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

Alexander Beznosikov
beznosikov.an@phystech.edu

March 13, 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

Federated Learning is a useful method to train models that require large amounts of data. Indeed, it is often the case that the data is distributed across multiple devices, like mobile phones [8], and it is not only costly to collect all the data in one place, but often also unacceptable due to the requirements of privacy. On the other hand, training the model only on the local data for a particular user is impossible, as the model is large. Hence, a need for a joint training procedure arises. We come to a setting with a server and a number of workers, where each worker has a local dataset. The goal is to train a model on the data from all the workers, without sharing the data between the workers or with the server.

The most obvious way to train a model in a federated setting is to use Stochastic Gradient Descent (SGD) [11] by passing the gradient to the server (add citation). However, when trasmitting the gradient itself, the communication cost is unaffordably high and the privacy of the data is compromised. To address this issue, one can use the Sign Stochastic Gradient Descent algorithm [1]. This algorithm transmits only the signs of the coordinates of the gradients, which is much cheaper in terms of communication.

Clipping the norm of gradient estimate before SGD step, which comprises the method called ClipSGD, is another great idea that demonstrates great empirical results [10, 4]. However, the clipping requires meticulous tuning of the clipping threshold, which was shown to depend on not only iteration number but also the

objective function and the noise [12, Theorem 3.1]. In case of federative learning, the search of the threshold is even more complicated, as the data is distributed across the workers and cannot be used for tuning, and the objective function is more complex due to sophistication and size of the model.

The natural simplification of the clipping is normalization of the gradient. SGD with additional normalization is called NSGD [5]. NSGD outperformed ClipSGD on the task of sequence labeling via LSTM Language Models [9]. However, NSGD, unlike ClipSGD requires larged batch sizes for convergence. This can be mended by applying momentum [2]. The major drawback of NSGD is the absence of proofs of convergence with high probability. Moreover, the method seems to be not private.

As to guarantee the privacy is our priority, we opted to base our algorithm on SignSGD. The convergence with high probability of SignSGD for the heavy-tailed noise was recently proved [7], which makes SignSGD a perfect candidate for federated learning. Simultaneously, SignSGD was already used as a base to create a differentially private algorithm [6]. However, in both cases, no proofs for the convergence of the algorithm in the modern federated learning setting were provided.

In this paper, we propose a modification of SignSGD that can be used to train LLMs on real data. We show that our algorithm converges with high probability, even in the presence of heavy-tailed noise, and does not require tuning. We also show that our algorithm is differentially private. Finally, we test the algorithm on the training of LLMs.

2 Problem statement.

First, we need to state that we work with stochastic optimization.

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

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{\xi \sim \mathcal{S}}[f(x, \xi)], \tag{1}$$

where random variable ξ 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 ξ [13].

The most popular algorithm to solve (1) is Stochastic Gradient Descent (SGD) [11]:

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

For non-convex functions, the algorithm must stop at the point with sufficiently small gradient norm. We will apply the algorithm to the federated optimization problem.

Federated optimization problem. Let $I = X \times Y$ be 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

$$F(w) = \frac{1}{|D|} \sum_{(x,y) \in D} l(w; (x, y)). \tag{2}$$

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

$$f_m(w) = \frac{1}{|D_m|} \sum_{(x_n, y_n) \in D_m} l(w; (x_n, y_n)), \tag{3}$$

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:

$$\min_{w \in \mathbb{R}^d} F(w) \quad \text{where} \quad F(w) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} f_m(w). \tag{4}$$

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

Now, let us introduce the requirements for the algorithm to solve the federated optimization problem.

Heavy-tailed noise. Noise has bounded κ-th moment for some $\kappa \in (1, 2]$, i.e. $\mathbb{E}_\xi[\|\nabla f(x, \xi) - \nabla f(x)\|_2^\kappa] \leq \sigma^\kappa$. In particular, the noise can have unbounded variance, i.e. $\kappa < 2$.

Differential privacy. Additionally, the algorithm must be private, which means it must satisfy (ϵ, δ)-local differential privacy [3].

High-probability convergence. The algorithm must have convergence guarantees which hold true with probability at least $1 - \delta, \delta \in (0, 1)$.

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] Jeremy Bernstein et al. "signSGD: Compressed Optimisation for Non-Convex Problems". In: *International Conference on Machine Learning*. 2018, pp. 560–569.

[2] Ashok Cutkosky and Harsh Mehta. "Momentum improves normalized sgd". In: *International conference on machine learning*. PMLR. 2020, pp. 2260–2268.

[3] 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.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[5] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. "Beyond convexity: Stochastic quasi-convex optimization". In: *Advances in Neural Information Processing Systems*. 2015, pp. 1594–1602.

[6]   Richeng Jin et al. "Stochastic-Sign SGD for Federated Learning with Theoretical Guarantees". In: *Part of this work is published in IEEE Transactions on Neural Networks and Learning Systems, 2024* 36.2 (Feb. 25, 2020), pp. 3834–3846. ISSN: 2162-2388. DOI: `10.1109/tnnls.2023.3345367`. arXiv: `2002.10940 [cs.LG]`.

[7]   Nikita Kornilov 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`.

[8]   Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Artificial Intelligence and Statistics*. 2017, pp. 1273–1282.

[9]   Stephen Merity, Nitish Shirish Keskar, and Richard Socher. "Regularizing and optimizing LSTM language models". In: *arXiv preprint arXiv:1708.02182* (2017).

[10]   Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *International conference on machine learning*. 2013, pp. 1310–1318.

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

[12]   Abdurakhmon Sadiev et al. "High-Probability Bounds for Stochastic Optimization and Variational Inequalities: the Case of Unbounded Variance". In: *arXiv preprint arXiv:2302.00999* (2023).

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