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

Nikita Kornilov, Ohad Shamir, Aleksandr Lobanov,Darina Dvinskikh, Alexander Gasnikov, Innokentiy Shibaev, Eduard Gorbunov, Samuel Horvath

Skoltech

1 December, 2023

Problem statement

Problem statement

We consider stochastic non-smooth convex optimization problem

$$\min_{x \in Q} \left\{ f(x) \stackrel{\text{def}}{=} \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right] \right\}, \tag{1}$$

where Q either $\mathbb{R}^d$ or convex compact and ξ is from unknown distribution $\mathcal{D}$.

Problem statement

Problem statement

We consider stochastic non-smooth convex optimization problem

$$\min_{x \in Q} \left\{ f(x) \overset{\text{def}}{=} \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right] \right\}, \tag{1}$$

where Q either $\mathbb{R}^d$ or convex compact and ξ is from unknown distribution $\mathcal{D}$.

Function assumptions

Lipschitz: $f(x, \xi)$ is $M_2(\xi)$-Lipschitz continuous for any $\xi \in \mathcal{D}$:

$$|f(x, \xi) - f(y, \xi)| \leq M_2(\xi) ||x - y||_2, \quad x, y \in Q$$

Convex: $f(x, \xi)$ is convex on Q w.r.t. x for any $\xi \in \mathcal{D}$.

Oracle assumptions

Zeroth-order two-point oracle

We are able only to request $f(x, \xi)$ and $f(y, \xi)$ with the same ξ.

Oracle assumptions

Zeroth-order two-point oracle

We are able only to request $f(x, \xi)$ and $f(y, \xi)$ with the same ξ.

Heavy tails

There exist $\alpha \in (1, 2]$ and $M_2 > 0$ such that

$$\mathbb{E}_\xi[M_2(\xi)^\alpha] \leq M_2^\alpha.$$

For differentiable M_2-Lipschitz functions $\|\nabla f(x)\|_2 \leq M_2$.

Motivation

Examples

- Choosing proportion of ingredient in new soda:
  Ask each participant ξ to rate two drinks with different proportions x, y — give ratings $f(x, \xi), f(y, \xi)$.

Motivation

Examples

- Choosing proportion of ingredient in new soda:
  Ask each participant ξ to rate two drinks with different proportions x, y — give ratings $f(x, \xi), f(y, \xi)$.
- Various medicinal, biological and physical applications: numerical simulation or real experiment.
- Reinforcement learning: black-box models, simulation
- Bandit optimization problem: a learner x vs adversary ξ.
- Hyperparameters optimization in different machine and deep learning models

Previous solutions

Previous solutions

- Clipping technique for dealing with heavy tails was developed only for first-order methods
- Accelerated first-order methods weren't adopted to gradient-free setup
- In algorithms for heavy-tails batching technique wasn't developed at all

Our contribution

Contribution:

1. We propose the batched accelerated algorithm that copes with heavy-tailed noise finds a ε-solution with *high probability* and batchsize B after

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

   Bounds are optimal in terms of ε dependency.
2. For optimization on convex compact Q we adopt Mirror Descent Algorithm
3. Introduce batching theory for heavy-tailed samples

Main pipeline

Pipeline

1. Implicitly build close smooth approximation $\hat{f}$ of f based on *Smoothing Technique*.
2. Obtain unbiased batched gradient estimation of $\hat{f}(x)$ via zeroth-order oracle.
3. Minimize smoothed function $\hat{f}(x)$ via proper first-order algorithms which are robust to heavy-tailed noise

Smoothing Technique

Smooth approximation

$$\hat{f}_\tau(x) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{u},\xi}[f(x+\tau\mathbf{u},\xi)], \quad \text{where } \mathbf{u} \sim Uni(B_2^d).$$

Close approximation

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

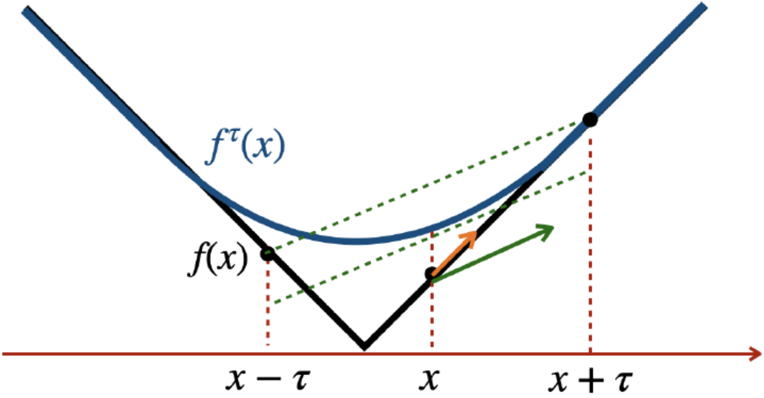
$$\sup_{x\in\mathbb{R}^d} |\hat{f}_\tau(x) - f(x)| \leq \tau M_2.$$

Smooth approximation

Function $\hat{f}_\tau(x)$ is differentiable and $\sqrt{d}M_2/\tau$ - smooth with the following gradient

$$\nabla \hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}}\left[\frac{d}{\tau} f(x+\tau\mathbf{e})\mathbf{e}\right], \quad \text{where } \mathbf{e} \sim Uni(S_2^d).$$

Smoothing example



Gradient estimation

Gradient estimation

We sample $\{\mathbf{e}_i\}_{i=1}^B \subset Uni(S_2^d)$ and $\{\xi_i\}_{i=1}^B \subset \mathcal{D}$ independently and construct

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

Gradient estimation

We sample $\{\mathbf{e}_i\}_{i=1}^B \subset Uni(S_2^d)$ and $\{\xi_i\}_{i=1}^B \subset \mathcal{D}$ independently and construct

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

Why $e \in Uni(S_2^d)$?

Lemma

Let $f(x)$ be M_2 Lipschitz continuous function w.r.t $\|\cdot\|_2$. If $\mathbf{e}$ is random and uniformly distributed on the Euclidean sphere and $\alpha \in (1, 2]$, then

$$\mathbb{E}_{\mathbf{e}}\left[\left(f(\mathbf{e}) - \mathbb{E}_{\mathbf{e}}[f(\mathbf{e})] \right)^{2\alpha} \right] \leq \left(\frac{bM_2^2}{\sqrt{2}} \right)^{\alpha}.$$

Frame Title

Boundness of α-th moment

$g^B(x, \{\xi\}, \{\mathbf{e}\})$ has bounded α-th moment, i.e., [1]

$$\mathbb{E}[\|g^B - \mathbb{E}[g^B]\|_2^\alpha] \leq \frac{2\sigma^\alpha}{B^{\alpha-1}},$$

where $\sigma^\alpha \overset{\text{def}}{=} \left(\sqrt{d}M_2/2^{\frac{1}{4}}\right)^\alpha$ — bound for one sample.

[1] Kornilov, N., Gasnikov, A., Dvurechensky, P., Dvinskikh, D. (2023). Gradient-free methods for non-smooth convex stochastic optimization with heavy-tailed noise on convex compact. Computational Management Science, 20(1), 37.

Batching Lemma

Lemma (Batching lemma)

For any sequence of i.i.d. r.vec. $X_1, \ldots, X_B \in \mathbb{R}^d$ *with the same* $\mathbb{E}[X_i] = x$ *and bounded* α*-th moment* $\mathbb{E}\|X_i - x\|_2^\alpha \leq \sigma^\alpha, \alpha \in (1, 2]$

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} X_i - x\right\|_2^\alpha\right] \leq \frac{2\sigma^\alpha}{B^{\alpha-1}}.$$

Dealing with heavy-tails

Clipping definition

Constant $\lambda > 0$ and update vector $g \in \mathbb{R}^d$

$$\texttt{clip}\left(g,\lambda\right) = \begin{cases} \frac{g}{\|g\|}\min\left(\|g\|,\lambda\right), & g \neq 0, \\ 0, & g = 0. \end{cases}$$

Lemma (Clipping properties)

Let G be a random vector in $\mathbb{R}^d$ with bounded α-th moment $\mathbb{E}[\|G-\mathbb{E}[G]\|^\alpha] \le \sigma^\alpha, \alpha \in (1,2]$. If $\widetilde{G} = \texttt{clip}(G,\lambda)$. Then,

1. $$\left\|\mathbb{E}[\widetilde{G}] - \mathbb{E}[G]\right\| \le \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}},$$
2. $$\mathbb{E}\left[\left\|\widetilde{G} - \mathbb{E}[\widetilde{G}]\right\|^2\right] \le 18\lambda^{2-\alpha}\sigma^\alpha.$$

SSTM

For optimization on whole space $\mathbb{R}^d$ we will use Stochastic Similar Triangles Method[2].

Algorithm SSTM (x^0, K, a, L)

1: Set $A_0 = \alpha_0 = 0, \;\; y^0 = z^0 = x^0$
2: **for** $k = 0, \dots, K-1$ **do**
3: Set $\alpha_{k+1} = {}^{k+2}/_{2aL}$, $A_{k+1} = A_k + \alpha_{k+1}$.
4: $x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}}$.
5: Get update vector g_{k+1}
6: $z^{k+1} = z^k - \alpha_{k+1} g_{k+1}$.
7: $y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}$.
8: **end for**

Output: y^K

---

[2] Gorbunov, E., Danilova, M., Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. Advances in Neural Information Processing Systems, 33, 15042-15053.

ZO-Clipped-SSTM

Algorithm ZO-clipped-SSTM

1: Set initial parameters of SSTM with $L = \sqrt{d}M_2/\tau$
2: **for** $k = 0, \ldots, K-1$ **do**
3: Sample $\{\xi_i^k\}_{i=1}^B \sim \mathcal{D}$ and $\{\mathbf{e}_i^k\}_{i=1}^B \sim S_2^d$ independently.
4: Compute $g^B(x^k, \xi^k, \mathbf{e}^k)$
5: Compute $\tilde{g}_k = \texttt{clip}\left(g^B(x^k, \xi^k, \mathbf{e}^k), \lambda_k\right)$
6: Perform a step of SSTM with update vector $\tilde{g}_k$
7: **end for**

Output: final point of SSTM

ZO-Clipped-SSTM Convergence

Theorem

ZO-Clipped-SSTM *with certain parameters* $\tau, a, \{\lambda_k\}$ *finds a* ε*-solution for convex* f*, batchsize* B *and* $R = \|x^0 - x^*\|$ *with high probability after* ($\tilde{O}$ *hides* log *factor*)

$$\sim \tilde{\mathcal{O}}\left(\max\left(\frac{M_2\sqrt[4]{d}R}{\varepsilon}, \frac{1}{B}\left(\frac{\sqrt{d}M_2R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right) \quad \textit{successive iterations,}$$

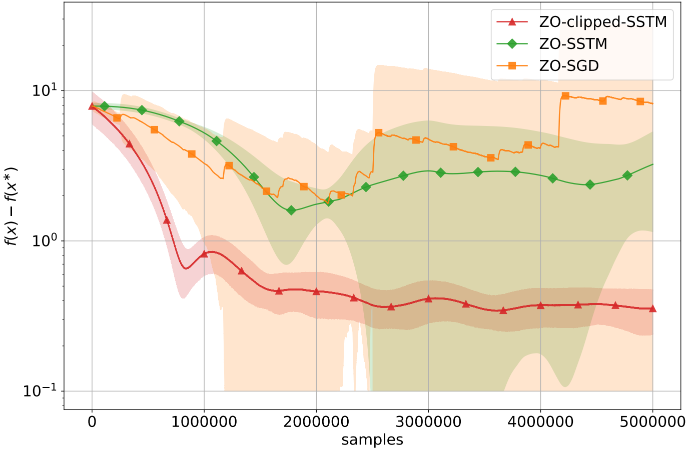
$$\sim \tilde{\mathcal{O}}\left(\frac{\sqrt{d}M_2R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}} \quad \textit{oracle calls.}$$

Corollaries:

1. Is dimension d dependency optimal?
2. Bound are optimal in terms of ε
3. Maximum $B = \left(\frac{\sqrt{d}M_2R}{\varepsilon}\right)^{\frac{1}{\alpha-1}}$

Experiments

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



μ-strongly convex case

μ-strongly convex

Function $f(x, \xi)$ is μ-strongly convex on Q for all $x \in Q$

$$\frac{\mu}{2}\|x - x^*\|_2^2 \leq f(x) - f(x^*)$$

Restarts (R-ZO-clipped-SSTM)

For N runs use output of ZO-clipped-SSTM as initial point
$\hat{x}^t = \text{ZO-clipped-SSTM}\left(\hat{x}^{t-1}, K_t, B_t, a_t, \tau_t, \{\lambda_k^t\}_{k=0}^{K_t-1}\right)$

R-ZO-Clipped-SSTM Convergence

Theorem

R-ZO-clipped-SSTM *with certain parameters* $N, K_t{}_{t=1}^{N}, \{a\}_{t=1}^{N}, \{\tau\}_{t=1}^{N}, \{\lambda_k\}$ *finds a* ε*-solution for* μ*-strongly convex* f, *batchsize* B *with high probability after* ($\tilde{O}$ *hides* log *factor)*

$$\widetilde{\mathcal{O}}\left(\max\left\{\sqrt{\frac{M_2^2\sqrt{d}}{\mu\varepsilon}}, \left(\frac{dM_2^2}{\mu\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right) \textit{ oracle calls.}$$

Adversarial noise

Zero-order oracle returns noisy approximation of $f(x, \xi)$

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

E.g., machine accuracy or accuracy of solving subtask

Boundedness of noise

There exists a constant $\Delta > 0$ such that $|\delta(x)| \leq \Delta$ for all $x \in Q$.

Gradient estimation

$$g^B(x, \xi, \mathbf{e}) = \frac{1}{B} \sum_{i=1}^{B} \frac{d}{2\tau} \left(f_\delta(x + \tau \mathbf{e}_i, \xi_i) - f_\delta(x - \tau \mathbf{e}_i, \xi_i) \right) \mathbf{e}_i.$$

Non-smooth case

Boundness of α-th moment

$$\mathbb{E}\left[\left\|g^B - \mathbb{E}[g^B]\right\|_2^\alpha\right] \leq \frac{2}{B^{\alpha-1}}\left(\frac{\sqrt{d}M_2}{2^{\frac{1}{4}}} + \frac{d\Delta}{\tau}\right)^\alpha.$$

Under which noise convergence rate remains the same? In theory $\tau \sim \frac{\varepsilon}{M_2}$, therefore

$$\text{Convex:} \quad \Delta \leq \frac{\varepsilon^2}{RM_2\sqrt{d}}$$

$$\mu\text{- strongly convex:} \quad \Delta \leq {}^{\mu^{1/2}\varepsilon^{3/2}}\!/\!\sqrt{d}M_2$$

Smooth case

L-smoothness

The function f is L-smooth, i.e., it is differentiable on Q and for all $x, y \in Q$ with $L > 0$:

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2$$

Approximation via Smoothing

$$\sup_{x \in Q} |\hat{f}_\tau(x) - f(x)| \leq \frac{L\tau^2}{2} \Longrightarrow \tau \sim \sqrt{\frac{\varepsilon}{L}}$$

Smooth case

Upper bounds for Δ

$$\text{Convex:} \quad \Delta \leq \frac{\varepsilon^{3/2}}{R\sqrt{dL}}$$

$$\mu\text{- strongly convex:} \quad \Delta \leq {\mu^{1/2}\varepsilon}/{\sqrt{dL}}$$

Corollaries:

1. Upper bounds are optimal in terms of ε
2. In smooth case finite coordinate-wise difference works better

Optimization on convex compact Q

Idea: exploit geometry of the convex compact Q to get better constants

Compact features

1. For $p \in [1,2]$, we use the l_p-norm, i.e. $\|x\|_p = \left(\sum_{k=1}^{d} |x_k|^p\right)^{1/p}$ and dual norm l_q where $1/q + 1/p = 1$.
2. We work with adversarial noise
3. Assumptions must hold true on a little bigger set than Q
4. Upper bound of α-th moment of g^B

$$\mathbb{E}[\|g^B - \mathbb{E}[g^B]\|_q^\alpha] \le \frac{2}{B^{\alpha-1}} \left(\frac{\sqrt{d} a_q M_2}{2^{\frac{1}{4}}} + \frac{d a_q \Delta}{\tau}\right)^\alpha = \frac{2}{B^{\alpha-1}} \sigma_q^\alpha$$

where $a_q \overset{\text{def}}{=} d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q-1}\}$.

Mirror descent

Let function $\Psi : \mathbb{R}^d \to \mathbb{R}$ be 1-stronfly convex w.r.t. the l_p-norm. Bregman divergence:

$$V_\Psi(y, x) = \Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y - x \rangle.$$

For a given stepsize ν and gradient g_{k+1}, the updates of SMD are defined as follows:

$$y_{k+1} = \nabla(\Psi^*)(\nabla\Psi(x_k) - \nu g_{k+1}), \quad x_{k+1} = \arg\min_{x \in Q} V_\Psi(x, y_{k+1}).$$

Using the assumptions on the function Ψ, it can be proved that the updates are well-defined and that $(\nabla\Psi)^{-1} = \nabla\Psi^*$, where Ψ^* is Fenchel conjugate.

ZO-Clipped-SMD

Algorithm ZO-Clipped-SMD (Ψ_p, T, ν)

1: $x_0 \leftarrow \arg\min_{x \in \mathcal{X}} \Psi_p(x)$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Compute $g^B(x^k, \xi^k, \mathbf{e}^k)$
4: Compute $\tilde{g}_k = \texttt{clip}\left(g^B(x^k, \xi^k, \mathbf{e}^k), \lambda_k\right)$
5: Perform a step of SMD $x_{k+1} = SMD\_Step(x_k, \tilde{g}_k)$
6: **end for**

Output: $\overline{x}_K \leftarrow \frac{1}{K} \sum_{k=0}^{K-1} x_k$

ZO-Clipped-SMD Convergence

Theorem

ZO-Clipped-SMD *with certain parameters* $\tau, \Psi_p, \nu, \{\lambda_k\}$ *for convex* f*, batchsize* B *with high probability after* K *iterations convergences as*

$$f(\overline{x}_K) - f(x^*) \leq 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}D_\Psi + \widetilde{O}\left(\frac{D_\Psi \sigma_q}{(B \cdot K)^{\frac{\alpha-1}{\alpha}}}\right),$$

where $D_\Psi^2 \overset{def}{=} 2 \sup\limits_{x,y \in Q} V_{\Psi_p}(x, y)$ *is compact's diameter.*

Convergence discussion I

1. Ball setup:
$$p = 2, \Psi_p(x) = \frac{1}{2}\|x\|_2^2$$
2. Entropy setup:
$$p = 1, \Psi_p(x) = (1+\gamma)\sum_{i=1}^{d}(x_i + \gamma/d)\log(x_i + \gamma/d), \gamma > 0$$

In order to achieve accuracy ε we have $K^{\frac{\alpha-1}{\alpha}}$ equals

1. For $\Delta_+^d = \{x \in^d : x \geq 0, \sum_i x_i = 1\}$ with Entropy setup — $\ln dM_2/\varepsilon$
2. For B_1^d with Entropy setup — $\ln dM_2/\varepsilon$
3. For B_2^d with Ball setup — $\sqrt{d}M_2/\varepsilon$
4. For B_∞^d with Ball setup — dM_2/ε

Convergence discussion II

Upper bound for Δ

$$\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$$

the same and optimal as before, Ball setup is always preferable

Restarts

Similarly to R-ZO-clipped-SSTM, we build restarted R-ZO-Clipped-SMD with the same oracle complexity and upper bounds for Δ. More details in the paper.

Non-linear multi-arm bandits

On step t we can choose strategy x_t from compact set $Q \subset \mathbb{R}^d$, e.g., probability of d arms.
Next adversary chooses action ξ_t and we receive reward $f_t(x_t, \xi_t)$, e.g. reward from pulling arms $\langle x_t, \xi_t \rangle$. But reward can be non-linear and heavy-tailed, and standard algorithms are not applicable.
We minimize pseudo-regret

$$\mathcal{R}_K(\{f_t(\cdot)\}, \{x_t\}) = \sum_{t=1}^{K} f_t(x_t) - \min_{x \in Q} \sum_{t=1}^{K} f_t(x).$$

ZO-Clipped-SMD for bandits

Idea: use ZO-Clipped-SMD with heavy-tailed losses

Features

- We have only one-point oracle and need additional assumption $\mathbb{E}_{\xi_t}[|f_t(x,\xi_t)|^\alpha] \leq G^\alpha$
- Gradient without batch

$$g_t(x_t, \xi_t, \mathbf{e}_t) = \frac{d}{\tau} f_t(x_t + \tau \mathbf{e}_t, \xi_t)\mathbf{e}_t$$

  with

$$\mathbb{E}[\|g^B\|_q^\alpha] \leq \left(\frac{da_q G}{\tau} + \frac{da_q \Delta}{\tau}\right)^\alpha$$

- Look at the rewards at points $x_t + \tau \mathbf{e}_t$, not x_t.
- Get optimal dependencies on ε

$$K = \tilde{O}\left(\frac{dG}{\varepsilon^2}\right)^{\frac{\alpha}{\alpha-1}}$$

Generalization

- r-growth functions
- Saddle-point problems
- Variational inequalities
- Composite problems
- Distributed problems

Questions?

Thank You For Your Attention!

References

- Kornilov, N., Shamir, O., Lobanov, A., Dvinskikh, D., Gasnikov, A., Shibaev, I., ... Horváth, S. (2023). Accelerated Zeroth-order Method for Non-Smooth Stochastic Convex Optimization Problem with Infinite Variance. arXiv preprint arXiv:2310.18763.
- Kornilov, N., Gasnikov, A., Dvurechensky, P., Dvinskikh, D. (2023). Gradient-free methods for non-smooth convex stochastic optimization with heavy-tailed noise on convex compact. Computational Management Science, 20(1), 37.
- Gorbunov, E., Danilova, M., Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. Advances in Neural Information Processing Systems, 33, 15042-15053.
- Dorn, Y., Nikita, K., Kutuzov, N., Nazin, A., Gorbunov, E., Gasnikov, A. (2023). Implicitly normalized forecaster with clipping for linear and non-linear heavy-tailed multi-armed bandits. arXiv preprint arXiv:2305.06743.