Accelerated Zeroth-order Method for Non-Smooth Stochastic Convex Optimization Problem with Infinite Variance

Nikita Kornilov[1,2], Ohad Shamir[3], Aleksandr Lobanov[1,4], Darina Dvinskikh[5,4], Alexander Gasnikov[1,4,2], Innokentiy Shibaev[1,6], Eduard Gorbunov[7], Samuel Horváth[7]

MIPT[1], Skoltech[2], Weizmann Institute of Science[3], ISP RAS[4], HSE University[5], IITP RAS[6], MBZUAI[7]

Problem Setup

We consider stochastic non-smooth convex optimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) \stackrel{\text{def}}{=} \mathbb{E}_{\xi \sim \mathcal{D}}\left[f(x, \xi)\right] \right\},$$

- $f(x, \xi)$ is $M_2(\xi)$-Lipschitz continuous in x w.r.t. Euclidean norm
- Samples ξ from unknown distribution $\mathcal{D}$ are available
- Zeroth-order two point oracle: for any $x, y \in \mathbb{R}^d$ we can compute $f(x, \xi)$ and $f(y, \xi)$ with the same ξ
- **Heavy-tailed noise:** oracle noise has bounded α-th moment, i.e., $\exists \alpha \in (1, 2], M_2 > 0$ such that $\mathbb{E}_\xi[M_2(\xi)^\alpha] \le M_2^\alpha$.

Motivation

- Various applications in medicine, biology, and physics: objective function is only computable through numerical simulation or the result of a real experiment
- Bandit optimization problem: the goal is to minimize average regret based only on observations of losses
- Reinforcement learning: black-box models parameters optimization via final reward of episode
- Hyperparameters optimization in the machine and deep learning models

Contributions

1. We propose the batched optimal accelerated algorithm that with
   - accuracy ε
   - problem dimension d
   - batchsize B
   - noise with bounded α-th moment
   - with high probability (e.i. $\forall \beta \in [0, 1]$ probability of achieving accuracy ε greater than $1 - \beta$)

   finds solution for convex function f after

   $$\sim \max\left(d^{\frac{1}{4}}/\varepsilon, \frac{1}{B}\left(\sqrt{d}/\varepsilon\right)^{\frac{\alpha}{\alpha-1}}\right) \quad \text{successive iterations,}$$

   $$\sim \left(\sqrt{d}/\varepsilon\right)^{\frac{\alpha}{\alpha-1}} \quad \text{oracle calls,}$$

   and for μ-strongly convex f after

   $$\sim \max\left(d^{\frac{1}{4}}/\left(\mu\varepsilon\right)^{\frac{1}{2}}, \frac{1}{B}\left(d/(\mu\varepsilon)\right)^{\frac{\alpha}{2(\alpha-1)}}\right) \quad \text{successive iterations,}$$

   $$\sim \left(d/(\mu\varepsilon)\right)^{\frac{\alpha}{2(\alpha-1)}} \quad \text{oracle calls.}$$

   Here we omitted $\log\frac{1}{\varepsilon}, \log\frac{1}{\beta}$ factors.
2. We prove a new batching result for the heavy-tailed noise case.

Methodology

Below, we overview the main steps in the construction of the optimal method

1. Implicitly build close smooth approximation $\hat{f}(x)$ for $f(x)$ based on *Randomized Smoothing*
2. Compute unbiased batched gradient estimation of $\hat{f}(x)$ via zeroth-order oracle
3. Minimize smoothed function $\hat{f}(x)$ via proper accelerated first-order algorithm
4. For μ-strongly convex functions we apply *restart technique*.

Randomized Smoothing [1]

Smooth approximation with parameter τ:

$$\hat{f}_\tau(x) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{u},\xi}[f(x + \tau\mathbf{u}, \xi)],$$

where $\mathbf{u} \sim U(B_2^d)$ is sampled from the uniform distribution on the unit Euclidean ball B_2^d.

1. Function $\hat{f}_\tau(x)$ is convex, M_2-Lipschitz , and satisfies

   $$\sup_{x \in \mathbb{R}^d} |\hat{f}_\tau(x) - f(x)| \le \tau M_2.$$

2. Function $\hat{f}_\tau(x)$ is differentiable with the following gradient

   $$\nabla \hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}}\left[\frac{d}{\tau} f(x + \tau\mathbf{e})\mathbf{e}\right], \tag{1}$$

   where $\mathbf{e} \sim U(S_2^d)$ is uniformly distributed on unit Euclidean Sphere S_2^d.

Batched gradient estimation:

$$g^B(x, \{\xi_i\}_i, \{\mathbf{e}_i\}_i) = \frac{d}{2B\tau}\sum_{i=1}^{B}(f(x + \tau\mathbf{e}_i, \xi_i) - f(x - \tau\mathbf{e}_i, \xi_i))\mathbf{e}_i. \tag{2}$$

In this setup, $g(x, \xi, \mathbf{e})$ has bounded (central) α-th moment (see [3]), i.e. $\mathbb{E}_{\xi,\mathbf{e}}[\|g(x, \xi, \mathbf{e}) - \mathbb{E}_{\xi,\mathbf{e}}[g(x, \xi, \mathbf{e})]\|^\alpha] \le \sigma^\alpha \stackrel{\text{def}}{=} \left(\sqrt{d}M_2/2^{\frac{1}{4}}\right)^\alpha$. To have a tight estimate of the (central) α-th moment of the batched estimate, we derive the following lemma.

Batching Lemma

For any sequence of i.i.d. random vectors $X_1, \dots, X_B \in \mathbb{R}^d$ with $\mathbb{E}[X_i] = x$ and bounded α-th moment $\mathbb{E}[\|X_i - x\|_2^\alpha] \le \sigma^\alpha, \alpha \in (1, 2]$ the next inequality holds

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} X_i - x\right\|_2^\alpha\right] \le \frac{\sigma^\alpha}{B^{\alpha-1}}.$$

Zeroth-order Algorithms

We use the Clipped Stochastic Similar Triangles Method (clipped-SSTM) from [2]. In order to cope with heavy-tailed noise it clips update vectors at a given level λ.

Algorithm 1 ZO-clipped-SSTM

Input: starting point x^0, number of iterations K, batch size B, stepsize $a > 0$, smoothing parameter τ, clipping levels $\{\lambda_k\}_{k=0}^{K-1}$.

Set $y^0 = z^0 = x^0$ and parameters $a, L = \sqrt{d}M_2/\tau$ of Clipped-SSTM

for $k = 0, \dots, K-1$ **do**

  Sample $\{\xi_i^k\}_{i=1}^B \sim \mathcal{D}$ and $\{\mathbf{e}_i^k\}_{i=1}^B \sim S_2^d$ independently.

  Compute $g^B(x^k, \xi^k, \mathbf{e}^k)$ as defined in (2).

  Perform a step of Clipped-SSTM with update vector g_k, clipping level λ_k and get points $x^{k+1}, y^{k+1}, z^{k+1}$

end for

Output: y^K

R-ZO-clipped-SSTM call ZO-clipped-SSTM with starting point $\hat{x}^t$, which is the output from the previous round for K_t iterations.

Algorithm 2 R-ZO-clipped-SSTM

Input: starting point x^0, number of restarts N, number of steps $\{K_t\}_{t=1}^N$, batchsizes $\{B_t\}_{t=1}^N$, stepsizes $\{a_t\}_{t=1}^N$, smoothing parameters $\{\tau_t\}_{t=1}^N$, clipping levels $\{\lambda_k^1\}_{k=0}^{K_1-1}, \dots, \{\lambda_k^N\}_{k=0}^{K_N-1}$

$\hat{x}^0 = x^0$.

for $t = 1, \dots, N$ **do**

  $\hat{x}^t =$ ZO-clipped-SSTM $\left(\hat{x}^{t-1}, K_t, B_t, a_t, \tau_t, \{\lambda_k^t\}_{k=0}^{K_t-1}\right)$.

end for

Output: $\hat{x}^N$

Deterministic noise:
We also allow deterministic absolutely bounded noise $\delta(x)$ with the following oracle

$$f_\delta(x, \xi) \stackrel{\text{def}}{=} f(x, \xi) + \delta(x), \quad |\delta(x)| \le \Delta.$$

Convergence rate remains the same if

- $\Delta \le \frac{\varepsilon^2}{M_2\sqrt{d}}$ for M_2-Lipschitz convex functions and $\Delta \le \frac{\mu^{1/2}\varepsilon^{3/2}}{\sqrt{d}M_2}$ for μ-strongly convex,
- $\Delta \le \frac{\varepsilon^{\frac{3}{2}}}{\sqrt{Ld}}$ for L-smooth convex functions and $\Delta \le \frac{\mu^{1/2}\varepsilon}{\sqrt{Ld}}$ for μ-strongly convex.

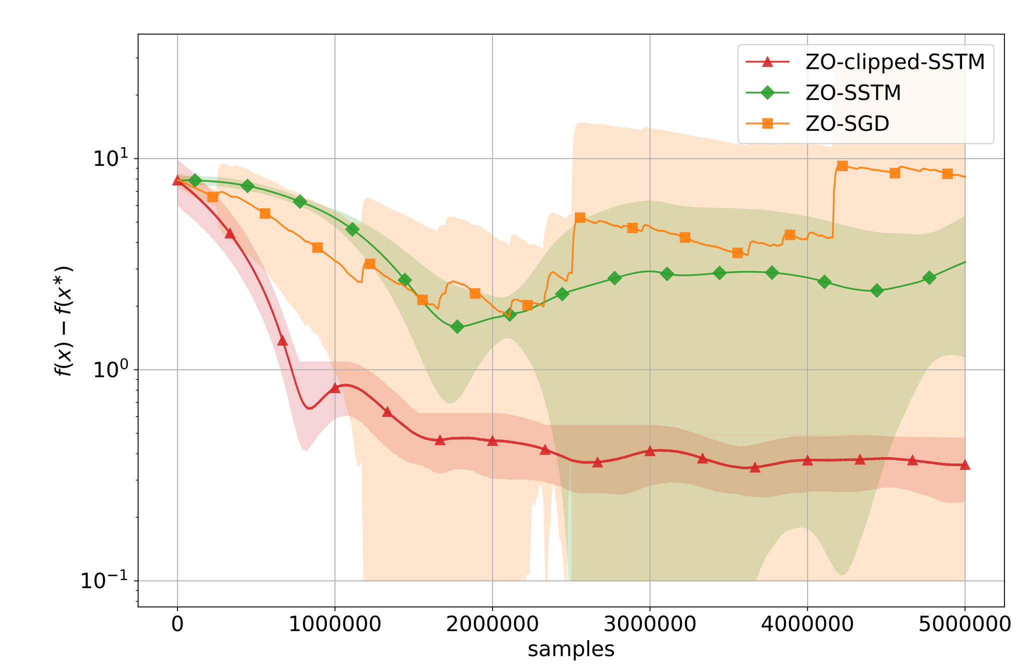
d dependency:

Open question: is the bound $\left(\sqrt{d}/\varepsilon\right)^{\frac{\alpha}{\alpha-1}}$ optimal in terms of the dependence on d?

Numerical Experiments

Methods ZO-SGD and ZO-SSTM are constructed from SGD and SSTM without clipping via the same methodology as ZO-clipped-SSTM.

The task was to minimize non-smooth $f(x) = \|Ax - b\|_2$ with heavy noise from symmetric Levy α-stable distribution with $\alpha = 3/2$.



Methods without clipping fail to converge due to the heavy tails in the distribution of the noise, while ZO-clipped-SSTM succeeds.

References

[1] Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu. The power of first-order smooth optimization for black-box non-smooth problems. *arXiv preprint arXiv:2201.12289*, 2022.

[2] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

[3] Nikita Kornilov, Alexander Gasnikov, Pavel Dvurechensky, and Darina Dvinskikh. Gradient free methods for non-smooth convex optimization with heavy tails on convex compact. *arXiv preprint arXiv:2304.02442*, 2023.

kornilov.nm@phystech.edu