Sign Operator for (L_0, L_1)-Smooth Optimization with Heavy-Tailed Noise

Ikonnikov Mark

Moscow Institute of physics and Technology

Course: My first scientific paper (Strijov's practice)/Group 206

Expert: Alexander Beznosikov

Consultant: Nikita Kornilov

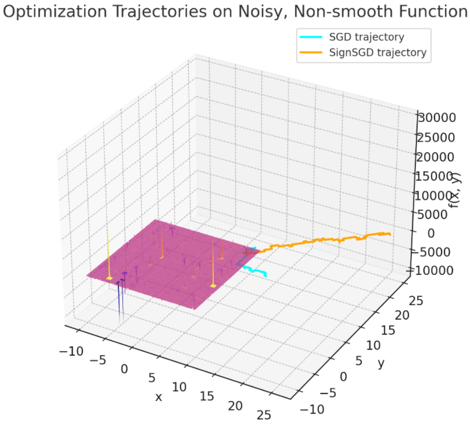
April 17, 2025

Goal of Research

Objectives

- Define (L_0, L_1)-smoothness.
- Develop sign-based methods (Sign-SGD, minibatch-SignSGD, momentum-SignSGD) for heavy-tailed (HT) noise.
- Establish theoretical convergence bounds under (L_0, L_1)-smoothness and HT noise.
- Validate results through computational experiments.

Idea



Convergence rates improve significantly with Sign-methods.

$\|\nabla^2 f(x)\|_2 \leq L_0 + L_1\|\nabla f(x)\|$

$\mathbb{E}_\xi[|\nabla f(x,\xi)_i - \nabla f(x)_i|^\kappa] \leq \sigma_i^\kappa,\ \kappa \in (1,2]$

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

Literature

Title	Year	Authors	Paper
Sign Operator for Coping with Heavy-Tailed Noise	2025	Kornilov et al.	arXiv
signSGD: Compressed Optimisation for Non-Convex Problems	2018	J. Bernstein et al.	PMLR
Methods for Convex (L0,L1)-Smooth Optimization	2024	Gorbunov et al.	arXiv
Robustness to Unbounded Smoothness of Generalized SignSGD	2022	M. Crawshaw et al.	NeurIPS

Hypothesis and Model

Hypothesis

Sign-based optimization methods outperform traditional gradient-based methods in (L_0, L_1)-smooth problems with heavy-tailed noise, achieving faster convergence and robustness.

Model

$f : \mathbb{R}^d \to \mathbb{R}$:

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

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ that is (L_0, L_1)-smooth:

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1 \|\nabla f(u)\|) \, \|x - y\|,$$

with gradient estimates $\nabla f(x, \xi)$ under HT noise:

- $\mathbb{E}_\xi[\nabla f(x, \xi)] = \nabla f(x)$,
- $\mathbb{E}_\xi[|\nabla f(x, \xi)_i - \nabla f(x)_i|^\kappa] \leq \sigma_i^\kappa$, $\kappa \in (1, 2]$.

Examples of (L_0, L_1)-Smooth Functions

The following functions illustrate (L_0, L_1)-smoothness:

- Let $f(x) = \|x\|^{2n}$, where n is a positive integer. Then, $f(x)$ is convex and $(2n, 2n-1)$-smooth. Moreover, $f(x)$ is not L-smooth for $n \geq 2$ and any $L \geq 0$.
- $f(x) = \log\left(1 + \exp(-a^\top x)\right)$, where $a \in^d$ is some vector. It is known that this function is L-smooth and convex with $L = \|a\|^2$. However, one can show that f is also (L_0, L_1)-smooth with $L_0 = 0$ and $L_1 = \|a\|$. For $\|a\| \gg 1$, both L_0 and L_1 are much smaller than L.

These are relevant to compressed sensing and machine learning.

Sign-SGD Algorithm

Algorithm 1 SignSGD

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

1: **for** $k = 1, \ldots, T$ **do**

2: Sample ξ^k and compute estimate $g^k = \nabla f(x^k, \xi^k)$;

3: Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(g^k)$;

4: **end for**

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

Sign-SGD-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 $\{\xi_i^k\}_{i=1}^{B_k}$ and compute gradient estimate $g^k = \sum_{i=1}^{B_k} \nabla f(x^k, \xi_i^k)/B_k$;

3: Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(g^k)$;

4: **end for**

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

Sign-SGD Momentum Algorithm

Algorithm 4 M-SignSGD

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

for $k = 1, \ldots, T$ **do**

  Sample ξ^k and compute estimate $g^k = \nabla f(x^k, \xi^k)$;

  Compute $m^k = \beta_k m^{k-1} + (1 - \beta_k) g^k$;

  Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(m^k)$;

end for

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

Lemma

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

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq (L_0 + L_1 \|\nabla f(y)\|_2) \exp(L_1 \|x - y\|_2) \|x - y\|_2. \tag{1}$$

Moreover, it implies

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

Lemma

Lemma (HT Batching Lemma)

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

$$\left[\left\|\frac{1}{B}\sum_{i=1}^{B} X_i\right\|_2^\kappa\right] \leq \frac{2}{B^\kappa}\sum_{i=1}^{B}[\|X_i\|_2^\kappa]. \tag{3}$$

Novel Lemma

Lemma (Sign Update Step Lemma (Ikonnikov))

*Let $x, m \in^d$ be arbitrary vectors, $A = diag(a_1, \ldots, a_d)$ be diagonal matrix and f be L-smooth function (As. **??**). Then for the update step*

$$x' = x - \gamma \cdot A \cdot (m)$$

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

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

$$\cdot \exp(\gamma L_1 \|A\|_F) \gamma^2 \|A\|_F^2.$$

Solution: Theoretical Part

Theorem (Complexity for minibatch-L0L1-SignSGD)

Consider lower-bounded $(L0, L1)$-smooth function f and HT gradient estimates. Then Alg. minibatch-SignSGD requires the sample complexity N to achieve $\frac{1}{T}\sum_{k=1}^{T}\|\nabla f(x^k)\|_1 \leq \varepsilon$ with probability at least $1-\delta$ for:

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 $B_k \equiv \max\left\{1, \left(\frac{16\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right\}$. $T = O\left(\frac{\Delta_1 L_1 \log\frac{1}{\delta} d^{\frac{3}{2}}}{\varepsilon}\right)$. The total number of oracle calls is:*

$$\varepsilon \geq \frac{8L_0}{L_1\sqrt{d}} \Rightarrow N = O\left(\frac{\Delta_1 L_1 \log(1\delta) d^{\frac{3}{2}}}{\varepsilon}\left[1+\left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right),$$

$$\varepsilon < \frac{8L_0}{L_1\sqrt{d}} \Rightarrow N = O\left(\frac{\Delta_1 L_0 \log(1\delta) d}{\varepsilon^2}\left[1+\left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right).$$

Solution: Theoretical Part

Theorem (Complexity for M-L0L1-SignSGD)

Consider lower-bounded (L_0, L_1)-smooth function f, and HT gradient estimates. Then SignSGD-M requires T iterations to achieve $\frac{1}{T}\sum_{k=1}^{T}\left[\|\nabla f(x^k)\|_1\right] \leq \varepsilon$ for:

Case $\varepsilon \geq \frac{3L_0}{cL_1}$**:** $\beta_k \equiv 1 - \min\left\{1, \left(\frac{c\Delta_1 L_1\sqrt{d}}{T\|\vec{\sigma}\|_\kappa}\right)^{\frac{\kappa}{2\kappa-1}}\right\}, \gamma_k \equiv \frac{1-\beta}{8c}\frac{1}{L_1 d}$

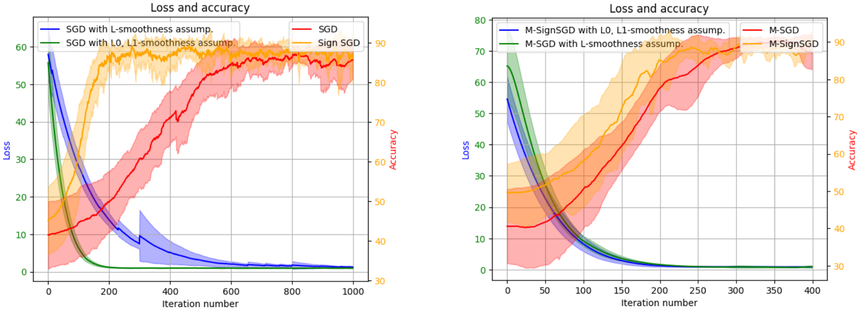
$$T = O\left(\frac{\Delta_1 L_1 d}{\varepsilon}\left(1 + \left(\frac{\sqrt{d}\|\vec{\sigma}\|_\kappa}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right)\right), \qquad (5)$$

Case $\varepsilon < \frac{3L_0}{L_1}$**:**

$1 - \beta_k \equiv 1 - \min\left\{1, \left(\frac{\Delta_1 L_0}{T\|\vec{\sigma}\|_\kappa^2}\right)^{\frac{\kappa}{3\kappa-2}}\right\}, \gamma_k \equiv \sqrt{\frac{\Delta_1(1-\beta_k)}{TL_0 d}}$

$$T = O\left(\frac{\Delta_1 L_1 d}{\varepsilon^2}\left(1 + \left(\frac{\sqrt{d}\|\vec{\sigma}\|_\kappa}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right)\right), \qquad (6)$$

Computational Experiment: Goals and Statistics



Convergence and accuracy rates improve significantly with Sign-methods.

Error Analysis

Error comparison

Method	Mean Loss	Mean Acc.	Loss Var.	Acc. Var.
M-SignSGD	3.63	82.86	73.56	135.77
M-SGD	7.72	73.46	209.46	341.58
SignSGD	6.71	79.12	155.10	140.47
SGD	16.44	62.96	234.20	70.55

Table: comparison of convergence of several methods under the assumptions

Results and Conclusions

Results

- Sign-based methods outperform SGD in convergence under (L_0, L_1)-smoothness and HT noise.
- Novel lemma is proven.
- Momentum-SignSGD and minibatch-SignSGD convergence are bounded and proved.

Conclusions

- (L_0, L_1)-smoothness enables better rates under (L_0, L_1) and HT-noise assumptions.
- Sign-based methods are noise-robust and communication-efficient.