BRAIn Lab · SBER AI Lab · Università di Genova · AARHUS UNIVERSITY · Mohamed bin Zayed University of Artificial Intelligence

Problem Setup

We consider the stochastic convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

- **Landscape:** function f is convex and L-smooth
- **Oracle**: for given $x \in \mathbb{R}^d$ returns stochastic gradient $\nabla f_\xi(x)$
- **Standard example**: $f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)]$, $f_\xi(x)$ – loss on the data sample ξ

Assumptions:

- **A1:** f is L-smooth, i.e., $\forall x, y \in \mathbb{R}^d \;\; \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
- **A2:** f is convex, i.e., $\forall x, y \in \mathbb{R}^d \;\; f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$
- **A3:** $\forall x \in \mathbb{R}^d \;\; \mathbb{E}[\nabla f_\xi(x)] = \nabla f(x)$ and $\mathbb{E}[\|\nabla f_\xi(x) - \nabla f(x)\|^\alpha] \leq \sigma^\alpha$ with $\alpha \in (1, 2]$

Why Heavy-Tailed Noise?

Heavy tails:

$$\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^\alpha\right] \leq \sigma^\alpha \quad \text{with} \quad \alpha \in (1, 2].$$

✓ Observed for image classification: Simsekli et al. (2019) [1]

✓ Validated for large language models (LLMs): Zhang et al. (2020) [2]

✓ Appears in reinforcement learning (RL): Garg et al. (2021) [3]

! Less restrictive than the classical assumption on light-tailed stochasticity [4], which is often unrealistic in practice:

$$\mathbb{E}\left[\exp\left(\|\nabla f_\xi(x) - \nabla f(x)\|^2 / \sigma^2\right)\right] \leq \exp(1)$$

Gradient Clipping: Tool for Handling Heavy-Tailed Noise

We investigate the following algorithm:

$$x_{t+1} = x_t - \gamma_t g_t, \quad g_t := \min\left\{1, \frac{\lambda_t}{\|\nabla f_\xi(x_t)\|}\right\} \nabla f_{\xi_t}(x_t) \qquad \text{(Clip-SGD)}$$

Intuition: clipping prevents SGD from large steps induced by large noise [2, 5]

✓ Has been extensively studied [2, 5, 6, 7]

Motivation

✓ Last-iterate convergence is more relevant in practice

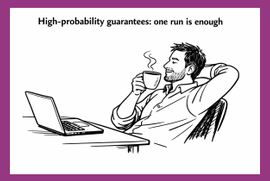
✗ High-probability last-iterate guarantees for Clip-SGD under heavy-tailed noise *have not been established previously.*

🤔 Almost all high-probability convergence results are established for constant stepsize that depend on the total number of iterations → can one develop a **horizon-free** analysis?

🤔 Can one derive **in-expectation** bounds from a high-probability analysis for methods with *bounded updates*?

Motivated by this, we present our main results!

High-Probability Bounds for the Last Iterate of Clipped SGD



TL;DR:

- **Horizon-free approach**
- **High-probability convergence**
- **Last-iterate bound**

Savelii Chezhegov
Daniella A. Parletta
Andrea Paudice
Eduard Gorbunov

Convergence Guarantees and Technical Novelty

ICLR

Main result: High-probability bounds

With appropriate choice of stepsizes γ_k and clipping thresholds λ_k, one can derive

$$f(x_K) - f^* = \widetilde{\mathcal{O}}\left(\frac{LR_0^2}{K} + \frac{D}{K^{2(\alpha-1)/3\alpha}}\right)$$

with probability at least $1 - \delta$.

! Convergence rate $\mathcal{O}\left(\frac{1}{K^{2(\alpha-1)/3\alpha}}\right)$ is *suboptimal* – interesting question for the future research

Highlights of the proof

- Potential function

$$\Phi_k = d_k(f(x_k) - f^*) + L\|x_k - x^*\|^2, \quad d_{k+1} = d_k + 2\gamma_k L, \; d_0 = 0$$

- Horizon-free choice of parameters:

$$\gamma_k, \lambda_k \sim K \quad \text{vs.} \quad \gamma_k, \lambda_k \sim k$$

- Modified clipping threshold: $\lambda_k \sim \frac{1}{\sqrt{d_k}}$ – new dependency for convex case
- Modified induction step:

Core changes

Logarithmically expanding compact set

Failure probability at step $k \sim \frac{\delta}{k^2}$

Additional result: In-expectation convergence

Key idea: high-probability bounds + bounded steps ⇒ in-expectation results!

$$\mathbb{E}[\text{Criterion(K)}] \leq (1 - \delta)(\text{High-prob bound}) + \delta(\text{bounded steps}) \qquad \text{for every } \delta$$

With certain choice of stepsizes γ_k, clipping thresholds λ_k and δ, one can derive

$$\mathbb{E}[f(x_K) - f^*] = \widetilde{\mathcal{O}}\left(\frac{LR_0^2}{K} + \frac{D}{K^{2(\alpha-1)/3\alpha}}\right).$$

Experiments

- Setup: logistic/linear regression and statistical learning
- Noise is sufficiently heavy-tailed in this case (validated in the paper)
- ✓ Last iterate is *better* and *easier to estimate* than the average iterate

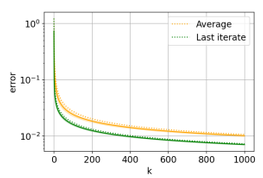


Figure 1: Logistic loss

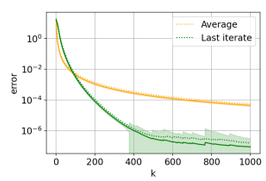


Figure 2: Quadratic loss

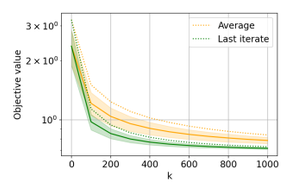


Figure 3: Statistical loss

References

1. Simsekli et al. *A tail-index analysis of stochastic gradient noise in deep neural networks* (ICML 2019)
2. Zhang et al. *Why gradient clipping accelerates training: A theoretical justification for adaptivity* (ICLR 2020)
3. Garg et al. *On proximal policy optimization's heavy-tailed gradients* (ICML 2021)
4. Nemirovski et al. *Robust stochastic approximation approach to stochastic programming* (SIOPT 2009).
5. Gorbunov et al. *Stochastic optimization with heavy-tailed noise via accelerated gradient clipping* (NeurIPS 2020)
6. Nguyen et al. *Improved convergence in high probability of clipped gradient methods with heavy tailed noise* (NeurIPS 2023)
7. Sadiev et al. *High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance* (ICML 2023).