

# Improved Exploiting Higher Order Smoothness in Derivative-free Optimization and Continuous Bandit

Vasilii Novitskii · Alexander Gasnikov

Received: date / Accepted: date

**Abstract** We consider  $\beta$ -smooth (satisfies the generalized Hölder condition with parameter  $\beta > 2$ ) stochastic convex optimization problem with zero-order one-point oracle. The best known result was [1]:

$$\mathbb{E}[f(\bar{x}_N) - f(x^*)] = \tilde{O}\left(\frac{n^2}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$$

in  $\gamma$ -strongly convex case, where  $n$  is the dimension. In this paper we improve this bound:

$$\mathbb{E}[f(\bar{x}_N) - f(x^*)] = \tilde{O}\left(\frac{n^{2-\frac{1}{\beta}}}{\gamma N^{\frac{\beta-1}{\beta}}}\right).$$

**Keywords** zeroth-order optimization · convex problem · stochastic optimization · one-point bandit · smoothing kernel

---

This work is based on results achieved by 63 Conference MIPT held in November 2020. The research of A. Gasnikov was partially supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, project no. 0714-2020-0005. The work of V. Novitskii was supported by Andrei M. Raigorodskii Scholarship in Optimization.

---

V. Novitskii  
Moscow Institute of Physics and Technology, Russia  
E-mail: vasilii.novitskiy@phystech.edu

A. Gasnikov  
Moscow Institute of Physics and Technology, Russia  
Institute for Information Transmission Problems RAS, Russia  
Weierstrass Institute for Applied Analysis and Stochastics, Germany

## 1 Introduction

We study the problem of zero-order stochastic optimization in which the aim is to minimize an unknown convex or strongly convex function where no gradient realization is given but a function value is available at each iteration with some additive noise  $\xi$ . We also study a closely related problem of continuous stochastic bandits. These problems have received significant attention in the literature (see [1–4, 6–9, 14, 15]) and are fundamental for many application where the derivative of function is not available or it is hard to calculate derivatives.

The goal of this paper is to exploit higher order smoothness of the function to improve the performance of projected gradient-like algorithms. Our approach is outlined in Algorithm 1, in which a sequential algorithm gets at each iteration two function values under some noise. At each iteration the algorithm gets function values at points  $x_k + \delta_k$  and  $x_k - \delta_k$ , where  $\delta_k = \tau_k r_k e_k$ . Here  $r_k$  is uniformly distributed random variable,  $e_k$  is uniformly distributed on the Euclidean sphere,  $\tau_k$  – is tunable parameter of the algorithm, the smaller  $\tau_k$  is, the smaller approximation error of the gradient  $\|\tilde{g}_k - \nabla f(x_k)\|$  is (in this article we use only the Euclidean norm) but the bigger variance of  $\|\tilde{g}_k\|$  is, so the trade-off between these terms is needed. Our approach uses kernel smoothing technique proposed by Polyak and Tsybakov in [11], this helps to exploit higher order smoothness.

---

### Algorithm 1 Zero-order Stochastic Projected Gradient

---

**Requires:** Kernel  $K : [-1, 1] \rightarrow \mathbb{R}$ , step size  $\alpha_k > 0$ , parameters  $\tau_k$ .

**Initialization:** Generate scalars  $r_1, \dots, r_N$  uniformly on  $[-1, 1]$  and vectors  $e_1, \dots, e_N$  uniformly on the Euclidean unit sphere  $S_n = \{e \in \mathbb{R}^n : \|e\| = 1\}$ .

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

1.  $y_k := f(x_k + \tau_k r_k e_k) + \xi_k$ ,  $y'_k := f(x_k - \tau_k r_k e_k) + \xi'_k$

2. Define  $\tilde{g}_k := \frac{n}{2\tau_k} (y_k - y'_k) e_k K(r_k)$

3. Update  $x_{k+1} := \Pi_Q(x_k - \alpha_k \tilde{g}_k)$

**end for**

**Output:**  $\{x_k\}_{k=1}^N$ .

---

In algorithms like Algorithm 1 the two possibilities are usually considered. The first one is to obtain a function value in one point with some noise ("one-point" multi-armed bandit), the second is to observe function values in two points with the same noise at each iteration ("two-point" multi-armed bandit). The use of three and more points do not make dramatic difference to the results for two points [5]. Note that despite our algorithm gets two function values for iteration, they are obtained with different noise  $\xi_k$  and  $\xi'_k$ , so it is correct to regard Algorithm 1 one-point and to compare it with one-point algorithms.

In this paper we study higher order smooth functions  $f$  functions satisfying the generalized Hölder condition with parameter  $\beta > 2$  (see inequality (1) below).

We address the question: what is the performance of Algorithm 1, namely the explicit dependency of the convergence rate on the main parameters  $n$  (dimension),  $N$ ,  $\gamma$  (strong convexity parameter for strongly convex functions),  $\beta$ . To handle this task we prove upper bound for Algorithm 1.

**Contributions.** Our main contributions can be summarized as follows:

1. For strongly-convex case: under an adversarial noise assumption (see Assumption 1) we establish for all  $\beta > 2$  the upper bound of order  $\mathcal{O}\left(\frac{n^{2-\frac{1}{\beta}}}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$  for the optimization error of Algorithm 1 for strongly convex case.
2. For convex case: under an adversarial noise assumption (see Assumption 1) we establish for all  $\beta > 2$  that after  $N(\varepsilon) = \mathcal{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$  iterations of Algorithm 1 for the regularized function  $f_\gamma(x) := f(x) + \frac{\varepsilon}{2R^2}\|x - x_0\|^2$  we achieve the optimization error less than or equal to  $\varepsilon$ .

For clarity we compare our results with state-of-the-art ones in Table 1 (dependence of optimization error  $\varepsilon$  on the number of iteration  $N$ , dimension  $n$  and  $\beta$ ,  $\gamma$ ) and Table 2 (dependence of the number of iteration  $N$  on the optimization error  $\varepsilon$ , dimension  $n$  and  $\beta$ ,  $\gamma$ ). To summarize the results we use  $\tilde{\mathcal{O}}()$ , where  $\tilde{\mathcal{O}}()$  coincides with  $\mathcal{O}()$  up to the logarithmic factor.

**Table 1** The dependence of optimization error ( $\varepsilon$ ) on  $N$  (number of iterations),  $n$  (dimension),  $\gamma$ ,  $\beta$

	strongly convex	convex
lower bound [1]	$\mathcal{O}\left(\min\left(\frac{n}{\gamma N^{\frac{\beta-1}{\beta}}}, \frac{n}{\sqrt{N}}\right)\right)$	$\mathcal{O}\left(\min\left(\frac{\sqrt{n}}{N^{\frac{\beta-1}{2\beta}}}, \frac{n}{\sqrt{N}}\right)\right)$
this work (2020)	$\tilde{\mathcal{O}}\left(\frac{n^{2-\frac{1}{\beta}}}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{1-\frac{1}{2\beta}}}{N^{\frac{\beta-1}{2\beta}}}\right)$
Akhavan, Pontil, Tsybakov (2020) [1]	$\tilde{\mathcal{O}}\left(\frac{n^2}{\gamma N^{\frac{\beta-1}{\beta}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{n}{N^{\frac{\beta-1}{2\beta}}}\right)$
Bach, Perchet (2016) [2]	$\mathcal{O}\left(\frac{n^{2-\frac{2}{\beta+1}}}{(\gamma N)^{\frac{\beta-1}{\beta+1}}}\right)$	$\mathcal{O}\left(\frac{n^{1-\frac{1}{\beta+1}}}{N^{\frac{\beta-1}{2(\beta+1)}}}\right)$
Gasnikov and al. (2015), $\beta = 2$ , [8]	$\tilde{\mathcal{O}}\left(\frac{n}{\sqrt{\gamma N}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$
Akhavan, Pontil, Tsybakov (2020), special case $\beta = 2$ [1]	$\tilde{\mathcal{O}}\left(\frac{n}{\sqrt{\gamma N}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$
Zhang and al. (2020) [15]	$\mathcal{O}\left(\frac{n}{\sqrt{\gamma N}}\right)$	$\mathcal{O}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$

**Table 2** The dependence of  $N$  (number of iterations) on  $\varepsilon$ ,  $n$  (dimension),  $\gamma$ ,  $\beta$ 

	strongly convex	convex
lower bound [1]	$\mathcal{O}\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)$	$\mathcal{O}\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)$
this work (2020)	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{1}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Akhavan, Pontil, Tsybakov (2020) [1]	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{2}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Bach, Perchet (2016) [2]	$\mathcal{O}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\gamma\varepsilon^{\frac{\beta+1}{\beta-1}}}\right)$	$\mathcal{O}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Gasnikov and al. (2015), $\beta = 2$ [8]	$\tilde{\mathcal{O}}\left(\frac{n^2}{\gamma\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^2}{\varepsilon^3}\right)$
Akhavan, Pontil, Tsybakov (2020), special case $\beta = 2$ [1]	$\tilde{\mathcal{O}}\left(\frac{n^2}{\gamma\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^2}{\varepsilon^3}\right)$
Zhang and al. (2020) [15]	$\mathcal{O}\left(\frac{n^2}{\gamma\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{n^2}{\varepsilon^3}\right)$

**Comments on Table 1 and Table 2.**

- Note that in Table 1 and Table 2 the right column equals to the central one by  $\gamma \sim \varepsilon$ .
- Note that the results of this work have better dependency  $\varepsilon(N)$  or  $N(\varepsilon)$  than Gasnikov's one-point method only if  $\beta > 2$  else another technique in Theorem 1 is better (see [8] or Theorem 5.1 in [1]). The result in this work is achieved using both kernel smoothing technique and measure concentration inequalities.
- The lower bound for strongly convex case is got under conditions  $\gamma \geq N^{-1/2+1/\beta}$  (otherwise it is better to use convex methods) and (see [1])  $2\gamma \leq \max_{x \in Q} \|\nabla f(x)\|$ .
- The bounds marked in blue are not given in this article and in references but they can be got.
- Too optimistic bounds  $\mathcal{O}\left(\frac{n^{2-\frac{4}{\beta+1}}}{(\gamma N)^{\frac{\beta-1}{\beta+1}}}\right)$  and  $\mathcal{O}\left(\frac{n^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$  were claimed in [2] instead of  $\mathcal{O}\left(\frac{n^{2-\frac{2}{\beta+1}}}{(\gamma N)^{\frac{\beta-1}{\beta+1}}}\right)$  and  $\mathcal{O}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$ , but Akhavan, Pontil and Tsybakov [1] found error in Lemma 2 in [2] where factor  $d$  of dimension ( $n$  in our notation) is missing.

**2 Preliminaries**

In this section we give the necessary notation, definitions and assumptions.

## 2.1 Notation

Let  $\langle \cdot, \cdot \rangle$  and  $\|\cdot\|$  be the standard inner product and Euclidean norm on  $\mathbb{R}^n$  respectively. For every closed convex set  $Q \subset \mathbb{R}^n$  and for every  $x \in \mathbb{R}^n$  let  $\Pi_Q(x)$  denote the Euclidean projection of  $x$  on  $Q$ .

## 2.2 Problem

We address the conditional minimization problem

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

where  $f : U_{\varepsilon_0}(Q) \rightarrow \mathbb{R}$  – function (convex or strongly convex),  $Q \subset \mathbb{R}^n$  – convex compact set (Euclidean metrics).

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

$$\frac{1}{N} \sum_{k=1}^N \mathbb{E}[f(x_k) - f(x^*)].$$

If the average regret is less than or equal to  $\varepsilon$  then the optimization error of averaged estimator  $\bar{x}_N = \frac{1}{N} \sum_{k=1}^N x_k$  is also less than or equal to  $\varepsilon$ :

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

## 2.3 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 noise  $\xi_k$  and  $\xi'_k$  respectively (see Algorithm 1). Recall that the Algorithm 1 is randomized: the scalars  $r_1, \dots, r_N$  are distributed uniformly on  $[-1, 1]$  and the vectors  $e_1, \dots, e_N$  are distributed uniformly on the Euclidean unit sphere  $S_n = \{e \in \mathbb{R}^n : \|e\| = 1\}$ .

**Assumption 1** For all  $k = 1, 2, \dots, N$  it holds that

1.  $\mathbb{E}[\xi_k^2] \leq \sigma^2$  and  $\mathbb{E}[\xi'_k{}^2] \leq \sigma^2$  where  $\sigma \geq 0$ ;
2. the random variables  $\xi_k$  and  $\xi'_k$  are independent from  $e_k$  and  $r_k$ , the random variables  $e_k$  and  $r_k$  are independent.

We do not assume here neither zero-mean of  $\xi_k$  and  $\xi'_k$  nor i.i.d of  $\{\xi_k\}_{k=1}^N$  and  $\{\xi'_k\}_{k=1}^N$  as condition 2 from assumption 1 allows to avoid that.

## 2.4 Higher order smoothness

Let  $l$  denote maximal integer number strictly less than  $\beta$ . Let  $\mathcal{F}_\beta(L)$  denote the set of all functions  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  which are differentiable  $l$  times and for all  $x, z \in U_{\varepsilon_0}(Q)$  satisfy Hölder condition:

$$\left| f(z) - \sum_{0 \leq |m| \leq l} \frac{1}{m!} D^m f(x) (z-x)^m \right| \leq L \|z-x\|^\beta, \quad (1)$$

where  $L > 0$ , the sum is over multi-index  $m = (m_1, \dots, m_n) \in \mathbb{N}^n$ , we use the notation  $m! = m_1! \cdots m_n!$ ,  $|m| = m_1 + \dots + m_n$  and we defined

$$D^m f(x) z^m = \frac{\partial^{|m|} f(x)}{\partial^{m_1} x_1 \dots \partial^{m_n} x_n} z_1^{m_1} \dots z_n^{m_n}, \quad \forall z = (z_1, \dots, z_n) \in \mathbb{R}^n.$$

Let  $\mathcal{F}_{\gamma, \beta}(L)$  denote the set of  $\gamma$ -strongly convex functions  $f \in \mathcal{F}_\beta(L)$ . Recall that  $f$  is called  $\gamma$ -strongly convex for some  $\gamma > 0$  if for all  $x, z \in \mathbb{R}^n$  it holds that  $f(z) \geq f(x) + \langle \nabla f(x), z-x \rangle + \frac{\gamma}{2} \|x-z\|^2$ .

## 2.5 Kernel

For gradient estimator  $\tilde{g}_k$  we use the kernel

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

satisfying

$$\mathbb{E}[K(r)] = 0, \quad \mathbb{E}[rK(r)] = 1, \quad \mathbb{E}[r^j K(r)] = 0, \quad j = 2, \dots, l, \quad \mathbb{E}[|r|^\beta |K(r)|] \leq \infty, \quad (2)$$

where  $r$  is a uniformly distributed on  $[-1, 1]$  random variable. This helps us to get better bounds on the gradient bias  $\|\tilde{g}_k - \nabla f(x_k)\|$  (see Theorem 1 for details).

A weighted sum of Legendre polynomials is an example of such kernels:

$$K_\beta(r) := \sum_{m=0}^{l(\beta)} p'_m(0) p_m(r), \quad (3)$$

where  $l(\beta)$  is maximal integer number strictly less than  $\beta$  and  $p_m(r) = \sqrt{2m+1} L_m(r)$ ,  $L_m(u)$  is Legendre polynomial. We have

$$\mathbb{E}[p_m p_{m'}] = \delta(m - m').$$

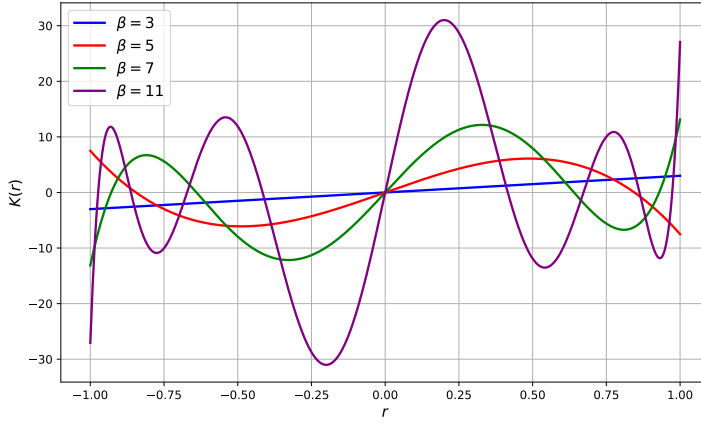
As  $\{p_m(r)\}_{m=0}^j$  is a basis for polynomials of degree less than or equal to  $j$  we can represent  $w^j := \sum_{m=0}^j b_m p_m(r)$  for some integers  $\{b_m\}_{m=0}^j$  (they depend on  $j$ ).

Let's calculate the expectation

$$\mathbb{E} [r^j K_\beta(r)] = \sum_{m=0}^j b_m p'_m(0) = (r^j)'|_{r=0} = \delta(j-1),$$

here  $\delta(0) = 1$  and  $\delta(x) = 0$  if  $x \neq 0$ . We proved that the presented  $K_\beta(r)$  satisfies (2). We have the following kernels for different betas (see Figure 1):

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



**Fig. 1** Examples of kernels from (3)

For Theorem 1 and Theorem 2 we need to introduce the constants

$$\kappa_\beta = \int |u|^\beta |K(u)| du \quad (4)$$

and

$$\kappa = \int K^2(u) du. \quad (5)$$

It is proved in [2] that  $\kappa_\beta$  and  $\kappa$  do not depend on  $n$ , they depend only on  $\beta$ :

$$\kappa_\beta \leq 2\sqrt{2}(\beta - 1), \quad (6)$$

$$\kappa \leq \sqrt{3}\beta^{3/2}. \quad (7)$$

### 3 Theorems

In this section we prove upper bounds on the optimization error of Algorithm 1 for strongly convex function (Theorem 1) and for convex function (Theorem 2).

**Theorem 1** *Let  $f \in \mathcal{F}_{\gamma, \beta}(L)$  with  $\gamma, L > 0$  and  $\beta > 2$ . Let Assumption 1 hold and let  $Q$  be a convex compact subset of  $\mathbb{R}^n$ . Let  $f$  be  $G$ -Lipschitz on the Euclidean  $\tau_1$ -neighborhood of  $Q$ .*

*Then the optimization error of averaged estimator  $\bar{x}_N = \frac{1}{N} \sum_{k=1}^N x_k$  where the points  $x_k$  are given by Algorithm 1 with parameters*

$$\tau_k = \left( \frac{3\kappa\sigma^2 n}{2(\beta-1)(\kappa_\beta L)^2} \right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}, \quad \alpha_k = \frac{2}{\gamma k}, \quad k = 1, \dots, N$$

*satisfies*

$$\mathbb{E} [f(\bar{x}_N) - f(x^*)] \leq \frac{1}{\gamma} \left( n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N} \right),$$

*where  $A_1 = 3\beta(\kappa\sigma^2)^{\frac{\beta-1}{\beta}} (\kappa_\beta L)^{\frac{2}{\beta}}$ ,  $A_2 = c^* \kappa G^2$ ,  $\kappa_\beta$  and  $\kappa$  are constants depending only on  $\beta$ , see (4) and (5).*

**Proof Step 1.** Fix an arbitrary  $x \in Q$ . As  $x_{k+1}$  is the Euclidean projection we have  $\|x_{k+1} - x\|^2 \leq \|x_k - \alpha_k \tilde{g}_k - x\|^2$  which is equivalent to

$$\langle \tilde{g}_k, x_k - x \rangle \leq \frac{\|x_k - x\|^2 - \|x_{k+1} - x\|^2}{2\alpha_k} + \frac{\alpha_k}{2} \|\tilde{g}_k\|^2. \quad (8)$$

By the strong convexity assumption we have

$$f(x_k) - f(x) \leq \langle \nabla f(x_k), x_k - x \rangle - \frac{\gamma}{2} \|x_k - x\|^2. \quad (9)$$

Combining the last two inequations we obtain

$$\begin{aligned} f(x_k) - f(x) &\leq \langle \nabla f(x_k) - \tilde{g}_k, x_k - x \rangle + \frac{\|x_k - x\|^2 - \|x_{k+1} - x\|^2}{2\alpha_k} \\ &\quad + \frac{\alpha_k}{2} \|\tilde{g}_k\|^2 - \frac{\gamma}{2} \|x_k - x\|^2. \end{aligned} \quad (10)$$

Taking conditional expectation given  $x_k$  with respect to  $r_k, \xi_k$  and  $\xi'_k$  we obtain

$$\begin{aligned} f(x_k) - f(x) &\leq \langle \nabla f(x_k) - \mathbb{E}[\tilde{g}_k | x_k], x_k - x \rangle + \frac{\alpha_k}{2} \mathbb{E}[\|\tilde{g}_k\|^2 | x_k] \\ &\quad + \frac{\|x_k - x\|^2 - \mathbb{E}[\|x_{k+1} - x\|^2 | x_k]}{2\alpha_k} - \frac{\gamma}{2} \|x_k - x\|^2. \end{aligned} \quad (11)$$



**Step 2 (Bounding bias term).** Our aim is to bound the first term in (11), namely  $\langle \nabla f(x_k) - \mathbb{E}[\tilde{g}_k | x_k], x_k - x \rangle$ . Using the Taylor expansion we have

$$\begin{aligned} f(x_k + \tau_k r_k e_k) &= f(x_k) + \langle \nabla f(x_k), \tau_k r_k e_k \rangle \\ &+ \sum_{2 \leq |m| \leq l} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)} f(x_k) e_k^m + R(\tau_k r_k e_k), \end{aligned} \quad (12)$$

where by assumption  $|R(\tau_k r_k e_k)| \leq L \|\tau_k r_k e_k\|^\beta = L(\tau_k \cdot |r_k|)^\beta$ . Thus,

$$\begin{aligned} \tilde{g}_k &= \left( \langle \nabla f(x_k), \tau_k r_k e_k \rangle + \sum_{2 \leq |m| \leq l, |m| \text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)} f(x_k) e_k^m \right. \\ &\left. + \frac{1}{2} R(\tau_k r_k e_k) - \frac{1}{2} R(-\tau_k r_k e_k) + \xi_k - \xi'_k \right) \frac{n}{\tau_k} K(r_k) e_k. \end{aligned} \quad (13)$$

Using the properties of the smoothing kernel  $K$ , independence of  $e_k$  and  $r_k$  (Assumption 1) and the fact that  $\mathbb{E}[e_k e_k^T] = \frac{1}{n} \mathbb{I}_{n \times n}$  we obtain

$$\mathbb{E}_{e_k, r_k} \left[ \langle \nabla f(x_k), \tau_k r_k e_k \rangle \frac{n}{\tau_k} K(r_k) e_k \mid x_k \right] = \nabla f(x_k). \quad (14)$$

Using the fact that  $\mathbb{E}[r_k^{|m|} K(r_k)] = 0$  if  $2 \leq |m| \leq l$  or  $|m| = 0$  and Assumption 1 we have

$$\left( \sum_{2 \leq |m| \leq l, |m| \text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)} f(x_k) e_k^m + \xi_k - \xi'_k \right) \frac{n}{\tau_k} K(r_k) e_k = 0. \quad (15)$$

Combining (13), (14) and (15) and using the definition of  $\kappa_\beta$  we obtain

$$\begin{aligned} |\langle \nabla f(x_k) - \mathbb{E}[\tilde{g}_k | x_k], x_k - x \rangle| &= \\ &= \left| \mathbb{E} \left[ \left( \frac{1}{2} R(\tau_k r_k e_k) - \frac{1}{2} R(-\tau_k r_k e_k) \right) \frac{n}{\tau_k} K(r_k) \langle e_k, x_k - x \rangle \mid x_k \right] \right| \\ &\leq L \tau_k^{\beta-1} \cdot \mathbb{E}_{r_k} [ |r_k|^\beta K(r_k) ] \cdot n |\mathbb{E}_{e_k} [\langle e_k, x_k - x \rangle \mid x_k]| \\ &\leq \kappa_\beta L \sqrt{n} \tau_k^{\beta-1} \|x_k - x\|, \end{aligned} \quad (16)$$

where in the last inequality the fact that  $|\mathbb{E}_e [\langle e, s \rangle]|^2 \leq \mathbb{E}_e [\langle e, s \rangle^2] = \frac{\|s\|^2}{n}$  was used (the fact from concentration measure theory). Applying the inequality  $ab \leq 1/2(a^2 + b^2)$  to the last expression in (16) we finally get

$$|\langle \nabla f(x_k) - \mathbb{E}[\tilde{g}_k | x_k], x_k - x \rangle| \leq \frac{(\kappa_\beta L)^2}{\gamma} n \tau_k^{2(\beta-1)} + \frac{\gamma}{4} \|x_k - x\|^2. \quad (17)$$

**Step 3 (Bounding second moment of gradient estimator).** Our aim is to estimate  $\mathbb{E}[\|\tilde{g}_k\|^2 | x_k]$  which is the second term in (11). The expectation

here is with respect to  $r_k$ ,  $\xi_k$  and  $\xi'_k$ . To lighten the presentation and without loss of generality we drop the lower script  $k$  in all quantities.

We have

$$\begin{aligned}\|\tilde{g}\|^2 &= \frac{n^2}{4\tau^2} \|(f(x + \tau re) - f(x - \tau re) + \xi - \xi')K(r)e\|^2 \\ &= \frac{n^2}{4\tau^2} ((f(x + \tau re) - f(x - \tau re) + \xi - \xi'))^2 K^2(r).\end{aligned}\quad (18)$$

Using the inequality  $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$  and Assumption 1 we get

$$\mathbb{E} [\|\tilde{g}\|^2 | x] \leq \frac{3n^2}{4\tau^2} (\mathbb{E} [(f(x + \tau re) - f(x - \tau re))^2 K^2(r) | x] + 2\kappa\sigma^2). \quad (19)$$

Lemma 9 in [13] states that for any function  $f$  which is  $G$ -Lipschitz with respect to 2-norm, it holds that if  $e$  is uniformly distributed on the Euclidean unit sphere, then

$$\sqrt{\mathbb{E} [(f(e) - \mathbb{E}[f(e)])^4]} \leq \frac{cG^2}{n}, \quad (20)$$

where  $c < 3$  is a positive numerical constant.

Using (20), symmetry of Euclidean unit sphere and the inequality  $(a + b)^2 \leq 2(a^2 + b^2)$  we obtain

$$\begin{aligned}\mathbb{E} [(f(x + e) - f(x - e))^2 | x] &= \mathbb{E}_e [(f(x + e) - f(x - e))^2] \\ &\leq \mathbb{E}_e [((f(x + e) - \mathbb{E}_e[f(x + e)]) - (f(x - e) - \mathbb{E}_e[f(x - e)]))^2] \\ &\leq 2\mathbb{E}_e [(f(x + e) - \mathbb{E}_e[f(x + e)])^2] + 2\mathbb{E}_e [(f(x - e) - \mathbb{E}_e[f(x - e)])^2] \\ &\leq 2\sqrt{\mathbb{E}_e [(f(x + e) - \mathbb{E}_e[f(x + e)])^4]} + 2\sqrt{\mathbb{E}_e [(f(x - e) - \mathbb{E}_e[f(x - e)])^4]} \\ &\leq \frac{4cG^2}{n},\end{aligned}\quad (21)$$

so we have

$$\mathbb{E} [(f(x + \tau re) - f(x - \tau re))^2 | x] \leq \frac{4c(\tau r)^2 G^2}{n} \leq \frac{4c\tau^2 G^2}{n}. \quad (22)$$

By substituting (22) into (19), using independence of  $e$  and  $r$  and returning the lower script  $k$  we finally get

$$\mathbb{E} [\|\tilde{g}_k\|^2 | x] \leq \kappa \left( c^* n G^2 + \frac{3(n\sigma)^2}{2\tau_k^2} \right), \quad (23)$$

where  $c^* = 3c$ .

**Step 4.** Let  $\rho_k^2$  denote  $\mathbb{E}[\|x_k - x\|^2]$ . Substituting (17) and (23) into (11), taking full expectation and summing over  $k$  we obtain

$$\begin{aligned} \sum_{k=1}^N \mathbb{E}[f(x_k) - f(x)] &\leq \sum_{k=1}^N \left( \frac{(\kappa_\beta L)^2}{\gamma} n \tau_k^{2(\beta-1)} + \frac{\alpha_k}{2} \kappa \left( c^* n G^2 + \frac{3(n\sigma)^2}{2\tau_k^2} \right) \right) \\ &\quad + \sum_{k=1}^N \left( \frac{\rho_k^2 - \rho_{k+1}^2}{2\alpha_k} - \left( \frac{\gamma}{2} - \frac{\gamma}{4} \right) \rho_k^2 \right). \end{aligned} \quad (24)$$

Let  $\rho_{N+1}^2 = 0$ . Then setting  $\alpha_k = \frac{2}{\gamma k}$  yields

$$\begin{aligned} \sum_{k=1}^N \left( \frac{\rho_k^2 - \rho_{k+1}^2}{2\alpha_k} - \frac{\gamma}{4} \rho_k^2 \right) &\leq \rho_1^2 \left( \frac{1}{2\alpha_1} - \frac{\gamma}{4} \right) + \sum_{k=2}^{N+1} \rho_k^2 \left( \frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}} - \frac{\gamma}{4} \right) \\ &= \rho_1^2 \left( \frac{\gamma}{4} - \frac{\gamma}{4} \right) + \sum_{k=2}^{N+1} \rho_k^2 \left( \frac{\gamma}{4} - \frac{\gamma}{4} \right) = 0. \end{aligned} \quad (25)$$

Substituting (25) into (24) with  $\alpha_k = \frac{2}{\gamma k}$  we obtain

$$\begin{aligned} \sum_{k=1}^N \mathbb{E}[f(x_k) - f(x)] &\leq \frac{1}{\gamma} \sum_{k=1}^N \left( (\kappa_\beta L)^2 n \tau_k^{2(\beta-1)} + \kappa \left( c^* n G^2 + \frac{3(n\sigma)^2}{2\tau_k^2} \right) \frac{1}{k} \right) \\ &= \frac{1}{\gamma} \sum_{k=1}^N \left( \left[ n \cdot (\kappa_\beta L)^2 \tau_k^{2(\beta-1)} + n^2 \cdot \frac{3\kappa\sigma^2}{2k\tau_k^2} \right] + \frac{c^* \kappa n G^2}{k} \right). \end{aligned} \quad (26)$$

If  $\sigma > 0$  then  $\tau_k = \left( \frac{3\kappa\sigma^2 n}{2(\beta-1)(\kappa_\beta L)^2} \right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}$  is the minimizer of square brackets. Plugging this  $\tau_k$  in (26) and using two inequalities: for the expression in square brackets  $\sum_{k=1}^N k^{-1+1/\beta} \leq \beta N^{1/\beta}$  (if  $\beta > 2$ ) and for the term after square brackets  $\sum_{k=1}^N \frac{1}{k} \leq 1 + \ln N$  we get

$$\sum_{k=1}^N \mathbb{E}[f(x_k) - f(x)] \leq \frac{1}{\gamma} \left( n^{2-\frac{1}{\beta}} A_1 N^{\frac{1}{\beta}} + A_2 n(1 + \ln N) \right) \quad (27)$$

with  $A_1$  and  $A_2$  from the formulation of Theorem 1. Due to the convexity of  $f$  we finally prove the theorem

$$\mathbb{E}[f(\bar{x}_N) - f(x^*)] \leq \frac{1}{\gamma} \left( n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1 + \ln N)}{N} \right). \quad (28)$$

□

We emphasize that the usage of kernel smoothing technique, measure concentration inequalities and the assumption that  $\xi_k$  is independent from  $e_k$  or  $r_k$  (Assumption 1) lead to the results better than the state-of-the-art ones for  $\beta > 2$  (see Table 1 and Table 2). The last assumption also allows us not to assume neither zero-mean of  $\xi_k$  and  $\xi'_k$  nor i.i.d of  $\{\xi_k\}_{k=1}^N$  and  $\{\xi'_k\}_{k=1}^N$ .

**Theorem 2** *Let  $f \in \mathcal{F}_\beta(L)$  with  $\gamma, L > 0$  and  $\beta > 2$ . Let Assumption 1 hold and let  $Q$  be a convex compact subset of  $\mathbb{R}^n$ . Let  $f$  be  $G$ -Lipschitz on the Euclidean  $\tau_1$ -neighborhood of  $Q$ . Let  $\bar{x}_N$  denote  $\frac{1}{N} \sum_{k=1}^N x_k$ .*

*Then we achieve the optimization error  $\mathbb{E}[f(\bar{x}_N) - f(x^*)] \leq \varepsilon$  after  $N(\varepsilon)$  steps of Algorithm 1 with settings from Theorem 1 for the regularized function:  $f_\gamma(x) := f(x) + \frac{\gamma}{2} \|x - x_0\|^2$ , where  $\gamma \leq \frac{\varepsilon}{R^2}$ ,  $R = \|x_0 - x^*\|$ ,  $x_0 \in Q$  - arbitrary point.*

$$N(\varepsilon) = \max \left\{ \left( R\sqrt{2A_1} \right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left( R\sqrt{2c'A_2} \right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}} \right\},$$

where  $A_1 = 3\beta(\kappa\sigma^2)^{\frac{\beta-1}{\beta}} (\kappa_\beta L)^{\frac{2}{\beta}}$ ,  $A_2 = c^* \kappa G^2$  - constants from Theorem 1,  $\rho > 0$  - arbitrarily small positive number.

**Proof Step 1.** Let  $x^*$  and  $x_\gamma^*$  denote  $\arg \min_{x \in Q} f(x)$  and  $\arg \min_{x \in Q} f_\gamma(x)$  respectively. Setting  $\gamma = \frac{\varepsilon}{R^2}$  and using the inequality  $f_\gamma(x_\gamma^*) \leq f_\gamma(x^*)$  we obtain

$$\begin{aligned} f(\bar{x}_N) - f(x^*) &= f_\gamma(\bar{x}_N) - f_\gamma(x^*) - \frac{\gamma}{2} \|\bar{x}_N - x_0\|^2 + \frac{\gamma}{2} \|x^* - x_0\|^2 \\ &\leq f_\gamma(\bar{x}_N) - f_\gamma(x^*) + \frac{\gamma}{2} \|x^* - x_0\|^2 \\ &\leq f_\gamma(\bar{x}_N) - f_\gamma(x_\gamma^*) + \frac{\varepsilon}{2}. \end{aligned} \quad (29)$$

**Step 2.** Now we apply Theorem 1 for  $f_\gamma(x)$  and bound RHS by  $\frac{\varepsilon}{2}$ :

$$\mathbb{E}[f_\gamma(\bar{x}_N) - f_\gamma(x^*)] \leq \frac{1}{\gamma} \left( n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N} \right) \leq \frac{\varepsilon}{2}. \quad (30)$$

The inequality (30) is done if ( $\gamma = \frac{\varepsilon}{R^2}$ )

$$\max \left\{ n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}}, A_2 \frac{n(1+\ln N)}{N} \right\} \leq \frac{\gamma\varepsilon}{2} = \frac{\varepsilon^2}{2R^2}. \quad (31)$$

It is true that  $1 + \ln N \leq c' N^{\frac{\rho}{\beta+1}}$  for some  $c' > 0$ . So the inequality (31) holds if

$$N \geq \max \left\{ \left( R\sqrt{2A_1} \right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left( R\sqrt{2c'A_2} \right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}} \right\}. \quad (32)$$

The inequalities (29) and (30) yield  $\mathbb{E}[f(\bar{x}_N) - f(x^*)] \leq \varepsilon$ .

□

## 4 Numerical Experiments

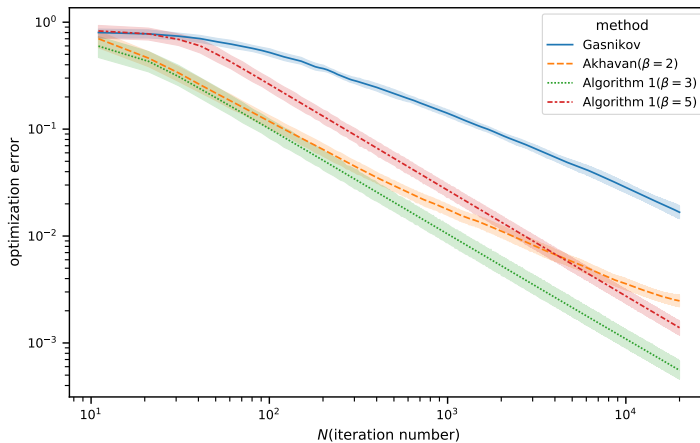
In our experiment [10] we compare the Algorithm 1 (with  $\beta = 3$  and  $\beta = 5$ ) proposed in this paper with Gasnikov's one-point method and with Akhavan's method for the special case  $\beta = 2$ .

We consider the problem of the minimization of the quadratic function

$$f(x) = \frac{1}{4}x_1^2 + x_2^2 + 4x_3^2$$

on the Euclidean ball  $Q = \{x \in \mathbb{R}^3 : \|x\| \leq 1\}$ .

The starting point is  $x_0$  with  $\|x_0\| = 1/2$ . The dependency of  $f(\bar{x}_N) - f(x^*)$  (optimization error) on  $N$  (iteration number) is presented on the Figure 2. The optimization error has its mean and 0.95-confidence interval. As the Lipschitz constants for the quadratic oracle are equal to zero, for the Algorithm 1 we choose  $L = 0.01$ .



**Fig. 2** Dependency of optimization error of Algorithm 1 on iteration

We see on the Figure 2 that the usage of higher-order smoothness by Algorithm 1 helps to overcome the methods which do not use this.

## 5 Discussion and related work

The results of this paper can be generalized for the saddle-point problems. Recently GANs and Reinforcement Learning caused a big interest for saddle-point problems, see [12].

Another possible generalization of this paper is obtaining the large probability bounds for optimization error. We cannot obtain upper bounds in terms of large deviation probability (not in terms of expectation) under the Assumption 1. The exploiting of higher order smoothness with the help of kernels

under rather general noise assumptions (non-zero mean) causes big variation  $\|\tilde{g}_k - \nabla f(x_k)\|$  and this can cause the problems with large deviation probability rates.

It remains an open question whether large deviation probability can be obtained under non-zero mean noise. And also it remains an open question whether better dependence of optimization error on the dimension  $n$  and strong convexity parameter  $\gamma$  can be obtained.

**Acknowledgements.** We would like to thank Alexandre B. Tsybakov for helpful remarks about Tables 1 and 2.

## References

1. Akhavan, A., Pontil, M., Tsybakov, A.B.: Exploiting higher order smoothness in derivative-free optimization and continuous bandits. arXiv preprint arXiv:2006.07862 (2020)
2. Bach, F., Perchet, V.: Highly-smooth zero-th order online optimization. In: Conference on Learning Theory, pp. 257–283 (2016)
3. Bubeck, S., Lee, Y.T., Eldan, R.: Kernel-based methods for bandit convex optimization. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pp. 72–85 (2017)
4. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. Society for Industrial and Applied Mathematics (2009). DOI 10.1137/1.9780898718768
5. Duchi, J.C., Jordan, M.I., Wainwright, M.J., Wibisono, A.: Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Transactions on Information Theory **61**(5), 2788–2806 (2015)
6. Gasnikov, A., Dvurechensky, P., Kamzolov, D.: Gradient and gradient-free methods for stochastic convex optimization with inexact oracle. arXiv preprint arXiv:1502.06259 (2015)
7. Gasnikov, A., Dvurechensky, P., Nesterov, Y.: Stochastic gradient methods with inexact oracle. arXiv preprint arXiv:1411.4218 (2014)
8. Gasnikov, A.V., Krymova, E.A., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case. Automation and remote control **78**(2), 224–234 (2017)
9. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numerica **28**, 287–404 (2019). DOI 10.1017/S0962492919000060
10. Novitskii, V.: Zeroth-order algorithms for smooth saddle-point problems (2020). URL <https://cutt.ly/bjxQHRY>
11. Polyak, B.T., Tsybakov, A.B.: Optimal order of accuracy of search algorithms in stochastic optimization. Problemy Peredachi Informatsii **26**(2), 45–53 (1990)
12. Sadiev, A., Beznosikov, A., Dvurechensky, P., Gasnikov, A.: Zeroth-order algorithms for smooth saddle-point problems. arXiv preprint arXiv:2009.09908 (2020)
13. Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. The Journal of Machine Learning Research **18**(1), 1703–1713 (2017)
14. Spall, J.C.: Introduction to Stochastic Search and Optimization, 1 edn. John Wiley & Sons, Inc., New York, NY, USA (2003)
15. Zhang, Y., Zhou, Y., Ji, K., Zavlanos, M.M.: Boosting one-point derivative-free online optimization via residual feedback. arXiv preprint arXiv:2010.07378 (2020)