Sign operator for (L_0, L_1)-smooth optimization

Mark Ikonnikov
ikonnikov.mi@phystech.edu

Nikita Kornilov
kornilov.nm@phystech.edu

March 20, 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 [2], 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 [4], 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 [3].

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 [2], 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 $\kappa \in (1, 2]$ [4]—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 [4]. We also leverage error feedback mechanisms [3] 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. [2] focus on (L_0, L_1)-smoothness without sign-based compression. Kornilov et al. [4] 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

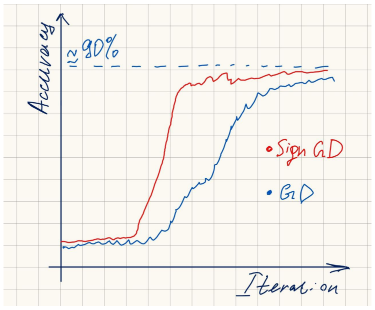

Figure 1: Performance of GD, SignGD

assumptions. Performance will be evaluated based on accuracy and iteration number, with minimal tuning to emphasize simplicity.

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.

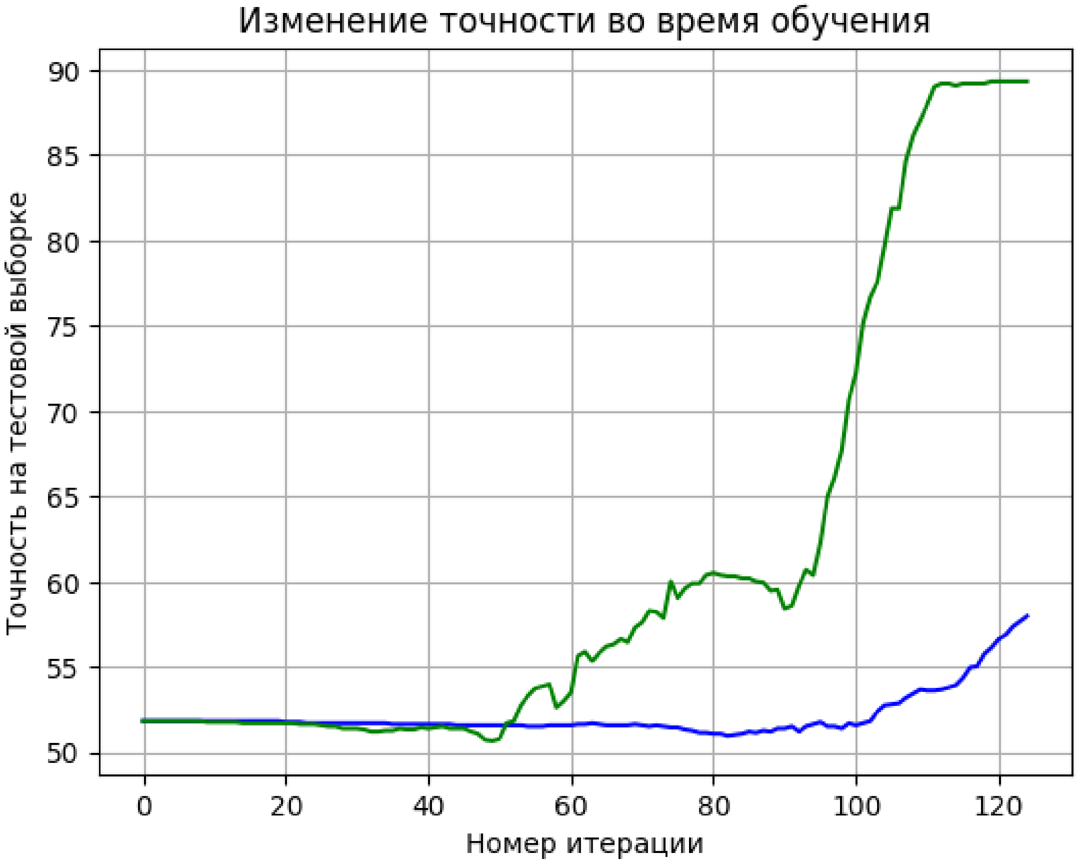

Figure 2: Logistic regression on Mushroom Dataset.
Performance of GD, SignGD with L_1-tuned step size

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] 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`.

[3] 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`.

[4] 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.