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

April 23, 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 and foremost, the whole process of learning must not jeopardize user data, hence be private to given extent. Second, the algorithm must correctly process real-world data, which, in case of Large Language Models (LLMs), means it must tolerate noise with unbounded variance. Third, to ensure applicability, the algorithm must converge under these conditions with high probability, not only in expectation. 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 defining a construction procedure which guarantees differential privacy. Then, we prove algorithm's high-probability convergence on data with heavy-tailed noise. Finally, we show the superior performance of the algorithm in training LLMs in distributed setting.

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

Idea: We add Gaussian noise and use privacy amplification by subsampling to make sign-SGD private.

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 is applicable to federated setting.

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 [**McMahan2017**], and it is not only costly to collect all the data in one place, but also often 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 and data of one user is insufficient. 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) [**Robbins1951**] by passing the gradient to the server. However, when trasmitting the gradient itself, the communication cost is unaffordably high, and again the data privacy might be violated. To address this issue, one can use the Sign Stochastic Gradient Descent (SignSGD)algorithm [**Bernstein2018**]. This algorithm transmits only the signs of the coordinates of the gradients, which is much cheaper in terms of communication.

Communication costs aside, clipping the norm of gradient estimate before SGD step, which comprises the method called ClipSGD, is an another great idea that demonstrates positive empirical results [**Pascanu2013**, **Goodfellow2016**]. Nonetheless, 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 [**Sadiev2023**]. In case of federated 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, which lies at the core of NSGD [**Hazan2015**]. NSGD outperformed ClipSGD on the task of sequence

labeling via LSTM Language Models [**Merity2017**]. Despite this fact, NSGD, unlike ClipSGD, requires larged batch sizes for convergence, which, however, can be mended by applying momentum [**Cutkosky2020**]. The major drawback of NSGD is the absence of proofs of convergence with high probability. Moreover, the method seems to be far from 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 has been recently proved [**Kornilov2025**], which makes SignSGD a perfect candidate for federated learning. Simultaneously, SignSGD has already been used as a base to create an allegedly differentially private algorithm [**Jin2020**] (the authors proved differential privacy of only one step of the algorithm). However, in both cases, no proofs for the convergence of the algorithm in the modern federated learning setting were provided.

In this paper, based on the concept of Rényi differential privacy, we develop the idea from [**Jin2020**] to derive truly differentially-private modification of SignSGD, which we naturally call DP-SIGNSGD. We show that our algorithm converges with high probability, even in the presence of heavy-tailed noise. 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 [**ShalevShwartz2014**].

The most popular algorithm to solve (1) is Stochastic Gradient Descent (SGD) [**Robbins1951**]:

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

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 when $\kappa < 2$.

Differential privacy. Additionally, the algorithm must be private. This is guaranteed by (ϵ, δ)-local differential privacy [**Dwork2014**]:

Definition 1. *Given a set of local datasets $\mathcal{D}$ provided with a notion of neighboring local datasets $\mathcal{N}_\mathcal{D} \subset \mathcal{D} \times \mathcal{D}$ that differ in only one data point. For a query function $f : \mathcal{D} \to \mathcal{X}$, a mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{O}$ to release the answer of the query is defined to be (ϵ, δ)-locally differentially private if for any measurable subset $\mathcal{S} \subseteq \mathcal{O}$ and two neighboring local datasets $(D_1, D_2) \in \mathcal{N}_\mathcal{D}$,*

$$P(\mathcal{M}(f(D_1)) \in \mathcal{S}) \leq e^\epsilon P(\mathcal{M}(f(D_2)) \in \mathcal{S}) + \delta. \tag{5}$$

A key quantity in characterizing local differential privacy for many mechanisms is the sensitivity of the query f in a given norm l_r, which is defined as

$$\Delta_r = \max_{(D_1,D_2)\in\mathcal{N}_\mathcal{D}} ||f(D_1) - f(D_2)||_r. \tag{6}$$

High-probability convergence. For given $\delta_{prob} \in (0, 1)$, the algorithm must converge with probability at least $1 - \delta_{prob}$.

3 Theory

3.1 The Algorithm and the compressor

The algorithm we are working with can be defined generally as follows:

Algorithm 1 Stochastic-Sign SGD with majority vote

Input: learning rate η, current hypothesis vector $w^{(t)}$, M workers each with an independent gradient $\boldsymbol{g}_m^{(t)}$, the 1-bit compressor $q(\cdot)$.
on server:
pull $q(\boldsymbol{g}_m^{(t)})$ **from** worker m.
push $\tilde{\boldsymbol{g}}^{(t)} = sign\Big(\frac{1}{M}\sum_{m=1}^{M} q(\boldsymbol{g}_m^{(t)})\Big)$ **to** all the workers.
on each worker:
update $w^{(t+1)} = w^{(t)} - \eta\tilde{\boldsymbol{g}}^{(t)}$.

Our goal is to find such compressor $q(\cdot)$ that the algorithm is private.

3.2 Incorrectness of earlier proposed DP-SIGN

Some time ago there was proposed a 1-bit compressor dp-sign [**Jin2020**]:

Definition 2. *For any given gradient $\boldsymbol{g}_m^{(t)}$, the compressor dp-sign outputs dp-sign$(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)$. The i-th entry of dp-sign$(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)$ is given by*

$$dp\text{-}sign(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)_i = \begin{cases} 1, & \text{with probability } \Phi\Big(\frac{(\boldsymbol{g}_m^{(t)})_i}{\sigma}\Big) \\ -1, & \text{with probability } 1 - \Phi\Big(\frac{(\boldsymbol{g}_m^{(t)})_i}{\sigma}\Big) \end{cases} \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the normalized Gaussian distribution; $\sigma = \frac{\Delta_2}{\epsilon}\sqrt{2\ln(\frac{1.25}{\delta})}$, where ϵ and δ are the differential privacy parameters and Δ_2 is the sensitivity measure.

As stated in Theorem 5 from [**Jin2020**], the *dp-sign* mechanism is (ϵ, δ)-differentially private for any ϵ and $\delta \in (0, 1)$.

Theorem 6 from [**Jin2020**] should establish probability of the dissimilarity of majority sign of the gradients and majority sign of the dp-signs:

Theorem 1. *Let $u_1, u_2, \cdots, u_M$ be M known and fixed real numbers. Further define random variables $\hat{u}_i = dp\text{-}sign(u_i, \epsilon, \delta), \forall 1 \leq i \leq M$. Then there always exist a constant σ_0 such that when $\sigma \geq \sigma_0$, $P(sign(\frac{1}{M}\sum_{m=1}^{M}\hat{u}_i) \neq sign(\frac{1}{M}\sum_{m=1}^{M}u_i)) < \left(1 - x^2\right)^{\frac{M}{2}}$, where $x = \frac{|\sum_{m=1}^{M}u_m|}{2\sigma M}$.*

However, as we recently found out, the theorem is fundamentally flawed, as it makes σ a parameter, while it follow from the definition of the *dp-sign* compressor that σ is a function of ϵ and δ.

Let us suppose that there exist ϵ and δ such that for $\sigma(\epsilon, \delta)$, the inequality on probability from the theorem holds true. Then, let us walk through the proof to find a precise lower bound for σ. If this bound is lower than $\sigma(\epsilon, \delta)$, the theorem is indeed correct.

In the proof of Theorem 6, the authors needed to find the lower bound for the following expression:

$$\frac{1}{\sqrt{2\pi}\sigma}\left[\left|\sum_{m=1}^{M}u_m\right|e^{-\frac{u_1^2}{2\sigma^2}} + \left|\sum_{m=2}^{M}u_m\right|\left[e^{-\frac{(\sum_{m=2}^{M}u_m)^2}{2\sigma^2}} - 1\right]\right]$$

Instead of taking limit $\sigma \to \infty$ as they did, we will use the well known relation $e^{-x^2} \geq 1 - x^2$. After applying it to the earlier mentioned expression and making some trivial transformations, we get the following:

$$\sigma^2 \geq \frac{7}{5}\left(u_1^2 + \left|\sum_{m=1}^{M}u_m\right|^2 + \frac{|u_1|^3}{\left|\sum_{m=1}^{M}u_m\right|}\right).$$

It is a sufficient condition, not a necessary one. However, it reflects the key features of the condition on σ. $\frac{|u_1|^3}{\left|\sum_{m=1}^{M}u_m\right|}$ is an extremely unreliable term, as $|u_1|$ may be high, especially for heavy-tailed noise, while $\left|\sum_{m=1}^{M}u_m\right|$ may be small (it is easy to construct an appropriate example for the case of 3 workers). Hence, the bound on σ is not only high, but also unstable. Consequently, we have no proofs whatsoever of the convergence of the algorithm.

Moreover, our misgivings are supported by the fact that in an updated version [**Jin2024**] of the article [**Jin2020**], the authors have removed all mentions of dp-sign.

The problems with convergence of dp-sign are not limited by the incorrect Theorem 6. The authors have overlooked another issue of utmost importance. Proving the (ϵ, δ) differential privacy of dp-sign, they did so for only one call of dp-sign operator. However,

according to [**Dwork2014**] (Theorem 3.16), the overall privacy of running dp-sign T times is $(T\epsilon, T\delta)$. Consequently, when we set target $\epsilon \approx 10$, $\delta \approx 10^{-}5$, as is recommended for machine learning purposes [**Ponomareva2023**], for $T = 100$ we get 100 times lower ϵ and δ for dp-sign. Judging by our experiments, this destroys convergence. That is intuitive: high noise for low (ϵ, δ) leads to poorer convergence, and this forces us to use still lower (ϵ, δ) to preserve the required level of privacy during the run of the algorithm.

To the best of our knowledge, there are no differentially private versions of sign-sgd reported except this incorrect dp-sign. As Gaussian noise is still the best mechanism to ensure differential privacy (Laplace noise and exponential mechanism complicate convergence), we preserve the idea of dp-sign. However, we drastically change the privacy analysis to make σ small enough for the algorithm to converge.

3.3 Rényi Differential privacy

To make an economical use of differential privacy, we have to introduce another type of differential privacy, which gives more tight bounds on composition.

Definition 3 (Rényi divergence [**Mironov2017**])**.** *Let P and Q be two distributions on $\mathcal{X}$ defined over the same probability space, and let p and q be their respective densities. The Rényi divergence of a finite order $\alpha \neq 1$ between P and Q is defined as*

$$\mathrm{D}_\alpha(P \parallel Q) \triangleq \frac{1}{\alpha - 1} \ln \int_{\mathcal{X}} q(x) \left(\frac{p(x)}{q(x)} \right)^\alpha \mathrm{d}x.$$

Rényi divergence at orders $\alpha = 1, \infty$ are defined by continuity.

Definition 4 (Rényi differential privacy (RDP) [**Mironov2017**])**.** *We say that a randomized mechanism $\mathcal{M} \colon \mathcal{S} \to \mathcal{R}$ satisfies (α, ε)-Rényi differential privacy (RDP) if for any two* adjacent *inputs $S, S' \in \mathcal{S}$ it holds that*

$$\mathrm{D}_\alpha(\mathcal{M}(S) \parallel \mathcal{M}(S')) \leq \varepsilon.$$

The notion of adjacency between input datasets is domain-specific and is usually taken to mean that the inputs differ in contributions of a single individual. In this work, we will use this definition and call two datasets S, S' to be adjacent if $S' = S \cup \{x\}$ for some x (or vice versa).

Definition 5 (Sampled Gaussian Mechanism (SGM) [**mironov2019SGM**])**.** *Let f be a function mapping subsets of $\mathcal{S}$ to $\mathbb{R}^d$. We define the Sampled Gaussian mechanism (SGM) parameterized with the sampling rate $0 < q \leq 1$ and the noise $\sigma > 0$ as*

$$\mathrm{SG}_{q,\sigma}(S) \triangleq f(\{x\colon x \in S \text{ is sampled with probability } q\}) + \mathcal{N}(0, \sigma^2 \mathbb{I}^d),$$

where each element of S is sampled independently at random with probability q without replacement, and $\mathcal{N}(0, \sigma^2 \mathbb{I}^d)$ is spherical d-dimensional Gaussian noise with per-coordinate variance σ^2.

From Section 3.3 (*"Numerically stable computations"*) of [**mironov2019SGM**], for each integer $\alpha \geq 1$, SGM applied to a function with $\Delta_2 \leq 1$ is (α, ε_R)-RDP for the following ε_R:

$$\varepsilon_R = \frac{1}{\alpha - 1} \log \left(\sum_{k=0}^{\alpha} \binom{\alpha}{k} (1-q)^{\alpha-k} q^k \exp\left(\frac{k^2 - k}{2\sigma^2} \right) \right) \tag{8}$$

Clearly, if $\Delta_2 > 1$ for a mechanism, $\sigma := \Delta_2 \sigma$ will deliver the same (α, ε_R)-RDP.

If an (α, ε_R)-RPD mechanism is applied T times, by adaptive sequential composition [**Mironov2017**], we get $(\alpha, T\varepsilon_R)$-RDP of the procedure. From *Proposition 3* from [**Mironov2017**], it follows that the procedure is also $(\varepsilon_R + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-differentially private for any $0 \leq \delta \leq 1$. Thus, the procedure is (ε, δ)-differntial privacy, if (α, ε_R) satisfy:

$$\varepsilon_R \leq \varepsilon / T - \frac{\log 1/\delta}{T(\alpha - 1)} \tag{9}$$

As convergence of the algorithm depends only on σ, while α is an internal parameter, we are interested in finding the minimal σ. This we achieve via grid search on integer $1 \leq \alpha \leq 20$ and $0.5 \leq \sigma \leq 3$. An example can be seen in Figure 1.

Judging by the chart, there is no need in searching among $\alpha \in \mathbb{R}$: *Case II. Fractional α* from [**mironov2019SGM**] gives a series expression to obtain a bound for ε_R, and the series is continuous on α. Continuous borderline on 1 implies that σ we get by the grid search is close to the optimum.

3.4 Constructing DP-SIGN

Let us define DP-SIGN 2. From the previous section, (ε, δ)-privacy of DP-SIGN immediately follows.

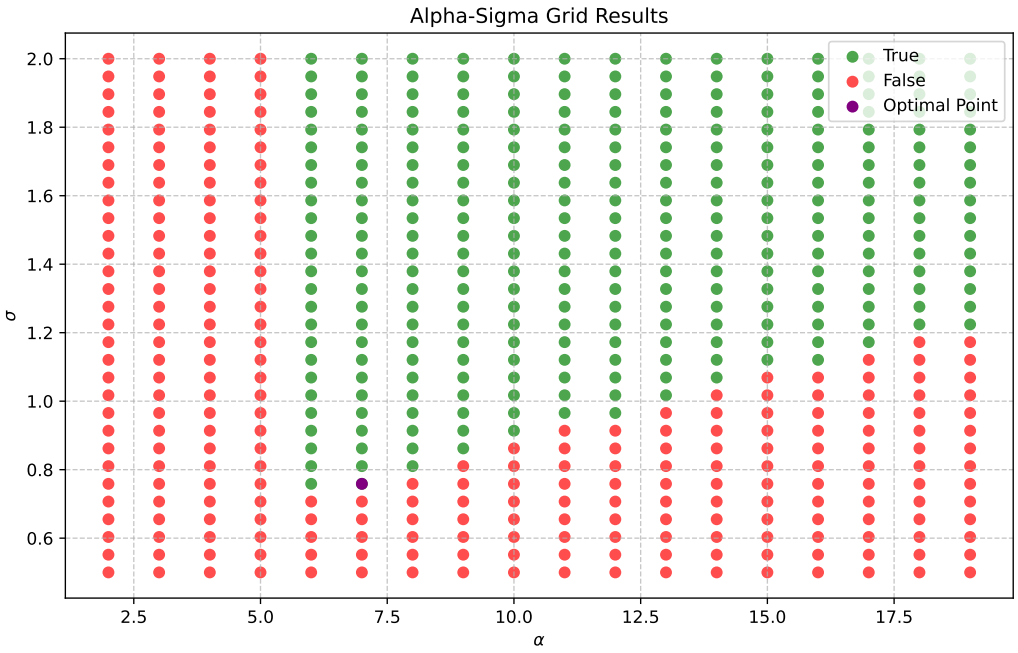

Figure 1: Finding the minimal σ for SGM with $\Delta_2 \leq 1$, sampling rate $q = 1/300$, and $T = 1000$

Theorem 2. *Let $l(\cdot;\cdot)$ be a loss function and let $\varepsilon > 0$, $\delta \in (0,1)$. Then applying T times dp-sign(w_i, l, D, (ε,δ), T, q, C) in different points w_i on a user database D with a fixed sampling rate q and fixed clipping level $C > 0$ is (ε,δ)-differentially private.*

Proof. $\boldsymbol{g}_{priv}$ is private due to the satisfaction of (8) and (9). As RDP preserves under post-processing ([**Mironov2017**]), $sign(\boldsymbol{g}_{priv})$ is also (α,ε_R)-RPD, which entails (ε,δ)-dp for all T iterations. □

4 Experiments

In this section, we present the experimental results for the methods we discussed in Sections 2 and 3. First, we appy the algorithms to a classic machine learning problem. Then, we test the algorithms on the training of Large Language Models.

Algorithm 2 DP-SIGN compressor

Input: coordinate w, loss function l, user database D, (ε, δ)-privacy requirement, number of iterations T, sampling rate q, clipping level C.
Prepare subsample S: add each element $(x, y) \in D$ with probability q.
Compute the gradient $\boldsymbol{g}$ of the subsample: $\frac{1}{|S|}\sum_{(x,y)\in S} l(w;(x,y))$. If S is empty, let $\boldsymbol{g} = 0$.

If $||\boldsymbol{g}||_2 > C$: $\boldsymbol{g} = C\frac{\boldsymbol{g}}{||\boldsymbol{g}||_2}$.
Grid search $\sigma(q, T, \varepsilon, \delta)$ to satisfy (8) and (9).
$\boldsymbol{g}_{priv} = \boldsymbol{g} + \mathcal{N}(0, (C\sigma)^2\mathbb{I}^d)$
Output: $sign(\boldsymbol{g}_{priv})$

4.1 Synthetic noise.

We test our algorithm on the method of logistic regression for UCI Mushroom Dataset. The dataset consists of 6,449 training samples and 1,625 testing samples. Each sample has $d = 112$ features, and represents a mushroom either poisonous or edible. We apply different algorithms to train the logistic regression model. The algorithms are run for $T = 100000$ iterations on 10 workers, with training data distributed between the workers equally. We compare the cases when there is no noise, when the noise is normal with $\sigma = 1/4$, and when the noise is Lévy stable with $\sigma_l = 1/4$, $\alpha_l = 1.6$, which corresponds to $\kappa = 1.5$, and $\beta_l = 0$ (this distribution is defined by its characteristic function $\varphi(t) = \exp\left(-0.25^{1.6}|t|^{1.6}\right)$). For private algorithms, we set $\epsilon \approx 10$, $\delta \approx 1/n^{1.1}$, where n is the size of local user database, as is recommended for machine learning purposes [**Ponomareva2023**]. We set the learning rate $\gamma = 1/\sqrt{dT}$ for SGD, sign-SGD, and DP-SIGNSGD. The results are presented in 2.

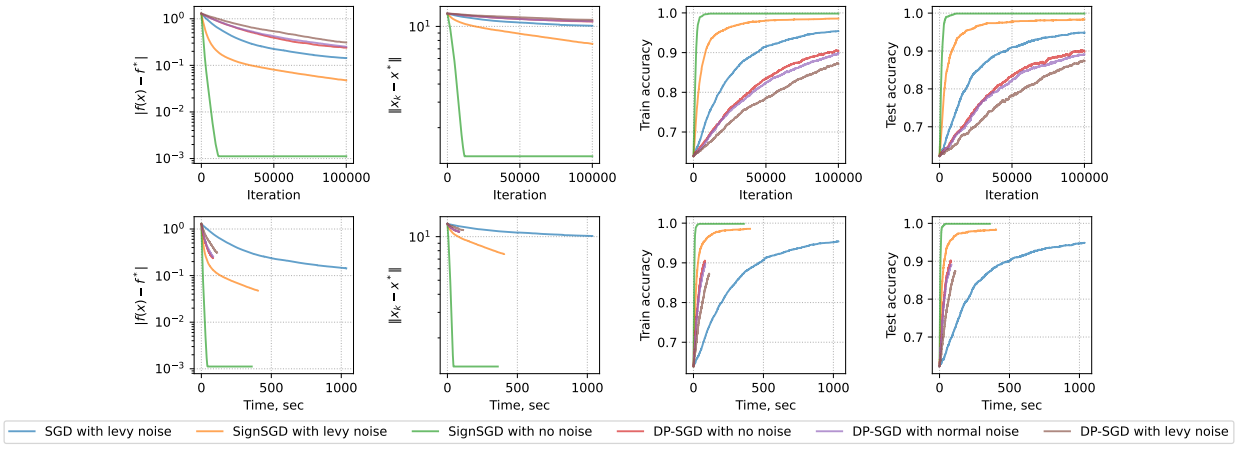

Figure 2: Logistic regression on UCI Mushroom Dataset.
Performance of SGD, SignSGD, and DP-SIGNSGD with majority voting.

As we see, DP-signSGD does converge, even under heavy-tailed noise.

4.2 DP-Sign (outdated)

Having verified Sign-SGD, we test dp-sign on the same dataset. We present three variants of the dp-sign:

1. DP-SignSGD from [**Jin2020**] with unrectified (ϵ, δ). The algorithm converges, but is not sufficiently private, as we highlighted earlier.

2. DP-SignSGD from [**Jin2020**] with rectified $(\epsilon/T, \delta/T)$, where T is 1000. The algorithm is now private, but does not converge.

3. Our modification of DP-SignSGD. We used Rényi differential privacy [**Dwork2014**]. This type of privacy can be easily converted to classical differential privacy, but has a more temperate deterioration of privacy during the sequential composition: the noise $\sigma \sim \sqrt{T}$ instead of $\sigma \sim T$. We will expand Theory and introduce there this algorithm, when it will be ready. For now, it as dp-sign does not converge.

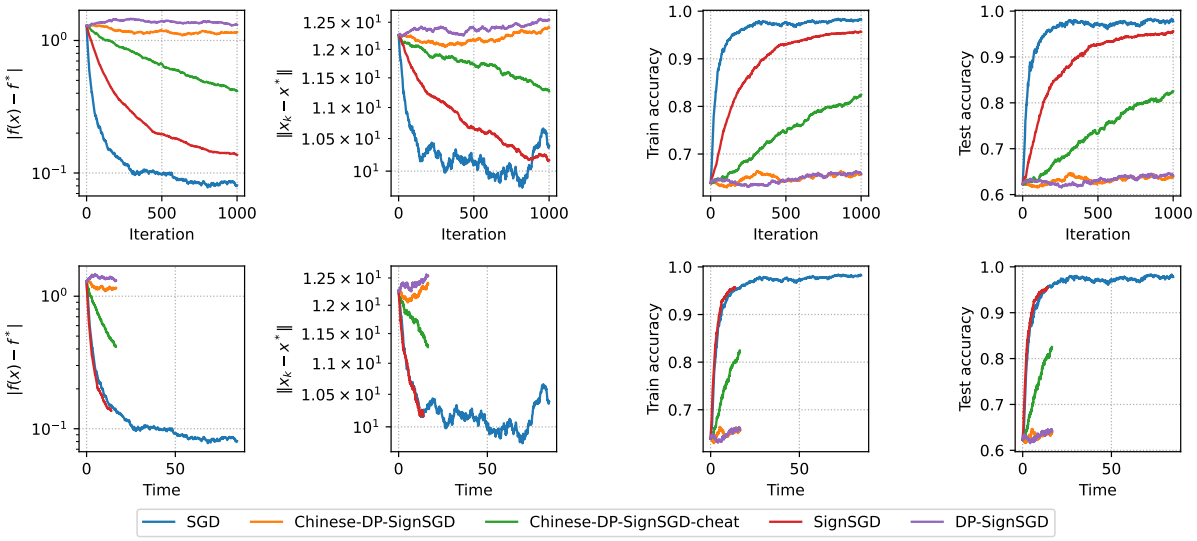

Figure 3: Logistic regression on UCI Mushroom Dataset.
Performance of SGD, SignSGD and different variants of DP-SignSGD with majority voting.

Thus, we have the following table 1.

Table 1: DP-Sign SGD methods

Method	Citation	Is (ϵ, δ) private	Converges
ChineseDP-SignSGD-cheat	[**Jin2020**]	No	Yes
ChineseDP-SignSGD rectified	[**Jin2020**]	Yes	No
Rényi DP-SignSGD	**This paper**	Yes	Yes

5 Conclusion

We have highlighted the deficiencies of the previous approach of working with dp-sign-SGD. With a use of Rényi differential privacy and advanced numerically-produced bounds for Sampled Gaussian Mechanism, we have consturcted a privacy accountant that makes dp-sign-SGD converge and private, which, to the best of our knowledge, has never been done before.

6 Acknowledgments

We will fill this section when we are done with the article.

A Additional Proofs

We will present them after we complete the tests and finalize the algorithm.