
Zeroth-order methods for noisy Hölder-gradient functions

Innokentiy Shibaev^{2,1}
Pavel Dvurechensky^{3,2}
Alexander Gasnikov^{1,2,3}

Abstract In this paper, we prove new complexity bounds for zeroth-order methods in non-convex optimization with inexact observations of the objective function values. We use the Gaussian smoothing approach of [14] and extend their results, obtained for optimization methods for smooth zeroth-order non-convex problems, to the setting of minimization of functions with Hölder-continuous gradient with noisy zeroth-order oracle, obtaining noise upper-bounds as well. We consider finite-difference gradient approximation based on normally distributed random Gaussian vectors and prove that gradient descent scheme based on this approximation converges to the stationary point of the smoothed function. We also consider convergence to the stationary point of the original (not smoothed) function and obtain bounds on the number of steps of the algorithm for making the norm of its gradient small. Additionally we provide bounds for the level of noise in the zeroth-order oracle for which it is still possible to guarantee that the above bounds hold. We also consider separately the case of $\nu = 1$ and show that in this case the dependence of the obtained bounds on the dimension can be improved.

Keywords gradient-free methods · zeroth-order optimization · non-convex problem · inexact oracle

I. Shibaev
E-mail: innokentiy.shibayev@phystech.edu

P. Dvurechensky
E-mail: pavel.dvurechensky@wias-berlin.de

A. Gasnikov
E-mail: gasnikov@yandex.ru

¹Moscow Institute of Physics and Technology, Moscow, Russia

²HSE University, Moscow, Russia

³Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

1 Introduction

The main advantage of zeroth-order (derivative-free) optimization methods [15, 8, 5, 17, 6, 11] is that computing function value is, in general, simpler than computing its gradient vector. On the one hand, zeroth-order methods usually have worse convergence rates, and may be inferior to gradient methods endowed with Fast Automatic Differentiation (FAD) technique, for which it is known [10, 1] that if there is a series of computational operations to evaluate the value of a function, then with at most four times large number of arithmetic operations it is possible to evaluate the gradient of this function. On the other hand, there are still a number of situations, when the objective is given as a black-box, there is no access to function derivatives and the FAD technique is not applicable. One of the many important recent examples is Reinforcement Learning problems, where the goal is to find an optimal control strategy by observing, in a stochastic environment, some black-box reward function values, see [18] for a review and examples. The problem can be even more complicated when one deals with computer simulation of some physical processes, e.g. satellite movement, since such models often have some noise in their outputs. Similarly, in Reinforcement Learning only noisy observations of the reward function are available. Moreover, the noise can be biased [19] and standard batch averaging may not help. Thus, it is important to analyze zeroth-order methods in the setting of possibly biased noisy observations of the objective function.

Another important application of zeroth-order methods with noisy observations is min-max or min-min problems, which are particular settings of bi-level optimization problems. For example, in [4] the authors consider the problem

$$\min_x \{f(x) = \max_y L(x, y)\},$$

where f has locally Hölder-continuous gradient and only inexact values of f and its gradient are available via inexact solution to the inner maximization problem in y . This leads to a non-convex minimization problem with inexact oracle and the authors focus on first-order inexact oracle. Motivated, in particular, by such problems, we consider in this paper the case when only noisy observations of the objective value f are available. Our bounds on the noise help to evaluate what accuracy of the solution to the inner problem is sufficient to solve the outer problem with some desired accuracy.

Related works. In [14], among other settings, the authors consider minimization of a non-convex function f on \mathbb{R}^n with exact values of f and used the Gaussian smoothing technique with parameter μ to prove convergence to a stationary point of a smoothed function $f_\mu(x)$, which is a uniform approximation to $f(x)$. The main idea is that the smoothed function $f_\mu(x)$ has better properties, e.g. it is smooth even if f is non-smooth. In the case when f has Lipschitz-continuous gradient, the authors of [14] prove that their method achieves $\mathbb{E} [\|\nabla f(x_N)\|_*^2] \leq \varepsilon_{\nabla f}$ after $N = O\left(\frac{n}{\varepsilon_{\nabla f}}\right)$ steps with 2 oracle calls in each step. When f is Lipschitz-continuous they estimate an ap-

appropriate value of the parameter μ such that the smoothed function $f_\mu(x)$ satisfies $|f_\mu(x) - f(x)| \leq \varepsilon_f$ for all $x \in \mathbb{R}^n$, and prove that in order to obtain $\mathbb{E} [\|\nabla f_\mu(x_N)\|_*^2] \leq \varepsilon_{\nabla f}$ it is sufficient to make $N = O\left(\frac{n^3}{\varepsilon_f \varepsilon_{\nabla f}^2}\right)$ steps of their method with 2 oracle calls in each step.

This technique was later used in the works [9] (RSGF algorithm) and [16] (RSPGF algorithm) to build an algorithm which finds so-called (ε, Λ) -solution i.e. a point x s.t. $\mathbb{P}\{\|\nabla f(x)\|_*^2 \leq \varepsilon_{\nabla f}\} \geq 1 - \Lambda$, in the case of Lipschitz-gradient function and stochastic oracle $F(x, \xi)$ s.t. $\mathbb{E}_\xi[F(x, \xi)] = f(x)$. They have shown that to find an $(\varepsilon_{\nabla f}, \Lambda)$ -solution it is sufficient to make $O\left(C_1 \frac{n}{\varepsilon_{\nabla f}} + C_2 \frac{n}{\varepsilon_{\nabla f}^2}\right)$ calls to the stochastic zeroth-order oracle (here the constants C_1, C_2 depend on Λ and other parameters of the problem, such as Lipschitz constant and diameter of the feasible set).

In the works [2, 3] the authors compare several types of gradient approximations $g(x)$, including Gaussian smoothing and smoothing based on uniform sampling on the Euclidean sphere, in terms of the number of calls to the inexact zeroth-order oracle $\hat{f}(x)$ for f which guarantees the approximation condition $\|g(x) - \nabla f(x)\|_* \leq \theta \|\nabla f(x)\|_*$, where $\theta \in [0, 1)$. They show that random-directions-based methods lose in theory to the standard finite differences approach, needing more oracle calls to ensure the above approximation condition. However, in the work [12] zeroth-order variants of stochastic variance reduction methods called ZO-SVRG are considered and a variant which uses random directions approach in the experiments required less number of oracle calls than the standard finite differences method (ZO-SVRG-Coord), despite having worse theoretical convergence rate. In this paper we do not rely on the above approximation condition, which allows to obtain better complexity bounds for the considered approach based on random directions and Gaussian smoothing.

Our contributions. The works listed above mainly focus on the setting when the objective f has Lipschitz-continuous gradient. The only paper, which considers non-smooth setting with f being Lipschitz continuous is [14], where the value of the objective f is assumed to be known exactly. Our main contribution consists in obtaining complexity bounds for zeroth-order methods with inexact values of the objective in the setting of f having Hölder-continuous gradient, i.e. for some $L_\nu > 0, \nu \in [0, 1]$, $\|\nabla f(x) - \nabla f(y)\|_* \leq L_\nu \|x - y\|^\nu$. This assumption is more general and includes as particular cases the previously considered settings of objectives f with Lipschitz-continuous gradient and objectives f which are differentiable and Lipschitz continuous. Our approach uses finite-difference gradient approximation based on normally distributed random Gaussian vectors u and we prove that a gradient descent scheme based on this approximation ensures

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_\mu(x_k)\|_*^2] \leq \varepsilon_{\nabla f} \text{ after } N = O\left(\frac{n^{\frac{7-3\nu}{2}}}{\frac{3-\nu}{1+\nu} \varepsilon_{\nabla f}}\right)$$

steps. Here x_k are the iterates, $f_\mu(\cdot)$ is a smoothed version of the objective f , $\mathcal{U} = (u_0, \dots, u_{N-1})$ is the history of the realizations of random Gaussian vector u . We also consider convergence to a stationary point of the initial (not smoothed) objective function f and prove that this scheme ensures

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x_k)\|_*^2] \leq \varepsilon_{\nabla f} \text{ after } N = O\left(\frac{n^{2+\frac{(1-\nu)}{2\nu}}}{\varepsilon_{\nabla f}^{\frac{1}{\nu}}}\right)$$

steps, when $\nu \in (0, 1]$. For both cases we obtain bounds for the maximum level of noise in the zeroth-order oracle which does not affect the above iteration complexity bounds. The main difference of our work from [14] is that we consider the inexact oracle setting, intermediate smoothness $\nu \in [0, 1]$ rather than the cases $\nu \in 0, 1$, which we also cover in a unified manner. We additionally provide a refined analysis for the case of $\nu = 1$ to achieve the complexity bound $N = O\left(\frac{n}{\varepsilon_{\nabla f}}\right)$ both for $\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_\mu(x_k)\|_*^2] \leq \varepsilon_{\nabla f}$ and $\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x_k)\|_*^2] \leq \varepsilon_{\nabla f}$, which is similar to the bound in [14] for this case.

The rest of the paper is organized as follows. The first section contains necessary definitions and some technical lemmas which extend or improve the corresponding bounds derived in [14]. In the second section, we consider a simple gradient descent process with Gaussian-sampling-based finite-difference gradient approximation and obtain complexity bounds for this method in terms of the gradient norm of the smoothed and of the non-smoothed function. We also analyze how the noise in the objective values influences the convergence and what level of inexactness can be tolerated without changing the convergence properties.

2 Gaussian smoothing, zeroth-order oracle

This section provides problem statement, technical preliminaries and properties of the function f_μ obtained from f by Gaussian smoothing, as well as the gradient of f_μ , and the estimates for the difference between f and f_μ as well as their gradients.

2.1 Definitions

We mostly follow the notation in [14] and [7], where a similar problem was considered from the point of view of inexact first-order oracle. We start with some definitions from [14]. For an n -dimensional space E , we denote by E^* its dual space. The value of a linear function $s \in E^*$ at point $x \in E$ is denoted by $\langle s, x \rangle$. We endow the spaces E and E^* with Euclidean norms

$$\|x\|^2 = \langle Bx, x \rangle, \quad \forall x \in E \quad \|s\|_*^2 = \langle s, B^{-1}s \rangle, \quad \forall s \in E^*, \quad (1)$$

where $B : E \rightarrow E^*$ is a linear operator s.t. $B \succ 0$.

In this paper we consider the problem of the form

$$\min_{x \in E} f(x) \quad (2)$$

under the two following assumptions.

Assumption 1 *The function $f(x)$ is equipped with an inexact zeroth-order oracle $\tilde{f}(x, \delta)$ with some $\delta > 0$ i.e. there exists $\delta > 0$ and one can calculate $\tilde{f}(x, \delta) \in \mathbb{R}$ satisfying, for all $x \in E$,*

$$|f(x) - \tilde{f}(x, \delta)| \leq \delta. \quad (3)$$

Assumption 2 *The function $f(x)$ is differentiable with Hölder-continuous gradient with some $\nu \in [0, 1]$ and $L_\nu \geq 0$ i.e.*

$$\|\nabla f(y) - \nabla f(x)\|_* \leq L_\nu \|y - x\|^\nu, \quad \forall x, y \in E. \quad (4)$$

The latter inequality gives a useful inequality

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{1 + \nu} \|y - x\|^{1+\nu}, \quad \forall x, y \in E. \quad (5)$$

Next, we consider the Gaussian smoothed version of $f(x)$ defined in [14].

Definition 1 Consider a function $f : E \rightarrow \mathbb{R}$. Its Gaussian approximation $f_\mu(x)$ is defined as

$$f_\mu(x) = \frac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du, \quad (6)$$

where

$$\kappa \stackrel{\text{def}}{=} \int_E e^{-\frac{1}{2}\|u\|^2} du = \frac{(2\pi)^{n/2}}{[\det B]^{1/2}}. \quad (7)$$

It can be shown, that (see [14] Section 2 for details)

$$\nabla f_\mu(x) = \frac{1}{\kappa} \int_E \frac{f(x + \mu u) - f(x)}{\mu} e^{-\frac{1}{2}\|u\|^2} B u du \quad (8)$$

$$= \frac{1}{\kappa} \int_E \frac{f(x + \mu u)}{\mu} e^{-\frac{1}{2}\|u\|^2} B u du \quad (9)$$

$$\nabla f(x) = \frac{1}{\kappa} \int_E \langle \nabla f(x), u \rangle e^{-\frac{1}{2}\|u\|^2} B u du, \quad (10)$$

where the latter equality holds when $f(x)$ is differentiable at x . If f is differentiable on E , then

$$\nabla f_\mu(x) = \frac{1}{\kappa} \int_E \nabla f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du. \quad (11)$$

The Gaussian approximation of the function $\tilde{f}(x, \delta)$ then takes the form

$$\tilde{f}_\mu(x, \delta) = \frac{1}{\kappa} \int_E \tilde{f}(x + \mu u, \delta) e^{-\frac{1}{2}\|u\|^2} du. \quad (12)$$

We also define the following vector which plays the role of the gradient of $\tilde{f}_\mu(x, \delta)$

$$\nabla \tilde{f}_\mu(x, \delta) = \frac{1}{\kappa} \int_E \frac{\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta)}{\mu} e^{-\frac{1}{2}\|u\|^2} B u du \quad (13)$$

$$= \frac{1}{\kappa} \int_E \frac{\tilde{f}(x + \mu u, \delta)}{\mu} e^{-\frac{1}{2}\|u\|^2} B u du. \quad (14)$$

For the case of $\delta = 0$ we have $\nabla \tilde{f}_\mu(x, \delta) = \nabla f_\mu(x)$. It is also worth noting that, in general, it is not possible to obtain for $\nabla \tilde{f}_\mu(x, \delta)$ a representation similar to (11) since the function $\tilde{f}(x, \delta)$ is not necessarily differentiable.

2.2 Basic results

As shown in Lemma 3 in [14], for $f(x)$ with Lipschitz-continuous gradient, it holds that

$$\|\nabla f_\mu(x) - \nabla f(x)\|_* \leq \frac{\mu L_1}{2} (n+3)^{3/2}.$$

This result was improved and extended (see A.1 in [2]) to the noisy case giving

$$\|\nabla \tilde{f}_\mu(x, \delta) - \nabla f(x)\|_* \leq \frac{\delta}{\mu} n^{1/2} + \mu L_1 n^{1/2}.$$

Extending it to the Hölder case we can show the following result.

Lemma 1 *Under Assumptions 1 and 2 it holds that*

$$\begin{aligned} \|\nabla \tilde{f}_\mu(x, \delta) - \nabla f_\mu(x)\|_* &\leq \frac{\delta}{\mu} n^{1/2} \\ \|\nabla f_\mu(x) - \nabla f(x)\|_* &\leq \mu^\nu L_\nu n^{\nu/2}, \end{aligned}$$

and, consequently,

$$\|\nabla \tilde{f}_\mu(x, \delta) - \nabla f(x)\|_* \leq \frac{\delta}{\mu} n^{1/2} + \mu^\nu L_\nu n^{\nu/2}. \quad (15)$$

Proof [Can be found in Appendix.](#)

It can be shown (assuming f is Lipschitz-continuous with constant L_0), that f_μ has Hölder-continuous gradient with $\nu = 1$ and $L = \frac{n^{1/2}}{\mu}L_0$ (Lemma 2 from [14]). Thus, we can obtain in this case that

$$|f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle| \leq \frac{L}{2} \|x - y\|^2. \quad (16)$$

Under more general Hölder condition we can obtain the following inexact version of (16).

Lemma 2 *Under Assumption 2 it holds that*

$$|f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle| \leq \frac{A_1}{2} \|y - x\|^2 + A_2,$$

where either

$$A_1 = \frac{L_\nu}{\mu^{1-\nu}} n^{\frac{1+\nu}{2}}, \quad A_2 = 0,$$

or

$$A_1 = \left[\frac{1}{\hat{\delta}} \right]^{\frac{1-\nu}{1+\nu}} \frac{2L_\nu}{\mu^{1-\nu}}, \quad A_2 = \hat{\delta} L_\nu \mu^{1+\nu} \quad \text{where } \hat{\delta} > 0.$$

Proof [Can be found in Appendix.](#)

One of the most important properties of the smoothed function $f_\mu(x)$ is that it provides a uniform approximation for f . For example, when f is Lipschitz-continuous with constant L_0 it can be shown (see Theorem 1 from [14]) that

$$|f_\mu(x) - f(x)| \leq \mu L_0 n^{1/2}.$$

For the more general case of Hölder-continuous gradient we obtain the following more general result.

Lemma 3 *Under Assumption 2 it can be shown that*

$$|f_\mu(x) - f(x)| \leq \frac{L_\nu}{1+\nu} \mu^{1+\nu} n^{\frac{1+\nu}{2}}.$$

Proof [Can be found in Appendix.](#)

From Lemma 1 we can obtain an upper bound which connects the gradient norm of f and gradient norm of its smoothed approximation f_μ . This will be the key to translate the convergence rate for the smoothed function gradient to the convergence rate of the original objective f gradient.

Lemma 4 *Under Assumption 2 it holds that*

$$\|\nabla f(x)\|_*^2 \leq 2\|\nabla f_\mu(x)\|_*^2 + 2\mu^{2\nu} L_\nu^2 n^\nu.$$

Proof Can be found in Appendix.

In the next section we consider a gradient descent method with gradient replaced with a random gradient estimation

$$g_\mu(x, u, \delta) = \frac{\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta)}{\mu} Bu, \quad (17)$$

where u is a Gaussian random vector with mean 0_n and identity covariance matrix I_n ($u \sim \mathcal{N}(0, I_n)$). Thus $\mathbb{E}_u [g_\mu(x, u, \delta)] = \nabla \tilde{f}_\mu(x, \delta)$. In what follows we need also one technical result about this estimation.

Lemma 5 *Under Assumptions 1 and 2 for the gradient estimation (17) it holds that*

$$\begin{aligned} \mathbb{E}_u [\|g_\mu(x, u, \delta)\|_*^2] &\leq 20(n+4)\|\nabla f_\mu(x)\|^2 + \\ &+ 5 \left(\frac{4\delta^2}{\mu^2}n + \frac{4L_\nu^2}{(1+\nu)^2}\mu^{2\nu}n^{2+\nu} + \frac{\mu^2 A_1^2}{4}(n+6)^3 + \frac{A_2^2}{\mu^2}n \right), \end{aligned}$$

where A_1, A_2 are constants equal to constants A_1, A_2 from Lemma 2.

Proof Can be found in Appendix.

3 Convergence rate analysis

We consider a gradient descent process

$$x_{k+1} = x_k - h_k B^{-1} g_\mu(x_k, u_k, \delta), \quad (18)$$

where u_k is normal random vector and $g_\mu(x_k, u_k, \delta)$ is defined in (17). We will consider two type of convergence – in the sense of $\|\nabla f(x_k)\|_*$ and $\|\nabla f_\mu(x_k)\|_*$. We start with proving the following result

Lemma 6 *Consider the process (18). Under Assumptions 1 and 2 it can be shown that after $N - 1$ iterations of this process*

$$\begin{aligned} \min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_\mu(x_k)\|_*^2] &\leq \frac{320(n+4)A_1(f_\mu(x_0) - f^*)}{ND} + \\ &+ \frac{D}{4(n+4)} \left(\frac{4\delta^2}{\mu^2}n + \frac{4L_\nu^2}{(1+\nu)^2}\mu^{2\nu}n^{2+\nu} + \frac{\mu^2(A'_1)^2}{4}(n+6)^3 + \frac{(A'_2)^2}{\mu^2}n \right) + \\ &+ \frac{320(n+4)A_1}{D} \left(A_2 + \frac{\delta^2}{2A_1\mu^2}n + \frac{\delta^2}{\mu^2}n \right), \end{aligned} \quad (19)$$

where $\mathcal{U} = (u_0, \dots, u_{N-1})$ is a random vector composed by i.i.d. $\{u_k\}_{k=0}^{N-1}$, A_1, A_2 and A'_1, A'_2 are the independent pair of constants from Lemma 2 and $D \in (0, 1]$.

Proof From Lemma 1, Lemma 2 and the fact that $ab \leq \frac{Ca^2}{2} + \frac{b^2}{2C}$ where $C > 0$, $a = \|y - x\|$ and $b = \frac{\delta}{\mu}n^{1/2}$ we obtain

$$\begin{aligned} & |f_\mu(y) - f_\mu(x) - \langle \nabla \tilde{f}_\mu(x, \delta), y - x \rangle| \leq \\ & \leq |f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle| + |\langle \nabla \tilde{f}_\mu(x, \delta) - \nabla f_\mu(x), y - x \rangle| \leq \\ & \leq \frac{A_1}{2} \|y - x\|^2 + A_2 + \frac{\delta}{\mu} n^{1/2} \|y - x\| \leq \\ & \leq \left(\frac{A_1}{2} + \frac{C}{2} \right) \|y - x\|^2 + A_2 + \frac{\delta^2}{2C\mu^2} n \stackrel{C=A_1}{=} A_1 \|y - x\|^2 + \left(A_2 + \frac{\delta^2}{2A_1\mu^2} n \right). \end{aligned}$$

Consider a gradient descent process (18). Substituting it and (17) into the last inequality and taking the expectation in u_k we obtain

$$\begin{aligned} \mathbb{E}_{u_k} [f_\mu(x_{k+1})] & \leq f_\mu(x_k) - h_k \|\nabla \tilde{f}_\mu(x_k, \delta)\|_*^2 + h_k^2 A_1 \mathbb{E}_{u_k} [\|g_\mu(x_k, u_k, \delta)\|_*^2] + \\ & \quad + \left(A_2 + \frac{\delta^2}{2A_1\mu^2} n \right). \end{aligned}$$

Now let's use the fact that $(a + b)^2 \leq 2a^2 + 2b^2$

$$\begin{aligned} \|\nabla f_\mu(x)\|_*^2 & \leq 2\|\nabla \tilde{f}_\mu(x, \delta)\|_*^2 + 2\|\nabla f_\mu(x) - \nabla \tilde{f}_\mu(x, \delta)\|_*^2 \leq \\ & \leq 2\|\nabla \tilde{f}_\mu(x, \delta)\|_*^2 + 2 \cdot \frac{\delta^2}{\mu^2} n \end{aligned}$$

thus

$$\begin{aligned} \mathbb{E}_{u_k} [f_\mu(x_{k+1})] & \leq f_\mu(x_k) - \frac{h_k}{2} \|\nabla f_\mu(x_k)\|_*^2 + h_k^2 A_1 \mathbb{E}_{u_k} [\|g_\mu(x_k, u_k, \delta)\|_*^2] + \\ & \quad + \left(A_2 + \frac{\delta^2}{2A_1\mu^2} n + \frac{\delta^2}{\mu^2} n \right). \end{aligned}$$

Substituting result of Lemma 5 (we rename constants from this lemma with A'_1, A'_2 because it is the second pair of constants, and it can be chosen independently from A_1, A_2) we obtain

$$\begin{aligned} \mathbb{E}_{u_k} [f_\mu(x_{k+1})] & \leq f_\mu(x_k) - \left(\frac{h_k}{2} - 20(n+4)h_k^2 A_1 \right) \|\nabla f_\mu(x_k)\|_*^2 + \\ & \quad + h_k^2 A_1 5 \left(\frac{4\delta^2}{\mu^2} n + \frac{4L_\nu^2}{(1+\nu)^2} \mu^{2\nu} n^{2+\nu} + \frac{\mu^2 (A'_1)^2}{4} (n+6)^3 + \frac{(A'_2)^2}{\mu^2} n \right) + \\ & \quad + \left(A_2 + \frac{\delta^2}{2A_1\mu^2} n + \frac{\delta^2}{\mu^2} n \right). \end{aligned}$$

Let's choose $h = h_k = \frac{D}{80(n+4)A_1}$ where $D \in (0, 1]$ then

$$\begin{aligned} \mathbb{E}_{u_k} [f_\mu(x_{k+1})] & \leq f_\mu(x_k) - \frac{D}{320(n+4)A_1} \|\nabla f_\mu(x_k)\|_*^2 + \\ & \quad + \frac{5D^2}{A_1(80(n+4))^2} \left(\frac{4\delta^2}{\mu^2} n + \frac{4L_\nu^2}{(1+\nu)^2} \mu^{2\nu} n^{2+\nu} + \frac{\mu^2 (A'_1)^2}{4} (n+6)^3 + \frac{(A'_2)^2}{\mu^2} n \right) + \\ & \quad + \left(A_2 + \frac{\delta^2}{2A_1\mu^2} n + \frac{\delta^2}{\mu^2} n \right) \end{aligned}$$

and after summing and taking expectations in \mathcal{U} it becomes

$$\begin{aligned} \mathbb{E}_{\mathcal{U}} [f_{\mu}(x_N)] &\leq f_{\mu}(x_0) - \frac{D}{320(n+4)A_1} \sum_{k=0}^{N-1} \mathbb{E}_{\mathcal{U}} [\|\nabla f_{\mu}(x_k)\|_*^2] + \\ &+ \frac{5ND^2}{A_1(80(n+4))^2} \left(\frac{4\delta^2}{\mu^2} n + \frac{4L_{\nu}^2}{(1+\nu)^2} \mu^{2\nu} n^{2+\nu} + \frac{\mu^2(A'_1)^2}{4} (n+6)^3 + \frac{(A'_2)^2}{\mu^2} n \right) + \\ &+ N \left(A_2 + \frac{\delta^2}{2A_1\mu^2} n + \frac{\delta^2}{\mu^2} n \right). \end{aligned}$$

Rearranging terms and using the fact that $f^* \leq \mathbb{E}_{\mathcal{U}} [f_{\mu}(x_N)]$ we finally obtain (19). \square

And now we will use it to obtain the rate of convergence and noise bounds for two cases.

3.1 Convergence in the sense of $\|\nabla f(x_k)\|_*$

Theorem 1 Consider the process (18) and Assumptions 1 and 2. Suppose we want to ensure

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x_k)\|_*^2] \leq \varepsilon_{\nabla f}$$

then it can be shown that with the right choice of the smoothing parameter μ this inequality holds after

$$N = O \left(\frac{n^{2+\frac{1-\nu}{2\nu}}}{\varepsilon_{\nabla f}^{\frac{1}{\nu}}} \right) \quad (20)$$

steps of the process (18) under the assumption that

$$\delta < \frac{\mu^{\frac{3+\nu}{2}}}{n^{\frac{3-\nu}{4}}} = O \left(\frac{\varepsilon_{\nabla f}^{\frac{3+\nu}{4\nu}}}{n^{\frac{3+7\nu}{4\nu}}} \right). \quad (21)$$

Proof We will use Lemma 4 to replace the gradient norm with the gradient norm of the smoothed function and then use Lemma 6

$$\begin{aligned} \min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x_k)\|_*^2] &\stackrel{Lem. 4}{\leq} \min_{k \in \{0, N-1\}} (2\mathbb{E}_{\mathcal{U}} [\|\nabla f_{\mu}(x_k)\|_*^2] + 2\mu^{2\nu} L_{\nu}^2 n^{\nu}) \stackrel{(19)}{\leq} \\ &\leq \frac{640(n+4)A_1(f_{\mu}(x_0) - f^*)}{ND} + \\ &+ \frac{D}{2(n+4)} \left(\frac{4\delta^2}{\mu^2} n + \frac{4L_{\nu}^2}{(1+\nu)^2} \mu^{2\nu} n^{2+\nu} + \frac{\mu^2(A'_1)^2}{4} (n+6)^3 + \frac{(A'_2)^2}{\mu^2} n \right) + \\ &+ \frac{640(n+4)A_1}{D} \left(A_2 + \frac{\delta^2}{2A_1\mu^2} n + \frac{\delta^2}{\mu^2} n \right) + 2\mu^{2\nu} L_{\nu}^2 n^{\nu}. \end{aligned}$$

As we can see, the best achievable power of μ is 2ν , so we can choose the remaining parameters based on this. Consider the case $A_1 = \frac{L_\nu}{\mu^{1-\nu}} n^{\frac{1+\nu}{2}}$, $A_2 = 0$ and $A'_1 = \left[\frac{1}{\delta}\right]^{\frac{1-\nu}{1+\nu}} \frac{2L_\nu}{\mu^{1-\nu}}$, $A'_2 = \hat{\delta} L_\nu \mu^{1+\nu}$ with $\hat{\delta} = (n+6)^{\frac{1+\nu}{2}}$ (this is chosen to equalize powers of n in second term):

$$\begin{aligned} \min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x)\|_*^2] &\leq \frac{640(n+4)L_\nu(f_\mu(x_0) - f^*)}{ND\mu^{1-\nu}} n^{\frac{1+\nu}{2}} + \quad (22) \\ &+ \frac{D\mu^{2\nu}}{2(n+4)} \left(\frac{4\delta^2}{\mu^{2+2\nu}} n + \frac{4L_\nu^2}{(1+\nu)^2} n^{2+\nu} + L_\nu^2(n+6)^{2+\nu} + L_\nu^2 n(n+6)^{1+\nu} \right) + \\ &+ \frac{640(n+4)L_\nu}{D\mu^{1-\nu}} n^{\frac{1+\nu}{2}} \left(0 + \frac{\delta^2}{2L_\nu n^{\frac{1+\nu}{2}} \mu^{1+\nu}} n + \frac{\delta^2}{\mu^2} n \right) + 2\mu^{2\nu} L_\nu^2 n^\nu. \end{aligned}$$

Now we see only terms with $\mu^{2\nu}$ and terms with δ^2 and some powers of μ . To ease assumptions on δ we can consider maximum possible $D = 1$. The bound for δ then has form of $\delta \leq \frac{\mu^\alpha}{n^\beta}$, where $\alpha = \frac{3+\nu}{2}$ (from the third term, we have $\mu^{2\alpha-(1-\nu)-2}$ and we want it to be $\mu^{2\nu}$) and $\beta = \frac{3-\nu}{4}$ to equalize powers of n in the second ($n^{1+\nu}$) and the third ($n^{2+\frac{1+\nu}{2}-2\beta}$) terms (therefore $\delta < \frac{\mu^{\frac{3+\nu}{2}}}{n^{\frac{3-\nu}{4}}}$):

$$\begin{aligned} \min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x)\|_*^2] &\leq \frac{640(n+4)L_\nu(f_\mu(x_0) - f^*)}{N\mu^{1-\nu}} n^{\frac{1+\nu}{2}} + \\ &+ \frac{\mu^{2\nu}}{2(n+4)} \left(4\mu^{1-\nu} n^{\frac{\nu-1}{2}} + \frac{4L_\nu^2}{(1+\nu)^2} n^{2+\nu} + L_\nu^2(n+6)^{2+\nu} + L_\nu^2 n(n+6)^{1+\nu} \right) + \\ &+ 320(n+4)\mu^{2\nu} \left(\mu^{1-\nu} n^{\frac{\nu-1}{2}} + 2L_\nu n^\nu \right) + 2\mu^{2\nu} L_\nu^2 n^\nu \end{aligned}$$

(notice, that $\mu^{1-\nu} \leq 1$ because $\mu < 1$ as the step of gradient estimation, so we will replace $\mu^{1-\nu}$ with 1 further). Consider $\mu \leq \mu_0 = (M \cdot n^{1+\nu})^{-\frac{1}{2\nu}} \varepsilon_{\nabla f}^{\frac{1}{2\nu}}$ where

$$\begin{aligned} M \cdot n^{1+\nu} &= \frac{4n^{\frac{\nu-1}{2}} + \frac{4L_\nu^2}{(1+\nu)^2} n^{2+\nu} + 2L_\nu^2(n+3)(n+6)^{1+\nu}}{4(n+4)} + \\ &+ 160(n+4) \left(n^{\frac{\nu-1}{2}} + 2L_\nu n^\nu \right) + \mu^{2\nu} L_\nu^2 n^\nu \end{aligned}$$

(thus $M = O(1 + L_\nu + L_\nu^2)$) and substituting it we obtain

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x_k)\|_*^2] \leq \frac{640(n+4)n^{(1-\nu) \cdot \frac{1+\nu}{2\nu}} L_\nu(f_\mu(x_0) - f^*)}{N \cdot M^{-\frac{1-\nu}{2\nu}} \varepsilon_{\nabla f}^{\frac{2-2\nu}{2\nu}}} n^{\frac{1+\nu}{2}} + \frac{\varepsilon_{\nabla f}}{2}.$$

That means that we need to make

$$N = O \left(\frac{n^{1+(1-\nu) \cdot \frac{1+\nu}{2\nu} + \frac{1+\nu}{2}}}{\varepsilon_{\nabla f}^{\frac{1}{2\nu}}} \right) = O \left(\frac{n^{2+\frac{1-\nu}{2\nu}}}{\varepsilon_{\nabla f}^{\frac{1}{2\nu}}} \right)$$

steps to ensure $\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_{\mu}(x_k)\|_*^2] \leq \varepsilon_{\nabla f}$. It's only left to substitute μ into upper bound for δ to obtain (21). \square

In case $\nu = 1$ the article [14] (Section 7) shows that the upper bound for the expected number of steps is $N = O\left(\frac{n}{\varepsilon^2}\right)$ where $\varepsilon^2 = \varepsilon_{\nabla f}$, while we show $N = O\left(\frac{n^2}{\varepsilon_{\nabla f}}\right)$, which is n times worse. This can be improved quite easily using the fact that for this case

$$\begin{aligned} \|\nabla f_{\mu}(y) - \nabla f_{\mu}(x)\|_* &= \left\| \frac{1}{\kappa} \int_E (\nabla f(y + \mu u) - \nabla f(x + \mu u)) e^{-\frac{1}{2}\|u\|^2} du \right\|_* \leq \\ &\leq \frac{1}{\kappa} \int_E L_1 \|y - x\| e^{-\frac{1}{2}\|u\|^2} du = L_1 \|y - x\| \end{aligned}$$

then this inequality can be used to set $A_1 = \frac{L_1}{2}$ and $A_2 = 0$ in (19), so the power of n in the first term will be 1 less and repeating following steps we will obtain $N = O\left(\frac{n}{\varepsilon_{\nabla f}}\right)$. This, however, cannot be easily extended to $\nu < 1$, because of $\|x - y\|^\nu$ term (see Lemma 2 proof for details).

3.2 Convergence in the sense of $\|\nabla f_{\mu}(x_k)\|_*$

The main problem of the previous result is that it doesn't work with $\nu = 0$ (which is normal because we cannot ensure gradient norm convergence when the gradient is only bounded) and convergence becomes infinitely slow when $\nu \rightarrow 0$. We will now consider the convergence in the sense of smoothed function gradient norm while keeping functional gap (Lemma 3) small.

Theorem 2 *Consider the process (18) and Assumptions 1 and 2. Suppose we want to ensure*

$$\begin{aligned} |f_{\mu}(x) - f(x)| &\leq \frac{L_{\nu}}{1+\nu} \mu^{1+\nu} n^{\frac{1+\nu}{2}} \leq \varepsilon_f \\ \min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_{\mu}(x_k)\|_*^2] &\leq \varepsilon_{\nabla f} \end{aligned}$$

where $\varepsilon_f \sim \varepsilon_{\nabla f}^{\frac{1+\nu}{2\nu+1}}$ then it can be shown that with the right choice of the smoothing parameter μ these inequalities hold after

$$N = O\left(\frac{n^{\frac{7-3\nu}{2}}}{\frac{3-\nu}{1+\nu} \varepsilon_{\nabla f}}\right) \quad (23)$$

steps of the process (18) under the assumption that

$$\delta < \frac{\mu^{\frac{5-\nu}{2}}}{n^{\frac{3-\nu}{4}}} = O\left(\frac{\varepsilon_{\nabla f}^{\frac{5-\nu}{2(1+\nu)}}}{n^{\frac{13-3\nu}{4}}}\right). \quad (24)$$

Proof Substituting the same A_1, A_2 and A'_1, A'_2 as in previous proof into (19) we will obtain almost (22) but without the fourth term and with a smaller constant:

$$\begin{aligned} \min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x)\|_*^2] &\leq \frac{320(n+4)L_\nu(f_\mu(x_0) - f^*)}{ND\mu^{1-\nu}} n^{\frac{1+\nu}{2}} + \\ &+ \frac{D\mu^{2\nu}}{4(n+4)} \left(\frac{4\delta^2}{\mu^{2+2\nu}} n + \frac{4L_\nu^2}{(1+\nu)^2} n^{2+\nu} + L_\nu^2(n+6)^{2+\nu} + L_\nu^2 n(n+6)^{1+\nu} \right) + \\ &+ \frac{320(n+4)L_\nu}{D\mu^{1-\nu}} n^{\frac{1+\nu}{2}} \left(\frac{\delta^2}{2L_\nu n^{\frac{1+\nu}{2}} \mu^{1+\nu}} n + \frac{\delta^2}{\mu^2} n \right). \end{aligned}$$

The difference now is that we are not restricted to use $\varepsilon_{\nabla f} \sim \mu^{2\nu}$, because we can select D to balance powers of μ (there is no fourth term with its invariable $\mu^{2\nu}$). Let's at first consider a case with $\delta = 0$. Suppose that $D = \mu^\alpha$, then

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f(x)\|_*^2] \leq O\left(\mu^{-(1-\nu+\alpha)}\right) + O(\mu^{2\nu+\alpha}) + 0$$

thus $\mu^{2\nu+\alpha} \sim \varepsilon_{\nabla f}$ (because like in previous proof we want to bound the second and the third terms with the $\varepsilon_{\nabla f}/2$) and from Lemma 3 we have

$$\varepsilon_f \geq \frac{L_\nu}{1+\nu} \mu^{1+\nu} n^{\frac{1+\nu}{2}} \sim \varepsilon_{\nabla f}^{\frac{1+\nu}{2\nu+\alpha}}.$$

Now, in previous subsection we had $\mu \sim \varepsilon_{\nabla f}^{\frac{1}{2}}$ (for the case of $\nu = 1$), so substituting it into Lemma 3 we would obtain $\varepsilon_{\nabla f} \sim \varepsilon_f$. So let's just consider this to be our case, then we can obtain $\frac{1+\nu}{2\nu+\alpha} = 1$ which gives us $\alpha = 1 - \nu$ (such reasoning combines results from this and previous sections in the case of $\nu = 1$).

Now, let's set $D = \mu^{1-\nu} < 1$ and $\delta < \frac{\mu^{\frac{5-\nu}{2}}}{n^{\frac{3-\nu}{4}}}$ (the power of n is chosen similar to the previous proof) then

$$\begin{aligned} \min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_\mu(x_k)\|_*^2] &\leq \frac{320(n+4)L_\nu(f_\mu(x_0) - f^*)}{N\mu^{2-2\nu}} n^{\frac{1+\nu}{2}} + \\ &+ \frac{\mu^{1+\nu}}{4(n+4)} \left(4\mu^{3-3\nu} n^{\frac{\nu-1}{2}} + \frac{4L_\nu^2}{(1+\nu)^2} n^{2+\nu} + 2L_\nu^2(n+3)(n+6)^{1+\nu} \right) + \\ &+ 160(n+4)\mu^{1+\nu} \left(\mu^{1-\nu} n^{\frac{\nu-1}{2}} + 2L_\nu n^\nu \right). \end{aligned}$$

Consider $\mu \leq \mu_0 = (M \cdot n^{1+\nu})^{-\frac{1}{1+\nu}} \varepsilon_{\nabla f}^{\frac{1}{1+\nu}} = \frac{1}{n \cdot M^{\frac{1}{1+\nu}}} \varepsilon_{\nabla f}^{\frac{1}{1+\nu}}$ where

$$\begin{aligned} M \cdot n^{1+\nu} &= \frac{4n^{\frac{\nu-1}{2}} + \frac{4L_\nu^2}{(1+\nu)^2} n^{2+\nu} + 2L_\nu^2(n+3)(n+6)^{1+\nu}}{8(n+4)} + \\ &+ 80(n+4) \left(n^{\frac{\nu-1}{2}} + 2L_\nu n^\nu \right) \end{aligned}$$

and substituting it we obtain

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_{\mu}(x_k)\|_*^2] \leq \frac{320(n+4)n^{2-2\nu}L_{\nu}(f_{\mu}(x_0) - f^*)}{N \cdot M^{-\frac{2-2\nu}{1+\nu}} \varepsilon_{\nabla f}^{\frac{2-2\nu}{1+\nu}}} n^{\frac{1+\nu}{2}} + \frac{\varepsilon_{\nabla f}}{2}.$$

That means that we need to make

$$N = O\left(\frac{n^{1+(2-2\nu)+\frac{1+\nu}{2}}}{\varepsilon_{\nabla f}^{\frac{3-\nu}{1+\nu}}}\right) = O\left(\frac{n^{\frac{7-3\nu}{2}}}{\varepsilon_{\nabla f}^{\frac{3-\nu}{1+\nu}}}\right) \quad (25)$$

steps to ensure $\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} [\|\nabla f_{\mu}(x_k)\|_*^2] \leq \varepsilon_{\nabla f}$. Substituting $\mu = \mu_0$ into Lemma 3 we obtain

$$|f_{\mu}(x) - f(x)| \leq \frac{L_{\nu}}{1+\nu} \mu^{1+\nu} n^{\frac{1+\nu}{2}} = \Theta\left(\frac{\varepsilon_{\nabla f}}{n^{\frac{1+\nu}{2}}}\right).$$

Thus we ensure $|f_{\mu}(x) - f(x)| \leq \varepsilon_f$ with $\varepsilon_f = \Theta\left(\frac{\varepsilon_{\nabla f}}{n^{\frac{1+\nu}{2}}}\right)$. The bound (24) can be obtained the same way as in previous theorem. \square

In case $\nu = 0$ [14] shows that $N = O\left(\frac{n^3}{\varepsilon_f \varepsilon_{\nabla f}^2}\right) \stackrel{\varepsilon_f = \Theta\left(\frac{\varepsilon_{\nabla f}}{n^{1/2}}\right)}{=} O\left(\frac{n^{\frac{7}{2}}}{\varepsilon_{\nabla f}^2}\right)$ which coincides with our result. In case $\nu = 1$ this result coincides with the result of the previous theorem, and we can repeat the reasoning at the end improving the result by making the iteration complexity to be proportional to n rather than n^2 .

We didn't discuss the question of what is the weakest possible bound on δ at which it is still possible to prove the convergence. It can be easily shown that if we remove powers of n from these δ upper bounds it won't change the fact of the convergence, however this will increase the powers of n in N bounds. For example in the end of the proof of the Theorem 2 we can choose $\mu_0 = \left(M \cdot n^{\frac{5+\nu}{2}}\right)^{-\frac{1}{1+\nu}} \varepsilon_{\nabla f}^{\frac{1}{1+\nu}}$ (this is the biggest power of n there) and then repeating the steps we obtain

$$N = O\left(\frac{n^{1+\frac{5+\nu}{2} \cdot \frac{1}{1+\nu} + \frac{1+\nu}{2}}}{\varepsilon_{\nabla f}^{\frac{3-\nu}{1+\nu}}}\right).$$

Changing the powers of $\varepsilon_{\nabla f}$ for noise bounds is harder though, and can be a topic of the further studies.

4 Conclusion

In this paper we extend the results of [14] to non-convex minimization problems with Hölder-continuous gradients and noisy zeroth-order oracle. Table 1 below summarizes our results for two types of the quality measures: norm of the

gradient of the smoothed version of the objective f_μ and norm of the gradient of the original objective function f . We provide an upper bound for the necessary number of iterations N and an upper bound on the oracle inexactness δ which can be tolerated and still allows to achieve the desired accuracy in terms of the corresponding criterion. We also show that in the case $\nu = 1$, the upper bounds for N can be improved by reducing the exponent of n to 1 (second part of the Table 1). The interesting fact is that for the case of $\nu = 1$ the upper bound for the noise level δ is linear in $\varepsilon_{\nabla f}$, and bounds on N and δ for both $\|\nabla f(x_k)\|_*$ and $\|\nabla f_\mu(x_k)\|_*$ coincide.

In future it would be interesting to explore in more details the trade-off between the oracle noise level δ and the iteration number N in terms of their dependence on n , which we briefly discussed after the proofs of Theorems 1 and 2. Another interesting question for future research is whether it is possible to obtain a bound for N which continuously depends on ν and for $\nu = 1$ gives the same bound as the bound in [14].

Table 1 Convergence properties for the different convergence types

Convergence type	N upper bound	δ upper bound	$ f_\mu(x) - f(x) $	Possible ν
$\mathbb{E}\ \nabla f(x_k)\ _*^2 \leq \varepsilon_{\nabla f}$	$O\left(\frac{n^{2+\frac{1-\nu}{2\nu}}}{\varepsilon_{\nabla f}^\nu}\right)$	$O\left(\frac{\frac{3+\nu}{4\nu}}{n^{\frac{3+7\nu}{4\nu}}}\right)$	—	$\nu \in (0, 1]$
$\mathbb{E}\ \nabla f_\mu(x_k)\ _*^2 \leq \varepsilon_{\nabla f}$	$O\left(\frac{n^{\frac{7-3\nu}{2}}}{\varepsilon_{\nabla f}^{\frac{3-\nu}{1+\nu}}}\right)$	$O\left(\frac{\frac{5-\nu}{2(1+\nu)}}{n^{\frac{13-3\nu}{4}}}\right)$	$\Theta\left(\frac{\varepsilon_{\nabla f}}{n^{\frac{1+\nu}{2}}}\right)$	$\nu \in [0, 1]$
$\mathbb{E}\ \nabla f(x_k)\ _*^2 \leq \varepsilon_{\nabla f}$	$O\left(\frac{n}{\varepsilon_{\nabla f}}\right)$	$O\left(\frac{\varepsilon_{\nabla f}}{n^{5/2}}\right)$	—	$\nu = 1$
$\mathbb{E}\ \nabla f_\mu(x_k)\ _*^2 \leq \varepsilon_{\nabla f}$	$O\left(\frac{n}{\varepsilon_{\nabla f}}\right)$	$O\left(\frac{\varepsilon_{\nabla f}}{n^{5/2}}\right)$	$\Theta\left(\frac{\varepsilon_{\nabla f}}{n}\right)$	$\nu = 1$

Acknowledgements The authors are grateful to K. Scheinberg and A. Beznosikov for several discussions on derivative-free methods.

References

1. Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M.: Automatic differentiation in machine learning: a survey. arxiv:1502.05767 (2018)
2. Berahas, A.S., Cao, L., Choromanski, K., Scheinberg, K.: A theoretical and empirical comparison of gradient approximations in derivative-free optimization. arxiv:1905.01332 (2019)
3. Berahas, A.S., Cao, L., Scheinberg, K.: Global convergence rate analysis of a generic line search algorithm with noise. arxiv:1910.04055 (2019)
4. Bolte, J., Glaudin, L., Pauwels, E., Serrurier, M.: A hölderian backtracking method for min-max and min-min problems. arxiv:2007.08810 (2020)
5. Brent, R.: Algorithms for Minimization Without Derivatives. Dover Books on Mathematics. Dover Publications (1973)

6. 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
7. Dvurechensky, P.: Gradient method with inexact oracle for composite non-convex optimization. arxiv:1703.09180 (2017)
8. Fabian, V.: Stochastic approximation of minima with improved asymptotic speed. Ann. Math. Statist. **38**(1), 191–200 (1967). DOI 10.1214/aoms/1177699070
9. Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization **23**(4), 2341–2368 (2013). DOI 10.1137/120880811. URL <https://doi.org/10.1137/120880811>
10. Kim, K., Nesterov, Y., Skokov, V., Cherkasskii, B.: Effektivnii algoritm vychisleniya proizvodnyh i ekstremalnye zadachi (efficient algorithm for calculation of derivatives and extreme problems). Ekonomika i matematicheskie metody **20**(2), 309–318 (1984)
11. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numerica **28**, 287–404 (2019). DOI 10.1017/S0962492919000060
12. Liu, S., Kaikhura, B., Chen, P.Y., Ting, P., Chang, S., Amini, L.: Zeroth-order stochastic variance reduction for nonconvex optimization. Advances in Neural Information Processing Systems **31**, 3727–3737 (2018)
13. Nesterov, Y.: Universal gradient methods for convex optimization problems. Mathematical Programming **152**(1), 381–404 (2015). DOI 10.1007/s10107-014-0790-0. URL <https://doi.org/10.1007/s10107-014-0790-0>
14. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Foundations of Computational Mathematics **17**(2), 527–566 (2015). DOI 10.1007/s10208-015-9296-2
15. Rosenbrock, H.H.: An automatic method for finding the greatest or least value of a function. The Computer Journal **3**(3), 175–184 (1960). DOI 10.1093/comjnl/3.3.175
16. Saeed Ghadimi, G.L., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Mathematical Programming **155** (2013). DOI 10.1007/s10107-014-0846-1
17. Spall, J.C.: Introduction to Stochastic Search and Optimization, 1 edn. John Wiley & Sons, Inc., New York, NY, USA (2003)
18. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
19. Wang, J., Liu, Y., Li, B.: Reinforcement learning with perturbed rewards. Proceedings of the AAAI Conference on Artificial Intelligence **34**, 6202–6209 (2020). DOI 10.1609/aaai.v34i04.6086

A Appendix

A.1 Proofs of Lemmas 2.1 —2.5

Proof (Lemma 1) From (1) we get $\|Bu\|_*^2 = \langle Bu, B^{-1}Bu \rangle = \langle Bu, u \rangle = \|u\|^2$. Using this and Lemma 7 we obtain

$$\begin{aligned}
& \|\nabla \tilde{f}_\mu(x, \delta) - \nabla f_\mu(x)\|_* \stackrel{(14)}{=} \\
& = \left\| \frac{1}{\kappa} \int_E \frac{\tilde{f}(x + \mu u, \delta) \pm f(x + \mu u)}{\mu} e^{-\frac{1}{2}\|u\|^2} B u du - \nabla f_\mu(x) \right\|_* \leq \\
& \leq \left\| \frac{1}{\kappa} \int_E \left(\frac{\tilde{f}(x + \mu u, \delta) - f(x + \mu u)}{\mu} \right) e^{-\frac{1}{2}\|u\|^2} B u du \right\|_* + \\
& + \left\| \frac{1}{\kappa} \int_E \frac{f(x + \mu u)}{\mu} e^{-\frac{1}{2}\|u\|^2} B u du - \nabla f_\mu(x) \right\|_* \stackrel{A.s.m. 1, (9)}{\leq} \\
& \leq \frac{1}{\kappa} \int_E \frac{\delta}{\mu} \|u\| e^{-\frac{1}{2}\|u\|^2} du + \|\nabla f_\mu(x) - \nabla f(x)\|_* \stackrel{Lem. 7}{\leq} \frac{\delta}{\mu} n^{1/2}
\end{aligned}$$

and

$$\begin{aligned}
& \|\nabla f_\mu(x) - \nabla f(x)\|_* \stackrel{(11)}{=} \left\| \frac{1}{\kappa} \int_E (\nabla f(x + \mu u) - \nabla f(x)) e^{-\frac{1}{2}\|u\|^2} du \right\|_* \stackrel{A.s.m. 2}{\leq} \\
& \leq \frac{1}{\kappa} \int_E L_\nu \|\mu u\|^\nu e^{-\frac{1}{2}\|u\|^2} du \stackrel{Lem. 7}{\leq} \mu^\nu L_\nu n^{\nu/2}
\end{aligned}$$

thus, finally

$$\begin{aligned}
& \|\nabla \tilde{f}_\mu(x, \delta) - \nabla f(x)\|_* \leq \|\nabla \tilde{f}_\mu(x, \delta) - \nabla f_\mu(x)\|_* + \|\nabla f_\mu(x) - \nabla f(x)\|_* \leq \\
& \leq \frac{\delta}{\mu} n^{1/2} + \mu^\nu L_\nu n^{\nu/2}. \quad \square
\end{aligned}$$

Proof (Lemma 2)

$$\begin{aligned}
& \|\nabla f_\mu(y) - \nabla f_\mu(x)\|_* \stackrel{(8)}{=} \\
& = \frac{1}{\kappa} \left\| \int_E \left(\frac{f(y + \mu u) - f(y)}{\mu} - \frac{f(x + \mu u) - f(x)}{\mu} \right) B u e^{-\frac{1}{2}\|u\|^2} du \right\|_* \leq \\
& \leq \frac{1}{\mu \kappa} \int_E \left| \int_0^1 \langle \nabla f(\mu u + ty + (1-t)x) - \nabla f(ty + (1-t)x), y - x \rangle dt \right| \|u\| e^{-\frac{1}{2}\|u\|^2} du \stackrel{A.s.m. 2}{\leq} \\
& \leq \frac{1}{\mu \kappa} \int_E L_\nu \mu^\nu \|y - x\| \|u\|^{1+\nu} e^{-\frac{1}{2}\|u\|^2} du \stackrel{Lem. 7}{\leq} \frac{L_\nu}{\mu^{1-\nu}} n^{\frac{1+\nu}{2}} \|y - x\|.
\end{aligned}$$

Integrating this we obtain

$$f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle \leq \frac{L_\nu}{2\mu^{1-\nu}} n^{\frac{1+\nu}{2}} \|y - x\|^2 \quad (26)$$

so using this way we proved lemma with $A_1 = \frac{L_\nu}{\mu^{1-\nu}} n^{\frac{1+\nu}{2}}$ and $A_2 = 0$.

The other way to obtain A_1 and A_2 is to directly upper bound $f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle$ applying Lemma 8:

$$\begin{aligned} f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle &\stackrel{(6,11)}{=} \\ &= \frac{1}{\kappa} \int_E (f(y + \mu u) - f(x + \mu u) - \langle \nabla f(x + \mu u), y - x \rangle) e^{-\frac{1}{2}\|u\|^2} du \stackrel{A_{sm.} 2}{\leq} \\ &\leq \frac{L_\nu}{1+\nu} \|y - x\|^{1+\nu} \stackrel{Lem. 8}{\leq} \frac{1}{2} \left[\frac{1-\nu}{1+\nu} \frac{2}{\tilde{\delta}} \right]^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} \|y - x\|^2 + \tilde{\delta}. \end{aligned}$$

Setting $\tilde{\delta} = \hat{\delta} \mu^{1+\nu} L_\nu$ and using upper bound $\left[2 \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} \leq 2$ we obtain

$$f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle \leq \left[\frac{1}{\hat{\delta}} \right]^{\frac{1-\nu}{1+\nu}} \frac{L_\nu}{\mu^{1-\nu}} \|y - x\|^2 + \hat{\delta} L_\nu \mu^{1+\nu} \quad (27)$$

so we proved lemma with $A_1 = \left[\frac{1}{\hat{\delta}} \right]^{\frac{1-\nu}{1+\nu}} \frac{2L_\nu}{\mu^{1-\nu}}$ and $A_2 = \hat{\delta} L_\nu \mu^{1+\nu}$.

Proof (Lemma 3) To proof this we should notice that

$$\frac{1}{\kappa} \int_E \langle \nabla f(x), u \rangle e^{-\frac{1}{2}\|u\|^2} du = 0$$

thus

$$\begin{aligned} |f_\mu(x) - f(x)| &\stackrel{(6)}{=} \left| \int_E (f(x + \mu u) - f(x)) e^{-\frac{1}{2}\|u\|^2} du \right| = \\ &= \left| \int_E (f(x + \mu u) - f(x) - \langle \nabla f(x), \mu u \rangle) e^{-\frac{1}{2}\|u\|^2} du \right| \stackrel{A_{sm.} 2}{\leq} \\ &\leq \frac{L_\nu}{1+\nu} \mu^{1+\nu} \int_E \|u\|^{1+\nu} e^{-\frac{1}{2}\|u\|^2} du \stackrel{Lem. 7}{\leq} \frac{L_\nu}{1+\nu} \mu^{1+\nu} n^{\frac{1+\nu}{2}}. \quad \square \end{aligned}$$

Proof (Lemma 4) From the fact that $a^2 \leq 2(a+b)^2 + 2b^2$:

$$\begin{aligned} \|\nabla f(x)\|_*^2 &\leq 2\|\nabla f_\mu(x)\|_*^2 + 2\|\nabla f(x) - \nabla f_\mu(x)\|_*^2 \stackrel{Lem. 1}{\leq} \\ &\leq 2\|\nabla f_\mu(x)\|_*^2 + 2\mu^{2\nu} L_\nu^2 n^{2\nu}. \quad \square \end{aligned}$$

Proof (Lemma 5)

$$\mathbb{E}_u [\|g_\mu(x, u, \delta)\|_*^2] = \frac{1}{\kappa} \int_E \left| \frac{\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta)}{\mu} \right|^2 \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du$$

let's bound $|\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta)|$:

$$\begin{aligned} |\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta)| &\leq 2\delta + |f(x + \mu u) - f(x)| \leq \\ &\leq 2\delta + |f(x + \mu u) - f_\mu(x + \mu u) - f(x) + f_\mu(x)| + |f_\mu(x + \mu u) - f_\mu(x)| \stackrel{Lem. 3}{\leq} \\ &\leq 2\delta + \frac{2L_\nu}{1+\nu} \mu^{1+\nu} n^{\frac{1+\nu}{2}} + |f_\mu(x + \mu u) - f_\mu(x) - \langle \nabla f_\mu(x), \mu u \rangle| + \\ &+ |\langle \nabla f_\mu(x), \mu u \rangle| \stackrel{Lem. 2}{\leq} \\ &\leq 2\delta + \frac{2L_\nu}{1+\nu} \mu^{1+\nu} n^{\frac{1+\nu}{2}} + \frac{\mu^2 A_1}{2} \|u\|^2 + A_2 + |\langle \nabla f_\mu(x), \mu u \rangle| \end{aligned}$$

thus from the fact that $\left(\sum_{i=1}^k a_i\right)^2 \leq k \left(\sum_{i=1}^k a_i^2\right)$

$$\begin{aligned} & |\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta)|^2 \leq \\ & \leq 5 \left(4\delta^2 + \frac{4L_\nu^2}{(1+\nu)^2} \mu^{2+2\nu} n^{1+\nu} + \frac{\mu^4 A_1^2}{4} \|u\|^4 + A_2^2 + \langle \nabla f_\mu(x), \mu u \rangle^2 \right) \end{aligned}$$

and applying Theorem 3 we finally obtain

$$\begin{aligned} \mathbb{E}_u [\|g_\mu(x, u, \delta)\|_*^2] & \leq 20(n+4) \|\nabla f_\mu(x)\|^2 + \\ & + 5 \left(\frac{4\delta^2}{\mu^2} n + \frac{4L_\nu^2}{(1+\nu)^2} \mu^{2\nu} n^{2+\nu} + \frac{\mu^2 A_1^2}{4} (n+6)^3 + \frac{A_2^2}{\mu^2} n \right) \quad \square \end{aligned}$$

A.2 External results

Lemma 7 (Lemma 1 from [14]) For $p \geq 0$, we have

$$\frac{1}{\kappa} \int_E \|u\|^p e^{-\frac{1}{2}\|u\|^2} du \leq \begin{cases} n^{p/2}, & p \in [0, 2] \\ (n+p)^{p/2}, & p > 2 \end{cases}$$

Lemma 8 (Lemma 2 from [13]) Let the function f satisfy Assumption 2. Then for any $\tilde{\delta} > 0$

$$\frac{L_\nu}{1+\nu} t^{1+\nu} \leq \frac{1}{2} \left[\frac{1-\nu}{1+\nu} \frac{2}{\tilde{\delta}} \right]^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} t^2 + \tilde{\delta} = \frac{L}{2} t^2 + \tilde{\delta}$$

Theorem 3 (Theorem 3 from [14]) If f is differentiable at x and u is a standard random normal vector, then

$$\mathbb{E}_u [\langle \nabla f(x), u \rangle^2 \|u\|^2] \leq (n+4) \|\nabla f(x)\|_*^2$$