УДК ?

Differentially private modification of SignSGD

A.Yu. Kravatskiy, A.A. Pliusnin, S.A. Chezhegov, and A.N. Beznosikov

Moscow Institute of Physics and Technology

Large models requiring more and more data for training, federating learning has become indispensable like never before. To use potentially sensitive user data for training the model, one has to guarantee its privacy. The gold standard for data privacy is (ε, δ)-differential privacy [1] of *mechanism*, which guarantees that that probabilities p_1 and p_2 of any response of the mechanism on datasets D_1 and D_2 respectively that differ only in one entry satisfy the relations $\frac{p_1}{p_2} \leq e^{\varepsilon} + \delta$ and $\frac{p_2}{p_1} \leq e^{\varepsilon} + \delta$. In the case of federated learning, the mechanism is all the outputs from the user's side, i.e. all the gradient data that they send to the server. For neural networks, recommended $\varepsilon = 1$, $\delta = 10^{-5}$.

Another constraint of federated learning is a high communication cost. To reduce it, one may send not the whole gradient, but only signs of each coordinate, thus reducing the cost in $32/2 = 16$ times. The standard algorithm that utilizes this technique is SignSGD [2]. This algorithm with *majority voting* (amidst workers that participate in the learning process) is not only communication-efficient, but also has proved to be resistant to heavy-tailed noise and converge with high probability [3].

Initially, we planned to provide theoretical guarantees of convergence, as in [3], for DPSignSGD, a differentially private modification of SignSGD [4]. However, on closer inspection, it turned out that that the algorithm is not truly differentially private. The authors have shown (ε, δ)-privacy of a single use of DPSignSGD, which means only $(T\varepsilon, T\delta)$-privacy for the whole procedure (T iterations). Moreover, they did not prove the convergence of DPSignSGD.

Thus, we had to construct our own DPSignSGD. We use the same Gaussian noise mechanism, adding $\mathcal{N}(0, \sigma^2 \mathbb{I}^d)$ noise, to ensure differential privacy. However, to make the most out of composition of mechanisms (the single mechanism being a sign of the gradient sent) and to enhance privacy, we add Bernoulli subsampling mechanism (one samples each of the elements with probability q) with proved (α, ε_R)-Rényi differential privacy [5], which is easily converted to (ε, δ)-privacy. We use the tighest proved bound for ε_R, which can be computed only numerically:

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

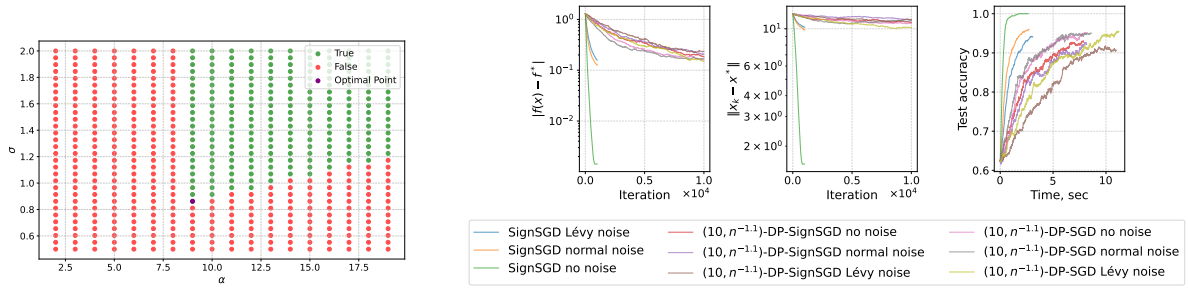
To preserve privacy, earlier defined ε_R and σ must satisfy:

$$\varepsilon_R \leq \varepsilon / T - \frac{\log 1/\delta}{T(\alpha - 1)} \tag{2}$$

To find the lowest σ, we perform grid search, like 1a.

Thus, we derive a new dp-sign compressor 1, which defines truly private DP-SignSGD. We test it on a regularized logistic regression problem for binary classification 1b.

Also we test it on MLP and CNN for hand-written digit classification from MNIST dataset. On a single worker, we reach 70% accuracy when we run the algorithm for 30 000 iterations, and 40% in 2 000 iterations. The main obstacle to reaching high accuracy is a Gaussian noise, inherent in private algorithms. Even with a skillful privacy accounting, it causes a tradeoff: precision vs max iteration number that still preserves privacy.



(a) Finding the minimal σ, sampling rate $q = 1/300$, and $T = 1000$

(b) Logistic regression on UCI Mushroom Dataset with SGD, SignSGD, DP-SIGNSGD and different type of noises

Figure 1: Experimental results

Algorithm 1 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 $\boldsymbol{g} = 0$.
Grid search $\sigma(q, T, \varepsilon, \delta)$ to satisfy (1) and (2).
$sign_{noised} = sign(\boldsymbol{g} + \mathcal{N}(0, \sigma)^2\mathbb{I}^d)$
Output: $sign(sign_{noised})$

That being said, DP-SignSGD promises to be far more private than we calculate. Indeed, the sign mechanism returns not all values in range [-1, 1], but only 1 and -1, which naturally should add more privacy guarantees. Moreover, even the gradients might be private to some extent in respect to the user dataset. Estimating the impact of these features of the mechanism might produce far more feasible algorithms.

Interestingly, DP-SignSGD can be easily adapted to any new tighter privacy guarante: lowering σ or raising T directly improves training results. At present, the pressing need concerning DP-SignSGD with σ noise and sampling probability q is to establish theoretical guarantees of its convergence. To the best of our knowledge, none of the existing works contain theoretical analysis of application of subsampling mechanism to optimization algorithms.

References

[1] C. Dwork et al. "The algorithmic foundations of differential privacy". In: Foundations and Trends® in Theoretical Computer Science, 9: 3–4: 211–407, 211–407.
[2] J. Bernstein et al. "signSGD: Compressed Optimisation for Non-Convex Problems". In: *International Conference on Machine Learning*. 2018, 560–569.
[3] N. Kornilov et al. Sign Operator for Coping with Heavy-Tailed Noise: High Probability Convergence Bounds with Extensions to Distributed Optimization and Comparison Oracle. 2025.
[4] R. 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: 3834–3846, 3834–3846. ISSN: 2162-2388.
[5] I. Mironov et al. Rényi Differential Privacy of the Sampled Gaussian Mechanism. 2019.