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

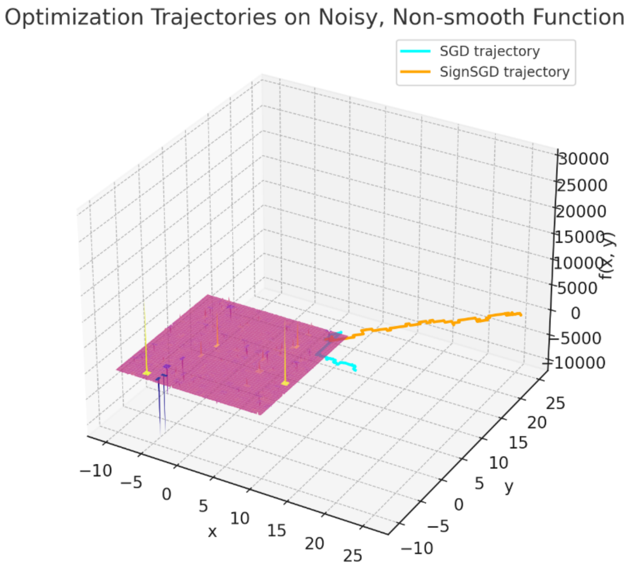
May 22, 2025

Research Objective: Robust Optimization with Sign-Based Methods

Objectives

1. Define (L_0, L_1)-smoothness.
2. Develop Sign-based methods for heavy-tailed (HT) noise.
3. Prove convergence bounds under (L_0, L_1)-smoothness and HT noise.
4. Validate theory via computational experiments.

Sign-Based Optimization Tackles Hard Settings



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

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

Sign methods show faster convergence with HT

Background and 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	Bernstein et al.	PMLR
Methods for Convex (L0,L1)-Smooth Optimization	2024	Gorbunov et al.	arXiv
Robustness to Unbounded Smoothness of Generalized SignSGD	2022	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.

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. Then for the update step

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

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

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

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

Our findings for minibatch-SignSGD

Smoothness	**Step size** γ
$(L_0, L_1),\ \varepsilon \geq \frac{L_0}{L_1\sqrt{d}}$	$\gamma = \Theta\left(\frac{1}{L_1 d\sqrt{d}}\right)$
$(L_0, L_1),\ \varepsilon < \frac{L_0}{L_1\sqrt{d}}$	same
L-smooth	$\gamma = \Theta\left(\frac{1}{L\sqrt{d}}\right)$

Iterations T
$T = \widetilde{O}\left(\frac{L_1 d^{3/2}}{\varepsilon}\right)$
$T = \widetilde{O}\left(\frac{L_0 d}{\varepsilon^2}\right)$
$T = \widetilde{O}\left(\frac{L\sqrt{d}}{\varepsilon}\left(1 + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right)\right)$

Table: Iteration complexity for minibatch-SignSGD

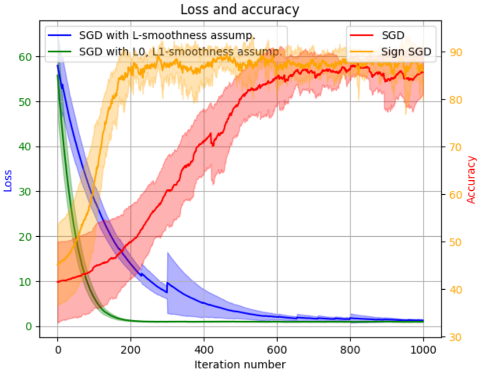
Our findings for M$_S$*ignSGD*

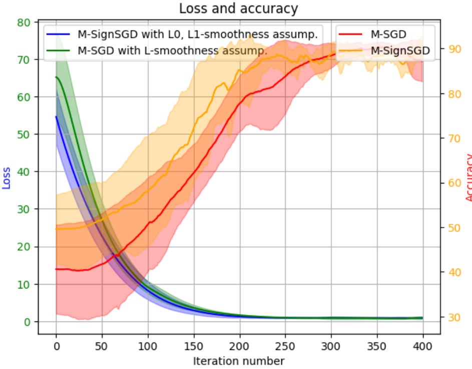
Smoothness	Step size γ_k
$(L_0, L_1),\ \varepsilon \geq \frac{3L_0}{cL_1}$	$\gamma_k = \frac{1-\beta_k}{8cL_1 d}$, $1-\beta_k = \min\left\{1, \left(\frac{c\Delta_1 L_1\sqrt{d}}{T\|\vec{\sigma}\|_\kappa}\right)^{\frac{\kappa}{2\kappa-1}}\right\}$
$(L_0, L_1),\ \varepsilon < \frac{3L_0}{L_1}$	$\gamma_k = \sqrt{\frac{\Delta_1(1-\beta_k)}{TL_0 d}}$, $1-\beta_k = \min\left\{1, \left(\frac{\Delta_1 L_0}{T\|\vec{\sigma}\|_\kappa^2}\right)^{\frac{\kappa}{3\kappa-2}}\right\}$
L-smooth	$\gamma = \frac{1}{L\sqrt{d}}$, $\beta = 1 - \Theta\left(\left(\frac{1}{T}\right)^{\frac{\kappa}{2\kappa-1}}\right)$

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

Table: Iteration complexity for M-SignSGD

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.