جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

Methods with Local Steps and Random Reshuffling for Generally Smooth Non-Convex Federated Optimization

Yury Demidovich* Petr Ostroukhov* Grigory Malinovsky Samuel Horváth Martin Takáč Peter Richtárik Eduard Gorbunov

King Abdullah University of Science and Technology (KAUST) Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) Moscow Institute of Physics and Technology (MIPT)

Problem Setup

We consider a standard distributed optimization problem

$$\min_{x\in\mathbb{R}^d}\left\{f(x) := \frac{1}{M}\sum_{m=1}^{M} f_m(x)\right\}. \quad (1)$$

- $[M] := \{1, 2, \ldots, M\}$ is a set of workers, $f_m : \mathbb{R}^d \to \mathbb{R}$ is a non-convex loss function, computed on the data available on client m for the current model $x \in \mathbb{R}^d$;
- workers compute $\nabla f_m(x)$ or $\nabla f_{mj}(x)$ (in this case we assume that functions $\{f_m\}_{m=1}^M$ have the finite-sum form).

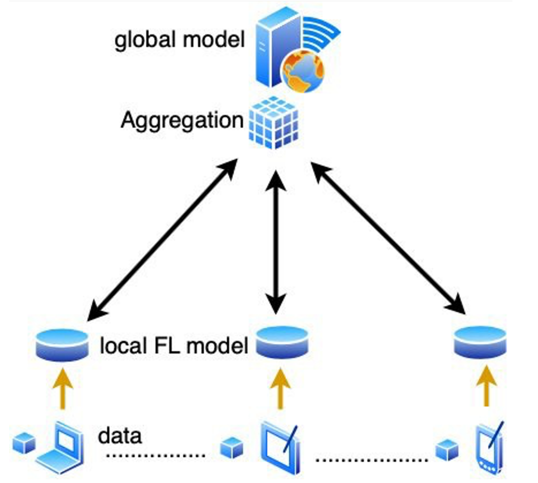


Figure: Federated Learning illustration

Assumption: Symmetric (L_0, L_1)-smoothness or general smoothness

The function $f(x)$ is *symmetrically* (L_0, L_1)*-smooth* (*generally smooth*) if

$$\|\nabla f(x) - \nabla f(y)\| \le \left(L_0 + L_1 \sup_{u\in[x,y]} \|\nabla f(u)\|\right)\|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \quad (2)$$

If f is twice-differentiable, this is equivalent to

$$\left\|\nabla^2 f(x)\right\| \le L_0 + L_1 \|\nabla f(x)\|, \ \forall x \in \mathbb{R}^d. \quad (3)$$

Motivation

- Often real-life problems do not suit under regular L-smoothness condition.
- In Figure to the right we show, that Hessian of x^4 can be bounded by $L_0 + L_1\|\nabla f(x)\|$, but can't be bounded by some L.
- In [1, 2] authors introduce concept of generalized smoothness and empirically show that it accurately represents real-world problems.
- Such problem class is largely unexplored in context of federated learning.
- Generalized smoothness shows strong connection with clipping.

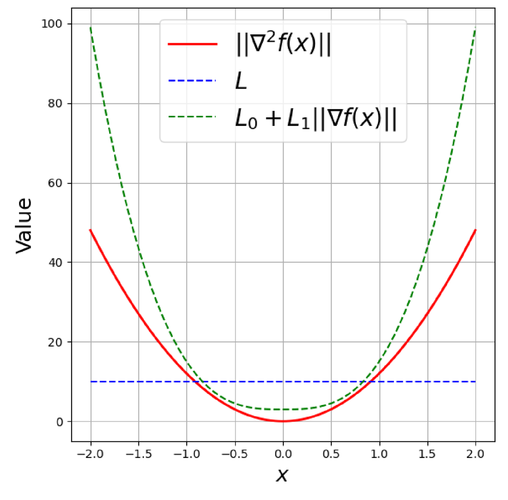


Figure: Gen. smoothness of x^4.

Generalized Smoothness and Clipping

Generalized Smoothness step size:

$$\gamma_k \equiv \frac{1}{L_0 + L_1\|\nabla f(x_k)\|} \le \min\left\{\frac{1}{2L_0}, \frac{1}{2L_1\|\nabla f(x_k)\|}\right\} = \frac{1}{2L_0}\min\left\{1, \frac{L_0}{L_1\|\nabla f(x_k)\|}\right\}. \quad (4)$$

Clipped step size:

$$\gamma_k \equiv \gamma \min\left\{1, \frac{\lambda}{\|\nabla f(x_k)\|}\right\}. \quad (5)$$

Main Contribution

Algorithm	Local Steps	Data Reshuffling	Client Participation	Server Step	Server LR
Clip-LocalGDJ	✓	-	Full	Aggregated	Clipped
CLERR	✓	Global	Full	Aggregated	Clipped
Clipped-RR-CLI	✓	Local	Partial	Aggregated	Clipped

Algorithm CLERR: Clipped once in an Epoch Random Reshuffling

```
Input: Starting point x_0 ∈ R^d, number of epochs T, constants c_0, c_1 > 0.
for t = 0, ..., T − 1 do                                    ▷ cycle over communication rounds
    Choose global stepsize γ_t = 1/(c_0 + c_1‖∇f(x_t)‖).    ▷ clipping of global stepsize
    Choose small inner stepsize α_t > 0.
    Sample a permutation π_t = {π_t(1), ..., π_t(N)}.      ▷ permute data once in a communication round
    for m = 1, ..., M do                                    ▷ cycle over clients
        x^m_{t,0} = x_t
        for j = 0, ..., N − 1 do                            ▷ cycle over data points
            x^m_{t,j+1} = x^m_{t,j} − α_t ∇f_{m,π_t(j)}(x^m_{t,j}).   ▷ update client point
        end for
        g^m_t = (1/(α_t N)) (x_t − x^m_{t,N})               ▷ aggregate gradient for m-th client
    end for
    g_t = (1/M) Σ_{m=1}^{M} g^m_t.                          ▷ aggregate gradient over all the M clients
    x_{t+1} = x_t − γ_t g_t.                                ▷ aggregated server step (jumping)
end for
```

Convergence Analysis

- If $T \ge \frac{256\delta_0}{\zeta\varepsilon}$ and α_t is small enough, then $\mathbb{E}\left[\min_{t=1\ldots T}\left\{\min\left\{\frac{\|\nabla f(x_t)\|^2}{L_0}, \frac{\|\nabla f(x_t)\|}{L_1}\right\}\right\}\right] \le \varepsilon$.
- In standard smooth case, we recover rate $O\left(\frac{L_0\delta_0}{\varepsilon}\right)$ of RR from [4].
- In standard smooth case with PL-condition, we recover $O\left(\frac{L_0}{\mu}\ln\frac{2\delta_0}{\varepsilon}\right)$ of RR from [4].

Theorem 1

Let $f \equiv \sum_{m=1}^{M} f_m(x)$, $f_m \equiv \sum_{j=0}^{N-1} f_{mj}(x)$ and $f_{mj}(x)$ be lower bounded and (L_0, L_1)-smooth. Choose small client stepsizes α_t, global stepsizes $\gamma_t : \frac{\zeta}{\hat{a}_t} \le \gamma_t \le \frac{1}{4\hat{a}_t}$, where $\hat{a}_t \equiv L_0 + L_1\|\nabla f(x_t)\|$ Then, the iterates $\{x_t\}_{t=0}^{T-1}$ of Algorithm 2 satisfy

$$\mathbb{E}\left[\min_{t=0,\ldots,T-1}\left\{\frac{\zeta}{8}\min\left\{\frac{\|\nabla f(x_t)\|^2}{L_0}, \frac{\|\nabla f(x_t)\|}{L_1}\right\}\right\}\right] \le \frac{8\left(1 + \frac{3\alpha_t^2\tilde{a}_t^3}{8\hat{a}_t}((N-1)(2N-1) + 2(N+1))\right)^T}{T}\delta_0 + \frac{6\alpha_t^2\tilde{a}_t^3}{\hat{a}_t}(N+1)\Delta^\star, \quad (6)$$

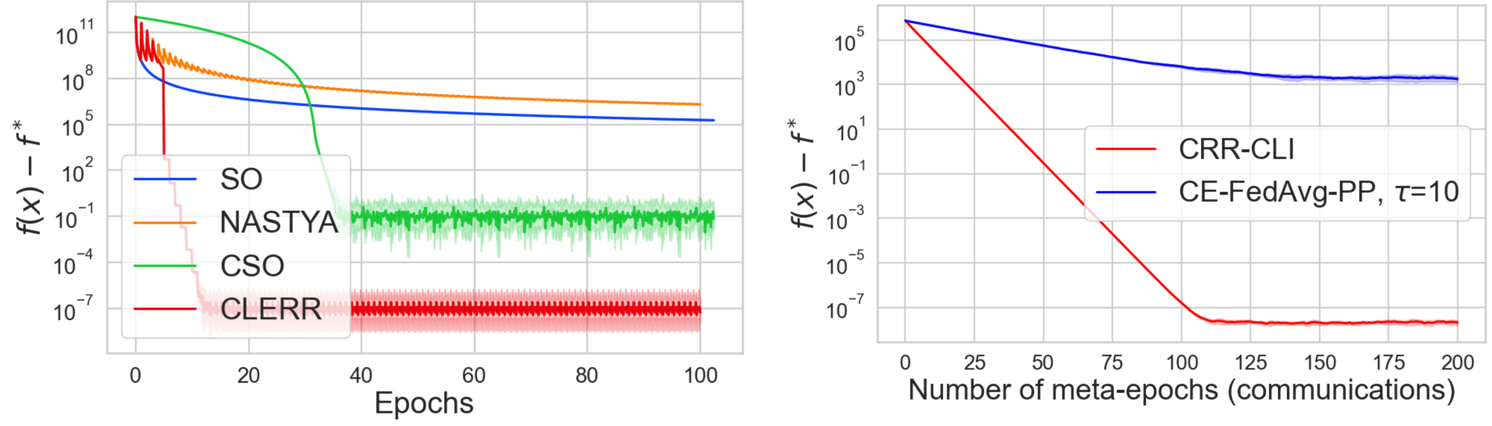
where $\hat{a}_t \equiv L_0 + L_1\|\nabla f(x_t)\|$, $a_t \equiv L_0 + L_1 \max_m \|\nabla f_m(x_t)\|$, $\tilde{a}_t \equiv L_0 + L_1 \max_{m,j}\|\nabla f_{mj}(x_t)\|$, $\Delta^* \equiv f^* - \frac{1}{M}\sum_{m=1}^{M} f_m^*$, $\delta_0 \equiv f(x_0) - f^*$.

* - **These authors contributed equally to this work.**

Experiments

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}\|x - x_i\|^4, \ x_i \in [-10, 10]^d \quad (7)$$

- Comparison of the Shuffle-Once (**SO**) methods, that shuffle data once before train loop, on generally-smooth (2) problem (7).
- Comparison of methods with local steps (**LS**) on (7).
- Comparison of methods with random reshuffling (**RR**), LS and partial participation (**PP**) on (7).
- Comparison of the SO methods on ResNet-18 on CIFAR-10 problem.



(a) SO methods on problem (7) (b) Methods with RR, LS (τ) and PP

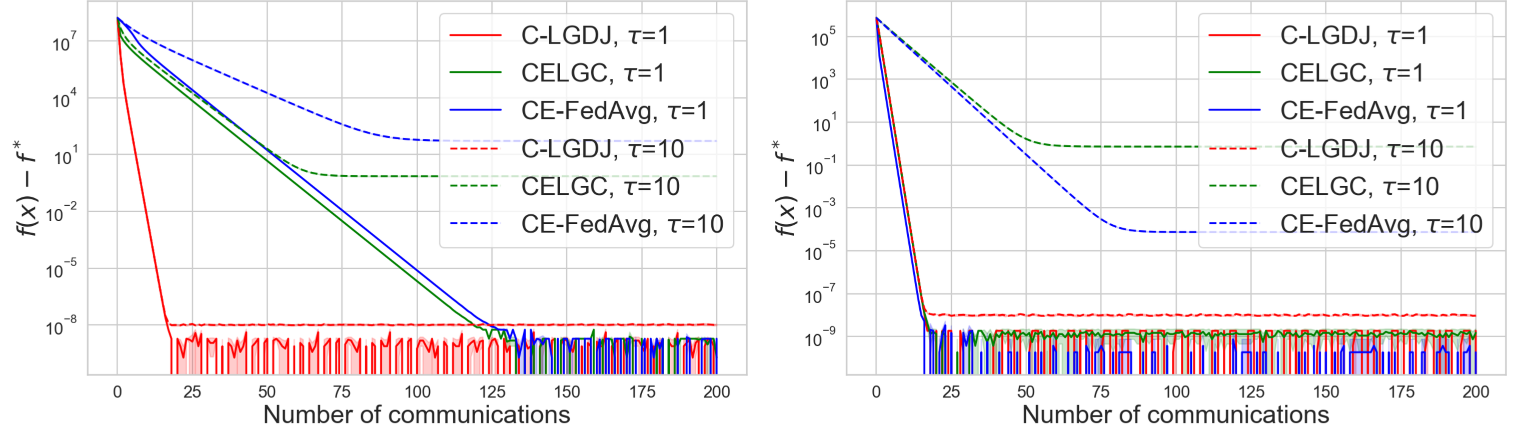


Figure: Problem (7) for $x_0 = (1, \ldots, 1)$ (left) and $x_0 = (10, \ldots, 10)$ (right) and different number of local steps τ

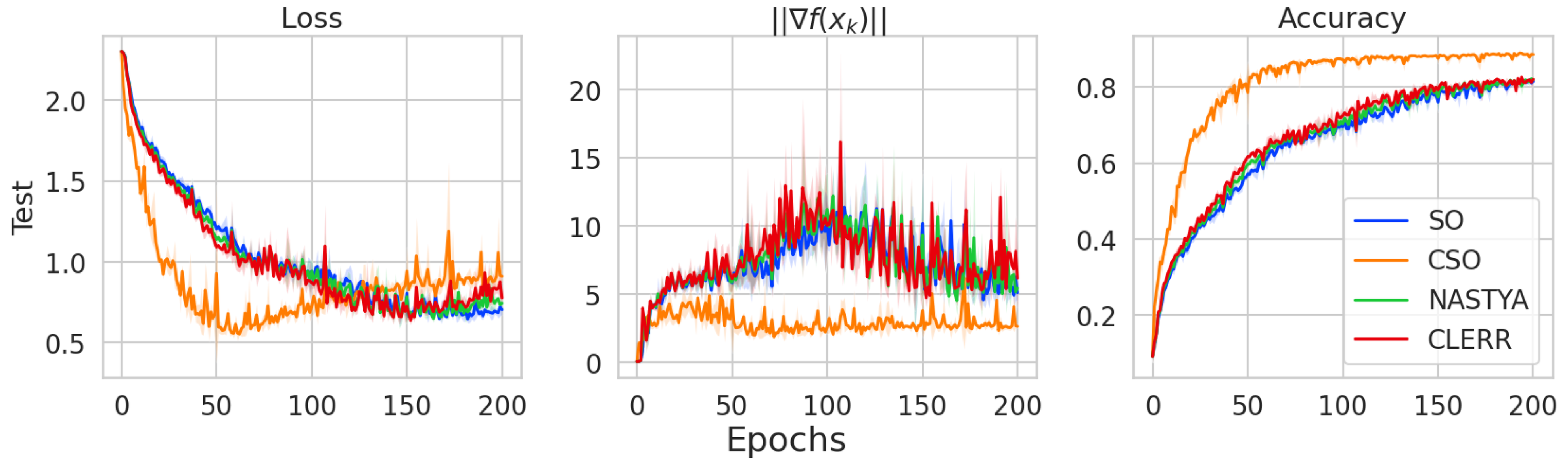


Figure: SO methods on ResNet-18 on CIFAR-10 problem

[1] Jingzhao Zhang, Tianxing He, Suvrit Sra, Ali Jadbabaie, Why gradient clipping accelerates training: A theoretical justification for adaptivity, 2020.
[2] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization, 2023.
[3] Grigory Malinovsky, Konstantin Mishchenko, and Peter Richtárik. Server-side stepsizes and sampling without replacement provably help in federated optimization, 2023.
[4] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. 2020.