IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

# Robust Analysis of Almost Sure Convergence of Zeroth-Order Mirror Descent Algorithm

Anik Kumar Paul, Arun D Mahindrakar and Rachel K Kalaimani

*Abstract*— **This paper presents an almost sure convergence of the zeroth-order mirror descent (ZOMD) algorithm. The algorithm admits non-smooth convex functions and a biased oracle which only provides noisy function value at any desired point. We approximate the subgradient of the objective function using Nesterov's Gaussian Approximation (NGA) with certain alternations suggested by some practical applications. We prove an almost sure convergence of the iterates' function value to the neighbourhood of optimal function value, which can not be made arbitrarily small, a manifestation of a biased oracle. This paper ends with a concentration inequality, which is a finite time analysis that predicts the likelihood that the function value of the iterates is in the neighbourhood of the optimal value at any finite iteration.**

*Index Terms*— **Almost sure convergence, subgradient approximation, mirror descent algorithm**

## I. Introduction

One of the earliest subfields of optimization is derivative-free optimization [1]–[3] or, more specifically zeroth-order optimization. It refers to an optimization problem with an oracle that only provides function value at a desired point and obtaining a subgradient may not be feasible at that point. As a result, we must approximate the function's subgradient from the noisy measurement of function value. Every step in the zeroth-order algorithm is similar to its first-order counterpart (such as gradient descent or mirror descent), except that the function's subgradient must be approximated at every point. There has recently been a surge of interest generated in different variants of zeroth-order optimization, for both convex and non-convex functions [4]–[8], where the subgradient is approximated by NGA [9]. For a full introduction of zeroth-order optimization and its various applications in diverse domains, see [10] (and the references therein).

In this paper, We extend the analysis of zeroth-order optimization focusing on the ZOMD algorithm, where the approximated subgradient established in [9] replaces the subgradient of the convex objective function in standard mirror descent algorithm [11]. Originally, the mirror descent algorithm generalizes the standard gradient descent algorithm in a more general non-Euclidean space [12]. The problem framework and analysis in this work differ significantly from the recent literature [4]–[8]. The main objective of this study is to show the almost sure convergence of the function value of iterates of the ZOMD algorithm to a neighbourhood of optimal value, as compared to the bulk of the literature, which focuses on showing that the expected error in function value converges to the neighbourhood of zero. An almost sure convergence guarantee to a neighbourhood of optimal value is more significant than the convergence in expectation since it describes what happens to the individual trajectory in each iteration. To the best of our knowledge, no prior work on almost sure convergence for zeroth-order optimization has been published. The problem framework in this study differs from most other works in that it includes a biased oracle that delivers only biased measurement of function value (the expectation of noise in the function measurement is non-zero) at any specified point. The motivation to consider "biased oracle" can be found in application of reinforcement learning and financial risk measurement (see [13] and references therein for more details). Furthermore, unlike other publications, we consider that the oracle returns distinct noise values for two different points. Lastly, in addition to showing almost sure convergence, we estimate the likelihood that the function value of the iterates will be in the neighbourhood of optimal value in any finite iteration. This analysis aids in determining the relationship between the convergence of the ZOMD algorithm and the various parameters of the approximated subgradient. The following list summarises the key contribution of this study.

1) We present ZOMD algorithm with a biased oracle. For the biased oracles, we re-evaluate the parameters of the approximated subgradient of the objective function at a specific location, which is calculated using NGA.
2) We prove that, under certain assumptions, the function values of the iterates of ZOMD algorithm almost surely converges to the neighbourhood of optimal function value. This neighbourhood is determined by several parameters, which are explored in this study.
3) Finally, we show that for any confidence level and a given neighbourhood, the function value of the iterate sequence should be in that neighbourhood after some finite iteration with that confidence. This analysis helps us in determining the convergence rate of the algorithm.

## II. Notation and Mathematical Preliminaries

Let $\mathbb{R}$ and $\mathbb{R}^n$ represent the set of real numbers, set of $n$ dimensional real vectors. Let $\|.\|$ denote any norm on $\mathbb{R}^n$. Given a norm $\|.\|$ on $\mathbb{R}^n$, the dual norm of $x \in \mathbb{R}$ is $\|x\|_* := \sup\{\langle x, y\rangle : \|y\| \leq 1, y \in \mathbb{R}^n\}$. $x[i]$ denotes the $i$th component of the vector $x$. $I_n$ is $n \times n$ identity matrix. A random vector $X \sim \mathcal{N}(0_n, I_n)$ denotes a $n$- dimensional

Anik is a Graduate student in the Department of Electrical Engineering, IIT Madras, Chennai-600036, India email: anikpaul42@gmail.com

Arun and Rachel are with the Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai-600036, India (email: arun_dm@iitm.ac.in, rachel@ee.iitm.ac.in)

Copyright ©2023 IEEE

normal random vector with zero-mean and unit standard-deviation. For two random variables $X$ and $Y$, $\sigma(X,Y)$ is the smallest sigma-algebra generated by random variables $X$ and $Y$. We consider $\|.\|_* \le \kappa_1 \|.\|_2$ because of the equivalence of norm in finite dimension.

Let $f:\mathbb{R}^n \to \mathbb{R}$ be a convex function. For $\delta \ge 0$, the vector $g_\delta \in \mathbb{R}^n$ is called a $\delta$-subgradient of $f$ at $x$ if and only if $f(y) \ge f(x) + \langle g_\delta, y-x\rangle - \delta \quad \forall\ y \in \mathbb{R}^n$ [9]. The set of all $\delta$-subgradients at a point $x$ is called the $\delta$-subdifferential of $f$, denoted by $\partial_\delta f(x)$. If $\delta = 0$, we simply write the notation $\partial f(x)$. If $f$ is differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$, gradient of $f$ at $x$. We say $f \in \mathcal{C}^{0,0}$ if $\exists\ L_0 > 0$ such that $\|f(x)-f(y)\| \le L_0\|x-y\|$ and $f \in \mathcal{C}^{1,1}$ if $f$ is continuously differentiable and $\exists\ L_1 > 0$ such that $\|\nabla f(x) - \nabla f(y)\| \le L_1\|x-y\|\ \forall\ x,y \in \mathbb{R}^n$.

If $f$ has directional derivative in all directions, then we can form the Gaussian approximation as follows: $f_\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int\limits_{\mathbb{R}^n} f(x+\mu u)e^{-\frac{1}{2}\|u\|^2}du$, where $\mu > 0$ is any constant. The function $f_\mu$ is differentiable at each $x \in \mathbb{R}^n$ and $\nabla f_\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\mu} \int\limits_{\mathbb{R}^n} uf(x+\mu u)e^{-\frac{1}{2}\|u\|^2}du$. It can also be seen that $\nabla f_\mu(x) \in \partial_\delta f(x)$, where, $\delta = \mu L_0\sqrt{n}$ if $f \in \mathcal{C}^{0,0}$ and $\delta = \frac{\mu^2}{2}L_1 n$ if $f \in \mathcal{C}^{1,1}$ [9].

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space. An event $A \in \mathcal{F}$ is occurred almost surely (a.s.) if $\mathbb{P}(A) = 1$. If $X \sim \mathcal{N}(0_n, I_n)$, it can be shown that $\mathbb{E}[\|X\|_2^p] \le n^{\frac{p}{2}}$ if $p \in [0,2]$ and $\mathbb{E}[\|X\|_2^p] \le (p+n)^{\frac{p}{2}}$ if $p > 2$. We will use the following two Lemmas in our analysis.

*Lemma 1 ( [14]):* Let $\{X_t\}_{t\ge1}$ be a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\ge1}$ such that $\mathbb{E}[\|X_t\|] < \infty$ and $\{\beta(t)\}$ is a non-decreasing sequence of positive numbers such that $\lim\limits_{t\to\infty}\beta(t) = \infty$ and $\sum\limits_{t\ge1}\frac{\mathbb{E}[\|X_t - X_{t-1}\|^2|\mathcal{F}_{t-1}]}{\beta(t)^2} < \infty$, then $\lim\limits_{t\to\infty}\frac{X_t}{\beta(t)} = 0$ a.s.

*Lemma 2:* If $\{X_t, \mathcal{F}_t\}_{t\ge1}$ is a non-negative submartingales, then, for any $\epsilon > 0$ we have $\mathbb{P}(\max\limits_{1\le t\le T} X(t) \ge \epsilon) \le \frac{E[X(T)]}{\epsilon}$.

## III. Problem Statement

Consider the following optimization problem

$$\min_{x\in\mathbb{X}} f(x) \qquad \text{(CP1)}$$

The constraint set $\mathbb{X}$ is a convex and compact subset of $\mathbb{R}^n$ with diameter $D$. The function $f:\mathbb{R}^n \to \mathbb{R}$ is a convex. Define $f^* = \min\limits_{x\in\mathbb{X}} f(x)$ and $\mathbb{X}^* = \{x^* \in \mathbb{X} | f(x^*) = f^*\}$. Observe that $\mathbb{X}^*$ is nonempty due to compactness of the constraint set $\mathbb{X}$ and continuity of $f$. We assume in this paper that we have an oracle which generates a noisy value of the function at a given point $x \in \mathbb{X}$. That is, at each point $x \in \mathbb{X}$, we have only the information $\hat{f}(x) = f(x) + e(x,\omega)$, where $e(x,\omega): \mathbb{R}^n \times \Omega \to \mathbb{R}$ is a random variable for each $x \in \mathbb{X}$ satisfying

$$\begin{aligned} &\mathbb{E}[e(x,\omega)] = b(x) \ \text{ with } \ \|b(x)\|_* \le B \\ &\text{and } \ \mathbb{E}[\|e(x,\omega)\|^2] \le \mathrm{V}^2 \end{aligned} \qquad (1)$$

where, $B$ is a non-negative constant, and $V$ can be any constant.

*Remark 1:* In the context of zeroth-order stochastic optimization problem [4]–[6], the objective is to solve the optimization problem: $\min\limits_{x\in\mathbb{X}} f(x) = \mathbb{E}[F(x,\omega)]$ and the oracle only provides $F(x,\omega)$ at any desired $x \in \mathbb{R}^n$. In such a situation, it is straightforward to verify that $\mathbb{E}[e(x,\omega)] = 0$, implying that $B = 0$. The assumption of positive $B$ makes the problem more generic than previous recent studies. In a broader sense, if $B = 0$, we call it an *unbiased oracle*. However, $B$ is non-zero in many applications (see [13] and references therein for further details), therefore the problem in this study is more general than in other recent works due to the presence of positive $B$. For sake of brevity, we henceforth use $e(x)$ to denote $e(x,\omega)$. In the next section, we discuss the zeroth-order mirror descent algorithm.

## IV. Zeroth-Order Mirror Descent Algorithm

Mirror descent algorithm is a generalization of standard subgradient method where the Euclidean norm is replaced with a more general Bergman divergence as a proximal function. Let $R$ be the $\sigma_R$-strongly convex function and differentiable over an open set that contains the set $\mathbb{X}$. The Bergman divergence $\mathbb{D}_R(x,y): \mathbb{X}\times\mathbb{X} \to \mathbb{R}$ is $\mathbb{D}_R(x,y) := R(x) - R(y) - \langle \nabla R(y), x-y\rangle\ \ \forall\ x,y \in \mathbb{X}$. It is clear from the definition of strong convexity that

$\mathbb{D}_R(x,y) \ge \frac{\sigma_R}{2}\|x-y\|^2$ and $\mathbb{D}_R(z,y) - \mathbb{D}_R(z,x) - \mathbb{D}_R(x,y) = \langle \nabla R(x) - \nabla R(y), z-x\rangle \quad \forall\ \ x,y,z \in \mathbb{X}$. We outline the steps of the mirror descent algorithm.

At iteration $t$, let $x_t$ be the iterates of the ZOMD algorithm. We approximate the subgradient of function $f(x)$ at $x = x_t$ as follows. We generate a normal random vector $u_t \sim \mathcal{N}(0_n, I_n)$. We use the zeroth-order oracles to get the noisy function values ($\hat{f}$) at two distinct values, that is,

$\hat{f}(x_t + \mu u_t) = f(x_t + \mu u_t) + e(x_t + \mu u_t, \omega_t^1)$ and

$\hat{f}(x_t) = f(x_t) + e(x_t, \omega_t^2)$. Note that $\omega_t^1$ and $\omega_t^2$ are two independent realizations from the sample space $\Omega$ according to the probability law $\mathbb{P}$. Hence, we approximate the subgradient of $f$ at $x = x_t$, denoted by $\tilde{g}(t)$ as $\tilde{g}(t) = \frac{\hat{f}(x_t+\mu u_t) - \hat{f}(x_t)}{\mu}u_t$. The next iterate $x_{t+1}$ is calculated as follows:

$$x_{t+1} = \arg\min_{x\in\mathbb{X}}\{\langle \tilde{g}(t), x - x_t\rangle\} + \frac{1}{\alpha(t)}\mathbb{D}_R(x,x_t))\} \qquad (2)$$

where, $\alpha(t)$ is the step-size of the algorithm. To show almost sure convergence, we consider weighted averaging akin to the recent work [15] in first-order algorithm as $z_t = \frac{\sum_{j=1}^{t}\alpha(j)x_j}{\sum_{k=1}^{t}\alpha(k)}$. The Bergman divergence should be chosen in such a way that a closed form solution to (2) is available [16].

*Assumption 1:* The step-size $\alpha(t)$ is a decreasing sequence which satisfies $\sum\limits_{t=1}^{\infty}\alpha(t) = \infty$ and $\sum\limits_{t=1}^{\infty}\alpha(t)^2 < \infty$.

From Assumption 1, we can conclude that $\lim\limits_{t\to\infty}\alpha(t) = 0$.

*Assumption 2:* Let the following hold.

1) The generating random vectors $u_t \in \mathbb{R}^n (\forall\ t \in \mathbb{N})$ are mutually independent and normally distributed and for each $t \in \mathbb{N}$ $u_t$ is independent of $x_t$.
2) The random variables $e(x_t, .) : \Omega \to \mathbb{R}$ and $e(x_t + \mu u_t, .) : \Omega \to \mathbb{R}$ ( $\forall\ t \in \mathbb{N}$) are mutually independent and identically distributed in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
3) The random variables $e(x_t + \mu u_t)$ and $e(x_t)$ are independent of $x_t$ and $u_t$.

Using Assumption 2 and (1), we can write $\mathbb{E}[e(x_t + \mu u_t)|\sigma\{x_t, u_t\}] = b(x_t + \mu u_t)$ and $\mathbb{E}[e(x_t)|\sigma\{x_t\}] = b(x_t)$, where, $\|b(x_t + \mu u_t)\|_* $ and $\|b(x_t)\|_* \leq B$ a.s. Similarly, $\mathbb{E}[\|e(x_t + \mu u_t)\|^2 |\sigma\{x_t, u_t\}] \leq \mathrm{V}^2$ and $\mathbb{E}[\|e(x_t)\|^2 |\sigma\{x_t\}] \leq \mathrm{V}^2$ a.s. For an unbiased oracle $B = 0$.

*Remark 2:* Note that, most recent literature on zeroth-order stochastic optimization (e.g. [4]–[9], [17], [18]) computes function values at two separate points $x_t$ and $x_t + \mu u_t$ under the assumption that the stochastic parameters $e(x_t)$ and $e(x_t + \mu u_t)$ are the same. For many applications, this is rather a stringent assumption. In this paper, we avoid such an assumption, which in turn leads to significant deviation in the properties of approximated subgradient and the pertinent properties will be discussed in the ensuing section.

## V. Main Result

In this section we discuss the properties of approximated subgradient, almost sure convergence and the finite time analysis. Before proceeding further, first define $\mathcal{F}_t = \sigma\{x_l | 1 \leq l \leq t\}\ \forall\ t \in \mathbb{N}$. Hence we get a filtration such as $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_t$. Observe that $\tilde{g}(t-1)$ is $\mathcal{F}_t$ measurable in view of (2) and also the Bergman divergence $\mathbb{D}_R(x, x_t)$ ($\forall\ \ x \in \mathbb{X}$) is $\mathcal{F}_t$ measurable. Define another filtration as $\{\mathcal{G}_t\}_{t \geq 1}$ such that $\mathcal{G}_{t-1} = \mathcal{F}_t$, which will be helpful in the subsequent analysis.

### *A. Properties of Approximated Subgradient*

The analysis in this subsection borrows some steps from [9]. However, our analysis contains significant deviations, most notably, the result concerning the properties of approximated subgradient, which is derived using the noisy information of the function value.

*Lemma 3:* $\mathbb{E}[\tilde{g}(t)|\mathcal{F}_t] = \nabla f_\mu(x_t) + \mathrm{B}(t)$ a.s. where, $\mathrm{B}(t)$ satisfies $\|\mathrm{B}(t)\|_* \leq \frac{2\kappa_1 \mathrm{B}}{\mu}\sqrt{n}$ a.s and we have (a.s.)
$\mathbb{E}[\|\tilde{g}(t)\|_*^2 |\mathcal{F}_t]$

$$\leq \begin{cases} \kappa_1^2(2L_0^2(n+4)^2 + 8\left(\frac{\mathrm{V}}{\mu}\right)^2 n) & \text{if } f \in \mathcal{C}^{0,0} \\ \kappa_1^2(\frac{3}{4}L_1\mu^2(n+6)^3 + 3G^2(n+4)^2 + 12\frac{\mathrm{V}^2}{\mu^2}n) & \text{if } f \in \mathcal{C}^{1,1}. \end{cases}$$

*Proof:* Consider the $\sigma$-algebra $\mathcal{H}_t$ defined as $\mathcal{H}_t = \sigma(\{x_k\}_{k=1}^t, u_t)$. Consider the term

$$\begin{aligned} &\mathbb{E}[\tilde{g}(t)|\mathcal{H}_t] \\ =&\nabla f_\mu(x_t) + \mathbb{E}\Big[\frac{e(x_t + \mu u_t) - e(x_t)}{\mu} u_t | \sigma(x_t, u_t)\Big] \text{ a.s.} \end{aligned} \tag{3}$$

Note that because of Assumption 2, $\mathbb{E}\Big[\frac{f(x_t + \mu u_t) - f(x_t)}{\mu} u_t | \mathcal{H}_t\Big] = \mathbb{E}\Big[\frac{f(x_t+\mu u_t) - f(x_t)}{\mu} u_t | \sigma(x_t, u_t)\Big]$ a.s. and

$$\begin{aligned} &\left\|\mathbb{E}\Big[\frac{e(x_t + \mu u_t) - e(x_t)}{\mu} u_t | \sigma(x_t, u_t)\Big]\right\|_* \\ \leq &\left\|\frac{b(x_t + \mu u_t) - b(x_t)}{\mu}\right\|_* \|u_t\|_* \leq \frac{2B}{\mu}\|u_t\|_* \text{ a.s.} \end{aligned}$$

Observe that $\mathcal{F}_t \subseteq \mathcal{H}_t$ and hence by using the Towering property we get $\mathbb{E}[\tilde{g}(t)|\mathcal{F}_t] = \mathbb{E}[\mathbb{E}[\tilde{g}(t)|\mathcal{H}_t]|\mathcal{F}_t] = \nabla f_\mu(x_t) + \mathrm{B}(t)$, Where, $\mathrm{B}(t) = \mathbb{E}\Big[\mathbb{E}\Big[\frac{e(x_t+\mu u_t) - e(x_t)}{\mu} u_t | \sigma(x_t, u_t)\Big] | \mathcal{F}_t\Big]$ satisfies $\|\mathrm{B}(t)\|_* \leq \frac{2\mathrm{B}\kappa_1}{\mu}\mathbb{E}[\|u_t\|_2 |\mathcal{F}_t] \leq \frac{2\mathrm{B}\kappa_1}{\mu}\sqrt{n}$ a.s.

Consider the term $\left\|\frac{f(x_t+\mu u_t) + e(x_t+\mu u_t) - f(x_t) - e(x_t)}{\mu} u_t\right\|_*^2 \leq 2\kappa_1^2 \left\|\frac{f(x_t+\mu u_t) - f(x_t)}{\mu} u_t\right\|_2^2 + 2\kappa_1^2 \left\|\frac{e(x_t+\mu u_t) - e(x_t)}{\mu} . u_t\right\|_2^2$.
Applying the definition of $\mathcal{C}^{0,0}$, we have

$$\begin{aligned} &\left\|\frac{f(x_t + \mu u_t) + e(x_t + \mu u_t) - f(x_t) - e(x_t)}{\mu} . u_t\right\|_*^2 \\ \leq &2\kappa_1^2 L_0^2 \|u_t\|_2^4 + 2\kappa_1^2 \left\|\frac{e(x_t + \mu u_t) - e(x_t)}{\mu} . u_t\right\|_2^2. \end{aligned} \tag{4}$$

Consider the term $\mathbb{E}\Big[\left\|\frac{e(x_t+\mu u_t) - e(x_t)}{\mu} u_t\right\|_2^2 | \mathcal{H}_t\Big]$

$$\begin{aligned} =&\mathbb{E}\Big[\left\|\frac{e(x_t + \mu u_t) - e(x_t)}{\mu} u_t\right\|_2^2 | \sigma(x_t, u_t)\Big] \\ \leq &\frac{2}{\mu^2}\Big(\mathbb{E}[(e(x_t + \mu u_t))^2 \|u_t\|_2^2 + e(x_t)^2 \|u_t\|_2^2 | \sigma(x_t, u_t)]\Big) \\ \leq &\frac{4\mathrm{V}^2}{\mu^2}\|u_t\|^2 \text{ a.s.} \end{aligned}$$

Hence, by applying Towering property in (4), we get the result. For $f \in \mathcal{C}^{1,1}$, $\left\|\frac{f(x_t+\mu u_t) + e(x_t+\mu u_t) - f(x_t) - e(x_t)}{\mu} u_t\right\|_*^2$

$$\begin{aligned} \leq &3\kappa_1^2 \left\|\frac{f(x_t + \mu u_t) - f(x_t) - \mu\langle\nabla f(x_t), u_t\rangle}{\mu} u_t\right\|_2^2 \\ &+ 3\kappa_1^2 \|\nabla f(x_t)\|_2^2 \|u_t\|_2^4 + 3\kappa_1^2 \left\|\frac{e(x_t + \mu u_t) - e(x_t)}{\mu} u_t\right\|_2^2 \end{aligned} \tag{5}$$

Note that $\left\|\frac{f(x_t+\mu u_t) - f(x_t) - \mu\langle\nabla f(x_t), u_t\rangle}{\mu} u_t\right\|_2^2 \leq \frac{L_1^2\mu^2}{4}\|u_t\|_2^6$ because of the definition of $\mathcal{C}^{1,1}$. Taking conditional expectation on (5) we get the result. ■

Using the similar procedure we can extend the analysis for $f \in \mathcal{C}^{2,2}$ and so on. It is important to note that because of consideration of more generic framework $\mathbb{E}[\|\tilde{g}(t)\|_*^2] = \mathcal{O}(\frac{1}{\mu^2})$ for small values of $\mu$, as opposed to [9] because of consideration of more general framework. This result plays a significant role in the subsequent discussion of this paper.

*Corollary 1:* For unbiased oracle, $\mathbb{E}[\tilde{g}_t|\mathcal{F}_t] = \nabla f_\mu(x_t)$ a.s.

### B. Almost Sure Convergence of the ZOMD Algorithm

Based on the discussion in Lemma 3, we redefine properties of biased subgradient as follows : $\mathbb{E}[\tilde{g}(t)|\mathcal{F}_t] = g_\delta(t) + \mathrm{B}(t)$, where, $g_\delta(t) \in \partial_\delta f(x)$ at $x = x_t$ and $\mathrm{B}(t)$ is $\mathcal{F}_t$ measurable and $\|\mathrm{B}(t)\|_* \leq B_1$ a.s. Then there exists a random vector $\zeta(t)$ such that (a.s.) $\tilde{g}(t) = g_\delta(t) + \mathrm{B}(t) + \zeta(t)$. Moreover, $\mathbb{E}[\zeta(t)|\mathcal{F}_t] = 0$ and $\mathbb{E}[\|\tilde{g}(t)\|_*^2 |\mathcal{F}_t] \leq \mathrm{K}$ a.s. Note that we can get an expression of $\delta$, $B_1$ and $K$ from Lemma 3 depending on the properties of the noise and the smoothness of $f$.

*Theorem 1:* Under Assumptions 1 and 2 and $\forall\, \epsilon > 0$, for the iterate sequence generated by ZOMD algorithm $\{x_t\}$, $\exists$ $\{x_{t_k}\} \subseteq \{x_t\}$ such that $f(x_{t_k}) - f^* \leq \delta + B_1 D + \epsilon$ a.s. For the iterate sequence $\{z_t\}$, $\exists\, t_0 \in \mathbb{N}$ such that $\forall\, t \geq t_0$ we have $f(z_t) - f^* \leq \delta + B_1 D + \epsilon$ a.s.

Before proving the Theorem 1, we need the following three Lemmas which we discuss here.

*Lemma 4:* $\sum\limits_{t\geq 1} \frac{\alpha(t)^2}{2\sigma_R} \|\tilde{g}(t)\|_*^2 < \infty$. a.s., where, $\sigma_R$ is the strong convexity parameter of the function $R$.

*Proof:* $\lim\limits_{t\to\infty} \mathbb{E}\Big[\sum\limits_{k=1}^{t} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2\Big] \leq \sum\limits_{t\geq 1} \frac{\alpha(t)^2}{2\sigma_R}\mathrm{K} < \infty$

By applying Fatou's Lemma we get

$$\mathbb{E}[\liminf_{t\to\infty} \sum_{k=1}^{t} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2] \leq \liminf_{t\to\infty} \mathbb{E}[\sum_{k=1}^{t} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2]$$

$< \infty$. Hence we can say $\sum\limits_{t\geq 1} \frac{\alpha(t)^2}{2\sigma_R} \|\tilde{g}(t)\|_*^2 < \infty$ a.s. ∎

*Lemma 5:* $\exists\, C > 0$ such that $\mathbb{E}[\|\zeta(t)\|_*^2 |\mathcal{F}_t] < C$ a.s.

*Proof:* From the definition of $\zeta(t)$ we get that

$$\|\zeta(t)\|_*^2 \leq 3\kappa_1^2(\|\tilde{g}(t)\|_2^2 + \|\mathrm{B}(t)\|_2^2 + \|g_\delta(t)\|_2^2). \quad (6)$$

Notice that $\exists\, K_1 > 0$ such that $\|g_\delta(t)\| \leq K_1\ \forall\, t$ because of compactness of $\mathbb{X}$. Taking expectation on both sides of (6), we get (a.s.) $\mathbb{E}[\|\zeta(t)\|_*^2 |\mathcal{F}_t] \leq 3\kappa_1^2(\mathrm{K} + B_1^2 + K_1) \triangleq C$. ∎

*Lemma 6:*

$$\frac{\sum\limits_{t\geq 1} \alpha(t) \langle \zeta(t), x - x_t\rangle}{\sum\limits_{t\geq 1} \alpha(t)} = 0 \ \text{ a.s. } \forall\ x \in \mathbb{X}.$$

*Proof:* Define $X(t) = \sum\limits_{k=1}^{t} \alpha(k) \langle \zeta(k), x - x_k\rangle$. In the light of definition of $\zeta(t)$ and since $X(t)$ is $\mathcal{F}_t$ measurable we get that $\mathbb{E}[X(t)|\mathcal{F}_t] = X(t-1)$. Hence $\{X(t), \mathcal{G}_t\}$ is a martingale. On the other hand, it can be seen that (a.s.)

$$\sum_{t\geq 1} \frac{\mathbb{E}[\|X(t) - X(t-1)\|^2 |\mathcal{F}_t]}{(\sum\limits_{k=1}^{t} \alpha(k))^2} \leq$$

$$\sum_{t\geq 1} \frac{\mathbb{E}[\alpha(t)^2 \|\zeta(t)\|_*^2 \|x - x_t\|^2 |\mathcal{F}_t]}{(\sum\limits_{k=1}^{t} \alpha(k))^2} \leq \sum_{t\geq 1} \frac{\alpha(t)^2 D^2 C}{(\sum\limits_{k=1}^{t} \alpha(t))^2} < \infty.$$

The last line is because of Lemma 5 and the diameter of the compact set $\mathbb{X}$. Hence by applying Lemma 1, the result follows. ∎

Now we are in a position to prove the main result.

*Proof:* The first-order optimality condition to (2) yields

$$\alpha(t) \langle \tilde{g}(t), x - x_{t+1}\rangle \geq -\langle \nabla R(x_{t+1}) - \nabla R(x_t), x - x_{t+1}\rangle$$
$$\geq \mathbb{D}_R(x_{t+1}, x_t) + \mathbb{D}_R(x, x_{t+1}) - \mathbb{D}_R(x, x_t). \quad (7)$$

The last inequality in (7) is because of the definition of Bergman Divergence. From the LHS of (7), we obtain

$$\alpha(t) \langle \tilde{g}(t), x - x_{t+1}\rangle = \alpha(t) \langle \tilde{g}(t), x - x_t + x_t - x_{t+1}\rangle$$
$$\leq \alpha(t) \langle \tilde{g}(t), x - x_t\rangle + \frac{\alpha(t)^2}{2\sigma_R} \|\tilde{g}(t)\|_*^2 + \frac{\sigma_R}{2} \|x_t - x_{t+1}\|^2.$$

The last inequality follows by applying the Young-Fenchel inequality to the term $\alpha(t) \langle \tilde{g}(t), x_t - x_{t+1}\rangle$. Hence from (7), we get that $\quad \mathbb{D}_R(x, x_{t+1})$

$$\leq \mathbb{D}_R(x, x_t) + \alpha(t) \langle \tilde{g}(t), x - x_t\rangle + \frac{\alpha(t)^2}{2\sigma_R} \|\tilde{g}(t)\|_*^2. \quad (8)$$

Notice that $\mathbb{D}_R(x_{t+1}, x_t) \geq \frac{\sigma_R}{2} \|x_{t+1} - x_t\|^2$. Consider the term

$$\alpha(t) \langle \tilde{g}(t), x - x_t\rangle = \alpha(t) \langle g_\delta(x_t) + \mathrm{B}(t) + \zeta(t), x - x_t\rangle$$

$$\leq \alpha(t)(f(x) - f(x_t) + \delta + B_1 D + \langle \zeta(t), x - x_t\rangle). \quad (9)$$

The last inequality in (9) is because of $\delta$-subgradient of function $f$ and the generalized Cauchy-Schwartz inequality, where, D is the diameter of the constraint set. Plugging (9) into (8) and on applying telescopic sum from $k = 1$ to $t$ we get

$$\mathbb{D}_R(x^*, x_{t+1}) \leq \mathbb{D}_R(x^*, x_1) + \sum_{k=1}^{t} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2 +$$
$$\sum_{k=1}^{t} \alpha(k)\Big(f^* - f(x_k) + \delta + B_1 D + \langle \zeta(k), x^* - x_k\rangle\Big). \quad (10)$$

Let $\epsilon > 0$ and define the sequence of stopping times $\{T_p\}_{p\geq 1}$ and $\{T^p\}_{p\geq 1}$ as follows:

$$T_1 = \inf\{f(x_t) - f^* \geq \delta + B_1 D + \epsilon\}$$
$$T^1 = \inf\{t \geq T_1 | f(x_t) - f^* < \delta + B_1 D + \epsilon\}$$
$$\vdots$$
$$T^p = \inf\{t \geq T_p | f(x_t) - f^* < \delta + B_1 D + \epsilon\}$$
$$T_{p+1} = \inf\{t \geq T^p | f(x_t) - f^* \geq \delta + B_1 D + \epsilon\ \}.$$

If $\exists\, p \in \mathbb{N}$ such that infimum does not exist, we assume that $T_p = \infty$ or $T^p = \infty$ .

Claim-1 - If $T_p < \infty$, then $T^p < \infty$ a.s. $\forall\ p \in \mathbb{N}$. Suppose, ad absurdum, $\exists\, p_0 \in \mathbb{N}$ such that $T_{p_0} < \infty$ but $T^{p_0} = \infty$ with probability (w.p.) $\eta$. Let $T_{p_0} = t_0$, then it implies that $\forall\, t \geq t_0$, $f(x_t) - f^* \geq \delta + B_1 D + \epsilon$ w.p. $\eta$. From (10), we deduce that $\forall\, t \geq t_0$ (w.p. $\eta$)

$$\mathbb{D}_R(x^*, x_{t+1}) \leq \mathbb{D}_R(x^*, x_{t_0}) + \sum_{k=t_{t_0}}^{t} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2.$$
$$+ \sum_{k=t_0}^{t} \alpha(k)\Big(-\epsilon + \langle \zeta(k), x^* - x_k\rangle\Big) \quad (11)$$

Let $t \to \infty$. Notice that in view of Lemma 6, $\sum_{k \geq t_0} \alpha(k)(-\epsilon + \langle \zeta(k), x^* - x_k \rangle) = -\infty$ and also in view of Lemma 4 $\sum_{k \geq t_0} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2 < \infty$ a.s. Hence, from (11) we get $\limsup_{t \to \infty} \mathbb{D}_R(x^*, x_t) = -\infty$ w.p. atleast $\eta$. But $\mathbb{D}_R(x^*, x_t) \geq 0$ because of strong convexity of $R$ which implies $\eta = 0$. Thus, $T^{p_0} < \infty$ a.s. This establishes Claim-1. Hence $\exists \{x_{t_k}\} \subseteq \{x_t\}$ such that $f(x(t_k)) - f^* \leq \delta + B_1 D + \epsilon$ a.s.

From the definition of convexity of $f$ we get that $\sum_{k=1}^{t} \alpha(k) f(z_t) \leq \sum_{j=1}^{t} \alpha(j) f(x_j)$. Hence, from (10) we get that

$$\mathbb{D}_R(x^*, x_{t+1}) \leq \mathbb{D}_R(x^*, x_1) + \sum_{k=1}^{t} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2 + \sum_{k=1}^{t} \alpha(k)\Big(f^* - f(z_t) + \delta + B_1 D + \langle \zeta(k), x^* - x_k \rangle\Big). \tag{12}$$

In a similar fashion, define the sequence of stopping times $\{\bar{T}_p\}_{p \geq 1}$ and $\{\bar{T}^p\}_{p \geq 1}$ as follows:

$$\begin{aligned}
&\bar{T}_1 = \inf\{f(z_t) - f^* \geq \delta + B_1 D + \epsilon\} \\
&\bar{T}^1 = \inf\{t \geq \bar{T}_1 | f(z_t) - f^* < \delta + B_1 D + \epsilon\} \\
&\quad \vdots \\
&\bar{T}^p = \inf\{t \geq T_p | f(z_t) - f^* < \delta + B_1 D + \epsilon\} \\
&\bar{T}_{p+1} = \inf\{t \geq T^p | f(z_t) - f^* \geq \delta + B_1 D + \epsilon\ \}.
\end{aligned}$$

If $\bar{T}_p < \infty$ then $\bar{T}^p < \infty$ a.s. The reason is similar to the proof of Claim-1.

Claim- 2: $\exists\, p_0 \in \mathbb{N}$ such that $\bar{T}_{p_0} = \infty$ a.s. If this claim is true, it proves the second part of the Theorem.

Otherwise, $\forall\, t_1 \in \mathbb{N}$, $\exists\, t > t_1$ such that $(f(z_t) - f^*) \geq \delta + B_1 D + \epsilon$ with some probability $\eta$. Hence, from (12) we get that (11) holds for that $t$. Letting $t \to \infty$ and using similar arguments we get $\liminf_{t \to \infty} \mathbb{D}_R(x^*, x_t) = -\infty$ w.p. atleast $\eta$, that means $\eta = 0$. Hence, the Claim-2 holds. ■

*Corollary 2 (ZOMD with unbiased and biased oracle):* For unbiased oracle, $\forall\, \epsilon > 0$, $\exists\, t_0 \in \mathbb{N}$ such that $\forall\, t \geq t_0$

$$f(z_t) - f^* \leq \begin{cases} \mu L_0 \sqrt{n} + \epsilon & \text{if } f \in \mathcal{C}^{0,0} \\ \frac{\mu^2}{2} L_1 n + \epsilon & \text{if } f \in \mathcal{C}^{1,1}. \end{cases} \quad \text{a.s.}$$

and for biased oracle, $\forall\, \epsilon > 0\ \exists\, t_0 \in \mathbb{N}$ such that $\forall\, t \geq t_0$ the following holds

$$f(z_t) - f^* \leq \begin{cases} \mu L_0 \sqrt{n} + \frac{2\kappa_1 B}{\mu} \sqrt{n} D + \epsilon & \text{if } f \in \mathcal{C}^{0,0} \\ \frac{\mu^2}{2} L_1 n + \frac{2\kappa_1 B}{\mu} \sqrt{n} D + \epsilon & \text{if } f \in \mathcal{C}^{1,1}. \end{cases} \quad \text{a.s.}$$

For unbiased oracle, by selecting a very small value of $\mu$, we can show function value of the iterate sequence converges to arbitrarily close neighbourhood around optimal value. However, this is not the case with biased oracle as suggested by Corollary 2.

### *C. Concentration Bound - Finite Time Analysis*

In the next Theorem, we discuss the adverse affect of choosing a very small $\mu$ for subgradient approximation. Gaussian random variable.

*Theorem 2:* Consider any $t_0 \in \mathbb{N}$ such that $\sum_{k=1}^{t_0} \alpha(k) \geq \frac{3}{\epsilon} D$. Then $\forall\ t \geq t_0$, $\mathbb{P}(f(z_t) - f^* \geq \delta + B_1 D + \epsilon) \leq \frac{3\mathrm{K}}{\epsilon} \frac{\sum_{k=1}^{t} \alpha(k)^2}{\sum_{k=1}^{t} \alpha(k)} + \frac{9CD^2 \sum_{k=1}^{t} \alpha(k)^2}{\epsilon^2 (\sum_{k=1}^{t} \alpha(k))^2}$.

*Proof:* The first order optimality condition to (2) yields (as we have proved in Theorem 1) $f(z_t) - f^* \leq \delta + B_1 D + \frac{\mathbb{D}_R(x^*, x_1)}{\sum_{k=1}^{t} \alpha(k)} + \frac{\sum_{k=1}^{t} \alpha(k) \langle \zeta(k), x^* - x_k \rangle}{\sum_{k=1}^{t} \alpha(k)} + \frac{\sum_{k=1}^{t} \alpha(k)^2 \|\tilde{g}(k)\|_*^2}{2\sigma_R \sum_{k=1}^{t} \alpha(k)}$.

Define $X(t) = \sum_{k=1}^{t} \alpha(k) \langle \zeta(k), x^* - x_k \rangle$ and $Y(t) = \sum_{k=1}^{t} \frac{\alpha(k)^2}{2\sigma_R} \|\tilde{g}(k)\|_*^2$. Choose a $t_0$ such that $\mathbb{D}_R(x^*, x_1) \leq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)\ \forall\, t \geq t_0$ and in view of Assumption 1, $t_0 < \infty$ . Consider any $t > t_0$ and from if $f(z_t) - f^* \geq B_1 D + \delta + \epsilon$ then atleast one of the following holds.
$X(t) \geq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)$ or, $Y(t) \geq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)$. That implies that $\forall\, t \geq t_0$; $\mathbb{P}(f(z_t) - f^* \geq \delta + B_1 D + \epsilon)$

$$\leq \mathbb{P}(X(t) \geq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)) + \mathbb{P}(Y(t) \geq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)). \tag{13}$$

It can be seen from the definition of $\tilde{g}(t)$ that $\mathbb{E}[Y(t)|\mathcal{F}_t] = Y(t-1) + \frac{\alpha(t)^2}{2\sigma_R} \|\tilde{g}(t)\|_*^2 \geq Y(t-1)$. Hence, $\{Y(t), \mathcal{G}_t\}$ is a non-negative sub-martingale.

Note that
$\mathbb{P}(Y(t) \geq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)) \leq \mathbb{P}(\max_{1 \leq j \leq t} Y(j) \geq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)))$.
Hence, by applying Lemma 2 we arrive at

$$\mathbb{P}(Y(t) \geq \frac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)) \leq \frac{3}{\epsilon} \frac{E[Y(t)]}{\sum_{k=1}^{t} \alpha(k)} \leq \frac{3\mathrm{K}}{\epsilon} \frac{\sum_{k=1}^{t} \alpha(k)^2}{\sum_{k=1}^{t} \alpha(k)}. \tag{14}$$

It has already been shown in the proof of Lemma 6 that $\{X(t), \mathcal{G}_t\}$ is a martingale, which implies $\{\|X(t)\|^2, \mathcal{G}_t\}$ is a sub-martingale. It can be seen that

$\mathbb{E}[X(t)^2] = \sum_{k=1}^{t} \mathbb{E}[\alpha(k)^2 \langle \zeta(k), x^* - x_k \rangle^2]$.

Noting that, $k < l$ then $\mathbb{E}[\langle \zeta(k), x^* - x_k \rangle . \langle \zeta(l), x^* - x_l \rangle] = \mathbb{E}[\mathbb{E}[\langle \zeta(k), x^* - x_k \rangle . \langle \zeta(l), x^* - x_l \rangle | \mathcal{F}_l]] = 0$. Hence, by applying Generalized Cauchy-Schwartz inequality we get $\mathbb{E}[X(t)^2] \leq CD^2 \sum_{k=1}^{t} \alpha(k)^2$. The last inequality is because of Lemma 5 and the diameter of the constraint set $\mathbb{X}$. In a similar fashion, by applying Lemma 2, we get

$$\mathbb{P}(X(t) \geq \tfrac{\epsilon}{3} \sum_{k=1}^{t} \alpha(k)) \leq \mathbb{P}(\|X(t)\|^2 \geq \tfrac{\epsilon^2}{9} (\sum_{k=1}^{t} \alpha(k))^2)$$

$$\leq \frac{9}{\epsilon^2} \frac{\mathbb{E}\left\|X(t)\right\|^2}{(\sum_{k=1}^{t} \alpha(k))^2} \leq \frac{9CD^2}{\epsilon^2} \frac{\sum_{k=1}^{t} \alpha(k)^2}{(\sum_{k=1}^{t} \alpha(k))^2}. \tag{15}$$

Hence, by plugging (14) and (15) into (13), we get Theorem 2. ∎

*Remark 3:* Notice that both K and $C$ are $\mathcal{O}(\frac{1}{\mu^2})$ from Lemma 3, this implies that an arbitrary small $\mu$ makes the convergence of the function value to the neighbourhood of the optimal solution slower. Hence, there is a trade-off between accuracy of the convergence to the optimal value and convergence speed of the algorithm in the choice of $\mu$. In the next Corollary, we capture this in detail.

*Corollary 3:* For any $\epsilon > 0$ and a confidence level $0 < p < 1$, let $p_1 = 1-p$. Define $t_1$ such that $\forall\, t \geq t_1$ $\sum_{k=1}^{t} \alpha(k) \geq \frac{6\mathrm{K}}{\epsilon p_1} \sum_{k=1}^{t} \alpha(k)^2$ and $(\sum_{k=1}^{t} \alpha(k))^2 \geq \frac{18CD^2}{\epsilon^2 p_1} \sum_{k=1}^{t} \alpha(k)^2$. Then $\forall\, t \geq \max\{t_0, t_1\}$ we obtain
$\mathbb{P}(f(z_t) - f^* < \delta + B_1 D + \epsilon) \geq p$. Where, $t_0$ is defined in Theorem 2. Notice that $t_1 < \infty$ due to Assumption 1.

## VI. Numerical Simulation

We illustrate the result on ZOMD algorithm using the standard nonsmooth test problem [9] defined as follows

$$\min_{x \in \mathbb{X}} \{|x[1] - \tfrac{1}{10}| + \sum_{i=1}^{9} |\tfrac{1}{10} + x[i+1] - 2x[i]|\}$$

where, $\mathbb{X} = \{x \in \mathbb{R}^{10} | \sum_{i=1}^{10} x[i] = 1, x[i] \geq 0\}$. The subgradient of the objective function at a desired point is approximated from the noisy function value (both biased and unbiased) at that point. The commonly used Bergman Divergence for the constraint set $\mathbb{X}$ is KL divergence defined as $\mathbb{D}_R(x,y) = \sum_{j=1}^{10} x[j] \log(\frac{x[j]}{y[j]})$. For the choice of Bergman divergence, the closed form solution [19] of (2) is as follows : $x_{(t+1)}[k] = \frac{x_t[k] \exp(-\alpha(t)\tilde{g}(t)[k])}{\sum_{l=1}^{10} x_t[l] \exp(-\alpha(t)\tilde{g}(t)[l])} \qquad k \in [10].$

The performance of the ZOMD algorithm is evaluated through the following metric: $Oe(t) \triangleq f(z_t) - f^*$. The performance of the ZOMD algorithm for both unbiased and biased oracle is illustrated in the Figures 1 and 2.

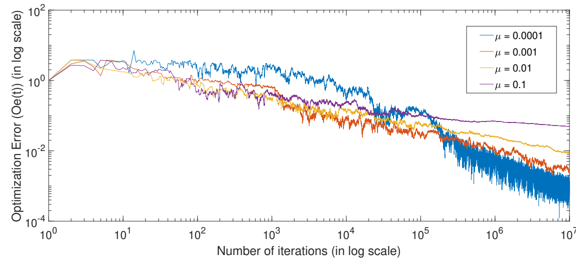


Fig. 1: Performance for Unbiased Oracle.

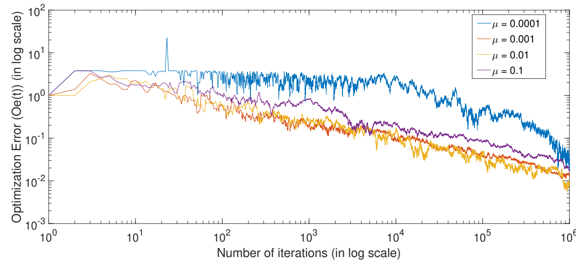


Fig. 2: Performance for Biased Oracle

## References


[1] H. H. Rosenbrock, "An automatic method for finding the greatest or least Value of a function," *The Computer Journal*, vol. 3, no. 3, pp. 175–184, 01 1960.

[2] V. Fabian, "Stochastic approximation of minima with improved asymptotic speed," *The Annals of Mathematical Statistics*, vol. 38, no. 1, pp. 191 – 200, 1967.

[3] R. Brent, *Algorithms for minimization without derivatives*. Prentice-Hall, 1973.

[4] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[5] K. Balasubramanian and S. Ghadimi, "Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points," *Foundations of Computational Mathematics*, vol. 22, no. 1, p. 35–76, Feb. 2022.

[6] D. Yuan and D. W. C. Ho, "Randomized gradient-free method for multiagent optimization over time-varying networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1342–1347, 2015.

[7] Y. Pang and G. Hu, "Exact convergence of gradient-free distributed optimization method in a multi-agent system," in *IEEE Conference on Decision and Control (CDC)*, Florida,USA, 2018, pp. 5728–5733.

[8] ——, "Randomized gradient-free distributed optimization methods for a multiagent system with unknown cost function," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 333–340, 2020.

[9] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[10] S. Liu, P.-Y. Chen, B. Kailkhura, G. Zhang, A. O. Hero III, and P. K. Varshney, "A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications," *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 43–54, 2020.

[11] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundation and Trends in Machine Learning*, vol. 8, no. 3–4, p. 231–357, Nov. 2015.

[12] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operation Research Letter*, vol. 31, no. 3, p. 167–175, May 2003.

[13] N. Bhavsar and L. A. Prashanth, "Non-asymptotic bounds for stochastic optimization with biased noisy gradient oracles," *IEEE Transactions on Automatic Control*, 2022.

[14] H. Robbins and D. Siegmund, "A convergence theorem for non-negative almost supermartingales and some applications," in *Optimizing Methods in Statistics*, J. S. Rustagi, Ed. Academic Press, 1971, pp. 233–257.

[15] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 84–107, 2014.

[16] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research.*, vol. 6, p. 1705–1749, Dec. 2005.

[17] Y. Wang, W. Zhao, Y. Hong, and M. Zamani, "Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks," *SIAM Journal on Control and Optimization*, vol. 57, no. 4, pp. 2821–2842, 2019.

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

[19] J. C. Duchi, "Introductory lectures on stochastic optimization," 2018.