Sign operator for (L_0, L_1)-smooth optimization

Mark Ikonnikov
ikonnikov.mi@phystech.edu

Nikita Kornilov
kornilov.nm@phystech.edu

Aleksandr Beznosikov
beznosikov.an@phystech.edu

May 17, 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 further research of optimization methods under broader assumptions. This paper explores the efficacy of sign-based methods, which address slow transmission by communicating only the sign of each stochastic gradient. We investigate these methods for (L_0, L_1)-smooth problems, which encompass a wider range of problems than the L-smoothness assumption. To address the problem of data accuracy, we introduce the convergence bounds for (L_0, L_1) *SignSGD* and *M-SignSGD* under heavy-tailed noise, defined as noise with bounded κ-th moment $\kappa \in (1, 2]$.

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

1 Introduction

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.

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. Majority vote aggregation enables 1-bit communication per worker and maintains variance reduction comparable to full-precision distributed SGD. The research for the majority vote method was conducted under the assumption on noise, i.e. noise in each component of the stochastic gradient is unimodal and symmetric about the mean (e.g. Gaussian).

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 $\kappa \in (1, 2]$. Their results extend to distributed optimization and the comparison oracle setting, highlighting the robustness of sign-based updates in non-ideal regimes. The results demonstrate that SignSGD achieves optimal sample complexity $\tilde{O} \ \left(\varepsilon^{\frac{3k-2}{k1}}\right)$ with high probability for attaining an average gradient norm accuracy of ε. Under HT conditions the upper bound $O\left(\varepsilon^{\frac{3k-2}{k1}}\right)$ 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, introducing adaptive, clipped, and accelerated variants of existing methods with new convergence guarantees.

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.

1.1 Contributions

In this project, we:

1. Investigated sign-based methods for communication-efficient distributed optimization under the assumptions above.

2. Developed high-probability convergence guarantees accounting for generalized conditions.

The experimental goals are to validate convergence under (L_0, L_1)-smoothness and heavy-tailed noise. The setup includes real-world datasets datasets satisfying (L_0, L_1)-smoothness with synthetic heavy-tailed noise and convex logistic regression models. The workflow compares sign-based methods against traditional methods, measuring convergence rates and including different accuracy scorings.

The nearest alternative to our research is Gorbunov et. al. ([3]). Our advantage is the extension to (L_0, L_1)-smoothness while maintaining 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

The object of this research is the stochastic optimization of 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., $\|\nabla f(x^*)\| \leq \varepsilon$, especially for non-convex objectives.

Traditional optimization often assumes L-smoothness (Lipschitz continuity of the gradient), but this may appear to be restrictive for real-world deep learning models like Transformers. Instead, we adopt the (L_0, L_1)-smoothness condition [3], 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.

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.,

$$\mathbb{E}_\xi[\|\nabla f(x, \xi) - \nabla f(x)\|_2^\kappa] \leq \sigma^\kappa, \kappa \in (1, 2]$$

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 assumption, 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 $\mathbb{R}^d$.

2 Theory

In this section, we present our novel convergence guarantees with high probability for existing sign-based methods for non-convex functions with heavy-tailed noise in gradient estimates. For two algorithms (minibatch SignSGD and M-SignSGD), we provide an explicit optimal tuning for the parameters. All proofs are located in Appendix **??**.

2.1 Assumptions

Assumption 1 (Lower bound)**.** *The objective function f is lower bounded by $f^* > -\infty$, i.e., $f(x) \geq f^*, \forall x \in \mathbb{R}^d$.*

Assumption 2 ((L_0, L_1)-Smoothness)**.** *The objective function f is differentiable and symmetrically (L_0, L_1)-smooth, i.e., for the non-negative constants (L_0, L_1) for all $x, y \in \mathbb{R}^d$*

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

Assumption 3 (Heavy-tailed noise in gradient estimates)**.** *The unbiased estimate $\nabla f(x, \xi)$ has bounded κ-th moment $\kappa \in (1, 2]$ for each coordinate, i.e., $\forall x \in \mathbb{R}^d$:*

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

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

2.2 SignSGD

We begin our analysis by presenting the simplest of sign-based methods, namely SignSGD (Alg. 1) formulated first in Bernstein et. al. [1].

Algorithm 1 SignSGD

```
Input: Starting point $x^1 \in \mathbb{R}^d$, number of iterations $T$, stepsizes $\{\gamma_k\}_{k=1}^T$.
for $k = 1, \ldots, T$ do
    Sample $\xi^k$ and compute estimate $g^k = \nabla f(x^k, \xi^k)$;
    Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(g^k)$;
end for
Output: uniformly random point from $\{x^1, \ldots, x^T\}$ . =0
```

In order to achieve accuracy ε, the noise $\|\vec{\sigma}\|_1$ have not to exceed ε. The first way to lower the noise is to use batching.

2.3 SignSGD with minibatching

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 $k = 1, \dots, T$ do
    Sample $\{\xi_i^k\}_{i=1}^{B_k}$
    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 $\{x^1, \dots, x^T\}$ . =0
```

Thus, under the assumptions above we formulate the following theorem:

Theorem 1 (**(L_0, L_1) HP complexity for minibatch-SignSGD**). *Consider lower-bounded (L_0, L_1)-smooth function f (As. 1, 2) and HT gradient estimates (As. 3). Then Alg. 2 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:*

Case** $\varepsilon \geq \frac{8L_0}{cL_1\sqrt{d}}$**: $T = O\left(\frac{c\Delta_1 L_1^\delta d^{\frac{3}{2}}}{\varepsilon}\right)$, $\gamma_k \equiv \frac{1}{48cL_1^\delta d^{\frac{3}{2}}}$, $B_k \equiv \max\left\{1, \left(\frac{16\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right\}$:

$$\varepsilon \quad \geq \quad \frac{8L_0}{cL_1\sqrt{d}} \quad \Rightarrow \quad N = O\left(\frac{c\Delta_1 L_1^\delta d^{\frac{3}{2}}}{\varepsilon}\left[1 + \left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right), \quad c \geq 1$$

Case** $\varepsilon < \frac{8L_0}{L_1\sqrt{d}}$**: $T = O\left(\frac{\Delta_1 L_0^\delta d}{\varepsilon^2}\right)$, $\gamma_k \equiv \sqrt{\frac{\Delta_1}{20L_0^\delta dT}}$, $B_k \equiv \max\left\{1, \left(\frac{16\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right\}$:

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

where $\Delta_1 = f(x^1) - f^*$, $L_0^\delta = L_0 \log(1/\delta)$, $L_1^\delta = L_1 \log(1/\delta)$.

2.4 SignSGD with momentum

Instead of variance reduction, one can use the momentum technique with the same sample complexity.

Algorithm 3 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 $\xi^k$ 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 $\{x^1, \ldots, x^T\}$ . =0
```

In M-SignSGD, the sign operator is applied to the momentum vector instead of the gradient estimate. The following theorem states in expectation convergence rates and parameters for M-SignSGD.

Theorem 2 (**(L_0, L_1) Complexity for M-SignSGD in expectation**). *Consider lower-bounded (L_0, L_1)-smooth function f (As. 1, 2) and HT gradient estimates (As. 3). Then Alg. 3 requires T iterations to achieve $\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}\left[\|\nabla f(x^k)\|_1\right] \leq \varepsilon$ for:*

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

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

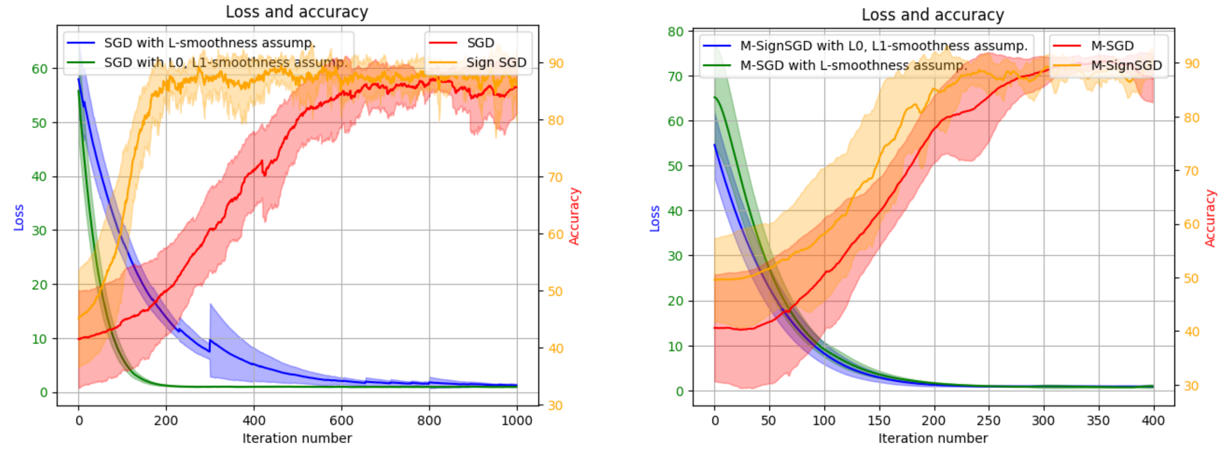
***Case** $\varepsilon < \frac{3L_0}{L_1}$**:** $1 - \beta_k \equiv 1 - \min\left\{1, \left(\frac{\Delta_1 L_0}{T\|\vec{\sigma}\|_\kappa^2}\right)^{\frac{\kappa}{3\kappa-2}}\right\}, \gamma_k \equiv \sqrt{\frac{\Delta_1(1-\beta_k)}{TL_0 d}}$*

$$T = O\left(\frac{\Delta_1 L_1 d}{\varepsilon^2}\left(1 + \left(\frac{\sqrt{d}\|\vec{\sigma}\|_\kappa}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right)\right), \tag{3}$$

where $\Delta_1 = f(x^1) - f^*$.

3 Computational experiment

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



((a)) GD and SignSGD with L_1-tuned step size.

((b)) Momentum SGD and M-SignGD with L_1-tuned step size.

Figure 1: Logistic regression on Mushroom Dataset.
Comparison of Sign-based optimization methods.

model, highlighting the advantages of sign-based methods under L_0, L_1 smoothness assumptions. Performance will be evaluated based on accuracy and iteration number, with minimal tuning to emphasize simplicity.

Table 1: Loss Comparison of Optimization Methods

Method	**Mean Loss**	**Sigma, gamma, Loss Variance**	**Convergence Rate**
M-SignSGD	3.6366	73.5606	0.0537
M-SGD	7.7299	209.4682	0.1289
SignSGD	6.7171	155.1072	0.1215
SGD	16.4469	234.2066	0.1166

Table 2: Accuracy Comparison of Optimization Methods

Method	**Mean Accuracy (%)**	**Accuracy Variance**
M-SignSGD	82.8688	135.7708
M-SGD	73.4684	341.5860
SignSGD	79.1219	140.4712
SGD	62.9615	70.5549

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.

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.

[6] Nikita Kornilov et al. "Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance". In: *Advances in Neural Information Processing Systems* 36 (2024).

[7] Florian Hübler, Ilyas Fatkhullin, and Niao He. "From Gradient Clipping to Normalization for Heavy Tailed SGD". In: *arXiv preprint arXiv:2410.13849* (2024).

[8] Tao Sun et al. "Momentum ensures convergence of signsgd under weaker assumptions". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 33077–33099.

4 Proofs

Proposition 1 (Norm Relation)**.** *For two norms ℓ_p and ℓ_q with $1 \le p \le q \le 2$, the following relation holds true:*

$$\|x\|_q \le \|x\|_p \le d^{\frac{1}{p}-\frac{1}{q}}\|x\|_q, \quad \forall x \in \mathbb{R}^d. \tag{4}$$

Proposition 2 (Jensen's Inequality)**.** *For scalar random variable ξ with bounded κ-th moment $\kappa \in (1,2]$, the following inequality holds true:*

$$\mathbb{E}[|\xi|] \le \left(\mathbb{E}[|\xi|^\kappa]\right)^{\frac{1}{\kappa}}. \tag{5}$$

Proposition 3 (Markov's Inequality)**.** *For scalar random variable ξ with bounded first moment, the following inequality holds true for any $a > 0$:*

$$\mathbb{P}(|\xi - \mathbb{E}[\xi]]| \ge a) \le \frac{\mathbb{E}[|\xi|]}{a}. \tag{6}$$

Lemma 1. *(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*

$$\|\nabla f(x) - \nabla f(y)\|_2 \le (L_0 + L_1\|\nabla f(y)\|_2)\exp(L_1\|x-y\|_2)\|x-y\|_2. \tag{7}$$

Moreover, it implies

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_0 + L_1\|\nabla f(x)\|_2}{2}\exp(L_1\|x-y\|_2)\|x-y\|_2^2. \tag{8}$$

To prove the HP bounds with the logarithmic dependency, we use the following measure concentration result (see, for example, [4, Lemma 1].

Lemma 2 (Measure Concentration Lemma)**.** *Let $\{D_k\}_{k=1}^T$ be a martingale difference sequence (MDS), i.e., $\mathbb{E}[D_k|D_{k-1},\ldots,D_1] = 0$ for all $k \in \overline{1,T}$. Furthermore, for each $k \in \overline{1,T}$, there exists positive $\sigma_k \in \mathbb{R}$, s.t. $\mathbb{E}\left[\exp\left(\frac{D_k^2}{\sigma_k^2}\right)|k\right] \le e$. Then the following probability bound holds true:*

$$\forall \lambda > 0, \delta \in (0,1): \quad \mathbb{P}\left(\sum_{k=1}^T D_k \le \frac{3}{4}\lambda\sum_{k=1}^T \sigma_k^2 + \frac{1}{\lambda}\log(1/\delta)\right) \ge 1 - \delta. \tag{9}$$

To control error reduction during batching, we use the following batching lemma for HT variables. Its modern proof for $d = 1$ was proposed in [5, Lemma 4.2] and then generalized for the multidimensional case in [6, 7].

Lemma 3 (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*

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^B X_i\right\|_2^\kappa\right] \le \frac{2}{B^\kappa}\sum_{i=1}^B \mathbb{E}[\|X_i\|_2^\kappa]. \tag{10}$$

We need the following lemma about changes after one update step of sign-based methods from [8, Lemma 1].

Lemma 4 (Sign Update Step Lemma)**.** *Let $x, m \in \mathbb{R}^d$ be arbitrary vectors, $A = diag(a_1, \dots, a_d)$ be diagonal matrix and f be L-smooth function (As. 2). Then for the update step*

$$x' = x - \gamma \cdot A \cdot \operatorname{sign}(m)$$

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

$$f(x') - f(x) \leq -\gamma \|A\nabla f(x)\|_1 + 2\gamma\|A\|_F\|\epsilon\|_2 + \frac{L_0 + L_1\|A\nabla f(x^k)\|_2}{2}\exp\left(\gamma L_1\|A\|_F\right)\gamma^2\|A\|_F^2. \tag{11}$$

Proof of Theorem. Consider the k-th step of SignSGD. We use (L_0, L_1) smoothness of function f (Lemma 1) to estimate:

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &\leq \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L_0 + L_1\|\nabla f(x^k)\|_2}{2}\exp(L_1\|x^{k+1} - x^k\|_2)\|x^{k+1} - x^k\|_2^2 \\
&= -\gamma_k\frac{\langle \nabla f(x^k), \operatorname{sign}(g^k)\rangle}{\|\nabla f(x^k)\|_1} \cdot \|\nabla f(x^k)\|_1 + \frac{L_0 d\gamma_k^2}{2}\exp(L_1\sqrt{d}\gamma_k) \\
&+ \frac{L_1 d\gamma_k \exp(L_1\sqrt{d}\gamma_k)}{2} \cdot \gamma_k\|\nabla f(x^k)\|_2 \\
&\leq -\gamma_k\frac{\langle \nabla f(x^k), \operatorname{sign}(g^k)\rangle}{\|\nabla f(x^k)\|_1} \cdot \|\nabla f(x^k)\|_1 + \frac{L_0 d\gamma_k^2}{2}\exp(L_1\sqrt{d}\gamma_k) \\
&+ \frac{L_1 d\gamma_k \exp(L_1\sqrt{d}\gamma_k)}{2} \cdot \gamma_k\|\nabla f(x^k)\|_1.
\end{aligned}$$

Let us choose $\gamma_k \leq \frac{1}{4L_1 d}$, then we have $L_1 d\gamma_k \exp(L_1\sqrt{d}\gamma_k) \leq \frac{1}{4}$ and

$$f(x^{k+1}) - f(x^k) \leq -\gamma_k\frac{\langle \nabla f(x^k), \operatorname{sign}(g^k)\rangle}{\|\nabla f(x^k)\|_1} \cdot \|\nabla f(x^k)\|_1 + L_0 d\gamma_k^2 + \frac{\gamma_k}{4}\|\nabla f(x^k)\|_1.$$

Consequently, after summing all T steps, we obtain:

$$\sum_{k=1}^{T}\gamma_k\left[\frac{\langle \nabla f(x^k), \operatorname{sign}(g^k)\rangle}{\|\nabla f(x^k)\|_1} - \frac{1}{4}\right] \cdot \|\nabla f(x^k)\|_1 \leq \underbrace{f(x^1) - f(x^*)}_{=\Delta_1} + L_0 d\sum_{k=1}^{T}\gamma_k^2. \tag{12}$$

We introduce the following terms $\phi_k := \frac{\langle \nabla f(x^k), \operatorname{sign}(g^k)\rangle}{\|\nabla f(x^k)\|_1} \in [-1, 1]$, $\psi_k := \mathbb{E}[\phi_k | x^k]$ and $D_k := -\gamma_k(\phi_k - \psi_k)\|\nabla f(x^k)\|_1$. We note that D_k is a martingale difference sequence ($\mathbb{E}[D_k | D_{k-1}, \dots, D_k] = 0$) and satisfies

$$\exp\left(\frac{D_k^2}{4\gamma_k^2\|\nabla f(x^k)\|_1^2}\right) = \exp\left(\frac{(\phi_k - \psi_k)^2}{4}\right) \leq e.$$

Applying Measure Concentration Lemma 2 to MSD D_k with $\sigma_k^2 = 4\gamma_k^2 \|\nabla f(x^k)\|_1^2$, we derive the bound for all $\lambda > 0$ with probability at least $1-\delta$:

$$\sum_{k=1}^{T} \gamma_k(\psi_k - 3\lambda\gamma_k \|\nabla f(x^k)\|_1 - {}^1\!/\!_4)\|\nabla f(x^k)\|_1 \leq \Delta_1 + L_0 d \sum_{k=0}^{T-1} \gamma_k^2 + \frac{1}{\lambda}\log({}^1\!/\!_\delta).$$

We use norm relation (4) and (L_0, L_1)-smoothness to estimate maximum gradient norm for all $k \in \overline{2, T+1}$:

$$\begin{array}{rcl} \|\nabla f(x^k)\|_1/\sqrt{d} & \leq & \|\nabla f(x^k)\|_2 \leq \|\nabla f(x^k) - \nabla f(x^{k-1}) + \nabla f(x^{k-1})\|_2 \\ & \leq & \|\nabla f(x^k) - \nabla f(x^{k-1})\|_2 + \|\nabla f(x^{k-1})\|_2 \\ & \leq & (L_0 + L_1\|\nabla f(x^{k-1})\|_2)\exp(L_1\|x^k - x^{k-1}\|_2)\|x^k - x^{k-1}\|_2 + \|\nabla f(x^{k-1})\|_2 \\ & \leq & (L_0 + L_1\|\nabla f(x^{k-1})\|_2)\exp(L_1\sqrt{d}\gamma_k)\sqrt{d}\gamma_k + \|\nabla f(x^{k-1})\|_2. \end{array}$$

At this point, we take $\gamma_k \leq \frac{1}{48 L_1 d \log\frac{1}{\delta}\sqrt{d}}$ to obtain

$$\begin{array}{rcl} \|\nabla f(x^k)\|_1/\sqrt{d} & \leq & 2L_0\sqrt{d}\gamma_k + \dfrac{\|\nabla f(x^{k-1})\|_2}{48 d \log\frac{1}{\delta}} + \|\nabla f(x^{k-1})\|_2 \\ & \leq & 2L_0\sqrt{d}\displaystyle\sum_{\tau=1}^{k-1}\gamma_\tau + \sum_{\tau=1}^{k-1}\dfrac{\|\nabla f(x^\tau)\|_2}{48 d\log\frac{1}{\delta}} + \|\nabla f(x^1)\|_2 \\ & \leq & 2L_0\sqrt{d}\displaystyle\sum_{\tau=1}^{k-1}\gamma_\tau + \sum_{\tau=1}^{k-1}\dfrac{\|\nabla f(x^\tau)\|_1}{48 d\log\frac{1}{\delta}} + \|\nabla f(x^1)\|_1. \end{array}$$

Hence, the choice $\lambda := \frac{1}{6d(\gamma^{max}\|\nabla f(x^1)\|_1 + \sum_{k=1}^T \frac{\gamma_k\|\nabla f(x^k)\|_1}{48 d \log\frac{1}{\delta}} + 2C_T L_0)}$ where $C_T := \max\limits_{k\in\overline{1,T}} \gamma_k \cdot \sum\limits_{\tau=1}^{k-1}\gamma_\tau$ and $\gamma^{max} := \max\limits_{k\in\overline{1,T}}\gamma_k$ yields with probability at least $1-\delta$:

$$\begin{array}{rcl} \displaystyle\sum_{k=1}^{T}\gamma_k\left(\psi_k - \frac{1}{2} - \frac{1}{4}\right)\|\nabla f(x^k)\|_1 & \leq & \Delta_1 + L_0 d\displaystyle\sum_{k=1}^{T}\gamma_k^2 + 6\sqrt{d}(\gamma^{max}\|\nabla f(x^1)\|_1 + 2C_T L_0)\log({}^1\!/\!_\delta) + \\ & & + \dfrac{6}{48}\displaystyle\sum_{k=1}^{T}\gamma_k\|\nabla f(x^k)\|_1, \\ \displaystyle\sum_{k=1}^{T}\gamma_k\left(\psi_k - \frac{1}{2} - \frac{1}{4} - \frac{1}{8}\right)\|\nabla f(x^k)\|_1 & \leq & \Delta_1 + L_0 d\displaystyle\sum_{k=1}^{T}\gamma_k^2 + 6\sqrt{d}(\gamma^{max}\|\nabla f(x^1)\|_1 + 2C_T L_0)\log({}^1\!/\!_\delta), \end{array}$$

Next, we estimate each term $\psi_k\|\nabla f(x^k)\|_1$ in the previous sum:

$$\begin{array}{rcl} \psi_k\|\nabla f(x^k)\|_1 & = & \mathbb{E}\left[\langle\nabla f(x^k), \mathrm{sign}(g^k)\rangle | x^k\right] \\ & = & \|\nabla f(x^k)\|_1 - \displaystyle\sum_{i=1}^{d} 2|\nabla f(x^k)|_i \cdot \mathbb{P}(\mathrm{sign}(\nabla f(x^k))_i \neq \mathrm{sign}(g^k)_i | x^k). \quad (13) \end{array}$$

For each coordinate, we have a bound derived from Markov's inequality (6) followed by Jensen's inequality (5):

$$\begin{aligned}\mathbb{P}(\text{sign}(\nabla f(x^k))_i \neq \text{sign}(g^k)_i | x^k) &\leq \mathbb{P}(|\nabla f(x^k)_i - g_i^k| \geq |\nabla f(x^k)_i| | x^k) \leq \frac{\mathbb{E}_{\xi^k}[|\nabla f(x^k)_i - g_i^k|]}{|\nabla f(x^k)_i|} \\ &\leq \frac{(\mathbb{E}_{\xi^k}[|\nabla f(x^k)_i - g_i^k|^\kappa])^{\frac{1}{\kappa}}}{|\nabla f(x^k)_i|} \leq \frac{\sigma_i}{|\nabla f(x^k)_i|}. \qquad (14)\end{aligned}$$

Hence, the whole sum can be bounded as

$$\sum_{i=1}^{d} 2|\nabla f(x^k)|_i \cdot \mathbb{P}(\text{sign}(\nabla f(x^k))_i \neq \text{sign}(g^k)_i | x^k) \quad \leq \quad 2\|\vec{\sigma}\|_1.$$

Finally, we put this bound in (13) and obtain:

$$\begin{aligned}\frac{1}{16}\sum_{k=1}^{T}\gamma_k\|\nabla f(x^k)\|_1 &\leq \Delta_1 + L_0 d\sum_{k=1}^{T}\gamma_k^2 + 2\sum_{k=1}^{T}\gamma_k\|\vec{\sigma}\|_1 \\ &+ 6d(\gamma^{max}\|\nabla f(x^1)\|_1 + 2C_T L_0)\log({}^1\!/\!_\delta). \qquad (15)\end{aligned}$$

Plugging in constant stepsizes $\gamma_k \equiv \gamma \leq \frac{1}{48L_1 d\log\frac{1}{\delta}\sqrt{d}}$ implies $C_T = T\gamma^2, \gamma^{max} = \gamma$ and the required bound:

$$\frac{1}{T}\sum_{k=1}^{T}\|\nabla f(x^k)\|_1 \leq \frac{4\Delta_1}{T\gamma} + 80L_0 d\gamma\log({}^1\!/\!_\delta) + 8\|\vec{\sigma}\|_1 + 24\frac{d\|\nabla f(x^1)\|_1}{T}\log({}^1\!/\!_\delta).$$

Case $\varepsilon \geq \frac{8L_0}{cL_1\sqrt{d}}$: We use stepsize $\gamma = \frac{1}{48cL_1 d\log\frac{1}{\delta}\sqrt{d}} \Rightarrow 80L_0 d\gamma\log({}^1\!/\!_\delta) \leq \varepsilon/2$ and batchsize $8\frac{\|\vec{\sigma}\|_1}{B^{\frac{\kappa-1}{\kappa}}} \leq \varepsilon/2 \Rightarrow B_k \equiv \max\left\{1, \left(\frac{16\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right\}$. The number of iterations T is chosen to bound the first term:

$$\frac{4\Delta_1}{T\gamma} = \frac{192c\Delta_1 L_1\log\frac{1}{\delta}d^{\frac{3}{2}}}{T} \leq \frac{\varepsilon}{2} \Rightarrow T = O\left(\frac{c\Delta_1 L_1\log\frac{1}{\delta}d^{\frac{3}{2}}}{\varepsilon}\right).$$

The total number of oracle calls is:

$$\varepsilon \quad \geq \quad \frac{8L_0}{cL_1\sqrt{d}} \quad \Rightarrow \quad N = O\left(\frac{c\Delta_1 L_1\log({}^1\!/\!_\delta)d^{\frac{3}{2}}}{\varepsilon}\left[1 + \left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right),$$

Case $\varepsilon < \frac{8L_0}{L_1\sqrt{d}}$: We set the same batchsize $8\frac{\|\vec{\sigma}\|_1}{B^{\frac{\kappa-1}{\kappa}}} \leq \varepsilon/2 \Rightarrow B_k \equiv \max\left\{1, \left(\frac{16\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right\}$. The stepsize γ is set to minimize the sum:

$$\min_{\gamma}\left[\frac{4\Delta_1}{T\gamma} + 80L_0 d\gamma\log({}^1\!/\!_\delta)\right] = 2\sqrt{\frac{320\Delta_1 L_0 d\log({}^1\!/\!_\delta)}{T}},$$

it means that the stepsize $\gamma = \sqrt{\frac{4\Delta_1}{80TL_0\log(1/\delta)d}}$. The number of iterations T is chosen to satisfy

$$2\sqrt{\frac{320\Delta_1 L_0\log(1/\delta)d}{T}} \leq \frac{\varepsilon}{2} \Rightarrow T = O\left(\frac{\Delta_1 L_0\log(1/\delta)d}{\varepsilon^2}\right).$$

We only need to check whether condition $\gamma \leq \frac{1}{48L_1 d\log\frac{1}{\delta}\sqrt{d}}$ holds:

$$\begin{aligned}\gamma &= \sqrt{\frac{4\Delta_1}{80TL_0\log(1/\delta)d}} = \sqrt{\frac{4\Delta_1}{T}\frac{1}{80L_0\log(1/\delta)d}}\\ &\leq \frac{\varepsilon}{4}\frac{1}{80L_0\log(1/\delta)d} \leq \frac{8L_0}{4L_1\sqrt{d}}\frac{1}{80L_0\log(1/\delta)d}\\ &\leq \frac{1}{48L_1 d\log\frac{1}{\delta}\sqrt{d}}.\end{aligned}$$

Hence, we have the following bound for sample complexity

$$\varepsilon \quad < \quad \frac{8L_0}{L_1\sqrt{d}} \quad \Rightarrow \quad N = O\left(\frac{\Delta_1 L_0\log(1/\delta)d}{\varepsilon^2}\left[1+\left(\frac{\|\vec{\sigma}\|_1}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right]\right). \tag{16}$$

□

4.1 Proof of Sign Update Step Lemma

Proof. The proof is identical to the proof from [8, Lemma 1] with changing the smoothness assumption according to Lemma 1. Using the (L_0, L_1)-smoothness of f, we have

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

Substitute $x' - x = -\gamma A\text{sign}(m)$:

$$\langle\nabla f(x), x'-x\rangle = -\gamma\langle\nabla f(x), A\text{sign}(m)\rangle.$$

We now decompose the inner product:

$$\langle\nabla f(x), A\text{sign}(m)\rangle = \langle\nabla f(x), A\text{sign}(\nabla f(x))\rangle + \langle\nabla f(x), A(\text{sign}(m) - \text{sign}(\nabla f(x)))\rangle.$$

Using the identity:

$$\langle\nabla f(x), A\text{sign}(\nabla f(x))\rangle = \|A\nabla f(x)\|_1,$$

and define $[\nabla f(x)]_i =: g_i$, then the second term becomes

$$\sum_{i=1}^{d} a_i g_i\left(\text{sign}(m_i) - \text{sign}(g_i)\right).$$

Now we analyze two cases for each i:

- If $\text{sign}(m_i) = \text{sign}(g_i)$, then the term is zero.
- Otherwise, $g_i \cdot m_i \leq 0$, hence $|g_i - m_i| \geq |g_i|$, and we define $\epsilon_i = m_i - g_i$. Then:

$$a_i g_i \left(\text{sign}(m_i) - \text{sign}(g_i)\right) \leq 2a_i|g_i| \leq 2a_i|\epsilon_i|.$$

So we have:

$$\langle \nabla f(x), A\text{sign}(\nabla f(x)) - A\text{sign}(m)\rangle \leq 2\sum_{i=1}^{d} a_i|\epsilon_i| \leq 2\|A\|_F\|\epsilon\|_2.$$

Thus,

$$\langle \nabla f(x), x' - x\rangle \leq -\gamma\|A\nabla f(x)\|_1 + 2\gamma\|A\|_F\|\epsilon\|_2.$$

Now, observe that

$$\|x' - x\| = \gamma\|A\text{sign}(m)\|_2 \leq \gamma\|A\|_F.$$

Substituting into the smoothness upper bound:

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

□

4.2 Proof of (L_0, L_1) M-SignSGD General Convergence

Proof. Consider the k-th step of SignSGD. We use (L_0, L_1) step update (Lemma 4) to estimate:

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L_0 + L_1\|\nabla f(x^k)\|_2}{2}\exp(L_1\|x^{k+1} - x^k\|_2)\|x^{k+1} - x^k\|_2^2 \\ &\leq -\gamma_k\|\nabla f(x^k)\|_1 + 2\gamma_k\sqrt{d}\|\epsilon^k\|_2 + \frac{L_0 d\gamma_k^2}{2}\exp(L_1\sqrt{d}\gamma_k) \\ &+ \frac{L_1 d\gamma_k\exp(L_1\sqrt{d}\gamma_k)}{2}\cdot\gamma_k\|\nabla f(x^k)\|_1. \end{aligned}$$

The inequality holds due to Lemma 4.

Since we set constant steps sizes and momentum, we denote them as $\gamma \equiv \gamma_k$ and $\beta \equiv \beta_k$, respectively. We use notations $\epsilon^k := m^k - \nabla f(x^k)$ and $\theta^k := g^k - \nabla f(x^k)$. Therefore, we have at k-th step values:

$$\begin{aligned} m^k &= \beta m^{k-1} + (1-\beta)g^k = \gamma(\epsilon^{k-1} + \nabla f(x^{k-1})) + (1-\gamma)(\theta^k + \nabla f(x^k)), \\ \epsilon^k &= m^k - \nabla f(x^k) = \beta\epsilon^{k-1} + \beta\underbrace{(\nabla f(x^{k-1}) - \nabla f(x^k))}_{=:s^k} + (1-\beta)\theta^k, \\ \epsilon^k &= m^k - \nabla f(x^k) = \beta\epsilon^{k-1} + \beta s^k + (1-\beta)\theta^k. \end{aligned}$$

Unrolling the recursion, we obtain an explicit formula (upper index of β is its power):

$$\epsilon^k \;=\; \beta^{k-1}\epsilon^1 + \sum_{i=2}^{k}\beta^{k-i+1}s^i + (1-\beta)\sum_{i=2}^{k}\beta^{k-i}\theta^i. \tag{17}$$

From $(L_0, L_1)-$smoothness of f (Lem. 1) follows the bound:

$$\|s^k\|_2 \leq (L_0+L_1\|\nabla f(x^k)\|_2)\exp(L_1\|x^k-x^{x+1}\|_2)\|x^k-x^{k+1}\|_2 = (L_0+L_1\|\nabla f(x^k)\|_2)\exp(L_1\gamma_k\sqrt{d})\gamma_k\sqrt{d}$$

Denote $\lambda := \exp(L_1\gamma_k\sqrt{d})\gamma_k\sqrt{d}$.

Hence, the norm of (17) can be bounded as:

$$\|\epsilon^k\|_2 \leq \beta^{k-1}\|\epsilon^1\|_2 + L_0\lambda\sum_{i=2}^{k}\beta^{k-i+1} + L_1\lambda\sum_{i=2}^{k}\beta^{k-i+1}\|\nabla f(x^k)\|_2 + (1-\beta)\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2.$$

We notice that variables $\{\theta_i\}$ are martingale difference sequence from Lemma 3 which we plan to use. Due to the formal definition of $\theta^i = g^i - \nabla f(x^i) = \nabla f(x^i,\xi_i) - \nabla f(x^i)$ and M-SinSGD step, the conditioning on $\theta^{i-1},\dots,\theta^1$ with randomness $\xi_1,\dots,\xi_{i-1}$ is equivalent to the conditioning on point s $x^i,\dots,x^2$. Hence, we show by definition of martingale difference sequence that

$$\mathbb{E}[\theta^i|\theta^{i-1},\dots,\theta^1] = \mathbb{E}[\theta^i|x^i,\dots,x^2] = \mathbb{E}[\nabla f(x^i,\xi_i) - \nabla f(x^i)|x^i,\dots,x^2] = 0.$$

To take math expectation from both sides, we first take it from the term

$$\mathbb{E}\left[\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2\right] \leq \left(\mathbb{E}\left[\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2^\kappa\right]\right)^{\frac{1}{\kappa}} \overset{\text{Lem. } 3}{\leq} \left(\sum_{i=2}^{k}2\mathbb{E}\left[\|\beta^{(k-i)}\theta^i\|_2^\kappa\right]\right)^{\frac{1}{\kappa}} \leq \left(\sum_{i=2}^{k}2\beta^{\kappa(k-i)}\mathbb{E}\left[\|\theta^i\|_2^\kappa\right]\right)^{\frac{1}{\kappa}}.$$

For each $i \in \overline{2,T}$, we estimate $\mathbb{E}\left[\|\theta^i\|_2^\kappa\right]$ as

$$\mathbb{E}\left[\|\theta^i\|_2^\kappa\right] \overset{(4)}{\leq} \mathbb{E}\left[\|\theta^i\|_\kappa^\kappa\right] = \mathbb{E}\left[\sum_{j=1}^{d}|g_j^k - \nabla f(x^k)_j|^\kappa\right] \overset{As. 3}{\leq} \sum_{j=1}^{d}\sigma_j^\kappa = \|\vec{\sigma}\|_\kappa^\kappa. \tag{18}$$

We continue bounding (18) with

$$(18) \leq \left(\sum_{i=2}^{k}2\beta^{\kappa(k-i)}\|\vec{\sigma}\|_\kappa^\kappa\right)^{\frac{1}{\kappa}} \leq \frac{2\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}}.$$

Therefore, the final math expectation can be calculated as:

$$\mathbb{E}\|\epsilon^k\|_2 \;\leq\; \beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 + \frac{L\sqrt{d}\gamma}{1-\beta} + \frac{2(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}}. \tag{19}$$

Now, we can update the step using Lemma 4:

$$\tag{20}$$

Then take math expectation:

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &\leq -\gamma\|\nabla f(x^k)\|_1 + 2\gamma\sqrt{d}\|\epsilon\|_2 + \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma_k}{2}\cdot\lambda\|\nabla f(x^k)\|_1 \\
&\leq -\gamma\|\nabla f(x^k)\|_1 + 2\gamma\sqrt{d}\cdot \\
&+ \left[\beta^{k-1}\|\epsilon^1\|_2 + L_0\lambda\sum_{i=2}^{k}\beta^{k-i+1} + L_1\lambda\sum_{i=2}^{k}\beta^{k-i+1}\|\nabla f(x^i)\| + (1-\beta)\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2\right] \\
&+ \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\cdot\lambda\|\nabla f(x^k)\|_1 \\
&\leq -\gamma\|\nabla f(x^k)\|_1 \\
&+ 2\gamma\sqrt{d}\left[\beta^{k-1}\|\epsilon^1\|_2 + L_0\lambda\frac{1}{1-\beta} + L_1\lambda\sum_{i=2}^{k}\beta^{k-i+1}\|\nabla f(x^i)\| + (1-\beta)\|\sum_{i=2}^{k}\beta^{k-i}\theta^i\|_2\right] \\
&+ \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\cdot\lambda\|\nabla f(x^k)\|_1
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[f(x^{k+1})] - \mathbb{E}[f(x^k)] &\leq -\gamma\mathbb{E}[\|\nabla f(x^k)\|_1] + 2\gamma\sqrt{d}\beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 \\
&+ L_0\lambda\frac{2\gamma\sqrt{d}}{1-\beta} + L_1\lambda 2\gamma\sqrt{d}\sum_{i=2}^{k}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^k)\|_1 + \frac{4\gamma\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}} \\
&+ \frac{L_0\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\lambda\mathbb{E}\|\nabla f(x^k)\|_1
\end{aligned} \tag{21}$$

Summing it over k, we derive

$$\begin{aligned}
f^* - f(x^1) &\leq -\gamma\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 + 2\gamma T\sqrt{d}\beta^{k-1}\mathbb{E}\|\epsilon^1\|_2 \\
&+ L_1\lambda 2\gamma\sqrt{d}\sum_{k=1}^{T}\sum_{i=2}^{k}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^i)\|_1 + \frac{4\gamma T\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{\frac{1}{\kappa}}} \\
&+ \frac{L_0 T\sqrt{d}\gamma}{2}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\lambda\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1
\end{aligned} \tag{22}$$

Changing the order of summation in the right part of the (23) we obtain:

$$\begin{aligned}
2\gamma L_1\lambda\sqrt{d}\sum_{k=1}^{T}\left(\sum_{i=2}^{k}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^i)\|_1\right) &= \\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\left(\sum_{k=i}^{T}\beta^{k-i+1}\mathbb{E}\|\nabla f(x^i)\|_1\right) &= \\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\beta^{-i}\left(\sum_{k=i}^{T}\beta^{k+1}\right)\mathbb{E}\|\nabla f(x^i)\|_1 &= \\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\beta^{-i+1}\beta^{i}\left(\frac{1-\beta^{T-i}}{1-\beta}\right)\mathbb{E}\|\nabla f(x^i)\|_1 &\leq \\
2\gamma L_1\lambda\sqrt{d}\sum_{i=2}^{T}\beta\left(\frac{1}{1-\beta}\right)\mathbb{E}\|\nabla f(x^i)\|_1 &
\end{aligned}$$

Finally

$$\begin{aligned}
f^* - f(x^1) &\leq -\gamma\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 + \frac{2\gamma\sqrt{d}\mathbb{E}\|\epsilon^1\|_2}{1-\beta} \\
&+ 2\gamma L_1\lambda\sqrt{d}\cdot\frac{\beta}{1-\beta}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 + \frac{4\gamma T\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{1/\kappa}} \\
&+ \frac{L_0T\sqrt{d}\gamma}{2(1-\beta)}\lambda + \frac{L_1\sqrt{d}\gamma}{2}\lambda\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1
\end{aligned}\tag{23}$$

$$\begin{aligned}
&\leq \left(-\gamma + \frac{2\gamma L_1\lambda\sqrt{d}\beta}{1-\beta} + \frac{L_1\sqrt{d}\gamma}{2}\lambda\right)\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 \\
&+ \frac{2\gamma\sqrt{d}\mathbb{E}\|\epsilon^1\|_2}{1-\beta} + \frac{4\gamma T\sqrt{d}(1-\beta)\|\vec{\sigma}\|_\kappa}{(1-\beta^\kappa)^{1/\kappa}} + \frac{L_0T\sqrt{d}\gamma}{2(1-\beta)}\lambda.
\end{aligned}\tag{24}$$

Let us set stepsize γ such that

$$\frac{2\gamma L_1\lambda\sqrt{d}\beta}{1-\beta} + \frac{L_1\sqrt{d}\gamma}{2}\lambda \leq \frac{3\gamma^2 L_1 d\exp(L_1 d\gamma)}{1-\beta} \leq \gamma/2 \Rightarrow \gamma \leq \frac{1-\beta}{8}\frac{1}{L_1 d}.$$

Thus, we obtain

$$f^* - f(x^1) \leq -\frac{\gamma}{2}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 + \frac{2\gamma\sqrt{d}\mathbb{E}\|\epsilon^1\|_2}{1-\beta} + 4\gamma T\sqrt{d}(1-\beta)^{\frac{\kappa-1}{\kappa}}\|\vec{\sigma}\|_\kappa + \frac{L_0Td\gamma^2}{(1-\beta)},$$

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 \leq \frac{2(f^* - f(x^1))}{\gamma T} + \frac{4\sqrt{d}\mathbb{E}\|\epsilon^1\|_2}{T(1-\beta)} + 8\sqrt{d}(1-\beta)^{\frac{\kappa-1}{\kappa}}\|\vec{\sigma}\|_\kappa + \frac{2L_0d\gamma}{(1-\beta)}.\tag{25}$$

Case $\varepsilon \geq \frac{3L_0}{L_1 c}$**.** We choose the stepsize $\gamma = \frac{1-\beta}{8}\frac{1}{L_1 dc} \leq \frac{1-\beta}{8}\frac{1}{L_1 d}$ and get:

$$\begin{aligned} \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 &\leq \frac{16c\Delta_1 L_1 d}{T(1-\beta)} + \frac{4\sqrt{d}\mathbb{E}\|\epsilon^1\|_2}{T(1-\beta)} + 8\sqrt{d}(1-\beta)^{\frac{\kappa-1}{\kappa}}\|\vec{\sigma}\|_\kappa + \frac{4L_0}{L_1 c} \\ &\leq \frac{16c(\Delta_1 L_1 + \mathbb{E}\|\epsilon^1\|_2)d}{T(1-\beta)} + 8\sqrt{d}(1-\beta)^{\frac{\kappa-1}{\kappa}}\|\vec{\sigma}\|_\kappa + \frac{4\varepsilon}{3}. \end{aligned}$$

Then, we choose $1-\beta = \min\left\{1, \left(\frac{c\Delta_1 L_1\sqrt{d}}{T\|\vec{\sigma}\|_\kappa}\right)^{\frac{\kappa}{2\kappa-1}}\right\}$ to obtain

$$\min_{\beta\in[0,1)}\left[\frac{16c\Delta_1 L_1 d}{T(1-\beta)} + 8\sqrt{d}(1-\beta)^{\frac{\kappa-1}{\kappa}}\|\vec{\sigma}\|_\kappa\right] \leq 24\sqrt{d}\left(\frac{c\Delta_1 L_1\sqrt{d}}{T}\right)^{\frac{\kappa-1}{2\kappa-1}}\|\vec{\sigma}\|_\kappa^{\frac{\kappa}{2\kappa-1}} + \frac{24c\Delta_1 L_1 d}{T}. \quad (26)$$

Finally, we choose number of iterations T to get:

$$24\sqrt{d}\left(\frac{c\Delta_1 L_1\sqrt{d}}{T}\right)^{\frac{\kappa-1}{2\kappa-1}}\|\vec{\sigma}\|_\kappa^{\frac{\kappa}{2\kappa-1}} + \frac{24c\Delta_1 L_1 d}{T} \leq \varepsilon \Rightarrow T = O\left(\frac{c\Delta_1 L_1 d}{\varepsilon}\left(1 + \left(\frac{\sqrt{d}\|\vec{\sigma}\|_\kappa}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right)\right).$$

Case $\varepsilon \leq \frac{3L_0}{L_1}$**.** We choose stepsize $\gamma = \sqrt{\frac{\Delta_1(1-\beta)}{TL_0 d}}$ to minimize the sum

$$\min_{\gamma}\left[\frac{2(f^* - f(x^1))}{\gamma T} + \frac{2L_0 d\gamma}{(1-\beta)}\right] = 4\sqrt{\frac{\Delta_1 L_0 d}{T(1-\beta)}},$$

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x^k)\|_1 \leq \frac{4\sqrt{d}\mathbb{E}\|\epsilon^1\|_2}{T(1-\beta)} + 4\sqrt{\frac{\Delta_1 L_0 d}{T(1-\beta)}} + 8\sqrt{d}(1-\beta)^{\frac{\kappa-1}{\kappa}}\|\vec{\sigma}\|_\kappa. \quad (27)$$

The first term is much smaller than the second one, hence we omit it. Next, we choose $1-\beta = \min\left\{1, \left(\frac{\Delta_1 L_0}{T\|\vec{\sigma}\|_\kappa^2}\right)^{\frac{\kappa}{3\kappa-2}}\right\}$ to minimize the last two terms:

$$\min_{\beta\in[0,1)}\left[4\sqrt{\frac{\Delta_1 L_0 d}{T(1-\beta)}} + 8\sqrt{d}(1-\beta)^{\frac{\kappa-1}{\kappa}}\|\vec{\sigma}\|_\kappa\right] \leq 12\sqrt{d}\left(\frac{\Delta_1 L_0}{T}\right)^{\frac{\kappa-1}{3\kappa-2}}\|\vec{\sigma}\|_\kappa^{\frac{\kappa}{3\kappa-2}} + 12\sqrt{\frac{\Delta_1 L_0 d}{T}}.$$

Finally, we choose number of iterations T to satisfy:

$$12\sqrt{d}\left(\frac{\Delta_1 L_0}{T}\right)^{\frac{\kappa-1}{3\kappa-2}}\|\vec{\sigma}\|_\kappa^{\frac{\kappa}{3\kappa-2}} + 12\sqrt{\frac{\Delta_1 L_0 d}{T}} \leq \frac{\varepsilon}{2} \Rightarrow T = O\left(\frac{\Delta_1 L_0 d}{\varepsilon^2}\left(1 + \left(\frac{\sqrt{d}\|\vec{\sigma}\|_\kappa}{\varepsilon}\right)^{\frac{\kappa}{\kappa-1}}\right)\right). \quad (28)$$

We only need to check that

$$\gamma = \sqrt{\frac{\Delta_1(1-\beta)}{TL_0 d}} = \sqrt{\frac{\Delta_1 L_0 d}{T(1-\beta)}}\frac{(1-\beta)}{L_0 d} \leq \frac{\varepsilon}{2\cdot 12}\frac{(1-\beta)}{L_0 d} \overset{\varepsilon\leq\frac{3L_0}{L_1}}{\leq} \frac{(1-\beta)}{L_1 d}.$$

□