Sign operator for (L_0, L_1)-smooth optimization

Mark Ikonnikov
ikonnikov.mi@phystech.edu

Nikita Kornilov
kornilov.nm@phystech.edu

April 15, 2025

Abstract

In Machine Learning, the non-smoothness of optimization problems, the high cost of communicating gradients between workers, and severely corrupted data during training necessitate generalized optimization approaches. This paper explores the efficacy of sign-based methods [1], which address slow transmission by communicating only the sign of each minibatch stochastic gradient. We investigate these methods within (L_0, L_1)-smooth problems [3], which encompass a wider range of problems than the L-smoothness assumption. Furthermore, under the assumptions above, we investigate techniques to handle heavy-tailed noise [8], defined as noise with bounded κ-th moment $\kappa \in (1, 2]$. This includes the use of SignSGD with Majority Voting in the case of symmetric noise. We then attempt to extend the findings to convex cases using error feedback [5].

Keywords: Sign-based methods, (L_0, L_1)-smoothness, high-probability convergence, heavy-tailed noise.

Highlights below to be fixed later (these are our hopes for the paper)

Highlights:

1. Proves convergence of sign-based methods for (L_0, L_1)-smooth optimization
2. Handles heavy-tailed noise with high-probability convergence guarantees
3. Extends sign-based optimization to convex functions using error feedback

1 Introduction

The object of this research is the stochastic optimization of a smooth, non-convex function $f : \mathbb{R}^d \to \mathbb{R}$, defined as $f(x) = \mathbb{E}_{\xi \sim \mathcal{S}}[f(x, \xi)]$, where ξ is a random variable sampled from an unknown distribution $\mathcal{S}$. In machine learning, $f(x, \xi)$ typically represents a loss function

evaluated on a sample ξ, and the gradient oracle provides unbiased estimates $\nabla f(x, \xi)$. The most popular approach for solving (**??**) is Stochastic Gradient Descent:

$$x^{k+1} = x^k - \gamma_k \cdot g^k, \quad g^k := \nabla f(x^k, \xi^k).$$

For non-convex functions, the main goal of stochastic optimization is to find a point with small gradient norm. Traditional optimization often assumes L-smoothness (Lipschitz continuity of the gradient), but this is often restrictive for real-world deep learning models like Transformers. Instead, we adopt the (L_0, L_1)-smoothness condition [3], where:

$$\|\nabla f(x) - \nabla f(y)\| \leq \left(L_0 + L_1 \sup_{u \in [x,y]} \|\nabla f(u)\| \right) \|x - y\|,$$

allowing for a broader class of functions encountered in practice.

A key challenge is the communication bottleneck in distributed machine learning, where gradients are exchanged between workers and a parameter server. For large-scale neural networks, this process is computationally expensive. Sign-based methods, such as SignSGD [1], compress gradients by transmitting only their signs, reducing communication to one bit per parameter. However, their convergence under non-smooth conditions and heavy-tailed noise—where noise has a bounded κ-th moment for $\kappa \in (1, 2]$ [8]—remains underexplored. This noise, prevalent in modern datasets, can destabilize optimization, necessitating robust techniques.

Our methodology builds on a literature review of stochastic optimization, including Stochastic Gradient Descent (SGD), sign-based methods [1], and recent advances in heavy-tailed noise [8]. We also leverage error feedback mechanisms [5] to extend these methods to convex problems. The project tasks are:

1. Investigate sign-based methods for communication-efficient distributed optimization under the assumptions above.

2. Develop high-probability convergence guarantees accounting for generalized conditions.

3. Extend findings to convex optimization via error feedback technique.

The proposed solution is a sign-based optimization framework for (L_0, L_1)-smooth functions, robust to heavy-tailed noise. Its novelty lies in providing convergence guarantees under generalized smoothness conditions and handling heavy-tailed noise with high-probability bounds. Advantages include reduced communication costs, robustness to noise, and flexibility for non-smooth problems. Recent works like Bernstein et al. [1] offer communication savings but assume standard smoothness, while Gorbunov et al. [3] focus on (L_0, L_1)-smoothness without sign-based compression. Kornilov et al. [8] tackle heavy-tailed noise but not communication efficiency. Our work unifies these aspects.

The experimental goals are to validate convergence under (L_0, L_1)-smoothness and heavy-tailed noise. The setup includes synthetic datasets satisfying (L_0, L_1)-smoothness, real-world datasets with heavy-tailed noise, non-convex neural networks, and convex logistic

regression models. The workflow compares sign-based SGD (with/without error feedback) against traditional SGD and other compression methods, measuring convergence rates and communication costs.

The nearest alternative to our reaserch is Bernstein et al. [1]. Our advantage is the extension to (L_0, L_1)-smoothness and robustness to heavy-tailed noise, with the distinguished characteristic of high-probability convergence bounds. Thus, the paper proposes a sign-based optimization method for (L_0, L_1)-smooth non-convex problems, providing communication efficiency and robustness to heavy-tailed noise, distinguished by high-probability convergence guarantees.

Problem Statement

We consider the stochastic optimization problem of minimizing a smooth, non-convex function $f : \mathbb{R}^d \to \mathbb{R}$:

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{\xi \sim \mathcal{S}}[f(x, \xi)],$$

where ξ is a random variable sampled from an unknown distribution $\mathcal{S}$, and the gradient oracle provides an unbiased estimate $\nabla f(x, \xi) \in \mathbb{R}^d$. In machine learning, $f(x, \xi)$ represents the loss on a sample ξ, and the goal is to find a point x^* with a small gradient norm, i.e., $\|\nabla f(x^*)\| \leq \epsilon$, especially for non-convex objectives.

The samples ξ are drawn from $\mathcal{S}$, representing data points (e.g., images, text) in a machine learning task. The data originates from real-world or synthetic sources, with the statistical hypothesis that gradients $\nabla f(x, \xi)$ exhibit heavy-tailed noise, i.e., $\mathbb{E}_\xi[\|\nabla f(x, \xi) - \nabla f(x)\|_2^\kappa] \leq \sigma^\kappa$ for $\kappa \in (1, 2]$. The model is a parameterized function (e.g., neural network) with parameters $x \in \mathbb{R}^d$, within the class of (L_0, L_1)-smooth functions, satisfying symmetric (L_0, L_1)-smoothness, relaxing traditional L-smoothness. The objective $f(x)$ is the expected loss, with $f(x, \xi)$ as the sample-wise loss (e.g., cross-entropy). Convergence is measured by the gradient norm, with high-probability bounds (probability $\geq 1 - \delta$, $\delta \in (0, 1)$). Solutions are unconstrained in $\mathbb{R}^d$, but we seek robustness to noise and communication efficiency. In distributed settings, we prioritize low communication costs (bits transmitted per iteration).

2 Computational experiment

THIS PART WILL BE DELETED

The computational experiment aims to compare the performance of standard gradient descent (GD) and sign-based gradient descent (Sign-GD) in training a logistic regression model, highlighting the advantages of sign-based methods under L0 and L1 smoothness assumptions. Performance will be evaluated based on accuracy and iteration number, with minimal tuning to emphasize simplicity.

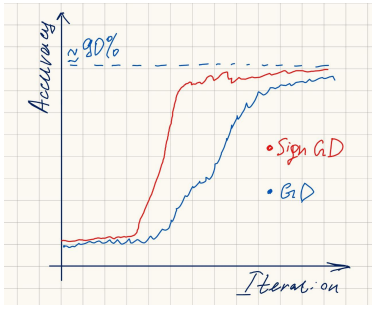

Figure 1: Performance of GD, SignGD

The experiment compares standard gradient descent (GD) and sign-based gradient descent (Sign-GD) on an open-source binary classification dataset "Mushrooms". Below is a mini-report based on expected outcomes. The preliminary plot (hand-drawn) shows Sign-GD achieving comparable or better performance with faster convergence, consistent with (L_0, L_1) smoothness assumptions.

Comments: The idea is based on the fact that logistic regression function $l(z, y) = \ln(1 + \exp(-yz)$ is both smooth and (L_0, L_1)-smooth, with $L = ||y||^2$ and $L_0 = 0, L_1 = ||y||$ which can be much smaller than L. Sign-GD slightly outperforms GD in accuracy and convergence time, suggesting that the sign-based update leverages smoothness assumptions effectively. The simplicity of the approach (no complex tuning) aligns with the minimal-effort goal. These results do not contradict the experiment's aim to showcase sign-based method advantages.

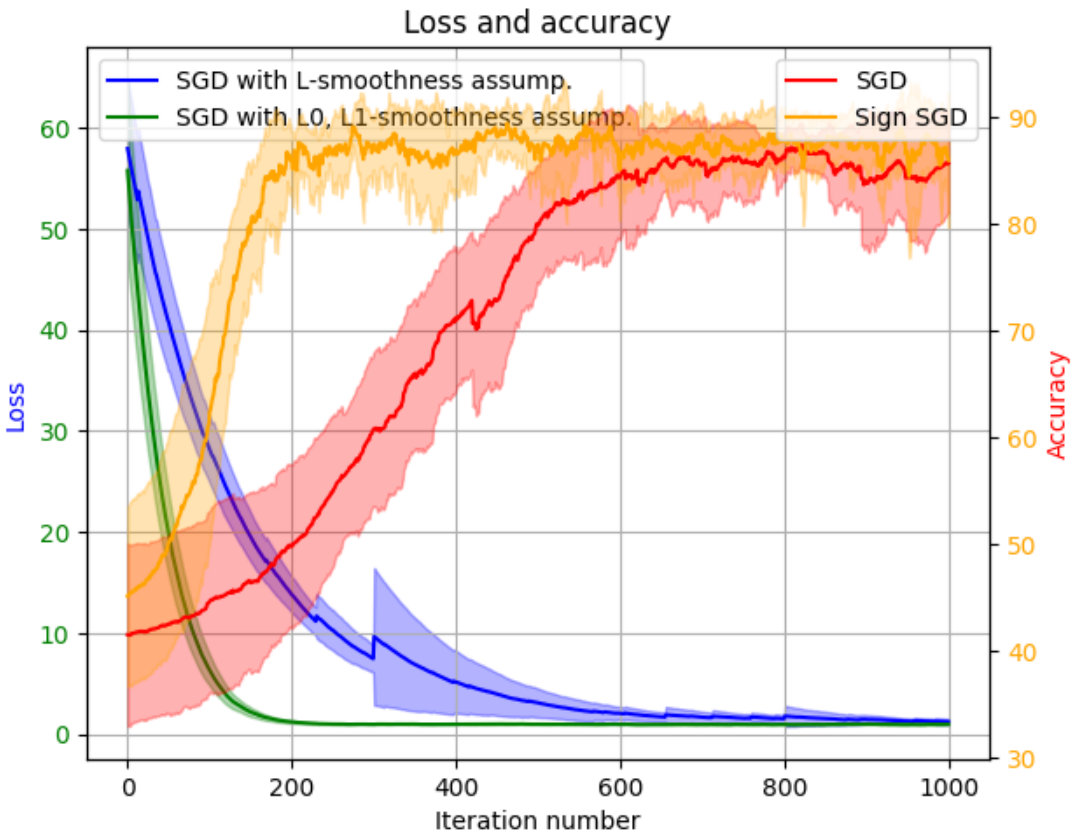

Figure 2: Logistic regression on Mushroom Dataset.
Performance of GD, SignGD with L_1-tuned step size

3 Theory

In this section, we present our novel convergence guarantees with high probability for existing sign-based methods for non-convex functions with heavy-tailed noise in gradient estimates. For each algorithm, we provide an explicit optimal tuning for the parameters. If function's smoothness constant and noise's characteristics are not given, we state the rates for arbitrary tuning. All proofs are located in Appendix **??**.

3.1 Assumptions

Assumption 1 (Lower bound)**.** *The objective function f is lower bounded by $f^* > -\infty$, i.e., $f(x) \geq f^*, \forall x \in \mathbb{R}^d$.*

Assumption 2 ((L_0, L_1)-Smoothness)**.** *The objective function f is differentiable and (L_0, L_1)-smooth, i.e., for the positive constants (L_0, L_1)*

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1 \|\nabla f(u)\|) \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 3 (Heavy-tailed noise in gradient estimates)**.** *The unbiased estimate* $\nabla f(x, \xi)$ *has bounded* κ*-th moment* $\kappa \in (1, 2]$ *for each coordinate, i.e.,* $\forall x \in \mathbb{R}^d$*:*

- $\mathbb{E}_\xi[\nabla f(x, \xi)] = \nabla f(x)$,
- $\mathbb{E}_\xi[|\nabla f(x, \xi)_i - \nabla f(x)_i|^\kappa] \leq \sigma_i^\kappa, i \in \overline{1, d}$,

where $\vec{\sigma} = [\sigma_1, \ldots, \sigma_d]$ *are non-negative constants. If* $\kappa = 2$*, then the noise is called a bounded variance.*

3.2 SignSGD and its HP convergence properties

We begin our analysis with the simplest of sign-based methods, namely SignSGD (Alg. 1) and prove a general lemma on its convergence with high probability.

Algorithm 1 SignSGD

Input: Starting point $x^1 \in \mathbb{R}^d$, number of iterations T, stepsizes $\{\gamma_k\}_{k=1}^T$.
0: **for** $k = 1, \ldots, T$ **do**
0: Sample ξ^k and compute estimate $g^k = \nabla f(x^k, \xi^k)$;
0: Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(g^k)$;
0: **end for**
Output: uniformly random point from $\{x^1, \ldots, x^T\}$. =0

In order to achieve accuracy ε, the noise $\|\vec{\sigma}\|_1$ have not to exceed ε. The first way to lower the noise is to use batching.

3.3 SignSGD with batching

Algorithm 2 minibatch-SignSGD

Input: Starting point $x^1 \in \mathbb{R}^d$, number of iterations T, stepsizes $\{\gamma_k\}_{k=1}^T$, batchsizes $\{B_k\}_{k=1}^T$.
1: **for** $k = 1, \ldots, T$ **do**
2: Sample $\{\xi_i^k\}_{i=1}^{B_k}$
3: Compute gradient estimate $g^k = \sum_{i=1}^{B_k} \nabla f(x^k, \xi_i^k)/B_k$;
4: Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(g^k)$;
5: **end for**
Output: uniformly random point from $\{x^1, \ldots, x^T\}$. =0

Algorithm 3 M-SignSGD

Input: Starting point $x^1 \in \mathbb{R}^d$, number of iterations K, stepsizes $\{\gamma_k\}_{k=1}^T$, momentums $\{\beta_k\}_{k=1}^T$.

1: **for** $k = 1, \ldots, T$ **do**
2: Sample ξ^k and compute estimate $g^k = \nabla f(x^k, \xi^k)$;
3: Compute $m^k = \beta_k m^{k-1} + (1-\beta_k) g^k$;
4: Set $x^{k+1} = x^k - \gamma_k \cdot \operatorname{sign}(m^k)$;
5: **end for**

Output: uniformly random point from $\{x^1, \ldots, x^T\}$. =0

3.4 SignSGD with momentum

Instead of variance reduction, one can use the momentum technique with the same sample complexity.

Theorem 1 (**Complexity for M-SignSGD in expectation**). *Consider lower-bounded L-smooth function f (As. 1, 2) and HT gradient estimates (As. 3). Then Alg. 3 requires T iterations to achieve $\frac{1}{T}\sum_{k=1}^T \mathbb{E}\left[\|\nabla f(x^k)\|_1\right] \leq \varepsilon$ for:*

arbitrary tuning: $T, \gamma_k \equiv \gamma_0 T^{-\frac{3}{4}}, \beta_k \equiv 1 - 1/\sqrt{T}$*:*

$$T = O\left(\frac{(\Delta_1/\gamma_0 + Ld\gamma_0)^4}{\varepsilon^4} + \left(\frac{\sqrt{d}\|\vec{\sigma}\|_\kappa}{\varepsilon}\right)^{\frac{2\kappa}{\kappa-1}}\right),$$

optimal tuning: $\gamma_k \equiv \sqrt{\frac{\Delta_1(1-\beta_k)}{4LdT}}, \beta_k \equiv 1 - \min\left\{1, \frac{1}{\|\vec{\sigma}\|_\kappa^2} \cdot \left(\frac{\Delta_1 L}{T}\right)^{\frac{\kappa}{3\kappa-2}}\right\}$:

$$T = O\left(\frac{\Delta_1 Ld}{\varepsilon^2} + \frac{\Delta_1 Ld}{\varepsilon^2}\left(\frac{\sqrt{d}\|\vec{\sigma}\|_\kappa}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right), \tag{1}$$

where $\Delta_1 = f(x^1) - f^*$.

4 (L_0, L_1) smoothness

Lemma 1. *(Symmetric (L_0, L_1)-smoothness) Function $f : \mathbb{R}^d \to \mathbb{R}$ is asymmetrically (L_0, L_1)-smooth, i.e., for all $x, y \in \mathbb{R}^d$, it holds*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq (L_0 + L_1\|\nabla f(y)\|_2)\exp(L_1\|x-y\|_2)\|x-y\|_2. \tag{2}$$

Moreover, it implies

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_0 + L_1\|\nabla f(x)\|_2}{2}\exp(L_1\|x-y\|_2)\|x-y\|_2^2. \tag{3}$$

To prove the HP bounds with the logarithmic dependency, we use the following measure concentration result (see, for example, [7, Lemma 1].

Lemma 2 (Measure Concentration Lemma)**.** *Let $\{D_k\}_{k=1}^T$ be a martingale difference sequence (MDS), i.e., $\mathbb{E}[D_k|D_{k-1},\dots,D_1]=0$ for all $k\in\overline{1,T}$. Furthermore, for each $k\in\overline{1,T}$, there exists positive $\sigma_k\in\mathbb{R}$, s.t. $\mathbb{E}\left[\exp\left(\frac{D_k^2}{\sigma_k^2}\right)|k\right]\leq e$. Then the following probability bound holds true:*

$$\forall\lambda>0,\delta\in(0,1):\quad \mathbb{P}\left(\sum_{k=1}^T D_k\leq\frac{3}{4}\lambda\sum_{k=1}^T\sigma_k^2+\frac{1}{\lambda}\log(1/\delta)\right)\geq 1-\delta. \tag{4}$$

To control error reduction during batching, we use the following batching lemma for HT variables. Its modern proof for $d=1$ was proposed in [2, Lemma 4.2] and then generalized for the multidimensional case in [6, 4].

Lemma 3 (HT Batching Lemma)**.** *Let $\kappa\in(1,2]$, and $X_1,\dots,X_B\in\mathbb{R}^d$ be a martingale difference sequence (MDS), i.e., $\mathbb{E}[X_i|X_{i-1},\dots,X_1]=0$ for all $i\in\overline{1,B}$. If all variables X_i have bounded $\kappa-$th moment, i.e., $\mathbb{E}[\|X_i\|_2^\kappa]<+\infty$, then the following bound holds true*

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^B X_i\right\|_2^\kappa\right]\leq\frac{2}{B^\kappa}\sum_{i=1}^B\mathbb{E}[\|X_i\|_2^\kappa]. \tag{5}$$

We need the following lemma about changes after one update step of sign-based methods from [9, Lemma 1].

Lemma 4 (Sign Update Step Lemma)**.** *Let $x,m\in\mathbb{R}^d$ be arbitrary vectors, $A=diag(a_1,\dots,a_d)$ be diagonal matrix and f be L-smooth function (As. 2). Then for the update step*

$$x'=x-\gamma\cdot A\cdot\operatorname{sign}(m)$$

with $\epsilon:=m-\nabla f(x)$, the following inequality holds true

$$f(x')-f(x)\leq-\gamma\|A\nabla f(x)\|_1+2\gamma\|A\|_F\|\epsilon\|_2+\frac{L_0+L_1\gamma\|A\nabla f(x^k)\|_2}{2}\exp\left(\gamma\|A\|_F\right)\gamma^2\|A\|_F^2. \tag{6}$$

4.1 Proof of (L_0, L_1) SignSGD General Convergence

Proof. Consider the k-th step of SignSGD. We use (L_0, L_1) smoothness of function f (Lemma 1) to estimate:

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &\leq \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L_0 + L_1 \|\nabla f(x^k)\|_2}{2} \exp(L_1 \|x^{k+1} - x^k\|_2) \|x^{k+1} - x^k\|_2^2 \\
&= -\gamma_k \frac{\langle \nabla f(x^k), \operatorname{sign}(g^k) \rangle}{\|\nabla f(x^k)\|_1} \cdot \|\nabla f(x^k)\|_1 + \frac{L_0 d \gamma_k^2}{2} \exp(L_1 \sqrt{d} \gamma_k) \\
&+ \frac{L_1 d \gamma_k \exp(L_1 \sqrt{d} \gamma_k)}{2} \cdot \gamma_k \|\nabla f(x^k)\|_2 \\
&= -\gamma_k \frac{\langle \nabla f(x^k), \operatorname{sign}(g^k) \rangle}{\|\nabla f(x^k)\|_1} \cdot \|\nabla f(x^k)\|_1 + \frac{L_0 d \gamma_k^2}{2} \exp(L_1 \sqrt{d} \gamma_k) \\
&+ \frac{L_1 \sqrt{d} \gamma_k \exp(L_1 \sqrt{d} \gamma_k)}{2} \cdot \gamma_k \|\nabla f(x^k)\|_1.
\end{aligned}$$

Consequently, after summing all T steps, we obtain the next equation. We introduce the following terms $\phi_k := \frac{\langle \nabla f(x^k), \operatorname{sign}(g^k) \rangle}{\|\nabla f(x^k)\|_1} \in [-1, 1]$, $\psi_k := \mathbb{E}[\phi_k | x^k]$ and $D_k := -\gamma_k (\phi_k - \psi_k) \|\nabla f(x^k)\|_1$. We note that D_k is a martingale difference sequence ($\mathbb{E}[D_k | D_{k-1}, \ldots, D_k] = 0$) and satisfies

$$\exp\left(\frac{D_k^2}{4\gamma_k^2 \|\nabla f(x^k)\|_1^2}\right) = \exp\left(\frac{(\phi_k - \psi_k)^2}{4}\right) \leq e.$$

Applying Measure Concentration Lemma to MSD we derive the bound for all $\lambda > 0$ with probability at least $1 - \delta$. We then use norm relation and (L_0, L_1)-smoothness to estimate maximum gradient norm for all $k \in \overline{2, T+1}$:

$$\begin{aligned}
\|\nabla f(x^k)\|_1 / \sqrt{d} &\leq \|\nabla f(x^k)\|_2 \leq \|\nabla f(x^k) - \nabla f(x^{k-1}) + \nabla f(x^{k-1})\|_2 \\
&\leq \|\nabla f(x^k) - \nabla f(x^{k-1})\|_2 + \|\nabla f(x^{k-1})\|_2 \\
&\leq (L_0 + L_1 \|\nabla f(x^{k-1})\|_2) \exp(L_1 \|x^k - x^{k-1}\|_2) \|x^k - x^{k-1}\|_2 + \|\nabla f(x^{k-1})\|_2 \\
&\leq (L_0 + L_1 \|\nabla f(x^{k-1})\|_2) \exp(L_1 \sqrt{d} \gamma_k) \sqrt{d} \gamma_k + \|\nabla f(x^{k-1})\|_2.
\end{aligned}$$

At this point, we take $\gamma_k \leq \frac{1}{48 L_1 d \log \frac{1}{\delta} \sqrt{d}}$ to obtain

$$\begin{aligned}
\|\nabla f(x^k)\|_1 / \sqrt{d} &\leq 2 L_0 \sqrt{d} \gamma_k + \frac{\|\nabla f(x^{k-1})\|_2}{48 d \log \frac{1}{\delta}} + \|\nabla f(x^{k-1})\|_2 \\
&\leq 2 L_0 \sqrt{d} \sum_{\tau=1}^{k-1} \gamma_\tau + \sum_{\tau=1}^{k-1} \frac{\|\nabla f(x^\tau)\|_2}{48 d \log \frac{1}{\delta}} + \|\nabla f(x^1)\|_2 \\
&\leq 2 L_0 \sqrt{d} \sum_{\tau=1}^{k-1} \gamma_\tau + \sum_{\tau=1}^{k-1} \frac{\|\nabla f(x^\tau)\|_1}{48 d \log \frac{1}{\delta}} + \|\nabla f(x^1)\|_1.
\end{aligned}$$

Hence, the choice $\lambda := \frac{1}{6d(\gamma^{max}\|\nabla f(x^1)\|_1+\sum_{k=1}^T \frac{\gamma_k\|\nabla f(x^k)\|_1}{48d\log\frac{1}{\delta}}+2C_TL_0)}$ where $C_T := \max\limits_{k\in\overline{1,T}} \gamma_k \cdot \sum\limits_{\tau=1}^{k-1}\gamma_\tau$ and $\gamma^{max} := \max\limits_{k\in\overline{1,T}}\gamma_k$ yields with probability at least $1-\delta$:

$$\begin{aligned}\sum_{k=1}^T \gamma_k\left(\psi_k - \frac{1}{2} - \frac{1}{4}\right)\|\nabla f(x^k)\|_1 &\leq \Delta_1 + L_0 d\sum_{k=1}^T\gamma_k^2 + 6\sqrt{d}(\gamma^{max}\|\nabla f(x^1)\|_1 + 2C_TL_0)\log(1/\delta) \\ &+ \frac{6}{48}\sum_{k=1}^T\gamma_k\|\nabla f(x^k)\|_1,\end{aligned} \tag{7}$$

$$\sum_{k=1}^T \gamma_k\left(\psi_k - \frac{1}{2} - \frac{1}{4} - \frac{1}{8}\right)\|\nabla f(x^k)\|_1 \leq \Delta_1 + L_0 d\sum_{k=1}^T\gamma_k^2 + 6\sqrt{d}(\gamma^{max}\|\nabla f(x^1)\|_1 + 2C_TL_0)\log(1/\delta),$$

Next, we estimate each term $\psi_k\|\nabla f(x^k)\|_1$ in the previous sum:

$$\begin{aligned}\psi_k\|\nabla f(x^k)\|_1 &= \mathbb{E}\left[\langle\nabla f(x^k), \operatorname{sign}(g^k)\rangle|x^k\right] \\ &= \|\nabla f(x^k)\|_1 - \sum_{i=1}^d 2|\nabla f(x^k)|_i\cdot\mathbb{P}(\operatorname{sign}(\nabla f(x^k))_i \neq \operatorname{sign}(g^k)_i|x^k).\end{aligned} \tag{8}$$

For each coordinate, we have a bound derived from Markov's inequality (**??**) followed by Jensen's inequality (**??**):

$$\begin{aligned}\mathbb{P}(\operatorname{sign}(\nabla f(x^k))_i \neq \operatorname{sign}(g^k)_i|x^k) &\leq \mathbb{P}(|\nabla f(x^k)_i - g_i^k| \geq |\nabla f(x^k)_i||x^k) \leq \frac{\mathbb{E}_{\xi^k}[|\nabla f(x^k)_i - g_i^k|]}{|\nabla f(x^k)_i|} \\ &\leq \frac{(\mathbb{E}_{\xi^k}[|\nabla f(x^k)_i - g_i^k|^\kappa])^{\frac{1}{\kappa}}}{|\nabla f(x^k)_i|} \leq \frac{\sigma_i}{|\nabla f(x^k)_i|}.\end{aligned} \tag{9}$$

Hence, the whole sum can be bounded as

$$\sum_{i=1}^d 2|\nabla f(x^k)|_i\cdot\mathbb{P}(\operatorname{sign}(\nabla f(x^k))_i \neq \operatorname{sign}(g^k)_i|x^k) \leq 2\|\vec{\sigma}\|_1.$$

Finally, we put this bound in (8) and obtain:

$$\begin{aligned}\frac{1}{16}\sum_{k=1}^T\gamma_k\|\nabla f(x^k)\|_1 &\leq \Delta_1 + L_0 d\sum_{k=1}^T\gamma_k^2 + 2\sum_{k=1}^T\gamma_k\|\vec{\sigma}\|_1 \\ &+ 6d(\gamma^{max}\|\nabla f(x^1)\|_1 + 2C_TL_0)\log(1/\delta).\end{aligned} \tag{10}$$

Plugging in constant stepsizes $\gamma_k \equiv \gamma \leq \frac{1}{48L_1 d\log\frac{1}{\delta}\sqrt{d}}$ implies $C_T = T\gamma^2, \gamma^{max} = \gamma$ and the required bound:

$$\frac{1}{T}\sum_{k=1}^T\|\nabla f(x^k)\|_1 \leq \frac{4\Delta_1}{T\gamma} + 80L_0 d\gamma\log(1/\delta) + 8\|\vec{\sigma}\|_1 + 24\frac{d\|\nabla f(x^1)\|_1}{T}\log(1/\delta).$$

Optimal tuning. In case $\varepsilon \geq \frac{8L_0}{L_1\sqrt{d}}$, we use stepsize $\gamma = \frac{1}{48L_1 d \log\frac{1}{\delta}\sqrt{d}} \Rightarrow 80L_0 d\gamma \log(1/\delta) \leq \varepsilon/2$ and batchsize $8\frac{\|\vec{\sigma}\|_1}{B^{\frac{\kappa-1}{\kappa}}} \leq \varepsilon/2 \Rightarrow B_k \equiv \max\left\{1, \left(\frac{16\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right\}$. The number of iterations T is chosen to bound the first term: $T = O\left(\frac{\Delta_1 L_1 \log\frac{1}{\delta} d^{\frac{3}{2}}}{\varepsilon}\right)$. The total number of oracle calls is:

$$\begin{aligned}
\varepsilon \;\geq\; \frac{8L_0}{L_1\sqrt{d}} \quad &\Rightarrow \quad N = O\left(\frac{\Delta_1 L_1 \log(1/\delta) d^{\frac{3}{2}}}{\varepsilon}\left[1 + \left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right), \\
\varepsilon \;<\; \frac{8L_0}{L_1\sqrt{d}} \quad &\Rightarrow \quad N = O\left(\frac{\Delta_1 L_0 \log(1/\delta) d}{\varepsilon^2}\left[1 + \left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right). \qquad (11)
\end{aligned}$$

□

4.2 Proof of Sign Update Step Lemma

Proof. The proof is identical to the proof from [9, Lemma 1] with changing the smoothness assumprion according to Lemma 1. □

4.3 Proof of (L_0, L_1) M-SignSGD General Convergence

Proof. Consider the k-th step of SignSGD. We use (L_0, L_1) step update (Lemma 4) to estimate:

$$\begin{aligned}
f(x^{k+1}) - f(x^k) \;&\leq\; \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L_0 + L_1\|\nabla f(x^k)\|_2}{2}\exp(L_1\|x^{k+1} - x^k\|_2)\|x^{k+1} - x^k\|_2^2 \\
&=\; -\gamma_k\langle \nabla f(x^k), \operatorname{sign}(m^k)\rangle + \frac{L_0 d\gamma_k^2}{2}\exp(L_1\sqrt{d}\gamma_k) \\
&+\; \frac{L_1 d\gamma_k \exp(L_1\sqrt{d}\gamma_k)}{2}\cdot\gamma_k\|\nabla f(x^k)\|_2 \\
&\leq\; -\gamma_k\langle \nabla f(x^k), \operatorname{sign}(m^k)\rangle + \frac{L_0 d\gamma_k^2}{2}\exp(L_1\sqrt{d}\gamma_k) \\
&+\; \frac{L_1 d\gamma_k \exp(L_1\sqrt{d}\gamma_k)}{2}\cdot\gamma_k\|\nabla f(x^k)\|_1. \\
&\leq\; \gamma_k\langle \nabla f(x^k), \operatorname{sign}(\nabla f(x^k)) - \operatorname{sign}(m^k)\rangle - \gamma_k\|\nabla f(x^k)\|_1 + \frac{L_0 d\gamma_k^2}{2}\exp(L_1\sqrt{d}\gamma_k) \\
&+\; \frac{L_1 d\gamma_k \exp(L_1\sqrt{d}\gamma_k)}{2}\cdot\gamma_k\|\nabla f(x^k)\|_1 \\
&\leq\; -\gamma_k\|\nabla f(x^k)\|_1 + 2\gamma_k\sqrt{d}\|\epsilon^k\|_2 + \frac{L_0 d\gamma_k^2}{2}\exp(L_1\sqrt{d}\gamma_k) \\
&+\; \frac{L_1 d\gamma_k \exp(L_1\sqrt{d}\gamma_k)}{2}\cdot\gamma_k\|\nabla f(x^k)\|_1.
\end{aligned}$$

The last inequality holds due to Lemma 4.

Since we set constant steps sizes and momentum, we denote them as $\gamma \equiv \gamma_k$ and $\beta \equiv \beta_k$, respectively. We use notations $\epsilon^k := m^k - \nabla f(x^k)$ and $\theta^k := g^k - \nabla f(x^k)$. Therefore, we have at k-th step values:

$$\begin{aligned} m^k &= \beta m^{k-1} + (1-\beta)g^k = \gamma(\epsilon^{k-1} + \nabla f(x^{k-1})) + (1-\gamma)(\theta^k + \nabla f(x^k)), \\ \epsilon^k &= m^k - \nabla f(x^k) = \beta\epsilon^{k-1} + \beta(\underbrace{\nabla f(x^{k-1}) - \nabla f(x^k)}_{=:s^k}) + (1-\beta)\theta^k, \\ \epsilon^k &= m^k - \nabla f(x^k) = \beta\epsilon^{k-1} + \beta s^k + (1-\beta)\theta^k. \end{aligned}$$

Unrolling the recursion, we obtain an explicit formula (upper index of β is its power):

$$\epsilon^k = \beta^{k-1}\epsilon^1 + \sum_{i=2}^{k} \beta^{k-i+1} s^i + (1-\beta)\sum_{i=2}^{k} \beta^{k-i}\theta^i. \tag{12}$$

From $(L_0, L_1)-$smoothness of f (Lem. 1) follows the bound:

$$\|s^k\|_2 \leq (L_0 + L_1\|\nabla f(x^k)\|_2)\exp(L_1\|x^k - x^{x+1}\|_2)\|x^k - x^{k+1}\|_2 = (L_0 + L_1\|\nabla f(x^k)\|_2)\exp(L_1\gamma_k\sqrt{d})\gamma_k\sqrt{d}$$

Denote $\lambda := \exp(L_1\gamma_k\sqrt{d})\gamma_k\sqrt{d}$.

Hence, the norm of (12) can be bounded as:

$$\|\epsilon^k\|_2 \leq \beta^{k-1}\|\epsilon^1\|_2 + L_0\lambda\sum_{i=2}^{k}\beta^{k-i+1} + L_1\lambda\sum_{i=2}^{k}\beta^{k-i+1}\|\nabla f(x^k)\|_2 + (1-\beta)\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2.$$

We notice that variables $\{\theta_i\}$ are martingale difference sequence from Lemma 3 which we plan to use. Due to the formal definition of $\theta^i = g^i - \nabla f(x^i) = \nabla f(x^i, \xi_i) - \nabla f(x^i)$ and M-SinSGD step, the conditioning on $\theta^{i-1}, \ldots, \theta^1$ with randomness $\xi_1, \ldots, \xi_{i-1}$ is equivalent to the conditioning on point s $x^i, \ldots, x^2$. Hence, we show by definition of martingale difference sequence that

$$\mathbb{E}[\theta^i|\theta^{i-1}, \ldots, \theta^1] = \mathbb{E}[\theta^i|x^i, \ldots, x^2] = \mathbb{E}[\nabla f(x^i, \xi_i) - \nabla f(x^i)|x^i, \ldots, x^2] = 0.$$

To take math expectation from both sides, we first take it from the term

$$\mathbb{E}\left[\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2\right] \leq \left(\mathbb{E}\left[\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2^\kappa\right]\right)^{\frac{1}{\kappa}} \overset{\text{Lem. 3}}{\leq} \left(\sum_{i=2}^{k} 2\mathbb{E}\left[\|\beta^{(k-i)}\theta^i\|_2^\kappa\right]\right)^{\frac{1}{\kappa}} \leq \left(\sum_{i=2}^{k} 2\beta^{\kappa(k-i)}\mathbb{E}\left[\|\theta^i\|_2^\kappa\right]\right)^{\frac{1}{\kappa}}.$$

For each $i \in \overline{2, T}$, we estimate $\mathbb{E}\left[\|\theta^i\|_2^\kappa\right]$ as

$$\mathbb{E}\left[\|\theta^i\|_2^\kappa\right] \overset{(??)}{\leq} \mathbb{E}\left[\|\theta^i\|_\kappa^\kappa\right] = \mathbb{E}\left[\sum_{j=1}^{d}|g_j^k - \nabla f(x^k)_j|^\kappa\right] \overset{As.3}{\leq} \sum_{j=1}^{d}\sigma_j^\kappa = \|\vec{\sigma}\|_\kappa^\kappa. \tag{13}$$

We continue bounding (13) with

$$(13) \leq \left(\sum_{i=2}^{k} 2\beta^{\kappa(k-i)}\|\vec{\sigma}\|_\kappa^\kappa\right)^{\frac{1}{\kappa}} \leq \frac{2\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}}.$$

Therefore, the final math expectation can be calculated as:

$$\mathbb{E}\|\epsilon^k\|_2 \leq \beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 + \frac{L\sqrt{d}\gamma}{1-\beta} + \frac{2(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}}. \tag{14}$$

Now, we can update the step using Lemma 4:

(15)

Then take math expectation:

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &\leq -\gamma\|\nabla f(x^k)\|_1 + 2\gamma\sqrt{d}\|\epsilon\|_2 + \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma_k}{2}\cdot\lambda\|\nabla f(x^k)\|_1 \\
&\leq -\gamma\|\nabla f(x^k)\|_1 + 2\gamma\sqrt{d}\cdot \\
&+ \left[\beta^{k-1}\|\epsilon^1\|_2 + L_0\lambda\sum_{i=2}^{k}\beta^{k-i+1} + L_1\lambda\sum_{i=2}^{k}\beta^{k-i+1}\|\nabla f(x^i)\| + (1-\beta)\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2\right] \\
&+ \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\cdot\lambda\|\nabla f(x^k)\|_1 \\
&\leq -\gamma\|\nabla f(x^k)\|_1 \\
&+ 2\gamma\sqrt{d}\left[\beta^{k-1}\|\epsilon^1\|_2 + L_0\lambda\frac{1}{1-\beta} + L_1\lambda\sum_{i=2}^{k}\beta^{k-i+1}\|\nabla f(x^i)\| + (1-\beta)\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2\right] \\
&+ \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\cdot\lambda\|\nabla f(x^k)\|_1
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[f(x^{k+1})] - \mathbb{E}[f(x^k)] &\leq -\gamma\mathbb{E}[\|\nabla f(x^k)\|_1] + 2\gamma\sqrt{d}\beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 \\
&+ L_0\lambda\frac{2\gamma\sqrt{d}}{1-\beta} + L_1\lambda 2\gamma\sqrt{d}\sum_{i=2}^{k}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^k)\|_1 + \frac{4\gamma\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}} \\
&+ \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\lambda\mathbb{E}\|\nabla f(x^k)\|_1
\end{aligned} \tag{16}$$

Summing it over k, we derive

$$
\begin{aligned}
f^* - f(x^1) &\leq -\gamma \sum_{k=1}^{T} \mathbb{E}\|\nabla f(x^k)\|_1 + 2\gamma T\sqrt{d}\beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 \qquad (17)\\
&+ L_1\lambda 2\gamma\sqrt{d}\sum_{k=1}^{T}\sum_{i=2}^{k}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^i)\|_1 + \frac{4\gamma T\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}}\\
&+ \frac{L_0 T\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\lambda\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1
\end{aligned}
$$

Changing the order of summation in the right part of the (18) we obtain:

$$
\begin{aligned}
2\gamma L_1\lambda\sqrt{d}\sum_{k=1}^{T}\left(\sum_{i=2}^{k}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^i)\|_1\right) &=\\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\left(\sum_{k=i}^{T}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^i)\|_1\right) &=\\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\beta^{-i}\left(\sum_{k=i}^{T}\beta^{k+1}\right)\mathbb{E}\|\nabla f(x^i)\|_1 &=\\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\beta^{-i+1}\beta^{i}\left(\frac{1-\beta^{T-i}}{1-\beta}\right)\mathbb{E}\|\nabla f(x^i)\|_1 &\leq\\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\beta\left(\frac{1}{1-\beta}\right)\mathbb{E}\|\nabla f(x^i)\|_1
\end{aligned}
$$

Finally

$$
\begin{aligned}
f^* - f(x^1) & \qquad (18)\\
&\leq -\gamma\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 + 2\gamma T\sqrt{d}\beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 \qquad (19)\\
&+ 2\gamma L_1\lambda\sqrt{d}\cdot\frac{\beta}{1-\beta}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 + \frac{4\gamma T\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{1/\kappa}}\\
&+ \frac{L_0 T\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\lambda\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1\\
&\leq \left(-\gamma + \frac{2\gamma L_1\lambda\sqrt{d}\beta}{1-\beta} + \frac{L_1\sqrt{d}\gamma}{2}\lambda\right)\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 \qquad (20)\\
&+ 2\gamma T\sqrt{d}\beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 + \frac{4\gamma T\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{1/\kappa}} + \frac{L_0 T\sqrt{d}\gamma}{2}\lambda.
\end{aligned}
$$

1) For arbitrary tuning 2) For optimal tuning □

References

[1] Jeremy Bernstein et al. "signSGD: Compressed Optimisation for Non-Convex Problems". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 560–569. URL: https://proceedings.mlr.press/v80/bernstein18a.html.

[2] Yeshwanth Cherapanamjeri et al. "Optimal mean estimation without a variance". In: *Conference on Learning Theory*. PMLR. 2022, pp. 356–357.

[3] Eduard Gorbunov et al. *Methods for Convex (L_0, L_1)-Smooth Optimization: Clipping, Acceleration, and Adaptivity*. 2024. arXiv: 2409.14989 [math.OC]. URL: https://arxiv.org/abs/2409.14989.

[4] Florian Hübler, Ilyas Fatkhullin, and Niao He. "From Gradient Clipping to Normalization for Heavy Tailed SGD". In: *arXiv preprint arXiv:2410.13849* (2024).

[5] Sai Praneeth Karimireddy et al. "Error Feedback Fixes SignSGD and other Gradient Compression Schemes". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 3252–3261. URL: https://proceedings.mlr.press/v97/karimireddy19a.html.

[6] Nikita Kornilov et al. "Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance". In: *Advances in Neural Information Processing Systems* 36 (2024).

[7] Xiaoyu Li and Francesco Orabona. "A high probability analysis of adaptive SGD with momentum". In: *arXiv preprint arXiv:2007.14294* (2020).

[8] Kornilov Nikita et al. *Sign Operator for Coping with Heavy-Tailed Noise: High Probability Convergence Bounds with Extensions to Distributed Optimization and Comparison Oracle*. 2025. DOI: 10.48550/ARXIV.2502.07923.

[9] Tao Sun et al. "Momentum ensures convergence of signsgd under weaker assumptions". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 33077–33099.