UDC 519.676, 519.673, 519.688

Sign operator for (L_0, L_1)-smooth optimization

M. I. Ikonnikov, N. M. Kornilov, A. N. Beznosikov
Moscow Institute of Physics and Technology (National Research University)

In Machine Learning, the non-smoothness of optimization problems, high communication cost between distributed workers, and stochastic gradient corruption with heavy-tailed noise motivate the study of new methods under weaker assumptions. This paper investigates sign-based optimization algorithms—specifically SignSGD and M-SignSGD—under (L_0, L_1)-smoothness and heavy-tailed (HT) noise models.

Originally, SignSGD was proposed by Bernstein et.al ([1]) as a communication-efficient alternative to SGD, offering convergence for non-convex problems by transmitting only the sign of gradients. In the paper it is proved that *SignSGD* can get the best of both worlds: compressed gradients and SGD-level convergence rate. Recent advancements have deepened the theoretical understanding of *sign-based optimization methods* under heavy-tailed noise conditions. In their high-probability analysis, Kornilov et. al.([2]) introduce convergence guarantees for SignSGD, Majority vote SignSGD and M-SignSGD under heavy-tailed stochastic noise and L-smoothness, assuming only a bounded κ-th moment for $\kappa \in (1, 2]$. The results demonstrate that SignSGD achieves optimal sample complexity $\tilde{O}\left(\varepsilon^{\frac{-3k-2}{k-1}}\right)$ with high probability for attaining an average gradient norm accuracy of ε. Under HT conditions the upper bound $O\left(\varepsilon^{\frac{-3k-2}{k-1}}\right)$ for convergence of M-SignSGD is provided. In convex settings, Gorbunov et. al. ([3]) develop a comprehensive framework for (L_0, L_1)-smooth optimization.

Collectively, these works motivate the continued exploration of sign-based methods for large-scale stochastic optimization, especially in the presence of $L_0, L_1)$-smoothness and noise with weak moment assumptions.

Definition. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be (L_0, L_1)-*smooth* if

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

This generalizes classical L-smoothness and captures many practical functions that appear in deep learning.

The noise model assumes that the stochastic gradients $\nabla f(x, \xi)$ have bounded κ-th moment:

$$\mathbb{E}_\xi[|\nabla f(x, \xi)_i - \nabla f(x)_i|^\kappa] \leq \sigma_i^\kappa, \quad \text{for all } i \in [d], \quad \kappa \in (1, 2].$$

Theoretical Contributions.

- **Step Update Lemma.** We prove that for $x' = x - \gamma A \cdot \text{sign}(m)$ where $m = \nabla f(x) + \epsilon$, and diagonal matrix A, the following holds:

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

- **Theorem (Minibatch SignSGD)** with probability $1 - \delta$ with ε small enough, sample complexity to reach $\mathbb{E}\|\nabla f(x^k)\|_1 \leq \varepsilon$ is

$$N = O\left(\frac{\Delta_1 L_0^\delta d}{\varepsilon^2}\left(1 + \left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right)\right).$$

- **Theorem (M-SignSGD)** in expectation with ε small enough:

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

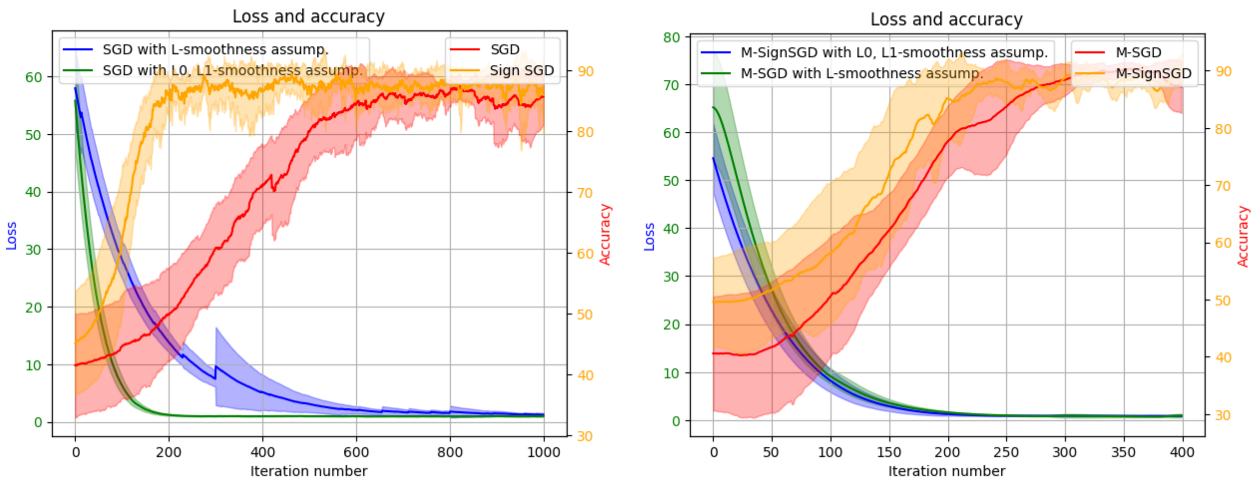

Figure 1: Left: GD vs SignSGD; Right: M-SGD vs M-SignSGD on logistic regression (Mushroom dataset).

Table 1: Comparison of Optimization Methods

Method	Mean Loss	Loss Var.	Mean Acc. (%)	Acc. Var.
M-SignSGD	3.64	73.56	82.87	135.77
M-SGD	7.73	209.47	73.47	341.59
SignSGD	6.72	155.11	79.12	140.47
SGD	16.45	234.21	62.96	70.55

Experiment. We validate the theoretical findings using logistic regression on the Mushroom dataset. The goal is to compare Sign-based methods (SignSGD, M-SignSGD) against standard SGD and momentum-SGD under minimal tuning and HT noise. Results are shown in Fig. 1 and Table 1.

Conclusion. We establish high-probability and expectation-based convergence guarantees for SignSGD variants under general (L_0, L_1)-smoothness and heavy-tailed gradient noise. The results confirm that sign-based methods can provide communication efficiency and noise resilience, supporting their use in distributed learning and low-resource systems.

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] 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.

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

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