# Sign SGD with Heavy-Tailed Noise and Differential Privacy

Alexey Kravatskiy
kravtskii.aiu@phystech.edu

Anton Plusnin
plusnin.aa@phystech.edu

Savelii Chezhegov
chezhegov.sa@phystech.edu

March 12, 2025

**Abstract**

In the era of large models, federated learning has become indispensable like never before. Sound modern federated learning must meet three key requirements. First, the method must process correctly real-world data, which, in case of Large Language Models, means the algorithm must tolerate noise with unbounded variance. Second, to ensure applicability, the algorithm must converge under these conditions with high probability, not only in expectation. Third, the whole procedure must not jeopardize user data. To address these natural requirements, we have constructed a novel modification of Sign version of Stochastic Gradient Descent. In this paper, we demonstrate that it meets all three earlier stated requisites. We start with proving algorithm's high-probability convergence on data with heavy-tailed noise. Then, we prove its differential privacy. Finally, we show the superior performance of the algorithm in training Large Language Models.

**Keywords:** Sign SGD, differential privacy, high-probability convergence, federated learning, heavy-tailed noise.

**Highlights:**

1. Sign Stochastic Gradient Descent can be used to train LLMs on real data.
2. Our modification of Sign Stochastic Gradient Descent keeps user data private.
3. Our modification of Sign Stochastic Gradient Descent does not require tuning.

## 1 Introduction

We start with formal statement of the problem

## 2 Problem statement.

**Federated optimization problem.** We consider the federated optimization problem in machine learning. $I = X \times Y$ is a sample space, where $X$ is a space of feature vectors and $Y$ is a label space. For the hypothesis space $\mathcal{W} \subseteq \mathbb{R}^d$, a loss function is defined as $l : \mathcal{W} \times I \to \mathbb{R}$ which measures the loss of the prediction on the data point $(x, y) \in I$ based on the hypothesis $w \in \mathcal{W}$. For a dataset $D \subset I$, the global loss function $F : \mathcal{W} \to \mathbb{R}$ is defined as

$$F(w) = \frac{1}{|D|} \sum_{(x,y)\in D} l(w; (x, y)). \tag{1}$$

In case of distributed optimization, the dataset is split between $M$ workers. Each worker $m$ has a local dataset $D_m \subset I$ and a local function $f_m : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f_m(w) = \frac{1}{|D_m|} \sum_{(x_n, y_n) \in D_m} l(w; (x_n, y_n)), \tag{2}$$

where $|D_m|$ is the size of worker $m$'s local dataset $D_m$.

Thus, our goal is to solve the following federated optimization problem:

$$\min_{w \in \mathbb{R}^d} F(w) \quad \text{where} \quad F(w) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} f_m(w). \tag{3}$$

We assume that the data are distributed over the workers uniformly, consequently, $\mathbb{E}[f_m(w)] = F(w)$ for workers' data distribution.

**Stochastic optimization problem.** The stochastic optimization problem for a smooth non-convex function $f : \mathbb{R}^d \to \mathbb{R}$ is:

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{\xi \sim \mathcal{S}}[f(x, \xi)], \tag{4}$$

where random variable $\xi$ is sampled from an unknown distribution $\mathcal{S}$. The gradient oracle returns unbiased gradient estimate $\nabla f(x, \xi) \in \mathbb{R}^d$. In machine learning, for instance, $f(x, \xi)$ is a loss function on a sample $\xi$ [3].

The most popular algorithm to solve (1) is Stochastic Gradient Descent (SGD) [2]:

$$x^{k+1} = x^k - \gamma_k \cdot g^k, \quad g^k := \nabla f(x^k, \xi^k).$$

For non-convex functions, the algorithm must stop at the point with sufficiently small gradient norm.

**Differential privacy.** Additionally, the algorithm must be private, which means it must satisfy $(\epsilon, \delta)$-local differential privacy [1].

# 3 Theory

Present your theoretical framework, definitions, lemmas, and proofs.

# 4 Experiments

Describe your experimental setup, methodology, and results.

# 5 Conclusion

Summarize your findings and discuss future work.

# 6 Acknowledgments

Optional acknowledgments section.

# A Additional Proofs and Results

Include detailed proofs and supplementary materials here.

# References


[1] C. Dwork, A. Roth, et al. "The algorithmic foundations of differential privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[2] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.

[3] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.