УДК 519.676, 519.688

Differentially private modification of SignSGD

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

Moscow Institute of Physics and Technology

With large models requiring more and more data for training, federated learning has become more indispensable than ever. 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 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 federated learning, the mechanism returns all the outputs from the user, i.e. all the gradient data that they send to the server. For neural networks, standard parameters are $\varepsilon = 1$, $\delta = 10^{-5}$.

Another constraint of federated learning is the high communication cost. To reduce it, one can send not the whole gradient, but only the sign of each coordinate, thereby reducing the cost by a factor of $32/2 = 16$. The standard algorithm that utilizes this technique is SignSGD [2]. This algorithm, which employs *majority voting* among the workers, not only is communication-efficient but has also been shown to be robust to heavy-tailed noise and to converge with high probability [3].

Initially, we aimed to provide theoretical guarantees of convergence for DP-SignSGD, a differentially private modification of SignSGD [4], following the approach in [3]. However, upon closer examination, we found that the algorithm is not truly differentially private. The authors have shown (ε, δ)-privacy of a single iteration of DP-SignSGD, which means only $(T\varepsilon, T\delta)$-privacy over T iterations. Furthermore, they did not establish any convergence guarantees for DP-SignSGD.

Consequently, we had to construct our own DP-SignSGD. Like in [4], we use *Gaussian mechanism*, that ensures differential privacy by adding $\mathcal{N}(0, \sigma^2 \mathbb{I}^d)$ noise. However, to make the most out of composition of mechanisms (a single mechanism returning a noised sign of the gradient) and to enhance privacy, we incorporate *Bernoulli subsampling*. In this scheme, at each iteration of the alogirthm, each entry of the dataset is sampled with probability q and the gradients are computed only for the subsampled data. The resulting mechanism satisfies (α, ε_R)-Rényi differential privacy [5], which is readily converted to (ε, δ)-privacy. We use the tighest bound for ε_R from [5], which has only numerical form:

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

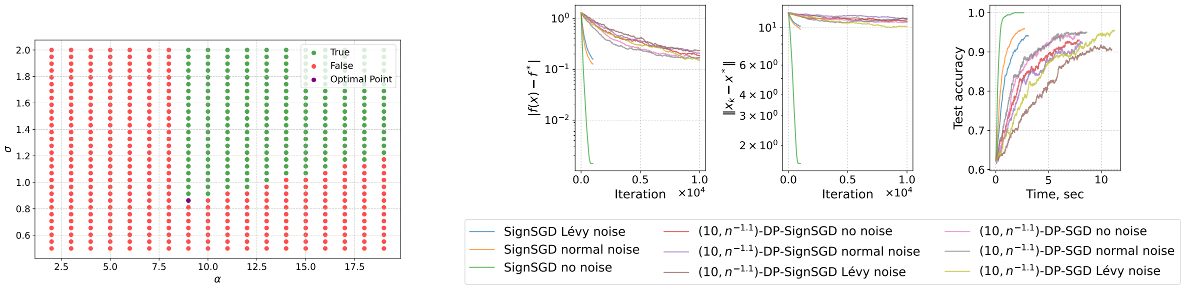
To ensure privacy, the previously defined $\varepsilon_R(q, \alpha, \sigma)$ must satisfy the following condition:

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

With q being fixed, we use a grid search to find the minimal σ, as shown in fig. 1a.

Thus, we propose a new DP-SIGN compressor (see algorithm 1) that defines a truly private variant of DP-SignSGD. We test it on a regularized logistic regression problem for binary classification. The results are presented in fig. 1b.

We also test the algorithm on an MLP and CNN for handwritten digit classification using the MNIST dataset. On a single worker, we achieve 70% accuracy after 30,000 iterations with large σ, and 40% accuracy after 2,000 iterations with small σ. The main



(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 types of noise

Figure 1: Experimental results

Algorithm 1 DP-SIGN compressor

```
Input: coordinate $w$, loss function $l$, user database $D$, $(\varepsilon, \delta)$-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: $sign(sign_{noised})$
```

obstacle to higher accuracy is the Gaussian noise inherent in private algorithms. Even with careful privacy accounting, this introduces a tradeoff between precision of an iteration and the maximum number of iterations that still preserve privacy.

That being said, DP-SignSGD is likely more private than our calculations suggest. The sign mechanism, for instance, does not return all values in the range $[-1, 1]$, but only 1 and -1, which inherently provides additional privacy guarantees. Furthermore, even the gradients may be partially private with respect to the user dataset. Estimating the impact of these aspects of the mechanism could lead to more feasible algorithms.

Interestingly, DP-SignSGD can be easily adapted to any new, tighter privacy guarantee: lowering σ or increasing T directly improves training results. At present, the primary challenge concerning DP-SignSGD with σ-noise and sampling probability q is establishing theoretical guarantees of its convergence. To the best of our knowledge, none of the existing works provide a theoretical analysis of the application of the 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.