IMPROVED ANALYSIS FOR SIGN-BASED METHODS WITH MOMENTUM UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents enhanced analysis for sign-based optimization algorithms with momentum updates. Traditional sign-based methods obtain a convergence rate of $\mathcal{O}(T^{-1/4})$ under the separable smoothness assumption, but they typically require large batch sizes or assume unimodal symmetric stochastic noise. To address these limitations, we demonstrate that signSGD with momentum can achieve the same convergence rate using constant batch sizes without additional assumptions. We also establish a convergence rate under the l_2-smoothness condition, improving upon the result of the prior momentum-based signSGD variant by a factor of $\mathcal{O}(d^{1/2})$, where d is the problem dimension. Furthermore, we explore sign-based methods with majority vote in distributed settings and show that the proposed momentum-based method yields convergence rates of $\mathcal{O}\left(d^{1/2}T^{-1/2} + dn^{-1/2}\right)$ and $\mathcal{O}\left(\max\{d^{1/4}T^{-1/4}, d^{1/10}T^{-1/5}\}\right)$, which outperform the previous results of $\mathcal{O}\left(dT^{-1/4} + dn^{-1/2}\right)$ and $\mathcal{O}\left(d^{3/8}T^{-1/8}\right)$, respectively. Numerical experiments also validate the effectiveness of the proposed methods.

1 INTRODUCTION

This paper investigates the stochastic optimization problem in the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth function. We assume that only noisy estimations of the gradient are available, denoted as $\nabla f(\mathbf{x}; \xi)$, where ξ is a random sample such that $\mathbb{E}[\nabla f(\mathbf{x}; \xi)] = \nabla f(\mathbf{x})$.

Problem (1) has been extensively studied in the literature (Duchi et al., 2011; Kingma & Ba, 2015; Loshchilov & Hutter, 2017; Fang et al., 2018; Wang et al., 2019). One of the most widely used methods for this problem is Stochastic Gradient Descent (SGD), which updates the parameters as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t; \xi_t), \tag{2}$$

where η is the learning rate and ξ_t is the random sample drawn at the t-th iteration. It is known that SGD achieves a convergence rate of $\mathcal{O}(T^{-1/4})$, where T is the number of iterations (Ghadimi & Lan, 2013). This rate is proved to be optimal under standard assumptions (Arjevani et al., 2023).

Instead of using the stochastic gradient to update, several works (Bernstein et al., 2018; 2019; Safaryan & Richtarik, 2021) propose to update using only the sign of the stochastic gradient, i.e.,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}\left(\nabla f(\mathbf{x}_t; \xi_t)\right), \tag{3}$$

which is particularly beneficial in distributed settings. In such scenarios, only the sign information needs to be transmitted between nodes, significantly reducing communication overhead.

Recently, several studies have investigated the convergence properties of signSGD and its variants. Bernstein et al. (2018) first prove that signSGD achieves a convergence rate of $\mathcal{O}(N^{-1/4})$ under the separable smoothness assumption, where N is the number of stochastic gradient calls. However, their analysis requires a large batch size of $\mathcal{O}(\sqrt{N})$ in each iteration. Later, Bernstein et al. (2019) demonstrate that signSGD can achieve the same convergence rate with constant batch sizes, but under the additional assumption that the noise is unimodal and symmetric. To avoid such extra

Table 1: Summary of convergence rates for sign-based algorithms. Here, T represents the number of stochastic gradient calls and $l_1\&l_2$ denotes mixed l_1-norm and weighted l_2-norm. We use stochastic gradient calls rather than iteration numbers to measure convergence, in order to provide a fairer comparison across different algorithms with varying batch sizes.

Method	Convergence	Assumptions	Measure	Additional Requirements
signSGD (Bernstein et al., 2018)	$\mathcal{O}\left(\frac{1}{T^{1/4}}\right)$	Assumptions 1, 2, 3	l_1	Large batch size of $\mathcal{O}(\sqrt{T})$
Signum (Bernstein et al., 2018)	$\tilde{\mathcal{O}}\left(\frac{1}{T^{1/4}}\right)$		l_1	Large batch size of $\mathcal{O}(\sqrt{T})$
signSGD (Bernstein et al., 2019)	$\mathcal{O}\left(\frac{1}{T^{1/4}}\right)$		$l_1\&l_2$	Unimodal symmetric noise
Theorem 1 **(this work)**	$\mathcal{O}\left(\frac{1}{T^{1/4}}\right)$		l_1	–
signSGD-SIM (Sun et al., 2023)	$\mathcal{O}\left(\frac{d}{T^{1/4}}\right)$	Assumptions 1, 4, 5	l_1	–
Theorem 2 **(this work)**	$\mathcal{O}\left(\frac{d^{1/2}}{T^{1/4}}\right)$			

assumptions, Sun et al. (2023) show that signSGD with momentum can achieve a convergence rate of $\mathcal{O}(dT^{-1/4})$ under the l_2-smoothness assumption. However, this dependence on d is unsatisfactory, leading to high sample complexity for high-dimensional problems.

In this paper, we re-examine signSGD with momentum and establish a convergence rate of $\mathcal{O}(T^{-1/4})$ under the separable smoothness condition. Compared with previous work (Bernstein et al., 2018; 2019), our analysis does not require large batch sizes or unimodal symmetric noise. Under the l_2-smoothness assumption, we also derive a convergence rate of $\mathcal{O}(d^{1/2}T^{-1/4})$, improving the previous result of $\mathcal{O}(dT^{-1/4})$ (Sun et al., 2023).

For distributed sign-based methods, each node typically transmits the sign of its gradient to the server, which then sends back the sign of the aggregated gradients for update. In this context, previous literature establishes convergence rates of $\mathcal{O}\left(\frac{d}{T^{1/4}}+\frac{d}{n^{1/2}}\right)$ (Sun et al., 2023) and $\mathcal{O}\left(\frac{d^{3/8}}{T^{1/8}}\right)$ (Jin et al., 2021), where n denotes the number of nodes. To improve these rates, we utilize an unbiased sign operation along with momentum updates, achieving convergence rates of $\mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}}+\frac{d}{n^{1/2}}\right)$, $\mathcal{O}\left(\frac{n^{1/2}}{T}+\frac{d}{n^{1/2}}\right)$ and $\mathcal{O}\left(\max\{\frac{d^{1/4}}{T^{1/4}},\frac{d^{1/10}}{T^{1/5}}\}\right)$, with different hyper-parameter settings and algorithm designs. In summary, this paper makes the following contributions:

- Under the separable smoothness assumption, we prove that signSGD with momentum can achieve a convergence rate of $\mathcal{O}(T^{-1/4})$ without additional assumptions. In contrast, existing analyses require either large batches or the assumption of unimodal symmetric noise.
- Under the l_2-smoothness assumption, we show that signSGD with momentum achieves a convergence rate of $\mathcal{O}(d^{1/2}T^{-1/4})$, improving upon the $\mathcal{O}(dT^{-1/4})$ result of the existing momentum-based signSGD method under the same conditions.
- In distributed settings, we derive convergence rates of $\mathcal{O}\left(\frac{n^{1/2}}{T}+\frac{d}{n^{1/2}}\right)$, $\mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}}+\frac{d}{n^{1/2}}\right)$ and $\mathcal{O}\left(\max\left\{\frac{d^{1/4}}{T^{1/4}},\frac{d^{1/10}}{T^{1/5}}\right\}\right)$, with the latter two substantially outperforming previous results of $\mathcal{O}\left(\frac{d}{T^{1/4}}+\frac{d}{n^{1/2}}\right)$ and $\mathcal{O}\left(\frac{d^{3/8}}{T^{1/8}}\right)$, respectively.

We compare our results with existing methods in Tables 1 and 2.

2 Related Work

In this section, we review the signSGD method and its variants, as well as sign-based methods with majority vote in distributed settings.

Table 2: Summary of results for sign-based algorithms in the distributed setting, where n represents the number of nodes and T denotes the iteration number.

Method	Convergence	Measure
MV-sto-signSGD-SIM (Sun et al., 2023)	$\mathcal{O}\left(\frac{d}{T^{1/4}}+\frac{d}{n^{1/2}}\right)$	l_1
Theorem 3 **(this work)**	$\mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}}+\frac{d}{n^{1/2}}\right)$	
Theorem 4 **(this work)**	$\mathcal{O}\left(\frac{n^{1/2}}{T}+\frac{d}{n^{1/2}}\right)$	
Sto-signSGD (Jin et al., 2021)	$\mathcal{O}\left(\frac{d^{3/8}}{T^{1/8}}\right)$	l_2
Theorem 5 **(this work)**	$\mathcal{O}\left(\max\left\{\frac{d^{1/4}}{T^{1/4}}, \frac{d^{1/10}}{T^{1/5}}\right\}\right)$	

2.1 SIGNSGD AND ITS VARIANTS

The convergence of signSGD is first analyzed by Bernstein et al. (2018), who obtain a rate of $\mathcal{O}(N^{-1/4})$ with a large batch size of $\mathcal{O}(\sqrt{N})$, where N is the number of stochastic gradient calls. They also show that the momentum version of signSGD, named Signum, achieves a convergence rate of $\mathcal{O}(N^{-1/4}\log N)$ with increasingly large batches. To avoid large batch sizes, Bernstein et al. (2019) attain the same convergence rate with a constant batch size, but rely on the strong assumption that the stochastic gradient noise is both unimodal and symmetric, which is not satisfied for many types of noise in practice.

Subsequently, Karimireddy et al. (2019) observe that signSGD with a constant batch size may not converge to optimal points for convex objectives and performs poorly compared to traditional SGD. To address this, they incorporate the compression error into the next update step and show that error feedback enhances practical performance. Rather than assuming unbiased estimation and bounded noise, Safaryan & Richtarik (2021) provide convergence guarantees under the success probability bounds assumption, which posits that the sign of the stochastic gradient matches that of the true gradient with a probability greater than $1/2$. Recently, Sun et al. (2023) analyze the momentum-based version of signSGD and achieve a convergence rate of $\mathcal{O}(dT^{-1/4})$ under standard assumptions. However, their dependence on d can be further improved, as demonstrated by our analysis.

Besides, several other variants have been proposed. For instance, ZO-signSGD (Liu et al., 2019) combines zeroth-order updates with sign information, ensuring gradient-free and communication compression. Jiang et al. (2024) incorporate variance reduction with sign operation, improving the convergence to $\mathcal{O}(T^{-1/3})$ under a stronger smoothness assumption and to $\mathcal{O}(d^{1/2}m^{1/4}T^{-1/2})$ for finite-sum problems, where m denotes the number of functions in the finite-sum structure.

2.2 SIGN-BASED METHODS WITH MAJORITY VOTE

The majority vote technique is employed to enable communication compression in distributed settings. In this framework, each node transmits only the sign of its gradient estimation to the parameter server, which then aggregates the information and sends the sign of the aggregated data back to each node for updating. In the homogeneous setting, Bernstein et al. (2018) first demonstrate that signSGD with majority vote can achieve a convergence rate of $\mathcal{O}(T^{-1/4})$ with large batch sizes. Later, Bernstein et al. (2019) further obtain the same rate with a constant batch size when the noise is unimodal and symmetric. For more challenging heterogeneous environments, the SSDM method (Safaryan & Richtarik, 2021) achieves a convergence rate of $\mathcal{O}(d^{1/2}T^{-1/4})$ under the success probability bounds assumption. However, SSDM only guarantees 1-bit compression in one direction, since the information sent back to each node is not the sign information anymore. To address this, Stochastic-Sign SGD (Jin et al., 2021) ensures 1-bit compression in both directions and achieves a convergence rate of $\mathcal{O}(d^{3/8}T^{-1/8})$ in terms of the l_2-norm. Later, Sun et al. (2023) propose the MV-sto-signSGD-SIM method, attaining a convergence rate of $\mathcal{O}\left(\frac{d}{T^{1/4}}+\frac{d}{n^{1/2}}\right)$. By

incorporating variance reduction techniques, Jiang et al. (2024) improve the convergence rates to $\mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}} + \frac{d}{n^{1/2}}\right)$ and $\mathcal{O}(d^{1/4}T^{-1/4})$, under a stronger average smoothness assumption.

3 SIGNSGD WITH MOMENTUM UPDATES

In this section, we first introduce the assumptions used to analyze sign-based methods and then present our convergence guarantees for signSGD with momentum. Due to space limitations, all proofs are deferred to the Appendix.

3.1 ASSUMPTIONS

We outline the assumptions commonly used to derive convergence guarantees for sign-based methods (Bernstein et al., 2018; 2019).

Assumption 1 *$f_* = \inf_x f(x) > -\infty$ and $f(\mathbf{x}_1) - f_* \leq \Delta_f$ for the initial solution $\mathbf{x}_1$.*

Assumption 2 *(Separable smoothness) The objective function f is separable smooth if there exist non-negative constants $[L_1, L_2, \ldots, L_d]$ such that*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2}\sum_{i=1}^{d} L_i(\mathbf{y}_i - \mathbf{x}_i)^2.$$

Assumption 3 *(Separable bounded noise) For non-negative constants $[\sigma_1, \sigma_2, \cdots, \sigma_d]$, we have*

$$\mathbb{E}_\xi \left[\left([\nabla f(\mathbf{x}; \xi)]_i - [\nabla f(\mathbf{x})]_i \right)^2 \right] \leq \sigma_i^2.$$

Instead of using Assumptions 2 and 3, other literature (Sun et al., 2023; Jiang et al., 2024) employs the following assumptions alternatively.

Assumption 4 *(l_2-smoothness) The objective function f is L-smooth if*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|.$$

Assumption 5 *(Bounded noise) The stochastic gradient noise is bounded such that*

$$\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|^2 \right] \leq \sigma^2.$$

Remark: To align with different literature, we provide two distinct theorems in the next subsection, derived under Assumptions 1, 2, 3 and Assumptions 1, 4, 5 respectively.

3.2 THE CONVERGENCE GUARANTEES

Here, we introduce the sign-based method with momentum updates and present the corresponding convergence guarantees. The traditional signSGD method uses the sign of the stochastic gradient for updates, in the form of equation (3). In contrast to the signSGD method, we track the gradient using a momentum estimator $\mathbf{v}_t$, defined as

$$\mathbf{v}_t = (1 - \beta)\mathbf{v}_{t-1} + \beta \nabla f(\mathbf{x}_t; \xi_t), \tag{4}$$

where β is the momentum parameter and we use $\mathbf{v}_1 = \nabla f(\mathbf{x}_1; \xi_1)$ for the first iteration. After computing the estimator $\mathbf{v}_t$, we update the decision variable using the sign of $\mathbf{v}_t$ as follows:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}(\mathbf{v}_t). \tag{5}$$

The full algorithm is outlined in Algorithm 1, which is called Signum in the previous work (Bernstein et al., 2018) (also named as signSGD-SIM by Sun et al. (2023)). Our contribution lies in the improved theoretical analysis. To compare with previous signSGD studies (Bernstein et al., 2018; 2019), we first provide guarantees under the separable smoothness assumption.

Algorithm 1 Signum

Input: iteration number T, initial point $\mathbf{x}_1$
for time step $t = 1$ **to** T **do**
  if $t == 1$ **then**
    Compute $\mathbf{v}_t = \nabla f(\mathbf{x}_t; \xi_t)$
  else
    Compute $\mathbf{v}_t = (1-\beta)\mathbf{v}_{t-1} + \beta \nabla f(\mathbf{x}_t; \xi_t)$
  end if
  Update the decision variable: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}(\mathbf{v}_t)$
end for
Select τ uniformly at random from $\{1, \ldots, T\}$
Return $\mathbf{x}_\tau$

Theorem 1 *Under Assumptions 1, 2 and 3, by setting $\beta = \mathcal{O}\left(T^{-1/2}\right)$ and $\eta = \mathcal{O}\left(T^{-3/4}\right)$, Algorithm 1 ensures that*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|_1\right] \leq \mathcal{O}\left(\frac{1}{T^{1/4}}\right).$$

Remark: The above rate implies a sample complexity of $\mathcal{O}(\epsilon^{-4})$, matching the state-of-the-art results for signSGD (Bernstein et al., 2018; 2019). However, our method does not require large batch sizes which can be as large as $\mathcal{O}(\epsilon^{-2})$ for signSGD (Bernstein et al., 2018), and avoids the unimodal symmetric noise assumption required by Bernstein et al. (2019).

Next, we also provide the theoretical guarantee under the l_2-smoothness assumption.

Theorem 2 *Under Assumptions 1, 4 and 5, by setting $\beta = \mathcal{O}\left(T^{-1/2}\right)$ and $\eta = \mathcal{O}\left(d^{-1/2}T^{-3/4}\right)$, Algorithm 1 ensures that*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|_1\right] \leq \mathcal{O}\left(\frac{d^{1/2}}{T^{1/4}}\right).$$

Remark: This rate implies a sample complexity of $\mathcal{O}(d^2\epsilon^{-4})$, an improvement over the $\mathcal{O}(d^4\epsilon^{-4})$ results of previous sign-based momentum methods (Sun et al., 2023). This improvement is especially significant when the dimension d is large.

Remark: In Theorem 1, by using the separable smoothness and separable bounded noise assumptions (Assumptions 2 and 3), we can directly analyze under the ℓ_1-norm and provide coordinate-wise bounds, thus avoiding the $d^{1/2}$ dependency.

Source of Theoretical Improvement: In the previous work (Bernstein et al., 2018), to bound the term $\sum_i |[\nabla f(\mathbf{x}_t)]_i| \cdot \mathbb{P}\left(\operatorname{sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{sign}([\mathbf{v}_t]_i)\right)$ appeared in the analysis, they apply $\mathbb{P}\left(\operatorname{sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{sign}(\nabla f(\mathbf{x}_t; \xi_t)_i)\right) \leq \frac{\sigma_i}{\sqrt{n_i}|[\nabla f(\mathbf{x}_t)]_i|}$, which inevitably requires *huge batch sizes* n_i to ensure convergence. Later work (Bernstein et al., 2019) assumes *unimodal symmetric noise* to deal with $\mathbb{P}\left(\operatorname{sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{sign}([\mathbf{v}_t]_i)\right)$. While in our analysis, we find that standard assumption is already adequate and we use $\sum_i |[\nabla f(\mathbf{x}_t)]_i| \cdot \mathbb{P}\left(\operatorname{sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{sign}([\mathbf{v}_t]_i)\right) \leq \sum_i |[\nabla f(\mathbf{x}_t)]_i - [\mathbf{v}_t]_i| \leq \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|_1$ in the analysis. Since we further provide a tighter bound for the estimation error $\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|_1$ compared to Sun et al. (2023), we achieve the state-of-the-art convergence rate without relying on additional assumptions.

3.3 Sharpness of the Obtained Rates

The convergence lower bound for stochastic optimization is $\Omega(T^{-1/4})$ in the l_2-norm (Arjevani et al., 2023). Since we know that $\|z\|_1 \geq \|z\|_2$, this lower bound also implies that $\mathbb{E}[\|\nabla f(\mathbf{x}_\tau)\|_1] \geq \mathbb{E}[\|\nabla f(\mathbf{x}_\tau)\|_2] \geq \Omega(T^{-1/4})$, indicating that our result is optimal with respect to T.

Regarding the $d^{1/2}$ factor in the convergence rate, several pieces of evidence suggest that this factor is inherent for l_1-norm convergence under the standard l_2-smoothness assumption:

- Jiang et al. (2025) establish an $\Omega\left(\sqrt{\frac{d\|L\|_\infty}{T}}+\frac{d^{1/4}\left(\sum_{i=1}^d \sigma_i\sqrt{L_i}\right)^{1/2}}{T^{1/4}}\right)$ lower bound for SGD when measured with the l_1-norm. Suppose that $\{\sigma_i\}$ and $\{L_i\}$ have the same value across coordinates such that $\sigma_i=\sigma/\sqrt{d}, L_i=L$, the lower bound becomes $\Omega\left(\frac{d^{1/2}}{T^{1/4}}\right)$, which confirms the $\sqrt{d}$ factor is required in the l_1-norm setting.
- Prior works (Bernstein et al., 2018; Dong et al., 2024) have already conducted extensive experiments on various vision and language tasks, and find that the ratio of the gradient norm $r=\|\nabla f(x)\|_1 / \|\nabla f(x)\|_2$ always stay close to the level of $\Theta(\sqrt{d})$, supporting the presence of the $\sqrt{d}$ factor in the l_1 measure from the empirical sense.
- Existing rates for sign-based methods under the l_1-norm and l_2-smoothness assumption also include the $\sqrt{d}$ dependency, or even worse (Jin et al., 2021; Sun et al., 2023). Our Theorem 2 already improves the d-dependency from Sun et al. (2023) under the same assumptions.

4 Majority Vote signSGD with Momentum Updates

We first present the problem formulation and the assumptions used. Then, we introduce the proposed method and establish the convergence guarantees.

4.1 Problem Formulation and Assumptions

Sign-based methods are highly communication-efficient in distributed settings, as they only require 1-bit sign information for updates. Previous literature (Bernstein et al., 2018; 2019; Jin et al., 2021; Sun et al., 2023) has analyzed sign-based methods with majority vote in distributed environments. To begin with, consider the following distributed learning problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \coloneqq \frac{1}{n}\sum_{j=1}^n f_j(\mathbf{x}), \quad f_j(\mathbf{x})=\mathbb{E}_{\xi^j\sim\mathcal{D}_j}\left[f_j(\mathbf{x};\xi^j)\right], \tag{6}$$

where $\mathcal{D}_j$ represents the data distribution on node j, and $f_j(\mathbf{x})$ is the corresponding loss function.

Early studies (Bernstein et al., 2018; 2019) focus on homogeneous settings, where $\mathcal{D}_j$ and f_j are identical across nodes. For the more difficult heterogeneous setting, Jin et al. (2021) derive a convergence rate of $\mathcal{O}\left(d^{3/8}T^{-1/8}\right)$ and Sun et al. (2023) achieve the rate of $\mathcal{O}\left(\frac{d}{T^{1/4}}+\frac{d}{n^{1/2}}\right)$. However, these rates can still be improved based on our analysis.

Next, we introduce the assumptions required in this section, which are standard and commonly used in previous works (Jin et al., 2021; Sun et al., 2023).

Assumption 6 *(Smoothness on node j) For each node $j\in[n]$, we suppose*

$$\|\nabla f_j(\mathbf{x})-\nabla f_j(\mathbf{y})\|\le L\|\mathbf{x}-\mathbf{y}\|.$$

Assumption 7 *(Bounded noise on node j) For each node $j\in[n]$, we have*

$$\mathbb{E}_\xi\left[\|\nabla f_j(\mathbf{x};\xi)-\nabla f_j(\mathbf{x})\|^2\right]\le\sigma^2.$$

Assumption 8 *(Bounded gradients) For each node $j\in[n]$, we assume* $\sup_{\mathbf{x}}\|\nabla f_j(\mathbf{x};\xi)\|_\infty\le G$.

Remark: The bounded gradients assumption is standard and widely employed for sign-based optimization in heterogeneous settings (Jin et al., 2021; Sun et al., 2023; Tang et al., 2024). Also note that our Assumption 8 is strictly weaker than the one used by Sun et al. (2023), which requires bounded gradients in the l_2-norm, i.e., $\sup_{\mathbf{x}}\|\nabla f_j(\mathbf{x};\xi)\|_2\le G$.

4.2 The Proposed Method

In this subsection, we introduce the proposed method for distributed environments and present the corresponding convergence guarantees. For distributed settings, the most straightforward approach

Algorithm 2 Majority vote signSGD with momentum (MVSM)

Input: iteration number T, initial point $\mathbf{x}_1$
for time step $t = 1$ **to** T **do**
  On node $j \in \{1, 2, \cdots, n\}$:
    Compute $\mathbf{v}_t^j = (1-\beta)\mathbf{v}_{t-1}^j + \beta \nabla f_j(\mathbf{x}_t; \xi_t^j)$
    Send $\mathrm{S}_G(\mathbf{v}_t^j)$ to the parameter server
  On parameter server:
    (v1) Send $\mathbf{v}_t = \operatorname{sign}\left(\frac{1}{n}\sum_{j=1}^n \mathrm{S}_G\left(\mathbf{v}_t^j\right)\right)$ to all nodes
    (v2) Send $\mathbf{v}_t = \mathrm{S}_1\left(\frac{1}{n}\sum_{j=1}^n \mathrm{S}_G\left(\mathbf{v}_t^j\right)\right)$ to all nodes
  On node $j \in \{1, 2, \cdots, n\}$:
    Update the decision variable $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$
end for
Select τ uniformly at random from $\{1, \ldots, T\}$
Return $\mathbf{x}_\tau$

is to apply the sign operation twice:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}\left(\frac{1}{n}\sum_{j=1}^n \operatorname{sign}(\mathbf{v}_t^j)\right), \tag{7}$$

where $\mathbf{v}_t^j$ is the gradient estimator at node j. In this formulation, each node transmits the sign of its gradient estimate $\operatorname{sign}(\mathbf{v}_t^j)$ to the server. The server then aggregates these sign values and broadcasts back the sign of the resulting information $\operatorname{sign}\left(\frac{1}{n}\sum_{j=1}^n \operatorname{sign}(\mathbf{v}_t^j)\right)$ to each node for updating. This approach ensures 1-bit communication in both directions. However, the sign operation introduces bias in the estimation, and applying it twice can significantly amplify this bias. To mitigate this, we introduce an unbiased sign operation (Sun et al., 2023) as stated below.

Definition 1 *For any vector* $\mathbf{v}$ *with* $\|\mathbf{v}\|_\infty \le R$*, the function* $\mathrm{S}_R(\mathbf{v})$ *is defined component-wise by:*

$$[\mathrm{S}_R(\mathbf{v})]_k = \begin{cases} 1, & \text{with probability } \frac{R+[\mathbf{v}]_k}{2R}, \\ -1, & \text{with probability } \frac{R-[\mathbf{v}]_k}{2R}. \end{cases} \tag{8}$$

Remark: This operation provides an unbiased estimate of $\mathbf{v}/R$, since $\mathbb{E}[\mathrm{S}_R(\mathbf{v})] = \mathbf{v}/R$.

We can now introduce our majority vote signSGD with momentum updates. First, we use the momentum gradient estimator at each node j as follows:

$$\mathbf{v}_t^j = (1-\beta)\mathbf{v}_{t-1}^j + \beta \nabla f_j(\mathbf{x}_t; \xi_t^j), \tag{9}$$

where β is the momentum parameter. Next, by communicating the gradient estimators with the unbiased sign operation $\mathrm{S}_G(\cdot)$, we update the decision variable as follows:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{Sign}\left(\frac{1}{n}\sum_{j=1}^n \mathrm{S}_G(\mathbf{v}_t^j)\right). \tag{10}$$

After applying $\mathrm{S}_G(\cdot)$, the output is a sign vector, which can be efficiently transmitted between nodes. The complete algorithm is described in Algorithm 2 (v1), called Majority Vote SignSGD with Momentum (MVSM). For $t = 1$, we initialize $\mathbf{v}_1^j = \nabla f_j(\mathbf{x}_1; \xi_1^j)$. MVSM-v1 is identical to MV-sto-signSGD-SIM (with $\alpha = 0$) from Sun et al. (2023). However, our analysis yields stronger convergence guarantees as stated below.

Theorem 3 *Under Assumptions 1, 6, 7 and 8, by setting that* $\beta = \frac{1}{2}$ *and* $\eta = \mathcal{O}\left(\frac{1}{T^{1/2}d^{1/2}}\right)$*, our MVSM (v1) method ensures the following convergence:*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|_1\right] \le \mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}} + \frac{d}{n^{1/2}}\right).$$

Remark: Our rate is superior to the previous result of $\mathcal{O}\left(\frac{d}{T^{1/4}}+\frac{d}{n^{1/2}}\right)$, indicating our significant improvement over prior work (Sun et al., 2023) in both d and T dependencies.

By adjusting the learning rate, we can also obtain the following convergence guarantee.

Theorem 4 *Under Assumptions 1, 6, 7 and 8, by setting $\beta=\frac{1}{2}$ and $\eta=\mathcal{O}(n^{-1/2})$, our MVSM (v1) method ensures:*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|_1\right] \leq \mathcal{O}\left(\frac{n^{1/2}}{T}+\frac{d}{n^{1/2}}\right).$$

Remark: This rate improves Theorem 3 when $T \geq \frac{n}{d}$, which is easily satisfied when d is large.

Source of Theoretical Improvement: The improvement for Algorithm 2 (v1) lies in deriving a tighter error bound for the gradient estimator, i.e., $\epsilon_t = \left\|\nabla f(\mathbf{x}_t)-\frac{1}{n}\sum_{j=1}^n \mathbf{v}_t^j\right\|_2^2$. By carefully analyzing the aggregated estimator $\frac{1}{n}\sum_{j=1}^n \mathbf{v}_t^j$, we obtain the recurrence: $\epsilon_{t+1} = (1-\beta)\epsilon_t + \frac{\sigma^2\beta^2}{n}+\frac{2L^2\eta^2 d}{\beta}$, which allows fast decay of ϵ_t with appropriate choices of β and η.

Although the above theorems achieve better convergence rates than previous methods, they do not converge to zero as T increases. To address this issue, we replace the sign operation in the server with the unbiased sign operation $\mathrm{S}_1(\cdot)$ as defined in equation (8) with $R=1$. The revised formulation for the update is:

$$\mathbf{v}_t = \mathrm{S}_1\left(\frac{1}{n}\sum_{j=1}^n \mathrm{S}_G\left(\mathbf{v}_t^j\right)\right). \tag{11}$$

The corresponding algorithm is presented in Algorithm 2 (v2), with the only modification in Step 8. We now present the convergence guarantee for this modified approach.

Theorem 5 *Under Assumptions 1, 6, 7 and 8, by setting that $\eta=\mathcal{O}\left(\min\left\{\frac{1}{T^{1/2}d^{1/2}},\frac{1}{T^{3/5}d^{1/5}}\right\}\right)$ and $\beta=\mathcal{O}\left(\eta^{2/3}d^{1/3}\right)$, the MVSM (v2) method ensures the following convergence:*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|_2\right] \leq \mathcal{O}\left(\max\left\{\frac{d^{1/4}}{T^{1/4}},\frac{d^{1/10}}{T^{1/5}}\right\}\right).$$

Remark: This convergence rate approaches zero as $T\to\infty$, and significantly improves upon the previous result of $\mathcal{O}\left(\frac{d^{3/8}}{T^{1/8}}\right)$ (Jin et al., 2021), in terms of both T and d.

Source of Theoretical Improvement: The improved rate stems from the unbiased estimation of the full gradient, allowing us to use $\mathbb{E}\left[S_1\left(\frac{1}{n}\sum_{j=1}^n S_G(\mathbf{v}_t^j)\right)\right]=\frac{1}{nG}\sum_{j=1}^n \mathbf{v}_t^j$ in the analysis. In contrast, prior works (Jin et al., 2021; Sun et al., 2023) use biased sign operators on the server, which leads to looser bounds and higher complexities.

5 Experiments

In this section, we evaluate the performance of our methods through numerical experiments. We first assess the Signum algorithm in a centralized setting, and then test the proposed MVSM method in the distributed learning environment. All experiments are conducted on NVIDIA GeForce RTX 3090 GPUs, and results are averaged over 10 runs, with shaded regions representing the standard deviation. For each optimizer, hyperparameters are determined through grid search. Specifically, the momentum parameter β is selected from the set $\{0.9, 0.5, 0.1, 0.01\}$, and the learning rate η is chosen from the set $\{0.5, 0.25, 0.1, 0.05, 0.025, 0.01\}\times 10^{-2}$.

5.1 Experiments in the Centralized Environment

We validate the effectiveness of sign-based methods on the image classification task. Specifically, we train a ResNet-18 model (He et al., 2016) on the CIFAR-10 dataset (Krizhevsky, 2009) and

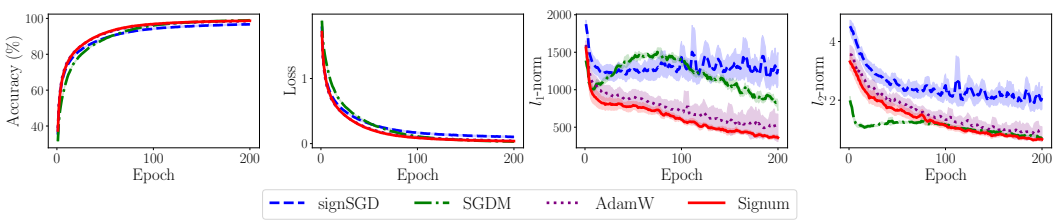

Figure 1: Results for CIFAR-10 dataset in the centralized environment.

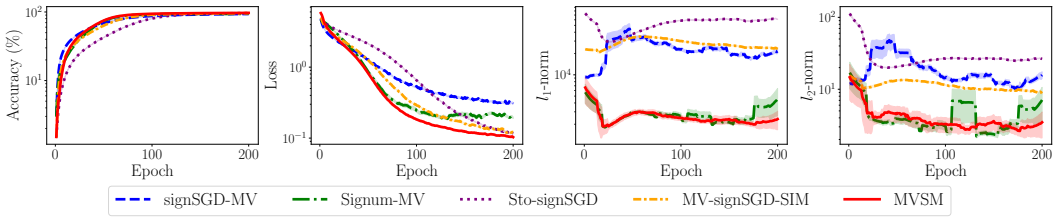

Figure 2: Results for CIFAR-100 dataset in the distributed environment.

compare our method against signSGD (Bernstein et al., 2018), SGDM (Sutskever et al., 2013), and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). For SGDM and AdamW, we use the official PyTorch implementations (Paszke et al., 2019).

Figure 1 reports the training loss, accuracy, and the l_1- and l_2-norms of the gradients. In terms of loss and accuracy, the Signum method converges fastest even among the algorithms that update with full gradient information. Additionally, the Signum method results in the most rapid reduction of both l_1 and l_2 gradient norms. These findings are consistent with our theoretical results, further highlighting the effectiveness of momentum-based sign methods in accelerating convergence and improving optimization efficiency.

5.2 Experiments in the Distributed Environment

Next, we evaluate our method in the distributed setting. We train a ResNet-50 model (He et al., 2016) on the CIFAR-100 dataset (Krizhevsky, 2009) across 8 nodes. We compare our MVSM method against signSGD (with Majority Vote) (Bernstein et al., 2018), Signum (with Majority Vote) (Bernstein et al., 2019), Sto-signSGD (Jin et al., 2021), and MV-signSGD-SIM (Sun et al., 2023).

Figure 2 presents the training loss, accuracy, and the l_1- and l_2-norms of the gradients. Our MVSM algorithm achieves the lowest loss and highest accuracy, while also exhibiting sparser gradients compared to other methods. In contrast, sign-based optimizers that do not incorporate momentum updates—specifically, signSGD-MV and Sto-signSGD—exhibit poor performance and produce significantly larger gradients. These results further underscore the advantage of integrating momentum into sign-based optimization methods.

6 Conclusion

In this paper, we demonstrate that signSGD with momentum update can achieve a convergence rate of $\mathcal{O}(T^{-1/4})$ without requiring large batch sizes or assuming unimodal symmetric noise. When analyzed under the l_2-smoothness assumption, our method achieves a convergence rate of $\mathcal{O}(d^{1/2}T^{-1/4})$, which improves upon the previous rate of $\mathcal{O}(dT^{-1/4})$. In distributed settings, we establish convergence rates of $\mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}} + \frac{d}{n^{1/2}}\right)$ and $\mathcal{O}\left(\max\left\{\frac{d^{1/4}}{T^{1/4}}, \frac{d^{1/10}}{T^{1/5}}\right\}\right)$, which significantly outperform prior results of $\mathcal{O}\left(\frac{d}{T^{1/4}} + \frac{d}{n^{1/2}}\right)$ and $\mathcal{O}\left(\frac{d^{3/8}}{T^{1/8}}\right)$. Finally, numerical experiments in different learning environments also validate the effectiveness of the proposed method.

Reproducibility statement

We provide clear explanations of all assumptions and include complete proofs of our theoretical claims in the appendix. For the experimental results, we specify the dataset, baseline methods, and hyperparameter choices.

The use of LLMs

We used large language models (LLMs) solely for minor language polishing of the manuscript. The LLMs did not contribute to research ideation, algorithm design, theoretical analysis, or experimental work. Their role was strictly limited to assisting with improving readability and grammar.

References

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake E. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199 (1-2):165–214, 2023.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 560–569, 2018.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019.

Yiming Dong, Huan Li, and Zhouchen Lin. Convergence rate analysis of lion. *ArXiv e-prints*, arXiv:2411.07724, 2024.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems 31*, pp. 689–699, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of adagrad over sgd: Upper and lower bounds in stochastic non-convex optimization. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, pp. 3124–3158, 2025.

Wei Jiang, Sifan Yang, Wenhao Yang, and Lijun Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. In *Advances in Neural Information Processing Systems 37*, pp. 33891–33932, 2024.

Richeng Jin, Yufan Huang, Xiaofan He, Huaiyu Dai, and Tianfu Wu. Stochastic-Sign SGD for federated learning with theoretical guarantees. *ArXiv e-prints*, arXiv:2002.10940, 2021.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3252–3261, 2019.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Masters Thesis, Deptartment of Computer Science, University of Toronto*, 2009.

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

Yuxing Liu, Rui Pan, and Tong Zhang. Adagrad under anisotropic smoothness. In *International Conference on Learning Representations*, 2025.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 2019.

Mher Safaryan and Peter Richtarik. Stochastic sign descent methods: New algorithms and better theory. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9224–9234, 2021.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 33077–33099, 2023.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1139–1147, 2013.

Zhiwei Tang, Yanmeng Wang, and Tsung-Hui Chang. z-signfedavg: A unified stochastic sign-based compression for federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15301–15309, 2024.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems 32*, pp. 2406–2416, 2019.

APPENDIX

A PROOF OF THEOREM 1

According to Assumption 2 and considering the update $\mathbf{x}_{t+1} - \mathbf{x}_t = -\eta \operatorname{Sign}(\mathbf{v}_t)$, we know that

$$\begin{aligned}
f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq & \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{1}{2} \sum_{i=1}^{d} L_i ([\mathbf{x}_{t+1}]_i - [\mathbf{x}_t]_i)^2 \\
\leq & - \langle \nabla f(\mathbf{x}_t), \eta \operatorname{Sign}(\mathbf{v}_t) \rangle + \frac{\eta^2}{2} \sum_{i=1}^{d} L_i ([\operatorname{Sign}(\mathbf{v}_t)]_i)^2 \\
\leq & - \langle \nabla f(\mathbf{x}_t), \eta \operatorname{Sign}(\nabla f(\mathbf{x}_t)) \rangle \\
& + \eta \langle \nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) - \operatorname{Sign}(\mathbf{v}_t) \rangle + \frac{\eta^2}{2} \sum_{i=1}^{d} L_i \\
\leq & - \eta \left\| \nabla f(\mathbf{x}_t) \right\|_1 + 2\eta \left\| \nabla f(\mathbf{x}_t) - \mathbf{v}_t \right\|_1 + \frac{\eta^2}{2} \sum_{i=1}^{d} L_i,
\end{aligned}$$

where the last inequality is due to

$$\begin{aligned}
& \langle \nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) - \operatorname{Sign}(\mathbf{v}_t) \rangle \\
= & \sum_{i=1}^{d} [\nabla f(\mathbf{x}_t)]_i \cdot (\operatorname{Sign}[\nabla f(\mathbf{x}_t)]_i - \operatorname{Sign}[\mathbf{v}_t]_i) \\
\leq & \sum_{i=1}^{d} 2 [\nabla f(\mathbf{x}_t)]_i \cdot \mathbb{I}(\operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{Sign}[\mathbf{v}_t]_i) \\
\leq & \sum_{i=1}^{d} 2 |[\nabla f(\mathbf{x}_t)]_i - [\mathbf{v}_t]_i| \cdot \mathbb{I}(\operatorname{Sign}[\nabla f(\mathbf{x}_t)]_i \neq \operatorname{Sign}[\mathbf{v}_t]_i) \\
\leq & \sum_{i=1}^{d} 2 |[\nabla f(\mathbf{x}_t)]_i - [\mathbf{v}_t]_i| = 2 \left\| \nabla f(\mathbf{x}_t) - \mathbf{v}_t \right\|_1.
\end{aligned}$$

Rearranging the obtained relation and summing up yields

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \frac{\Delta_f}{\eta T} + 2\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|_1 \right] + \frac{\eta}{2} \sum_{i=1}^{d} L_i, \tag{12}$$

where we define $\Delta_f = f(\mathbf{x}_1) - f_*$.

Next, we proceed to bound the error term $\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|_1 \right]$. For convenience, we define the following notations:

$$\boldsymbol{\epsilon}_t := \mathbf{v}_t - \nabla f(\mathbf{x}_t), \quad \mathbf{n}_t := \nabla f(\mathbf{x}_t; \xi_t) - \nabla f(\mathbf{x}_t), \quad \mathbf{s}_t := \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t).$$

By definition, we have

$$\begin{aligned}
\boldsymbol{\epsilon}_t = & \mathbf{v}_t - \nabla f(\mathbf{x}_t) = (1-\beta)\mathbf{v}_{t-1} + \beta \nabla f(\mathbf{x}_t; \xi_t) - \nabla f(\mathbf{x}_t) \\
= & (1-\beta)(\mathbf{v}_{t-1} - \nabla f(\mathbf{x}_{t-1})) + (1-\beta)(\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t)) + \beta(\nabla f(\mathbf{x}_t; \xi_t) - \nabla f(\mathbf{x}_t)) \\
= & (1-\beta)\boldsymbol{\epsilon}_{t-1} + (1-\beta)\mathbf{s}_t + \beta \mathbf{n}_t.
\end{aligned}$$

Performing this recursively yields

$$\boldsymbol{\epsilon}_t = (1-\beta)^{t-1}\mathbf{n}_1 + \beta \sum_{k=2}^{t} (1-\beta)^{t-k} \mathbf{n}_k + \sum_{k=2}^{t} (1-\beta)^{t-k+1} \mathbf{s}_k,$$

where we use the fact that $\boldsymbol{\epsilon}_1 = \mathbf{v}_1 - \nabla f(\mathbf{x}_1) = \nabla f(\mathbf{x}_1;\xi_1) - \nabla f(\mathbf{x}_1) = \mathbf{n}_1$. We bound $\boldsymbol{\epsilon}_t$ via two terms $\mathtt{A_t}$ and $\mathtt{B_t}$ as follows:

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] \le \underbrace{\mathbb{E}\left[\left\|(1-\beta)^{t-1}\mathbf{n}_1 + \beta\sum_{k=2}^{t}(1-\beta)^{t-k}\mathbf{n}_k\right\|_1\right]}_{\mathtt{A_t}} + \underbrace{\mathbb{E}\left[\left\|\sum_{k=2}^{t}(1-\beta)^{t-k+1}\mathbf{s}_k\right\|_1\right]}_{\mathtt{B_t}}$$

Firstly, we cope with $\mathtt{A_t}$ following the similar procedure as in Liu et al. (2025, Lemma E.2). We denote the i-th element of the vector $\mathbf{n}_t$ by $\mathbf{n}_{t,i}$. By the Cauchy–Schwarz inequality, for any $\lambda_1, \cdots, \lambda_d > 0$, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left\|(1-\beta)^{t-1}\mathbf{n}_1 + \beta\sum_{k=2}^{t}(1-\beta)^{t-k}\mathbf{n}_k\right\|_1^2\right] \\
&\le \left(\sum_{i=1}^{d}\lambda_i\right)\sum_{i=1}^{d}\frac{1}{\lambda_i}\mathbb{E}\left[(1-\beta)^{t-1}\mathbf{n}_{1,i} + \beta\sum_{k=2}^{t}(1-\beta)^{t-k}\mathbf{n}_{k,i}\right]^2 \\
&= \left(\sum_{i=1}^{d}\lambda_i\right)\sum_{i=1}^{d}\frac{1}{\lambda_i}\left((1-\beta)^{2t-2}\mathbb{E}\left[\mathbf{n}_{1,i}^2\right] + \beta^2\sum_{k=2}^{t}(1-\beta)^{2(t-k)}\mathbb{E}\left[\mathbf{n}_{k,i}^2\right]\right) \\
&\le \left(\sum_{i=1}^{d}\lambda_i\right)\sum_{i=1}^{d}\frac{1}{\lambda_i}\left((1-\beta)^{2t-2}\sigma_i^2 + \beta^2\sum_{k=2}^{t}(1-\beta)^{2(t-k)}\sigma_i^2\right) \\
&\le \left(\sum_{i=1}^{d}\lambda_i\right)\sum_{i=1}^{d}\frac{\sigma_i^2}{\lambda_i}\left((1-\beta)^{2t-2} + \frac{\beta}{2-\beta}\right),
\end{aligned}$$

where the equality is due to $\mathbb{E}\left[\mathbf{n}_{s,i}\cdot\mathbf{n}_{t,i}\right] = 0, \forall s < t \in [T], \forall i \in [d]$; the second inequality is due to Assumption 3. Denoting by $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_d]^\top$ and setting $\lambda_i = \sigma_i$, we obtain

$$\begin{aligned}
\mathtt{A_t} &\le \sqrt{\mathbb{E}\left[\left\|(1-\beta)^{t-1}\mathbf{n}_1 + \beta\sum_{k=2}^{t}(1-\beta)^{t-k}\mathbf{n}_k\right\|_1^2\right]} \\
&\le \sqrt{\left((1-\beta)^{2t-2} + \frac{\beta}{2-\beta}\right)\|\boldsymbol{\sigma}\|_1^2} \le \left((1-\beta)^{t-1} + \sqrt{\frac{\beta}{2-\beta}}\right)\|\boldsymbol{\sigma}\|_1,
\end{aligned}$$

where we make use of $\mathbb{E}^2[X] \le \mathbb{E}[X^2]$ and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}, \forall a, b \ge 0$.

Secondly, we cope with $\mathtt{B_t}$ as the following:

$$\mathtt{B_t} \le \sum_{k=2}^{t}(1-\beta)^{t-k+1}\mathbb{E}\left[\|\mathbf{s}_k\|_1\right] \le 2\eta\|\vec{L}\|_1\sum_{k=2}^{t}(1-\beta)^{t-k+1} \le \frac{2(1-\beta)\eta\|\vec{L}\|_1}{\beta},$$

where the second inequality uses

$$\mathbb{E}\left[\|\mathbf{s}_k\|_1\right] = \|\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t)\|_1 = \|\nabla f(\mathbf{x}_t + \eta\,\mathrm{Sign}\left(\mathbf{v}_{t-1}\right)) - \nabla f(\mathbf{x}_t)\|_1 \le 2\eta\|\vec{L}\|_1,$$

which is due to the following lemma.

Lemma 1 *(Lemma F.3. in Bernstein et al. (2018)) Under Assumption 2, for any sign vector* $\mathbf{s} \in \{-1, 1\}^d$*, any* $\mathbf{x} \in \mathbb{R}^d$ *and any* η

$$\|\nabla f(\mathbf{x} + \eta s) - \nabla f(\mathbf{x})\|_1 \le 2\eta\|\vec{L}\|_1.$$

Now it suffices to combine the bounds for $\mathtt{A_t}, \mathtt{B_t}$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] \le \frac{1}{T}\sum_{t=1}^{T}\left(\mathtt{A_t} + \mathtt{B_t}\right) \le \left(\frac{1}{T\beta} + \sqrt{\frac{\beta}{2-\beta}}\right)\|\boldsymbol{\sigma}\|_1 + \frac{2(1-\beta)\eta\|\vec{L}\|_1}{\beta},$$

where we make use of $\sum_{t=1}^{T}(1-\beta)^{t-1} \leq 1/\beta$. Plugging this relation into equation 12 yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|_1\right] \leq \frac{\Delta_f}{\eta T} + \frac{\eta\|\vec{L}\|_1}{2} + 2\|\boldsymbol{\sigma}\|_1\left(\frac{1}{T\beta} + \sqrt{\frac{\beta}{2-\beta}}\right) + \frac{4(1-\beta)\eta\|\vec{L}\|_1}{\beta}$$

Setting $\eta = \sqrt{\frac{\Delta_f}{\|\vec{L}\|_1}}\cdot T^{-3/4},\ \beta = \frac{1}{\sqrt{T}}$, we obtain

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_\tau)\|_1\right] \leq \sqrt{\|\vec{L}\|_1\Delta_f}\left(\frac{5}{T^{1/4}} + \frac{1}{2T^{3/4}}\right) + 2\|\boldsymbol{\sigma}\|_1\left(\frac{1}{T^{1/4}} + \frac{1}{\sqrt{T}}\right) = \mathcal{O}\left(\frac{1}{T^{1/4}}\right).$$

B Proof of Theorem 2

Firstly, due to the l_2-smoothness assumption (Assumption 4), we have that

$$\begin{aligned}
&f(\mathbf{x}_{t+1}) \\
&\leq f(\mathbf{x}_t) + \langle\nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= f(\mathbf{x}_t) - \eta\langle\nabla f(\mathbf{x}_t), \operatorname{Sign}(\mathbf{v}_t)\rangle + \frac{\eta^2 L}{2}\|\operatorname{Sign}(\mathbf{v}_t)\|^2 \\
&\leq f(\mathbf{x}_t) + \eta\langle\nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) - \operatorname{Sign}(\mathbf{v}_t)\rangle - \eta\langle\nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t))\rangle + \frac{\eta^2 Ld}{2} \\
&= f(\mathbf{x}_t) + \eta\langle\nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) - \operatorname{Sign}(\mathbf{v}_t)\rangle - \eta\|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 Ld}{2} \\
&= f(\mathbf{x}_t) + \eta\sum_{i=1}^{d}\langle[\nabla f(\mathbf{x}_t)]_i, \operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) - \operatorname{Sign}([\mathbf{v}_t]_i)\rangle - \eta\|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 Ld}{2} \\
&\leq f(\mathbf{x}_t) + \eta\sum_{i=1}^{d} 2\left|[\nabla f(\mathbf{x}_t)]_i\right| \cdot \mathbb{I}\left(\operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{Sign}([\mathbf{v}_t]_i)\right) - \eta\|\nabla f(\mathbf{x}_t)\|_1 \\
&\quad + \frac{\eta^2 Ld}{2} \\
&\leq f(\mathbf{x}_t) + \eta\sum_{i=1}^{d} 2|[\nabla f(\mathbf{x}_t)]_i - [\mathbf{v}_t]_i| \cdot \mathbb{I}\left(\operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{Sign}([\mathbf{v}_t]_i)\right) - \eta\|\nabla f(\mathbf{x}_t)\|_1 \\
&\quad + \frac{\eta^2 Ld}{2} \\
&\leq f(\mathbf{x}_t) + \eta\sum_{i=1}^{d} 2|[\nabla f(\mathbf{x}_t)]_i - [\mathbf{v}_t]_i| - \eta\|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 Ld}{2} \\
&= f(\mathbf{x}_t) + 2\eta\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|_1 - \eta\|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 Ld}{2} \\
&\leq f(\mathbf{x}_t) + 2\eta\sqrt{d}\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| - \eta\|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 Ld}{2}.
\end{aligned} \tag{13}$$

Summing up and rearranging the equation (13), we derive:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|_1\right] \\
\leq&\frac{f(\mathbf{x}_1) - f(\mathbf{x}_{T+1})}{\eta T} + 2\sqrt{d}\cdot\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|\right] + \frac{\eta Ld}{2} \\
\leq&\frac{\Delta_f}{\eta T} + 2\sqrt{d}\cdot\sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2\right]} + \frac{\eta Ld}{2}.
\end{aligned} \tag{14}$$

where we define $\Delta_f = f(\mathbf{x}_1) - f_*$, and the second inequality is due to Jensen's Inequality.

Next, we can bound the term $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2\right]$ as follows.

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t+1}) - \mathbf{v}_{t+1}\|^2\right] &= \mathbb{E}\left[\|(1-\beta)\mathbf{v}_t + \beta\nabla f(\mathbf{x}_{t+1};\xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})\|^2\right] \\
&= \mathbb{E}\left[\|(1-\beta)(\mathbf{v}_t - \nabla f(\mathbf{x}_t)) + \beta\left(\nabla f(\mathbf{x}_{t+1};\xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})\right)\right. \\
&\qquad \left.+(1-\beta)\left(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1})\right)\|^2\right] \\
&= (1-\beta)^2\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1})\|^2\right] \\
&\qquad + \beta^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t+1};\xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})\|^2\right] \\
&\leq (1-\beta)^2(1+\beta)\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t))\|^2\right] \\
&\quad + (1-\beta)^2(1+\frac{1}{\beta})\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1})\|^2\right] + \beta^2\sigma^2 \\
&\leq (1-\beta)\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2\right] + \frac{2L^2}{\beta}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \beta^2\sigma^2 \\
&\leq (1-\beta)\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2\right] + \frac{2\eta^2L^2d}{\beta} + \beta^2\sigma^2,
\end{aligned}
$$

where the third equality is due to the fact $\mathbb{E}\left[\nabla f(\mathbf{x}_{t+1};\xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})\right] = 0$. Summing up, we can ensure

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2\right] &\leq \frac{\mathbb{E}\left[\|\mathbf{v}_1 - \nabla f(\mathbf{x}_1)\|^2\right]}{\beta T} + \frac{2\eta^2L^2d}{\beta^2} + \beta\sigma^2 \\
&\leq \frac{\sigma^2}{\beta T} + \frac{2\eta^2L^2d}{\beta^2} + \beta\sigma^2.
\end{aligned} \tag{15}
$$

Incorporating the above into equation (14) and setting that $\beta = \mathcal{O}\left(T^{-1/2}\right)$, $\eta = \mathcal{O}\left(d^{-1/2}T^{-3/4}\right)$, we observe:

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|_1\right] &\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d}\cdot\sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2\right]} + \frac{\eta L d}{2} \\
&\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d}\cdot\sqrt{\frac{\sigma^2}{\beta T} + \frac{2\eta^2L^2d}{\beta^2} + \beta\sigma^2} + \frac{\eta L d}{2} \\
&= \mathcal{O}\left(\frac{\left(1 + \Delta_f + \sigma + L\right)d^{1/2}}{T^{1/4}}\right) \\
&= \mathcal{O}\left(\frac{d^{1/2}}{T^{1/4}}\right),
\end{aligned}
$$

which finishes the proof of Theorem 2.

C Proof of Theorem 3 and 4

Since the overall objective function $f(\mathbf{x})$ is L-smooth, we have the following:

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) \leq & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
\leq & f(\mathbf{x}_t) - \eta \left\langle \nabla f(\mathbf{x}_t), \operatorname{Sign}\left(\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle + \frac{\eta^2 L d}{2} \\
= & f(\mathbf{x}_t) + \eta \left\langle \nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) - \operatorname{Sign}\left(\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle \\
& - \eta \langle \nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) \rangle + \frac{\eta^2 L d}{2} \\
= & f(\mathbf{x}_t) + \eta \left\langle \nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) - \operatorname{Sign}\left(\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle \\
& - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 L d}{2} \\
\leq & f(\mathbf{x}_t) + 2\eta R \sqrt{d} \left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right\| - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 L d}{2},
\end{aligned}
\tag{16}
$$

where the last inequality is because of

$$
\begin{aligned}
& \left\langle \nabla f(\mathbf{x}_t), \operatorname{Sign}(\nabla f(\mathbf{x}_t)) - \operatorname{Sign}\left(\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle \\
= & \sum_{i=1}^{d} \left\langle [\nabla f(\mathbf{x}_t)]_i, \operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) - \operatorname{Sign}\left(\left[\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right]_i \right) \right\rangle \\
\leq & \sum_{i=1}^{d} 2 \left| [\nabla f(\mathbf{x}_t)]_i \right| \cdot \mathbb{I}\left(\operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{Sign}\left(\left[\frac{1}{n} \sum_{j=1}^{n} S(\mathbf{v}_t^j) \right]_i \right) \right) \\
\leq & \sum_{i=1}^{d} 2R \left| \frac{[\nabla f(\mathbf{x}_t)]_i}{R} \right| \cdot \mathbb{I}\left(\operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{Sign}\left(\left[\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right]_i \right) \right) \\
\leq & \sum_{i=1}^{d} 2R \left| \frac{[\nabla f(\mathbf{x}_t)]_i}{R} - \left[\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right]_i \right| \cdot \mathbb{I}\left(\operatorname{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \operatorname{Sign}\left(\left[\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right]_i \right) \right) \\
\leq & \sum_{i=1}^{d} 2R \left| \frac{[\nabla f(\mathbf{x}_t)]_i}{R} - \left[\frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right]_i \right| \\
= & 2R \left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right\|_1 \\
\leq & 2R\sqrt{d} \left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{n} \sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right\|.
\end{aligned}
\tag{17}
$$

Rearranging and taking the expectation over equation (16), we have:

$$
\begin{aligned}
&\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)\right] \\
\leq& 2\eta G\sqrt{d}\mathbb{E}\left[\left\|\frac{\nabla f(\mathbf{x}_t)}{G} - \frac{1}{n}\sum_{j=1}^{n}\mathrm{S}_G(\mathbf{v}_t^j)\right\|\right] - \eta\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] + \frac{\eta^2 Ld}{2} \\
\leq& 2\eta G\sqrt{d}\mathbb{E}\left[\left\|\frac{\nabla f(\mathbf{x}_t)}{G} - \frac{1}{nG}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|\right] + 2\eta G\sqrt{d}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\mathrm{S}_G(\mathbf{v}_t^j) - \frac{\mathbf{v}_t^j}{G}\right)\right\|\right] \\
&- \eta\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] + \frac{\eta^2 Ld}{2} \\
\leq& 2\eta\sqrt{d}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|\right] + 2\eta G\sqrt{d}\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\mathrm{S}_G(\mathbf{v}_t^j) - \frac{\mathbf{v}_t^j}{G}\right)\right\|^2\right]} \\
&- \eta\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] + \frac{\eta^2 Ld}{2} \\
\leq& 2\eta\sqrt{d}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|\right] + 2\eta G\sqrt{d}\sqrt{\frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\left(\mathrm{S}_G(\mathbf{v}_t^j) - \frac{\mathbf{v}_t^j}{G}\right)\right\|^2\right]} \\
&- \eta\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] + \frac{\eta^2 Ld}{2} \\
\leq& 2\eta\sqrt{d}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|\right] + 2\eta G\sqrt{d}\sqrt{\frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\mathrm{S}_G(\mathbf{v}_t^j)\right\|^2\right]} \\
&- \eta\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] + \frac{\eta^2 Ld}{2} \\
\leq& 2\eta\sqrt{d}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|\right] + \frac{2\eta dG}{\sqrt{n}} - \eta\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] + \frac{\eta^2 Ld}{2},
\end{aligned}
\tag{18}
$$

where the third inequality is due to the fact that $\left(\mathbb{E}\left[X\right]\right)^2 \leq \mathbb{E}\left[X^2\right]$, and the forth inequality is because of $\mathbb{E}\left[S_G\left(\mathbf{v}_t^j\right)\right] = \frac{\mathbf{v}_t^j}{G}$, as well as the S_G operation in each node is independent.

Rearranging the terms and summing up, we have:

$$
\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] &\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d}\mathbb{E}\left[\frac{1}{T}\sum_{i=1}^{T}\left\|\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|\right] + \frac{2dG}{\sqrt{n}} + \frac{\eta Ld}{2} \\
&\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d}\sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{i=1}^{T}\left\|\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|^2\right]} + \frac{2dG}{\sqrt{n}} + \frac{\eta Ld}{2},
\end{aligned}
$$

where the last inequality is due to Jensen's inequality.

For each worker j, we have the following according to the definition of $\mathbf{v}_t^j$:

$$
\begin{aligned}
\mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1}) =& (1-\beta)\left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right) + \beta\left(\nabla f_j(\mathbf{x}_{t+1};\xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1})\right) \\
&+ (1-\beta)\left(\nabla f_j(\mathbf{x}_t) - \nabla f_j(\mathbf{x}_{t+1})\right).
\end{aligned}
$$

Averaging over $\{n\}$ and noting that $\nabla f(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x})$, we can obtain:

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_{t+1}^j - \nabla f(\mathbf{x}_{t+1}) = \frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1})\right)\\
=&(1-\beta)\frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right) + \beta\frac{1}{n}\sum_{j=1}^{n}\left(\nabla f_j(\mathbf{x}_{t+1};\xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1})\right)\\
&+ (1-\beta)\frac{1}{n}\sum_{j=1}^{n}\left(\nabla f_j(\mathbf{x}_t) - \nabla f_j(\mathbf{x}_{t+1})\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_{t+1}^j - \nabla f(\mathbf{x}_{t+1})\right\|^2\right]\\
\leq&(1-\beta)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right)\right\|^2\right] + \beta^2\frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla f_j(\mathbf{x}_{t+1};\xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1})\right\|^2\right]\\
&+ \frac{2}{\beta n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla f_j(\mathbf{x}_{t+1}) - \nabla f_j(\mathbf{x}_t)\right\|^2\right]\\
\leq&(1-\beta)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right)\right\|^2\right] + \frac{\beta^2\sigma^2}{n} + \frac{2L^2}{\beta}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2\\
\leq&(1-\beta)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j - \nabla f(\mathbf{x}_t)\right\|^2\right] + \frac{\beta^2\sigma^2}{n} + \frac{2L^2\eta^2 d}{\beta}.
\end{aligned}$$

By summing up and rearranging, we observe

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j - \nabla f(\mathbf{x}_t)\right\|^2\right] \leq&\frac{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_1^j - \nabla f(\mathbf{x}_1)\right\|^2\right]}{\beta T} + \frac{\beta\sigma^2}{n} + \frac{2L^2\eta^2 d}{\beta^2}\\
\leq&\frac{\sigma^2}{n\beta T} + \frac{\sigma^2\beta}{n} + \frac{2L^2\eta^2 d}{\beta^2}.
\end{aligned} \tag{19}$$

Finally, we can ensure that

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T}\|\nabla f(\mathbf{x}_t)\|_1 \leq& \frac{\Delta_f}{\eta T} + \frac{2dG}{\sqrt{n}} + \frac{\eta L d}{2} + 2\sqrt{d}\sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{i=1}^{T}\left\|\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_t^j\right\|^2\right]}\\
\leq& \frac{\Delta_f}{\eta T} + \frac{2dG}{\sqrt{n}} + \frac{\eta L d}{2} + 2\sqrt{d}\sqrt{\frac{\sigma^2}{n\beta T} + \frac{\sigma^2\beta}{n} + \frac{2L^2\eta^2 d}{\beta^2}}.
\end{aligned}$$

By setting $\beta = \frac{1}{2}$ and $\eta = \mathcal{O}\left(T^{-1/2}d^{-1/2}\right)$, we have

$$\frac{1}{T}\sum_{i=1}^{T}\|\nabla f(\mathbf{x}_t)\|_1 = \mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}} + \frac{d}{n^{1/2}}\right).$$

By setting $\beta = \frac{1}{2}$ and $\eta = \mathcal{O}\left(n^{-1/2}\right)$, we have

$$\frac{1}{T}\sum_{i=1}^{T}\|\nabla f(\mathbf{x}_t)\|_1 = \mathcal{O}\left(\frac{n^{1/2}}{T} + \frac{d}{n^{1/2}}\right).$$

D PROOF OF THEOREM 5

Due to the fact that the overall objective function $f(\mathbf{x})$ is L-smooth, we have the following:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) \leq & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
\leq & f(\mathbf{x}_t) - \eta \left\langle \nabla f(\mathbf{x}_t), \mathrm{S}_1 \left(\frac{1}{n}\sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle + \frac{\eta^2 L d}{2} \\
= & f(\mathbf{x}_t) + \eta \left\langle \nabla f(\mathbf{x}_t), \frac{\nabla f(\mathbf{x}_t)}{G} - \mathrm{S}_1 \left(\frac{1}{n}\sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle \\
& - \eta \left\langle \nabla f(\mathbf{x}_t), \frac{\nabla f(\mathbf{x}_t)}{G} \right\rangle + \frac{\eta^2 L d}{2} \\
= & f(\mathbf{x}_t) + \eta \left\langle \nabla f(\mathbf{x}_t), \frac{\nabla f(\mathbf{x}_t)}{G} - \mathrm{S}_1 \left(\frac{1}{n}\sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle - \frac{\eta}{G}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta^2 L d}{2}.
\end{aligned}$$

Taking expectations leads to:

$$\begin{aligned}
& \mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)\right] \\
\leq & \eta \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}_t), \frac{1}{G}\nabla f(\mathbf{x}_t) - \mathrm{S}_1 \left(\frac{1}{n}\sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right) \right\rangle \right] - \frac{\eta}{G}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \frac{\eta^2 L d}{2} \\
= & \eta \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}_t), \frac{1}{G}\nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n} \mathrm{S}_G(\mathbf{v}_t^j) \right\rangle \right] - \frac{\eta}{G}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \frac{\eta^2 L d}{2} \\
= & \eta \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}_t), \frac{1}{G}\nabla f(\mathbf{x}_t) - \frac{1}{nG}\sum_{j=1}^{n} \mathbf{v}_t^j \right\rangle \right] - \frac{\eta}{G}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \frac{\eta^2 L d}{2} \\
\leq & \eta \mathbb{E}\left[\frac{1}{2G}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2G}\left\| \nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n} \mathbf{v}_t^j \right\|^2 \right] - \frac{\eta}{G}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \frac{\eta^2 L d}{2} \\
= & \frac{\eta}{2G}\mathbb{E}\left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n} \mathbf{v}_t^j \right\|^2 \right] - \frac{\eta}{2G}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \frac{\eta^2 L d}{2}.
\end{aligned} \tag{20}$$

Rearranging the terms and summing up:

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2] &\leq \frac{2\Delta_f G}{\eta T} + \mathbb{E}\left[\frac{1}{T}\sum_{i=1}^{T} \left\| \nabla f(\mathbf{x}_t) - \frac{1}{n}\sum_{j=1}^{n} \mathbf{v}_t^j \right\|^2 \right] + \eta L d G \\
&\leq \frac{2\Delta_f G}{\eta T} + \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \frac{1}{T}\sum_{i=1}^{T} \left\| \nabla f_j(\mathbf{x}_t) - \mathbf{v}_t^j \right\|^2 \right] + \eta L d G.
\end{aligned}$$

For each worker j, according to the definition of $\mathbf{v}_t^j$, we have:

$$\begin{aligned}
\mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1}) = (1-\beta)\left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right) + \beta\left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1})\right) \\
+ (1-\beta)\left(\nabla f_j(\mathbf{x}_t) - \nabla f_j(\mathbf{x}_{t+1})\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1})\right\|^2\right] \\
\leq&(1-\beta)\mathbb{E}\left[\left\|\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right\|^2\right] + \beta^2\mathbb{E}\left[\left\|\nabla f_j(\mathbf{x}_{t+1};\xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1})\right\|^2\right] \\
&+ \frac{2}{\beta}\mathbb{E}\left[\left\|\nabla f_j(\mathbf{x}_{t+1}) - \nabla f_j(\mathbf{x}_t)\right\|^2\right] \\
\leq&(1-\beta)\mathbb{E}\left[\left\|\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right\|^2\right] + \beta^2\sigma^2 + \frac{2L^2}{\beta}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2 \\
\leq&(1-\beta)\mathbb{E}\left[\left\|\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right\|^2\right] + \beta^2\sigma^2 + \frac{2L^2\eta^2 d}{\beta}.
\end{aligned}$$

As a result, we know that

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n\frac{1}{T}\sum_{t=1}^T\left\|\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t)\right\|^2\right] \leq \frac{\sigma^2}{\beta T} + \sigma^2\beta + \frac{2L^2\eta^2 d}{\beta^2}.$$

Finally, we can obtain the final bound:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{i=1}^T\left\|\nabla f(\mathbf{x}_t)\right\|\right] &\leq \sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{i=1}^T\left\|\nabla f(\mathbf{x}_t)\right\|^2\right]} \\
&\leq \sqrt{\frac{2\Delta_f G}{\eta T} + \eta L d G + \frac{\sigma^2}{\beta T} + \sigma^2\beta + \frac{2L^2\eta^2 d}{\beta^2}}.
\end{aligned}$$

That is to say, by setting $\beta = \eta^{2/3}d^{1/3}$, $\eta = \mathcal{O}\left(\min\left\{\frac{1}{T^{1/2}d^{1/2}}, \frac{1}{T^{3/5}d^{1/5}}\right\}\right)$, we can obtain the convergence rate of $\mathcal{O}\left(\max\left\{\frac{d^{1/4}}{T^{1/4}}, \frac{d^{1/10}}{T^{1/5}}\right\}\right)$.