Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский физико-технический институт (национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Направление подготовки: 03.04.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

Использование гладкости высоких порядков в безградиентных методах с неточным оракулом для задач выпуклой стохастической оптимизации

(магистерская диссертация)

Студент:
Новицкий Василий Геннадьевич

____
(подпись обучающегося)

Научный руководитель:
Доктор физико-математических наук
Гасников Александр Владимирович

____
(подпись научного руководителя)

Москва 2021

Аннотация

В диссертации изучается задача выпуклой стохастической оптимизации с безградиентным оракулом в условиях гладкости высоких порядков. Оптимизируемая функция удовлетворяет обобщенному условию Гёльдера, другими словами исследуются функции, у которых как минимум липшицев Гессиан или производные более высоких порядков.

В этой диссертации представлен алгоритм, который позволяет получить верхние оценки скорости сходимости, превышающие существующие как в сильно выпуклом случае, так и в выпуклом случае. Также результат был обобщен на седловые задачи, где получены новые рекордные оценки скорости сходимости для выпукло–вогнутых функций и для сильно выпукло–сильно вогнутых функций.

Contents

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 ξ. We also study a closely related problem of continuous stochastic bandits. These problems have received significant attention in the literature (see [1–7, 10, 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 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, τ_k – is tunable parameter of the algorithm, the smaller τ_k is, the smaller approximation error of the gradient $\|\widetilde{g}_k - \nabla f(x_k)\|$ is (in this article we use only the Euclidean norm) but the bigger variance of $\|\widetilde{g}_k\|$ is, so the trade-off between these terms is needed. Our approach uses kernel smoothing technique proposed by Polyak and Tsybakov in [12], this helps to exploit higher order smoothness.

Algorithm 1 Zero-order Stochastic Projected Gradient

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

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

for $k = 1, \ldots, N$ **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 $\widetilde{g}_k := \frac{n}{2\tau_k}(y_k - y'_k)e_k K(r_k)$

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

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 does not make dramatic difference to the results for two points [11]. Note that despite our algorithm gets two function values for iteration, they are obtained with different noise ξ_k and ξ_k', so it is correct to regard Algorithm 1 one-point and to compare it with one-point algorithms.

Here we study 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, γ (strong convexity parameter for strongly convex functions), β. To handle this task we prove an 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 ε.

3 We have also generalized the results of minimization for the saddle-point problem $\min\limits_{x\in\mathcal{X}} \max\limits_{y\in\mathcal{Y}} \varphi(x,y)$. Upper bounds for both strongly-convex-strongly-concave case and convex-concave case coincide with the minimization upper bounds.

For clarity we compare our results with state-of-the-art ones in Table 1 (dependence of optimization error ε on the number of iteration N, dimension n and β, γ) and Table 2 (dependence of the number of iteration N on the optimization error ε, dimension n and β, γ). 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 (ε) on N (number of iterations), n (dimension), γ, β

	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), $\beta = 2$, [15]	$\tilde{\mathcal{O}}\left(\frac{n}{\sqrt{\gamma N}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$
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) [3]	$\mathcal{O}\left(\frac{n}{\sqrt{\gamma N}}\right)$	$\mathcal{O}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$

Table 2: The dependence of N (number of iterations) on ε, n (dimension), γ, β

	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), $\beta = 2$ [15]	$\tilde{\mathcal{O}}\left(\frac{n^2}{\gamma\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^2}{\varepsilon^3}\right)$
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) [3]	$\mathcal{O}\left(\frac{n^2}{\gamma\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{n^2}{\varepsilon^3}\right)$

Comments on Table 1 and Table 2.

1 Note that in Table 1 and Table 2 the right column equals to the central one by $\gamma \sim \varepsilon$.

2 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 [15] or Theorem 5.1 in [1]). The result in this work is achieved using both kernel smoothing technique and measure concentration inequalities.

3 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]) $2\gamma \leq \max_{x\in Q} \|\nabla f(x)\|$.

4 The bounds marked in blue are not given in this article and in references but they can be got.

5 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

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

where $f : U_{\varepsilon_0}(Q) \to \mathbb{R}$ – 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:

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

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

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

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

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

1 $\mathbb{E}[\xi_k^2] \le \sigma^2$ and $\mathbb{E}[\xi'^2_k] \le \sigma^2$ where $\sigma \ge 0$;

2 the random variables ξ_k and ξ'_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 ξ_k and ξ'_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 β. Let $\mathcal{F}_\beta(L)$ denote the set of all functions $f : \mathbb{R}^n \to \mathbb{R}$ which are differentiable l times and for all $x, z \in U_{\varepsilon_0}(Q)$ satisfy Hölder condition:

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

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

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

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

2.5 Kernel

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

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

satisfying

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

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

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

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

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

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

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

Let's calculate the expectation

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

here $\delta(0) = 1$ and $\delta(x) = 1$ 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):

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

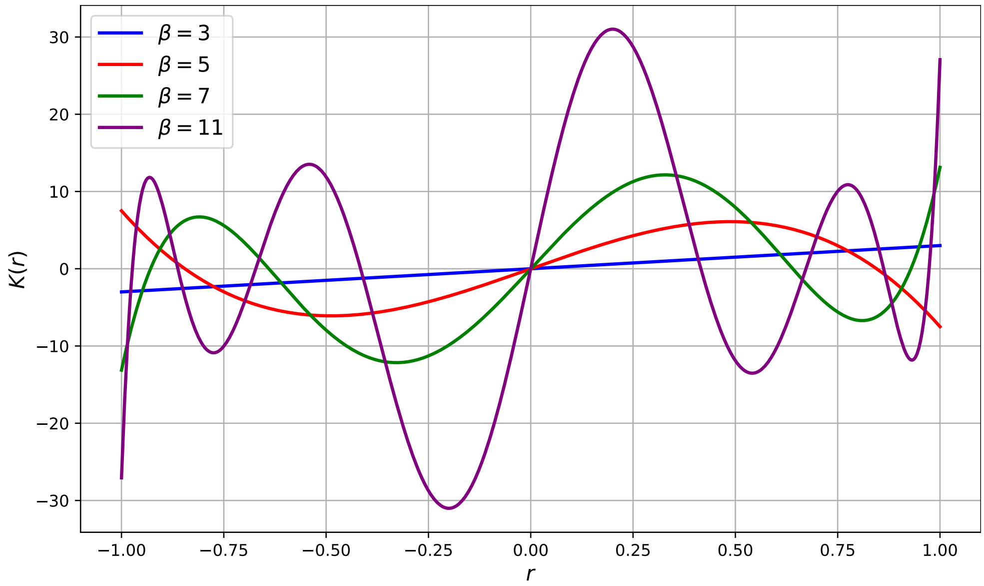

Figure 1: Examples of kernels from (3)

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

$$\kappa_\beta = \int |u|^\beta |K(u)|\, du \tag{4}$$

and

$$\kappa = \int K^2(u)\, du. \tag{5}$$

It is proved in [2] that κ_β and κ do not depend on n, they depend only on β:

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

$$\kappa \leq \sqrt{3}\beta^{3/2}. \tag{7}$$

3 Minimization problem

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

Theorem 1. Let $f \in \mathcal{F}_{\gamma,\beta}(L)$ with γ, $L > 0$ 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 τ_1-neighborhood of Q.

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

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

satisfies

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

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

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

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

By the strong convexity assumption we have

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

Combining the last two inequations we obtain

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

Taking conditional expectation given x_k with respect to r_k, ξ_k and ξ_k' we obtain

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

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

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

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,

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

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

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

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

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

Combining (13), (14) and (15) and using the definition of κ_β we obtain

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

where in the last inequality the fact that $\left|\mathbb{E}_e\left[\langle e, s\rangle\right]\right|^2 \leq \mathbb{E}_e\left[\langle e, s\rangle^2\right] = \frac{\|s\|^2}{n}$ 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

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

Step 3 (Bounding second moment of gradient estimator). Our aim is to estimate $\mathbb{E}\left[\|\widetilde{g}_k\|^2 | x_k\right]$ which is the second term in (11). The expectation here is with respect to r_k, ξ_k and ξ_k'. To lighten the presentation and withous loss of generality we drop the lower script k in all quantities.

We have

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

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

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

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

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

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

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

so we have

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

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

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

where $c^* = 3c$.

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

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

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

$$\sum_{k=1}^{N} \left(\frac{\rho_k^2 - \rho_{k+1}^2}{2\alpha_k} - \frac{\gamma}{4}\rho_k^2 \right) \le \rho_1^2 \left(\frac{1}{2\alpha_1} - \frac{\gamma}{4} \right) + \sum_{k=2}^{N+1} \rho_k^2 \left(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}} - \frac{\gamma}{4} \right) = \rho_1^2 \left(\frac{\gamma}{4} - \frac{\gamma}{4} \right) + \sum_{k=2}^{N+1} \rho_k^2 \left(\frac{\gamma}{4} - \frac{\gamma}{4} \right) = 0. \tag{25}$$

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

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

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

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

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

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

$\square$

We emphasize that the usage of kernel smoothing technique, measure concentration inequalities and the assumption that ξ_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 ξ_k and ξ_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 γ, $L > 0$ 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 τ_1-neighborhood of Q. Let $\overline{x}_N$ denote $\frac{1}{N}\sum\limits_{k=1}^{N} x_k$.

Then we achieve the optimization error $\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leq \varepsilon$ 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_γ^* denote $\arg\min\limits_{x\in Q} f(x)$ and $\arg\min\limits_{x\in Q} f_\gamma(x)$ respectively. Setting $\gamma = \frac{\varepsilon}{R^2}$ and using the inequality $f_\gamma(x_\gamma^*) \leq f_\gamma(x^*)$ we obtain

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

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

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

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

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

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

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

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

□

4 Saddle-point problem

Recently GANs and Reinforcement Learning caused a big interest for saddle-point problems, see [8, 9, 16]. So in this section we generalize the results for minimization problems to saddle-point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\varphi(x,y). \tag{33}$$

As mentioned earlier, we consider the saddle-point problem (33), where $\varphi(\cdot, y)$ is convex function defined on compact convex set $\mathcal{X} \subset \mathbb{R}^{n_x}$, $\varphi(x, \cdot)$ is concave function defined on compact convex set $\mathcal{Y} \subset \mathbb{R}^{n_y}$. For convenience, we denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and then $z \in \mathcal{Z} \subset \mathbb{R}^n$ means $z \overset{\text{def}}{=} (x, y)$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$. When we use $\varphi(z)$, we mean $\varphi(z) = \varphi(x, y)$.

Definition (γ-strong convexity–strong concavity). Function $\varphi(z)$ is γ-strongly-convex-strongly-concave in $\mathcal{Z}$ with $\gamma > 0$ when $\varphi(\cdot, y)$ is γ-strongly-convex for all y and $\varphi(x, \cdot)$ is γ-strongly-concave for all x.

In this paragraph we study higher-order smooth functions φ functions satisfying so called generalized Hölder condition with parameter $\beta > 2$.

Let $\Phi_\beta(L)$ denote the set of all functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ which satisfy Hölder condition (1) Let $\Phi_{\gamma,\beta}(L)$ denote the set of γ-strongly-convex-strongly-concave functions $\varphi \in \Phi_\beta(L)$.

For the saddle-point problem (33) we propose to use Algorithm 2 which is a slightly modified version of Algorithm 1.

Algorithm 2 Zero-order Stochastic Projected Gradient

Requires: Kernel $K : [-1,1] \to \mathbb{R}$, step size $\gamma_k > 0$, parameters τ_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 $u_k := \varphi(z_k + \tau_k r_k e_k) + \xi_k$, $u_k' := \varphi(z_k - \tau_k r_k e_k) + \xi_k'$

2 Define $\widetilde{g}_k := \frac{n}{2\tau_k}(u_k - u_k') \begin{pmatrix} (e_k)_x \\ -(e_k)_y \end{pmatrix} K(r_k)$

3 Update $z_{k+1} := \Pi_Z(z_k - \gamma_k \widetilde{g}_k)$

end for

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

Theorem 3. Let $\varphi \in \Phi_{\gamma,\beta}(L)$ with γ, $L > 0$ and $\beta > 2$. Let Assumption 1 hold and let $\mathcal{Z}$ be a convex compact subset of $\mathbb{R}^n$. Let φ be G-Lipschitz on the Euclidean τ_1-neighborhood of $\mathcal{Z}$ (see τ_k below).

Then the rate of convergence is given by Algorithm 2 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

$$\begin{aligned} \mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] &\le \max_{y \in \mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x \in \mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \\ &\le \frac{1}{\gamma}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N} \right), \end{aligned}$$

where $\overline{z}_N = \frac{1}{N}\sum_{k=1}^{N} z_k$, $A_1 = 3\beta(\kappa\sigma^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L)^{\frac{2}{\beta}}$, $A_2 = 9\kappa G^2$, κ_β and κ are constants depending only on β, see (4) and (5).

Proof. Step 1. Fix an arbitrary $z \in \mathcal{Z}$. As z_{k+1} is the Euclidean projection we have $\|z_{k+1} - z\|^2 \le \|z_k - \gamma_k \widetilde{g}_k - z\|^2$ which is equivalent to

$$\langle \widetilde{g}_k, z_k - z \rangle \le \frac{\|z_k - z\|^2 - \|z_{k+1} - z\|^2}{2\gamma_k} + \frac{\gamma_k}{2}\|\widetilde{g}_k\|^2. \tag{34}$$

Using the strong convexity-concavity and combining x and y parts of the argument z together we have

$$
\begin{aligned}
\varphi(x_k, y) - \varphi(x, y_k) =& \varphi(x_k, y) - \varphi(x_k, y_k) + \varphi(x_k, y_k) - \varphi(x, y_k) \\
\leq& \langle -\nabla_y \varphi(x_k, y_k), y_k - y\rangle - \frac{\gamma}{2}\|y_k - y\|^2 \\
+& \langle -\nabla_x \varphi(x_k, y_k), x_k - x\rangle - \frac{\gamma}{2}\|x_k - x\|^2 \\
=& \langle \widetilde{\nabla}\varphi(z_k), z_k - z\rangle - \frac{\gamma}{2}\|z_k - z\|^2.
\end{aligned} \tag{35}
$$

Combining the last two inequations we obtain

$$
\begin{aligned}
\varphi(x_k, y) - \varphi(x, y_k) \leq& \langle \widetilde{\nabla}\varphi(z_k) - \widetilde{g_k}, z_k - z\rangle + \frac{\|z_k - z\|^2 - \|z_{k+1} - z\|^2}{2\gamma_k} \\
&+ \frac{\gamma_k}{2}\|\widetilde{g_k}\|^2 - \frac{\gamma}{2}\|z_k - z\|^2.
\end{aligned} \tag{36}
$$

Taking conditional expectation given z_k with respect to r_k, ξ_k and ξ_k' we obtain

$$
\begin{aligned}
\varphi(x_k, y) - \varphi(x, y_k) \leq& \langle \widetilde{\nabla}\varphi(z_k) - \mathbb{E}\left[\widetilde{g_k}|z_k\right], z_k - z\rangle + \frac{\gamma_k}{2}\mathbb{E}\left[\|\widetilde{g_k}\|^2|z_k\right] \\
&+ \frac{\|z_k - z\|^2 - \mathbb{E}\left[\|z_{k+1} - z\|^2|z_k\right]}{2\gamma_k} - \frac{\gamma}{2}\|z_k - z\|^2.
\end{aligned} \tag{37}
$$

Step 2 (Bounding bias term). Our aim is to bound the first term in (37), namely $\langle \widetilde{\nabla}\varphi(z_k) - \mathbb{E}\left[\widetilde{g_k}|z_k\right], z_k - z\rangle$. Using the Taylor expansion we have

$$
\begin{aligned}
\varphi\left(z_k + \tau_k r_k e_k\right) =& \varphi(z_k) + \langle \nabla\varphi(z_k), \tau_k r_k e_k\rangle \\
&+ \sum_{2\leq|m|\leq l} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)}\varphi(z_k) e_k^m + R(\tau_k r_k e_k),
\end{aligned} \tag{38}
$$

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,

$$
\begin{aligned}
\widetilde{g_k} =& \Big(\langle \nabla\varphi(z_k), \tau_k r_k e_k\rangle + \sum_{2\leq|m|\leq l, |m| \text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)}\varphi(z_k) e_k^m \\
&+ \frac{1}{2}R(\tau_k r_k e_k) - \frac{1}{2}R(-\tau_k r_k e_k) + \xi_k - \xi_k' \Big) \frac{n}{\tau_k} K(r_k) \begin{pmatrix} (e_k)_x \\ -(e_k)_y \end{pmatrix}.
\end{aligned} \tag{39}
$$

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

$$
\mathbb{E}_{e_k, r_k}\left[\langle \nabla\varphi(z_k), \tau_k r_k e_k\rangle \frac{n}{\tau_k} K(r_k) \begin{pmatrix} (e_k)_x \\ -(e_k)_y \end{pmatrix} \middle| z_k \right] = \widetilde{\nabla}\varphi(z_k). \tag{40}
$$

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

$$\mathbb{E}\left[\Big(\sum_{2\le|m|\le l,|m|\text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)}\varphi(z_k)e_k^m + \xi_k - \xi_k'\Big)\frac{n}{\tau_k}K(r_k)\begin{pmatrix}(e_k)_x \\ -(e_k)_y\end{pmatrix}\Bigg| x_k\right] = 0. \tag{41}$$

Substituting (39), (40) and (41) in the first term in (37) and using the definition of κ_β (see (4)) we obtain

$$\begin{aligned}
&\left|\langle\widetilde{\nabla}\varphi(z_k) - \mathbb{E}\left[\widetilde{g}_k|z_k\right], z_k - z\rangle\right| = \\
&= \left|\mathbb{E}\left[\left(\frac{1}{2}R(\tau_k r_k e_k) - \frac{1}{2}R(-\tau_k r_k e_k)\right)\frac{n}{\tau_k}K(r_k)\left\langle\begin{pmatrix}(e_k)_x \\ -(e_k)_y\end{pmatrix}, z_k - z\right\rangle\Bigg| z_k\right]\right| \\
&\qquad\le L\tau_k^{\beta-1}\cdot\mathbb{E}_{r_k}\left[|r_k|^\beta K(r_k)\right]\cdot n\left|\mathbb{E}_{e_k}\left[\langle e_k, z_k - z\rangle|z_k\right]\right| \\
&\qquad\qquad\le \kappa_\beta L\sqrt{n}\tau_k^{\beta-1}\|z_k - z\|,
\end{aligned} \tag{42}$$

where in the last two inequalities the symmetry of Euclidean sphere and the fact from concentration measure theory that $|\mathbb{E}_e\left[\langle e, s\rangle\right]|^2 \le \mathbb{E}_e\left[\langle e, s\rangle^2\right] = \frac{\|s\|^2}{n}$ were used . Applying the inequality $ab \le {}^1/_2(a^2 + b^2)$ to the last expression in (42) we finally get

$$\left|\langle\widetilde{\nabla}\varphi(z_k) - \mathbb{E}\left[\widetilde{g}_k|z_k\right], z_k - z\rangle\right| \le \frac{(\kappa_\beta L)^2}{\gamma}n\tau_k^{2(\beta-1)} + \frac{\gamma}{4}\|z_k - z\|^2. \tag{43}$$

Step 3 (Bounding second moment of gradient estimator). Our aim is to estimate $\mathbb{E}\left[\|\widetilde{g}_k\|^2|z_k\right]$ which is the second term in (37). The expectation here is with respect to r_k, ξ_k and ξ_k'. To lighten the presentation and without loss of generality we drop the lower script k in all quantities.

We have

$$\begin{aligned}
\|\widetilde{g}\|^2 =& \frac{n^2}{4\tau^2}\left\|(\varphi(z+\tau re) - \varphi(z-\tau re) + \xi - \xi')K(r)\begin{pmatrix}e_x \\ -e_y\end{pmatrix}\right\|^2 \\
=& \frac{n^2}{4\tau^2}\left((\varphi(z+\tau re) - \varphi(z-\tau re) + \xi - \xi')\right)^2 K^2(r).
\end{aligned} \tag{44}$$

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

$$\mathbb{E}\left[\|\widetilde{g}\|^2|z\right] \le \frac{3n^2}{4\tau^2}\left(\mathbb{E}\left[(\varphi(z+\tau re) - \varphi(z-\tau re))^2K^2(r)\big|z\right] + 2\kappa\sigma^2\right). \tag{45}$$

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

$$
\begin{aligned}
\mathbb{E}\left[(\varphi(z+e)-\varphi(z-e))^2 \middle| z\right] &= \mathbb{E}_e\left[(\varphi(z+e)-\varphi(z-e))^2\right] \\
&\le \mathbb{E}_e\left[((\varphi(z+e)-\mathbb{E}_e[\varphi(z+e)])-(\varphi(z-e)-\mathbb{E}_e[\varphi(z-e)]))^2\right] \\
&\le 2\mathbb{E}_e\left[(\varphi(z+e)-\mathbb{E}_e[\varphi(z+e)])^2\right]+2\mathbb{E}_e\left[(\varphi(z-e)-\mathbb{E}_e[\varphi(z-e)])^2\right] \\
&\le 2\sqrt{\mathbb{E}_e\left[(\varphi(z+e)-\mathbb{E}_e[\varphi(z+e)])^4\right]}+2\sqrt{\mathbb{E}_e\left[(\varphi(z-e)-\mathbb{E}_e[\varphi(z-e)])^4\right]} \\
&\le \frac{12G^2}{n}, \qquad (46)
\end{aligned}
$$

where in the last inequality (20) was used, so we have

$$
\mathbb{E}\left[(\varphi(z+\tau re)-\varphi(z-\tau re))^2 \middle| z\right] \le \frac{12(\tau r)^2G^2}{n} \le \frac{12\tau^2G^2}{n}. \qquad (47)
$$

By substituting (47) into (45), using independence of e and r and returning the lower script k we finally get

$$
\mathbb{E}\left[\|\widetilde{g}_k\|^2|z_k\right] \le \kappa\left(9nG^2+\frac{3(n\sigma)^2}{2\tau_k^2}\right). \qquad (48)
$$

Step 4. Let ρ_k^2 denote full expectation $\mathbb{E}[\|z_k-z\|^2]$. Substituting (43) and (48) into (37), taking full expectation we obtain

$$
\begin{aligned}
\mathbb{E}[\varphi(x_k,y)-\varphi(x,y_k)] \le &\frac{(\kappa_\beta L)^2}{\gamma}n\tau_k^{2(\beta-1)}+\frac{\gamma_k}{2}\kappa\left(9nG^2+\frac{3(n\sigma)^2}{2\tau_k^2}\right) \\
&+\frac{\rho_k^2-\rho_{k+1}^2}{2\gamma_k}-\left(\frac{\gamma}{2}-\frac{\gamma}{4}\right)\rho_k^2.
\end{aligned} \qquad (49)
$$

Using the convexity-concavity of φ and (49) we have

$$
\begin{aligned}
\mathbb{E}\left[\varphi\left(\overline{x}_N,y\right)-\varphi\left(x,\overline{y}_N\right)\right] &\le \frac{1}{N}\sum_{k=1}^N\varphi\left(x_k,y\right)-\frac{1}{N}\sum_{k=1}^N\varphi\left(x,y_k\right) \\
&\le \frac{1}{N}\sum_{k=1}^N\left(\frac{(\kappa_\beta L)^2}{\gamma}n\tau_k^{2(\beta-1)}+\frac{\gamma_k}{2}\kappa\left(9nG^2+\frac{3(n\sigma)^2}{2\tau_k^2}\right)\right) \\
&+\frac{1}{N}\sum_{k=1}^N\left(\frac{\rho_k^2-\rho_{k+1}^2}{2\gamma_k}-\frac{\gamma}{4}\rho_k^2\right).
\end{aligned} \qquad (50)
$$

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

$$\begin{aligned}\sum_{k=1}^{N}\left(\frac{\rho_k^2-\rho_{k+1}^2}{2\gamma_k}-\frac{\gamma}{4}\rho_k^2\right) &\le \rho_1^2\left(\frac{1}{2\gamma_1}-\frac{\gamma}{4}\right)+\sum_{k=2}^{N+1}\rho_k^2\left(\frac{1}{2\gamma_k}-\frac{1}{2\gamma_{k-1}}-\frac{\gamma}{4}\right)\\ &= \rho_1^2\left(\frac{\gamma}{4}-\frac{\gamma}{4}\right)+\sum_{k=2}^{N+1}\rho_k^2\left(\frac{\gamma}{4}-\frac{\gamma}{4}\right)=0.\end{aligned} \tag{51}$$

Substituting (51) into (49) with $\gamma_k = \frac{2}{\gamma k}$ we obtain

$$\begin{aligned}&\mathbb{E}[\varphi\left(\overline{x}_N, y\right)-\varphi\left(x,\overline{y}_N\right)]\\ &\qquad\le \frac{1}{\gamma N}\sum_{k=1}^{N}\left((\kappa_\beta L)^2 n\tau_k^{2(\beta-1)}+\kappa\left(9nG^2+\frac{3(n\sigma)^2}{2\tau_k^2}\right)\frac{1}{k}\right)\\ &\qquad= \frac{1}{\gamma N}\sum_{k=1}^{N}\left(\left[n\cdot(\kappa_\beta L)^2\tau_k^{2(\beta-1)}+n^2\cdot\frac{3\kappa\sigma^2}{2k\tau_k^2}\right]+\frac{9\kappa nG^2}{k}\right).\end{aligned} \tag{52}$$

If $\sigma > 0$ then $\tau_k = \left(\dfrac{3\kappa\sigma^2 n}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}$ is the minimizer of square brackets. Plugging this τ_k in (52) and using two inequalities: for the expression in square brackets $\sum\limits_{k=1}^{N} k^{-1+1/\beta} \le \beta N^{1/\beta}$ (if $\beta > 2$) and for the term after square brackets $\sum\limits_{k=1}^{N}\frac{1}{k} \le 1+\ln N$ we get

$$\mathbb{E}[\varphi\left(\overline{x}_N, y\right)-\varphi\left(x,\overline{y}_N\right)] \le \frac{1}{\gamma}\left(n^{2-\frac{1}{\beta}}\frac{A_1}{N^{\frac{\beta-1}{\beta}}}+A_2\frac{n(1+\ln N)}{N}\right).$$

with A_1 and A_2 from the formulation of Theorem 3.

Taking the minimum over x and the maximum over y we finally obtain

$$\begin{aligned}\mathbb{E}\left[\varphi(\overline{x}_N, y^*)-\varphi(x^*,\overline{y}_N)\right] &\le \max_{y\in\mathcal{Y}}\mathbb{E}\left[\varphi(\overline{x}_N,y)\right]-\min_{x\in\mathcal{X}}\mathbb{E}\left[\varphi(x,\overline{y}_N)\right]\\ &\le \frac{1}{\gamma}\left(n^{2-\frac{1}{\beta}}\frac{A_1}{N^{\frac{\beta-1}{\beta}}}+A_2\frac{n(1+\ln N)}{N}\right).\end{aligned}$$

Theorem 4. Let $\varphi \in \Phi_\beta(L)$ with $L > 0$ and $\beta > 2$. Let Assumption 1 hold and let $\mathcal{Z}$ be a convex compact subset of $\mathbb{R}^n$. Let φ be G-Lipschitz on the Euclidean τ_1-neighborhood of $\mathcal{Z}$ (τ_k is parameter from Theorem 3 for the regularized function $\varphi_\gamma(z)$ whose description is given below). Let $\overline{z}_N$ denote $\frac{1}{N}\sum\limits_{k=1}^{N} z_k$.

Proof. Let's define $N(\varepsilon)$:

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

where $A_1 = 3\beta(\kappa\sigma^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L)^{\frac{2}{\beta}}$, $A_2 = 9\kappa G^2$ – constants from Theorem 3, $\rho > 0$ – arbitrarily small positive number, c' – constant which depends on ρ.

Then the rate of convergence is given by the following expression:

$$\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] \leq \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \leq \varepsilon \quad (53)$$

after $N(\varepsilon)$ steps of Algorithm 2 with settings from Theorem 3 for the regularized function: $\varphi_\gamma(z) := \varphi(z) + \frac{\gamma}{2}\|x - x_0\|^2 - \frac{\gamma}{2}\|y - y_0\|^2$, where $\gamma \leq \frac{\varepsilon}{R^2}$, $R = \|z_0 - z^*\|$, $z_0 \in \mathcal{Z}$ – arbitrary point.

Proof. Step 1. Let $z^* = (x^*, y^*)$ and $z^*_\gamma = (x^*_\gamma, y^*_\gamma)$ denote the solutions of the saddle-point problems for functions $\varphi(z)$ and $\varphi_\gamma(z)$ respectively. Let $\mathring{\overline{x}}_N$ denote $\overline{x}_N - x_0$, $\mathring{\overline{y}}_N$ denote $\overline{y}_N - y_0$ respectively. Let $\mathring{x}$ denote $x - x_0$, $\mathring{y}$ denote $y - y_0$ and $\mathring{z}$ denote $z - z_0$, where $z = (x, y)$, $z_0 = (x_0, y_0)$ and so on.

Setting $\gamma = \frac{\varepsilon}{R^2}$ and using the inequality $\varphi_\gamma(\overline{x}_N, y^*) - \varphi_\gamma(x^*, \overline{y}_N) \leq \varphi_\gamma\left(\overline{x}_N, y^*_\gamma\right) - \varphi_\gamma\left(x^*_\gamma, \overline{y}_N\right)$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[\varphi(\overline{x}_N, y^*)\right] - \mathbb{E}\left[\varphi(x^*, \overline{y}_N)\right] \leq \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \\
&= \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\gamma(\overline{x}_N, y) - \varphi_\gamma(x, \overline{y}_N) - \frac{\gamma\mathring{\overline{x}}_N^2}{2} + \frac{\gamma\mathring{y}^2}{2} + \frac{\gamma\mathring{x}^2}{2} - \frac{\gamma\mathring{\overline{y}}_N^2}{2}\right] \\
&\leq \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\gamma(\overline{x}_N, y) - \varphi_\gamma(x, \overline{y}_N) + \frac{\gamma\mathring{z}^2}{2}\right] \\
&\leq \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\gamma(\overline{x}_N, y) - \varphi_\gamma(x, \overline{y}_N)\right] + \frac{\varepsilon}{2} \\
&= \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\gamma(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi_\gamma(x, \overline{y}_N)\right] + \frac{\varepsilon}{2}
\end{aligned} \quad (54)$$

Step 2. Now we apply Theorem 3 for $\varphi_\gamma(z)$ until function error is not greater than $\frac{\varepsilon}{2}$:

$$\max_{y\in\mathcal{Y}}\mathbb{E}\left[\varphi_\gamma(\overline{x}_N,y)\right]-\min_{x\in\mathcal{X}}\mathbb{E}\left[\varphi_\gamma(x,\overline{y}_N)\right]\leq\frac{1}{\gamma}\left(n^{2-\frac{1}{\beta}}\frac{A_1}{N^{\frac{\beta-1}{\beta}}}+A_2\frac{n(1+\ln N)}{N}\right)\leq\frac{\varepsilon}{2}. \tag{55}$$

Using that $\gamma=\frac{\varepsilon}{R^2}$ the inequality (55) is done if

$$\max\left\{n^{2-\frac{1}{\beta}}\frac{A_1}{N^{\frac{\beta-1}{\beta}}},A_2\frac{n(1+\ln N)}{N}\right\}\leq\frac{\gamma\varepsilon}{2}=\frac{\varepsilon^2}{2R^2}. \tag{56}$$

It is true that $1+\ln N\leq c'N^{\frac{\rho}{\rho+1}}$ for some $c'>0$. So the inequality (56) holds if

$$N\geq\max\left\{\left(R\sqrt{2A_1}\right)^{\frac{2\beta}{\beta-1}}\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}},\left(R\sqrt{2c'A_2}\right)^{2(1+\rho)}\frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}}\right\}. \tag{57}$$

The inequalities (54) and (55) yield (53).

□

5 Numerical experiment

In our experiment we compare the Algorithm 1 (with $\beta=3$ and $\beta=5$) proposed here with Gasnikov's one-point method for the minimization problem.

We consider the problem of the minimization of the following function

$$f(x)=\frac{1}{2}x^TAx+\frac{1}{10}\sum_{k=1}^{50}|x_k|^4$$

on the Euclidean ball $Q=\{x\in\mathbb{R}^{50}:\ \|x\|\leq 1\}$.

The starting point is x_0 with $\|x_0\|=1/2$. The dependency of $f(\overline{x}_N)-f(x^*)$ (optimization error) on N (iteration number) is presented on Figure 2. The optimization error has its mean and 0.95-confidence interval. As the constant L for Algorithm 1 with $\beta=5$ is equal to zero we choose $L=0.001$.

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

Moreover, from Table 3 we see that the dependence of ε on N is better when we use higher-order smoothness.

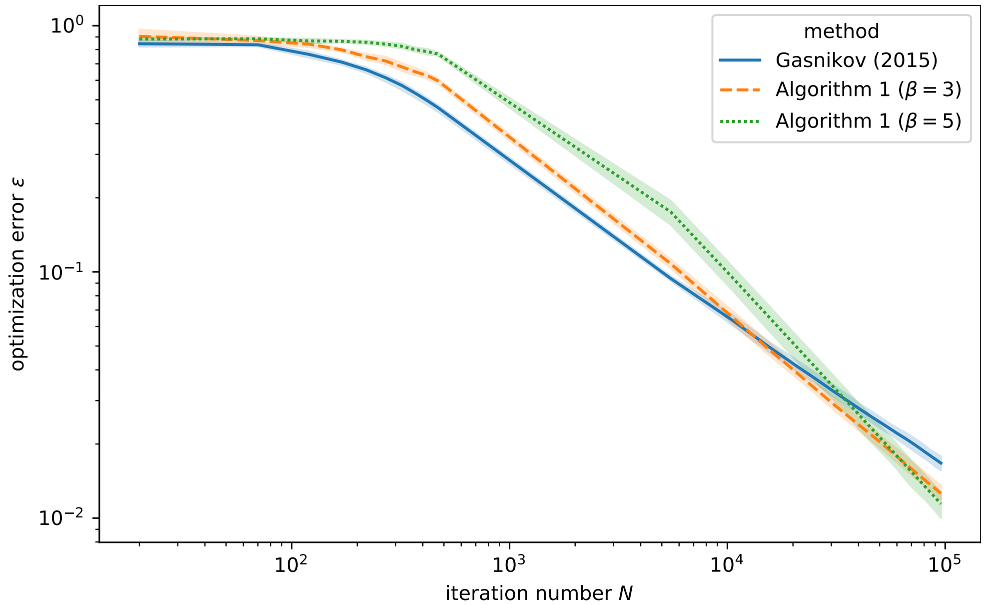

Figure 2: Dependency of optimization error of Algorithm 1 on iteration

Table 3: The dependence of optimization error ε on iteration number N

	theory	experiment
Gasnikov, $\beta = 2$, 2015	$\varepsilon \sim N^{-0.5}$	$\varepsilon \sim N^{-0.61}$
Algorithm 1, $\beta = 3$, 2020	$\varepsilon \sim N^{-2/3}$	$\varepsilon \sim N^{-0.73}$
Algorithm 1, $\beta = 5$, 2020	$\varepsilon \sim N^{-4/5}$	$\varepsilon \sim N^{-0.91}$

6 Conclusion

We have got to the best of out knowledge the best upper bounds (dependence of ε on n (dimension)) among optimal dependencies of ε on N for the problem of minimization of strongly-convex and convex functions. However, the lower bound is not achieved yet.

Possible genelization 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 $\|\widetilde{g}_k - \nabla f(x_k)\|$ and this can causes 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 dimenstion n and strong convexity parameter γ can be obtained.

Bibliography

[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 // Conference on Learning Theory. "— 2016. "— Pp. 257–283.

[3] Boosting one-point derivative-free online optimization via residual feedback / Y. Zhang, Y. Zhou, K. Ji, M. M. Zavlanos // *arXiv preprint arXiv:2010.07378*. "— 2020.

[4] *Bubeck S., Lee Y. T., Eldan R.* Kernel-based methods for bandit convex optimization // Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. "— 2017. "— Pp. 72–85.

[5] *Conn A. R., Scheinberg K., Vicente L. N.* Introduction to Derivative-Free Optimization. "— Society for Industrial and Applied Mathematics, 2009.

[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] *Goodfellow I.* Nips 2016 tutorial: Generative adversarial networks // *arXiv preprint arXiv:1701.00160*. "— 2016.

[9] *Jin Y., Sidford A.* Efficiently solving MDPs with stochastic mirror descent // Proceedings of the 37th International Conference on Machine

Learning / Ed. by H. D. III, A. Singh. "— Vol. 119 of *Proceedings of Machine Learning Research.* "— PMLR, 2020. "— 13–18 Jul. "— Pp. 4890–4900.

[10] *Larson J., Menickelly M., Wild S. M.* Derivative-free optimization methods // *Acta Numerica.* "— 2019. "— Vol. 28. "— P. 287–404.

[11] Optimal rates for zero-order convex optimization: The power of two function evaluations / J. C. Duchi, M. I. Jordan, M. J. Wainwright, A. Wibisono // *IEEE Transactions on Information Theory.* "— 2015. "— Vol. 61, no. 5. "— Pp. 2788–2806.

[12] *Polyak B. T., Tsybakov A. B.* Optimal order of accuracy of search algorithms in stochastic optimization // *Problemy Peredachi Informatsii.* "— 1990. "— Vol. 26, no. 2. "— Pp. 45–53.

[13] *Shamir O.* An optimal algorithm for bandit and zero-order convex optimization with two-point feedback // *The Journal of Machine Learning Research.* "— 2017. "— Vol. 18, no. 1. "— Pp. 1703–1713.

[14] *Spall J. C.* Introduction to Stochastic Search and Optimization. "— 1 edition. "— New York, NY, USA: John Wiley & Sons, Inc., 2003.

[15] Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case / A. V. Gasnikov, E. A. Krymova, A. A. Lagunovskaya et al. // *Automation and remote control.* "— 2017. "— Vol. 78, no. 2. "— Pp. 224–234.

[16] Zeroth-order algorithms for smooth saddle-point problems / A. Sadiev, A. Beznosikov, P. Dvurechensky, A. Gasnikov // *arXiv preprint arXiv:2009.09908.* "— 2020.