Les Cahiers du GERAD ISSN: 0711–2440

Sequential stochastic blackbox optimization with zeroth order gradient estimator

C. Audet, J. Bigeon, R. Couderc, M. Kokkolaras

G–2023–16

May 2023

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Sequential stochastic blackbox optimization with zeroth order gradient estimator

Charles Audet [a, b]

Jean Bigeon[a, c]

Romain Couderc[a, d]

Michael Kokkolaras[a, e]

[a] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[b] *Département de mathématiques et génie industriel, École Polytechnique de Montréal, Montréal (Qc), Canada, H3T 1J4*

[c] *Nantes University, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France*

[d] *Université Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France*

[e] *Department of Mechanical Engineering, McGill University, Montréal (Qc), Canada, H3A 0G4*

charles.audet@gerad.ca
jean.bigeon@ls2n.fr
romain.couderc@grenoble-inp.fr
michael.kokkolaras@mcgill.ca

May 2023
Les Cahiers du GERAD
G–2023–16

Abstract : This work considers stochastic optimization problems in which the objective function values can only be computed by a blackbox corrupted by some random noise following an unknown distribution. The proposed method is based on sequential stochastic optimization (SSO): the original problem is decomposed into a sequence of subproblems. Each of these subproblems is solved using a zeroth order version of a sign stochastic gradient descent with momentum algorithm (ZO-Signum) and with an increasingly fine precision. This decomposition allows a good exploration of the space while maintaining the efficiency of the algorithm once it gets close to the solution. Under Lipschitz continuity assumption on the blackbox, a convergence rate in expectation is derived for the ZO-signum algorithm. Moreover, if the blackbox is smooth and convex or locally convex around its minima, a convergence rate to an ϵ-optimal point of the problem may be obtained for the SSO algorithm. Numerical experiments are conducted to compare the SSO algorithm with other state-of-the-art algorithms and to demonstrate its competitiveness.

Acknowledgements: This work was financed by the IVADO Fundamental Research Projects Grant PRF–2019–8079623546 and by the NSERC Alliance grant 544900–19 in collaboration with Huawei-Canada.

1 Introduction

Stochastic blackbox optimization aims at solving the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}}\left[F(\mathbf{x},\boldsymbol{\xi})\right], \tag{1}$$

where $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$ is a blackbox [3] and $\boldsymbol{\xi}$ is an uncertain vector whose the distribution is unknown. Ω is the sample space of a probability space. This optimization problem finds application mainly in two different fields. Firstly, in a machine learning framework where the loss function's gradient is unavailable or difficult to compute, for instance in optimizing neural network architecture [33], design of adversarial attacks [13], or game content generation [42]. Secondly, when the function F is evaluated by means of a computational procedure [24]. In many cases, the function value is noisy as well. This noise may be due to environmental conditions, costs, or effects of repair actions that are unknown [35]. Another source of uncertainty also appears when the optimization is conducted at the early stages of the design process, where knowledge, information, and data can be very limited.

1.1 Related work

Stochastic derivative-free optimization has been the subject of research for many years. Traditional derivative-free methods may be divided into two categories [14]: direct search-based methods and model-based methods. Algorithms corresponding to both methods have been adapted to noisy objective functions. Examples of these works may be the stochastic Nelder-Mead algorithm [11] or the stochastic versions of the MADS algorithm [2, 4] for the direct search methods. For model-based methods, most work consider extensions of the trust region method [12, 15, 30]. A major shortcoming of these methods is their difficulty to scale to large problems.

Recently, another class of methods, named zeroth-order (ZO) methods, has been attracting increasing attention. These methods use stochastic gradient estimators, which are based on the seminal work in [22, 34] and have been extended in [18, 31, 36, 39]. These estimators have the appealing property of being able to estimate the gradient with only one or two function evaluations regardless of the problem size. Zeroth-order methods take advantage of this property to extend first-order methods. For instance, the well known first-order methods Conditional Gradient (CG), Sign Gradient Descent (SGD) [6] and ADAptive Momentum (ADAM) [23] have been extended to ZSCG [5], ZO-SGD [27] and ZO-ADAMM [13], respectively. More methods, not only based on first-order algorithms, have also emerged to solve regularized optimization problem [9] or stochastic composition optimization problem [19]. For an overview on ZO methods, readers may consult [28].

1.2 Motivation

Formally, stochastic gradient estimators involve a smoothed functional f^β (see Chapter 7.6 in [36]) which is a convolution product between f and a kernel $h^\beta(\mathbf{v})$

$$f^\beta(\mathbf{x}) := \int_{-\infty}^{\infty} h^\beta(\mathbf{u}) f(\mathbf{x}-\mathbf{u}) d\mathbf{u} = \int_{-\infty}^{\infty} h^\beta(\mathbf{x}-\mathbf{u}) f(\mathbf{u}) d\mathbf{u}. \tag{2}$$

In order for the smoothed functional to have interesting properties, the kernel must fulfill a set of conditions [pp. 263, [36]]:

1. $h^\beta(\mathbf{u}) = \frac{1}{\beta^n} h(\frac{\mathbf{u}}{\beta})$ is a piecewise differentiable function;
2. $\lim_{\beta\to 0} h^\beta(\mathbf{u}) = \delta(\mathbf{u})$, where $\delta(v)$ is Dirac's delta function;
3. $\lim_{\beta\to 0} f^\beta(\mathbf{x}) = f(\mathbf{x})$, if $\mathbf{x}$ is a point of continuity of f;
4. The kernel $h^\beta(\mathbf{u})$ is probability density function (p.d.f.), that is $f^\beta(\mathbf{x}) = \mathbb{E}_{\mathbf{U}\sim h^\beta(\mathbf{u})}[f(\mathbf{x}-\mathbf{U})] = \mathbb{E}_{\mathbf{U}\sim h(\mathbf{u})}[f(\mathbf{x}-\beta U)]$.

Frequently used kernels include the Gaussian distribution and the uniform distribution on a unit ball. Three properties about smoothed functional are worth noting. First, the smoothed functional may be interpreted as a local weighted average of the function values in the neighborhood of $\mathbf{x}$. As the kernel satisfies Condition 3, it is possible to obtain a minimum arbitrarily close to a local minimum f^*. Second, the smoothed functional is infinitely differentiable as a consequence of the convolution product, regardless of the degree of smoothness of f. Moreover, according to the chosen kernel, stochastic gradient estimators may be calculated. These estimators are unbiased estimators of the gradient of f^β and may be constructed on the basis of observations of $F(\mathbf{x}, \boldsymbol{\xi})$ alone. Finally, the smoothed functional allows the convexification of the original function f. Previous studies [37, 40] show that greater values of β result in better convexification, as illustrated in Figure 1. Additionally, a larger β leads to greater exploration of the space during the calculation of the gradient estimator. It has also been demonstrated in [29] that if the smoothing parameter is too small, the difference in function values is too small to accurately represent the function differential, particularly when the noise level is significant.

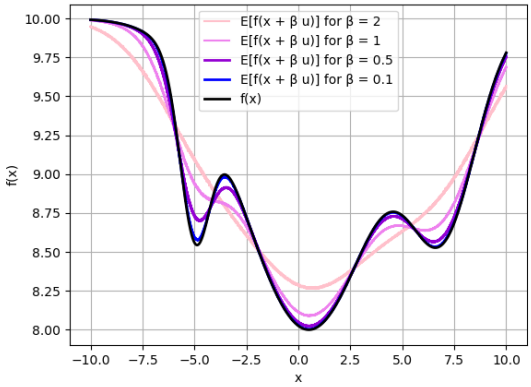

Figure 1: Curves of f^β for $u \sim \mathcal{N}(0,1)$ and different values of β.

Although the first two properties of the smoothed functional are exploited by ZO methods, the last property has not been utilized since the work in [40]. This may be because the convexification phenomenon becomes insignificant when dealing with high-dimensional problems[1]. However, for problems of relatively small size ($n \simeq 10$), this property can be useful. In the work in [40], the authors use an iterative algorithm to minimize the sequence of subproblems

$$\min_{\mathbf{x}\in\mathbb{R}^n} f^{\beta^i}(\mathbf{x}), \tag{3}$$

where β^i belongs to a finite prescaled sequence of scalars. This approach is limited because the sequence β^i does not necessarily converge to 0 and the number of iterations to pass from subproblem i to $i+1$ is arbitrarily fixed and a priori. Furthermore, neither a convergence proof nor a convergence rate are provided for the algorithm. Finally, although promising, numerical results are only presented for analytical test problems. These shortcomings motivated the research presented here.

1.3 Contributions

The main contributions of this paper can be summarized as follows.

- A sequential stochastic (SSO) optimization algorithm is developed to solve the sequence of subproblems in Equation (3). In the inner loop, a subproblem is solved according to zeroth order

[1]Note that a blackbox optimization problem with dimensions ranging from 100 to 1000 may be considered large, while problems with $n \geq 10000$ may be considered very large.

version of the Signum algorithm [6]. The stopping criterion is based on the norm of the momentum which must be below a certain threshold. In the outer loop, the sequence of β^i is proportional to the threshold needed to consider a subproblem solved and is driven to 0. Therefore, the smaller the value of β^i is (and thus better is the approximation given by f^{β^i}), the larger the computational budget granted to the resolution of the subproblem.

- A theoretical analysis of this algorithm is conducted. First, the norm of the momemtum is proved to converge to 0 in expectation, with a convergence rate that depends on the step sizes. Then, the convergence rate in expectation of the ZO-Signum algorithm to a stationary point of f^β is derived under Lipschitz continuity of the function F. Finally, if the function F is smooth, and f^β is convex or become convex around its local minima, a convergence rate to an ϵ-optimal point is derived for the SSO algorithm.
- Numerical experiments are conducted to evaluate the performance of the proposed algorithm in two applications. Firstly, a comparison is made with traditional derivative-free algorithms on the optimization of the storage cost of a solar thermal power plant model, which is a low-dimensional problem. Secondly, a comparison is made with other ZO algorithms in order to generate blackbox adversarial attacks, which are large size problems.

The remainder of this paper is organized as follows. In Section 2, the main assumptions and the Gaussian gradient estimator are described. In Section 3, the sequential optimization algorithm is presented and its convergence properties are studied in Section 4. Section 5 presents numerical results and Section 6 draws conclusions and discusses future work.

2 Gaussian Gradient estimator

The assumptions concerning the stochastic blackbox function F are as follows.

Assumption 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

a. The function satisfies $F(\cdot, \boldsymbol{\xi}) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $f(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}}[F(\mathbf{x}, \boldsymbol{\xi})]$, for all $\mathbf{x} \in \mathbb{R}^n$.
b. $F(\cdot, \boldsymbol{\xi})$ is Lipschitz continuous for any $\boldsymbol{\xi}$, with constant $L_0(F) > 0$.

Assumption 1.a implies that the expectation of $F(\mathbf{x}, \boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ is well defined on $\mathbb{R}^n$ and that the estimator $F(\mathbf{x}, \boldsymbol{\xi})$ is unbiased. Assumption 1.b is commonly used to ensure convergence and to bound the variance of the noisy objective function. It is worth noticing that no assumption is made on the differentiability of the objective function f or of its estimate F with respect to $\mathbf{x}$, contrary to most work on zeroth order methods.

Under Assumption 1, a smooth approximation of the function f may be constructed by its convolution with a Gaussian random vector. Let $\mathbf{u}$ be an n-dimensional standard Gaussian random vector and $\beta > 0$ be the smoothing parameter. Then, a smooth approximation of f is defined as

$$f^\beta(\mathbf{x}) := \frac{1}{(2\pi)^{\frac{n}{2}}} \int f(\mathbf{x} + \beta\mathbf{u}) e^{-\frac{||u||^2}{2}} d\mathbf{u} = \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \beta\mathbf{u})]. \tag{4}$$

This estimator has been studied in the literature (especially in [31]) and benefits of several appealing properties. The properties used in this work are summarized in the following Lemma.

Lemma 2.1. Under Assumption 1, the following statements hold for any integrable function $f : \mathbb{R}^n \to \mathbb{R}$ and its approximation f^β parameterized by $\beta > 0$.

1. f^β is infinitely differentiable: $f^\beta \in \mathcal{C}^\infty$.
2. A one-sided unbiased estimator of ∇f^β is

$$\tilde{\nabla} f^\beta(\mathbf{x}) := \frac{\mathbf{u}(f(\mathbf{x} + \beta\mathbf{u}) - f(\mathbf{x}))}{\beta}. \tag{5}$$

3. Let $\beta^1, \beta^2 \geq 0$, then $\forall \mathbf{x} \in \mathbb{R}^n$

$$|f^{\beta^1}(\mathbf{x}) - f^{\beta^2}(\mathbf{x})| \leq L_0(F)|\beta^1 - \beta^2|\sqrt{n}.$$

   Moreover, for $\beta > 0$, then f^β is $L_1(f^\beta)$-smooth, i.e, $f^\beta \in \mathcal{C}^{1+}$ with $L_1(f^\beta) = \frac{2\sqrt{n}}{\beta} L_0(F)$.
4. If f is convex, then f^β is also convex.

Proof.

1. It is a consequence of the convolution product between an integrable function and an infinitely differentiable kernel.
2. See [31].
3. If $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then

$$\begin{aligned} |f^{\beta^1}(\mathbf{x}) - f^{\beta^2}(\mathbf{x})| &= |\mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \beta^1\mathbf{u})] - \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \beta^2\mathbf{u})]| \\ &\leq \mathbb{E}_{\mathbf{u}}[|f(\mathbf{x} + \beta^1\mathbf{u}) - f(\mathbf{x} + \beta^2\mathbf{u})|] \\ &\leq L_0(F)|\beta^1 - \beta^2|\mathbb{E}_{\mathbf{u}}[||\mathbf{u}||] \\ &\leq L_0(F)|\beta^1 - \beta^2|\sqrt{n}, \end{aligned}$$

   where the first inequality comes from the Jensen's inequality, the second one comes from the Lipschitz continuity of f and the last one from [31, Lemma 1]. The second part is obtained from [31, Lemma 2].
4. See [31]. □

Lemma 2.1.3 establishes that at the point $\mathbf{x}$, the difference between the values of $f^{\beta^1}(\mathbf{x})$ and $f^{\beta^2}(\mathbf{x})$ is a factor of the Lipschitz constant L, the square root of the dimension n and the absolute difference between β^1 and β^2. As L and n are constants for a given problem, this difference may be arbitrarily reduced according to the values of β^1 and β^2 chosen. A particularly interesting corollary of Lemma 2.1.3 establishes that the difference between the optimal values of f^{β^1} and f^{β^2} may be bounded in the same way.

Corollary 2.2. Under the assumptions of Lemma 2.1.3

$$|f_*^{\beta^1} - f_*^{\beta^2}| \leq L_0(F)|\beta^1 - \beta^2|\sqrt{n}$$

where $f_*^{\beta^1}$ and $f_*^{\beta^2}$ denote respectively the minimum of the functions f^{β^1} and f^{β^2}.

Proof. Let $\mathbf{x}_*^{\beta^1} \in \operatorname{argmin} f^{\beta^1}(\mathbf{x})$ and $\mathbf{x}_*^{\beta^2} \in \operatorname{argmin} f^{\beta^2}(\mathbf{x})$, it follows that

$$f^{\beta^1}(\mathbf{x}_*^{\beta^1}) \leq f^{\beta^1}(\mathbf{x}_*^{\beta^2}) \leq f^{\beta^2}(\mathbf{x}_*^{\beta^2}) + L_0(F)|\beta^1 - \beta^2|\sqrt{n},$$

where the first inequality comes from the fact that $\mathbf{x}_*^{\beta^1}$ is a minimum of f^{β^1} and the second one from Lemma 2.1.4. A similar inequality may be obtained f^{β^2} for:

$$f^{\beta^2}(\mathbf{x}_*^{\beta^2}) \leq f^{\beta^2}(\mathbf{x}_*^{\beta^1}) \leq f^{\beta^1}(\mathbf{x}_*^{\beta^1}) + L_0(F)|\beta^1 - \beta^2|\sqrt{n},$$

leading to the absolute value in the corollary. □

The estimator obtained in Equation (5) may be adapted to the noisy objective function F. For instance a one-sided (mini-batch) estimator of the noised function F is

$$\tilde{\nabla} f^\beta(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{q} \sum_{j=1}^{q} \frac{\mathbf{u}^j (F(\mathbf{x} + \beta\mathbf{u}^j, \boldsymbol{\xi}^j) - F(\mathbf{x}, \boldsymbol{\xi}^j))}{\beta}, \tag{6}$$

where $(u^j)_{j=1}^q$ and $(\boldsymbol{\xi}^j)_{j=1}^q$ are q Gaussian random direction vectors and their associated q estimates values of the function F. This is still an unbiased estimator of ∇f^β because

$$\mathbb{E}_{\mathbf{u},\boldsymbol{\xi}}[\tilde{\nabla} f^\beta(\mathbf{x},\boldsymbol{\xi})] = \mathbb{E}_{\mathbf{u}}[\mathbb{E}_{\boldsymbol{\xi}}[\tilde{\nabla} f^\beta(\mathbf{x},\boldsymbol{\xi})|\mathbf{u}]] = \nabla f^\beta(\mathbf{x}). \tag{7}$$

The result of Corollary 2.2 is essential to understand why solving a sequence of optimization problems defined in Equation (3) may be efficient while it might seem counterproductive at first sight. Below are examples of the advantages of treating the problem with sequential smoothed function optimization.

- The subproblems are approximations of the original problem and it is not necessary to solve them exactly. Thus, an appropriate procedure for solving these problems with increasingly fine precision can be used. Moreover, as seen in Corollary 2.2, the best value obtained in a subproblem is close to the one of the following subproblem. The computational effort to find a solution to the second subproblem from the solution of the first should therefore not be important.
- The information collected during the optimization process of a subproblem may be reused in the subsequent subproblems since they are similar.
- A specific interest in the case of smoothed functional is the ability of using a larger value of β during the solving of the first subproblems. It allows for a better exploration of the space and convexification phenomenon of the function (see Figure 1). Moreover, the new step size may be used for each subproblem, it allows to increase the step size momentarily, in the hope of having a greater chance of escaping a local minimum.

3 A Sequential Stochastic Optimization (SSO) algorithm

Section 3.1 presents a zeroth-order version of the Signum algorithm [6] to solve Subproblem (3) for a given β^i and Section 3.2 presents the complete algorithm used to solve the sequential optimization problem.

3.1 The Zeroth-order Signum algorithm

Algorithm 1 ZO-Signum (ZOS) algorithm to solve subproblem $i \in \mathbb{N}$

Input: $\mathbf{x}^{i,0}, \mathbf{m}^{i,0}, \beta^i, s_1^{i,0}, s_2^{i,0}$, L, q, M
Set $k = 0$
Define stepsize sequences $s_1^{i,k} = \frac{s_1^{i,0}}{(k+1)^{\alpha_1}}$ and $s_2^{i,k} = \frac{s_2^{i,0}}{(k+1)^{\alpha_2}}$
while $||\mathbf{m}^{i,k}|| > \frac{L\beta^i}{4\beta^0}$ or $k \leq M$ **do**
    Draw q samples $\mathbf{u}^k$ from the Gaussian distribution $\mathcal{N}(\mathbf{0}, I)$
    Calculate the average of the q Gaussian estimate $\tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,k}, \boldsymbol{\xi}^{i,k})$ from Equation (6)
    Update:

$$\mathbf{m}^{i,k+1} = s_2^{i,k} \tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,k}, \boldsymbol{\xi}^k) + (1 - s_2^{i,k}) \mathbf{m}^{i,k} \tag{8}$$

$$x_j^{i,k+1} = x_j^{i,k} - s_1^k \text{sign}(m_j^{i,k+1}) \; \forall j \in [1, n] \tag{9}$$

    $k \leftarrow k + 1$
end while
Return $\mathbf{m}^{i,k}$ and $\mathbf{x}^{i,k}$

In order to solve a subproblem, a zeroth-order version of the Signum algorithm (Algorithm 2 of [6]) is used. The Signum algorithm is a momentum version of the sign-SGD algorithm. In [27], the authors extended the original sign-SGD algorithm to a zeroth-order version of this algorithm. However, a zeroth-order of Signum is not studied in the work of [27]. As the Signum algorithm has been shown to be competitive with the Adam algorithm [6], a zeroth-order version of this algorithm seems interesting to consider. There is an important difference between the original Signum algorithm and its zeroth

order version presented in Algorithm 1. Indeed, while the step size of the momentum $1 - s_2^{i,k}$ is kept constant in the work of [6], it is driven to 1 in our work. This leads to two consequences. First, the variance is reduced since the gradient is averaged on a longer time horizon, without using mini-batch sampling. Second, as it has been demonstrated in other stochastic approximation works [7, section 3.3], [38], with carefully chosen step sizes the norm of the momentum goes to 0 with probability one. In the ZO-Signum algorithm, the norm of the momemtum is thus used as a stopping criterion.

3.2 The SSO algorithm

The optimization of the sequence of subproblems described in Equation (3) is driven by the sequential stochastic (SSO) algorithm presented in Algorithm 2. The value of β plays a critical role in this algorithm as it serves as both the smoothing parameter and the stopping criterion for Algorithms 1 and 2. The process of Algorithm 2 is inspired by the MADS algorithm [1]. It is based on two steps: a search step and a local step. The search step is optional and may consist in any heuristics and is required only on problem with relatively small dimensions. In Algorithm 2, an example of a search is given which consists of updating $\mathbf{x}$ after M iterations of the ZO-Signum algorithm with the best known $\mathbf{x}$ found so far. The local step is then used: Algorithm 1 is launched on each subproblem i with specific values of β^i and step-size sequences. Once Algorithm 1 meets the stopping criterion (which depends on the value of β^i), the value of β^i and the initial step-sizes $s_1^{i,0}$ and $s_2^{i,0}$ are reduced, and the algorithm proceeds to the next subproblem. The convergence is garanteed by the local step since the search step is run only a finite number of times.

Algorithm 2 Sequential Stochastic Optimization (SSO) algorithm

1: **Initialization:**
2: Set $\mathbf{x}^{0,0} \in \mathbb{R}^n$, $\beta^0 > 0$ and N the maximum number of function calls for the search step
3: Set q the number of gradient estimates at each iteration of ZO-Signum algorithm
4: Set M the minimum number of iterations made by the ZO-Signum algorithm on a subproblem
5: $\mathcal{C}$ the cache containing all the evaluated points
6: Set $\mathbf{m}^{0,0} = \tilde{\nabla} f^{\beta^0}(\mathbf{x}^{0,0}, \boldsymbol{\xi}^0)$ and $L = +\infty$
7: Set $s_1^{0,0} > 0$ and $s_2^{0,0} > 0$
8: Set $i = 0$
9: **Search step (optional):**
10: **while** $M(i+1)q \leq N$: **do**
11: Solve subproblem i with Algorithm 1:

$$\mathbf{m}^{i+1,0} = ZOS(\mathbf{x}^{i,0}, \mathbf{m}^{i,0}, \beta^i, s_1^{i,0}, s_2^{i,0}, L, q, M)$$
$$\mathbf{x}^{i+1,0} \in \operatorname*{argmin}_{\mathbf{x} \in \mathcal{C}} F(\mathbf{x}, \boldsymbol{\xi})$$

12: Update β^i, $s_1^{i,0}$ and $s_2^{i,0}$ as in step 18
13: **end while**
14: $L = ||\mathbf{m}^{0,0}||$
15: **Local step:**
16: **while** $\beta^i > \epsilon$ **do**
17: Solve subproblem i with Algorithm 1:

$$\mathbf{m}^{i+1,0}, \mathbf{x}^{i+1,0} = ZOS(\mathbf{x}^{i,0}, \mathbf{m}^{i,0}, \beta^i, s_1^{i,0}, s_2^{i,0}, L, q, M)$$

18: Update:

$$\beta^i = \frac{\beta^0}{(i+1)^2}, s_1^{i,0} = \frac{s_1^{0,0}}{(i+1)^{\frac{3}{2}}}, \quad s_2^{i,0} = \frac{s_2^{0,0}}{i+1}$$
$$i \leftarrow i+1$$

19: **end while**
20: Return $\mathbf{x}^i$

It is worth noting that the decrease rate of β^i is chosen to be so that the difference between subproblems i and $i+1$ is not significant. Therefore, the information collected in subproblem i, through the momentum vector $\mathbf{m}$, can be used in subproblem $i+1$. Furthermore, the initial step-

sizes $s_1^{i,0}$ and $s_2^{i,0}$ decrease with each iteration, allowing us to focus our efforts quickly towards a local optimum when $s_1^{0,0}$ and β^0 are chosen to be relatively large.

4 Convergence analysis

The convergence analysis is conducted in two steps : first the convergence in expectation is derived for Algorithm 1 and then the convergence for Algorithm 2 is derived.

4.1 Convergence of the ZOS algorithm

The analysis of Algorithm 1 follows the general methodology given in Appendix E in [6]. In the following subsection, the main result in [6] is recalled for completeness. The next subsections are devoted to bound the variance and bias terms when $\lim_{k\to\infty} s_2^{i,k} = 0$. Finally, these results are used to obtain the convergence rate in expectation of Algorithm 1 in the non convex and convex case. The last subsection is devoted to a theoretical comparison with other ZO methods of the literature. The subproblem index i is kept constant throughout this section.

4.1.1 Preliminary result [6]

The following proposition uses the Lipschitz continuity of the function f^{β^i} (proved in Lemma 2.1) to bound the gradient at the kth iteration.

Proposition 4.1 ([6])**.** For the subproblem $i \in \mathbb{N}$, under Assumption 1 and in the setting of Algorithm 1, we have

$$\begin{aligned} s_1^{i,k}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq & \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,k+1})] + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2 \\ & + 2s_1^{i,k}\underbrace{\mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1]}_{\text{bias}} + 2s_1^{i,k}\sqrt{n}\sqrt{\underbrace{\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||_2^2]}_{\text{variance}}} \end{aligned} \tag{10}$$

where $\bar{m}_j^{i,k+1}$ is defined recursively as $\bar{m}_j^{i,k+1} = s_2^{i,k}\nabla f^{\beta^i}(\mathbf{x}^{i,k}) + (1 - s_2^{i,k})\bar{m}_j^{i,k}$.

Proof. See Appendix A. □

Now, it remains to bound the three terms on the right side of Inequality (10).

4.1.2 Bound on the variance term

The three following lemmas are consecrated to bound the variance term. Unlike the work reported in [6], the variance reduction is conducted by driving the step size of the momemtum to 0. It avoids to sample an increasing number of stochastic gradients at each iteration, which may be problematic as noted in [27]. To achieve this, the variance term is first decomposed in term of expectation of the squared norm of the stochastic gradient estimators $\tilde{g}$.

Lemma 4.2. For the subproblem $i \in \mathbb{N}$, let $k \in \mathbb{N}$ and $j \in [1, n]$, we have

$$\begin{aligned} \mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2] \leq & (s_2^{i,k})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,k}||^2] + \sum_{r=0}^{k-1}(s_2^{i,r})^2\prod_{t=r}^{k-1}(1 - s_2^{i,t+1})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,r}||^2] \\ & + \prod_{t=0}^{k}(1 - s_2^{i,t})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,0}||^2], \end{aligned}$$

where $\tilde{g}_j^{i,r} = \tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,r}, \boldsymbol{\xi}^r), \forall r \in [0, k]$ is defined in Equation (6) and the norm is $||\cdot||_2$.

Proof. Let $k \in \mathbb{N}$, by definition of $\mathbf{m}^{i,k}$ and $\bar{\mathbf{m}}^{i,k}$, it follows that

$$\begin{aligned}||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2 =&(s_2^{i,k})^2||\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||^2 + (1-s_2^{i,k})^2||\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k}||^2 \\ &+ 2s_2^{i,k}(1-s_2^{i,k})(\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k}))^T(\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k}).\end{aligned}$$

The expectation of this expression is

$$\begin{aligned}\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2] =&(s_2^{i,k})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||^2] + (1-s_2^{i,k})^2\mathbb{E}[||\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k}||^2] \quad (11)\\ &+ 2s_2^{i,k}(1-s_2^{i,k})\mathbb{E}[(\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(x^{i,k}))^T(\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k})]. \quad (12)\end{aligned}$$

Now, introducing the associated sigma field of the process $\mathcal{F}^{i,k} = \sigma(\mathbf{x}^{j,t}, \mathbf{m}^{j,t}, \bar{\mathbf{m}}^{j,t}; j \leq i, t \leq k)$ by the law of total expectation, it follows that

$$\begin{aligned}\mathbb{E}[(\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(x^{i,k}))^T(\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k})] &= \mathbb{E}[\mathbb{E}[(\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k}))^T(\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k})|\mathcal{F}^{i,k}] \\ &= \mathbb{E}[(\mathbb{E}[\tilde{\mathbf{g}}^{i,k}|\mathcal{F}^{i,k}] - \nabla f^{\beta^i}(x^{i,k}))^T(\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k})] \\ &= 0,\end{aligned}$$

where the second equality holds because $\mathbf{m}^{i,k}, \bar{\mathbf{m}}^{i,k}$ and $\nabla f^{\beta^i}(\mathbf{x}^{i,k})$ are fixed conditioned on $\mathcal{F}^{\rangle,\|}$ and because $\mathbb{E}[\tilde{\mathbf{g}}^{i,k}|\mathbf{x}^{i,k}] = \nabla f^{\beta^i}(\mathbf{x}^{i,k})$ as $\tilde{\mathbf{g}}^{i,k}$ is an unbiased estimator of the gradient by Equation (7). By substituting this result in (12), it follows that

$$\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2] = (s_2^{i,k})^2\mathbb{E}[||\tilde{g}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||^2] + (1-s_2^{i,k})^2\mathbb{E}[||\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k}||^2].$$

By repeating this process iteratively, we obtain

$$\begin{aligned}\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2] =&(s_2^{i,k})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||^2] \\ &+ \sum_{r=0}^{k-1}(s_2^{i,r})^2\prod_{t=r}^{k-1}(1-s_2^{i,t+1})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,r} - \nabla f^{\beta^i}(\mathbf{x}^{i,r})||^2] \\ &+ \prod_{t=0}^{k}(1-s_2^{i,t})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,0} - \nabla f^{\beta^i}(x^{i,0})||^2].\end{aligned} \quad (13)$$

Finally, by observing that $\forall r \in [0,k], \mathbb{E}[\tilde{\mathbf{g}}^{i,r}|\mathbf{x}^{i,r}] = \nabla f^{\beta^i}(x^{i,r})$ and by the law of total expectation, we obtain

$$\begin{aligned}\mathbb{E}[||\tilde{\mathbf{g}}^{i,r} - \nabla f^{\beta^i}(x^{i,r})||^2] &= \mathbb{E}[||\tilde{\mathbf{g}}^{i,r} - \mathbb{E}[\tilde{\mathbf{g}}^{i,r}|\mathbf{x}^{i,r}]||^2] \\ &= \mathbb{E}[||\tilde{\mathbf{g}}^{i,r}||^2] - \mathbb{E}[||\nabla f^{\beta^i}(x^{i,r})||^2] \\ &\leq \mathbb{E}[||\tilde{\mathbf{g}}^{i,r}||^2].\end{aligned}$$

Introducing this inequality in Equation (13) completes the proof. □

Second, the expectation of the squared norm of the stochastic gradient estimators are bounded by a constant depending quadratically of the dimension.

Lemma 4.3. Let $i \in \mathbb{N}$, $r \in [0,k]$, $j \in [1,n]$, then under Assumption 1, we have

$$\mathbb{E}[||\tilde{\mathbf{g}}^{i,r}||^2] \leq L_0(F)^2(n+4)^2$$

where $L_0(F)$ is the Lipschitz constant of F.

Proof. By Equation (6) with $q = 1$, it follows that

$$\mathbb{E}[||\tilde{\mathbf{g}}^{i,r}||^2] = \mathbb{E}\left[\frac{||u||^2}{(\beta^i)^2}\left(F(\mathbf{x}^{i,r} + \beta^i\mathbf{u}, \boldsymbol{\xi}) - F(\mathbf{x}^{i,r}, \boldsymbol{\xi})\right)^2\right]$$

$$\leq L_0(F)^2\mathbb{E}[||u||^4]$$
$$\leq L_0(F)^2(n+4)^2$$

where the first inequality follows from Assumption 1.b and the second by [31, Lemma 1]. □

Finally, a technical lemma bounds the second term of the decomposition of the Lemma 4.2 by a decreasing sequence. It allows to obtain the same rate of convergence than in the work in [6] without sampling any stochastic gradient.

Lemma 4.4. For the subproblem $i \in \mathbb{N}$, let $s_2^{i,k}$ defined such that $s_2^{i,k} = \frac{s_2^{i,0}}{(k+1)^{\alpha_2}}$ with $\alpha_2 \in (0,1)$ and $s_2^{i,0} \in (0,1)$, then for k such that

$$\frac{k}{(k+1)^{\alpha_2}} \geq \frac{\ln(s_2^{i,0}) + (1+\alpha_2)\ln(k)}{s_2^{i,0}} \tag{14}$$

the following inequality holds

$$\sum_{r=0}^{k-1}(s_2^{i,r})^2 \prod_{t=r}^{k-1}(1-s_2^{i,t+1})^2 \leq \frac{9s_2^{i,0}}{k^{\alpha_2}}. \tag{15}$$

Proof. Let $k \in \mathbb{N}$; as in [6],the strategy consists of breaking up the sum in order to bound the both terms separately.

$$\begin{aligned}
\sum_{r=0}^{k-1}(s_2^{i,r})^2 \prod_{t=r}^{k-1}(1-s_2^{i,t+1})^2 &= \sum_{r=0}^{\lfloor k/2\rfloor-1}(s_2^{i,r})^2 \prod_{t=r}^{k-1}(1-s_2^{i,t+1})^2 + \sum_{r=\lfloor k/2\rfloor}^{k-1}(s_2^{i,r})^2 \prod_{t=r}^{k-1}(1-s_2^{i,t+1})^2 \\
&\leq (1-s_2^{i,k})^{2\lfloor k/2\rfloor}\sum_{r=0}^{\lfloor k/2\rfloor-1}(s_2^{i,r})^2 + (s_2^{i,\lfloor k/2\rfloor-1})^2 \sum_{r=\lfloor k/2\rfloor}^{k-1}(1-s_2^{i,k})^{2(k-r-1)} \\
&\leq (s_2^{i,0})^2\lfloor k/2\rfloor(1-s_2^{i,k})^{2\lfloor k/2\rfloor} + \frac{8(s_2^{i,0})^2}{k^{2\alpha_2}}\sum_{r=0}^{\lfloor k/2\rfloor}(1-s_2^{i,k})^{2r} \\
&\leq (s_2^{i,0})^2 k(1-s_2^{i,k})^{2\lfloor k/2\rfloor} + \frac{8(s_2^{i,0})^2}{k^{2\alpha_2}(1-(1-s_2^{i,k})^2)} \\
&\leq (s_2^{i,0})^2 k(1-s_2^{i,k})^{2\lfloor k/2\rfloor} + \frac{8s_2^{i,0}}{k^{\alpha_2}(2-s_2^{i,k})}.
\end{aligned}$$

Now, we are looking for k such that

$$s_2^{i,0}k(1-s_2^{i,k})^{2\lfloor k/2\rfloor} \leq \frac{1}{k^{\alpha_2}} \Leftrightarrow e^{2\lfloor k/2\rfloor \ln(1-s_2^{i,k})} \leq \frac{1}{(s_2^{i,0})k^{1+\alpha_2}}.$$

As, $\ln(1-x) \leq -x$, it is sufficient to find k such that

$$\begin{aligned}
e^{-s_2^{i,0}\frac{k}{(k+1)^{\alpha_2}}} &\leq \frac{1}{(s_2^{i,0})k^{1+\alpha_2}} \\
\Leftrightarrow \frac{k}{(k+1)^{\alpha_2}} &\geq \frac{\ln(s_2^{i,0}) + (1+\alpha_2)\ln(k)}{s_2^{i,0}}.
\end{aligned}$$

Taking such a k allows to complete the proof. □

Combining the three previous Lemmas allows to bound the variance term in the Proposition 4.1.

Proposition 4.5. In the setting of Lemmas 4.3 and 4.4 and under assumption 1.b, the variance term of Proposition 4.1 is bounded by

$$\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||_2^2] \leq \frac{9 s_2^{i,0} L_0(F)^2 (n+4)^2}{k^{\alpha_2}} + o\left(\frac{1}{k^{\alpha_2}}\right).$$

Proof. By Lemmas 4.2 and 4.3, it follows that

$$\begin{aligned}\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2] &\leq (s_2^{i,k})^2 \mathbb{E}[||\tilde{\mathbf{g}}^{i,k}||^2] + \sum_{r=0}^{k-1} (s_2^{i,r})^2 \prod_{t=r}^{k-1} (1 - s_2^{i,t+1})^2 \mathbb{E}[||\tilde{\mathbf{g}}^{i,r}||^2] \\ &\quad + \prod_{t=0}^{k} (1 - s_2^{i,t})^2 \mathbb{E}[||\tilde{\mathbf{g}}^{i,0}||^2] \\ &\leq \left((s_2^{i,k})^2 + \sum_{r=0}^{k-1} (s_2^{i,r})^2 \prod_{t=r}^{k-1} (1 - s_2^{i,t+1})^2 + \prod_{t=0}^{k} (1 - s_2^{i,t})^2 \right) L_0(F)^2 (n+4)^2.\end{aligned}$$

Now as $(s_2^{i,k})^2 = o\left(\frac{1}{k^{\alpha_2}}\right)$ and $\prod_{t=0}^{k} (1 - s_2^{i,t})^2 = o\left(\frac{1}{k^{\alpha_2}}\right)$, the result follows from Lemma 4.4. □

4.1.3 Bound on the bias term

First, the biasterm is bounded by a sum depending on s_1^k and s_2^k.

Lemma 4.6. For the subproblem $i \in \mathbb{N}$ and at iteration $k \in \mathbb{N}$ of the algorithm 1, we have

$$\mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq 2nL_1(f^{\beta^i}) \left(\sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) \right).$$

Proof. First of all, observe that the quantity

$$S^{i,k} := \begin{cases} 1 & \text{if } k = 0 \\ s_2^{i,k} + \sum_{r=0}^{k-1} s_2^{i,r} \prod_{t=r}^{k-1} (1 - s_2^{i,r+1}) + \prod_{t=0}^{k} (1 - s_2^{i,t}) & \text{otherwise,} \end{cases} \tag{16}$$

may be written recursively as

$$S^{i,k} = \begin{cases} 1 & \text{if } k = 0 \\ s_2^{i,k} + (1 - s_2^{i,k}) S^{k-1} & \text{otherwise.} \end{cases}$$

Therefore, $S^{i,k} = 1$ for all k. By definition of $\bar{m}_j^{i,k}$, we have

$$\begin{aligned}\bar{\mathbf{m}}^{i,k} &= s_2^{i,k} \nabla f^{\beta^i}(\mathbf{x}^{i,k}) + \sum_{r=0}^{k-1} s_2^{i,r} \prod_{t=r}^{k-1} (1 - s_2^{i,t+1}) \nabla f^{\beta^i}(\mathbf{x}^{i,r}) + \prod_{t=0}^{k} (1 - s_2^{i,t}) \nabla f^{\beta^i}(\mathbf{x}^{i,0}) \\ \nabla f^{\beta^i}(\mathbf{x}^{i,k}) &= \left(s_2^{i,k} + \sum_{r=0}^{k-1} s_2^{i,r} \prod_{t=r}^{k-1} (1 - s_2^{i,t+1}) + \prod_{t=0}^{k} (1 - s_2^{i,t}) \right) \nabla f^{\beta^i}(\mathbf{x}^{i,k}).\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}\mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq &\sum_{r=0}^{k-1} s_2^{i,r} \prod_{t=r}^{k-1} (1 - s_2^{i,t+1}) \mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,r}) - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \\ &+ \prod_{t=0}^{k} (1 - s_2^{i,t}) \mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,0}) - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1].\end{aligned} \tag{17}$$

By the smoothness of the function f^{β^i}, Lemma F.3 of [6] ensures that $\forall r \in [0, k-1]$

$$||\nabla f^{\beta^i}(\mathbf{x}^{i,r}) - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1 \leq \sum_{l=r}^{k-1} ||\nabla f^{\beta^i}(\mathbf{x}^{i,l+1}) - \nabla f^{\beta^i}(\mathbf{x}^{i,l})||_1 \;\leq\; 2nL_1(f^{\beta^i}) \sum_{l=r}^{k-1} s_1^{i,l}.$$

Substituting this inequality in Equation (17) gives

$$\mathbb{E}[||\bar{m}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq 2nL_1(f^{\beta^i}) S_1^{i,k} \tag{18}$$

where

$$S_1^{i,k} = \sum_{r=0}^{k-1} s_2^{i,r} \sum_{l=r}^{k-1} s_1^{i,l} \prod_{t=r}^{k-1} (1 - s_2^{i,t+1}) + \sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=0}^{k} (1 - s_2^{i,t}).$$

Reordering the terms in S_1^k, we obtain

$$\begin{aligned}
S_1^{i,k} &= \sum_{l=0}^{k-1} s_1^{i,l} \left(\sum_{r=0}^{l} s_2^{i,r} \prod_{t=r}^{k-1} (1 - s_2^{i,t+1}) + \prod_{t=0}^{k} (1 - s_2^{i,t}) \right) \\
&= \sum_{l=0}^{k-1} s_1^{i,l} \left(s_2^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) + \sum_{r=0}^{l-1} s_2^{i,r} \prod_{t=r}^{k-1} (1 - s_2^{i,t+1}) + \prod_{t=0}^{k} (1 - s_2^{i,t}) \right) \\
&= \sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) \underbrace{\left(s_2^{i,l} + \sum_{r=0}^{l-1} s_2^{i,r} \prod_{t=r}^{l-1} (1 - s_2^{i,t+1}) + \prod_{t=0}^{l} (1 - s_2^{i,t}) \right)}_{S^{i,l}=1} \\
&= \sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}),
\end{aligned}$$

which completes the proof. □

Second, the sum may be bounded by a term decreasing with k.

Lemma 4.7. For the subproblem $i \in \mathbb{N}$ and let $s_2^{i,k} = \frac{s_2^{i,0}}{(k+1)^{\alpha_2}}$ and $s_1^{i,k} = \frac{s_1^{i,0}}{(k+1)^{\alpha_1}}$ with $s_1^{i,0} \in (0,1), s_2^{i,0} \in (0,1)$ and $0 < \alpha_2 < \alpha_1 < 1$, then for k such that

$$\frac{k}{(k+1)^{\alpha_2}} \geq \frac{2\left(\ln(s_2^{i,0}) + (1 + \alpha_1 - \alpha_2)\ln(k)\right)}{s_2^{i,0}} \tag{19}$$

the following inequality holds

$$\sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) \leq \frac{5 s_1^{i,0}}{s_2^{i,0} k^{\alpha_1 - \alpha_2}}. \tag{20}$$

Proof. The proof follows the proof of Lemma 4.4. The sum is partitioned as follows:

$$\begin{aligned}
\sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) &= \sum_{l=0}^{\lfloor k/2 \rfloor - 1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) + \sum_{l=\lfloor k/2 \rfloor - 1}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) \\
&\leq (1 - s_2^{i,k})^{\lfloor k/2 \rfloor} \sum_{l=0}^{\lfloor k/2 \rfloor - 1} s_1^{i,l} + s_1^{i,\lfloor k/2 \rfloor - 1} \sum_{l=\lfloor k/2 \rfloor - 1}^{k-1} (1 - s_2^{i,k})^{k-r-1} \\
&\leq s_1^{i,0} k (1 - s_2^{i,k})^{\lfloor k/2 \rfloor} + \frac{4 s_1^{i,0}}{k^{\alpha_1}(1 - (1 - s_2^{i,k}))}
\end{aligned}$$

$$= \frac{s_1^{i,0} s_2^{i,0} k (1 - s_2^{i,k})^{\lfloor k/2 \rfloor}}{s_2^{i,0}} + \frac{4 s_1^{i,0}}{s_2^{i,0} k^{\alpha_1 - \alpha_2}}.$$

Now, as in Lemma 4.4 taking k such that

$$\frac{k}{(k+1)^{\alpha_2}} \geq \frac{2\left(\ln(s_2^{i,0}) + (1 + \alpha_1 - \alpha_2)\ln(k)\right)}{s_2^{i,0}}$$

ensures that $s_2^{i,0} k (1 - s_2^{i,k})^{\lfloor k/2 \rfloor} \leq \frac{1}{k^{\alpha_1 - \alpha_2}}$, which completes the proof. □

Finally, using the two previous Lemmas allows to bound the bias term.

Proposition 4.8. In the setting of Lemma 4.7, the bias term of Proposition 4.1 is bounded by

$$\mathbb{E}[||\bar{m}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq 10 n L_1(f^{\beta^i}) \frac{s_1^{i,0}}{s_2^{i,0} k^{\alpha_1 - \alpha_2}}.$$

Proof. The proof is a straightforward consequence of Lemmas 4.6 and 4.7. □

4.1.4 Convergence in expectation of the ZOS algorithm

As the different terms in the inequality of the Proposition 4.1 have been bounded, the main result of this section may be derived in the following theorem.

Theorem 4.9. For a subproblem $i \in \mathbb{N}$ and under Assumption 1, let $\alpha_1 \in (0,1)$, $\alpha_2 \in (0, \alpha_1)$, $0 < s_1^{i,0}, s_2^{i,0} < 1$ and $K > C$ where $C \in \mathbb{N}$ satisfies Equations (14) and (19), we have

$$\begin{aligned} \mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||_1] \leq & \frac{1}{K^{1-\alpha_1} - \frac{C}{K^{\alpha_1}}} \left(\frac{D_f^i}{s_1^{i,0}} + \frac{n\sqrt{n} L_0(F) s_1^{i,0}}{\beta^i} \sum_{k=C}^{K} \frac{1}{k^{2\alpha_1}} \right. \\ & \left. + 6\sqrt{s_2^{i,0}} L_0(F) \sqrt{n}(n+4) \sum_{k=C}^{K} \frac{1}{k^{\alpha_1 + \frac{\alpha_2}{2}}} + \frac{40 L_0(F) s_1^{i,0} n\sqrt{n}}{s_2^{i,0} \beta^i} \sum_{k=C}^{K} \frac{1}{k^{2\alpha_1 - \alpha_2}} \right), \end{aligned} \tag{21}$$

where $f^{\beta^i}(\mathbf{x}^{i,C}) - \min_{\mathbf{x}} f^{\beta^i}(\mathbf{x}) \leq D_f^i$, $L_0(F)$ is the Lipschitz constant of F and R is randomly picked from a uniform distribution in $[C, K]$.

Proof. Let $C \in \mathbb{N}$ satisfying Equations (14) and (19) and sum over the inequality in Proposition 4.1, it follows that

$$\begin{aligned} \sum_{k=C}^{K} s_1^{i,k} \mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq & \mathbb{E}[f^{\beta^i}(x^{i,C}) - f^{\beta^i}(\mathbf{x}^{i,K+1})] + \frac{n L_1(f^{\beta^i})}{2} \sum_{k=C}^{K} (s_1^{i,k})^2 \\ & + 2\sqrt{n} \sum_{k=C}^{K} s_1^{i,k} \sqrt{\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||_2^2]} \\ & + 2 \sum_{k=C}^{K} s_1^{i,k} \mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1]. \end{aligned}$$

By substituting the results of Proposition 4.5 and 4.8 in the previous inequality, we obtain

$$\begin{aligned} \sum_{k=C}^{K} s_1^{i,k} \mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq & \mathbb{E}[f^{\beta^i}(x^{i,C}) - f^{\beta^i}(\mathbf{x}^{i,K+1})] + \frac{n L_1(f^{\beta^i})}{2} \sum_{k=C}^{K} (s_1^{i,k})^2 \\ & + 6\sqrt{s_2^{i,0}} L_0(F)(n+4)\sqrt{n} \sum_{k=C}^{K} \frac{s_1^{i,0}}{k^{\alpha_1 + \frac{\alpha_2}{2}}} + \frac{20 L_1(f^{\beta^i}) s_1^{i,0} n}{s_2^{i,0}} \sum_{k=C}^{K} \frac{s_1^{i,0}}{k^{2\alpha_1 - \alpha_2}}. \end{aligned}$$

Dividing both sides by $s_1^{i,0}K^{-\alpha_1}(K-C)$, picking R randomly uniformly in $[C,K]$ and using the definition of D_f^i given that $\min_{\mathbf{x}} f(\mathbf{x}) \leq f(\mathbf{x})$ for all $\mathbf{x}$, we get

$$\begin{aligned}\mathbb{E}[||\nabla f^{\beta^i}(x^{i,R})||_1] =& \frac{1}{K-C}\sum_{k=C}^{K}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq \frac{1}{K-C}\sum_{k=C}^{K}\frac{K^{\alpha_1}}{k^{\alpha_1}}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \\ \leq& \frac{1}{K^{1-\alpha_1}-\frac{C}{K^{\alpha_1}}}\left(\frac{D_f^i}{s_1^{i,0}} + \frac{nL_1(f^{\beta^i})s_1^{i,0}}{2}\sum_{k=C}^{K}\frac{1}{k^{2\alpha_1}} + 6\sqrt{s_2^{i,0}}L_0(F)(n+4)\sqrt{n}\sum_{k=C}^{K}\frac{1}{k^{\alpha_1+\frac{\alpha_2}{2}}}\right. \\ &\left. + \frac{20L_1(f^{\beta^i})s_1^{i,0}n}{s_2^{i,0}}\sum_{k=C}^{K}\frac{1}{k^{2\alpha_1-\alpha_2}}\right).\end{aligned}$$

Recalling that $L_1(f^{\beta^i}) = \frac{2\sqrt{n}L_0(F)}{\beta^i}$ completes the proof. □

This theorem allows to prove the convergence in expectation of the norm of the gradient when α_1 and α_2 are chosen adequately. In particular, the following corollary provides the convergence when $\alpha_1 = \frac{3}{4}$ and $\alpha_2 = \frac{1}{2}$.

Corollary 4.10. Under the same setting of Theorem with $\beta^i \approx 1$ $\alpha_1 = \frac{3}{4}$, $\alpha_2 = \frac{1}{2}$, $s_1^{i,0} = \frac{1}{n^{\frac{3}{4}}}$ and $s_2^{i,0} \approx 1$, we have

$$\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||_2] = O\left(\frac{n^{\frac{3}{2}}}{K^{1/4}}\ln(K)\right). \tag{22}$$

Proof. The result is a direct consequence of Theorem 4.9 with the specified constant and by noting that $||\cdot||_2 \leq ||\cdot||_1$ in $\mathbb{R}^n$. □

In [13, 18, 27], the function F is assumed to be smooth with Lipschitz continuous gradient. In the present work, F is only assumed to be Lipschitz continuous. This has two main consequences on the result of convergence: the dependence of the dimension on the convergence rate is larger. Furthermore, while β must be chosen relatively small in the smooth case, it is interesting to note that it does not have to be this way in the nonsmooth case.

4.1.5 The convex case

The convergence results of the ZOS algorithm has been derived in the non-convex case. In the next theorem, convergence results are derived when the function f^{β^i} is convex.

Theorem 4.11. Under Assumption 1, suppose moreover that f^{β^i} is convex and there exists ρ such that $\rho = \max_{k,k'\in\mathbb{N}}||\mathbf{x}^{i,k} - \mathbf{x}^{i,k'}||$, then by setting

$$\beta^i \leq \frac{1}{\sqrt{n}K^{\frac{1}{3}}}, s_1^{i,k} = \frac{2\rho}{(k+1)},\ s_2^{i,k} = \frac{1}{(k+1)^{\frac{2}{3}}} \text{ and } \Gamma^k := \prod_{l=2}^{k}\left(1-\frac{2}{k+1}\right) = \frac{2}{k(k+1)} \text{ with } \Gamma^1 = 1, \tag{23}$$

it follows that

$$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,K}) - f^{\beta^i}(\mathbf{x}^*)] \leq \frac{4\rho n\sqrt{n}L_0(F)}{\beta^i K^{\frac{1}{3}}}. \tag{24}$$

and

$$\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||] \leq \frac{2L_0(F)}{K^2} + \frac{4n\sqrt{n}L_0(F)}{\beta^i K^{\frac{1}{3}}} \tag{25}$$

where R is a random variable in $[0, K-1]$ whose the probability distribution is given by

$$\mathbb{P}(R = k) = \frac{s_1^{i,k}/\Gamma^{k+1}}{\sum_{k=0}^{K-1} s_1^{i,k}/\Gamma^{k+1}}.$$

Proof. Under the assumptions in the statement of the Theorem, it follows by Proposition 4.1 that

$$\begin{aligned}\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,*})] \leq & \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*})] - s_1^{i,k}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||] + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2 \\ & + 2s_1^{i,k}\mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] + 2s_1^{i,k}\sum_{j=1}^{n}\sqrt{\mathbb{E}[(m_j^{i,k+1} - \bar{m}_j^{i,k+1})^2]} \qquad (26) \\ \leq & \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*})] - s_1^{k}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||] + \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i(k+1)^{\frac{4}{3}}},\end{aligned}$$

where the last inequality follows thanks to Propositions 4.5, 4.8 with $L_1(f^{\beta^i}) = \frac{2L_0(F)\sqrt{n}}{\beta^i}$ and the values of $s_1^{i,k}$ and $s_2^{i,k}$. Now, by convexity assumption of f^{β^i} and the bound on the maximal distance between two iterates, the following holds

$$\begin{aligned}f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*}) &\leq \nabla f^{\beta^i}(\mathbf{x}^{i,k})^T(\mathbf{x}^{i,k} - \mathbf{x}^{i,*}) \\ &\leq ||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||\,||\mathbf{x}^{i,k} - \mathbf{x}^{i,*}|| \\ &\leq \rho||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||.\end{aligned}$$

Thus, by substituting this result into Equation (26), it follows that

$$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,*})] \leq \left(1 - \frac{2}{(k+1)}\right)\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*})] + \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i(k+1)^{\frac{4}{3}}}.$$

Now by dividing by Γ^{k+1} both sides of the equation and summing up the inequalities, it follows that

$$\begin{aligned}\frac{\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,K}) - f^{\beta^i}(\mathbf{x}^{i,*})]}{\Gamma^K} &\leq \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i}\sum_{k=0}^{K-1}\frac{1}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}} \\ &\leq \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i}\sum_{k=0}^{K-1}(k+1)^{\frac{2}{3}}.\end{aligned}$$

Thus

$$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,K}) - f^{\beta^i}(\mathbf{x}^{i,*})] \leq \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i}\Gamma^K\sum_{k=0}^{K-1}(k+1)^{\frac{2}{3}} \leq \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i K^{\frac{1}{3}}}.$$

Now, the second part of the proof may be demonstrated. By Equation (26), it follows also that

$$s_1^{i,k}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||] \leq \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*})] - \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,*})] + \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i(k+1)^{\frac{4}{3}}}.$$

As in the previous part, by dividing both sides by Γ^{k+1}, summing up the inequalities and noting $\bar{f^k} = \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*})]$, we obtain

$$\sum_{k=0}^{K-1}\frac{s_1^{i,k}}{\Gamma^{k+1}}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||] \leq \sum_{k=0}^{K-1}\frac{\bar{f^k} - \bar{f^{k+1}}}{\Gamma^{k+1}} + \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i}\sum_{k=0}^{K-1}\frac{1}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}}.$$

Then, again by dividing both sides by $\sum_{k=0}^{K-1}\frac{s_1^{i,k}}{\Gamma^{k+1}}$ it follows that

$$\begin{aligned}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||] &= \frac{\sum_{k=0}^{K-1}\frac{s_1^{i,k}}{\Gamma^{k+1}}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||]}{\sum_{k=0}^{K-1}\frac{s_1^{i,k}}{\Gamma^{k+1}}} \\ &\leq \frac{1}{\sum_{k=0}^{K-1}\frac{s_1^{i,k}}{\Gamma^{k+1}}}\left(\sum_{k=0}^{K-1}\frac{\mathbb{E}[\bar{f^k} - \bar{f^{k+1}}]}{\Gamma^{k+1}} + \frac{2\rho n\sqrt{n}L_0(F)}{\beta^i}\sum_{k=0}^{K-1}\frac{1}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}}\right),\end{aligned}$$

where R is a random variable whose the distribution is given in the statement of the theorem. Now, as in Equation (2.21) of [5], the following hold

$$\sum_{k=0}^{K-1} \frac{\bar{f^k} - \bar{f^{k+1}}}{\Gamma^{k+1}} \leq \bar{f^0} + \sum_{k=1}^{K-1} \frac{2}{\Gamma^{k+1}(k+1)} \bar{f^k} \qquad \text{and} \qquad \sum_{k=0}^{K-1} \frac{s_1^{i,k}}{\Gamma^{k+1}} = \frac{\rho}{\Gamma^K}.$$

Thus, by substituting these in the inequality involving the expectation we obtain

$$\begin{aligned}
\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||] &\leq \frac{\Gamma^K}{\rho} \left(\mathbb{E}[\bar{f^0}] + \sum_{k=1}^{K-1} \frac{2}{\Gamma^{k+1}(k+1)} \mathbb{E}[\bar{f^k}] + \frac{2\rho n\sqrt{n} L_0(F)}{\beta^i} \sum_{k=0}^{K-1} \frac{1}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}} \right) \\
&\leq \frac{\Gamma^K}{\rho} \left(\mathbb{E}[\bar{f^0}] + 4\rho n\sqrt{n} L \sum_{k=0}^{K-1} \frac{1}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}} \right) \\
&\leq \frac{2L_0(F)}{K^2} + \frac{4n\sqrt{n} L_0(F)}{\beta^i K^{\frac{1}{3}}},
\end{aligned}$$

where the second inequality follows from Equation (24). □

4.1.6 Summary of convergence rates and complexity guarantees

The result obtained in Equation (22) is consistent with the convergence results of other ZO methods. To gain a better understanding of its performance, this result is compared with those of four other algorithms from different perspectives: the assumptions, the measure used, the convergence rate, and the function query complexity. All methods seek a solution to a stochastic optimization problem; the comparison is presented in Table 1. Since the convergence rate of the ZO-Signum and ZO-signSGD algorithms is measured using $||\nabla f(\mathbf{x})||$, but $||\nabla f(x)||^2$ is used by ZO-adaMM and ZO-SGD, Jensen's inequality is used to rewrite convergence rates in term of gradient norm.

- for ZO-SGD [18]

$$\mathbb{E}[||\nabla f(\mathbf{x})||] \leq \sqrt{\mathbb{E}[||\nabla f(\mathbf{x})||^2]} \leq \sqrt{O\left(\frac{\sigma\sqrt{n}}{\sqrt{K}} + \frac{n}{K}\right)} \leq O\left(\frac{\sqrt{\sigma} n^{\frac{1}{4}}}{K^{\frac{1}{4}}} + \frac{\sqrt{n}}{\sqrt{K}}\right),$$

- for ZO-adaMM [13]

$$\begin{aligned}
\mathbb{E}[||\nabla f(\mathbf{x})||] \leq \sqrt{\mathbb{E}[||\nabla f(\mathbf{x})||^2]} &\leq \sqrt{O\left(\left(\frac{n}{\sqrt{K}} + \frac{n^2}{K}\right)\sqrt{\ln(K) + \ln(n)}\right)} \\
&\leq O\left(\left(\frac{\sqrt{n}}{K^{\frac{1}{4}}} + \frac{n}{\sqrt{K}}\right)(\ln(K) + \ln(n))^{\frac{1}{4}}\right),
\end{aligned}$$

where the third inequalities are due to $\sqrt{a^2 + b^2} \leq a + b$, for $a, b \geq 0$. For ZO-signSGD, unless the value of b depends on K, the algorithm's convergence is only guaranteed within some ball around the solution, making it difficult to compare with other methods. Thus in the non convex case, after this transformation, it becomes apparent that ZO-signum has a convergence rate of $O\left(\frac{n^{\frac{3}{4}}}{\sqrt{\sigma}}\right)$ and $O(\sqrt{n})$ worse than that of ZO-SGD and ZO-adaMM, respectively. This may be attributed to the milder assumption made on the function F in the present work, which also explains why the convergence is relative to f^β. In the convex case, ZO-signum has a convergence rate of $O\left(\frac{nK^{\frac{1}{6}}}{\sigma}\right)$ worse than ZSCG and $O\left(\sqrt{n}K^{\frac{1}{6}}\right)$ worse than ZO-SGD. This may be explained because the $sign(\cdot)$ operator looses the magnitude information of the gradient when it applied. This problem may be fixed as in [21] but it outside the scope of this work. Finally, all methods but ZO-signSGD are momentum-based versions of

the original ZO-SGD method. Although the momemtum-based versions are mostly used in practice, it is interesting to notice that none of these methods possess a better convergence rate than the original ZO-SGD method. The next section provides some clues on the interests of the momemtum-based method.

Table 1: Summary of convergence rate and query complexity of various ZO-algorithms given K iterations.

Method	Assumptions	Measure	Convergence rate	Queries
ZO-SGD [18]	$F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{1+}$ $\mathbb{E}[\|\|\nabla F(\mathbf{x},\boldsymbol{\xi}) - \nabla f(\mathbf{x})\|\|^2] \leq \sigma^2$	$\mathbb{E}[\|\|\nabla f(\mathbf{x}^R)\|\|_2]$	$O\left(\frac{\sqrt{\sigma} n^{\frac{1}{4}}}{K^{\frac{1}{4}}} + \frac{\sqrt{n}}{\sqrt{K}}\right)$	$O(K)$
ZO-signSGD [27]	$F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{0+}$ $F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{1+}$ $\|\|\nabla F(\mathbf{x},\boldsymbol{\xi})\|\|_2 \leq \eta$	$\mathbb{E}[\|\|\nabla f(\mathbf{x}^R)\|\|_2]$	$O\left(\frac{\sqrt{n}}{\sqrt{K}} + \frac{\sqrt{n}}{\sqrt{b}} + \frac{n}{\sqrt{bq}}\right)$	$O(bqK)$
ZO-adaMM [13]	$F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{0+}$ $F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{1+}$ $\|\|\nabla F(\mathbf{x},\boldsymbol{\xi})\|\|_\infty \leq \eta$	$\mathbb{E}[\|\|\nabla f(\mathbf{x}^R)\|\|_2]$	$O\left(\left(\frac{\sqrt{n}}{K^{\frac{1}{4}}} + \frac{n}{\sqrt{K}}\right)(\ln(K) + \ln(n))^{\frac{1}{4}}\right)$	$O(K)$
ZO-Signum	$F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{0+}$	$\mathbb{E}[\|\|\nabla f^\beta(\mathbf{x}^R)\|\|_2]$	$O\left(\frac{n}{K^{\frac{1}{4}}}\ln(K)\right)$	$O(K)$
ZO-Signum	$F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{0+}$, f convex	$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,K}) - f^{\beta^i}(\mathbf{x}^{i,*})]$	$O\left(\frac{n\sqrt{n}}{K^{\frac{1}{3}}}\right)$	$O(K)$
ZO-SGD [31]	$F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{0+}$, f convex	$\mathbb{E}[f(\mathbf{x}^{i,K}) - f(\mathbf{x}^{i,*})]$	$O\left(\frac{n}{\sqrt{K}}\right)$	$O(K)$
Modified ZSCG [5]	$F(\cdot,\boldsymbol{\xi}) \in \mathcal{C}^{1+}$, F convex $\mathbb{E}[\|\|\nabla F(\mathbf{x},\boldsymbol{\xi}) - \nabla f(\mathbf{x})\|\|^2] \leq \sigma^2$	$\mathbb{E}[f(\mathbf{x}^{i,K}) - f(\mathbf{x}^{i,*})]$	$O\left(\frac{\sigma\sqrt{n}}{\sqrt{K}}\right)$	$O(K)$

4.2 Convergence of the SSO algorithm

The convergence analysis from the previous subsection is in expectation, i.e., it establishes the expected convergence performance over many executions of the ZO-Signum algorithm. As in [18], we now focus on the performance of a single run. Unlike [18], our analysis is based on a sequential optimization framework rather than a post-optimization process. Our SSO algorithm uses the norm of the momentum as an indicator of the quality of the current solution. In order to analyze the rate of convergence of this algorithm, the following additional assumptions are made on the function F. The first assumption concerns the smoothness of the function F.

Assumption 2. The function $F(\cdot,\boldsymbol{\xi})$ has $L_1(F)$-Lipschitz continuous gradient.

The second assumption concerns the local convexity of the function f^β.

Assumption 3. Let $(\mathbf{x}^{i,0})$ be a sequence of points produced by Algorithm 2 and $\mathbf{x}^{i,*}$ a sequence of local minima of f^{β^i}. We assume that there exists a threshold $I \in \mathbb{N}$ and a radius $\rho > 0$ such that $\forall i \geq I$:

1. f^{β^i} is convex on the ball $\mathcal{B}_\rho(\mathbf{x}^{i,*}) := \{\mathbf{x} \in \mathbb{R}^n : ||x - x^{i,*}|| < \rho\}$;
2. $\mathbf{x}^{i,0} \in \mathcal{B}_\rho(\mathbf{x}^{i,*})$.

Under these assumptions, we will prove that if the norm of the momentum vector $\mathbf{m}$ is below some threshold, then this threshold can be used to bound the norm of the gradient. Second, an estimate for the number of iterations required to reduce the norm of $\mathbf{m}$ below the threshold is provided. The next lemma is simply technical and demonstrates the link between $\bar{\mathbf{m}}$ and $\mathbf{m}$.

Lemma 4.12. For any subproblem $i \in \mathbb{N}$ and iteration $k \geq 1$, the following equality holds

$$\mathbb{E}[\mathbf{m}^{i,k}|\mathbf{x}^{i,k-1}] = \mathbb{E}[\bar{\mathbf{m}}^{i,k}|\mathbf{x}^{i,k-1}],$$

where $\bar{\mathbf{m}}^{i,k}$ is defined recursively in Proposition 4.1.

Proof. The proof is conducted by induction on k. For $k = 1$, setting $\mathbf{m}^{i,0} = \tilde{\nabla} f^{\beta^i}(\mathbf{x}^i, 0)$ implies

$$\mathbf{m}^{i,1} = s_2^{i,0}\tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,0}) + (1 - s_2^{i,0})\mathbf{m}^{i,0} = \tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,0}).$$

In the same way, $\bar{\mathbf{m}}^{i,1} = \nabla f^{\beta^i}(\mathbf{x}^{i,0})$. Thefore, we have

$$\mathbb{E}[\mathbf{m}^{i,1}|\mathbf{x}^{i,0}] = \mathbb{E}[\tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,0})|x^{i,0}] = \nabla f^{\beta^i}(\mathbf{x}^{i,0}) = \mathbb{E}[\nabla f^{\beta^i}(\mathbf{x}^{i,0})|\mathbf{x}^{i,0}] = \mathbb{E}[\bar{\mathbf{m}}^{i,1}|\mathbf{x}^{i,0}].$$

Now, suppose that the induction assumption is true for a given $k \in \mathbb{N}$, then

$$\mathbb{E}[\mathbf{m}^{i,k+1}|\mathbf{x}^{i,k}] = s_2^{i,k}\nabla f^{\beta^i}(\mathbf{x}^{i,k}) + (1 - s_2^{i,k})\mathbb{E}[\mathbf{m}^{i,k}|\mathbf{x}^{i,k}].$$

Now, by the law of total expectation

$$\begin{aligned}
\mathbb{E}[\mathbf{m}^{i,k}|\mathbf{x}^{i,k}] &= \mathbb{E}[\mathbb{E}[\mathbf{m}^{i,k}|\mathbf{x}^{i,k}, \mathbf{x}^{i,k-1}]|\mathbf{x}^{i,k}] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{m}^{i,k}|\mathbf{x}^{i,k-1}]|\mathbf{x}^{i,k}] \\
&= \mathbb{E}[\mathbb{E}[\bar{\mathbf{m}}^{i,k}|\mathbf{x}^{i,k-1}]|\mathbf{x}^{i,k}] \quad \text{(by the induction assumption)} \\
&= \mathbb{E}[\bar{\mathbf{m}}^{i,k}|\mathbf{x}^{i,k}].
\end{aligned}$$

Thus as $\mathbb{E}[\nabla f^{\beta^i}(\mathbf{x}^{i,k})|\mathbf{x}^{i,k}] = \nabla f^{\beta^i}(\mathbf{x}^{i,k})$, it follows that

$$\begin{aligned}
\mathbb{E}[\mathbf{m}^{i,k+1}|\mathbf{x}^{i,k}] &= s_2^{i,k}\nabla f^{\beta^i}(\mathbf{x}^{i,k}) + (1 - s_2^{i,k})\mathbb{E}[\mathbf{m}^{i,k}|\mathbf{x}^{i,k}] \\
&= s_2^{i,k}\mathbb{E}[\nabla f^{\beta^i}(\mathbf{x}^{i,k})|\mathbf{x}^{i,k}] + (1 - s_2^{i,k})\mathbb{E}[\bar{\mathbf{m}}^{i,k}|\mathbf{x}^{i,k}] \\
&= \mathbb{E}[\bar{\mathbf{m}}^{i,k+1}|\mathbf{x}^{i,k}],
\end{aligned}$$

which completes the proof. □

The following lemma shows that if $||\mathbf{m}||$ is below a certain threshold, then this threshold can be used to bound the norm of the gradient almost surely.

Lemma 4.13. For a subproblem $i \in \mathbb{N}$, let $K_i \in \mathbb{N}$ denote the first iteration in Algorithm 1 for which $||\mathbf{m}^{i,K_i}|| \leq \frac{L\beta^i}{4\beta^0}$, then, under Assumption 2 the norm of the gradient of the function f^{β^i} at $\mathbf{x}^{i,K}$ may be bounded as follows

$$||\nabla f^{\beta^i}(\mathbf{x}^{i,K_i})|| \leq \frac{L\beta^i}{4\beta^0} + 10nL_1(F)\frac{s_1^{i,0}}{s_2^{i,0}K_i^{\alpha_1-\alpha_2}}.$$

Moreover, if the problem $i+1$ is considered, the gradient of the function $f^{\beta^{i+1}}$ may be bounded at the point $\mathbf{x}^{i,K} = \mathbf{x}^{i+1,0}$ as follows

$$||\nabla f^{\beta^{i+1}}(\mathbf{x}^{i+1,0})|| \leq ||\nabla f^{\beta^i}(\mathbf{x}^{i,K_i})|| + L_0(F)\sqrt{n}|\beta^{i+1} - \beta^i|.$$

Proof. Let K_i be taken as in the statement of the proposition. The norm of the gradient may be bounded as follows

$$\begin{aligned}
||\nabla f^{\beta^i}(\mathbf{x}^{i,K_i})|| &\leq ||\mathbb{E}[\mathbf{m}^{i,K_i}|\mathbf{x}^{i,K_i}]|| + ||\nabla f^{\beta^i}(\mathbf{x}^{i,K_i}) - \mathbb{E}[\mathbf{m}^{i,K_i}|\mathbf{x}^{i,K_i}]|| \\
&\leq \mathbb{E}[||\mathbf{m}^{i,K_i}||\ |\mathbf{x}^{i,K_i}] + ||\nabla f^{\beta^i}(\mathbf{x}^{i,K_i}) - \mathbb{E}[\bar{\mathbf{m}}^{i,K_i}|\mathbf{x}^{i,K_i}]||,
\end{aligned}$$

where the second inequality follows from Jensen's inequality and Lemma 4.12. Now, using $||\mathbf{m}^{i,K_i}|| \leq \frac{L\beta^i}{4\beta^0}$, $\mathbb{E}[\nabla f^{\beta^i}(\mathbf{x}^{i,K})|\mathbf{x}^{i,K_i}] = \nabla f^{\beta^i}(\mathbf{x}^{i,K_i})$, $L_1(f^{\beta^i}) \leq L_1(F)$ and the result of Proposition 4.8 complete the first part of the proof

$$||\nabla f^{\beta^i}(\mathbf{x}^{i,K_i})|| \leq \frac{L\beta^i}{4\beta^0} + \mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,K_i}) - \bar{\mathbf{m}}^{i,K_i}||\ |\mathbf{x}^{i,K_i}]$$

$$\leq \frac{L\beta^i}{4\beta^0} + 10nL_1(F)\frac{s_1^{i,0}}{s_2^{i,0}K_i^{\alpha_1-\alpha_2}}.$$

The second part of the proof follows directly by applying the triangular inequality and the result in Lemma 2.1.3. □

Under Assumption 3, the expected difference between the values of f^{β^i} at $\mathbf{x}^{i,0}$ and its optimal value is bounded in the next Lemma.

Lemma 4.14. Let I be the threshold from Assumption 3. If $i \geq I$, then

$$\mathbb{E}[f^{\beta^{i+1}}(\mathbf{x}^{i+1,0}) - f^{\beta^{i+1}}(\mathbf{x}^{i+1,*})] \leq \rho\left(\frac{L\beta^i}{4\beta^0} + 10nL_1(F)\frac{s_1^{i,0}}{s_2^{i,0}K_i^{\alpha_1-\alpha_2}} + L_0(F)\sqrt{n}|\beta^{i+1}-\beta^i|\right). \quad (27)$$

Proof. Convexity of the function f^{β^i} on the ball $\mathcal{B}_\rho(\mathbf{x}^{i,*})$ implies

$$\begin{aligned}\mathbb{E}[f^{\beta^{i+1}}(\mathbf{x}^{i+1,0}) - f^{\beta^{i+1}}(\mathbf{x}^{i+1,*})] &\leq \mathbb{E}[\langle \nabla f^{\beta^{i+1}}(\mathbf{x}^{i+1,0}), \mathbf{x}^{i+1,0} - \mathbf{x}^{i+1,*}\rangle] \\ &\leq \mathbb{E}[||\nabla f^{\beta^{i+1}}(\mathbf{x}^{i+1,0})||||\mathbf{x}^{i+1,0} - \mathbf{x}^{i+1,*}||].\end{aligned}$$

The result follows using the Lemmma 4.13 and since $\mathbf{x}^{i+1,0}$ belongs to the ball $\mathcal{B}^\epsilon(\mathbf{x}^{i,*})$. □

Moreover, an estimate on the number of iterations required to reduce the norm of the gradient below some threshold may be given.

Lemma 4.15. Under Assumptions 1, 2 and 3, for a subproblem $i > I$ and in the setting of Algorithm 2, let $s_2^{i,0} \in \mathbb{R}^+$ be such that $k = 1$ in Equations (14) and (19), $L = \max(L_0(F), L_1(F))$, $\alpha_1 = \frac{3}{4}$ and $\alpha_2 = \frac{1}{2}$. Then, for a uniformly randomly chosen $R \in [0, K_i]$, it follows that

$$\mathbb{P}\left(||\nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) \leq \frac{4\beta^0}{\beta^i K_i^{\frac{1}{4}}}(A^i + B^i),$$

where A^i and B^i are defined in Equation (28).

Proof. Markov's inequality implies that

$$\mathbb{P}\left(||\nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) \leq \frac{4\beta^0\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||]}{L\beta^i}.$$

Now, given the result of Theorem 4.9 with the specified value of α_1 and α_2 and the fact that $L_1(f^{\beta^i}) \leq L_1(F)$ together with Lemma 4.14, it follows that

$$\frac{4\beta^0\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||]}{L\beta^i} \leq \frac{4\beta^0}{\beta^i K_i^{\frac{1}{4}}}(A^i + B^i),$$

where

$$\begin{aligned}A^i &= \frac{\rho}{s_1^{i,0}}\left(\frac{\beta^{i-1}}{4\beta^0} + 10n\frac{s_1^{i-1,0}}{s_2^{i-1,0}K_{i-1}^{\frac{1}{4}}} + \sqrt{n}|\beta^{i-1}-\beta^i|\right) \\ B^i &= \frac{ns_1^{i,0}}{2}H_k^{(-\frac{3}{2})} + \ln(K_i)\left(6\sqrt{s_2^{i,0}}(n+4)\sqrt{n} + \frac{20ns_1^{i,0}}{s_2^{i,0}}\right),\end{aligned} \quad (28)$$

and K_i is the iteration number for subproblem i and $H_k^{(-\frac{3}{2})}$ is the generalized harmonic number. □

The following Lemma provides an estimate on the number of iterations required to bound the norm of the difference between $\mathbf{m}$ and the gradient below a certain threshold.

Lemma 4.16. For a subproblem $i \in \mathbb{N}$ and in the setting of Algorithm 2, let $s_2^{i,0} \in \mathbb{R}^+$ be such that $k = 1$ in Equations (14) and (19), $L = \max(L_0(F), L_1(F))$, $\alpha_1 = \frac{3}{4}$ and $\alpha_2 = \frac{1}{2}$. Then, for a uniformly randomly chosen $R \in [0, K_i]$, it follows that

$$\mathbb{P}\left(||\mathbf{m}^{i,R} - \nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) \leq \frac{4\beta^0}{\beta^i K_i^{\frac{1}{4}}}\left(3\sqrt{s_2^{i,0}}(n+4)\sqrt{n} + \frac{10ns_1^{i,0}}{s_2^{i,0}}\right).$$

Proof. By Markov's inequality, it follows that

$$\begin{aligned}\mathbb{P}\left(||\mathbf{m}^{i,R} - \nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) &\leq \frac{4\beta^0 \mathbb{E}[||\mathbf{m}^{i,R} - \nabla f^{\beta^i}(\mathbf{x}^{i,R})||]}{L\beta^i} \\ &= \frac{4\beta^0}{L\beta^i K_i}\sum_{k=0}^{K_i}\mathbb{E}[||\mathbf{m}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||] \\ &\leq \frac{4\beta^0}{\beta^i K_i^{\frac{1}{4}}}\left(3\sqrt{s_2^{i,0}}(n+4)\sqrt{n} + \frac{10ns_1^{i,0}}{s_2^{i,0}}\right),\end{aligned}$$

where the last inequality holds by Proposition 4.5 and 4.8 with $\alpha_1 = \frac{3}{4}$ and $\alpha_2 = \frac{1}{2}$. □

Finally, the main theorem of this section may be stated.

Theorem 4.17. Let Assumptions 1, 2 and 3 hold and let I be the threshold from Assumption 3. For $i \in \mathbb{N}$, set

$$\beta^i = \frac{1}{\sqrt{n}(i+1)^2}, s_1^{i,0} = \frac{1}{6n(i+1)^{3/2}} \text{ and } s_2^{i,0} = \frac{s_2}{(i+1)}$$

with s_2 so that Equations (14) and (19) are satisfied for $k = 1$. Moreover, let denote K_i the first iteration for which $||\mathbf{m}^{i,K_i}|| \leq \frac{L\beta^i}{4\beta^0}$ with $L = \max(L_0(F), L_1(F))$ and $\epsilon > 0$ be a desired accuracy and let $i^* \geq \sqrt{\frac{L}{\epsilon}} \geq I$. If for any $i \geq I, K_i \geq (i+1)^6$, then after at most

$$O\left(\frac{n^6 L^{7/2}}{\epsilon^{7/2}}\right)$$

function evaluations, the following inequality holds

$$||\nabla f^{\beta^{i^*}}(x^{i^*,0})|| \leq \epsilon. \tag{29}$$

Furthermore, when for any $i \in \mathbb{N}$, f^{β^i} is convex then under the same setting that Theorem 4.11 given in Equation (23), it follows that after at most

$$O\left(\frac{n^{\frac{9}{2}} L^{7/2}}{\epsilon^{7/2}}\right)$$

function evaluations, the inequality in Equation (29) holds.

Proof. For a subproblem $i \in \mathbb{N}$, a probabilistic upper bound on the iteration $K_i \in \mathbb{N}$ such that $||\mathbf{m}^{i,K_i}|| \leq \frac{L\beta^i}{4\beta^0}$ may be provided. We have

$$\begin{aligned}||\mathbf{m}^{i,K_i}|| &= \min_{k\in[0,K_i]}||\mathbf{m}^{i,k}|| \\ &\leq ||\mathbf{m}^{i,R}|| \\ &\leq ||\mathbf{m}^{i,R} - \nabla f^{\beta^i}(\mathbf{x}^{i,R})|| + ||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||,\end{aligned} \tag{30}$$

where $R \sim \mathcal{U}[0, K_i]$. Now, probabilistic upper bounds on the number K_i required to obtain that both terms in the right-hand side of the previous inequality are below $\frac{L\beta^i}{4\beta^0}$. For the first term of the right-hand side in Equation (30), using the specified value of $s_1^{i,0}$, $s_2^{i,0}$ and β^i, Lemma 4.16 ensures that

$$\begin{aligned}\mathbb{P}\left(||\mathbf{m}^{i,R} - \nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) &\leq \frac{4\beta^0}{\beta^i K_i^{\frac{1}{4}}}\left(3\sqrt{s_2^{i,0}}(n+4)\sqrt{n} + \frac{10ns_1^{i,0}}{s_2^{i,0}}\right) \\ &\leq O\left(\frac{n\sqrt{n}(i+1)^{\frac{3}{2}}}{K_i^{\frac{1}{4}}}\right).\end{aligned}$$

The second term of the right-hand side in Equation (30) depends on the value of I. For subproblems $i \leq I$, it follows by Markov's inequality and Theorem 4.9 that

$$\begin{aligned}\mathbb{P}\left(||\nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) &\leq \frac{4\beta^0}{L\beta^i}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||] \\ &\leq \frac{4\beta^0}{\beta^i}\left(\frac{D_f^i}{s_1^{i,0}} + \frac{ns_1^{i,0}}{2}H_k^{(-\frac{3}{2})} + \ln(K_i)\left(6\sqrt{s_2^{i,0}}(n+4)\sqrt{n} + \frac{40s_1^{i,0}n}{s_2^{i,0}}\right)\right) \\ &\leq O\left(\frac{\max\left(\frac{n(i+1)^{\frac{7}{2}}}{L}, n\sqrt{n}\ln(K_i)(i+1)^{\frac{3}{2}}\right)}{K_i^{\frac{1}{4}}}\right).\end{aligned}$$

For subproblems $i > I$, Lemma 4.15 ensures that

$$\mathbb{P}\left(||\nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) \leq \frac{4\beta^0}{\beta^i K^{\frac{1}{4}}}(A^i + B^i),$$

where A^i and B^i are given in Equation (28). Now, given condition on K_i, it follows that

$$\begin{aligned}A^i &= \rho n(i+1)^{3/2}\left(\frac{1}{i^2} + \frac{6}{s_2 i^2} + \frac{2}{i^2(i+1)}\right) \text{ and} \\ B^i &= \frac{H_k^{(-\frac{3}{2})}}{2(i+1)^{3/2}} + \ln(K_i)\left(\frac{6n\sqrt{n+3}\sqrt{s_2}}{\sqrt{i+1}} + \frac{12}{s_2\sqrt{i+1}}\right).\end{aligned}$$

Thus, we obtain

$$\mathbb{P}\left(||\nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) \leq O\left(\frac{n\sqrt{n}(i+1)^{\frac{3}{2}}\ln(K_i)}{K_i^{\frac{1}{4}}}\right). \tag{31}$$

Therefore, to obtain $||\mathbf{m}^{i,K_i}|| \leq \frac{L\beta^i}{4\beta^0}$, it takes at most

$$K_i = \begin{cases} O\left(\max\left(n^4(i+1)^{14}, n^6(i+1)^6\right)\right) & \text{if } i \leq I \\ O\left(\left(n^6(i+1)^6\right)\right) & \text{otherwise,}\end{cases}$$

iterations. Thus, by taking $i^* \geq \sqrt{\frac{L}{\epsilon}}$, it follows that the number of iterations needed to reach the subproblem i^* is

$$\begin{aligned}\sum_{i=1}^{i^*} K_i &= \sum_{i=1}^{I} K_i + \sum_{i=I+1}^{i^*} K_i \\ &= O\left(\max\left(n^4(I+1)^{15}, n^6(I+1)^7\right)\right) + O(n^6(i^*)^7) \\ &= O\left(\frac{n^6 L^{7/2}}{\epsilon^{7/2}}\right),\end{aligned} \tag{32}$$

where I is a constant with respect to ϵ. Once this number of iterations is reached, it follows that $||\mathbf{m}^{i^*,0}|| \leq \frac{L}{(i^*+1)^2} \leq \epsilon$ and by Lemma 4.13

$$||\nabla f^{\beta^{i^*}}(\mathbf{x}^{i^*,K_{i^*}})|| \leq \frac{L}{(i^*+1)^2} + \frac{L}{\sqrt{i^*+1}(i^*)^{\frac{3}{2}}} \leq 2\epsilon.$$

For the second part of the proof, the bounds on Equation (30) does not depend on the value of I since f^{β^i} is assumed convex for any $i \in \mathbb{N}$. With the setting in Equation (23), it follows that

$$\mathbb{P}(||\nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}) \leq \frac{4\beta^0}{\beta^i}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,R})||] \leq 16\frac{n\sqrt{n}(i+1)^2}{K_i^{\frac{1}{3}}} \text{ and}$$

$$\mathbb{P}\left(||\mathbf{m}^{i,R} - \nabla f^{\beta^i}(\mathbf{x}^{i,R})|| \geq \frac{L\beta^i}{4\beta^0}\right) \leq \frac{4\beta^0}{L\beta^i} n\sqrt{n} L \frac{\sum_{k=0}^{K_i-1} \frac{2\rho}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}}}{\sum_{k=0}^{K_i-1} \frac{2\rho}{\Gamma^{k+1}(k+1)}} \leq 8\frac{n\sqrt{n}(i+1)^2}{K_i^{\frac{1}{3}}},$$

where the first inequality follows by Theorem 4.11 and the second one by the definition of the probabilty density of R together with Propositions 4.5 and 4.8. Therefore it takes at most $K_i = O(n^{\frac{9}{2}}(i+1)^6)$ to obtain $||\mathbf{m}^{i,K_i}|| \leq \frac{L\beta^i}{4\beta^0}$. Thus, by taking $i^* \geq \sqrt{\frac{L}{\epsilon}}$, it follows that the number of iterations neaded to reach the subproblem i^* is

$$\sum_{i=1}^{i^*} K_i = O(n^{\frac{9}{2}}(i^*)^7) = O\left(\frac{n^{\frac{9}{2}}L^{\frac{7}{2}}}{\epsilon^{\frac{7}{2}}}\right).$$

It remains to apply the Lemma 4.13 as previously to complete the proof. □

We would like to make few remarks about this theorem. First, in the algorithm, one approach to satisfy the condition $K_i \geq (i+1)^6$ for any $i \in \mathbb{N}$ is to incorporate it into the stopping criterion of Algorithm 1. However, due to the limited number of iterations in practice, this condition is typically replaced by a weaker one, $K_i \geq M$, where $M > 0$ is a constant. Second, the main result of Theorem 4.13 establishes an ϵ convergence rate for a single run of the SSO algorithm, which is the first of its kind to the best of our knowledge. This was made possible by decomposing the problem given in Equation (1) into a sequence of subproblems, each of which is solved using carefully chosen stopping criteria and step sizes. It is worth noting that, in [18], the (ϵ, Λ)-solution of the norm of the gradient is obtained after at most $O\left(\frac{nL^2\sigma^2}{\epsilon^4}\right)$. Although this bound has a weaker dependence on n and L, it is worse in terms of ϵ. Third, the first term in Equation (32) may be significant even if it is a fixed constant, particularly if the ball where the function becomes convex is difficult to reach, indeed this constant disappears when f^{β^i} is convex for any i. Nevertheless, the bounds given are the worst one, they may be considerably smaller in practice. Moreover, a way to decrease this term is to decrease the power on i in the denominator of β^i, $s_1^{i,0}$ and $s_2^{i,0}$ but it decrease the asymptotic rate of convergence. Finally, the process used in the SSO algorithm may be extended to other momentum-based methods and give an appealing property for these methods compared to the classical SGD.

5 Numerical experiments

The numerical experiments are conducted for two bounded constrained blackbox optimization problems. In order to handle the bound constraints $\mathbf{x} \in [\ell, \mathbf{u}] \subset \mathbb{R}^n$, the update in Equation (9) is simply projected such that $\mathbf{x} = \min(\ell, \max(\mathbf{x}, \mathbf{u}))$.

5.1 Application to a solar thermal powerplant

The first stochastic is SOLAR[2] [17], which simulates a thermal solar power plant and contains several instances allowing to choose the number of variables, the types of constraints and the objective function

[2]https://github.com/bbopt/solar

to optimize. All the instances of SOLAR are stochastics have nonconvex constraints and integer variables. In this work, the algorithms developed does not deal with this type of problem. Therefore the problem is slightly modified: the integer variables are fixed to their initial value and the problem aims to obtain a feasible solution by optimizing the remainding variables. Numerical experiments are conducted for the second instance of the SOLAR framework, which considers 12 variables (2 integers) and 12 constraints:

$$\min_{\mathbf{x}\in[0,1]^{12}} \mathbb{E}\left[\sum_{j=1}^{m} \max(0, c_j(\mathbf{x}, \boldsymbol{\xi}))^2\right]$$

where the c_j are the original stochastic constraints and the bound constraints have been normalized. The second instance of SOLAR is computationally expensive; a run may take between several seconds and several minutes. Therefore, the maximum number of function evaluations is set to 1000. Four algorithms are used:

- SSO, whose the hyperparameters values are given in Table 4. The search step given in Algorithm 2 is used for this experiment. A truncated version of the Gaussian gradient based estimate is used for this experiment.
- ZO-AdaMM [13] which is a the zeroth order version of the original Adam algorithm. This algorithm appears as one of the most effective according to [13, 28] in terms of distortion value, number of function evaluations and success rate. The default parameters defined experimentally in [13] are used on this problem except that $\beta = 0.05$ and the learning rate is equal to 0.3. Moreover, the same gradient estimator that ZO-Signum is used to eliminate its impact on the performance.
- CMA-ES [20] an algorithm based on biological inspired operators. Its name comes from the adaptation of the covariance matrix of the multivariate normal distribution used during the mutation. The version of CMA-ES used is the one of the pymoo [8] library with the default setting.
- The NOMAD software [26], which is based on the Mesh Adaptive Direct Search (MADS) [1] algorithm, a popular blackbox optimization solver.

The results are presented in Figure 2, which plots the average best result obtained by each algorithm with five different seeds. In this experiment, SSO obtains similar performance to NOMAD and CMAES which are state of the art algorithms for this type of problem. ZO-adaMM has difficulty to converge even though it is a ZO algorithm.

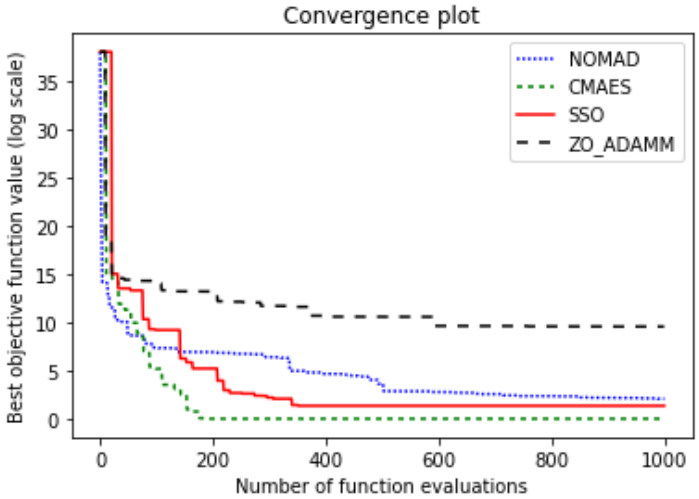

Figure 2: Average of 5 different seed runs for the NOMAD, CMAES, SSO and ZO adaMM algorithms.

5.2 Application to blackbox adversarial attack

This section demonstrates the competitiveness of the SSO algorithm through experiments involving the generation of blackbox adversarial examples for Deep Neural Networks (DNNs) [43]. Generating an adversarial example for a DNN involves adding a well-designed perturbation to the original legal input to cause the DNN to misclassify it. In this work, the attacker considers the DNN model to be unknown, hence the term blackbox. Adversarial attacks against DNNs are not just theoretical, they pose a real safety issue [32]. Having an algorithm that generates effective adversarial examples enables modification of DNN architecture to enhance its robustness against such attacks. An ideal adversarial example is one that can mislead a DNN to recognize it as any target image label, while appearing visually similar to the original input, making the perturbations indiscernible to human eyes. The similarity between the two inputs is typically measured by an ℓ_p norm. Mathematically, a blackbox adversarial attack can be formalized as follows. Let $(\mathbf{y}, \ell)$ denote a legitimate image $\mathbf{y}$ with the true label $\ell \in [1, M]$, where M is the total number of image classes. Let $\mathbf{x}$ denote the adversarial perturbation; the adversarial example is then given by $\mathbf{y}' = \mathbf{y} + \mathbf{x}$, and the goal is to solve the problem [13]

$$\begin{aligned} &\min_{\mathbf{x}} \lambda f(\mathbf{y}+\mathbf{x}) + ||\mathbf{x}||_2 \\ &\text{subject to } (\mathbf{y}+\mathbf{x}) \in [-0.5, 0.5]^n, \end{aligned}$$

where $\lambda > 0$ is a regularization parameter and f is the blackbox attack loss function. In our experiments, $\lambda = 10$ and the loss function is defined for untargeted attack [10], i.e,

$$f(\mathbf{y}') = \max\{Z(\mathbf{y}')_\ell - \max_{j \neq \ell} Z(\mathbf{y})_j, 0\},$$

where $Z(\mathbf{y}')_k$ denotes the prediction score of class k given the input $\mathbf{y}'$. Thus, the minimum value of 0 is reached as the perturbation succeeds to fool the neural network. The experiments of generating blackbox adversarial examples will be performed firstly on an adapted AlexNet [25] under the dataset Cifar10 and secondly on InceptionV3 [41] under the dataset ImageNet [16]. Since the NOMAD algorithm is not suitable to the size of this problem, three algorithm are compared : SSO (without search), ZO-ADAMM and CMA-ES. In the experiments, the hyperparameters of the algorithm ZO-adaMM are taken as in [13], those of SSO are given in Table 4 and the uniform gradient based estimate is used for both algorithms. Moreover, for the Cifar10 dataset, different initial learning rates for ZO-adaMM are used to observe its influence on the success rate. The experiments are conducted for 100 randomly selected images with a starting point corresponding to a null distortion, the maximum number of function queries is set to 5000. Thus, as the iteration increases, the attack loss decreases until it converges to 0 (indicating a successful attack) while the norm of the distortion could increase. In this sense, the best attack performance should correspond to the best trade-off between a fast convergence to a 0 attack loss in term of function evaluations, a high rate of success, and a low distortion (evaluated by the ℓ_2-norm).

The results for the Cifar10 dataset are given in Table 2. Except for ZO-adaMM with an initial learning rate equal to 0.01, all the algorithms achieve to have success rate above 95%. Among these algorithms, ZO-adaMM with a learning rate equal to 0.05, has the best convergence rate in terms of function evaluations but has the worst value of distortion. On the contrary, CMA-ES obtains the best value of distortion but has the worst convergence rate. The SSO algorithm obtains balanced results, and is the only one that reaches full success rate.

For the ImageNet dataset, the results are given in Table 3. Only two algorithms are compared since the dimension is too large to consider inverting the covariance matrix in CMA-ES. For this dataset, ZO-adaMM and SSO have the same convergence rate. However, SSO outperforms ZO-adaMM in term of success rate while having a slightly higher level of distortion.

Table 2: Results of blackbox adversarial attack for the Cifar10 dataset ($n = 3 \times 32 \times 32$)

Method	Attack success rate	$\|\|\ell_2\|\|$ first success	Average # of function evaluations
ZO-adaMM $lr = 0.01$	79 %	0.14	582
ZO-adaMM $lr = 0.03$	96%	0.97	310
ZO-adaMM $lr = 0.05$	98%	2.10	215
CMAES $\sigma = 0.005$	99%	0.33	862
SSO	100%	0.55	442

Table 3: Results of blackbox adversarial attack for the ImageNet dataset ($n = 3 \times 299 \times 299$)

Method	Attack success rate	$\|\|\ell_2\|\|$ first success	Average # of function evaluations
ZO-adaMM $lr = 0.01$	59 %	19	1339
SSO	73 %	33	1335

6 Concluding remarks

This paper presents a method for stochastic blackbox optimization in scenarios where function evaluations are computionally expensive. The approach relies on using zeroth-order gradient estimates, which offer three main advantages. Firstly, they require only a small number of function evaluations to estimate the gradient, regardless of the problem's dimension. Secondly, under mild conditions on the noised objective function, the problem can be formulated as optimizing a smoothed functional. Thirdly, the smoothed functional, with respect to the value of the smoothing parameter, may appear to be locally convexified near local minima.

Based on these three features, the SSO algorithm was developed. This algorithm is a sequential one and comprises two steps. The first is an optional search step that improves the exploration of the decision variable space and the algorithm's efficiency. The second is a local search which ensures the convergence of the algorithm. In this step, the original problem is decomposed into subproblems solved by a ZO-version of a sign stochastic descent with momentum algorithm. As the momentum is an exponential moving average of the gradient estimates, it is used to consider a subproblem approximately solved. More specifically, when the momentum's norm falls below a certain threshold that depends on the smoothing parameter, the subproblem is considered solved. The smoothing parameter's value is then decreased, and the SSO algorithm moves on to the next subproblem.

A theoretical analysis of the algorithm is conducted. Firstly, under Lipschitz continuity of the noisy function, a convergence rate in expectation of the ZO-Signum algorithm is derived. Secondly, under additional assumptions of smoothness and convexity or local convexity of objective function near its minima, an convergence rate of the SSO algorithm to an ϵ-optimal point of the problem is derived, which is, to the best of our knowledge, the first of its kind.

Finally, numerical experiments are conducted on a solar power plant simulation and adversarial blackbox attacks. Both examples are computationally expensive, the former is a lower size problem ($n \approx 10$) and the latter is larger size problem (up to $n \approx 10^5$). The results demonstrate the SSO algorithm's competitiveness in both performance and convergence rate compared to state-of-the-art algorithms. Further work will be devoted to extending this approach to constrained stochastic optimization.

Appendix A Proof of Proposition 4.1

Proposition A.1 ([6]). For the subproblem $i \in \mathbb{N}$, under Assumption 1 and in the setting of Algorithm 1, we have

$$\begin{aligned} s_1^{i,k}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] \leq & \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,k+1})] + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2 \\ & + 2s_1^{i,k}\underbrace{\mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1]}_{\text{bias}} + 2s_1^{i,k}\sqrt{n}\sqrt{\underbrace{\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||_2^2]}_{\text{variance}}} \end{aligned} \tag{33}$$

where $\bar{m}_j^{i,k+1}$ is defined recursively as $\bar{m}_j^{i,k+1} = s_2^{i,k}\nabla f^{\beta^i}(\mathbf{x}^{i,k}) + (1 - s_2^{i,k})\bar{m}_j^{i,k}$.

Proof. By $L_1(f^{\beta^i})$-Lipschitz smoothness of f^{β^i} (see Lemma 2.1.3) , it follows that

$$\begin{aligned} f^{\beta^i}(\mathbf{x}^{i,k+1}) \leq & f^{\beta^i}(\mathbf{x}^{i,k}) + \langle \nabla f^{\beta^i}(\mathbf{x}^{i,k}), \mathbf{x}^{i,k+1} - \mathbf{x}^{i,k}\rangle + \frac{L_1(f^{\beta^i})}{2}||\mathbf{x}^{i,k+1} - \mathbf{x}^{i,k}||_2^2 \\ = & f^{\beta^i}(\mathbf{x}^{i,k}) - s_1^{i,k}\langle \nabla f^{\beta^i}(\mathbf{x}^{i,k}), \text{sign}(\mathbf{m}^{i,k})\rangle + \frac{L_1(f^{\beta^i})(s_1^{i,k})^2}{2}||\text{sign}(\mathbf{m}^{i,k})||_2^2 \\ = & f^{\beta^i}(\mathbf{x}^{i,k}) - s_1^{i,k}||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1 + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2 \\ & + 2s_1^{i,k}\sum_{j=1}^{n}|\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\mathbf{1}\{\text{sign}(m_j^{i,k+1}) \neq \text{sign}(\nabla_j f^{\beta^i}(\mathbf{x}^{i,k}))\}, \end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Now, as in [6, 27], the expected improvement conditioned on $\mathbf{x}^{i,k}$ is given by

$$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,k})|\mathbf{x}^{i,k}] \leq -s_1^{i,k}||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1 + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2 \tag{34}$$

$$+ 2s_1^{i,k}\sum_{j=1}^{n}|\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\mathbb{E}[\mathbf{1}\{\text{sign}(m_j^{i,k+1}) \neq \text{sign}(\nabla_j f^{\beta^i}(\mathbf{x}^{i,k}))\}|\mathbf{x}^{i,k}]. \tag{35}$$

Again, as in [6, 27], the expectation that the sign of $m_j^{i,k+1}$ be different of the sign of $\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})$ is relaxed by considering that the set

$$\{m_j^{i,k+1} : \text{sign}(m_j^{i,k+1}) \neq \text{sign}(\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})\} \subset \{m_j^{i,k+1} : \ |m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})| \geq |\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\}.$$

Therefore, it follows that

$$\mathbb{E}[\mathbf{1}\{\text{sign}(m_j^{i,k+1}) \neq \text{sign}(\nabla_j f^{\beta^i}(\mathbf{x}^{i,k}))\}|\mathbf{x}^{i,k}] \leq \mathbb{E}[\mathbf{1}\{|m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})| \geq |\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\}|\mathbf{x}^{i,k}] \tag{36}$$

$$\leq \frac{\mathbb{E}[|m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\ |\mathbf{x}^{i,k}]}{|\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|}, \tag{37}$$

where the second inequality comes from conditional Markov's inequality. Substituting Equation (37) into Equation (35) and taking expectation over all the randomness we obtain

$$\begin{aligned} \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,k})] \leq & -s_1^{i,k}\mathbb{E}[||\nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1] + \frac{nL}{2}(s_1^{i,k})^2 \\ & + 2s_1^{i,k}\sum_{j=1}^{n}\mathbb{E}[|m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|]. \end{aligned} \tag{38}$$

Moreover, by adding and subtracting $\bar{\mathbf{m}}^{i,k+1}$ in the terms of the sum of Equation (38),

$$\sum_{j=1}^{n}\mathbb{E}[|m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|] = \mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1} + \bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1]$$

$$\leq \sqrt{n}\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||_2] + \mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1]$$
$$\leq \sqrt{n}\sqrt{\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||_2^2]} + \mathbb{E}[||\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||_1],$$

where the first inequality comes from $||\cdot||_1 \leq \sqrt{n}||\cdot||_2$ and the second one from Jensen's inequality. Finally, incorporating the last inequality in Equation (38) completes the proof. □

Appendix B List of hyperparameters for the SSO algorithm

Table 4: List of hyperparameters for the SSO algorithm

Problem	β^i	$s_1^{i,k}$	$s_2^{i,k}$	M	q
Cifar10	$\frac{0.005}{(i+1)^2}$	$\frac{0.005}{(i+1)^{\frac{3}{2}}\sqrt{k+1}}$	$\frac{0.9}{(i+1)(k+1)^{\frac{1}{4}}}$	60	10
ImageNet	$\frac{0.001}{(i+1)^2}$	$\frac{0.003}{(i+1)^{\frac{3}{2}}\sqrt{k+1}}$	$\frac{0.7}{(i+1)(k+1)^{\frac{1}{4}}}$	100	10
Solar	$\frac{0.3}{(i+1)^2}$	$\frac{0.1}{(i+1)^{\frac{3}{2}}\sqrt{k+1}}$	$\frac{0.5}{(i+1)(k+1)^{\frac{1}{4}}}$	5	10

References

[1] Audet, C., Dennis, Jr., J.: Mesh Adaptive Direct Search Algorithms for Constrained Optimization. SIAM Journal on Optimization 17(1), 188–217 (2006). DOI 10.1137/040603371. URL http://dx.doi.org/doi:10.1137/040603371

[2] Audet, C., Dzahini, K.J., Kokkolaras, M., Le Digabel, S.: Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. Computational Optimization and Applications 79(1), 1–34 (2021). DOI 10.1007/s10589-020-00249-0

[3] Audet, C., Hare, W.: Derivative-Free and Blackbox Optimization. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, Switzerland (2017). DOI 10.1007/978-3-319-68913-5. URL https://dx.doi.org/10.1007/978-3-319-68913-5

[4] Audet, C., Ihaddadene, A., Le Digabel, S., Tribes, C.: Robust optimization of noisy blackbox problems using the Mesh Adaptive Direct Search algorithm. Optimization Letters 12(4), 675–689 (2018). DOI 10.1007/s11590-017-1226-6. URL https://doi.org/10.1007/s11590-017-1226-6

[5] Balasubramanian, K., Ghadimi, S.: Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. Foundations of Computational Mathematics pp. 1–42 (2022)

[6] Bernstein, J., Wang, Y.X., Azizzadenesheli, K., Anandkumar, A.: Signsgd: Compressed optimisation for non-convex problems. In: International Conference on Machine Learning, pp. 560–569. PMLR (2018)

[7] Bhatnagar, S., Prasad, H., Prashanth, L.: Stochastic Recursive Algorithms for Optimization, *Lecture Notes in Control and Information Sciences*, vol. 434. Springer London, London (2013). DOI 10.1007/978-1-4471-4285-0. URL http://link.springer.com/10.1007/978-1-4471-4285-0

[8] Blank, J., Deb, K.: pymoo: Multi-objective optimization in python. IEEE Access 8, 89497–89509 (2020)

[9] Cai, H., Mckenzie, D., Yin, W., Zhang, Z.: Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. SIAM Journal on Optimization 32(2), 687–714 (2022)

[10] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy (sp), pp. 39–57. IEEE (2017)

[11] Chang, K.H.: Stochastic nelder–mead simplex method–a new globally convergent direct search method for simulation optimization. European journal of operational research 220(3), 684–694 (2012)

[12] Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models. Mathematical Programming 169(2), 447–487 (2018)

[13] Chen, X., Liu, S., Xu, K., Li, X., Lin, X., Hong, M., Cox, D.: Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. Advances in Neural Information Processing Systems 32 (2019)

[14] Conn, A., Scheinberg, K., Vicente, L.: Introduction to Derivative-Free Optimization. MOS-SIAM Series on Optimization. SIAM, Philadelphia (2009). DOI 10.1137/1.9780898718768. URL http://dx.doi.org/10.1137/1.9780898718768

[15] Curtis, F.E., Scheinberg, K., Shi, R.: A stochastic trust region algorithm based on careful step normalization. Informs Journal on Optimization 1(3), 200–220 (2019)

[16] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee (2009)

[17] Garneau, M.L.: Modelling of a solar thermal power plant for benchmarking blackbox optimization solvers. Master's thesis, École Polytechnique de Montréal (2015). URL https://publications.polymtl.ca/1996/. Available at https://publications.polymtl.ca/1996/

[18] Ghadimi, S., Lan, G.: Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization 23(4), 2341–2368 (2013)

[19] Ghadimi, S., Ruszczynski, A., Wang, M.: A single timescale stochastic approximation method for nested stochastic optimization. SIAM Journal on Optimization 30(1), 960–979 (2020)

[20] Hansen, N.: The CMA Evolution Strategy: A Comparing Review. In: J. Lozano, P. Larrañaga, I. Inza, E. Bengoetxea (eds.) Towards a New Evolutionary Computation, *Studies in Fuzziness and Soft Computing*, vol. 192, pp. 75–102. Springer Berlin Heidelberg (2006). DOI 10.1007/3-540-32494-1_4. URL http://dx.doi.org/10.1007/3-540-32494-1_4

[21] Karimireddy, S.P., Rebjock, Q., Stich, S., Jaggi, M.: Error feedback fixes signsgd and other gradient compression schemes. In: International Conference on Machine Learning, pp. 3252–3261. PMLR (2019)

[22] Kiefer, J., Jacob, J.W., et al.: Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23(3), 462–466 (1952)

[23] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[24] Kokkolaras, M., Mourelatos, Z.P., Papalambros, P.Y.: Impact of uncertainty quantification on design: an engine optimisation case study. International Journal of Reliability and Safety 1 (2006). DOI https://doi.org/10.1504/IJRS.2006.010786

[25] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6), 84–90 (2017)

[26] Le Digabel, S.: Algorithm 909: NOMAD: Nonlinear Optimization with the MADS algorithm. ACM Transactions on Mathematical Software 37(4), 44:1–44:15 (2011). DOI 10.1145/1916461.1916468. URL http://dx.doi.org/10.1145/1916461.1916468

[27] Liu, S., Chen, P.Y., Chen, X., Hong, M.: sign-sgd via zeroth-order oracle. In: International Conference on Learning Representations (2018)

[28] Liu, S., Chen, P.Y., Kailkhura, B., Zhang, G., Hero III, A.O., Varshney, P.K.: A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. IEEE Signal Processing Magazine 37(5), 43–54 (2020)

[29] Liu, S., Kailkhura, B., Chen, P.Y., Ting, P., Chang, S., Amini, L.: Zeroth-order stochastic variance reduction for nonconvex optimization. Advances in Neural Information Processing Systems 31 (2018)

[30] Maggiar, A., Wachter, A., Dolinskaya, I.S., Staum, J.: A derivative-free trust-region algorithm for the optimization of functions smoothed via gaussian convolution using adaptive multiple importance sampling. SIAM Journal on Optimization 28(2), 1478–1507 (2018)

[31] Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Foundations of Computational Mathematics 17(2), 527–566 (2017)

[32] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp. 506–519 (2017)

[33] Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y.L., Tan, J., Le, Q.V., Kurakin, A.: Large-scale evolution of image classifiers. In: International Conference on Machine Learning, pp. 2902–2911. PMLR (2017)

[34] Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)

[35] Rockafellar, R.T., Royset, J.O.: Risk measures in engineering design under uncertainty. In: Proc. International Conf. on Applications of Statistics and Probability in Civil Engineering (2015)

[36] Rubinstein, R.Y. (ed.): Simulation and the Monte Carlo Method. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA (1981). DOI 10.1002/9780470316511. URL http://doi.wiley.com/10.1002/9780470316511

[37] Rubinstein, R.Y.: Simulation and the Monte Carlo Method, 1st edn. John Wiley & Sons, Inc., USA (1981)

[38] Ruszczynski, A., Syski, W.: Stochastic approximation method with gradient averaging for unconstrained problems. IEEE Transactions on Automatic Control 28(12), 1097–1105 (1983)

[39] Spall, J.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control 37(3), 332–341 (1992). DOI 10.1109/9.119632. URL http://ieeexplore.ieee.org/document/119632/

[40] Styblinski, M., Tang, T.S.: Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing. Neural Networks 3(4), 467–483 (1990)

[41] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)

[42] Volz, V., Schrum, J., Liu, J., Lucas, S.M., Smith, A., Risi, S.: Evolving mario levels in the latent space of a deep convolutional generative adversarial network. In: Proceedings of the genetic and evolutionary computation conference, pp. 221–228 (2018)

[43] Xu, K., Liu, S., Zhao, P., Chen, P.Y., Z., H., Fan, Q., Erdogmus, D., Wang, Y., Lin, X.: Structured adversarial attack: Towards general implementation and better interpretability. In: International Conference on Learning Representations (2019). URL https://openreview.net/forum?id=BkgzniCqY7