Tight Lower Bounds and Optimal Algorithms for Stochastic Nonconvex Optimization with Heavy-Tailed Noise

Adrien Fradin Abdurakhmon Sadiev Laurent Condat Peter Richtárik

King Abdullah University of Science and Technology

ÉCOLE POLYTECHNIQUE

INSTITUT POLYTECHNIQUE DE PARIS

Heavy-Tailed Stochastic Optimization

$$\min_{x\in\mathbb{R}^d} \quad F(x) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f(x,\xi)\right].$$

A.1. p-th Bounded Central Moment (p-BCM) for gradients, i.e., $\exists\ \sigma > 0$ such that $\mathbb{E}\left[\nabla f(x,\xi)\right] = \nabla F(x)$,

$$p \in (1,2], \qquad \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|^p\right] \le {\sigma_1}^p.$$

Example: Two-sided Pareto distribution: $\Pr(|\xi| \ge s) \sim s^{-\alpha}$, $p < \alpha$.

A.2. q-th Bounded Central Moment (q-BCM) for Hessian, i.e., $\exists\ \sigma_h > 0$ such that for any $x \in \mathbb{R}^d$ we have $\mathbb{E}\left[\nabla^2 f(x,\xi)\right] = \nabla^2 F(x)$,

$$q \in [1,2], \qquad \mathbb{E}\left[\|\nabla^2 f(x,\xi) - \nabla^2 F(x)\|_{\text{op}}^q\right] \le {\sigma_2}^q.$$

Remark: **A.2.** is motivated by the work of Sadiev et al. [2025].

A.3. q-Weak Average Smoothness (q-WAS), i.e., $\exists\ \bar{L} \ge 0$ such that for any $x \in \mathbb{R}^d$ we have

$$q \in [1,2], \qquad \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla f(y,\xi)\|^q\right] \le \bar{L}^q \|x-y\|^q.$$

Remark: When $q = 2$, **A.3.** is a standard assumption for MVR/STORM.

A.4. (q,δ)-Similarity, i.e., $\exists\ q \in [1,2]$, and $\delta \ge 0$ such that, for all $x, y \in \mathbb{R}^d$ we have

$$\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla f(y,\xi) - [\nabla F(x) - \nabla F(y)]\|_{\text{op}}^q\right] \le \delta^q \|x-y\|^q.$$

Remark: When $q = 2$, **A.4.** is the well-known expected similarity assumption, which is popular in distributed optimization.

Motivation

Question

Can we achieve tight lower bounds and optimal sample complexity for stochastic nonconvex optimization under heavy-tailed noise by leveraging generalized q-WAS and (q,δ)-Similarity frameworks?

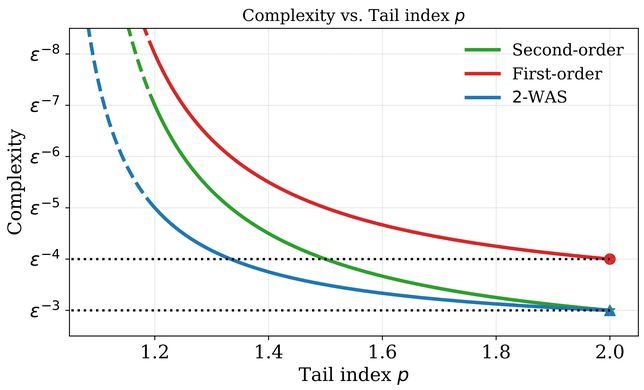

Figure 1: Sample complexity for FOSO, SOSO and 2-WAS vs. tail index p.

Main Contribution I

A.5. Lower Boundedness, i.e., the objective F is lower bounded:

$$F^* := \inf_{x\in\mathbb{R}^d} F(x) > -\infty$$

A.6. L_1-smoothness, i.e., $\exists\ L_1 \ge 0$ such that for all $x, y \in \mathbb{R}^d$

$$\|\nabla F(x) - \nabla F(y)\| \le L_1 \|x-y\|.$$

Lower Bounds

Theorem 1. There exists an optimization problem satisfying **A.1.**, **A.3.**, **A.5.**, where $F(x_1) - F^* \le \Delta$ and $\varepsilon \le \mathcal{O}(\sqrt{\bar{L}\Delta})$, such that for any first-order algorithm with multiple queries the complexity for finding a first order stationary point (i.e., $\mathbb{E}\left[\|\nabla F(x)\|\right] \le \varepsilon$) is lower bounded as

$$\Omega\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

Theorem 2. There exists an optimization problem satisfying **A.1.**, **A.4.**, **A.5.**, **A.6.**, where $F(x_1) - F^* \le \Delta$ and $\varepsilon \le \mathcal{O}(\sqrt{L_1\Delta})$, such that for any first-order algorithm with multiple queries the complexity for finding a first order stationary point (i.e., $\mathbb{E}\left[\|\nabla F(x)\|\right] \le \varepsilon$) is lower bounded as

$$\Omega\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{(L_1+\delta)\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

An important case is when $q = p$, then our lower bound is

$$\Omega\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{1}{(p-1)}}\right).$$

Can we design an algorithm handling heavy-tailed noise with the same optimal complexity ?

NSGD-MVR

$$x_{t+1} = x_t - \gamma\frac{g_t}{\|g_t\|};$$
$$g_{t+1} = (1-\alpha)\left(g_t + \nabla f(x_{t+1},\xi_{t+1}) - \nabla f(x_t,\xi_{t+1})\right) + \alpha\nabla f(x_{t+1},\xi_{t+1})$$

Remark. Normalization is needed to handle heavy-tailed noise [Hübler et al., 2025]. MVR accelerates convergence compared to Polyak's momentum.

Upper Bounds

Theorem 3. Under **A.1.**, **A.3.** and **A.5.**, **NSGD-MVR** guarantees that $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \le \varepsilon$ with total sample complexity

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

Theorem 4. Under **A.1.**, **A.4.**, **A.5.** and **A.6.**, **NSGD-MVR** guarantees that $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \le \varepsilon$ with total sample complexity

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{(L_1+\delta)\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

Main Contribution II

A.7. L_2-smoothness, i.e., $\exists\ L_2 \ge 0$ such that for all $x, y \in \mathbb{R}^d$

$$\|\nabla^2 F(x) - \nabla^2 F(y)\| \le L_2 \|x-y\|.$$

Lower Bound

Theorem 5. There exists an optimization problem satisfying **A.1.**, **A.2.**, **A.5.**, **A.6.**, **A.7.**, where $\varepsilon \le \mathcal{O}(\min\{\sqrt{L_1\Delta}, \sqrt[3]{L_2\Delta^2}\})$ and $F(x_1) - F^* \le \Delta$, such that for any second-order algorithm the complexity for finding a first-order stationary point (i.e., $\mathbb{E}\left[\|\nabla F(x)\|\right] \le \varepsilon$) is lower bounded as

$$\Omega\left(\min\left\{\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\sigma_2\Delta}{\varepsilon^2} + \frac{\sigma_2\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}} + \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}\right).$$

For simplicity, we present only the stochastic component of the complexity.

The first, second, and third terms of the lower bound correspond to **NSGDM**, **NIGT**, and **NSGDHess** (assuming $p = q$), respectively.

In the limit as $L_2 \to +\infty$ and $q = p$, the bound becomes:

$$\Omega\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\sigma_2\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2} + \frac{\sigma_2\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{1}{(p-1)}}\right).$$

NSGDHess

$$x_{t+1} = x_t - \gamma\frac{g_t}{\|g_t\|};$$
$$g_{t+1} = (1-\alpha)\left(g_t + \nabla^2 f(\hat{x}_{t+1},\hat{\xi}_{t+1})(x_{t+1}-x_t)\right) + \alpha\nabla f(x_{t+1},\xi_{t+1}),$$

where we sample $q_t \sim \mathcal{U}([0,1])$ and compute $\hat{x}_{t+1} = q_t x_{t+1} + (1-q_t)x_t$.

Effective implementation of $\nabla^2 f(x,\xi)\cdot v$:

$$\nabla_x\langle\nabla f(x,\xi), v\rangle = \nabla^2 f(x,\xi)\cdot v$$

Upper Bound

Theorem 6. Under **A.1.**, **A.2.**, **A.5.**, **A.6.** and **A.7.**, **NSGDHess** guarantees that $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \le \varepsilon$ with total sample complexity

$$\mathcal{O}\left(\frac{L_1\Delta}{\varepsilon^2} + \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}} + \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\sigma_2\Delta}{\varepsilon^2} + \frac{\sigma_2\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}} + \frac{L_2^{1/2}\Delta\sigma_1^{1/4}}{\varepsilon^{7/4}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right).$$

References

Zhang, J., Karimireddy, S.P., Veit, A., Kim, S., Reddi, S., Kumar, S. and Sra, S., 2020. Why are adaptive methods good for attention models?.In NeurIPS 2020.

Hübler, F., Fatkhullin, I. and He, N., 2025. From Gradient Clipping to Normalization for Heavy Tailed SGD. In AISTATS 2025.

Sadiev, A., Richtárik, P. and Fatkhullin, I., Second-order Optimization under Heavy-Tailed Noise: Hessian Clipping and Sample Complexity Limits. In NeurIPS 2025.