Sign operator for $(L_0, L_1)-$smooth optimization

Ikonnikov Mark
Phystech School of Applied Mathematics and Informatics
Moscow Institute of Physics and Technology

April 8, 2025

Overview

This presentation explores (L_0, L_1)-smooth optimization, a generalization of traditional smoothness for functions with sparse or structured gradients. Topics include:

- Definition of (L_0, L_1)-smoothness and its implications.
- Key algorithms, such as Sign-SGD and its variants.
- Theoretical results under heavy-tailed (HT) noise.
- Novel theoretical bounds for the methods under (L_0, L_1)-assumption.

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, 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, 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, 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.

Definition of (L_0, L_1)-Smoothness

A function $f : \mathbb{R}^d \to \mathbb{R}$ is (L_0, L_1)-smooth if, for all $x, y \in \mathbb{R}^d$, its gradient satisfies:

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

where:

- $L_0 \geq 0$: Base Lipschitz constant.
- $L_1 \geq 0$: Gradient-dependent smoothness factor.
- $[x, y]$: Line segment between x and y.

This captures non-uniform gradient behavior in sparse or noisy optimization.

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 $a \in \mathbb{R}^d$ 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.

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 $x^{k+1} = x^k - \gamma_k \cdot \mathrm{sign}(g^k)$;

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 $\{B_k\}_{k=1}^T$.

for $k = 1, \ldots, T$ **do**
    Sample $\{\xi_i^k\}_{i=1}^{B_k}$ and compute gradient estimate $g^k = \sum_{i=1}^{B_k} \nabla f(x^k, \xi_i^k)/B_k$;
    Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(g^k)$;
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 $\{\beta_k\}_{k=1}^T$.

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 $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(m^k)$;
5: **end for**

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

HT Noise Definition

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

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

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

Assumptions

Assumption (Lower Bound)

The function f is lower bounded: $f(x) \geq f^* > -\infty$, $\forall x \in \mathbb{R}^d$.

Assumption (Smoothness)

The function f is differentiable and (L_0, L_1)*-smooth:*

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

Assumption (Heavy-Tailed Noise)

The gradient estimate $\nabla f(x, \xi)$ *is unbiased with bounded* κ*-th moments:*

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

where $\kappa \in (1, 2]$, $\sigma_i \geq 0$.

Lemma

Lemma

(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

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

Moreover, it implies

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

Lemma

Lemma (HT Batching Lemma)

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

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

Theorem (**HP 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 $\frac{1}{T}\sum_{k=1}^{T}\|\nabla f(x^k)\|_1 \le \varepsilon$ with probability at least $1-\delta$ for:

Optimal tuning. *In case $\varepsilon \ge \frac{8L_0}{L_1\sqrt{d}}$, we use stepsize*

$\gamma = \frac{1}{48L_1 d \log\frac{1}{\delta}\sqrt{d}} \Rightarrow 80L_0 d\gamma \log(1/\delta) \le \varepsilon/2$ *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:*

$$\begin{aligned} \varepsilon &\ge \frac{8L_0}{L_1\sqrt{d}} &\Rightarrow\quad N = O\left(\frac{\Delta_1 L_1 \log(1/\delta) d^{\frac{3}{2}}}{\varepsilon}\left[1 + \left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right), \\ \varepsilon &< \frac{8L_0}{L_1\sqrt{d}} &\Rightarrow\quad N = O\left(\frac{\Delta_1 L_0 \log(1/\delta) d}{\varepsilon^2}\left[1 + \left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right). \end{aligned} \tag{4}$$

Proof Scketch

Theorem (HP complexity for minibatch-L0L1-SignSGD proof sketch)

- *Consider the k-th step and use the Lemma.*
- *After summing T steps, introduce the following terms* $\phi_k := \frac{\langle \nabla f(x^k), \mathrm{sign}(g^k) \rangle}{\|\nabla f(x^k)\|_1} \in [-1, 1]$, $\psi_k := \mathbb{E}[\phi_k | x^k]$ *and* $D_k := -\gamma_k(\phi_k - \psi_k)\|\nabla f(x^k)\|_1$. D_k *is a martingale difference sequence.*
- *Applying Measure Concentration Lemma to MSD we derive the bound for all* $\lambda > 0$ *with probability at least* $1 - \delta$.
- *use norm relation and* (L_0, L_1)*-smoothness to estimate maximum gradient norm for all* $k \in \overline{2, T+1}$:
- *We take* $\gamma_k \leq \frac{1}{48 L_1 d \log \frac{1}{\delta} \sqrt{d}}$ *to obtain the estimate for* $\|\nabla f(x^k)\|_1 / \sqrt{d} \leq ...$
- *We estimate each term* $\psi_k \|\nabla f(x^k)\|_1$ *using Markov's inequality followed by Jensen's inequality*
- *We put this bound in telescopic sum and obtain our bound.*

Theorem (HP complexity for momentum-L0L1-SignSGD)

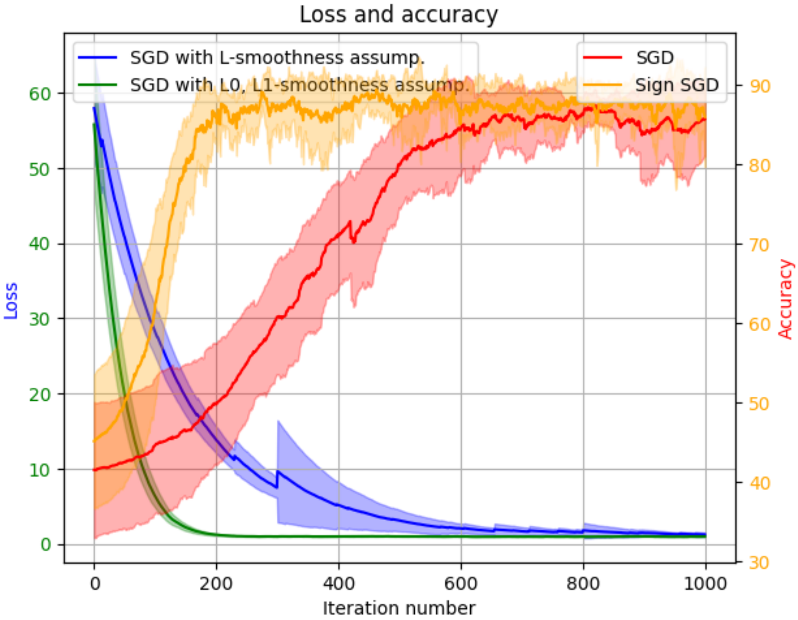
Consider lower-bounded $(L0, L1)$-smooth function f and HT gradient estimates. Then Alg. minibatch-SignSGD requires the sample complexity N to achieve $\frac{1}{T}\sum_{k=1}^{T} \|\nabla f(x^k)\|_1 \leq \varepsilon$ with probability at least $1 - \delta$ for:

Optimal tuning. *In progress ...*

Proof Scketch

Theorem (HP complexity for minibatch-L0L1-SignSGD proof sketch)

- *Consider the k-th step and use the Lemma.*
- *After summing T steps, introduce the terms $\epsilon^k := m^k - \nabla f(x^k)$ and $\theta^k := g^k - \nabla f(x^k)$ and note that $\{\theta_i\}$ is a martingale difference sequence.*
- *use norm relation and (L_0, L_1)-smoothness to estimate maximum gradient norm for all $k \in \overline{2, T+1}$:*
- *Applying Measure Concentration Lemma and HT Batching Lemma to MSD we derive the bound for the expected value with probability at least $1 - \delta$.*
- *We finally obtain that components of the telescopic sum split to L_0 and L_1-dependent*



Loss and accuracy
SGD with L-smoothness assump.
SGD with L0, L1-smoothness assump.
SGD
Sign SGD
Loss
Accuracy
Iteration number
0
10
20
30
40
50
60
0
200
400
600
800
1000
30
40
50
60
70
80
90

Plans

These are the plans for the time we have left:

- Obtain the optimal tuning convergence bound for M-SignSGD and finish the proof.
- Modify the proof of SingSGD with minibatch to obtain the bound for MajorityVote-SignSGD.
- Validate theoretical findings with numerical experiments.
- (Optional) Explore the changing parameters.
- (Optional) Consider the methods under the convexity assumption.

Key Articles for Research

Topic	Title	Year	Authors	Paper	Summary
Key article 1	Sign Operator for Coping with Heavy-Tailed Noise	2025	Kornilov et al.	arXiv	Proofs for heavy-tailed noise
Key article 2	signSGD: Compressed Optimisation for Non-Convex Problems	2018	J. Bernstein et al.	PMLR	3 Sign-based methods
Key article 3	Methods for Convex (L0,L1)-Smooth Optimization: Clipping, Acceleration, and	2024	Gorbunov et al.	arXiv	New convergence guarantees for existing methods

Additional Papers for Methods and Proofs (Part 1)

Topic	Title	Year	Authors	Paper	Summary
Additional theory	Robustness to Unbounded Smoothness of Generalized SignSGD	2022	M. Crawshaw et al.	Curran Associates	$L_0, L1SignSG$
Additional theory	Error Feedback Fixes SignSGD and other	2019	Karimireddy et al.	PMLR	Check for convex case

Additional Papers for Methods and Proofs (Part 2)

Topic	Title	Year	Authors	Paper	Summary
Proofs	From Gradient Clipping to Normalization for Heavy Tailed SGD	2024	Hubler et al.	arXiv	Heavy Tailed SGD
Additional theory	Why gradient clipping accelerates training: A theoretical justification for adaptivity	2020	Zhang et al.	arXiv	Intro to (L0,L1)-smoothness assump.