# Sign-SGD with Heavy Tails and Differential Privacy

Speaker: Alexey Kravatsky, 3rd student, MIPT

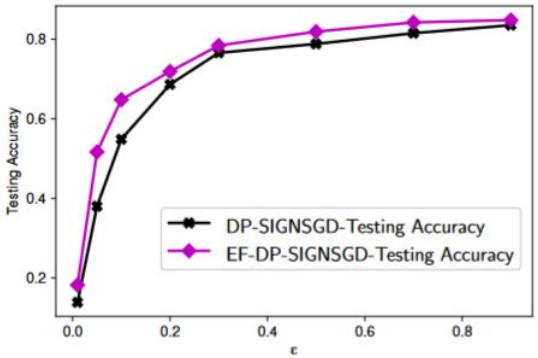
Advisor: Savelii Chezhegov

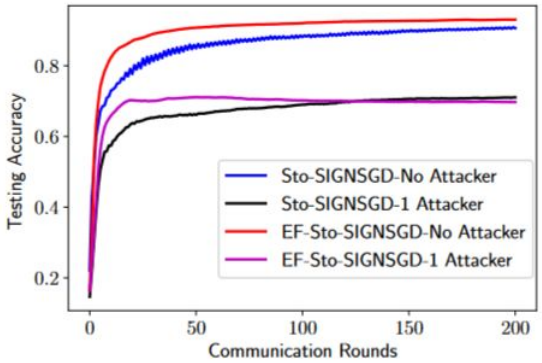
03/10/2025, MIPT

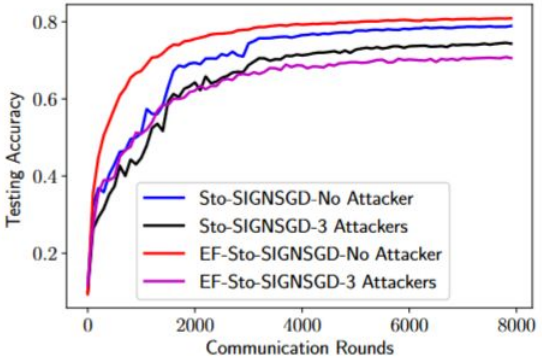
# Meld the approaches of two papers

- Paper 1: relatively general proofs for vanilla Sign-SGD convergence
- Paper 2: weaker proofs, but theoretical and experimental analysis of the application to the federated learning

We want to mix it to create an algorithm to securely train LLMs on user data. However, we will start from MNIST to test our ideas.



**Algorithm 1** SignSGD

**Input:** Starting point $x^1 \in \mathbb{R}^d$, number of iterations $T$, stepsizes $\{\gamma_k\}_{k=1}^T$.

1: **for** $k = 1, \ldots, T$ **do**
2: Sample $\xi^k$ and compute estimate $g^k = \nabla f(x^k, \xi^k)$;
3: Set $x^{k+1} = x^k - \gamma_k \cdot \text{sign}(g^k)$;
4: **end for**

**Output:** uniformly random point from $\{x^1, \ldots, x^T\}$ .

**Algorithm 1** Stochastic-Sign SGD with majority vote

**Input**: learning rate $\eta$, 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** $w^{(t+1)} = w^{(t)} - \eta\tilde{\boldsymbol{g}}^{(t)}$.

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

$$dp\text{-}sign(\boldsymbol{g}_m^{(t)}, \epsilon, \delta)_i = \begin{cases} 1, & \text{with probability } \Phi\big(\frac{(\boldsymbol{g}_m^{(t)})_i}{\sigma}\big) \\ -1, & \text{with probability } 1 - \Phi\big(\frac{(\boldsymbol{g}_m^{(t)})_i}{\sigma}\big) \end{cases} \tag{15}$$

**Assumption 1** (Lower bound). *The objective function $f$ is lower bounded by $f^* > -\infty$, i.e., $f(x) \geq f^*, \forall x \in \mathbb{R}^d$.*

**Assumption 2** (Smoothness). *The objective function $f$ is differentiable and $L$-smooth, i.e., for the positive constant $L$*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^d.$$

**Assumption 3** (Heavy-tailed noise in gradient estimates). *The unbiased estimate $\nabla f(x, \xi)$ has bounded $\kappa$-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 $\vec{\sigma} = [\sigma_1, \ldots, \sigma_d]$ are non-negative constants. If $\kappa = 2$, then the noise is called a bounded variance.*

The idea for our proofs is to be borrowed from Kornilov et al.

**Lemma 1** (SignSGD **Convergence Lemma**). *Consider lower-bounded $L$-smooth function $f$ (As. 1, 2) and HT gradient estimates (As. 4). Then Alg. 1 after $T$ iterations with constant stepsizes $\gamma_k \equiv \gamma$ achieves with probability at least $1-\delta$ starting with $\Delta_1 = f(x^1) - f^*$:*

$$\frac{1}{T}\sum_{k=1}^{T}\|\nabla f(x^k)\|_1 \leq \frac{2\Delta_1}{T\gamma} + 16Ld\gamma\log(1/\delta) + 4\|\vec{\sigma}\|_1 + 12\frac{d\|\nabla f(x^1)\|_1}{T}\log(1/\delta). \quad (2)$$

# From China with love

**Definition 3.** *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 $(\epsilon, \delta)$-locally differentially private if for any measurable subset $\mathcal{S} \subseteq \mathcal{O}$ and two neighboring local datasets $(D_1, D_2) \in \mathcal{N}_{\mathcal{D}}$,*

$$P(\mathcal{M}(f(D_1)) \in \mathcal{S}) \leq e^{\epsilon} P(\mathcal{M}(f(D_2)) \in \mathcal{S}) + \delta. \tag{21}$$

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

$$\Delta_r = \max_{(D_1, D_2) \in \mathcal{N}_{\mathcal{D}}} ||f(D_1) - f(D_2)||_r. \tag{22}$$

**Theorem 6.** *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 $\sigma_0$ such that when $\sigma \geq \sigma_0$, $P(sign(\frac{1}{M}\sum_{m=1}^{M} \hat{u}_i) \neq sign(\frac{1}{M}\sum_{m=1}^{M} u_i)) < \left[\left(1 - x^2\right)\right]^{\frac{M}{2}}$, where $x = \frac{|\sum_{m=1}^{M} u_m|}{2\sigma M}$.*

# This is impractical. Our goal is to condense it.

**Theorem 2.** *Suppose Assumptions 1, 2 and 4 are satisfied, and the learning rate is set as* $\eta = \frac{1}{\sqrt{Td}}$. *Then by running Algorithm 1 with* $q(\boldsymbol{g}_m^{(t)}) = \textit{sto-sign}(\nabla f_m(w^{(t)}), \boldsymbol{b})$ *(termed as* Sto-SIGNSGD*) for* $T$ *iterations, we have*

$$\frac{1}{T}\sum_{t=1}^{T}||\nabla F(w^{(t)})||_1 \leq \frac{1}{c}\left[\frac{\mathbb{E}[F(w^{(0)}) - F(w^{(T+1)})]\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + \frac{2}{T}\sum_{t=1}^{T}\sum_{i=1}^{d}|\nabla F(w^{(t)})_i|\mathbb{1}_{p_i^{(t)} > \frac{1-c}{2}}\right]$$
$$\leq \frac{1}{c}\left[\frac{(F(w^{(0)}) - F^*)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2\sum_{i=1}^{d} b_i \Delta(M)\right], \tag{10}$$

*where* $0 < c < 1$ *is some positive constant,* $p_i^{(t)}$ *is the probability that the aggregation on the* $i$*-coordinate of the gradient is wrong during the* $t$*-th communication round, and* $\Delta(M)$ *is the solution to* $\left(1 - x^2\right)^{\frac{M}{2}} = \frac{1-c}{2}$. *The second inequality is due to the fact that* $p_i^{(t)} > \frac{1-c}{2}$ *only if* $\frac{|\nabla F(w^{(t)})_i|}{b_i} \leq \Delta(M)$.

# Bibliography


- Jin et al., 2020: Chinese article with compressors and guarantees of differential privacy
- Kornilov et al., 2025: Russian preprint with proofs of high-probability convergence in the case of heavy-tailed noise
- These articles are key to our research. Others will be used only as a reference.