Breaking the Heavy-Tailed Noise Barrier in Stochastic Optimization Problems

Nikita Puchkin [*, 1, 2] Eduard Gorbunov [*, 3] Nikolay Kutuzov [4] Alexander Gasnikov [4, 5, 6]

[*] Equal contribution

[1]HSE University [2]IITP RAS [3]MBZUAI [4]MIPT [5]University Innopolis [6]ISP RAS

Stochastic Optimization

Unconstrained minimization

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

with stochastic first-order oracle:

$\nabla f_\xi(x)$ – an estimate of $\nabla f(x)$

Example: expectation minimization

$$\min_{x \in \mathbb{R}^d} \{f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)]\}$$

Standard noise models:

- Sub-Gaussian noise:

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

- Bounded variance:

$$\mathbb{E}\|\nabla f_\xi(x) - \nabla f(x)\|^2 \le \sigma^2$$

- Bounded α-th moment, $\alpha \in (1, 2]$:

$$\mathbb{E}\|\nabla f_\xi(x) - \nabla f(x)\|^\alpha \le \sigma^\alpha$$

Assumptions on the Objective

We introduce all assumptions on $B_{3R}(x^*) := \{x \in \mathbb{R}^d \mid \|x - x^*\| \le 3R\}$ and $R \ge \|x^0 - x^*\|$.

L-smoothness: $\forall x, y \in B_{3R}(x^*)$

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$$
$$\|\nabla f(x)\|^2 \le 2L(f(x) - f(x^*))$$

For accelerated case we also need

μ-strong convexity: $\forall x, y \in B_{3R}(x^*)$

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$$

For non-accelerated case we need

μ-quasi strong convexity: $\forall x \in B_{3R}(x^*)$

$$f(x^*) \ge f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2$$

High-Probability Convergence

In-expectation guarantees:

$$\mathbb{E}[f(x) - f(x^*)] \le \varepsilon$$

High-probability guarantees:

$$\mathbb{P}\{f(x) - f(x^*) \le \varepsilon\} \ge 1 - \delta$$

- more accurate than in-expectation ones

Optimal high-probability complexity, bounded α-th moment noise, $\alpha \in (1, 2]$:

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}} + \left(\frac{\sigma^2}{\mu\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right) \quad [1]$$

But can we have **better complexities** for heavy-tailed noise?

Warmup: Symmetric Noise

Assumption 1

For all $u \in \mathbb{R}$ and $j = 1, \ldots, d$

- p_j – PDF of the j-th component of the noise: $v_j = [v]_j = [\nabla f_\xi(x) - \nabla f(x)]_j$
- $\mathsf{p}_j(u) = \mathsf{p}_j(-u)$
- $\mathsf{p}_j(u) \le {}^{B}/_{\max\{1, |u|^{\beta+1}\}}$, $B > 0$, $\beta > 0$

Cauchy distribution meets Assumption 1:

$$\mathsf{s}_j(u) = \frac{1}{\pi} \cdot \frac{1}{1 + u^2}, \quad \beta = 1$$

Median properties

Fix any $j \in \{1, \ldots, d\}$ and assume that the marginal density of v_j satisfies Assumption 1. Let $v_{j,1}, \ldots, v_{j,(2m+1)}$ be independent copies of v_j. If $m > 3/\beta$, then $\mathbb{E}\,\mathtt{Med}(\nu_{j,1}, \ldots, \nu_{j,(2m+1)}) = 0$ and $\mathbb{E}\,\mathtt{Med}\left(v_{j,1}, \ldots, v_{j,(2m+1)}\right)^2$ is finite.

Main contributions

- **Novel stochastic optimization setup**
  - informal: heavy-tailed symmetric part + antisymmetric part with bounded variance
  - we cover the case of $\mathbb{E}_\xi\|\nabla f_\xi(x)\| = +\infty$
- **High-probability complexities breaking the lower bounds**
  - new high-probability upper bounds for versions of clipped-SGD and clipped-SSTM
  - key idea: use median / smoothed median of means in clipped-SGD and clipped-SSTM
  - our results match SOTA ones under the bounded variance assumption for symmetric noise
- **New non-asymptotic results for the smooth median of means**

New Setup

Assumption 2

For all $u \in \mathbb{R}$ and $j = 1, \ldots, d$

$$\int_{-\infty}^{+\infty} u\mathsf{r}_j(u)\,du = 0, \quad \int_{-\infty}^{+\infty} u^2|\mathsf{r}_j(u)|\,du \le M_j \quad \text{and} \quad \mathsf{s}_j^{*k}(u) \le \frac{Bk}{k^{(\beta+1)/\beta} + |u|^{1+\beta}}, \quad \text{where}$$

- p_j – PDF of the j-th component of the noise: $v_j = [v]_j = [\nabla f_\xi(x) - \nabla f(x)]_j$
- Antisymmetric part of PDF – $\mathsf{r}_j(u)$, symmetric part of PDF – $\mathsf{s}_j(x)$:

$$\mathsf{r}_j(u) = \frac{\mathsf{p}_j(u) - \mathsf{p}_j(-u)}{2} \quad \text{and} \quad \mathsf{s}_j(x) = \frac{\mathsf{p}_j(x) + \mathsf{p}_j(-x)}{2}$$

- Convolution: $g^{*k}(x) = \underbrace{g * \ldots * g}_{k \text{ times}}(x)$, where $g * h(x) = \int_{\mathbb{R}^d} g(x - y)h(y)\,dy$
- $M \ge 0$, $B > 0$, $\beta \ge 1$

Distributions satisfying Assumption 2:

- Cauchy distribution + any distribution with bounded second moment (e.g., Pareto): $\beta = 1$
- Symmetric α-stable distribution with $\alpha \in [1, 2]$: $\beta = 1$

Assumptions 1 and 2 allow $\mathbb{E}\|\nabla f_\xi(x)\| = +\infty$

Smoothed Median of Means: Properties

SMoM properties

Suppose that, for any $j \in \{1, \ldots, d\}$ and any $x \in \mathbb{R}^d$, the density of $v_j = \nabla f_\xi(x) - \nabla f(x)$ meets Assumption 2. Then, if $m > 2 + 3/\beta$ and $\theta^2 n \ge (2 \vee m^2)M$, it holds that

$$\mathbb{E}\left\|\mathtt{SMoM}_{m,n}\left(\nabla f_\xi(x), \theta\right) - \nabla f(x)\right\|^2 \lesssim md\left\{(1 + \theta^2) + \left(\frac{M}{\theta n}\right)^2\right\} + md\left\{\left(\frac{2^\beta B}{\beta n^{\beta-1}}\right)^{2/\beta} + \left(\frac{BM}{\theta n^\beta}\right)^{2/(\beta+1)}\right\},$$

$$\left\|\mathbb{E}\mathtt{SMoM}_{m,n}\left(\nabla f_\xi(x), \theta\right) - \nabla f(x)\right\| \lesssim \frac{(1 + \theta)mM\sqrt{d}}{\theta^2 n} + \frac{mM\sqrt{d}}{\theta^2 n}\left(\frac{2^\beta B}{n^{\beta-1}}\right)^{1/\beta}$$

Main Results for Stochastic Optimization

Convergence of clipped-SGD

Assumption 1 + L-smoothness + μ-quasi strong convexity: $f(\overline{x}^K) - f(x^*) \le \varepsilon$ with prob. $\ge 1 - \delta$ after

$$\tilde{\mathcal{O}}\left(\frac{L}{\mu} + \frac{\sigma^2}{\mu\varepsilon}\right) \text{ iterations}$$

- $\nabla f_\Xi(x)$ is $\mathtt{Med}$ of $\mathcal{O}(3/\beta)$ i.i.d. samples

Assumption 2 + L-smoothness + μ-quasi strong convexity: $f(\overline{x}^K) - f(x^*) \le \varepsilon$ with prob. $\ge 1 - \delta$ after

$$\tilde{\mathcal{O}}\left(\frac{L}{\mu} + \frac{(1 + \theta^2)d + D}{\mu\varepsilon}\right) \text{ iterations}$$

- $\nabla f_\Xi(x)$ is $\mathtt{SMoM}$ of $\mathcal{O}(1/\varepsilon)$ i.i.d. samples
- D – some constant depending on M, B, β, d, n

Convergence of clipped-SSTM

Assumption 1 + L-smoothness + μ-strong convexity: $f(\hat{x}^\tau) - f(x^*) \le \varepsilon$ with prob. $\ge 1 - \delta$ after

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}} + \frac{\sigma^2}{\mu\varepsilon}\right) \text{ iterations}$$

- $\nabla f_\Xi(x)$ is $\mathtt{Med}$ of $\mathcal{O}(3/\beta)$ i.i.d. samples

Assumption 2 + L-smoothness + μ-strong convexity: $f(\hat{x}^\tau) - f(x^*) \le \varepsilon$ with prob. $\ge 1 - \delta$ after

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}} + \frac{(1 + \theta^2)d + D}{\mu\varepsilon}\right) \text{ iterations}$$

- Stage t: $\nabla f_\Xi(x)$ is $\mathtt{SMoM}$ of $\mathcal{O}(2^t)$ i.i.d. samples; # of stages: $\tau = \mathcal{O}(\log({}^{\mu R^2}/_{\varepsilon}))$
- D – some constant depending on M, B, β, d, n

Smoothed Median of Means

Let ζ be a random element in $\mathbb{R}^d$ and let $\theta > 0$ be an arbitrary number. For any positive integers m and n, the smoothed median of means $\mathtt{SMoM}_{m,n}(\zeta, \theta)$ is defined as follows:

$$\mathtt{SMoM}_{m,n}(\zeta, \theta) = \mathtt{Med}\left(v_1, \ldots, v_{2m+1}\right),$$

where, for each $j \in \{0, \ldots, 2m\}$,

$$v_j = \mathtt{Mean}(\zeta_{jn+1}, \ldots, \zeta_{(j+1)n}) + \theta\,\eta_{j+1},$$

$\zeta_1, \ldots, \zeta_{(2m+1)n}$ are i.i.d. copies of ζ, and $\eta_1, \ldots, \eta_{2m+1} \sim \mathcal{N}(0, \mathbf{I}_d)$ are independent standard Gaussian random vectors.

Algorithms

clipped-SGD [2]

$$x^{k+1} = x^k - \gamma_k \mathtt{clip}_{\lambda_k}(\nabla f_{\Xi^k}(x^k))$$

clipped-SSTM [3]

$$x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}},$$
$$z^{k+1} = z^k - \alpha_{k+1}\mathtt{clip}_{\lambda_{k+1}}(\nabla f_{\Xi^k}(x^{k+1})),$$
$$y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}$$

R-clipped-SSTM = Restarted clipped-SSTM

Gradient clipping:

$$\mathtt{clip}_\lambda(x) = \begin{cases} 0, & \text{if } x = 0, \\ \min\left\{1, \frac{\lambda}{\|x\|}\right\} x, & \text{if } x \ne 0 \end{cases}$$

Oracle:

- Assumption 1: $\nabla f_\Xi(x) = \mathtt{Med}$ of $\mathcal{O}(m)$ samples
- Assumption 2: $\nabla f_\Xi(x) = \mathtt{SMoM}$ (see above)

Experiments

Problem:

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}x^\top \mathbf{A}x, \quad \nabla f_\xi(x) = \mathbf{A}x + \xi$$

Noise models:

- Cauchy distribution
- $0.7\times$ Cauchy + $0.3\times$ exponential
- $0.7\times$ Cauchy + $0.3\times$ Pareto

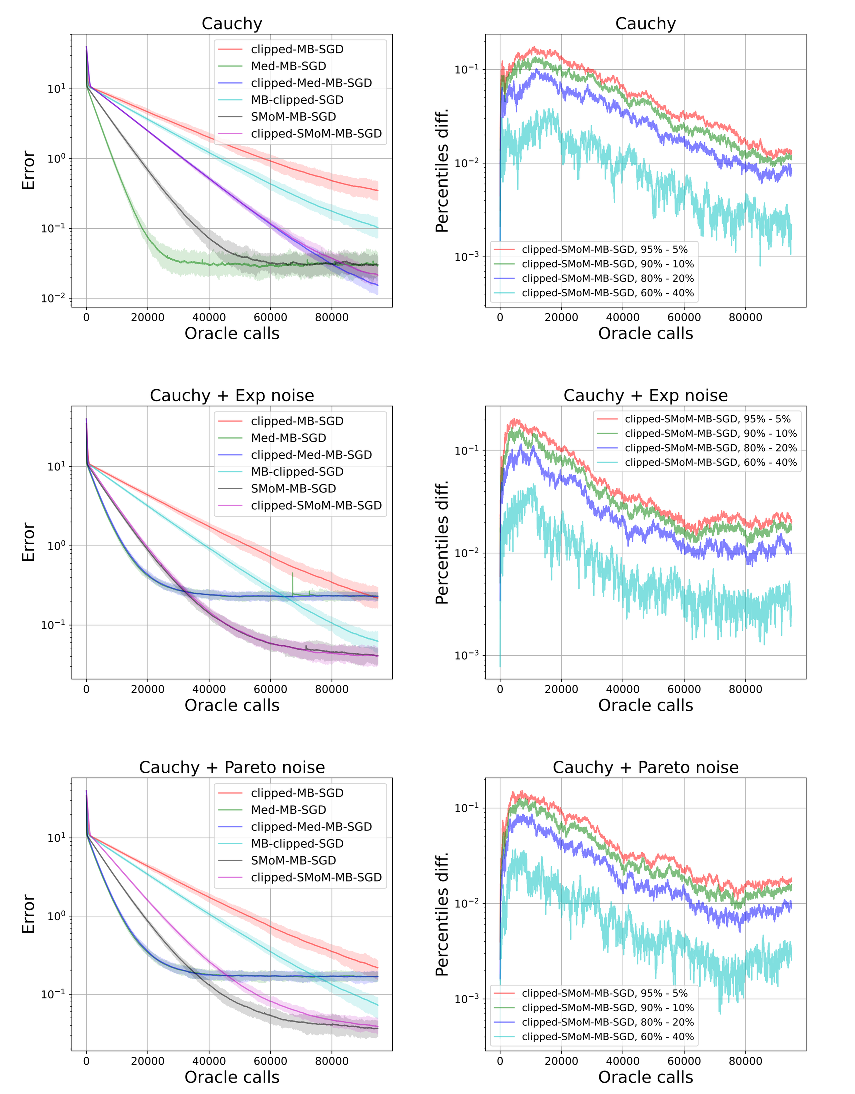

Figure: Left column: the mean error with a 95th and 5th percentile bounds. Right column: the confidence interval width for the error of mini-batched SGD with clipped smoothed median of means.

References

[1] A. Sadiev, M. Danilova, E. Gorbunov, S. Horváth, G. Gidel, P. Dvurechensky, A. Gasnikov, P. Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. *ICML 2023*.

[2] R. Pascanu, T. Mikolov, Y. Bengio. On the difficulty of training recurrent neural networks. *ICML 2013*.

[3] E. Gorbunov, M. Danilova, A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *NeurIPS 2020*.