Titre: Title:	Sequential stochastic blackbox optimization with zeroth-order gradient estimators
Auteurs: Authors:	Charles Audet, Jean Bigeon, Romain Couderc, & Michael Kokkolaras
Date:	2023
Type:	Article de revue / Article
Référence: Citation:	Audet, C., Bigeon, J., Couderc, R., & Kokkolaras, M. (2023). Sequential stochastic blackbox optimization with zeroth-order gradient estimators. AIMS Mathematics, 8(11), 25922-25956. https://doi.org/10.3934/math.20231321

AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(11): 25922–25956.
DOI:10.3934/math.20231321
Received: 23 May 2023
Revised: 11 August 2023
Accepted: 21 August 2023
Published: 08 September 2023

Research article

Sequential stochastic blackbox optimization with zeroth-order gradient estimators

Charles Audet[1], **Jean Bigeon**[2], **Romain Couderc**[1,3,*] **and Michael Kokkolaras**[4]

[1] GERAD, Department of Mathematical and Industrial Engineering, École Polytechnique de Montréal, Montréal, Québec, Canada

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

[3] GSCOP, Department of Industrial Engineering, Grenoble-Alpes University, Grenoble, France

[4] GERAD and Department of Mechanical Engineering, McGill University, Montréal, Canada

* **Correspondence:** Email: romain.couderc@polymtl.ca; Tel: +330646850267.

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), i.e., the original problem is decomposed into a sequence of subproblems. Each subproblem is solved by using a zeroth-order version of a sign stochastic gradient descent with momentum algorithm (i.e., ZO-signum) and with 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 the Lipschitz continuity assumption on the blackbox, a convergence rate in mean is derived for the ZO-signum algorithm. Moreover, if the blackbox is smooth and convex or locally convex around its minima, the rate of convergence 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.

Keywords: stochastic blackbox optimization; gradient approximation; sequential optimization; momentum-based method; convergence rate analysis
Mathematics Subject Classification: 65K05, 90C15, 90C30, 90C56, 90C90

1. Introduction

The present work targets stochastic blackbox optimization problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(x) \qquad \text{where} \qquad f(\mathbf{x}) := \mathbb{E}_{\xi}\left[F(\mathbf{x},\xi)\right], \tag{1.1}$$

and $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a blackbox [3] that takes two inputs: a vector of design variables $\mathbf{x} \in \mathbb{R}^n$ and a vector $\boldsymbol{\xi} \in \mathbb{R}^m$ that represents uncertainties with unknown distributions. The function F is called a stochastic zeroth-order oracle [20]. The objective function f is obtained by taking the expectation of F over all possible values of the uncertainties $\boldsymbol{\xi}$. This optimization problem can be found in two different fields. The first is in a machine learning framework wherein the loss function's gradient is unavailable or difficult to compute, such as in the optimization of neural network architecture [36], design of adversarial attacks [15] or game content generation [44]. The second field is when the function F is evaluated by means of a computational procedure [27]. In many cases, it depends on an uncertainty vector $\boldsymbol{\xi}$ due to environmental conditions, costs or effects of repair actions that are unknown [38]. Another source of uncertainty appears when the optimization is conducted at the early stages of the design process, where knowledge, information and data are 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 [16]: direct search and model-based methods. Algorithms corresponding to both methods have been adapted to a stochastic ZO oracle. Examples include the stochastic Nelder-Mead algorithm [13] and the stochastic versions of the mesh adaptive direct search algorithm [2, 4] for the direct search methods. For model-based methods, most studies consider extensions of the trust region method [14, 17, 33]. A major shortcoming of these methods is their difficulty to scale to large problems.

Recently, another class of methods, named ZO methods, has been attracting increasing amounts of attention. These methods use stochastic gradient estimators, which are based on the seminal work in [24,37], and they have been extended in [20,34,39,41]. 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. ZO methods take advantage of this property to extend first-order methods. For instance, the well known first-order methods conditional gradient, sign stochastic gradient descent (signSGD) [6] and adaptive momentum (ADAM) [26] have been extended to ZSCG [5], ZO-signSGD [30] and ZO-adaMM [15], respectively. More methods, not only based on first-order algorithms, have also emerged to solve regularized optimization problems [11], for very high dimensional blackbox optimization problems [9] and stochastic composition optimization problems [21]. Methods using second-order information based limited function queries have been developed [25]. Some methods handle situations in which the optimizer has only access to a comparison oracle that indicates which of two points has the highest value [10]. For an overview on ZO methods, readers may consult [31].

1.2. Motivation

Formally, stochastic gradient estimators involve a smooth approximation f^β (see Chapter 7.6 in [39]) which is a convolution product between f and a kernel $h^\beta(\mathbf{u})$

$$f^\beta(\mathbf{x}) := \int_{-\infty}^{\infty} h^\beta(\mathbf{u}) f(\mathbf{x}-\mathbf{u}) d\mathbf{u} = \int_{-\infty}^{\infty} h^\beta(\mathbf{x}-\mathbf{u}) f(\mathbf{u}) d\mathbf{u}. \tag{1.2}$$

The kernel must fulfill a set of conditions [39, pp. 263]:

(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 a probability density function, 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 of the smooth approximation are worth noting. First, the smooth approximation may be interpreted as a local weighted average of the function values in the neighborhood of $\mathbf{x}$. Condition 1.2 implies that it is possible to obtain a solution that is arbitrarily close to a local minimum f^*. Second, the smooth approximation 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 ∇f^β and may be constructed on the basis of observations of $F(\mathbf{x}, \xi)$ alone. Finally, the smooth approximation allows convexification of the original function f. Previous studies [39, 42] 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 [32] that if the smoothing parameter is too small, the difference in function values cannot be used to accurately represent the function differential, particularly when the noise level is significant.

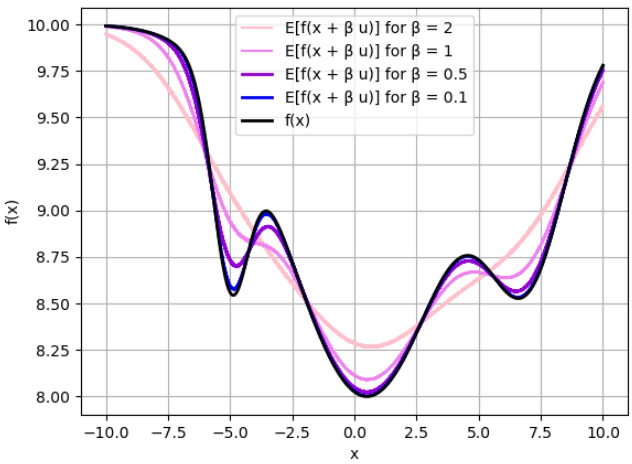

Figure 1. Curves of f^β for $u \sim \mathcal{N}(0, 1)$ and different values of β.

Although the two first properties of the smooth approximation are exploited by ZO methods, the last property has not been utilized since the work in [42]. This may be because the convexification phenomenon becomes insignificant when dealing with high-dimensional problems *. However, for problems of relatively small size ($n \simeq 10$), this property can be useful. The authors of [42] use an iterative algorithm to minimize the sequence of subproblems

$$\min_{\mathbf{x} \in \mathbb{R}^n} f^{\beta^i}(\mathbf{x}), \tag{1.3}$$

*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.

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 go from subproblem i to $i+1$ is arbitrarily fixed 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 motivate the research presented here.

1.3. Contributions

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

- A sequential stochastic optimization (SSO) algorithm is developed to solve the sequence of subproblems in Eq (1.3). In the inner loop, a subproblem is solved according to the ZO 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 it is driven to 0. Therefore, the smaller the value of β^i (and thus better the approximation given by f^{β^i}), the larger the computational budget allotted for the resolution of the subproblem.
- A theoretical analysis of this algorithm is conducted. First, the expectation of the norm of the momentum is proved to converge to 0, with a convergence rate that depends on the step sizes. Then, the convergence rate in mean of the ZO-signum algorithm toward 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 becomes convex around its local minima, the rate of convergence to an ϵ-optimal point is derived for the SSO algorithm.
- Numerical experiments were conducted to evaluate the performance of the proposed algorithm for two applications. First, a comparison is made with traditional derivative-free algorithms in terms of the optimization of the storage cost of a solar thermal power plant model, which is a low-dimensional problem. Second, a comparison is made with other ZO algorithms in order to generate blackbox adversarial attacks, which are large-sized 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 fixed value of* $\boldsymbol{\xi} = (\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$*, with the constant* $L_0(F) > 0$*, that is*

$$|F(\mathbf{x}, \boldsymbol{\xi}^1) - F(\mathbf{y}, \boldsymbol{\xi}^2)| \leq L_0(F)\|\mathbf{x} - \mathbf{y}\|.$$

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 bound the variance of the stochastic ZO oracle. It is worth noticing that no assumption is made regarding the differentiability of the objective function f or of its estimate F with respect to $\mathbf{x}$, contrary to most work on ZO methods.

Under Assumption 1, a smooth approximation of the function f may be constructed via 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

$$f^{\beta}(\mathbf{x}) := \frac{1}{(2\pi)^{\frac{n}{2}}} \int f(\mathbf{x} + \beta \mathbf{u}) e^{-\frac{\|u\|^2}{2}} d\mathbf{u} = \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \beta \mathbf{u})]. \tag{2.1}$$

This estimator has been studied in the literature (especially in [34]); it has the 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: $f^{\beta} \in C^{\infty}$.

(2) A one-sided unbiased estimator of ∇f^{β} is

$$\tilde{\nabla} f^{\beta}(\mathbf{x}) := \frac{\mathbf{u}(f(\mathbf{x} + \beta \mathbf{u}) - f(\mathbf{x}))}{\beta}. \tag{2.2}$$

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

$$\|\nabla f^{\beta^1}(\mathbf{x}) - \nabla f^{\beta^2}(\mathbf{x})\| \leq L_1(f^{\beta^1})(\beta^2 - \beta^1)(n+3)^{\frac{3}{2}}.$$

Moreover, for $\beta > 0$, f^{β} is $L_1(f^{\beta})$-smooth, i.e., $f^{\beta} \in C^{1+}$ 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 [34, Eq (22)].

(3) If $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, define the following for all $\mathbf{x} \in \mathbb{R}^n$

$$g(\mathbf{x}) = f^{\beta^1}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \beta^1 \mathbf{u})].$$

Let $\mu = \beta^2 - \beta^1 \geq 0$; it follows that for all $\mathbf{x} \in \mathbb{R}^n$

$$g^{\mu}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}}[g(\mathbf{x} + \mu \mathbf{u})] = \mathbb{E}_{\mathbf{u}}[f^{\beta^1}(\mathbf{x} + \mu \mathbf{u})] = \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \mu \mathbf{u} + \beta^1 \mathbf{u})] = \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \beta^2 \mathbf{u})] = f^{\beta^2}(\mathbf{x}).$$

Then, since by [34, Lemma 2] under Assumption 1, f^{β^1} is Lipschitz continuously differentiable, [34, Lemma 3] may be applied to the function g and it follows that

$$\|\nabla f^{\beta^1}(\mathbf{x}) - \nabla f^{\beta^2}(\mathbf{x})\| = \|\nabla g(\mathbf{x}) - \nabla g^{\mu}(\mathbf{x})\| \leq L_1(f^{\beta^1})\mu(n+3)^{\frac{3}{2}} = L_1(f^{\beta^1})(\beta^2 - \beta^1)(n+3)^{\frac{3}{2}}.$$

(4) See [34, pp. 5]. □

The estimator obtained in Eq (2.2) may be adapted to the stochastic ZO oracle F. For instance, a one-sided (mini-batch) estimator of the noised function F is

$$\tilde{\nabla} f^{\beta}(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{q} \sum_{j=1}^{q} \frac{\mathbf{u}^{j}(F(\mathbf{x} + \beta \mathbf{u}^{j}, \boldsymbol{\xi}^{j}) - F(\mathbf{x}, \boldsymbol{\xi}^{0}))}{\beta}, \tag{2.3}$$

where $\{u^j\}_{j=1}^q$ and $\{\boldsymbol{\xi}^j\}_{j=0}^q$ are q Gaussian random directional vectors and their associated q estimate values of the function F, respectively. This is still an unbiased estimator of ∇f^{β} because

$$\mathbb{E}_{\mathbf{u},\boldsymbol{\xi}}[\tilde{\nabla} f^{\beta}(\mathbf{x}, \boldsymbol{\xi})] = \mathbb{E}_{\mathbf{u}}[\mathbb{E}_{\boldsymbol{\xi}}[\tilde{\nabla} f^{\beta}(\mathbf{x}, \boldsymbol{\xi})|\mathbf{u}]] = \nabla f^{\beta}(\mathbf{x}). \tag{2.4}$$

The result of Lemma 2.1(3) is essential to understand why solving a sequence of optimization problems defined by Eq (1.3) may be efficient, although 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 Lemma 2.1(3), the norm of the gradient 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 for a subproblem may be reused in the subsequent subproblems since they are similar.
- A specific interest in the case of smooth approximation is the ability of using a larger value of β to solve 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 for an increase in the step size momentarily, in the hope of having a greater chance of escaping a local minimum.

3. SSO algorithm

Section 3.1 presents a ZO version of the signum algorithm [6] to solve Subproblem (1.3) for a given β^i and Section 3.2 presents the complete algorithm used to solve the sequential optimization problem.

3.1. The ZO signum algorithm

A ZO version of the signum algorithm (Algorithm 2 of [6]) is used to solve the subproblems. The signum algorithm is a momentum version of the sign-SGD algorithm. In [30], the authors extended the original sign-SGD algorithm to a ZO version of this algorithm. However, a ZO version of signum is not studied in the work of [30]. As the signum algorithm has been shown to be competitive with the ADAM algorithm [6], a ZO version of this algorithm seems interesting to consider. For completeness, the versions of the sign-SGD and the signum algorithms as they originally appeared in [6] are given in Appendix 6. There is an important difference between the original signum algorithm and its ZO version presented in Algorithm 1. Indeed, while the step size of the momentum $1 - s_2$ is kept constant in the work of [6], it is driven to 0 in our work.

Algorithm 1 ZO-signum 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 step-size 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 Eq (2.3)
  Update:

$$\mathbf{m}^{i,k+1} = s_2^{i,k} \tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,k}, \boldsymbol{\xi}^k) + (1 - s_2^{i,k})\mathbf{m}^{i,k} \tag{3.1}$$

$$x_j^{i,k+1} = x_j^{i,k} - s_1^k \operatorname{sign}(m_j^{i,k+1}) \; \forall j \in [1, n] \tag{3.2}$$

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

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] and [40]), 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 momentum is thus used as a stopping criterion.

3.2. The SSO algorithm

The optimization of the subproblem sequence described in Eq (1.3) is driven by the SSO algorithm presented in Algorithm 2. The value of β plays a critical role, as it serves as both the smoothing parameter and the stopping criterion for Algorithms 1 and 2. Algorithm 2 is inspired by the MADS algorithm [1] as it is based on two steps: a search step and a local step. The search step is optional, may use any heuristics and is required only for problems with relatively small dimensions. In Algorithm 2, an example of a search is given; it 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 for 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 guaranteed by the local step, since the search step is run only a finite number of times.

It is worth noting that the decreasing rate of β^i is chosen 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 toward a local optimum when $s_1^{0,0}$ and β^0 are chosen to be relatively large.

Algorithm 2 SSO algorithm.

1: **Initialization:**
2: Set $\mathbf{x}^{0,0} \in \mathbb{R}^n, \beta^0 > 0$ and N as the maximum number of function calls for the search step
3: Set q as the number of gradient estimates at each iteration of ZO-signum (ZOS) algorithm
4: Set M the minimum number of iterations made by the ZOS algorithm on a subproblem
5: C is the cache containing all of 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:

$$\begin{aligned} &\mathbf{m}^{i+1,0} = ZOS\,(\mathbf{x}^{i,0}, \mathbf{m}^{i,0}, \beta^i, s_1^{i,0}, s_2^{i,0}, L, q, M) \\ &\mathbf{x}^{i+1,0} \in \underset{\mathbf{x} \in C}{\operatorname{argmin}}\, F(\mathbf{x}, \boldsymbol{\xi}) \end{aligned}$$

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:

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

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

4. Convergence analysis

The convergence analysis is conducted in two steps: first the convergence rate in mean is derived for Algorithm 1 and then the rate of convergence to an ϵ-optimal point is derived for Algorithm 2.

4.1. Convergence rate of the ZO-signum algorithm

The analysis of Algorithm 1 follows the general methodology given in [6, Appendix E]. 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 mean 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. In order to better convey the convergence analysis of the ZO-signum algorithm, a hierarchical workflow of the different theoretical results is presented in Table 1. The main results are presented in Theorem 4.1 and its corollary for the non-convex case, and Theorem 4.2 for the convex case.

Table 1. Workflow of lemmas/propositions/theorems for the ZO-signum convergence analysis.

Assumptions on F	Preliminary results	Intermediate results	Main results	When f^β is convex
Assumption 1 which implies that $L_1(f^{\beta^i}) = \frac{2\sqrt{n}L_0(F)}{\beta^i}$		Proposition 4.1	Theorem 4.1 Corollary 4.1	Theorem 4.2
	Lemma 4.1 Lemma 4.2 Lemma 4.3	Proposition 4.2		
	Lemma 4.4 Lemma 4.5	Proposition 4.3		

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*

$$\begin{aligned} s_1^{i,k}\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] \leq \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,k+1})] + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2 \\ + 2s_1^{i,k}\underbrace{\mathbb{E}[\|\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1]}_{bias} + 2s_1^{i,k}\sqrt{n}\sqrt{\underbrace{\mathbb{E}[\|\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}\|_2^2]}_{variance}}, \end{aligned} \tag{4.1}$$

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 B. □

Now, it remains to bound the three terms on the right side of Inequality (4.1).

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 momentum to 0. It avoids the need to sample an increasing number of stochastic gradients at each iteration, which may be problematic, as noted in [30]. To achieve this, the variance term is first decomposed in terms of expectation of the squared norm of the stochastic gradient estimators $\tilde{g}$.

Lemma 4.1. *For the subproblem $i \in \mathbb{N}$, let $k \in \mathbb{N}$ and $j \in [1, n]$; we have*

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

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 Eq (2.3) 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

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

The expectation of this expression is

$$\begin{aligned}\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2] &= (s_2^{i,k})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||^2] + (1 - s_2^{i,k})^2\mathbb{E}[||\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k}||^2] \\ &\quad + 2s_2^{i,k}(1 - s_2^{i,k})\mathbb{E}[(\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(x^{i,k}))^T(\mathbf{m}^{i,k} - \bar{\mathbf{m}}^{i,k})]. \end{aligned} \tag{4.2}$$

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 $\mathcal{F}^{i,k}$ and because $\mathbb{E}[\tilde{\mathbf{g}}^{i,k}|\mathbf{x}^{i,k}] = \nabla f^{\beta^i}(\mathbf{x}^{i,k})$ where $\tilde{\mathbf{g}}^{i,k}$ is an unbiased estimator of the gradient by Eq (2.4). By substituting this result in (4.2), it follows that

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

By repeating this process iteratively, we obtain

$$\begin{aligned}\mathbb{E}[||\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}||^2] &= (s_2^{i,k})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,k} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})||^2] \\ &\quad + \sum_{r=0}^{k-1}(s_2^{i,r})^2\prod_{t=r}^{k-1}(1 - s_2^{i,t+1})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,r} - \nabla f^{\beta^i}(\mathbf{x}^{i,r})||^2] \\ &\quad + \prod_{t=0}^{k}(1 - s_2^{i,t})^2\mathbb{E}[||\tilde{\mathbf{g}}^{i,0} - \nabla f^{\beta^i}(x^{i,0})||^2].\end{aligned} \tag{4.3}$$

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 to Eq (4.3) completes the proof. □

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

Lemma 4.2. *Let $i \in \mathbb{N}$, $r \in [0, k]$ and $j \in [1, n]$; 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 Eq (2.3) with $q = 1$, it follows that

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{g}}^{i,r}\|^2] &= \mathbb{E}\left[\frac{\|u\|^2}{(\beta^i)^2}\left(F(\mathbf{x}^{i,r} + \beta^i \mathbf{u}, \boldsymbol{\xi}^1) - F(\mathbf{x}^{i,r}, \boldsymbol{\xi}^0)\right)^2\right] \\ &\leq L_0(F)^2 \mathbb{E}[\|u\|^4] \\ &\leq L_0(F)^2(n+4)^2\end{aligned}$$

where the first inequality follows from Assumption 1(b) and the second by [34, Lemma 1]. □

Finally, a technical lemma bounds the second term of the decomposition of Lemma 4.1 by a decreasing sequence. It achieves the same rate of convergence as in [6] without sampling any stochastic gradient.

Lemma 4.3. *For the subproblem $i \in \mathbb{N}$, let $s_2^{i,k}$ be 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*

$$\frac{k}{(k+1)^{\alpha_2}} \geq \frac{\ln(s_2^{i,0}) + (1+\alpha_2)\ln(k)}{s_2^{i,0}}, \tag{4.4}$$

the following inequality holds

$$\sum_{r=0}^{k-1} (s_2^{i,r})^2 \prod_{t=r}^{k-1} (1 - s_2^{i,t+1})^2 \leq \frac{9 s_2^{i,0}}{k^{\alpha_2}}. \tag{4.5}$$

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

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

$$\leq (s_2^{i,0})^2 k (1 - s_2^{i,k})^{2\lfloor k/2 \rfloor} + \frac{8 s_2^{i,0}}{k^{\alpha_2}(2 - s_2^{i,k})}.$$

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 us to complete the proof. □

Combining the three previous lemmas allows the bounding of the variance term in Proposition 4.1.

Proposition 4.2. *In the setting of Lemmas 4.2 and 4.3 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 4.1 and 4.2, it follows that

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

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 4.3. □

4.1.3. Bound on the bias term

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

Lemma 4.4. *For the subproblem $i \in \mathbb{N}$ and at iteration $k \in \mathbb{N}$ of the Algorithm 1, we have*

$$\mathbb{E}[\|\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] \leq 2n L_1(f^{\beta^i}) \left(\sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1} (1 - s_2^{i,t+1}) \right).$$

Proof. Foremost, observe that the quantity

$$S^{i,k} := \begin{cases} 1, & \text{if } k = 0, \\ s_2^{i,k} + \sum_{r=0}^{k-1} s_2^{i,r} \prod_{t=r}^{k-1}(1 - s_2^{i,t+1}) + \prod_{t=0}^{k}(1 - s_2^{i,t}), & \text{otherwise,} \end{cases} \tag{4.6}$$

may be written recursively as

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

Note that in its second expression $S^{i,k} = 1$ for all k. Therefore, by definition of $\bar{m}_j^{i,k}$ and the previous result on $S^{i,k}$, it follows that

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

Thus

$$\begin{aligned}
\mathbb{E}[\|\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] \leq& \sum_{r=0}^{k-1} s_2^{i,r} \prod_{t=r}^{k-1}(1 - s_2^{i,t+1})\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,r}) - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] \\
&+ \prod_{t=0}^{k}(1 - s_2^{i,t})\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,0}) - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1].
\end{aligned} \tag{4.7}$$

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 into Eq (4.7) gives

$$\mathbb{E}[\|\bar{m}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] \leq 2nL_1(f^{\beta^i})S_1^{i,k}, \tag{4.8}$$

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

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

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

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

which completes the proof. □

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

Lemma 4.5. *For the subproblem $i \in \mathbb{N}$, 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*

$$\frac{k}{(k+1)^{\alpha_2}} \geq \frac{2\left(\ln(s_2^{i,0}) + (1+\alpha_1-\alpha_2)\ln(k)\right)}{s_2^{i,0}}, \tag{4.9}$$

the following inequality holds

$$\sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1}(1-s_2^{i,t+1}) \leq \frac{5s_1^{i,0}}{s_2^{i,0}k^{\alpha_1-\alpha_2}}. \tag{4.10}$$

Proof. The proof follows the proof of Lemma 4.3. The sum is partitioned as follows:

$$\begin{aligned}
\sum_{l=0}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1}(1-s_2^{i,t+1}) &= \sum_{l=0}^{\lfloor k/2\rfloor-1} s_1^{i,l} \prod_{t=l}^{k-1}(1-s_2^{i,t+1}) + \sum_{l=\lfloor k/2\rfloor-1}^{k-1} s_1^{i,l} \prod_{t=l}^{k-1}(1-s_2^{i,t+1}) \\
&\leq (1-s_2^{i,k})^{\lfloor k/2\rfloor} \sum_{l=0}^{\lfloor k/2\rfloor-1} s_1^{i,l} + s_1^{i,\lfloor k/2\rfloor-1} \sum_{l=\lfloor k/2\rfloor-1}^{k-1} (1-s_2^{i,k})^{k-r-1} \\
&\leq s_1^{i,0}k(1-s_2^{i,k})^{\lfloor k/2\rfloor} + \frac{4s_1^{i,0}}{k^{\alpha_1}(1-(1-s_2^{i,k}))} \\
&= \frac{s_1^{i,0}s_2^{i,0}k(1-s_2^{i,k})^{\lfloor k/2\rfloor}}{s_2^{i,0}} + \frac{4s_1^{i,0}}{s_2^{i,0}k^{\alpha_1-\alpha_2}}.
\end{aligned}$$

Now, as in Lemma 4.3 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 for bounding of the bias term.

Proposition 4.3. *In the setting of Lemma 4.5, the bias term of Proposition 4.1 is bounded by*

$$\mathbb{E}[\|\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] \leq 10nL_1(f^{\beta^i})\frac{s_1^{i,0}}{s_2^{i,0}k^{\alpha_1-\alpha_2}}.$$

Proof. The proof is a straightforward consequence of Lemmas 4.4 and 4.5. □

4.1.4. Convergence rate in mean of the ZO-signum algorithm

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

Theorem 4.1. *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 Eqs (4.4) and (4.9); we have*

$$\begin{aligned}\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,R})\|_1] \leq & \frac{1}{K^{1-\alpha_1} - \frac{C}{K^{\alpha_1}}}\left(\frac{D_f^i}{s_1^{i,0}} + \frac{n\sqrt{n}L_0(F)s_1^{i,0}}{\beta^i}\sum_{k=C}^{K}\frac{1}{k^{2\alpha_1}} + 6\sqrt{s_2^{i,0}}L_0(F)\sqrt{n}(n+4)\sum_{k=C}^{K}\frac{1}{k^{\alpha_1+\frac{\alpha_2}{2}}}\right.\\ &\left. + \frac{40L_0(F)s_1^{i,0}n\sqrt{n}}{s_2^{i,0}\beta^i}\sum_{k=C}^{K}\frac{1}{k^{2\alpha_1-\alpha_2}}\right),\end{aligned} \tag{4.11}$$

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 $C \in \mathbb{N}$ satisfy Eqs (4.4) and (4.9) and, summing over the inequality in Proposition 4.1, it follows that

$$\begin{aligned}\sum_{k=C}^{K} s_1^{i,k}\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] \leq & \mathbb{E}[f^{\beta^i}(x^{i,C}) - f^{\beta^i}(\mathbf{x}^{i,K+1})] + \frac{nL_1(f^{\beta^i})}{2}\sum_{k=C}^{K}(s_1^{i,k})^2\\ & + 2\sqrt{n}\sum_{k=C}^{K} s_1^{i,k}\sqrt{\mathbb{E}[\|\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}\|_2^2]} + 2\sum_{k=C}^{K} s_1^{i,k}\mathbb{E}[\|\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1].\end{aligned}$$

By substituting the results of Propositions 4.2 and 4.3 in the previous inequality, we obtain

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

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 $L_1(f^{\beta^i}) = \frac{2\sqrt{n}L_0(F)}{\beta^i}$ (see [34, Lemma 2]) completes the proof. □

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

Corollary 4.1. *Under the same setting of Theorem 4.1 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*

$$\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,R})\|_2] = O\left(\frac{n^{\frac{3}{2}}}{K^{1/4}}\ln(K)\right). \tag{4.12}$$

Proof. The result is a direct consequence of Theorem 4.1 with the specified constant, and it can be obtained by noting that $\|\cdot\|_2 \leq \|\cdot\|_1$ in $\mathbb{R}^n$. □

In [15, 20, 30], the function F is assumed to be smooth with a 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 rate results for the ZO-signum algorithm has been derived in the non-convex case. In the next theorem, they are derived for the case when the function f^{β^i} is convex.

Theorem 4.2. *Under Assumption 1, suppose moreover that* f^{β^i} *is convex and there exists* ρ *such that* $\rho = \max_{k\in\mathbb{N}} \|\mathbf{x}^{i,k} - \mathbf{x}^{i,*}\|$*; then, by setting*

$$s_1^{i,k} = \frac{2\rho}{(k+1)},\ s_2^{i,k} = \frac{1}{(k+1)^{\frac{2}{3}}} \text{ and } \Gamma^k := \prod_{l=2}^{k}\left(1 - \frac{2}{k+1}\right) = \frac{2}{k(k+1)} \text{ with } \Gamma^1 = 1, \tag{4.13}$$

it follows that

$$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,K}) - f^{\beta^i}(\mathbf{x}^*)] \leq \frac{4\rho^2 n\sqrt{n}L_0(F)}{\beta^i K^{\frac{1}{3}}} \tag{4.14}$$

and

$$\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,R})\|] \leq \frac{2L_0(F)}{K^2} + \frac{4\rho n\sqrt{n}L_0(F)}{\beta^i K^{\frac{1}{3}}}, \tag{4.15}$$

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 Theorem 4.2, it follows by Proposition 4.1 that

$$\begin{aligned}\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,*})] &\leq \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*})] - s_1^{i,k}\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,k})\|] + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2 \\ &\quad + 2s_1^{i,k}\mathbb{E}[\|\bar{\mathbf{m}}^{i,k+1} - \nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] + 2s_1^{i,k}\sqrt{n}\sqrt{\mathbb{E}[\|\mathbf{m}^{i,k+1} - \bar{\mathbf{m}}^{i,k+1}\|^2]} \\ &\leq \mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k}) - f^{\beta^i}(\mathbf{x}^{i,*})] - s_1^k\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,k})\|] + \frac{4\rho^2 n\sqrt{n}L_0(F)}{\beta^i(k+1)^{\frac{4}{3}}},\end{aligned} \tag{4.16}$$

where the last inequality follows thanks to Propositions 4.2 and 4.3 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 ρ, 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 Eq (4.16), it follows that

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

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

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

Thus

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

Now, the second part of the proof may be demonstrated. By Eq (4.16), it also follows that

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

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,R})\|] &\leq \frac{\Gamma^K}{\rho}\left(\mathbb{E}[\bar{f}^0] + \sum_{k=1}^{K-1} \frac{2}{\Gamma^{k+1}(k+1)}\mathbb{E}[\bar{f}^k] + \frac{4\rho^2 n\sqrt{n}L_0(F)}{\beta^i}\sum_{k=0}^{K-1}\frac{1}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}}\right) \\
&\leq \frac{\Gamma^K}{\rho}\left(\mathbb{E}[\bar{f}^0] + \frac{8\rho n\sqrt{n}L_0(F)}{\beta^i}\sum_{k=0}^{K-1}\frac{1}{\Gamma^{k+1}(k+1)^{\frac{4}{3}}}\right) \\
&\leq \frac{2L_0(F)}{K^2} + \frac{8\rho n\sqrt{n}L_0(F)}{\beta^i K^{\frac{1}{3}}},
\end{aligned}$$

where the second inequality follows from Eq (4.14). □

4.1.6. Summary of convergence rates and complexity guarantees

The result obtained in Eq (4.12) 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 2. Since the convergence rates of the ZO-signum and ZO-signSGD algorithms are measured by using $\|\nabla f(\mathbf{x})\|$, although $\|\nabla f(x)\|^2$ is used for ZO-adaMM and ZO-SGD, Jensen's inequality is used to rewrite the convergence rates in terms of the gradient norm.

- for ZO-SGD [20]

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

- for ZO-adaMM [15]

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

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 that is $O\left(\frac{n^{\frac{3}{4}}}{\sqrt{\sigma}}\right)$ and $O(\sqrt{n})$

worse than those 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 that is $O\left(\frac{nK^{\frac{1}{6}}}{\sigma}\right)$ worse that of ZSCG and $O\left(\sqrt{n}K^{\frac{1}{6}}\right)$ worse that of ZO-SGD. This may be explained by the $sign(\cdot)$ operator loosing the magnitude information of the gradient when it is applied. This problem may be fixed as in [23] but it outside the scope of this work. Finally, all methods except ZO-signSGD are momentum-based versions of the original ZO-SGD method. Although the momentum-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 about the interests of the momentum-based method.

Table 2. Summary of convergence rates and query complexity for various ZO algorithms given K iterations.

Method	Assumptions	Measure	Convergence rate	Queries
ZO-SGD [20]	$F(\cdot,\boldsymbol{\xi}) \in C^{1+}$ $\mathbb{E}[\lVert\nabla F(\mathbf{x},\boldsymbol{\xi}) - \nabla f(\mathbf{x})\rVert^2] \leq \sigma^2$	$\mathbb{E}[\lVert\nabla f(\mathbf{x}^R)\rVert_2]$	$O\left(\frac{\sqrt{\sigma}n^{\frac{1}{4}}}{K^{\frac{1}{4}}} + \frac{\sqrt{n}}{\sqrt{K}}\right)$	$O(K)$
ZO-signSGD [30]	$F(\cdot,\boldsymbol{\xi}) \in C^{0+}$ $F(\cdot,\boldsymbol{\xi}) \in C^{1+}$ $\lVert\nabla F(\mathbf{x},\boldsymbol{\xi})\rVert_2 \leq \eta$	$\mathbb{E}[\lVert\nabla f(\mathbf{x}^R)\rVert_2]$	$O\left(\frac{\sqrt{n}}{\sqrt{K}} + \frac{\sqrt{n}}{\sqrt{b}} + \frac{n}{\sqrt{bq}}\right)$	$O(bqK)$
ZO-adaMM [15]	$F(\cdot,\boldsymbol{\xi}) \in C^{0+}$ $F(\cdot,\boldsymbol{\xi}) \in C^{1+}$ $\lVert\nabla F(\mathbf{x},\boldsymbol{\xi})\rVert_\infty \leq \eta$	$\mathbb{E}[\lVert\nabla f(\mathbf{x}^R)\rVert_2]$	$O\left(\left(\frac{\sqrt{n}}{K^{\frac{1}{4}}} + \frac{n}{\sqrt{K}}\right)(\ln(K) + \ln(n))^{\frac{1}{4}}\right)$	$O(K)$
ZO-Signum	$F(\cdot,\boldsymbol{\xi}) \in C^{0+}$	$\mathbb{E}[\lVert\nabla f^{\beta}(\mathbf{x}^R)\rVert_2]$	$O\left(\frac{n\sqrt{n}}{K^{\frac{1}{4}}}\ln(K)\right)$	$O(K)$
ZO-Signum	$F(\cdot,\boldsymbol{\xi}) \in C^{0+}$, 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 [34]	$F(\cdot,\boldsymbol{\xi}) \in C^{0+}$, 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]	$F(\cdot,\boldsymbol{\xi}) \in C^{1+}$, F convex $\mathbb{E}[\lVert\nabla F(\mathbf{x},\boldsymbol{\xi}) - \nabla f(\mathbf{x})\rVert^2] \leq \sigma^2$	$\mathbb{E}[f(\mathbf{x}^{i,K}) - f(\mathbf{x}^{i,*})]$	$O\left(\frac{\sigma\sqrt{n}}{\sqrt{K}}\right)$	$O(K)$

4.2. Convergence rate of the SSO algorithm

The convergence analysis from the previous subsection is in mean, i.e., it establishes the expected convergence performance over many executions of the ZO-signum algorithm. As in [20], we now focus on the performance of a single run. A second hierarchical workflow of the different theoretical results is presented in Table 3.

Table 3. Workflow of Lemmas/Propositions/Theorems for the SSO convergence analysis.

Assumptions on F	Preliminary results			Intermediate results	Main result	When f^{β} is convex
Assumptions 1, 2 and 3 which imply $L_1(f^{\beta^i}) \leq L_1(f)$	Lemma 4.6 Proposition 4.3 Lemma 2.1(3)	Lemma 4.7	Lemma 4.8	Lemma 4.9	Theorem 4.3 (i)	Theorem 4.3 (ii)
	Theorem 4.1 Proposition 4.2 Proposition 4.3			Lemma 4.10		

Unlike [20], 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 regarding the function F. The first assumption concerns the smoothness of the function F. The assumption of smoothness is used only to guarantee that $L_1(f^{\beta^i})$ is a constant with respect to β^i, contrary to the non-smooth case (see [34, Eq (12)]).

Assumption 2. *The function $F(\cdot,\boldsymbol{\xi})$ has a $L_1(F)$-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 4.6. *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 $\mathbf{m}^{i,0} = \tilde{\nabla} f^{\beta^i}(\mathbf{x}^{i,0},\boldsymbol{\xi}^0)$ implies that

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

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

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

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.

Lemma 4.7. *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 3 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 $\mathbf{x}^{i,K_i} = \mathbf{x}^{i+1,0}$ as follows:

$$\|\nabla f^{\beta^{i+1}}(\mathbf{x}^{i+1,0})\| \leq \|\nabla f^{\beta^i}(\mathbf{x}^{i,K_i})\| + L_1(F)(n+3)^{\frac{3}{2}}(\beta^i - \beta^{i+1}).$$

Proof. Let K_i be taken as in the statement of the lemma. 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 4.6. 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 4.3 completes the first part of the proof

$$\begin{aligned}\|\nabla f^{\beta^i}(\mathbf{x}^{i,K_i})\| &\leq \frac{L\beta^i}{4\beta^0} + \mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,K_i}) - \bar{\mathbf{m}}^{i,K_i}\| \,|\mathbf{x}^{i,K_i}] \\ &\leq \frac{L\beta^i}{4\beta^0} + 10nL_1(F)\frac{s_1^{i,0}}{s_2^{i,0}K_i^{\alpha_1-\alpha_2}}.\end{aligned}$$

The second part of the proof follows directly by applying the triangular inequality and the result in Lemma 2.1(3) because $\mathbf{x}^{i,K_i} = \mathbf{x}^{i+1,0}$. □

Under Assumption 2, 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 4.8. *Let I be the threshold from Assumption 2. If $i \geq I$, then*

$$\mathbb{E}[f^{\beta^{i+1}}(\mathbf{x}^{i+1,0}) - f^{\beta^{i+1}}(\mathbf{x}^{i+1,*})] \leq \rho\left(\frac{L\beta^i}{4\beta^0} + 10nL_1(F)\frac{s_1^{i,0}}{s_2^{i,0}K_i^{\alpha_1-\alpha_2}} + L_1(F)(n+3)^{\frac{3}{2}}(\beta^i - \beta^{i+1})\right). \tag{4.17}$$

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

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

The result follows by using Lemma 4.7 and because $\mathbf{x}^{i+1,0}$ belongs to the ball $\mathcal{B}^\epsilon(\mathbf{x}^{i,*})$. □

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

Lemma 4.9. *Under Assumptions 1–3, 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 Eqs (4.4) and (4.9); assume that $L = \max(L_0(F), L_1(F))$, $\alpha_1 = \frac{3}{4}$ 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 Eq (4.18).

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 4.1 with the specified values of α_1 and α_2 and the fact that $L_1(f^{\beta^i}) \leq L_1(F)$ together with Lemma 4.8, 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

$$\begin{aligned} A^i &= \frac{\rho}{s_1^{i,0}}\left(\frac{\beta^{i-1}}{4\beta^0} + 10n\frac{s_1^{i-1,0}}{s_2^{i-1,0}K_{i-1}^{\frac{1}{4}}} + (n+3)^{\frac{3}{2}}(\beta^i - \beta^{i+1})\right), \\ B^i &= \frac{ns_1^{i,0}}{2}H_k^{(-\frac{3}{2})} + \ln(K_i)\left(6\sqrt{s_2^{i,0}}(n+4)\sqrt{n} + \frac{20ns_1^{i,0}}{s_2^{i,0}}\right), \end{aligned} \tag{4.18}$$

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 of the number of iterations required to bound the norm of the difference between **m** and the gradient below a certain threshold.

Lemma 4.10. *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 Eqs (4.4) and (4.9); assume that $L = \max(L_0(F), L_1(F))$, $\alpha_1 = \frac{3}{4}$ 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 Propositions 4.2 and 4.3 with $\alpha_1 = \frac{3}{4}$ and $\alpha_2 = \frac{1}{2}$. □

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

Theorem 4.3. *Let Assumptions 1–3 hold and let I be the threshold from Assumption 3.*

(i) 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 Eqs (4.4) and (4.9) are satisfied for $k = 1$. Moreover, let us denote K_i as the first iteration for which $\|\mathbf{m}^{i,K_i}\| \leq \frac{L\beta^i}{4\beta^0}$ and that without loss of generality $L = \max(L_0(F), L_1(F))$. Let $\epsilon > 0$ be the 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

$$\|\nabla f^{\beta^{i^*}}(x^{i^*,0})\| \leq \epsilon. \tag{4.19}$$

(ii) Furthermore, when for every $i \in \mathbb{N}$, f^{β^i} is convex; then, under the same setting as Theorem 4.2 given in Eq (4.13), it follows that after at most

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

function evaluations, the inequality given by Eq (4.19) 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

$$\|\mathbf{m}^{i,K_i}\| = \min_{k \in [0,K_i]} \|\mathbf{m}^{i,k}\| \leq \|\mathbf{m}^{i,R}\| \leq \|\mathbf{m}^{i,R} - \nabla f^{\beta^i}(\mathbf{x}^{i,R})\| + \|\nabla f^{\beta^i}(\mathbf{x}^{i,R})\|, \tag{4.20}$$

where $R \sim \mathcal{U}[0, K_i]$. Now, probabilistic upper bounds on the number K_i are required to obtain that both terms on 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 Eq (4.20), using the specified values of $s_1^{i,0}$, $s_2^{i,0}$ and β^i, Lemma 4.10 ensures that

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

The second term of the right-hand side in Eq (4.20) depends on the value of I. For subproblems $i \leq I$, it follows by Markov's inequality and Theorem 4.1 that

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

$$\leq O\left(\frac{\max\left(\frac{n(i+1)^{\frac{7}{2}}}{L}, n\sqrt{n}\ln(K_i)(i+1)^{\frac{3}{2}}\right)}{K_i^{\frac{1}{4}}}\right).$$

For subproblems $i > I$, Lemma 4.9 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 by Eq (4.18). Now, given the condition on K_i, it follows that

$$A^i = \rho n(i+1)^{3/2}\left(\frac{1}{i^2} + \frac{10}{s_2 i^2} + \frac{2(n+3)}{i^2(i+1)}\right)$$

and

$$B^i = \frac{H_k^{(-\frac{3}{2})}}{2(i+1)^{3/2}} + \ln(K_i)\left(\frac{6n\sqrt{n+3}\sqrt{s_2}}{\sqrt{i+1}} + \frac{12}{s_2\sqrt{i+1}}\right).$$

Thus, we obtain

$$\mathbb{P}\left(\|\nabla f^{\beta^i}(\mathbf{x}^{i,R})\| \geq \frac{L\beta^i}{4\beta^0}\right) \leq O\left(\frac{n\sqrt{n}(i+1)^{\frac{3}{2}}\ln(K_i)}{K_i^{\frac{1}{4}}}\right). \tag{4.21}$$

Therefore, to obtain that $\|\mathbf{m}^{i,K_i}\| \leq \frac{L\beta^i}{4\beta^0}$, it takes at most the following number of iterations:

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

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

$$\sum_{i=1}^{i^*} K_i = \sum_{i=1}^{I} K_i + \sum_{i=I+1}^{i^*} K_i = O\left(\max\left(n^4(I+1)^{15}, n^6(I+1)^7\right)\right) + O(n^6(i^*)^7) = O\left(\frac{n^6 L^{7/2}}{\epsilon^{7/2}}\right), \tag{4.22}$$

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 4.7

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

For the second part of the proof, the bounds on Eq (4.20) do not depend on the value of I since f^{β^i} is assumed convex for every $i \in \mathbb{N}$. With the setting in Eq (4.13), it follows that

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

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 4.2 and the second one by the definition of the probability density of R together with Propositions 4.2 and 4.3. Therefore, it takes at most $K_i = O(n^{\frac{9}{2}}(i+1)^6)$ iterations 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 needed 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 Lemma 4.7 as previously done to complete the proof. □

We would like to make a few remarks about this theorem. First, 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 4.7 establishes the rate of convergence to an ϵ-optimal point 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 Eq (1.1) into a sequence of subproblems, each of which is solved by using carefully chosen stopping criteria and step sizes. It is worth noting that, in [20], the (ϵ, Λ)-solution of the norm of the gradient is obtained after at most $O\left(\frac{nL^2\sigma^2}{\epsilon^4}\right)$ function evaluations. Although this bound has a weaker dependence on n and L, it is worse in terms of ϵ. Third, the first term in Eq (4.22) may be significant even if it is fixed, particularly if the region where the function is convex is difficult to reach; indeed, this constant disappears when f^{β^i} is convex for every index i. Nevertheless, the bounds given represent the worst set and may be considerably smaller in practice. 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 this would also 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 given by Eq (3.2) is simply projected such that $\mathbf{x} \leftarrow \max(\ell, \min(\mathbf{x}, \mathbf{u}))$.

5.1. Application to a solar thermal power plant

The first stochastic test problem is SOLAR [19], which simulates a thermal solar power plant and contains several instances allowing for selection of the number of variables, the types of constraints and the objective function to optimize. All of the instances of SOLAR are stochastic and have non-convex constraints and integer variables. In this work, the algorithms developed do not deal with integer variables. Therefore, the problem is altered: all integer variables are fixed to their initial values and the problem is to obtain a feasible solution by optimizing the expectation of constraint violations over the remaining variables. Numerical experiments were 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 c_j denotes 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 was set to 1000. Four algorithms were used

- SSO, whose hyperparameters values are given in Table 4. The search step given in Algorithm 2 was used for this experiment. A truncated version of the Gaussian gradient based estimate was used for this experiment.

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

- ZO-adaMM [15] which is a ZO version of the original Adam algorithm. This algorithm appears as one of the most effective according to [15, 31] in terms of distortion value, number of function evaluations and success rate. The default parameters defined experimentally in [15] were used for this problem, except that $\beta = 0.05$ and the learning rate was equal to 0.3. Moreover, the same gradient estimator as that for ZO-signum was used to eliminate its impact on the performance.

- CMA-ES [22] an algorithm based on biologically 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 was the one of the pymoo [8] library with the default setting.

- The NOMAD 3.9.1 software [29], based on the 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 achieved similar performance to NOMAD and CMA-ES which are state-of-the-art algorithms for this type of problem. ZO-adaMM had difficulty converging even though it is a ZO algorithm.

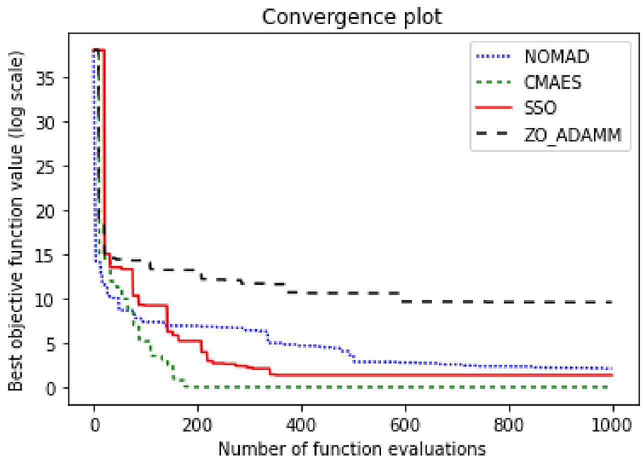

Figure 2. Average of five different seed runs for the NOMAD, CMAES, SSO and ZO-adaMM algorithms.

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) [45]. 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 [35]. 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 [15]

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

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 an untargeted attack [12], 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 for 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 were first performed by using an adapted AlexNet [28] on the dataset Cifar10 and then by using InceptionV3 [43] on the dataset ImageNet [18]. Since the NOMAD algorithm is not recommended for large problems, three algorithms are compared: SSO (without search), ZO-adaMM and CMA-ES. In the experiments, the hyperparameters of the ZO-adaMM algorithm were taken as in [15], and those of SSO are given in Table 4; the uniform gradient based estimate is used for both algorithms. Moreover, for the Cifar10 dataset, different initial learning rates for ZO-adaMM were used to observe its influence on the success rate. Experiments were conducted for 100 randomly selected images with a starting point corresponding to a null distortion; the maximum number of function queries was 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.

The best attack performance involves a trade-off between a fast convergence to a 0 attack loss in terms 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 5.

Table 5. 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

Except for ZO-adaMM with an initial learning rate equal to 0.01, all algorithms achieved a success rate above 95%. Among these algorithms, ZO-adaMM with a learning rate equal to 0.05, had the best convergence rate in terms of function evaluations but it had the worst value of distortion. On the contrary, CMA-ES obtained the best value of distortion but had the worst convergence rate. The SSO algorithm achieved balanced results, and it was the only one to reach full success rate.

Table 6 displays the results for the ImageNet dataset. Only two algorithms are compared since dimensions were too large to invert the covariance matrix in CMA-ES. For this dataset, ZO-adaMM and SSO had the same convergence rate. However, SSO outperformed ZO-adaMM in terms of success rate while having a slightly higher level of distortion.

Table 6. 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 computationally expensive stochastic blackbox optimization. The approach uses ZO gradient estimates, which provides three advantages. First, they require few function evaluations to estimate the gradient, regardless of the problem's dimensions. Second, under mild conditions on the noised objective function, the problem is formulated as optimization of a smooth approximation. Third, the smooth approximation may appear to be locally convexified near a local minima.

Based on these three features, the SSO algorithm was proposed. 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. More specifically, when the momentum's norm falls below a specified 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 has been conducted. Under Lipschitz continuity of the stochastic ZO oracle, a convergence rate in mean of the ZO-signum algorithm is derived. Under additional assumptions of smoothness and convexity or local convexity of the objective function near its minima, the rate of convergence of the SSO algorithm to an ϵ-optimal point of the problem has been derived, which is, to the best of our knowledge, the first of its kind.

Finally, numerical experiments were conducted based on a solar power plant simulation and adversarial blackbox attacks. Both examples were computationally expensive, the former was a small-sized problem ($n \approx 10$) and the latter was a large-sized problem (up to $n \approx 10^5$). The results demonstrate the SSO algorithm's competitiveness in terms of both performance and convergence rate compared to state-of-the-art algorithms. Further work will extend this approach to constrained stochastic optimization.

Use of AI tools declaration

The authors declare that they have not used artificial intelligence tools in the creation of this article.

Acknowledgements

The authors are grateful to the anonymous reviewers for their helpful comments and the English editor for improving the English quality of this text. 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.

Conflict of interest

All authors declare no conflicts of interest that may influence the publication of this paper.

References

1. C. Audet, J. Dennis, Mesh adaptive direct search algorithms for constrained optimization, *SIAM J. Optimiz.*, **17** (2006), 188–217. http://dx.doi.org/10.1137/040603371
2. C. Audet, K. Dzahini, M. Kokkolaras, S. Le Digabel, Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates, *Comput. Optim. Appl.*, **79** (2021), 1–34. http://dx.doi.org/10.1007/s10589-020-00249-0

3. C. Audet, W. Hare, *Derivative-free and blackbox optimization*, Cham: Springer, 2017. http://dx.doi.org/10.1007/978-3-319-68913-5

4. C. Audet, A. Ihaddadene, S. Le Digabel, C. Tribes, Robust optimization of noisy blackbox problems using the mesh adaptive direct search algorithm, *Optim. Lett.*, **12** (2018), 675–689. http://dx.doi.org/10.1007/s11590-017-1226-6

5. K. Balasubramanian, S. Ghadimi, Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality, and saddle points, *Found. Computat. Math.*, **22** (2022), 35–76. http://dx.doi.org/10.1007/s10208-021-09499-8

6. J. Bernstein, Y. Wang, K. Azizzadenesheli, A. Anandkumar, SignSGD: compressed optimisation for non-convex problems, *Proceedings of International Conference on Machine Learning*, 2018, 560–569.

7. S. Bhatnagar, H. Prasad, L. Prashanth, *Stochastic recursive algorithms for optimization*, London: Springer, 2013. http://dx.doi.org/10.1007/978-1-4471-4285-0

8. J. Blank, K. Deb, Pymoo: multi-objective optimization in Python, *IEEE Access*, **8** (2020), 89497–89509. http://dx.doi.org/10.1109/ACCESS.2020.2990567

9. H. Cai, Y. Lou, D. McKenzie, W. Yin, A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization, *Proceedings of the 38th International Conference on Machine Learning*, 2021, 1193–1203.

10. H. Cai, D. McKenzie, W. Yin, Z. Zhang, A one-bit, comparison-based gradient estimator, *Appl. Comput. Harmon. Anal.*, **60** (2022), 242–266. http://dx.doi.org/10.1016/j.acha.2022.03.003

11. H. Cai, D. Mckenzie, W. Yin, Z. Zhang, Zeroth-order regularized optimization (zoro): approximately sparse gradients and adaptive sampling, *SIAM J. Optim.*, **32** (2022), 687–714. http://dx.doi.org/10.1137/21M1392966

12. N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, *Proceedings of 2017 IEEE Symposium on Security and Privacy*, 2017, 39–57. http://dx.doi.org/10.1109/SP.2017.49

13. K. Chang, Stochastic nelder-mead simplex method-a new globally convergent direct search method for simulation optimization, *Eur. J. Oper. Res.*, **220** (2012), 684–694. http://dx.doi.org/10.1016/j.ejor.2012.02.028

14. R. Chen, M. Menickelly, K. Scheinberg, Stochastic optimization using a trust-region method and random models, *Math. Program.*, **169** (2018), 447–487. http://dx.doi.org/10.1007/s10107-017-1141-8

15. X. Chen, S. Liu, K. Xu, X. Li, X. Lin, M. Hong, et al., Zo-adamm: zeroth-order adaptive momentum method for black-box optimization, *Proceedings of 33rd Conference on Neural Information Processing Systems*, 2019, 1–12.

16. A. Conn, K. Scheinberg, L. Vicente, *Introduction to derivative-free optimization*, Philadelphia: SIAM, 2009. http://dx.doi.org/10.1137/1.9780898718768

17. F. Curtis, K. Scheinberg, R. Shi, A stochastic trust region algorithm based on careful step normalization, *Informs Journal on Optimization*, **1** (2019), 200–220. http://dx.doi.org/10.1287/ijoo.2018.0010

18. J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: a large-scale hierarchical image database, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848

19. M. Garneau, Modelling of a solar thermal power plant for benchmarking blackbox optimization solvers, Ph. D Thesis, École Polytechnique de Montréal, 2015.

20. S. Ghadimi, G. Lan, Stochastic first-and zeroth-order methods for nonconvex stochastic programming, *SIAM J. Optim.*, **23** (2013), 2341–2368. http://dx.doi.org/10.1137/120880811

21. S. Ghadimi, A. Ruszczynski, M. Wang, A single timescale stochastic approximation method for nested stochastic optimization, *SIAM J. Optim.*, **30** (2020), 960–979. http://dx.doi.org/10.1137/18M1230542

22. N. Hansen, The CMA evolution strategy: a comparing review, In: *Towards a new evolutionary computation*, Berlin: Springer, 2006, 75–102. http://dx.doi.org/10.1007/3-540-32494-1_4

23. S. Karimireddy, Q. Rebjock, S. Stich, M. Jaggi, Error feedback fixes signsgd and other gradient compression schemes, *Proceedings of the 36th International Conference on Machine Learning*, 2019, 3252–3261.

24. J. Kiefer, J. Wolfowitz, Stochastic estimation of the maximum of a regression function, *Ann. Math. Statist.*, **23** (1952), 462–466. http://dx.doi.org/10.1214/aoms/1177729392

25. B. Kim, H. Cai, D. McKenzie, W. Yin, Curvature-aware derivative-free optimization, arXiv:2109.13391.

26. D. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980.

27. M. Kokkolaras, Z. Mourelatos, P. Papalambros, Impact of uncertainty quantification on design: an engine optimisation case study, *International Journal of Reliability and Safety*, **1** (2006), 225–237. http://dx.doi.org/10.1504/IJRS.2006.010786

28. A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. http://dx.doi.org/10.1145/3065386

29. S. Le Digabel, Algorithm 909: NOMAD: nonlinear optimization with the MADS algorithm, *ACM T. Math. Software*, **37** (2011), 1–15. http://dx.doi.org/10.1145/1916461.1916468

30. S. Liu, P. Chen, X. Chen, M. Hong, Sign-SGD via zeroth-order oracle, *Proceedings of International Conference on Learning Representations*, 2019, 1–24.

31. S. Liu, P. Chen, B. Kailkhura, G. Zhang, A. Hero, P. Varshney, A primer on zeroth-order optimization in signal processing and machine learning: principals, recent advances, and applications, *IEEE Signal Proc. Mag.*, **37** (2020), 43–54. http://dx.doi.org/10.1109/MSP.2020.3003837

32. S. Liu, B. Kailkhura, P. Chen, P. Ting, S. Chang, L. Amini, Zeroth-order stochastic variance reduction for nonconvex optimization, *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, 3731–3741.

33. A. Maggiar, A. Wachter, I. Dolinskaya, J. Staum, A derivative-free trust-region algorithm for the optimization of functions smoothed via gaussian convolution using adaptive multiple importance sampling, *SIAM J. Optim.*, **28** (2018), 1478–1507. http://dx.doi.org/10.1137/15M1031679

34. Y. Nesterov, V. Spokoiny, Random gradient-free minimization of convex functions, *Found. Comput. Math.*, **17** (2017), 527–566. http://dx.doi.org/10.1007/s10208-015-9296-2

35. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, A. Swami, Practical black-box attacks against machine learning, *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, 506–519. http://dx.doi.org/10.1145/3052973.3053009

36. E. Real, S. Moore, A. Selle, S. Saxena, Y. Suematsu, J. Tan, et al., Large-scale evolution of image classifiers, *Proceedings of the 34th International Conference on Machine Learning*, 2017, 2902–2911.

37. H. Robbins, S. Monro, A stochastic approximation method, *Ann. Math. Statist.*, **22** (1951), 400–407. http://dx.doi.org/10.1214/aoms/1177729586

38. R. Rockafellar, J. Royset, Risk measures in engineering design under uncertainty, *Proceedings of International Conference on Applications of Statistics and Probability*, 2015, 1–8. http://dx.doi.org/10.14288/1.0076159

39. R. Rubinstein, *Simulation and the Monte Carlo method*, Hoboken: John Wiley & Sons Inc., 1981. http://dx.doi.org/10.1002/9780470316511

40. A. Ruszczynski, W. Syski, Stochastic approximation method with gradient averaging for unconstrained problems, *IEEE T. Automat. Contr.*, **28** (1983), 1097–1105. http://dx.doi.org/10.1109/TAC.1983.1103184

41. J. Spall, Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, *IEEE T. Automat. Contr.*, **37** (1992), 332–341. http://dx.doi.org/10.1109/9.119632

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

43. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 2818–2826. http://dx.doi.org/10.1109/CVPR.2016.308

44. V. Volz, J. Schrum, J. Liu, S. Lucas, A. Smith, S. Risi, Evolving mario levels in the latent space of a deep convolutional generative adversarial network, *Proceedings of the Genetic and Evolutionary Computation Conference*, 2018, 221–228. http://dx.doi.org/10.1145/3205455.3205517

45. K. Xu, S. Liu, P. Zhao, P. Chen, H. Zhang, Q. Fan, et al., Structured adversarial attack: towards general implementation and better interpretability, *Proceedings of International Conference on Learning Representations*, 2019, 1–21.

Appendix

Appendix A. Notations

The following list describes symbols used within the body of the document. Throughout the paper, when a symbol is shown in bold then it is a vector; otherwise, it is a scalar.

n	The dimension of the space of the design variables
Ω	The sample space of $\boldsymbol{\xi}$, i.e., the set of all possible outcomes of $\boldsymbol{\xi}$
$\boldsymbol{\xi} : \Omega \to \mathbb{R}^m$	The vector of uncertainties
$\mathbb{E}_{\boldsymbol{\xi}}[\cdot]$	The expectation with respect to the random vector $\boldsymbol{\xi}$
$F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$	The stochastic zeroth-order oracle that takes into account the uncertainty $\boldsymbol{\xi}$
$f : \mathbb{R}^n \to \mathbb{R}$	The expectation of F with respect to $\boldsymbol{\xi}$
$\beta \in \mathbb{R}^{+*}$	A strictly positive scalar for use as a smoothing parameter
$\mathbf{u} \in \mathbb{R}^n$	A Gaussian random vector
$f^\beta = \mathbb{E}[f(\mathbf{x} + \beta\mathbf{u})]$	A smooth approximation of a function f
$L_0(f)$	The Lipschitz constant associated with a function f
$L_1(f)$	The Lipschitz constant associated with the gradient of a function f
∇f	The gradient of a function f
$\tilde{\nabla} f$	An estimator of the gradient of a function f
$\tilde{\mathbf{g}}$	An estimator of the gradient of a function f based on outputs of the stochastic zeroth-order oracle $F(\mathbf{x}, \boldsymbol{\xi})$
$j \in [1, n]$	The counter associated with the dimension
$i \in \mathbb{N}$	The outer iteration counter associated with a subproblem
$k \in \mathbb{N}$	The inner iteration counter
$\mathbf{m} \in \mathbb{R}^n$	The momentum vector
$s_2^{i,k} \in (0, 1)$	The step size associated with the momentum
$s_1^{i,k} \in (0, 1)$	The step size associated with $\mathbf{x}$
$L \in \mathbb{R}^{+*}$	An approximation of the Lispchitz constant
$q \in \mathbb{N}$	The size of the mini batch used to estimate $\tilde{\nabla}$
$M \in \mathbb{N}$	The minimum number of iterations used in the ZO-signum algorithm
$H_k^{(\alpha)}$	The generalized harmonic number of order α
C^{0+}	Class of Lipschitz continuous functions
C^{1+}	Class of differentiable functions whose gradient is Lipschitz
C^∞	Class of infinitely differentiable functions

Appendix B. Proof of Proposition 4.1

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

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

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

$$
f^{\beta^i}(\mathbf{x}^{i,k+1}) \le f^{\beta^i}(\mathbf{x}^{i,k}) + \langle \nabla f^{\beta^i}(\mathbf{x}^{i,k}), \mathbf{x}^{i,k+1} - \mathbf{x}^{i,k} \rangle + \frac{L_1(f^{\beta^i})}{2}\|\mathbf{x}^{i,k+1} - \mathbf{x}^{i,k}\|_2^2
$$

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

where $\mathbf{1}\{\cdot\}$ is the indicator function. Now, as in [6, 30], the expected improvement conditioned on $\mathbf{x}^{i,k}$ is given by

$$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,k})|\mathbf{x}^{i,k}] \leq -s_1^{i,k}\|\nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1 + \frac{nL_1(f^{\beta^i})}{2}(s_1^{i,k})^2$$
$$+ 2s_1^{i,k}\sum_{j=1}^{n}|\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\mathbb{E}[\mathbf{1}\{\operatorname{sign}(m_j^{i,k+1}) \neq \operatorname{sign}(\nabla_j f^{\beta^i}(\mathbf{x}^{i,k}))\}|\mathbf{x}^{i,k}]. \tag{B.2}$$

Again, as in [6, 30], the expectation that the sign of $m_j^{i,k+1}$ is different from the sign of $\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})$ is relaxed by considering that the set

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

Therefore, it follows that

$$\begin{aligned}\mathbb{E}[\mathbf{1}\{\operatorname{sign}(m_j^{i,k+1}) \neq \operatorname{sign}(\nabla_j f^{\beta^i}(\mathbf{x}^{i,k}))\}|\mathbf{x}^{i,k}] &\leq \mathbb{E}[\mathbf{1}\{|m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})| \geq |\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\}|\mathbf{x}^{i,k}]\\ &\leq \frac{\mathbb{E}[|m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|\ |\mathbf{x}^{i,k}]}{|\nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|},\end{aligned} \tag{B.3}$$

where the second inequality comes from the conditional Markov's inequality. Substituting Eq (B.3) into Eq (B.2) and taking the expectation over all of the randomness we obtain

$$\mathbb{E}[f^{\beta^i}(\mathbf{x}^{i,k+1}) - f^{\beta^i}(\mathbf{x}^{i,k})] \leq -\ s_1^{i,k}\mathbb{E}[\|\nabla f^{\beta^i}(\mathbf{x}^{i,k})\|_1] + \frac{nL}{2}(s_1^{i,k})^2 + 2s_1^{i,k}\sum_{j=1}^{n}\mathbb{E}[|m_j^{i,k+1} - \nabla_j f^{\beta^i}(\mathbf{x}^{i,k})|]. \tag{B.4}$$

Moreover, by adding and subtracting $\bar{\mathbf{m}}^{i,k+1}$ in terms of the sum of Eq (B.4), one gets

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

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 Eq (B.4) completes the proof. □

Appendix C. Original signSGD and signum algorithms

Below are the original versions of the signSGD and signum algorithms.

Algorithm 3 signSGD algorithm.

Input: $\mathbf{x}^0$, $s_1 \in (0, 1)$
for $k = 0, 1, \ldots$ **do**
Calculate an estimate of the stochastic gradient $\tilde{\nabla} f(\mathbf{x}^k)$ and update:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - s_1 \mathrm{sign}(\tilde{\nabla} f(\mathbf{x}^k))$$

end for
Return $\mathbf{x}^k$

Algorithm 4 Signum algorithm.

Input: $\mathbf{x}^0, \mathbf{m}^0$, $s_1 \in (0, 1)$, $s_2 \in (0, 1)$
for $k = 0, 1, \ldots$ **do**
Calculate an estimate of the stochastic gradient $\tilde{\nabla} f(\mathbf{x}^k)$ and update:

$$\mathbf{m}^{k+1} = s_2 \mathbf{m}^k + (1 - s_2)\tilde{\nabla} f(\mathbf{x}^k)$$
$$\mathbf{x}^{k+1} = \mathbf{x}^k - s_1 \mathrm{sign}(\mathbf{m}^{k+1})$$

end for
Return $\mathbf{x}^k$