Smoothed Gradient Clipping and Error Feedback for Decentralized Optimization under Symmetric Heavy-Tailed Noise

Shuhua Yu
Electrical and Computer Engineering
Carnegie Mellon University

INFORMS 2024
Robustness at the Intersection of Optimization, Statistics, and Machine Learning

Collaborators

Dušan Jakovetić
Univ. of Novi Sad

Soummya Kar
CMU

Distributed systems

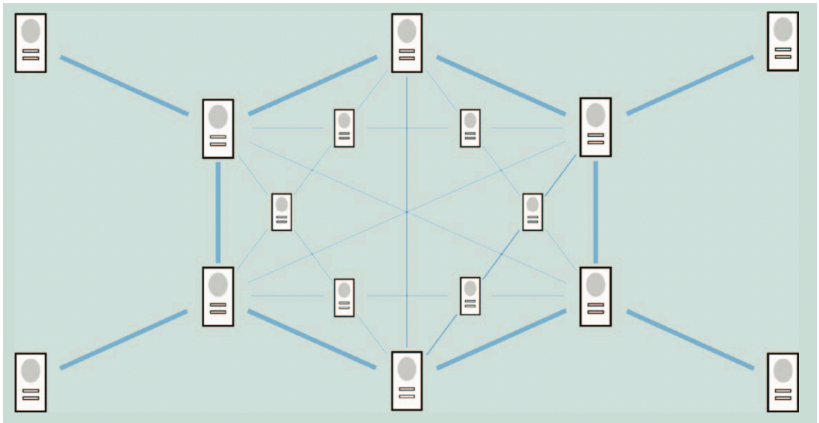


- Cyber-physical systems: power grids, sensor networks.
- Cloud centered devices: smartphones, wearable devices.
- Autonomous vehicle systems: sensors, actuators, multi-vehicle coordination.
- ...
- Goal: robust information processing over distributed systems.

Two distributed schemes

Server/client model

Server coordinates the *global* and *local* information exchange

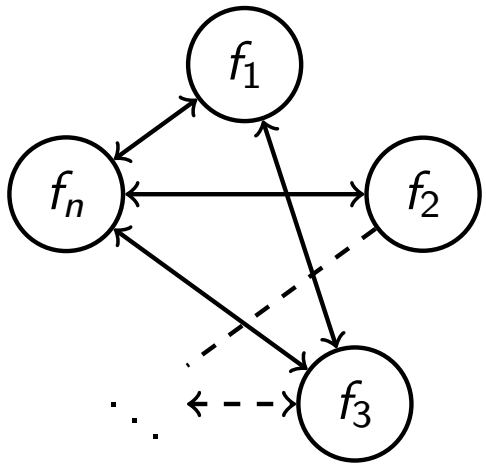


Decentralized model

Agents exchange *local* information with direct neighbors over a graph

- Data are distributed over multiple agents due to privacy and scalability.
- We focus on the more general decentralized model.
  - Flexible: no central server is required.
  - Less communication: communication with neighbors only.

Decentralized optimization



- Consider a network of agents $i = 1, \ldots, n$.
- Agent i holds a local data distribution $\mathcal{D}_i$, on which we define

$$f_i(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}_i \sim \mathcal{D}_i} \ell(\boldsymbol{x}, \boldsymbol{\xi}_i).$$

  for some loss function ℓ. Examples include: least-squares, logistic-regression, neural networks.
- Agents communicate over a graph to minimize $f(\boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x})$.

Decentralized SGD

A basic decentralized gradient method:

- Agent i holds a local decision variable $\boldsymbol{x}_i$;
- Agent i computes a stochastic gradient with random noise $\boldsymbol{\xi}_i$,

$$g_i(\boldsymbol{x}_i) = \nabla f_i(\boldsymbol{x}_i) + \boldsymbol{\xi}_i;$$

- Agent i employs some weight w_{ij}, $w_{ij} > 0$ if agent j is the neighbor of agent i;
- Decentralized stochastic gradient descent (SGD), for some stepsize η,

$$\boldsymbol{x}_i^+ \leftarrow \sum_{j=1}^{n} w_{ij} \left[\boldsymbol{x}_i - \eta(\nabla f_i(\boldsymbol{x}_i) + \boldsymbol{\xi}_i)\right].$$

- $\mathbb{E}[\boldsymbol{\xi}_i|\boldsymbol{x}_i] = \mathbf{0}$ and $\mathbb{E}[\|\boldsymbol{\xi}_i\|^2|\boldsymbol{x}_i] \leq \sigma^2$ for some $\sigma > 0$.

Q: $\boldsymbol{\xi}_i$ is typically assumed to be light-tailed, what if it is heavy-tailed?

Heavy-tailed noise

- Deep learning training may involve heavy-tailed gradient noise (Simsekli, Sagun, and Gurbuzbalaban 2019)

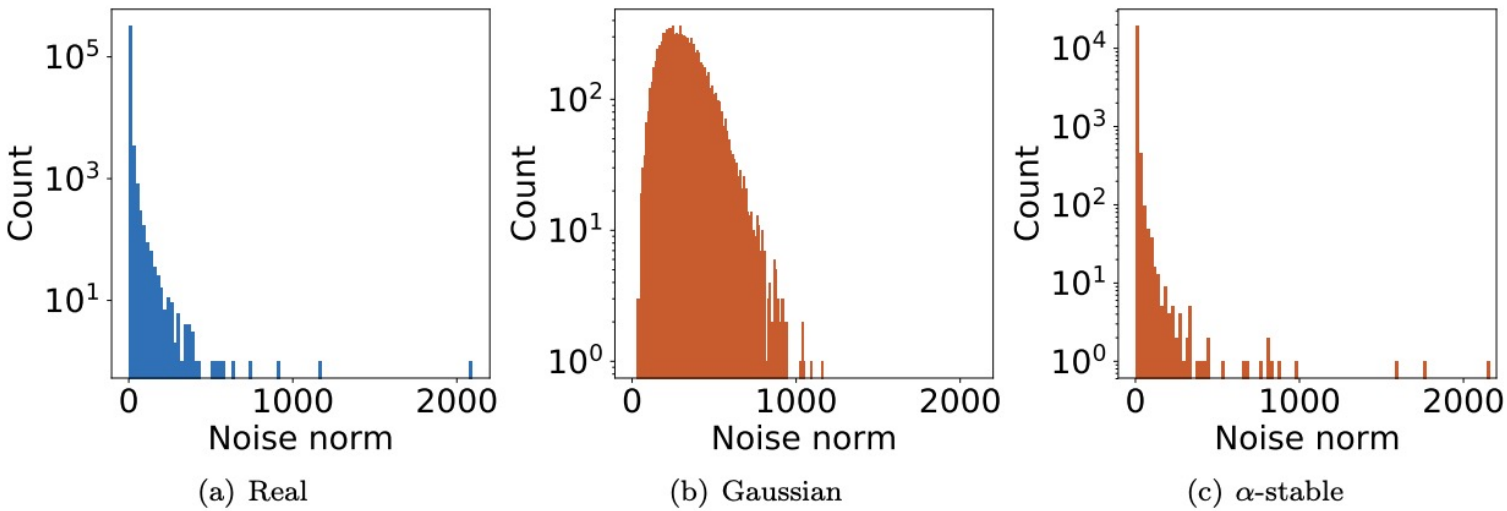
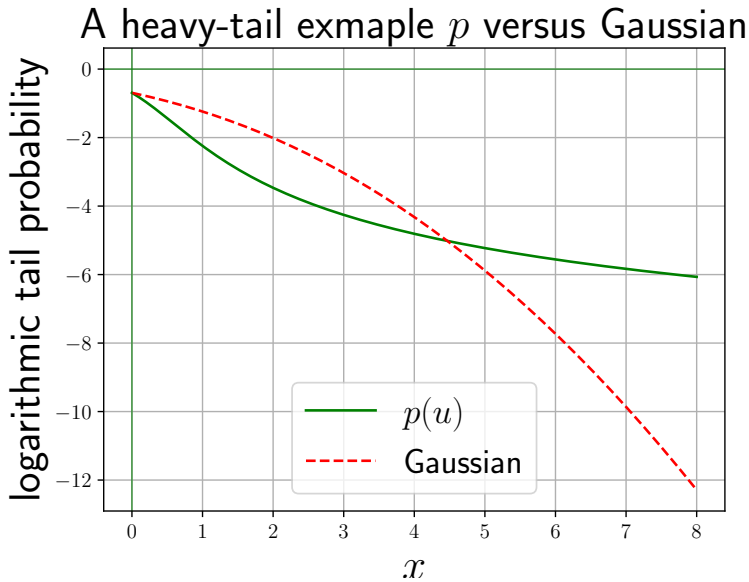


Figure: (a) The histogram of the norm of the gradient noises computed with AlexNet on CIFAR10. (b) and the histograms of the norms of (scaled) Gaussian and α-stable random variables.

Heavy-tailed noise

- A random variable X is heavy-tailed if $\forall a > 0$, $\limsup_{x \to \infty} \mathbb{P}(X > x) e^{ax} = \infty$.



- Alternatively, a random variable $\boldsymbol{\xi}$ is called heavy-tailed if it satisfies the *bounded moment condition (BM)*: $\mathbb{E}\|\boldsymbol{\xi}\|^{\alpha} \leq \sigma^{\alpha}$ for $\alpha < 2$ and $\sigma > 0$.
- In non-distributed case, vanilla SGD, i.e., $\boldsymbol{x}^{+} \leftarrow \boldsymbol{x} - \eta g(\boldsymbol{x})$, diverges under heavy-tailed noise (Zhang et al. 2020).

Non-distributed case

- Clipped SGD with step η, ℓ_2-norm threshold τ, stochastic gradient $g(\boldsymbol{x})$:

$$\boldsymbol{x}^+ \leftarrow \boldsymbol{x} - \eta \operatorname{clip}_\tau(g(\boldsymbol{x})), \ \operatorname{clip}_\tau(g(\boldsymbol{x})) = \min\left\{\frac{\tau}{\|g(\boldsymbol{x})\|}, 1\right\}$$

- Gradient clipping is effective in handling heavy-tailed gradient noise (Zhang et al. 2020).
- More general nonlinearities Ψ are effective in handling heavy-tailed noise (Jakovetić et al. 2023)

$$\boldsymbol{x}^+ \leftarrow \boldsymbol{x} - \eta \Psi\big(g(\boldsymbol{x})\big).$$

  Examples of Ψ include:
  - gradient clipping (Nguyen et al. 2023),
  - component-wise clipping (Zhang et al. 2020),
  - component-wise quantization,
  - sign,
  - normalized gradient (Hübler, Fatkhullin, and He 2024).

Fixed point issue

Q: How can we extend the gradient clipping to the distributed case?

- A direct extension of gradient clipping to server/client case: $\boldsymbol{x}^+ \leftarrow \boldsymbol{x} - \frac{\eta}{n}\sum_{i=1}^n \mathrm{clip}_\tau(g_i(\boldsymbol{x}))$.
- Consider minimizing $f(x) = \frac{1}{3}\sum_{i=1}^3 f_i(x), f_i(x) = \frac{1}{2}(x - a_i)^2$ where $a_1 = a_2 = -2, a_3 = 7$. Then $x^* = 1$, but for $\tau = 1$,

$$\frac{1}{3}\sum_{i=1}^3 \mathrm{clip}_\tau(f_i'(1)) = \frac{1}{3}.$$

  The global minimizer $x^* = 1$ is not a fixed point.
- Consider minimizing $f(x) = \frac{1}{2}\sum_{i=1}^2 f_i(x), f_i(x) = \frac{1}{2}(x - a_i)^2$ where $a_1 = 2, a_2 = -2$. Then $x^* = 0$, but for $\tau = 1$,

$$\frac{1}{2}\sum_{i=1}^2 \mathrm{clip}_\tau(f_i'(1)) = 0.$$

  A Non-minimizer $x = 1$ is a fixed point.

Distributed methods

Comparisons with existing distributed methods robust to heavy-tailed noise.

- Distributed setup: SC (sever/client), D (decentralized network).
- Gradient and noise condition: UG (unbounded gradient).
- c_s is a constant independent of moment index α.

	SC?	D?	UG?	α	Rate	Cvg
Yang, Qiu, and Liu 2022	✓	✗	✗	$\alpha > 1$	$O(1/t^{2-2/\alpha})$	MSE
gorbunov2023high	✓	✗	✓	$\alpha > 1$	$O(1/t^{2-2/\alpha})$	h.p.
Sun 2023	✓	✓	✗	$\alpha > 1$	no rate	a.s.
Ours	✓	✓	✓	$\alpha \geq 1$	$\mathcal{O}(1/t^{\min(c_s,\ 2/5)})$	MSE

Our method

Q: How to deal with heavy-tailed noise in decentralized gradient method?

SClip-EF-Network

Require: $\varphi_t, \epsilon_t, \beta_t, \eta_t; \boldsymbol{x}_i^0 = \boldsymbol{x}_j^0,\ \forall i, j$; component-wise: $\boldsymbol{\Psi}_t(\boldsymbol{x}) = \frac{\varphi_t}{\sqrt{\boldsymbol{x}^2+\epsilon_t}}$.

1: $\boldsymbol{m}_i^0 \leftarrow \boldsymbol{0},\ \forall i = 1, \ldots, n$
2: **for** $t = 0, \ldots, T-1$ **do**
3: **for** node $i = 1, \ldots, n$ in parallel **do**
4: $\boldsymbol{m}_i^{t+1} \leftarrow \beta_t \boldsymbol{m}_i^t + (1-\beta_t)\boldsymbol{\Psi}_t(g_i(\boldsymbol{x}^t) - \boldsymbol{m}_i^t)$
5: Send $\boldsymbol{x}_i^t - \eta_t \boldsymbol{m}_i^{t+1}$ to all neighbors of agent i
6: $\boldsymbol{x}_i^{t+1} \leftarrow \sum_{j=1}^n w_{ij}(\boldsymbol{x}_j^t - \eta_t \boldsymbol{m}_j^{t+1})$
7: **end for**
8: **end for**

Main ideas

- Local gradient estimator $\boldsymbol{m}_i^t$:

$$\boldsymbol{m}_i^{t+1} \leftarrow \beta_t \boldsymbol{m}_i^t + (1-\beta_t)\boldsymbol{\Psi}_t(g_i(\boldsymbol{x}^t) - \boldsymbol{m}_i^t).$$

- Use estimation error as a feedback to compensate the current estimate.
- Apply $\boldsymbol{\Psi}_t$ on the estimation error, as $g_i(\boldsymbol{x}^t) - \boldsymbol{m}_i^t \to \boldsymbol{0}$.
- Use smooth clipping operator to avoid discontinuity/non-differentiability.

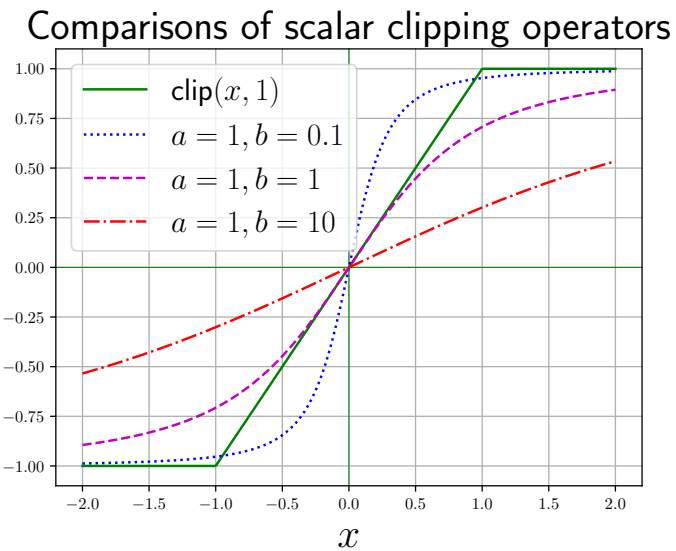


Figure: An illustration of $\Psi(x) = \frac{ax}{\sqrt{x^2+b}}$

Problem model

Functions:

- $\forall i,\ \mu \boldsymbol{I} \preceq \nabla^2 f_i(\boldsymbol{x}) \preceq L\boldsymbol{I}$ for some constants $L \geq \mu > 0$.
- Heterogeneity: Let $\boldsymbol{x}^* := \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$. $\|\nabla f_i(\boldsymbol{x}^*)\|_\infty \leq c_* < \infty$.

Noise:

- $\{\{\boldsymbol{\xi}_i^t\}_{i\in[n]}\}_{t\geq 0}$ are independently distributed over all iterations $t \geq 0$. Every component of $\boldsymbol{\xi}_i^t$ has marginal density p.
  - p is symmetric: $\forall u \in \mathbb{R}, p(u) = p(-u)$.
  - p has bounded first absolute moment: $\int_{-\infty}^{\infty} |u|p(u)du \leq \sigma$ for some $\sigma > 0$.

Network:

- Directed and strongly connected network that admits a non-negative and doubly stochastic weight matrix $\boldsymbol{W} = [w_{ij}]$.

Main result

Theorem (Yu, Jakovetic, and Kar 2023)

Let the above assumptions hold and $\kappa := L/\mu$. Take clipping parameters φ_t, ϵ_t, error feedback wight β_t, and step size η_t as

$$\varphi_t = \frac{c_\varphi}{\sqrt{t+1}},\ \epsilon_t = \tau(t+1)^{3/5},\ \beta_t = \frac{c_\beta}{\sqrt{t+1}},\ \eta_t = \frac{c_\eta}{(t+1)^{1/5}}, \tag{1}$$

where constants $c_\varphi, \tau, c_\beta, c_\eta$ satisfy that $c_\eta^2 c_\varphi^2 = \Theta(\sigma^2/\mu^2)$, $0 \le c_\beta < 1$, and $\tau = \Theta(d\sigma^2\kappa^{5/2})$. Then, there exists some constant $c_s = \Omega\big(d^{-1/2}\kappa^{-5/4}\big)$ such that, for each $i \in [n]$, the iterates $\{\boldsymbol{x}_i^t\}$ satisfy that for any $0 < \delta < (c_s, 2/5)$, $\lim_{t\to\infty}(t+1)^\delta \mathbb{E}\big[\|\boldsymbol{x}_i^t - \boldsymbol{x}^*\|^2\big] = 0$.

Corollary

Let $\boldsymbol{W} = (1/n)\mathbf{1}\mathbf{1}^\top$, and $\boldsymbol{x}^t = \boldsymbol{x}_i^t, \forall i \in [n]$, then the update of $\boldsymbol{x}^t$ specializes into the server/client setup and the above result applies.

Key idea

- Linearize the nonlinear operator:

$$\boldsymbol{\Phi}_t(\nabla f_i(\boldsymbol{x}_i^t) - \boldsymbol{m}_i^t) = \boldsymbol{H}_{\Phi,i}^t(\nabla f_i(\boldsymbol{x}_i^t) - \boldsymbol{m}_i^t).$$

- Let $\zeta_{k,i}^t := [\nabla f_i(\boldsymbol{x}_i^t) - \boldsymbol{m}_i^t]_k$. The k-th diagonal $[\boldsymbol{H}_{\Phi,i}^t]_{kk}$ is

$$\int_{-\infty}^{\infty} \frac{c_\varphi}{\sqrt{t+1}} \frac{p(u)}{\sqrt{\left(\zeta_{k,i}^t + u\right)^2 + \epsilon_t}} du + \frac{1}{\zeta_{k,i}^t} \int_{-\infty}^{\infty} \frac{c_\varphi}{\sqrt{t+1}} \frac{up(u)}{\sqrt{\left(\zeta_{k,i}^t + u\right)^2 + \epsilon_t}} du.$$

- All the diagonal entries of $\boldsymbol{H}_{\Phi,i}^t$ fall into the interval $\left[c_1/(t+1)^{4/5}, c_2/(t+1)^{4/5}\right]$ for sufficiently large t and $c_1 < c_2$.

Discussions

Contributions:

- As a decentralized heavy-tailed noise robust method, in contrast to the prior art (Sun and Chen 2024):
  - Relax the bounded gradient condition $\|\nabla f_i(\boldsymbol{x}_i^t)\| \leq C_0$ for some $C_0 > 0$.
  - Allow any moment bound $\alpha \geq 1$.
  - Provide convergence rate analysis.
- For distributed gradient clipping in general, we provide the first MSE convergence method without assuming bounded gradient or bounded noise (Gorbunov et al. 2024).

Limitations and future directions:

- Noise symmetry may not hold in general.
- Convergence for non-convex functions.

Symmetric noise

- Empirical resemblance between layer norm and stable α-distribution (Barsbey et al. 2021).

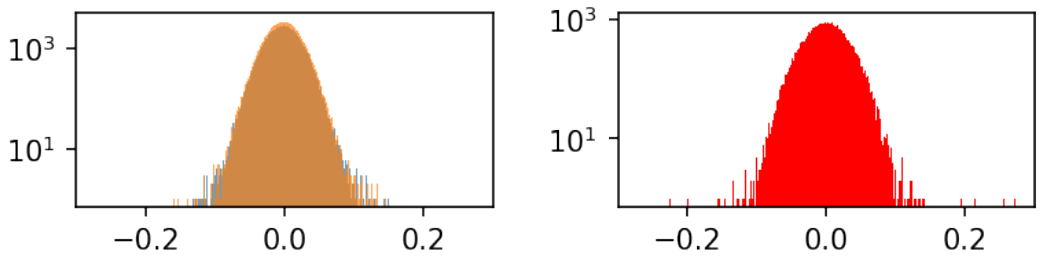
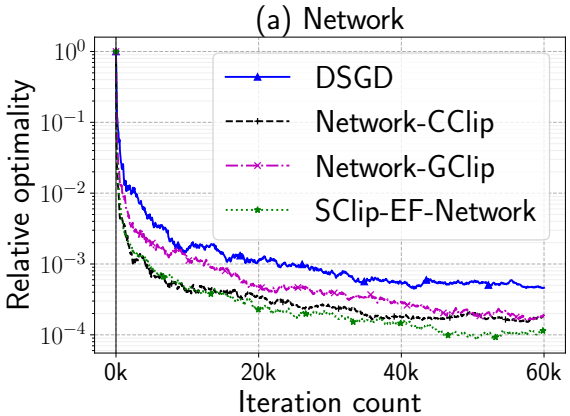


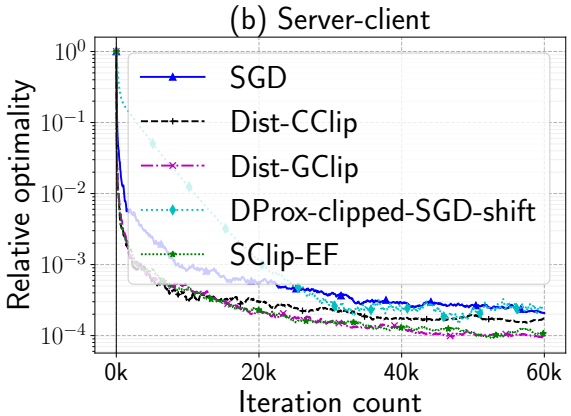
Figure: Empirical distribution of a CNN layer trained on MNIST. (Left) Overlaid histograms of a random partition of the weights, showing an identical distribution. (Right) Comparing the network weights to samples simulated i.i.d. from a symmetric α-stable distribution with the same tail index $\alpha \approx 1.95$.

- Theoretical evidence: gradient noise converges to a heavy-tailed α-stable random variable using generalized CLT (Simsekli, Sagun, and Gurbuzbalaban 2019).

Experiments

- Function: $\min_{\boldsymbol{x}\in\mathbb{R}^d} \sum_{i=1}^n \left(\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{b}_i^\top \boldsymbol{x}\right)$.
- Noise: $p(u) = \frac{c_p}{(u^2+2)\ln^2(u^2+2)}$ for normalization constant c_p.





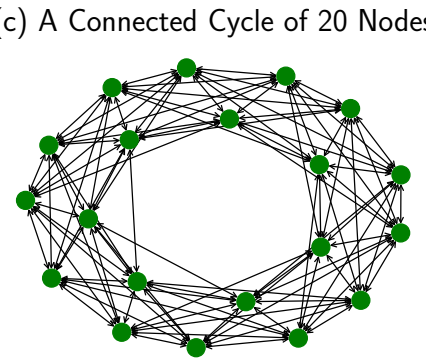


Figure: Average relative optimality $\log_{10}(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))/(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*))$ out of 10 runs in network and server-client cases, and network graph, from left to right.

Thanks

- Shuhua Yu, Dusan Jakovetic, and Soummya Kar. "Smoothed Gradient Clipping and Error Feedback for Distributed Optimization under Heavy-Tailed Noise." arXiv preprint arXiv:2310.16920 (2023).
- Aleksandar Armacki, Shuhua Yu, Sharma, Pranay Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, Soummya Kar "Nonlinear Stochastic Gradient Descent and Heavy-tailed Noise: A Unified Framework and High-probability Guarantees." arXiv preprint arXiv:2410.13954 (2024).