

Differentially private modification of SignSGD

Alexey Kravatskiy

Moscow Institute of Physics and Technology

Course: My first scientific paper
(Strijov's practice) & Innovative Practicum /Group 205

Expert: A. N. Beznosikov

Consultant: S. A. Chezhegov

2025

Distributed, Private, and Noise-resistant

Goal

A communication-efficient and private algorithm for distributed optimization converging under heavy-tailed noise (noise with unbounded variance).

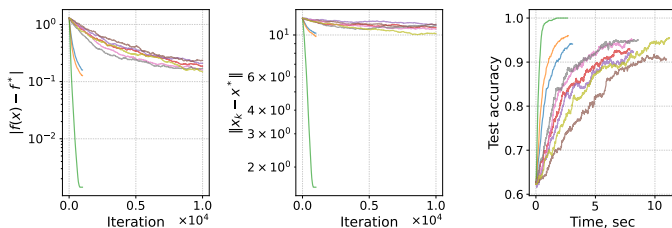
Problem

The only proposed private sign-based algorithm DP-SignSGD either is not private or does not converge.

Solution



Rényi differential privacy and Bernoulli subsampling afford lower noise to ensure convergence.

Differential privacy of DP-SignSGD



DP-SignSGD with our DP-SIGN compressor is (ϵ, δ) -private and converges under heavy-tailed noise. It behaves very much like DP-SGD with the same privacy mechanism.

Literature

-  Jin, Richeng et al. (Feb. 25, 2020). “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, pp. 3834–3846. ISSN: 2162-2388. DOI: 10.1109/tnnls.2023.3345367. arXiv: 2002.10940 [cs.LG].
-  Mironov, Ilya (Aug. 2017). “Rényi Differential Privacy”. In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE, pp. 263–275. DOI: 10.1109/csf.2017.11. URL: <http://dx.doi.org/10.1109/CSF.2017.11>.
-  Mironov, Ilya, Kunal Talwar, and Li Zhang (2019). *Rényi Differential Privacy of the Sampled Gaussian Mechanism*. arXiv: 1908.10530 [cs.LG]. URL: <https://arxiv.org/abs/1908.10530>.

Differential Privacy

Definition

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 $f : \mathcal{D} \rightarrow \mathcal{X}$, a mechanism $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{O}$ 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 $(D_1, D_2) \in \mathcal{N}_{\mathcal{D}}$,

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

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

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

Incorrect version of DP-SIGN (Jin et al. 2020)

Definition

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

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

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

Theorem 6 from (Jin et al. 2020)

Theorem

Let u_1, u_2, \dots, u_M be M known and fixed real numbers. Further define random variables $\hat{u}_i = \text{dp-sign}(u_i, \epsilon, \delta), \forall 1 \leq i \leq M$. Then there always exist a constant σ_0 such that when $\sigma \geq \sigma_0$,

$$P(\text{sign}(\frac{1}{M} \sum_{m=1}^M \hat{u}_i) \neq \text{sign}(\frac{1}{M} \sum_{m=1}^M u_i)) < (1 - x^2)^{\frac{M}{2}}, \text{ where}$$
$$x = \frac{|\sum_{m=1}^M u_m|}{2\sigma M}.$$

Fault From the proof, it follows that the constant σ_0 depends on the values of $\{u_i\}$, which precludes from constructing DP-SIGN compressor.

Fault The authors have not guaranteed the overall (ϵ, δ) -privacy.

Fault The authors have not proved the convergence of their DP-SignSGD.

Rényi divergence (Mironov 2017)

Definition (Rényi divergence)

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

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

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

Rényi differential privacy (Mironov 2017)

Definition (Rényi differential privacy (RDP))

We say that a randomized mechanism $\mathcal{M}: \mathcal{S} \rightarrow \mathcal{R}$ satisfies (α, ε) -Rényi differential privacy (RDP) if for any two *adjacent* inputs $S, S' \in \mathcal{S}$ it holds that

$$D_{\alpha}(\mathcal{M}(S) \parallel \mathcal{M}(S')) \leq \varepsilon.$$

Bernoulli sampling + Gaussian Mechanism (Mironov, Talwar, and Zhang 2019)

Definition (Sampled Gaussian Mechanism (SGM))

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

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

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 .

Criterion of a private algorithm

Following the procedure from Mironov, Talwar, and Zhang 2019, we get:

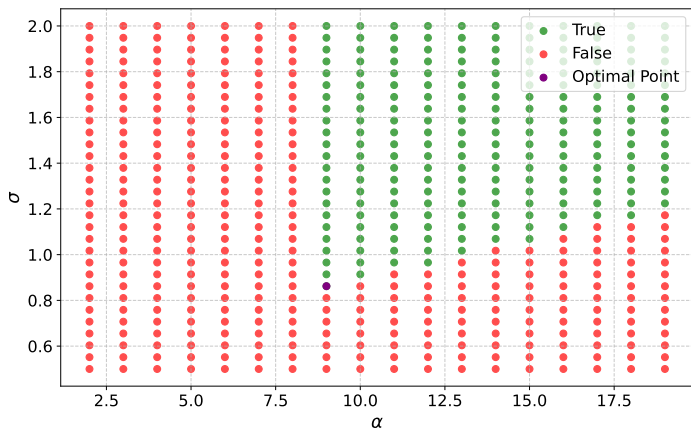
$$\varepsilon_R = \frac{1}{\alpha - 1} \log \left(\sum_{k=0}^{\alpha} \binom{\alpha}{k} (1 - q)^{\alpha - k} q^k \exp \left(\frac{k^2 - k}{2\sigma^2} \right) \right)$$

While according to the advanced composition theorem and conversion from Rényi privacy to (ε, δ) -privacy, ε_R must satisfy:

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

.

Grid Search to find minimal σ



(α, σ) that guarantee (ϵ, δ) -dp of T subsamplings for $\epsilon = 1$, $\delta = 1/n^{1.1}$, $T = 1000$, $q = 1/n$, where $n = 649$ (10% of the Mushroom dataset).

Our version of DP-SIGN compressor

Input: coordinate w , loss function l , user database D , (ε, δ) -privacy requirement, number of iterations T , sampling rate q .

Prepare subsample S : add each element $(x, y) \in D$ with probability q .

Compute the gradient \mathbf{g} of the subsample for $\frac{1}{|S|} \sum_{(x,y) \in S} l(w; (x, y))$. If S is empty, let $\mathbf{g} = 0$.

Grid search $\sigma(q, T, \varepsilon, \delta)$ to satisfy 2 expressions for ε_R stated earlier.

$$\text{sign}_{\text{noised}} = \text{sign}(\mathbf{g}) + \mathcal{N}(0, (2\sqrt{d}\sigma)^2 \mathbb{I}^d)$$

Output: $\text{sign}(\text{sign}_{\text{noised}})$

UCI Mushroom Dataset

There are 6,449 training samples equally distributed over 10 workers. Test consists of 1,625 samples. Each sample has $d = 112$ features. $q = 1/n$.

We solve the binary classification problem with $\lambda = 10^{-3}$ regularization.

Optimization problem:

$$\min_{x \in \mathbb{R}^n} \left(\frac{1}{m} \sum_{i=1}^m \ln(1 + \exp(-b_i \langle a_i, x \rangle)) + \frac{\lambda}{2} \|x\|^2 \right)$$

Classification rule: $b(a) := \text{sign}(\langle a, x \rangle)$

For each algorithm (SignSGD, DP-SGD, and DP-SignSGD), we set the learning rate $\eta = \frac{1}{\sqrt{100d}}$. The goal is $(10, 1/n^{1.1})$ privacy.

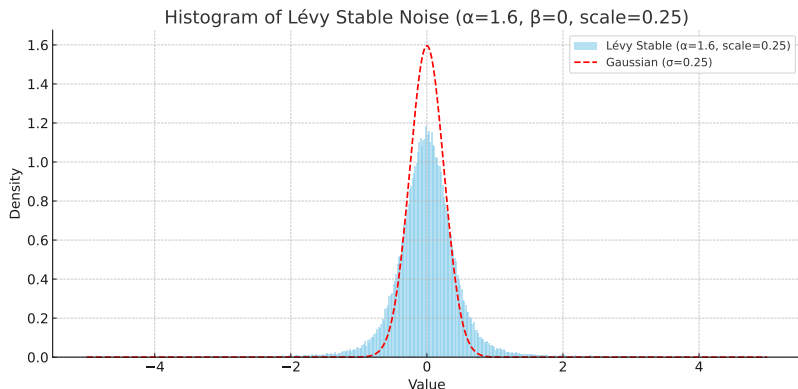
Heavy-tailed noise in gradient estimates

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

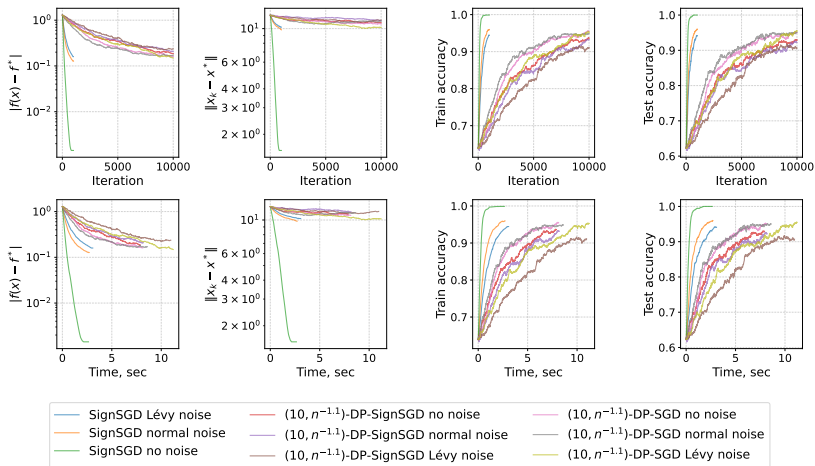
Synthetic heavy-tailed noise



We add to the gradients coordinatewise noise with Lévy stable distribution 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 $\varphi(t) = \exp(-0.25^{1.6}|t|^{1.6})$).

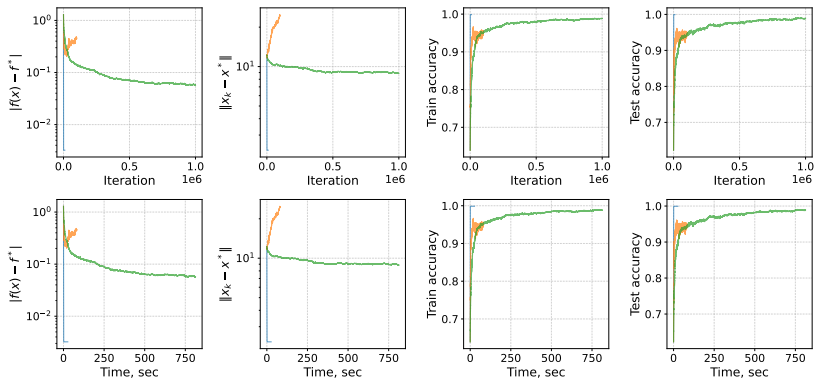
We compare settings with no noise, $\mathcal{N}(0, 0.25^2 \mathbb{I}^d)$ noise and this noise.

Computational experiment



DP-SignSGD converges very slowly, but it depends on a dataset.
Lower q leads to much lower σ and better convergence.

Constant vs Dynamic learning rate



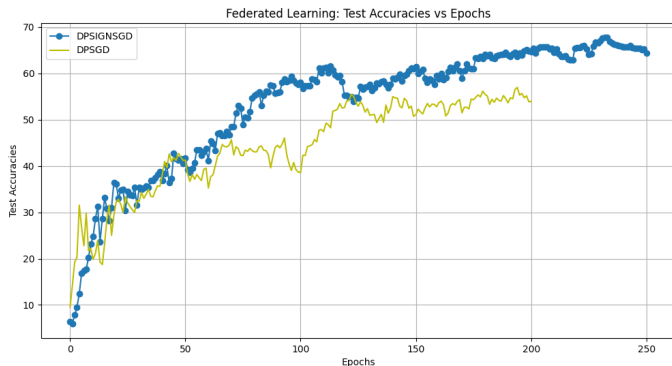
— SignSGD no noise

— $(10, n^{-1.1})$ -DP-SignSGD no noise const $lr = \frac{1}{10\sqrt{d}}$

— $(10, n^{-1.1})$ -DP-SignSGD $lr(k) = \frac{1}{10\sqrt{d}k^{1/5}}$

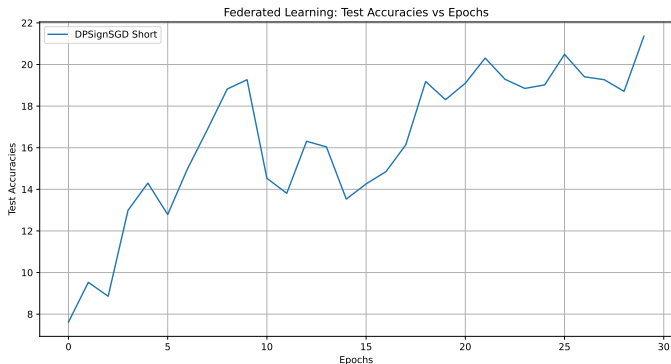
DP-SignSGD converges better, when the learning rate is dynamic.
 $T^{-1/5}$ instead of $T^{-1/2}$ factor is required, because σ depends on T .

Torch experiments: MLP on MNIST



Tradeoff: number of iterations vs precision of a single iteration.

Torch experiments: CNN on MNIST



Effective simulation of federated learning via torch.multiprocessing (10x speedup: up to 10 workers in parallel).

Experiment setting: 3 workers, $q = 0.25\%$,

$$\text{lr} = \min\left(1, \frac{t+1}{200}\right) \frac{0.01}{(t+1)^{1/3}}$$

Kith and kin of DP-SignSGD

- ▶ Add Top-k compressor
- ▶ σ s depend on the importance of the coordinate
- ▶ Repeat n times and then send the average
- ▶ Several local steps & send $\text{sign}(\Delta W)$

Summer plans

- ▶ Prove convergence of DP-SignSGD
- ▶ Test CNN on CIFAR-10 (code is almost ready)
- ▶ Research $\sigma(q, T)$ function: better understand noise
- ▶ Try DPSignSGD based on other DPSGD
- ▶ Try different compressor operators to tighten privacy accounting

DP-SignSGD with Bernoulli sampling $q = 1/n$

- ▶ $(\epsilon, 1/n^{1.1})$ differentially-private
- ▶ communication-effective
- ▶ empirically converges on logistic regression problem even with heavy-tailed noise
- ▶ trains MLP and CNN on MNIST to 70% accuracy
- ▶ empirically converges with the same type of convergence like DP-SGD with Bernoulli subsampling
- ▶ does not require clipping