experimental results of test accuracies for training **MLP** using Federated MV-STO-SIGNSGD-SIM on **MNIST** in **Non-IID** setting with different batch sizes and number of users.

Figure 8. The experimental results in both **Non-IID** and **Non-IID** settings demonstrate that incorporating more users in the training procedure improves the speed and stability of Federated MV-STO-SIGNSGD-SIM, particularly when the batch size is small.