Improved Exploitation of Higher Order Smoothness in Derivative-free Optimization

Vasilii Novitskii

vasiliy.novitskiy@phystech.edu

Moscow Institute of Physics and Technology

July 17, 2021

Problem

Consider stochastic **convex** optimization problem with zero-order oracle

$$\min_{x \in Q} f(x)$$

We study functions $f : U_{\varepsilon_0}(Q) \to \mathbb{R}$ satisfying:

- $\|\nabla f(x) - \nabla f(y)\| \leqslant G\|x - y\|$ for all $x, y \in U_{\varepsilon_0}(Q)$ (Lipschitz continuity)
- generalized Hölder condition with parameter β (see section **Higher Order Smoothness**)
- $Q \subset \mathbb{R}^n$– closed convex set (Euclidean metrics)

Problem

The optimization problem can be formulated as follows: find the sequence $\{x_k\}_{k=1}^N \subset Q$ minimizing the average regret:

$$\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leqslant \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[f(x_k) - f(x^*)\right] \leqslant \varepsilon.$$

The main question: how does the number of iterations $N(\varepsilon)$ (oracle complexity) depend on accuracy ε, dimension n, strong convexity parameter γ (in the case the problem is **strongly convex**), higher order smoothness parameter β (**Higher Order Smoothness**)?

Noise

Zeroth Order Oracle and Noise

The function values $f(x_k + \tau_k r_k e_k)$ and $f(x_k - \tau_k r_k e_k)$ are given with additive **random** noise ξ_k^+ and ξ_k^-

$$\widetilde{g_k} = \frac{n\left(f(x_k + \tau_k r_k e_k) + \xi_k^+ - f(x_k - \tau_k r_k e_k) - \xi_k^-\right)}{2\tau_k} K(r_k) e_k$$

One-point oracle

We call the oracle one-point as $\xi_k^+ \neq \xi_k^-$.

Assumptions on Noise and Randomness

1. $\mathbb{E}[\xi_k^{+2}] \leqslant \sigma^2$ and $\mathbb{E}[\xi_k^{-2}] \leqslant \sigma^2$ where $\sigma \geqslant 0$;
2. random variables ξ_k^+ and ξ_k^- are independent from e_k and r_k
3. random variables e_k and r_k are independent.

Algorithm

Algorithm 1

Input: x_0, step size $\{\alpha_k\}_{k=0}^{N-1}$, parameters $\{\tau_k\}_{k=0}^{N-1}$

for $k = 0, \ldots, N-1$ **do**

1. generate uniformly $r_k \in U[-1;1]$, $e_k \in S_1^n(0)$
2. $$\widetilde{g_k} = \frac{n\left(f(x_k + \tau_k r_k e_k) + \xi_k^+ - f(x_k - \tau_k r_k e_k) - \xi_k^-\right)}{2\tau_k} K(r_k) e_k$$
3. $x_{k+1} := \mathrm{Proj}_Q(x_k - \alpha_k \widetilde{g_k})$

end

Return: $\overline{x}_N = \frac{1}{N}\sum_{k=1}^{N} x_k$

Higher Order Smoothness

Generalized Hölder condition

Let $\mathcal{F}_\beta(L)$ ($L > 0$) denote the set of all **convex** functions $f : \mathbb{R}^n \to \mathbb{R}$ which are differentiable l times ($l = \max\{k \in \mathbb{N} : k < \beta\}$) and which for all $x, z \in U_{\varepsilon_0}(Q)$ satisfy so called generalized Hölder condition with parameter β:

$$\left| f(z) - \sum_{0 \leqslant |m| \leqslant l} \frac{1}{m!} D^m f(x)[z - x]^m \right| \leqslant L\|z - x\|^\beta.$$

Higher Order Smoothness

$$\widetilde{g_k} = \frac{n\left(f(x_k + \tau_k r_k e_k) + \xi_k^+ - f(x_k - \tau_k r_k e_k) - \xi_k^-\right)}{2\tau_k} K(r_k) e_k$$

Smoothing Kernels

For gradient estimator $\widetilde{g_k}$ we use the kernel

$$K : [-1, 1] \to \mathbb{R},$$

satisfying

$$\mathbb{E}[K(r)] = 0, \mathbb{E}[rK(r)] = 1, \mathbb{E}[r^j K(r)] = 0, j = 2, ..., l, \mathbb{E}\left[|r|^\beta |K(r)|\right] \leqslant \infty,$$

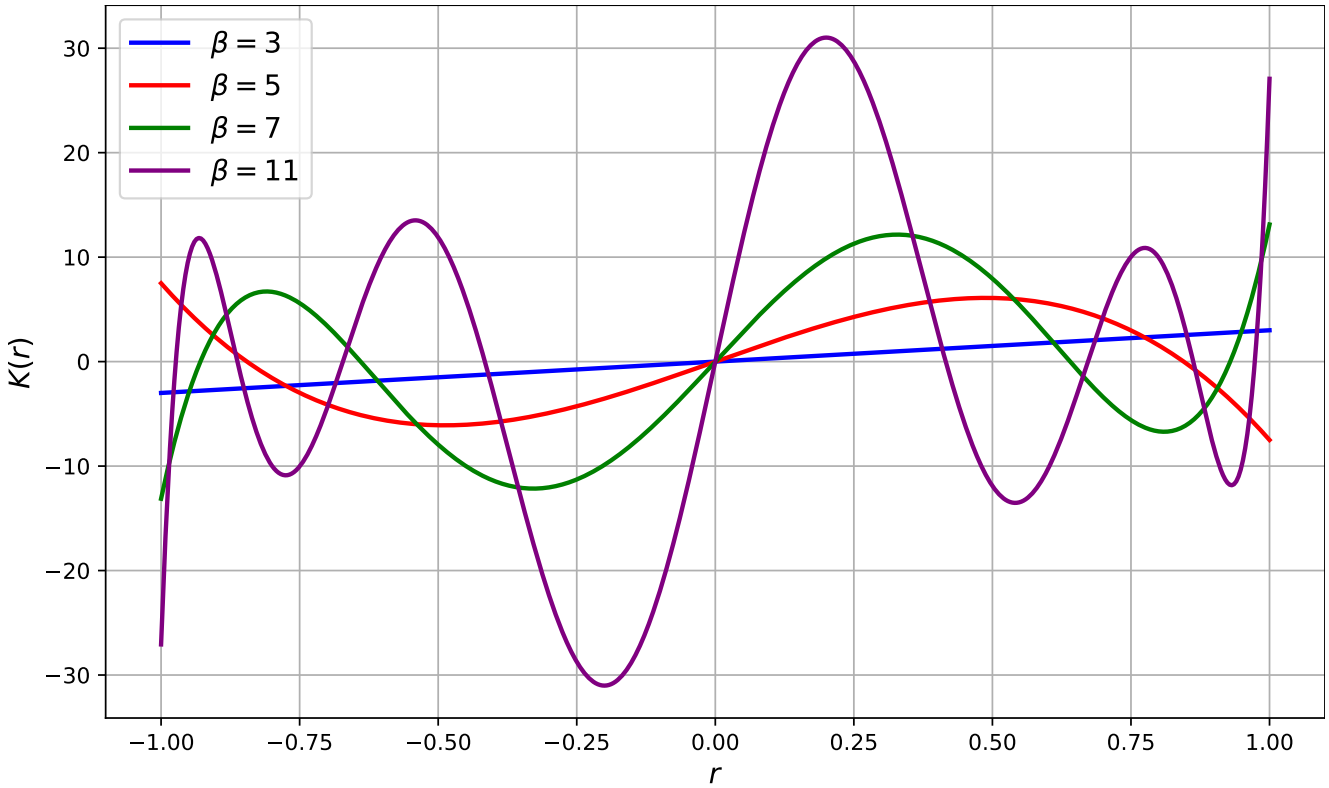
where r is a uniformly distributed on $[-1, 1]$ random variable. This helps us to get better bounds on the gradient bias $\|\widetilde{g_k} - \nabla f(x_k)\|$.

Kernels

We have the following kernels for different betas:

$$K_\beta(r) = 3r, \qquad \beta \in [2, 3],$$
$$K_\beta(r) = \frac{15r}{4}(5 - 7r^2), \qquad \beta \in (3, 5],$$
$$K_\beta(r) = \frac{105r}{64}(99r^4 - 126r^2 + 35), \qquad \beta \in (5, 7].$$

Kernels



Examples of kernels

Results

Table: The dependence of ε (optimization error) on N (number of iterations), n (dimension), γ, β

	strongly convex	convex
lower bound Akhavan, Pontil, Tsybakov (2020)	$\mathcal{O}\left(\min\left(\frac{n}{\gamma N^{\frac{\beta-1}{\beta}}}, \frac{n}{\sqrt{N}}\right)\right)$	$\mathcal{O}\left(\min\left(\frac{\sqrt{n}}{N^{\frac{\beta-1}{2\beta}}}, \frac{n}{\sqrt{N}}\right)\right)$
Novitskii (this work, 2020)	$\mathcal{O}\left(\frac{n^{2-\frac{1}{\beta}}}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$	$\mathcal{O}\left(\frac{n^{1-\frac{1}{2\beta}}}{N^{\frac{\beta-1}{2\beta}}}\right)$
Akhavan, Pontil, Tsybakov (2020)	$\mathcal{O}\left(\frac{n^2}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$	$\mathcal{O}\left(\frac{n}{N^{\frac{\beta-1}{2\beta}}}\right)$
Gasnikov and al. (2017), $\beta = 2$	$\tilde{\mathcal{O}}\left(\frac{n}{\sqrt{\gamma N}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$

The bounds marked in blue are not given in the references but they can be got.

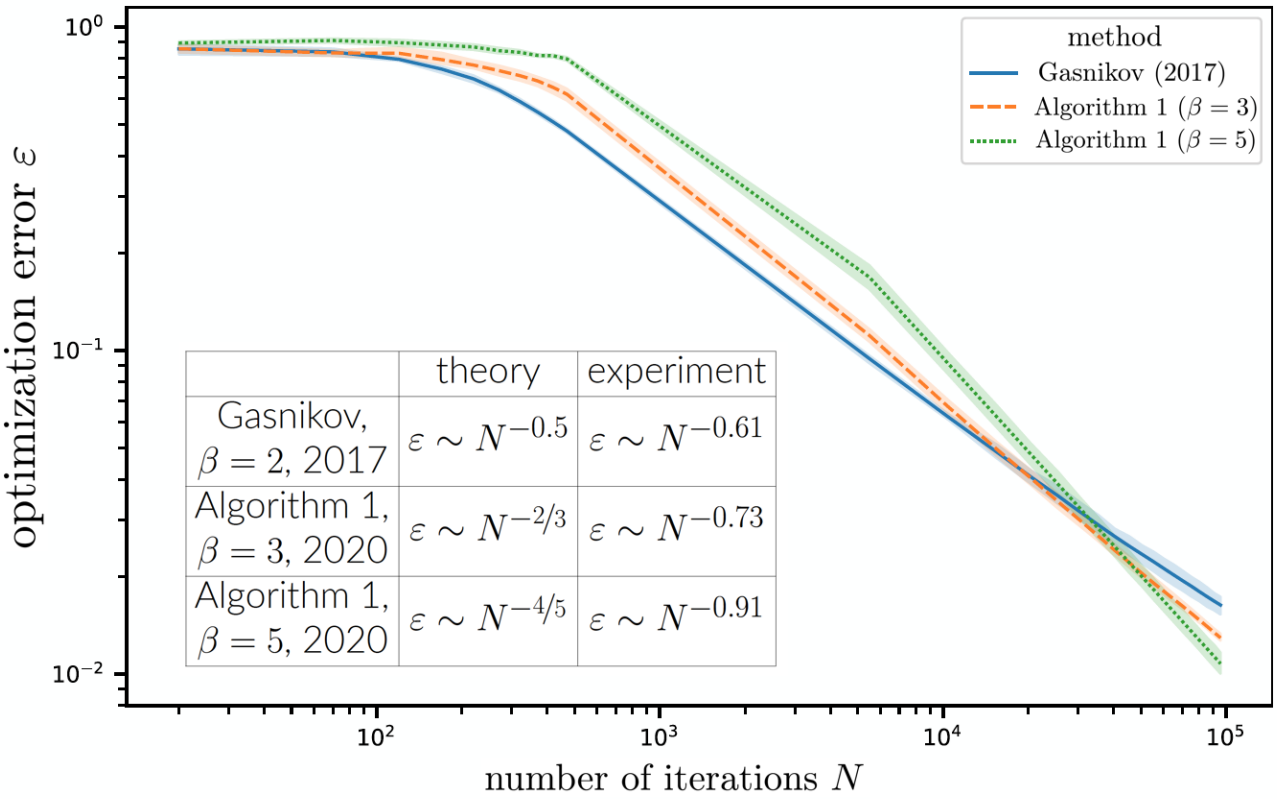
Numerical Experiment

We consider the problem of the minimization of the regularized quadratic function A with eigenvalues 0.25, 1, 4:

$$f(x) = \frac{1}{2}x^T Ax + \frac{1}{10}\sum_{k=1}^{50}|x_k|^4 \to \min_{x \in B_1(0)}.$$

The starting point is x_0 with $\|x_0\| = 1/2$.

Numerical Experiment



	theory	experiment
Gasnikov, $\beta = 2$, 2017	$\varepsilon \sim N^{-0.5}$	$\varepsilon \sim N^{-0.61}$
Algorithm 1, $\beta = 3$, 2020	$\varepsilon \sim N^{-2/3}$	$\varepsilon \sim N^{-0.73}$
Algorithm 1, $\beta = 5$, 2020	$\varepsilon \sim N^{-4/5}$	$\varepsilon \sim N^{-0.91}$

Dependence of optimization error ε of Algorithm 1 on iteration number N

Results

Table: The dependence of ε (optimization error) on N (number of iterations), n (dimension), γ, β

	strongly convex	convex
lower bound Akhavan, Pontil, Tsybakov (2020)	$\mathcal{O}\left(\min\left(\frac{n}{\gamma N^{\frac{\beta-1}{\beta}}}, \frac{n}{\sqrt{N}}\right)\right)$	$\mathcal{O}\left(\min\left(\frac{\sqrt{n}}{N^{\frac{\beta-1}{2\beta}}}, \frac{n}{\sqrt{N}}\right)\right)$
Novitskii (this work, 2020)	$\mathcal{O}\left(\frac{n^{2-\frac{1}{\beta}}}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$	$\mathcal{O}\left(\frac{n^{1-\frac{1}{2\beta}}}{N^{\frac{\beta-1}{2\beta}}}\right)$
Akhavan, Pontil, Tsybakov (2020)	$\mathcal{O}\left(\frac{n^2}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$	$\mathcal{O}\left(\frac{n}{N^{\frac{\beta-1}{2\beta}}}\right)$
Gasnikov and al. (2017), $\beta = 2$	$\tilde{\mathcal{O}}\left(\frac{n}{\sqrt{\gamma N}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$

The bounds marked in blue are not given in the references but they can be got.