Zeroth-order methods for noisy Hölder-gradient functions

Innokentiy Shibaev
Moscow Institute of Physics and Technology
Dolgoprudnyi 141700, Moscow Oblast, Russia
shibaev.kesha@gmail.com

Pavel Dvurechensky
Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstr. 39, 10117 Berlin, Germany
pavel.dvurechensky@gmail.com

Alexander Gasnikov
Moscow Institute of Physics and Technology
Dolgoprudnyi 141700, Moscow Oblast, Russia
gasnikov@yandex.ru

Abstract
In this paper, we prove new complexity bounds for zeroth-order methods of non-convex optimization. We use the Gaussian smoothing approach from [4] and extend their results on zeroth-order non-convex optimization to Hölder-gradient functions with noisy zeroth-order oracle, obtaining noise upper-bounds as well. We consider gradient estimation based on normally distributed random Gaussian vectors and prove \( \delta < \frac{\mu - \nu}{n^{\frac{3 + \nu}{2}}} = O \left( \frac{\varepsilon \sqrt{f}}{n^{\frac{3 + \nu}{2}}} \right) \) (where \( \mu \) is a smoothing parameter, \( n \) is a dimension and \( \nu \) is an exponent of Hölder condition and \( \varepsilon \sqrt{f} \) is the desired squared gradient norm) bound for noise for which it is still possible to prove convergence to stationary point of smoothed function. We also consider convergence to stationary point of the initial (not smoothed) function and prove \( \delta < \frac{\mu}{n^{\frac{3 + \nu}{2}}} = O \left( \frac{\varepsilon \sqrt{f}}{n^{\frac{3 + \nu}{2}}} \right) \) noise bound for it for the case \( \nu > 0 \).

1 Introduction
The main advantage of zeroth-order (derivative-free) optimization methods is that computing function value is always simpler than computing its gradient vector. However such methods generally have lower convergence rates, and with inventing of Fast Differentiation, which allows computing gradient vector with at most four times bigger computation complexity they became impractical. Nevertheless, they are few other choices for black-box types of problems, where we have no access to function derivatives (there are other methods, e.g. genetic algorithms [2]). This became more common with Reinforcement Learning problems (basically function optimization can be considered as a continuous multi-armed bandit problem) where rewards functions commonly have a black-box structure and the only thing we can get is a function value (see [5] for examples). The problem became even worse when we are dealing with computer simulation of some physical processes (satellite movement for example) - such models often deal with some noise in their outputs.
so it is important to prove convergence of optimization process in conditions of noisy function values. Sometimes it can be solved quite easily using batch averaging, but this implies that the noise is unbiased, which is not always holds (for RL examples see [6]).

In [4], the authors used the Gaussian smoothing technique to prove convergence to a stationary point of a smoothed function. Smoothing allowed them to consider function with better properties, which is exactly what we want here (because adding noise can make function lose all of it good properties, such as smoothness, convexity, etc.).

In this work, we extend their results to Hölder-gradient functions with noisy zeroth-order oracle. We mostly follow notation of [1] where the same problem (but for the case of the first-order noisy oracle) is considered.

The rest of the paper is organized as follows. The first section contains mostly definitions and some lemmas that mostly are extended variants of those in [4]. In the second section, we consider a simple gradient descent process with Gaussian gradient estimation (like that in [4]) and obtain complexity bounds in terms of smoothed and non-smoothed function gradient norm as well as noise upper bounds.

2 Gaussian smoothing, zeroth-order oracle

2.1 Definitions

We start with some definitions (from [4]). For a finite-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

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

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

**Definition 2.1.** We say that a function $f(x)$ is equipped with an inexact zeroth-order oracle on a set $X$ if there exists $\delta > 0$ and one can calculate $\hat{f}(x, \delta) \in \mathbb{R}$ and satisfying

$$| f(x) - \hat{f}(x, \delta) | \leq \delta$$

(2)

**Definition 2.2.** We say that a differentiable function $f(x)$ has Hölder-continuous gradient with some $\nu \in [0, 1]$ and $L_\nu > 0$ if

$$\| \nabla f(y) - \nabla f(x) \|_* \leq L_\nu \| y - x \|^\nu, \forall x, y \in X$$

(3)

If $f(x)$ has Hölder-continuous gradient with some $\nu, L_\nu$ then

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

(4)

Now, let’s consider Gaussian approximation of $f(x)$ from [4]:

**Definition 2.3.** Consider a function $f : E \to \mathbb{R}$. Its Gaussian approximation $f_\mu(x)$ is defined as

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

(5)

where

$$\kappa \overset{\text{def}}{=} \int_E e^{-\frac{1}{2} \| u \|^2} du = \frac{(2\pi)^{n/2}}{|\det B|^{1/2}},$$

(6)

then (see [4] Section 2 for details)

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

(7)

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

(8)

the last is true when $f(x)$ is differentiable at $x$. 

2
2.2 Basic results

As shown in Lemma 3 in [4] for $f(x)$ with Lipschitz-continuous gradient
\[ \| \nabla f_\mu (x) - \nabla f (x) \|_* \leq \frac{\mu L}{2} (n + 3)^{3/2} \]

We can extend this lemma for our case:

**Lemma 2.1.** Consider differentiable $f(x)$ with Hölder-continuous gradient with some $\nu \in [0, 1]$ and $L_\nu \geq 0$, and assume that $f(x)$ equipped with inexact zeroth-order oracle $\tilde{f}(x, \delta)$, $\delta > 0$. Let’s define
\[ \nabla \tilde{f}_\mu (x, \delta) = \frac{1}{\kappa} \int_{E} \left( \frac{\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta)}{\mu} - \frac{1}{2} \| u \|^2 \right) B u d u \]

then it can be shown that
\[ \| \nabla \tilde{f}_\mu (x, \delta) - \nabla f_\mu (x) \|_* \leq \frac{2\nu}{\mu} n^{1/2} \]
\[ \| \nabla f_\mu (x) - \nabla f (x) \|_* \leq \frac{\mu L_\nu}{1 + \nu} (n + 2 + \nu)^{1 + \frac{1}{2} \nu} \]

and, consequently
\[ \| \nabla \tilde{f}_\mu (x, \delta) - \nabla f (x) \|_* \leq \frac{2\nu}{\mu} n^{1/2} + \frac{\mu L_\nu}{1 + \nu} (n + 2 + \nu)^{1 + \frac{1}{2} \nu} \]

**Proof.** Can be found in Appendix

It can be shown, that (assuming $f$ is Lipschitz-continuous with $L_0$) that $f_\mu$ has Hölder-continuous gradient with $\nu = 1$ and $L = \frac{2^{3/2}}{\mu L_0}$ (Lemma 2 from [4]). Thus we can get
\[ |f_\mu (y) - f_\mu (x) - \langle \nabla f_\mu (x), y - x \rangle| \leq \frac{L}{2} \| x - y \|^2 \]

If the set $X$ is bounded with $\text{diam}(X) \leq R$ one can proof that this is true for $f_\mu (x)$ with Hölder-continuous gradient with $L = \frac{n^{1/2} \mu L_\nu}{\mu} L_\nu$ (using the fact that)
\[ \| \nabla f (y) - \nabla f (x) \|_* \leq L_\nu \| x - y \|^\nu \leq L_\nu R^\nu \Rightarrow |f_\mu (y) - f_\mu (x)| \leq L_\nu R^\nu \| x - y \| \]

yet without that assumption we still can obtain reasonable upper bound

**Lemma 2.2.** Consider differentiable $f(x)$ with Hölder-continuous gradient with some $\nu \in [0, 1]$ and $L_\nu \geq 0$ then it can be shown that
\[ |f_\mu (y) - f_\mu (x) - \langle \nabla f_\mu (x), y - x \rangle| \leq \frac{A_1}{2} \| y - x \|^2 + A_2 \]
with various possible $A_1, A_2$, specifically
\[ A_1 = \frac{L_\nu}{\mu^{1-\nu}} n^{\frac{1+\nu}{2}}, A_2 = 0 \]
and
\[ A_1 = \left[ \frac{1}{\delta} \right]^{\frac{1+\nu}{2}} \frac{2L_\nu}{\mu^{1-\nu}}, A_2 = \hat{d} L_\nu \mu^{1+\nu} \text{ where } \hat{d} > 0 \]

**Proof.** Can be found in Appendix

One of the best properties of smoothed function $f_\mu (x)$ is that $|f_\mu (x) - f (x)|$ is small, for example when $f$ is Lipschitz-continuous with $L_0$ it can be shown (see Theorem 1 from [4]) that
\[ |f_\mu (x) - f (x)| \leq \mu L_0 n^{1/2} \]
Lemma 2.3. Consider differentiable \( f(x) \) with Hölder-continuous gradient with some \( \nu \in [0, 1] \) and \( L_\nu \geq 0 \) then 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 \( \Box \)

From Lemma 2.1 we can obtain an upper bound for gradient norm

Lemma 2.4. Consider differentiable \( f(x) \) with Hölder-continuous gradient with some \( \nu \in [0, 1] \) and \( L_\nu \geq 0 \) then it can be shown that

\[
\|\nabla f(x)\|^2 \leq 2\|\nabla f_\mu(x)\|^2 + \frac{2\mu^{2\nu}}{(1+\nu)^2}L_\nu^2(n+2+\nu)^{2+\nu}
\]

Proof. Can be found in Appendix \( \Box \)

3 Convergence rate analysis

3.1 Obtaining initial upper-bound

Now consider \( \nu \in [0, 1] \). From Lemma 2.2 and the fact that \( ab \leq \frac{C_\mu^2}{\mu} + \frac{\mu^2}{2} \) where \( C > 0 \), \( a = \|y-x\| \) and \( b = \frac{2\delta}{\mu}n^{1/2} \) we obtain

\[
|f_\mu(y) - f_\mu(x) - (\nabla f_\mu(x, \delta), y-x)| \leq
\]

\[
\leq |f_\mu(y) - f_\mu(x) - (\nabla f_\mu(x), y-x)| + |(\nabla f_\mu(x, \delta) - \nabla f_\mu(x), y-x)| \leq
\]

\[
\leq \frac{A_1}{2}||y-x||^2 + A_2 + \frac{2\delta}{\mu}n^{1/2}||y-x|| \leq
\]

\[
\leq \left( \frac{A_1}{2} + C \right)||y-x||^2 + A_2 + \frac{4\delta^2}{2C\mu^2}n \]

Consider a gradient descent process \( x_{k+1} = x_k - h_kB^{-1}g_\mu(x_k, u_k, \delta) \) where \( u_k \) is normal random vector and \( g_\mu(x_k, u_k, \delta) = f_{\mu(x_k+\mu u_k, \delta)} - f_{\mu(x_k, \delta)} \) \( B \). Taking the expectation in \( u_k \) we obtain

\[
E_{u_k}f_\mu(x_{k+1}) \leq f_\mu(x_k) - h_k\|\nabla \tilde{f}_\mu(x_k, \delta)\|^2 + h_k^2A_1E_{u_k}\|g_\mu(x_k, u_k, \delta)\|^2 +
\]

\[
+ \left( A_2 + \frac{2\delta^2}{A_1\mu^2}n \right)
\]

now let’s use the fact that (from \( (a+b)^2 \leq 2a^2 + 2b^2 \))

\[
\|\nabla f_\mu(x)\|^2 \leq 2\|\nabla \tilde{f}_\mu(x, \delta)\|^2 + 2\|\nabla f_\mu(x) - \nabla \tilde{f}_\mu(x, \delta)\|^2 \leq
\]

\[
\leq 2\|\nabla \tilde{f}_\mu(x, \delta)\|^2 + 2\cdot \frac{4\delta^2}{\mu^2}n
\]

thus

\[
E_{u_k}f_\mu(x_{k+1}) \leq f_\mu(x_k) - \frac{h_k}{2}\|\nabla f_\mu(x_k)\|^2 + h_k^2A_1E_{u_k}\|g_\mu(x_k, u_k, \delta)\|^2 +
\]

\[
+ \left( A_2 + \frac{2\delta^2}{A_1\mu^2}n + \frac{4\delta^2}{\mu^2}n \right)
\]

It’s only left to bound \( E_{u_k}\|g_\mu(x_k, u_k, \delta)\|^2 \):

\[
E_{u_k}\|g_\mu(x_k, u_k, \delta)\|^2 = \frac{1}{\kappa} \int_E \left| \frac{\tilde{f}(x_k + \mu u_k, \delta) - \tilde{f}(x_k, \delta)}{\mu} \right|^2 \|u_k\|^2 e^{-\frac{u_k^2}{2}} du
\]
let’s bound $|\hat{f}(x + \mu, \delta) - \hat{f}(x, \delta)|$:

$$
|\hat{f}(x + \mu, \delta) - \hat{f}(x, \delta)| \leq 2\delta + |f(x + \mu) - f(x)| \leq 2\delta + |f(x + \mu) - f_\mu(x) + f_\mu(x) - f(x)| + |f_\mu(x + \mu) - f_\mu(x)|
$$

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

$$
\leq 2\delta + \frac{2L_\nu}{1 + \nu} \mu^{1+\nu} n^{1+\nu} + \frac{\mu^2 A_1^2}{2} \|u\|^2 + A_2^2 + |\langle \nabla f_\mu(x), \mu u \rangle|
$$

(there $A_1^2$ and $A_2^2$ are the second pair of constants, and can be chosen independently from $A_1, A_2$)

thus

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

and using Theorem A.3 we finally obtain

$$
\mathbb{E}_{\mu_k} \|g_\mu(x_k, u_k, \delta)\|^2 \leq 20(n + 4) \|\nabla f_\mu(x_k)\|^2 + 5 \left( 4\delta^2 + \frac{4L_\nu^2}{(1 + \nu)^2} \mu^{2+2\nu} n^{1+\nu} + \frac{\mu^4 A_1^2}{4} \|u\|^4 + (A_2^2)^2 + \frac{4\delta^2}{\mu^2 n} \right)
$$

substituting this

$$
\mathbb{E}_{\mu_k} f_\mu(x_{k+1}) \leq f_\mu(x_k) - \left( \frac{h_k}{2} - 20(n + 4) h_k^2 A_1 \right) \|\nabla f_\mu(x_k)\|^2 + h_k^2 A_1 5 \left( 4\delta^2 + \frac{4L_\nu^2}{(1 + \nu)^2} \mu^{2+2\nu} n^{1+\nu} + \frac{\mu^4 A_1^2}{4} \|u\|^4 + (A_2^2)^2 + \frac{4\delta^2}{\mu^2 n} \right) + \left( A_2 + \frac{2\delta^2}{A_1 \mu^2 n} + \frac{4\delta^2}{\mu^2 n} \right)
$$

Let’s choose $h = h_k = \frac{D}{320(n + 4) A_1}$ where $D \in (0, 1]$ then

$$
\mathbb{E}_{\mu_k} f_\mu(x_{k+1}) \leq f_\mu(x_k) - \frac{D}{320(n + 4) A_1} \|\nabla f_\mu(x_k)\|^2 + \frac{5D^2}{A_1 (80(n + 4))^2} \left( 4\delta^2 \mu^2 n + \frac{4L_\nu^2}{(1 + \nu)^2} \mu^{2+2\nu} n^{1+\nu} + \frac{\mu^4 A_1^2}{4} \|u\|^4 + (A_2^2)^2 + \frac{4\delta^2}{\mu^2 n} \right) + \left( A_2 + \frac{2\delta^2}{A_1 \mu^2 n} + \frac{4\delta^2}{\mu^2 n} \right)
$$

and after summing and taking expectations in $\mathcal{U}$ (where $\mathcal{U} = \{u_0, ..., u_{N-1}\}$ is a random vector composed by i.i.d. $\{u_k\}_{k=0}^{N-1}$) it becomes

$$
\mathbb{E}_{\mathcal{U}} f_\mu(x_N) \leq f_\mu(x_0) - \frac{D}{320(n + 4) A_1} \sum_{k=0}^{N-1} \mathbb{E}_{\mathcal{U}} \|\nabla f_\mu(x_k)\|^2 + \frac{5ND^2}{A_1 (80(n + 4))^2} \left( 4\delta^2 \mu^2 n + \frac{4L_\nu^2}{(1 + \nu)^2} \mu^{2+2\nu} n^{1+\nu} + \frac{\mu^4 A_1^2}{4} \|u\|^4 + (A_2^2)^2 + \frac{4\delta^2}{\mu^2 n} \right) + \left( A_2 + \frac{2\delta^2}{A_1 \mu^2 n} + \frac{4\delta^2}{\mu^2 n} \right)
$$

thus (using the fact that $f^* \leq \mathbb{E}_{\mathcal{U}} f_\mu(x_N)$)

$$
\min_{k \in \{0, N-1\}} \mathbb{E}_{\mathcal{U}} \|\nabla f_\mu(x_k)\|^2 \leq \frac{320(n + 4) A_1 (f_\mu(x_0) - f^*)}{ND} + \frac{D}{4(n + 4)} \left( 4\delta^2 \mu^2 n + \frac{4L_\nu^2}{(1 + \nu)^2} \mu^{2+2\nu} n^{1+\nu} + \frac{\mu^4 A_1^2}{4} \|u\|^4 + (A_2^2)^2 + \frac{4\delta^2}{\mu^2 n} \right) + \frac{320(n + 4) A_1}{D} \left( A_2 + \frac{2\delta^2}{A_1 \mu^2 n} + \frac{4\delta^2}{\mu^2 n} \right)
$$

(12)
3.2 Convergence in the sense of $\|\nabla f(x_k)\|_*$

Suppose we want to ensure

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\xi_k} \|\nabla f(x_k)\|^2 \leq \varepsilon \nabla f$$

Using Lemma 2.4 we can obtain a bound for $\min_{k \in \{0, N-1\}} \mathbb{E}_{\xi_k} \|\nabla f(x)\|^2$ from (12):

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\xi_k} \|\nabla f(x)\|^2 \leq \frac{2\mu^{2\nu}}{(1+\nu)^2} L^2_{\nu}(n+2+\nu)^{2+\nu} \leq \frac{320(n+4)A_1(f_\mu(x_0) - f^*)}{ND} +$$

$$+ \frac{D}{4(n+4)} \left( \frac{4\delta^2}{\mu^2} n + \frac{4L^2_{\nu}}{(1+\nu)^2} \mu^{2\nu} n^{2+\nu} + \frac{\mu^2(A_1')^2}{4} (n+6)^3 + \left( \frac{A_1'}{\mu^2} n \right) \right) +$$

$$+ \frac{320(n+4)A_1}{D} \left( A_2 + \frac{2\delta^2}{A_1} \mu^{2\nu} n + \frac{4\delta^2}{\mu^2} \right) + \frac{2\mu^{2\nu}}{(1+\nu)^2} L^2_{\nu}(n+2+\nu)^{2+\nu}$$

As we can see, the best achievable power of $\mu$ is $2\nu$, so we can choose the remaining parameters based on this. Consider the case $A_1 = \frac{1}{\mu^{1+\nu}} n^{1+\nu}$, $A_2 = 0$ and $A_1' = \left[ \frac{1}{8} \right] \frac{2\delta^2}{\mu^2}$. $A'_2 = \delta L_{\nu} \mu^{1+\nu}$

with $\delta = (n+6)^{1+\nu}$ (this is chosen to equalize powers of $n$ in second term):

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\xi_k} \|\nabla f(x)\|^2 \leq \frac{320(n+4)L^2_{\nu}(f_\mu(x_0) - f^*)}{ND\mu^{1-\nu}} +$$

$$+ \frac{D\mu^{2\nu}}{4(n+4)} \left( \frac{4\delta^2}{\mu^{2+2\nu}} n + \frac{4L^2_{\nu}}{(1+\nu)^2} \mu^{2\nu} n^{2+\nu} + L^2_{\nu}(n+6)^{2+\nu} + L^2_{\nu}(n+6)^{1+\nu} \right) +$$

$$+ \frac{320(n+4)L^2_{\nu}}{D\mu^{1-\nu}} \left( 0 + \frac{2\delta^2}{L_{\nu} \mu^{1+\nu} \mu^{2\nu} n} + \frac{4\delta^2}{\mu^2} \right) +$$

$$+ \frac{2\mu^{2\nu}}{(1+\nu)^2} L^2_{\nu}(n+2+\nu)^{2+\nu}$$

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

$$\min_{k \in \{0, N-1\}} \mathbb{E}_{\xi_k} \|\nabla f(x)\|^2 \leq \frac{320(n+4)L^2_{\nu}(f_\mu(x_0) - f^*)}{N\mu^{1-\nu}} n^{1+\nu} +$$

$$+ \frac{\mu^{2\nu}}{4(n+4)} \left( 4\mu^{1-\nu} n^{1+\nu} + \frac{4L^2_{\nu}}{(1+\nu)^2} n^{2+\nu} + L^2_{\nu}(n+6)^{2+\nu} + L^2_{\nu}(n+6)^{1+\nu} \right) +$$

$$+ 320(n+4)\mu^{2\nu} (2\mu^{1-\nu} n^{1+\nu} + 4L_{\nu} n^{1+\nu}) + \frac{2\mu^{2\nu}}{(1+\nu)^2} L^2_{\nu}(n+2+\nu)^{2+\nu}$$

(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 $\mu \leq \mu_0 = (M \cdot n^{2+\nu})^{-\frac{1}{2\nu}} \varepsilon \sqrt{\nu}$

$$M \cdot n^{2+\nu} = \frac{4n^{1+\nu} + \frac{4L^2_{\nu}}{(1+\nu)^2} n^{2+\nu} + 2L^2_{\nu}(n+3)(n+6)^{1+\nu}}{8(n+4)} +$$

$$+ 160(n+4) \left( 2n^{1+\nu} + 4L_{\nu} n^{1+\nu} \right) + \frac{L^2_{\nu}(n+2+\nu)^{2+\nu}}{(1+\nu)^2}$$
we have

so we need to make

steps to ensure

In case \( \nu = 1 \) the article \([4]\) (Section 7) shows that the upper bound for the expected number of steps is \( N = O \left( \frac{n}{\epsilon^2} \right) \) when \( \epsilon^2 = \nu \varepsilon_f \), while we show \( N = O \left( \frac{n^2}{\nu \epsilon_f} \right) \), which is \( n \) times worse. This can be improved quite easily using the fact that for the case \( \nu = 1 \)

then this inequality can be used to set \( A_1 = \frac{L}{\nu} \) and \( A_2 = 0 \) in (12), so the power of \( n \) in the first term will be 1 less and repeating following steps we will obtain \( N = O \left( \frac{n}{\nu \epsilon_f} \right) \). This, however, cannot be easily extended to \( \nu < 1 \), because of \( \|x - y\|^{\nu} \) term (see Lemma 2.2 proof for details).

### 3.3 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 \( \nu \to 0 \). We will now consider the convergence in the sense of smoothed function gradient norm while keeping functional gap (Lemma 2.3) small:

Substituting the same \( A_1, A_2 \) and \( A_1', A_2' \) into (12) we will obtain (13) but without the fourth term:

The difference now is that we are not bound to use \( \varepsilon_f \sim L \varepsilon_f \), because we can select \( D \) to balance powers of \( \mu \) (there is no fourth term with its invariable \( \mu^{2\epsilon} \)). Let’s at first consider a case with \( \delta = 0 \). Suppose that \( D = \mu^{\epsilon} \), then

thus \( \mu^{2\epsilon + \alpha} \sim \varepsilon_f \) and from Lemma 2.3 we have

Now, in previous subsection we had \( \mu \sim \frac{\varepsilon_f^{1/\nu}}{\nu} \) (for the case of \( \nu = 1 \)), so substituting it into Lemma 2.3 we would obtain \( \varepsilon_f \sim \varepsilon_f^{1/\nu} \). So let’s just consider this to be our case, then we can obtain

(Thus \( M = O(1 + L + L^2) \)) and substituting it we obtain

Substituting the same

Because we cannot ensure gradient norm convergence when the gradient is only bounded) and conver-

Using the fact that for the case \( \nu = 1 \)

In case \( \nu = 1 \) the article \([4]\) (Section 7) shows that the upper bound for the expected number of steps is \( N = O \left( \frac{n}{\epsilon^2} \right) \) when \( \epsilon^2 = \nu \varepsilon_f \), while we show \( N = O \left( \frac{n^2}{\nu \epsilon_f} \right) \), which is \( n \) times worse. This can be improved quite easily using the fact that for the case \( \nu = 1 \)

then this inequality can be used to set \( A_1 = \frac{L}{\nu} \) and \( A_2 = 0 \) in (12), so the power of \( n \) in the first term will be 1 less and repeating following steps we will obtain \( N = O \left( \frac{n}{\nu \epsilon_f} \right) \). This, however, cannot be easily extended to \( \nu < 1 \), because of \( \|x - y\|^{\nu} \) term (see Lemma 2.2 proof for details).

### 3.3 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 \( \nu \to 0 \). We will now consider the convergence in the sense of smoothed function gradient norm while keeping functional gap (Lemma 2.3) small:

Substituting the same \( A_1, A_2 \) and \( A_1', A_2' \) into (12) we will obtain (13) but without the fourth term:

Substituting the same

Because we cannot ensure gradient norm convergence when the gradient is only bounded) and conver-

Using the fact that for the case \( \nu = 1 \)

In case \( \nu = 1 \) the article \([4]\) (Section 7) shows that the upper bound for the expected number of steps is \( N = O \left( \frac{n}{\epsilon^2} \right) \) when \( \epsilon^2 = \nu \varepsilon_f \), while we show \( N = O \left( \frac{n^2}{\nu \epsilon_f} \right) \), which is \( n \) times worse. This can be improved quite easily using the fact that for the case \( \nu = 1 \)

then this inequality can be used to set \( A_1 = \frac{L}{\nu} \) and \( A_2 = 0 \) in (12), so the power of \( n \) in the first term will be 1 less and repeating following steps we will obtain \( N = O \left( \frac{n}{\nu \epsilon_f} \right) \). This, however, cannot be easily extended to \( \nu < 1 \), because of \( \|x - y\|^{\nu} \) term (see Lemma 2.2 proof for details).

Thus \( \mu^{2\epsilon + \alpha} \sim \varepsilon_f \) and from Lemma 2.3 we have

Now, in previous subsection we had \( \mu \sim \frac{\varepsilon_f^{1/\nu}}{\nu} \) (for the case of \( \nu = 1 \)), so substituting it into Lemma 2.3 we would obtain \( \varepsilon_f \sim \varepsilon_f^{1/\nu} \). So let’s just consider this to be our case, then we can obtain
\[ \frac{\nu}{2
u} = 1 \] 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 \( \delta = \frac{\nu}{n^{\nu}} \) (the power of \( n \) is chosen to minimize powers of \( n \) in the right side of inequality) then

\[
\min_{k \in \{0,N-1\}} \mathbb{E}_{d_k} \| \nabla f_{\mu}(x_k) \|_2^2 \leq \frac{320(n+4)\mu_\nu(f_{\mu}(x_0) - f^*)}{N \mu^{2-2\nu} n^{\nu \frac{\nu}{n^{\nu}}} + 
\]

\[
+ \frac{\mu^{1+\nu}}{4(n+4)} \left( 4\mu^{3-3\nu} n^{\nu \frac{\nu}{n^{\nu}}} + \frac{4L^2}{(1+\nu)^2} n^{2+\nu} + 2L^2(n+3)(n+6)^{1+\nu} \right) +
\]

\[
+ 320(n+4)\mu^{1+\nu} \left( 2\mu^{1-\nu} n^{\nu \frac{\nu}{n^{\nu}}} + 4\mu\nu \right)
\]

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

\[
M \cdot n^{1+\nu} = \frac{4n^{\nu-1} + \frac{4L^2}{(1+\nu)^2} n^{2+\nu} + 2L^2(n+3)(n+6)^{1+\nu}}{8(n+4)} +
\]

\[
+ 160(n+4) \left( 2n^{\nu-1} + 4\mu\nu \right)
\]

and substituting it we obtain

\[
\min_{k \in \{0,N-1\}} \mathbb{E}_{d_k} \| \nabla f_{\mu}(x_k) \|_2^2 \leq \frac{320(n+4)n^{2-2\nu}L_\nu(f_{\mu}(x_0) - f^*)}{N \cdot M^{2-\frac{2\nu}{1+\nu}} n^{\frac{\nu}{n^{\nu}}} + \frac{\nu f}{2}}
\]

so we need to make

\[
N = O \left( \frac{n^{1+\nu} + \frac{1}{n^{\nu}}}{\frac{\nu f}{2}} \right) = O \left( \frac{n^{1+\nu} + \frac{1}{n^{\nu}}}{\frac{\nu f}{2}} \right) (15)
\]

steps to ensure \( \min_{k \in \{0,N-1\}} \mathbb{E}_{d_k} \| \nabla f_{\mu}(x_k) \|_2^2 \leq 2\nu f \). Substituting \( \mu = \mu_0 \) into Lemma 2.3 we obtain

\[
|f_{\mu}(x) - f(x)| \leq \frac{L_\nu}{1+\nu} \mu^{1+\nu} n^{\frac{\nu}{n^{\nu}}} = \Theta \left( \frac{\nu f}{n^{\nu}} \right)
\]

we ensure \( |f_{\mu}(x) - f(x)| \leq \nu f \) with \( \nu = \Theta \left( \frac{\nu f}{n^{\nu}} \right) \).

In case \( \nu = 0 \) [4] shows that \( N = O \left( \frac{n^3}{\nu f} \right) = \Theta \left( \frac{\nu f}{n^3} \right) \) which coincides with our result. In case \( \nu = 1 \) this result coincides with the result of the previous section, and we can repeat the reasoning at the end improving the result in this case by 1 power of \( n \).
| Convergence type | $N$ upper bound | $\delta$ upper bound | $|f_\mu(x) - f(x)|$ |
|------------------|-----------------|----------------------|-----------------|
| $\|\nabla f(x_k)\|_\ast$ | $O\left(\frac{n^{2+\frac{1}{\nu}}}{\varepsilon \nabla f}\right)$ | $O\left(\frac{(\frac{3n^{1+\nu}}{\varepsilon \nabla f})}{n^{\frac{1}{2}}}\right)$ | — |
| $\|\nabla f_\mu(x_k)\|_\ast$ | $O\left(\frac{n^{2+\frac{1}{\nu}}}{\varepsilon \nabla f}\right)$ | $O\left(\frac{(\frac{3n^{1+\nu}}{\varepsilon \nabla f})}{\frac{n}{4}}\right)$ | $\Theta\left(\frac{\varepsilon \nabla f}{n^{\frac{1}{2}}}\right)$ |

## 4 Conclusion

In this article we have extended the results of [4] on zeroth-order non-convex optimization to Hölder-gradient functions with noisy zeroth-order oracle. The Table 1 contains the $N$ and $\delta$ upper bounds for both types of convergence (the $\delta$ upper bounds in terms of $n$ and $\varepsilon \nabla f$ are obtained substituting $\mu$ into corresponding bounds). For the case of $\nu = 1$, the $N$ upper bound can be improved by reducing the power of $n$ by 1 in both cases. The interesting fact is that for the case of $\nu = 1$ the required noise bound is linear on $\varepsilon \nabla f$.

## References

[1] Pavel Dvurechensky. Gradient method with inexact oracle for composite non-convex optimization. 2017.

[2] Zbigniew Michalewicz and Cezary Z. Janikow. Genetic algorithms for numerical optimization. *Statistics and Computing*, 1(2):75–91, dec 1991. doi: 10.1007/bf01889983.

[3] Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, Aug 2015. ISSN 1436-4646. doi: 10.1007/s10107-014-0790-0. URL https://doi.org/10.1007/s10107-014-0790-0.

[4] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, nov 2015. doi: 10.1007/s10208-015-9296-2.

[5] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[6] Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. *CoRR*, abs/1810.01032, 2018. URL http://arxiv.org/abs/1810.01032.
A Appendix

A.1 Proofs of Lemmas 2.1—2.4

**Proof of Lemma 2.1.** From (1) we get \( \|Bu\|_*^2 = \langle Bu, B^{-1}Bu \rangle = \langle Bu, u \rangle = \|u\|^2 \). Using this and Lemma A.1 we obtain

\[
\|\nabla \tilde{f}_\mu(x, \delta) - \nabla f_\mu(x)\|_* \overset{\text{Def. 2.3}}{=} \left\| \frac{1}{\kappa} \int_E \left( \frac{\tilde{f}(x + \mu u, \delta) - \tilde{f}(x, \delta) - f(x + \mu u) - f(x)}{\mu} \right) Bue^{-\frac{1}{2}\|u\|^2} du \right\|_* \leq \frac{1}{\kappa} \int_E \left( \frac{|f(x + \mu u, \delta) - f(x + \mu u)|}{\mu} + \frac{|\tilde{f}(x, \delta) - f(x)|}{\mu} \right) \|Bu\|_* e^{-\frac{1}{2}\|u\|^2} du \leq 2 \frac{\delta}{\mu} n^{1/2}
\]

**Proof of Lemma 2.2.**

\[
\|\nabla f_\mu(y) - \nabla f_\mu(x)\|_* \overset{\text{Def. 2.3}}{=} \left\| \frac{1}{\kappa} \int_E \left( \frac{f(y + \mu u) - f(y) - f(x + \mu u) - f(x)}{\mu} \right) Bue^{-\frac{1}{2}\|u\|^2} du \right\|_* \leq \frac{1}{\kappa} \int_E \left( \frac{1}{0} \langle \nabla f(\mu u + ty + (1 - t)x) - \nabla f(ty + (1 - t)x), y - x \rangle dt \right) \|u\| e^{-\frac{1}{2}\|u\|^2} du \overset{\text{Def. 2.2}}{=} \left\| L_{\nu} \mu^{\nu} \|y - x\|^{1+\nu} e^{-\frac{1}{2}\|u\|^2} du \right\|_* \leq \frac{L_{\nu}}{\mu^{1-\nu}} n^{1+\nu \frac{1}{2}} \|y - x\|.
\]

Integrating this we obtain

\[
f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle \leq \frac{L_{\nu}}{2\mu^{1-\nu}} n^{1+\frac{\nu}{2}} \|y - x\|^2 \quad (16)
\]
so using this way we proved lemma with $A_1 = \frac{L_\nu}{\mu^{1+\nu}} n^{1+\frac{1}{2}}$ 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 A.2:

$$f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle \overset{\text{Def. 2.3}}{=} \frac{1}{\kappa} \int_E \left( f(y + \mu u) - f(x + \mu u) - \langle \nabla f(x + \mu u), y - x \rangle \right) e^{-\frac{1}{2} \|u\|^2} du \overset{\text{Def. 2.2}}{\leq} L_\nu \|y - x\|^{1+\nu} \overset{\text{A.2}}{\leq} \frac{1 - \nu}{2} \frac{L_\nu}{1 + \nu} \|y - x\|^2 + \tilde{\delta}$$

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

$$f_\mu(y) - f_\mu(x) - \langle \nabla f_\mu(x), y - x \rangle \leq \left[ \frac{1}{\delta} \right]^{1+\nu} \frac{L_\nu}{\mu^{1+\nu}} \|y - x\|^2 + \tilde{\delta} \mu L_\nu^{1+\nu} \quad (17)$$

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

**Proof of Lemma 2.3.** To prove 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

$$|f_\mu(x) - f(x)| \overset{\text{Def. 2.3}}{=} \left| \int_E \left( f(y + \mu u) - f(x + \mu u) - \langle \nabla f(x + \mu u), y - x \rangle \right) e^{-\frac{1}{2} \|u\|^2} du \right| =$$

$$= \left| \int_E \left( f(x + \mu u) - f(x) - \langle \nabla f(x), \mu u \rangle \right) e^{-\frac{1}{2} \|u\|^2} du \right| \overset{\text{Def. 2.2}}{\leq} L_\nu \|y - x\|^{1+\nu} \overset{\text{Def. 2.2}}{\leq} \frac{L_\nu}{1 + \nu} \|y - x\|^{1+\nu} \mu^{1+\nu} \overset{\text{A.1}}{\leq} \frac{L_\nu}{1 + \nu} \mu^{1+\nu} n^{1+\nu} \quad \Box$$

**Proof of Lemma 2.4.** To prove this we use a fact that

$$\nabla f(x) = \frac{1}{\kappa} \int_E \langle \nabla f(x), u \rangle B u e^{-\frac{1}{2} \|u\|^2} du$$
and that $a^2 \leq 2(a + b)^2 + 2b^2$:
\[
\|\nabla f(x)\|^2 \overset{\text{Def. 2.3}}{=} \frac{1}{\kappa} \int_E \langle \nabla f(x), u \rangle \mu e^{-\frac{1}{2}\|u\|^2} du \leq 
\]
\[
\leq 2 \left( \frac{1}{\kappa} \int_E \left( f(x + \mu u) - f(x) \right) \mu e^{-\frac{1}{2}\|u\|^2} du \right)^2 + 2 \left( \frac{1}{\kappa} \int_E \left( f(x + \mu u) - \langle \nabla f(x), \mu u \rangle \right) \mu e^{-\frac{1}{2}\|u\|^2} du \right)^2 \overset{\text{Def. 2.2}}{\leq} 
\]
\[
\leq 2\|\nabla f_\mu(x)\|^2 + 2 \left( \frac{1}{\mu \kappa} \int_E \left( L_{\nu} \mu^{1+\nu} \|u\|^{1+\nu} \right) \mu e^{-\frac{1}{2}\|u\|^2} du \right)^2 \overset{\text{Def. 2.2}}{\leq} 
\]
\[
\leq 2\|\nabla f_\mu(x)\|^2 + \frac{2\mu^{2\nu}}{(1 + \nu)^2} L_{\nu}^2 (n + 2 + \nu)^{2+\nu} 
\]

A.2 External results

Lemma A.1 (Lemma 1 from [4]). For $p \geq 0$, we have
\[
\frac{1}{\kappa} \int_E \|u\|^{p-\frac{1}{2}} \|u\|^2 du \leq \begin{cases} 
np^{p/2}, & p \in [0, 2] \\
(n + p)^{p/2}, & p > 2
\end{cases} \tag{18}
\]

Lemma A.2 (Lemma 2 from [3]). Let the function $f$ satisfy Definition 2.2. Then for any $\tilde{\delta} > 0$
\[
\frac{L_{\nu}}{1 + \nu} \mu^{1+\nu} \leq \frac{1}{2} \left( \frac{1 - \nu}{1 + \nu} \right)^{\frac{1+\nu}{\nu}} \frac{L_{\nu}^2}{L_{\nu}^2} \mu^2 + \tilde{\delta} = \frac{L_{\nu}^2}{2} \mu^2 + \tilde{\delta} \tag{19}
\]

Theorem A.3 (Theorem 3 from [4]). If $f$ is differentiable at $x$ and $u$ is a standard random normal vector, then
\[
\mathbb{E}_u \left( \langle \nabla f(x), u \rangle^2 \right) \leq (n + 4)\|\nabla f(x)\|^2 \tag{20}
\]