Записки научных
семинаров ПОМИ
Том 540, 2024 г.

**A. Lobanov, A. Gasnikov**

# IMPROVED MAXIMUM NOISE LEVEL ESTIMATION IN BLACK-BOX OPTIMIZATION PROBLEMS

Abstract. In black-box optimization, accurately estimating the maximum noise level is crucial for robust performance. In this work, we propose a novel approach for improving maximum noise level estimation, focusing on scenarios where only function values (possibly with bounded adversarial noise) are available. Leveraging gradient-free optimization algorithms, we introduce a new noise constraint based on the Lipschitz assumption, enhancing the noise level estimate (or improving error floor) for non-smooth and convex functions. Theoretical analysis and numerical experiments demonstrate the effectiveness of our approach, even for smooth and convex functions. This advancement contributes to enhancing the robustness and efficiency of black-box optimization algorithms in diverse domains such as machine learning and engineering design, where adversarial noise presents a significant challenge.

## §1. Introduction

Adversarial noise poses a significant challenge in various computational tasks, particularly in the context of optimization problems where accurate estimation is crucial for achieving robust solutions. Adversarial noise usually refers to the perturbations or disturbances introduced to input data intentionally to mislead or deceive computational models. In the realm of black-box optimization, where the underlying objective function is either unknown or expensive to evaluate directly, adversarial noise can severely impact the performance and reliability of optimization algorithms. Understanding and effectively mitigating adversarial noise are paramount for ensuring the success of optimization techniques across diverse domains, including machine learning, engineering design, finance, and beyond. The presence of adversarial noise can lead to misdirection of search algorithms, and ultimately suboptimal or unreliable solutions.

*Key words and phrases*: noise level estimation, black-box optimization, adversarial noise.

This research was supported by the Russian Science Foundation, project no. 21-71-30005, `https://rscf.ru/en/project/21-71-30005/`.

In this work, we focus on the following general optimization problem:

$$\min_{x \in Q \subseteq \mathbb{R}^d} \{f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} f(x, \xi)\}, \tag{1}$$

where $f : Q \to \mathbb{R}$ is a non-smooth, convex, and possibly stochastic function. This problem formulation is widely used and has many applications in machine learning. However, we consider a subclass of this problem, assuming that the oracle has access only to the value of the objective function (possibly with some limited adversarial noise), i.e., we have access only to the zero-order oracle [13]. This problem setting is actively studied in the literature, where the authors often assign it to the class of black-box optimization problems [14]. Indeed, this class can be formally understood as a black-box optimization problem, where a zero-order oracle acts as the black box. Although we consider a narrower class of problems than (1), the black-box optimization problem is studied in applications such as machine learning [15], deep learning [16], reinforcement learning [17], federated learning [18], online optimization [19], multi-armed bandits [20, 21], hyperparameter tuning [22], "perfect" product creation [23], and more.

Gradient-free optimization algorithms are often used to solve such problems. Several concepts stand out in the literature for creating algorithms that use only function values (zero-order oracle) and do not have access to more complex information about the function, such as the n-th order of the derivative. One such concept, the one most amenable to theoretical analysis, uses the "power" of higher-order algorithms to create already gradient-free algorithms by using a gradient approximation instead of the true gradient. There is a classification of gradient approximations depending on the problem statement [24]. For example, in optimization problems where the function has increased smoothness, authors of the works [25, 26, 8, 27] use kernel approximation, where it is the kernel function that takes advantage of the increased smoothness. In the other direction, namely when the function is simply smooth, the works [29, 30, 28] use $l_1$ or $l_2$ randomization gradient approximation. Finally, when the function is non-smooth, the works [1, 31] use a smoothing scheme to apply $l_1$ or $l_2$ randomization instead of the true gradient. In this work, we focus on the non-smooth case, so we also use a smoothing scheme with $l_2$ randomization, but we add a discussion of the results in the smooth case too.

Recently, the performance of gradient-free algorithms has been evaluated using three optimality criteria. The first two criteria are also widely used in higher-order optimization, namely the number of iterations $N$ to

achieve the desired accuracy (iteration complexity) and the number of oracle calls $T$ (oracle complexity). But the third optimality criterion is specific to gradient-free algorithms and represents an estimate of the maximum noise level $\Delta$ at which the desired accuracy can still be achieved. In other words, the third optimality criterion represents a certain threshold up to which the algorithm behaves as if there were no noise at all, demonstrating adaptability to adversarial noise. But exceeding this threshold, the algorithm worsens convergence (and may not converge at all).

There are many works investigating the maximum level of deterministic noise. In the non-smooth formulation of the black-box optimization problem (1), it is shown (see [1, 30, 2]) that the optimal estimate of the maximum noise level is $\Delta \lesssim \varepsilon^2/\sqrt{d}$. Other works [24, 5] have shown that this estimator can be improved by imposing additional assumptions on the function. For example, if the function is *non-smooth and $\mu$-strongly convex* [32], the estimate can be improved to $\Delta \lesssim \mu^{1/2}\varepsilon^{3/2}/\sqrt{d}$. In case when the function is *already smooth and convex* [24], the estimate would change as follows $\Delta \lesssim \varepsilon^{3/2}/\sqrt{d}$. Finally, if the function is *smooth and $\mu$-strongly convex* [5], then the estimate of the maximum deterministic noise level can be represented as $\mu^{1/2}\varepsilon/\sqrt{d}$. In this work, we consider another way to improve the original maximum noise estimate when the function is *non-smooth and convex*. Instead of using the assumption that the noise is bounded in absolute value, we assume that the Lipschitz constraint is satisfied (see Assumption 4), obtaining the following estimate for the maximum noise level $\Delta \lesssim \varepsilon/\sqrt{d}$. Moreover, we show theoretically and numerically that this improvement holds even if the objective function is smooth and convex.

## §2. Notation and Assumptions

In this section, we present the final formulation of the problem considered in this work, imposing constraints on the objective function and adversarial deterministic noise. But before going into the assumptions we present the notations that are used throughout the paper.

*Notation.* We use $\langle x, y\rangle := \sum_{i=1}^d x_i y_i$ to denote the standard inner product of $x, y \in \mathbb{R}^d$, where $x_i$ and $y_i$ are the $i$-th components of $x$ and $y$ respectively. We denote $l_p$-norms (for $p \geqslant 1$) in $\mathbb{R}^d$ as $\|x\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$. In particular, for the $l_2$-norm in $\mathbb{R}^d$ it follows that $\|x\|_2 := \sqrt{\langle x, x\rangle}$.

We denote the $l_p$-ball as $B_p^d(r) := \left\{x \in \mathbb{R}^d : \|x\|_p \leqslant r\right\}$ and $l_p$-sphere as $S_p^d(r) := \left\{x \in \mathbb{R}^d : \|x\|_p = r\right\}$. Operator $\mathbb{E}[\cdot]$ denotes the expectation. To denote the distance between the initial point $x^0$ and the solution of the initial problem $x_*$ we introduce $R := \tilde{\mathcal{O}}\left(\|x^0 - x_*\|_p\right)$, where we use notation $\tilde{\mathcal{O}}(\cdot)$ to hide logarithmic factors.

We can now present the main assumptions used in this work.

*Assumptions on the objective function.* Throughout the paper, we assume that the objective function $f$ is $M$-Lipschitz continuous.

**Assumption 1.** *The function $f(x, \xi)$ is an $M$-Lipschitz continuous function in the $l_p$-norm, i.e for all $x, y \in Q$ we have*

$$|f(y, \xi) - f(x, \xi)| \leqslant M(\xi)\|y - x\|_p.$$

*Moreover, there exists a positive constant $M$ such that $\mathbb{E}\left[M^2(\xi)\right] \leqslant M^2$. In particular, for $p = 2$ we use the notation $M_2$ for the Lipschitz constant.*

We use the following assumption when we specify our results into the class of convex smooth functions.

**Assumption 2** (Smoothness of function)**.** *The function $f$ is smooth, that is, differentiable on $Q$ and such that for all $x, y \in Q$ with $L > 0$ we have*

$$\|\nabla f(y) - \nabla f(x)\|_q \leqslant L\|y - x\|_p.$$

**Assumption 3** (Convexity on the set $Q_\gamma$)**.** *Let $\gamma > 0$ be a small number to be defined later and let $Q_\gamma := Q + B_p^d(\gamma)$; then the function $f$ is convex on the set $Q_\gamma$.*

The assumptions presented in this subsection are not unique, they are actively used in the optimization community. For example, Assumption 1 is a standard assumption for works that obtain theoretical estimates when smoothness is not available. Assumption 2 is probably one of the most common assumptions in the field of numerical optimization methods and beyond. Finally, Assumption 3, introduced to obtain theoretical estimates, is common in works that use smoothing schemes to develop gradient-free optimization algorithms.

*Assumptions on the deterministic noise.* Before proceeding to the assumptions we introduce definitions for the zero-order oracle and gradient approximation.

**Definition 1** (Zero-order oracle)**.** *The oracle returns a function value $f(x,\xi)$ at the requested point $x$ with some adversarial deterministic noise, i.e., for all $x \in Q$*

$$f_\delta(x,\xi) := f(x,\xi) + \delta(x).$$

Now, using the definition of a zero-order oracle, we present the key assumption of this paper.

**Assumption 4** (Lipschitz bounded noise)**.** *The noise function $\delta(x)$ is a $\Delta$-Lipschitz continuous function, i.e., $\forall x, y \in Q$ we have*

$$|\delta(y) - \delta(x)| \leqslant \Delta \|y - x\|_2.$$

Definition 1 is a common definition in the literature describing the possibility of obtaining inexact information from a (black box) oracle $f_\delta(x)$. However, the noise $\delta(x)$ that is presented in Definition 1 must be bounded to guarantee the convergence of the algorithms. The noise constraint itself (Assumption 4) is narrower than other works when the noise was constrained in absolute value, $|\delta(x)| \leqslant \Delta$.

## §3. Idea of the smoothing scheme

In this section, we give a brief explanation of how to create gradient-free algorithms for the original optimization problem (1) in various settings using a smoothing scheme via $l_2$ randomization, assuming that there is no adversarial noise $\Delta = 0$.

The main idea of the smoothing scheme is to replace the original problem (1) by a smooth problem. For this purpose we introduce a smooth approximation of the non-smooth function $f$:

$$f_\gamma(x) := \mathbb{E}_{\tilde{e}} \left[ f(x + \gamma \tilde{e}) \right], \tag{2}$$

where $\gamma > 0$ is a smoothing parameter and $\tilde{e}$ is a random vector uniformly distributed on $B_2^d(1)$. Hereinafter, for simplicity, we denote $f(x) := \mathbb{E}\left[f(x,\xi)\right]$. Now we write down the properties of the smoothed function $f_\gamma$ that will help us understand how the two problems are related: the original non-smooth problem and the already smoothed problem.

**Lemma 1.** *Suppose that Assumptions* 1, 3 *hold; then for all $x \in Q$ we have*

$$f(x) \leqslant f_\gamma(x) \leqslant f(x) + \gamma M_2.$$

**Proof.** For the first inequality we use the convexity of the function $f(x)$

$$f_\gamma(x) = \mathbb{E}_{\tilde{e}}\left[f(x+\gamma\tilde{e})\right] \geqslant \mathbb{E}_{\tilde{e}}\left[f(x) + \langle \nabla f(x), \gamma\tilde{e}\rangle)\right] = \mathbb{E}_{\tilde{e}}\left[f(x)\right] = f(x).$$

For the second inequality we have

$$\begin{aligned} |f_\gamma(x) - f(x)| = |\mathbb{E}_{\tilde{e}}\left[f(x+\gamma\tilde{e})\right] - f(x)| &\leqslant \mathbb{E}_{\tilde{e}}\left[|f(x+\gamma\tilde{e}) - f(x)|\right] \\ &\leqslant \gamma M_2 \mathbb{E}_{\tilde{e}}\left[\|\tilde{e}\|_2\right] \leqslant \gamma M_2, \end{aligned}$$

using the fact that $f$ is an $M_2$-Lipschitz function. □

**Remark 1.** Indeed, knowing the relationship between the functions $f$ and $f_\gamma$ we can tell how the two problems are related: to achieve an $\varepsilon$-accuracy solution in a non-smooth problem, we need to address the corresponding smooth problem with $(\varepsilon/2)$-accuracy. Here, $\varepsilon$-suboptimality denotes the accuracy of the solution in terms of expectation:

$$\begin{aligned} f(x^{N+1}) - f(x_*) \leqslant f(x^{N+1}) - f(x_*(\gamma)) &\overset{①}{\leqslant} f_\gamma(x^{N+1}) - f(x_*(\gamma)) \\ &\overset{②}{\leqslant} f_\gamma(x^{N+1}) - f_\gamma(x_*(\gamma)) + \gamma M_2 \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

where ① means the first inequality of Lemma 1 and ② means the second inequality of Lemma 1.

We now write down the remaining properties of the smoothed function $f_\gamma$, namely that the function is also $M$-Lipschitz continuous in $l_p$ norm, and also that the function now has a Lipschitz gradient constant.

**Lemma 2.** *Suppose that Assumptions* 1, 3 *hold; then for* $f_\gamma(x)$ *from* (2) *we have*

$$|f_\gamma(y) - f_\gamma(x)| \leqslant M\|y - x\|_p, \quad \forall x, y \in Q.$$

**Proof.** Using $M$-Lipschitz continuity of function $f$ we obtain

$$|f_\gamma(y) - f_\gamma(x)| \leqslant \mathbb{E}_{\tilde{e}}\left[|f(y+\gamma\tilde{e}) - f(x+\gamma\tilde{e})|\right] \leqslant M\|y - x\|_p. \qquad \square$$

**Lemma 3** ([1, Theorem 1])**.** *Suppose that Assumptions* 1, 3 *hold; then* $f_\gamma(x)$ *has a* $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$*-Lipschitz gradient,*

$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leqslant L_{f_\gamma}\|y - x\|_p, \quad \forall x, y \in Q.$$

## §4. Main Results

In this section, we present the main result of this paper, namely an improved estimate on the maximum level of adversarial deterministic noise. But before moving on to the result, we conclude the explanation we started in the previous section about creating gradient-free algorithms to solve the original optimization problem in different settings. Now that we understand how the two problems are related, we only need to choose the optimal optimization algorithm (often accelerated and batched) and use it to solve the smooth optimization problem with $\varepsilon/2$ accuracy, in order to obtain the optimal algorithm to solve the original problem (1) with $\varepsilon$ accuracy.

**Remark 2.** However, the true gradient is still not available to us, so the gradient $f_\gamma(x,\xi)$ can be estimated by the following approximation (also known as $l_2$ randomization):

$$\nabla f_\gamma(x,\xi,e) = \frac{d}{2\gamma}\left(f_\delta(x+\gamma e,\xi) - f_\delta(x-\gamma e,\xi)\right)e, \tag{3}$$

where $f_\delta(x,\xi)$ is the gradient-free oracle from Definition 1 and $e$ is a random vector uniformly distributed on $S_2^d(\gamma)$.

It is not hard to see that the gradient approximation (3) uses a zero-order oracle that produces a noisy value of the objective function (see Definition 1). In order to guarantee "good" convergence for a gradient-free optimization algorithm, we need to find the maximum noise level. Typically, adversarial noise accumulates in two places: in the variance and in the bias. So we will look at each estimator to find the maximum noise level. And we start with the second moment (variance) of the gradient approximation.

**Lemma 4** (Second moment)**.** *Suppose that Assumptions* 1 *and* 4 *hold; then for all $x \in Q$ the gradient approximation $\nabla f_\gamma(x,\xi,e)$ via $l_2$ randomization* (3) *has the following variance (second moment):*

$$\mathbb{E}_{\xi,e}\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] \leqslant \kappa(p,d)\left(M_2^2 + \Delta^2\right),$$

*where $1/p + 1/q = 1$ and $\kappa(p,d) = \sqrt{2}\min\{q, \ln d\}\, d^{2-\frac{2}{p}}$.*

**Proof.** By definition we have

$$\begin{aligned}\mathbb{E}_{\xi,e}\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] &= \mathbb{E}_{\xi,e}\left[\left\|\frac{d}{2\gamma}(f_\delta(x+\gamma e,\xi)-f_\delta(x-\gamma e,\xi))e\right\|_q^2\right]\\ &= \frac{d^2}{4\gamma^2}\mathbb{E}_{\xi,e}[\|e\|_q^2(f(x+\gamma e,\xi)+\delta(x+\gamma e)\\ &\quad -f(x-\gamma e,\xi)+\delta(x-\gamma e))^2]\\ &\leqslant \frac{d^2}{2\gamma^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\gamma e,\xi)-f(x-\gamma e,\xi)\right)^2\right]\\ &\quad +\frac{d^2}{2\gamma^2}\mathbb{E}_e\left[\|e\|_q^2\left(\delta(x+\gamma e)-\delta(x-\gamma e)\right)^2\right], \quad (4)\end{aligned}$$

where we used the fact that for all $a, b, (a+b)^2 \leqslant 2a^2+2b^2$. For the first term in (4), the following holds with an arbitrary parameter $\alpha$ given the symmetric distribution $e$:

$$\begin{aligned}&\frac{d^2}{2\gamma^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\gamma e,\xi)-f(x-\gamma e,\xi)\right)^2\right]\\ &\quad=\frac{d^2}{2\gamma^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left((f(x+\gamma e,\xi)-\alpha)-(f(x-\gamma e,\xi)-\alpha)\right)^2\right]\\ &\quad\leqslant\frac{d^2}{\gamma^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\gamma e,\xi)-\alpha\right)^2+\left(f(x-\gamma e,\xi)-\alpha\right)^2\right]\\ &\quad=\frac{d^2}{\gamma^2}\left(\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\gamma e,\xi)-\alpha\right)^2\right]+\mathbb{E}_{\xi,e}\left[\left(f(x-\gamma e,\xi)-\alpha\right)^2\right]\right)\\ &\quad=\frac{2d^2}{\gamma^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\gamma e,\xi)-\alpha\right)^2\right]\\ &\quad\overset{①}{\leqslant}\frac{2d^2}{\gamma^2}\mathbb{E}_{\xi}\left[\sqrt{\mathbb{E}\left[\|e\|_q^4\right]}\sqrt{\mathbb{E}_e\left[\left(f(x+\gamma e,\xi)-\alpha\right)^4\right]}\right]\\ &\quad\overset{②}{\leqslant}\frac{2d^2\kappa'(p,d)}{\gamma^2}\mathbb{E}_{\xi}\left[\sqrt{\mathbb{E}_e\left[\left(f(x+\gamma e,\xi)-\alpha\right)^4\right]}\right], \quad (5)\end{aligned}$$

where in ① and ② we used the Cauchy-Schwarz inequality and the fact that $\sqrt{\mathbb{E}\left[\|e\|_q^4\right]} \leqslant \kappa'(p,d)$, where $\kappa'(p,d) = \min\{q, \ln d\}\, d^{2/q-1}$. Now we

perform similar operations for the second term in (4):

$$\begin{aligned}
&\frac{d^2}{2\gamma^2}\mathbb{E}_e\left[\|e\|_q^2\left(\delta(x+\gamma e)-\delta(x-\gamma e)\right)^2\right] \\
&\quad=\frac{d^2}{2\gamma^2}\mathbb{E}_e\left[\|e\|_q^2\left((\delta(x+\gamma e)-\beta)-(\delta(x-\gamma e)-\beta)\right)^2\right] \\
&\quad\leqslant\frac{d^2}{\gamma^2}\mathbb{E}_e\left[\|e\|_q^2\left(\delta(x+\gamma e)-\beta\right)^2+\left(\delta(x-\gamma e)-\beta\right)^2\right] \\
&\quad=\frac{d^2}{\gamma^2}\left(\mathbb{E}_e\left[\|e\|_q^2\left(\delta(x+\gamma e)-\beta\right)^2\right]+\mathbb{E}_e\left[\left(\delta(x-\gamma e)-\beta\right)^2\right]\right) \\
&\quad=\frac{2d^2}{\gamma^2}\mathbb{E}_e\left[\|e\|_q^2\left(\delta(x+\gamma e)-\beta\right)^2\right] \\
&\quad\leqslant\frac{2d^2}{\gamma^2}\sqrt{\mathbb{E}\left[\|e\|_q^4\right]}\sqrt{\mathbb{E}_e\left[\left(\delta(x+\gamma e)-\beta\right)^4\right]} \\
&\quad\leqslant\frac{2d^2\kappa'(p,d)}{\gamma^2}\sqrt{\mathbb{E}_e\left[\left(\delta(x+\gamma e)-\beta\right)^4\right]}. \qquad (6)
\end{aligned}$$

Now applying Lemma 4 of [2] to the $\gamma M_2(\xi)$-Lipschitz function $f(x+\gamma e,\xi)$ with respect to $e$ in terms of the $l_2$ norm from (5) and to the $\gamma\Delta$-Lipschitz function $\delta(x+\gamma e)$ from (6) we obtain the original statement of the Lemma:

$$\begin{aligned}
\mathbb{E}_{\xi,e}\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] &\leqslant \frac{2d^2\kappa'(p,d)}{\gamma^2}\mathbb{E}_\xi\left[\sqrt{\mathbb{E}_e\left[\left(f(x+\gamma e,\xi)-\alpha\right)^4\right]}\right] \\
&\quad+\frac{2d^2\kappa'(p,d)}{\gamma^2}\sqrt{\mathbb{E}_e\left[\left(\delta(x+\gamma e)-\beta\right)^4\right]} \\
&\leqslant \frac{2d^2\kappa'(p,d)}{\gamma^2}\left(\frac{\gamma^2\mathbb{E}_\xi\left[M_2^2(\xi)\right]}{\sqrt{2}d}+\frac{\gamma^2\Delta^2}{\sqrt{2}d}\right) \\
&= \kappa(p,d)\left(M_2^2+\Delta^2\right),
\end{aligned}$$

where $\kappa(p,d)=\sqrt{2}d\kappa'(p,d)=\sqrt{2}\min\{q,\ln d\}d^{2-\frac{2}{p}}$. □

Now in order to provide an estimate on the bias of the gradient approximation (3), we first cite several known facts in the form of lemmas.

**Lemma 5** ([3]). *The function $f_\gamma(x)$ is differentiable with the following gradient with $l_2$-randomization:*

$$\nabla f_\gamma(x) = \mathbb{E}_e \left[ \frac{d}{\gamma} f(x + \gamma e) e \right].$$

**Lemma 6** ([4]). *Let vector $e$ be a random unit vector from the Euclidean unit sphere $\{e : \|e\|_2 = 1\}$. Then for all $r \in \mathbb{R}^d$ it follows that*

$$\mathbb{E}[|\langle e, r\rangle|] \leqslant \frac{\|r\|_2}{\sqrt{d}}.$$

We can now present an estimate on the bias of the gradient approximation via $l_2$ randomization (3).

**Lemma 7** (Bias). *Suppose that Assumption* (4) *holds; then the gradient approximation $\nabla f_\gamma$ has the following bias:*

$$\langle \mathbb{E}_{\xi,e} [\nabla f_\gamma(x, \xi, e)] - \nabla f_\gamma(x), r\rangle \lesssim \sqrt{d}\Delta\|r\|_2, \qquad \forall r \in \mathbb{R}^d.$$

**Proof.** By definition of gradient approximation we have:

$$\begin{aligned}
\nabla f_\gamma(x, \xi, e) &= \frac{d}{2\gamma}(f_\delta(x + \gamma e, \xi) - f_\delta(x - \gamma e, \xi))e \\
&= \frac{d}{2\gamma}(f(x + \gamma e, \xi) + \delta(x + \gamma e) - f(x - \gamma e, \xi) - \delta(x - \gamma e))e \\
&= \frac{d}{2\gamma}((f(x + \gamma e, \xi) - f(x - \gamma e, \xi))e + (\delta(x + \gamma e) - \delta(x - \gamma e))e).
\end{aligned}$$

It follows from this equality that

$$\begin{aligned}
\mathbb{E}_{\xi,e} [\langle \nabla f_\gamma(x, \xi, e), r\rangle] = \frac{d}{2\gamma}\mathbb{E}_{\xi,e}[\langle (f(x + \gamma e, \xi) - f(x - \gamma e, \xi))e, r\rangle] \\
+ \frac{d}{2\gamma}\mathbb{E}_e[\langle (\delta(x + \gamma e) - \delta(x - \gamma e))e, r\rangle]. \qquad (7)
\end{aligned}$$

Applying Lemma 5 to the first term in (7), we get

$$\begin{aligned}
&\frac{d}{2\gamma}\mathbb{E}_{\xi,e}[\langle (f(x + \gamma e, \xi) - f(x - \gamma e, \xi))e, r\rangle] \\
&\qquad = \frac{d}{2\gamma}\mathbb{E}_{\xi,e}[\langle f(x + \gamma e, \xi)e, r\rangle] + \frac{d}{2\gamma}\mathbb{E}_{\xi,e}[\langle f(x - \gamma e, \xi)e, r\rangle] \\
&\qquad = \frac{d}{\gamma}\mathbb{E}_e \left[\langle \mathbb{E}_\xi \left[f(x + \gamma e, \xi)\right] e, r\rangle\right] = \frac{d}{\gamma}\mathbb{E}_e[\langle f(x + \gamma e)e, r\rangle] \\
&\qquad = \langle \nabla f_\gamma(x), r\rangle . \qquad (8)
\end{aligned}$$

For the second term in (7), with Assumption 4 satisfied, we obtain

$$\frac{d}{\gamma}\mathbb{E}_e[\langle(\delta(x+\gamma e)-\delta(x-\gamma e))e,r\rangle] \geqslant -\frac{d}{\gamma}\Delta\|\gamma e\|_2\mathbb{E}_e[|\langle e,r\rangle|] = -d\Delta\mathbb{E}_e[|\langle e,r\rangle|]. \tag{9}$$

Substituting (8) and (9) into expression (7), we get that

$$\mathbb{E}_{\xi,e}\left[\langle\nabla f_\gamma(x,\xi,e),r\rangle\right] \geqslant \langle\nabla f_\gamma(x),r\rangle - d\Delta\mathbb{E}_e[|\langle e,r\rangle|]. \tag{10}$$

Applying the statement of Lemma (6) to expression (10), we obtain the original statement of the Lemma. □

We are now ready to present the main result of this work.

**Theorem 1.** *Suppose that Assumptions* 1, 3, 4 *are satisfied. Then algorithm* $\boldsymbol{A}(L,\sigma^2)$ *obtained by applying smoothing schemes via* $l_2$ *randomization* (*see Section* 3) *based on the first order method has the following noise level:*

$$\Delta \lesssim \frac{\varepsilon}{\sqrt{d}},$$

*where* $\varepsilon$ *is the accuracy of the solution to problem* (1), $\mathbb{E}[f(x_N)] - f^* \leqslant \varepsilon$.

**Proof.** Since adversarial noise only accumulates in the variance and bias of the gradient approximation via $l_2$ randomization, the following conditions must be satisfied in order to guarantee "good" convergence of the algorithm $\mathbf{A}(L,\sigma^2)$ (while maintaining optimal estimates of iterative and oracle complexity):

- *for the second moment*; it follows from the statement of Lemma 4 that
$$\Delta^2 \leqslant M_2^2. \tag{11}$$
- *for the bias*; it follows from the statement of Lemma 7, given that $R = \|x_0 - x^*\|_2 = \|r\|_2$, that:
$$\sqrt{d}\Delta R \leqslant \varepsilon \qquad \Rightarrow \qquad \Delta \leqslant \frac{\varepsilon}{\sqrt{d}R}. \tag{12}$$

Since the estimate on the maximum noise level (12) is more influential than the estimate obtained from the variance (11), we obtain the original statement of the theorem. □

By the results of Theorem 1, we see that it is indeed possible to improve existing estimates (see [2]) of the maximum level of adversarial deterministic noise in the non-smooth convex problem. Moreover, estimation

presented in Theorem 1 shows that applying the noise constraint achieves a better estimate in the non-smooth case than the existing one, namely this estimate outperforms existing estimates in the following settings: non-smooth strongly convex function $\sim \mu^{1/2}\varepsilon^{3/2}/\sqrt{d}$, smooth convex function $\sim \varepsilon^{3/2}/\sqrt{d}$ and finally smooth strongly convex function $\sim \mu^{1/2}\varepsilon/\sqrt{d}$ (see Section 4 [5]). In addition, we point out that the technique presented in this work achieves the same estimate as the maximum level of adversarial stochastic noise in the non-smooth convex case $\sim \varepsilon/\sqrt{d}$ (see [6]). Finally, this estimate is fully consistent with the maximum deterministic noise estimate in the same Assumption 4, but in the setting of a non-smooth saddle optimization problem [30]. In the following, we present a development of these results in the form of remarks for the case of a smooth setting of the initial optimization problem (1), as well as the variation of the estimate depending on the chosen randomization.

**Remark 3** (Smooth setting)**.** If Assumption 2 is satisfied, Lemma 1 takes the following form: $f(x) \leqslant f_\gamma(x) \leqslant f(x) + \gamma^2 L^2$. The modified Lemma 1 in turn affects the smoothing parameter, i.e., instead of $\gamma = \frac{\varepsilon}{2M_2}$ (see Section 3), the smoothing parameter becomes $\sqrt{\varepsilon/L}$. However, all conclusions of Lemmas 4 and 7 are independent of the smoothing parameter because of the fulfillment of the Lipschitz noise Assumption 4. Thus, the result of Theorem 1 holds even in the smooth formulation of the original problem.

**Remark 4** ($l_1$ randomization)**.** Considering a smoothing scheme with $l_1$ randomization and performing the same operations as in this work, Lemma 4 will not change conceptually (i.e., the noise level term will not change), but Lemma 7 (which can be obtained using Assumption 4 in Lemma 8 of [2]) will deteriorate compared to $l_2$ randomization, giving the following estimate for the maximum noise level: $\Delta \leqslant \varepsilon/d$.

## §5. Experiments

In this section, we focus on the verification of the theoretical results obtained in Section 4. To demonstrate the effectiveness of the proposed method for improving the maximum noise level, we consider a smooth formulation of the original problem (1), namely the solution of a system of $p$ nonlinear equations [7, 8]:

$$\min_{x\in\mathbb{R}^d} f(x) := \|g(x)\|_2^2,$$

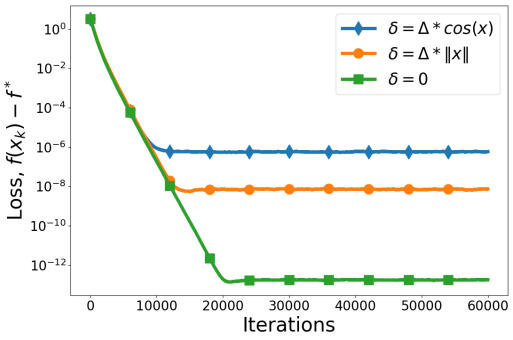


Figure 1. Effect of the adversarial noise concept on the error floor. Here we optimize $f(x)$ with parameters: $d = 128$ (dimensional), $p = 16$ (equations number), $\gamma = 0.01$ (smoothing parameter), $\eta = 0.01$ (fixed step size), $B = 10$ (batch size), $\Delta = 10^{-4}$ (maximum noise level).

where $g(x) = 0$ is a system of $p$ nonlinear equations such that $p \leqslant d$,

$$g(x) = C\sin(x) + D\cos(x) - b,$$

$x \in \mathbb{R}^d$, $C, D \in \mathbb{R}^{p\times d}$ and $b \in \mathbb{R}^p$.

This formulation of the problem also satisfies a wider class of functions, namely functions satisfying the Polyak–Lojasiewicz condition [10, 9]. This class also includes the convex functions considered in this work. We have chosen stochastic gradient descent as the optimization algorithm. Despite the fact that this algorithm is not accelerated, it is shown (see [11]) that unaccelerated algorithms are already optimal in a smooth class of problems satisfying the Polyak–Lojasiewicz condition.

Figure 1 shows the effect of adversarial deterministic noise (noise level $\Delta$) on the error floor. We compare the performance of mini-batch stochastic gradient descent in three settings on deterministic adversarial noise:

- the deterministic adversarial noise $\delta = \Delta\|x\|_2$ proposed in this paper, which satisfies Assumption 4;
- the deterministic adversarial noise $\delta = \Delta\cos(x)$, which satisfies modulus constraint, namely $|\delta(x)| \leqslant \Delta$;
- case where the adversarial noise $\delta = 0$ is a finite mantissa.

It is easy to observe that indeed stochastic gradient descent with adversarial deterministic noise that satisfies Assumption 4 significantly outperforms the same algorithm that is subject to deterministic adversarial noise (with bounded absolute value), thus confirming our theoretical results. It is also worth noting that despite the fact that we did not artificially add noise (the case where $\delta = 0$), the algorithm still converges to the error floor, as if the algorithm was also subject to noise. This phenomenon can be explained by the presence of computational error (finite mantissa). Finally, it is easy to see that in all cases of noise, the algorithm has the same convergence rate, thus showing that adversarial noise is directly related to the error floor.

## §6. Conclusion

This work is devoted to the study of improving the estimation of the maximum noise level for a non-smooth convex stochastic black-box optimization problem. We have introduced the technique of creating gradient-free algorithms using a smoothing scheme via $l_2$ randomization. By assuming that the noise is bounded by Lipschitz continuity, we were able to improve the estimation of the noise level compared to the standard absolute value constraint. We have shown in theory and practice that this advantage carries over to the smooth formulation of the black-box optimization problem.

## References

1. A. Gasnikov, A. Novitskii, V. Novitskii, F. Abdukhakimov, D. Kamzolov, A. Beznosikov, M. Takac, P. Dvurechensky, and B. Gu, *The power of first-order smooth optimization for black-box non-smooth problems*, in: International Conference on Machine Learning (2022), pp. 7241–7265. PMLR.
2. A. Lobanov, B. Alashqar, D. Dvinskikh, and A. Gasnikov, *Gradient-Free Federated Learning Methods with $l_1$ and $l_2$-Randomization for Non-Smooth Convex Stochastic Optimization Problems*, ArXiv preprint arXiv:2211.10783 (2022).
3. A.D. Flaxman, A.T. Kalai, and H.B. McMahan, *Online convex optimization in the bandit setting: gradient descent without a gradient*, in: Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (2005), pp. 385–394.
4. P. Dvurechensky, E. Gorbunov, and A. Gasnikov, *An accelerated directional derivative method for smooth stochastic convex optimization.* — Eur. J. Oper. Res. **290**(2) (2021), pp. 601–621.

5. N. Kornilov, O. Shamir, A. Lobanov, D. Dvinskikh, A. Gasnikov, I. Shibaev, E. Gorbunov, and S. Horváth, *Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance*, in: Advances in Neural Information Processing Systems, vol. 36 (2024).
6. A. Lobanov, *Stochastic adversarial noise in the "black box" optimization problem*, in: International Conference on Optimization and Applications (2023), pp. 60–71. Springer.
7. I.A. Kuruzov, F.S. Stonyakin, and M.S. Alkousa, *Gradient-type methods for optimization problems with Polyak-Łojasiewicz condition: Early stopping and adaptivity to inexactness parameter*, in: International Conference on Optimization and Applications (2022), pp. 18–32. Springer.
8. A. Lobanov, A. Gasnikov, and F. Stonyakin, *Highly smoothness zero-order methods for solving optimization problems under PL condition*, ArXiv preprint arXiv:2305.15828 (2023).
9. S. Lojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in: Les équations aux dérivées partielles, vol. 117 (1963), pp. 87–89.
10. B.T. Polyak, *Gradient methods for the minimisation of functionals.* — USSR Comput. Math. Math. Phys. **3**(4) (1963), pp. 864–878.
11. P. Yue, C. Fang, and Z. Lin, *On the lower bound of minimizing Polyak-Łojasiewicz functions*, in: The Thirty Sixth Annual Conference on Learning Theory (2023), pp. 2948–2968. PMLR.
12. R. Dang-Nhu, G. Singh, P. Bielik, and M. Vechev, *Adversarial attacks on probabilistic autoregressive forecasting models*, in: International Conference on Machine Learning (2020), pp. 2356–2365. PMLR.
13. H. Rosenbrock, *An automatic method for finding the greatest or least value of a function.* — Comput. J. **3**(3) (1960), pp. 175–184.
14. C. Audet and W. Hare, *Derivative-free and blackbox optimization*, Springer (2017).
15. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z.B. Celik, and A. Swami, *Practical black-box attacks against machine learning*, in: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (2017), pp. 506–519.
16. J. Gao, J. Lanchantin, M.L. Soffa, and Y. Qi, *Black-box generation of adversarial text sequences to evade deep learning classifiers*, in: 2018 IEEE Security and Privacy Workshops (SPW) (2018), pp. 50–56.
17. H. Mania, A. Guy, and B. Recht, *Simple random search of static linear policies is competitive for reinforcement learning*, in: Advances in Neural Information Processing Systems, vol. 31 (2018).
18. K.K. Patel, A. Saha, L. Wang, and N. Srebro, *Distributed Online and Bandit Convex Optimization*, in: OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop) (2022).
19. A. Akhavan, E. Chzhen, M. Pontil, and A. Tsybakov, *A gradient estimator via $L_1$-randomization for online zero-order optimization with two point feedback*, in: Advances in Neural Information Processing Systems, vol. 35 (2022), pp. 7685–7696.
20. O. Shamir, *An optimal algorithm for bandit and zero-order convex optimization with two-point feedback.* — J. Mach. Learn. Res. **18**(1) (2017), pp. 1703–1713.

21. T. Lattimore and A. Gyorgy, *Improved regret for zeroth-order stochastic convex bandits*, in: Conference on Learning Theory (2021), pp. 2938–2964. PMLR.
22. A. Nguyen and K. Balasubramanian, *Stochastic Zeroth-Order Functional Constrained Optimization: Oracle Complexity and Applications.* — INFORMS J. Optim. (2022).
23. A. Lobanov, A. Gasnikov, and A. Krasnov, *The Order Oracle: A New Concept in The Black Box Optimization Problems*, ArXiv preprint arXiv:2402.09014 (2024).
24. A. Gasnikov, D. Dvinskikh, P. Dvurechensky, E. Gorbunov, A. Beznosikov, and A. Lobanov, *Randomized gradient-free methods in convex optimization*, ArXiv preprint arXiv:2211.13566 (2022).
25. F. Bach and V. Perchet, *Highly-smooth zero-th order online optimization*, in: Conference on Learning Theory (2016), pp. 257–283. PMLR.
26. A. Akhavan, M. Pontil, and A. Tsybakov, *Exploiting higher order smoothness in derivative-free optimization and continuous bandits*, in: Advances in Neural Information Processing Systems, vol. 33 (2020), pp. 9017–9027.
27. A. Lobanov, N. Bashirov, and A. Gasnikov, *The Black-Box Optimization Problem: Zero-Order Accelerated Stochastic Method via Kernel Approximation*, ArXiv preprint arXiv:2310.02371 (2023).
28. A. Lobanov and A. Gasnikov, *Accelerated zero-order SGD method for solving the black box optimization problem under "overparametrization" condition*, in: International Conference on Optimization and Applications (2023), pp. 72–83. Springer.
29. A. Akhavan, M. Pontil, and A. Tsybakov, *Distributed zero-order optimization under adversarial noise*, in: Advances in Neural Information Processing Systems, vol. 34 (2021), pp. 10209–10220.
30. D. Dvinskikh, V. Tominin, I. Tominin, and A. Gasnikov, *Noisy zeroth-order optimization for non-smooth saddle point problems*, in: International Conference on Mathematical Optimization Theory and Operations Research (2022), pp. 18–33. Springer.
31. A. Lobanov, A. Anikin, A. Gasnikov, A. Gornov, and S. Chukanov, *Zero-order stochastic conditional gradient sliding method for non-smooth convex optimization*, in: International Conference on Mathematical Optimization Theory and Operations Research (2023), pp. 92–106. Springer.
32. N. Kornilov, A. Gasnikov, P. Dvurechensky, and D. Dvinskikh, *Gradient-free methods for non-smooth convex stochastic optimization with heavy-tailed noise on convex compact.* — Comput. Manag. Sci. **20**(1) (2023), p. 37.

Поступило 15 ноября 2024 г.

Moscow Institute of
Physics and Technology, Dolgoprudny, Russia;
Skolkovo Institute of Science and Technology;
ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
*E-mail*: lobanov.av@mipt.ru

Moscow Institute of Physics and Technology, Dolgoprudny, Russia;
Innopolis University, Innopolis, Russia;
Steklov Mathematical Institute of RAS, Moscow, Russia
*E-mail*: gasnikov.av@mipt.ru