

Constrained stochastic blackbox optimization using probabilistic estimates

C. Audet, K. J. Dzahini,
S. Le Digabel, M. Kokkolaras

G-2019-30

April 2019

La collection *Les Cahiers du GERAD* est constituée des travaux de recherche menés par nos membres. La plupart de ces documents de travail a été soumis à des revues avec comité de révision. Lorsqu'un document est accepté et publié, le pdf original est retiré si c'est nécessaire et un lien vers l'article publié est ajouté.

Citation suggérée : C. Audet, K. J. Dzahini, S. Le Digabel, M. Kokkolaras (Avril 2019). Constrained stochastic black-box optimization using probabilistic estimates, Rapport technique, Les Cahiers du GERAD G-2019-30, GERAD, HEC Montréal, Canada.

Avant de citer ce rapport technique, veuillez visiter notre site Web (<https://www.gerad.ca/fr/papers/G-2019-30>) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

The series *Les Cahiers du GERAD* consists of working papers carried out by our members. Most of these pre-prints have been submitted to peer-reviewed journals. When accepted and published, if necessary, the original pdf is removed and a link to the published article is added.

Suggested citation: C. Audet, K. J. Dzahini, S. Le Digabel, M. Kokkolaras (April 2019). Constrained stochastic black-box optimization using probabilistic estimates, Technical report, Les Cahiers du GERAD G-2019-30, GERAD, HEC Montréal, Canada.

Before citing this technical report, please visit our website (<https://www.gerad.ca/en/papers/G-2019-30>) to update your reference data, if it has been published in a scientific journal.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2019
– Bibliothèque et Archives Canada, 2019

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2019
– Library and Archives Canada, 2019

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

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

Constrained stochastic blackbox optimization using probabilistic estimates

Charles Audet^{a,b}

Kwassi Joseph Dzahini^{a,b}

Michael Kokkolaras^{b,c}

Sébastien Le Digabel^{a,b}

^a *Department of Mathematics and Industrial Engineering, Polytechnique Montréal (Québec) Canada, H3C 3A7*

^b *GERAD, Montréal (Québec), Canada, H3T 2A7*

^c *Department of Mechanical Engineering, McGill University, Montreal, Quebec, Canada*

charles.audet@gerad.ca
kwassi-joseph.dzahini@polymtl.ca
michael.kokkolaras@mcgill.ca
sebastien.le.digabel@gerad.ca

April 2019

Les Cahiers du GERAD

G–2019–30

Copyright © 2019 GERAD, Audet, Dzahini, Le Digabel, Kokkolaras

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs. Les auteurs conservent leur droit d'auteur et leurs droits moraux sur leurs publications et les utilisateurs s'engagent à reconnaître et respecter les exigences légales associées à ces droits. Ainsi, les utilisateurs:

- Peuvent télécharger et imprimer une copie de toute publication du portail public aux fins d'étude ou de recherche privée;
- Ne peuvent pas distribuer le matériel ou l'utiliser pour une activité à but lucratif ou pour un gain commercial;
- Peuvent distribuer gratuitement l'URL identifiant la publication.

Si vous pensez que ce document enfreint le droit d'auteur, contactez-nous en fournissant des détails. Nous supprimerons immédiatement l'accès au travail et enquêtrons sur votre demande.

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*. Copyright and moral rights for the publications are retained by the authors and the users must commit themselves to recognize and abide the legal requirements associated with these rights. Thus, users:

- May download and print one copy of any publication from the public portal for the purpose of private study or research;
- May not further distribute the material or use it for any profit-making activity or commercial gain;
- May freely distribute the URL identifying the publication.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Abstract: This work introduces Sto-MADS, a stochastic variant of the mesh adaptive direct search (MADS) algorithm designed for deterministic blackbox optimization. Sto-MADS considers the constrained optimization of an objective function f whose values can only be computed through a stochastic blackbox with some random noise of an unknown distribution. The proposed algorithm uses an algorithmic framework similar to that of MADS and utilizes random estimates of true function values obtained from their stochastic observations to ensure improvements since the exact deterministic computable version of f is not available. Such estimates are required to be accurate with a sufficiently large but fixed probability and satisfy a variance condition. The ability of the proposed algorithm to generate an asymptotically dense set of search directions is then exploited to show that it converges to a Clarke-hypertangent stationary point of f with probability one, with the help of martingale theory.

Acknowledgments: The authors are grateful to Erick Delage and Richard Labib for valuable discussions and constructive suggestions. Work supported by the NSERC CRD RDCPJ 490744-15 grant and by an InnovÉÉ grant, both in collaboration with Hydro-Québec and Rio Tinto.

1 Introduction

Blackbox optimization (BBO), a specific form of nonlinear optimization, is the study of design and analysis of algorithms that assume the objective and/or constraint functions are given by blackboxes, i.e., any processes whose inner workings are analytically unavailable and which return an output, when provided an input [11].

This work introduces a stochastic variant of the mesh adaptive direct search (MADS) algorithm [7] for deterministic BBO and analyzes it, using theories inspired from [7, 14, 19, 36]. It aims to solve the following constrained stochastic blackbox optimization problem, also referred to as simulation optimization problem in the literature (see e.g. [3] and references therein):

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^n} \left\{ f(x) = \mathbb{E}_\eta [f_\eta(x)] \right\} \quad \text{s.t.} \quad c(x) \leq 0, \quad (1)$$

in which $\mathcal{X}$ is a subset of $\mathbb{R}^n$, η is a random variable obeying some unknown distribution, $\mathbb{E}_\eta$ denotes the expectation with respect to η , $f: \mathcal{X} \mapsto \mathbb{R} \cup \{+\infty\}$ and $c: \mathcal{X} \mapsto (\mathbb{R} \cup \{+\infty\})^m$ with $c = (c_1, c_2, \dots, c_m)^\top$ are functions, and f_η denotes the blackbox, noisy computable version of the numerically unavailable objective function f . In other words, the values of $f(x)$ can only be computed numerically with some random noise. In the convergence analysis, the objective function is assumed to be locally L -Lipschitz continuous and bounded from below. Let $\mathcal{D} = \{x \in \mathcal{X}: c(x) \leq 0\}$ denote the feasible set. To treat the constraints of Problem (1), the *extreme barrier* approach [7] is used by defining the barrier functions $f_{\eta, \mathcal{D}} = f_\eta + \chi_{\mathcal{D}}$ and $f_{\mathcal{D}} = f + \chi_{\mathcal{D}}$, where $\chi_{\mathcal{D}}$ is the indicator function for $\mathcal{D}$, i.e., it equals zero on $\mathcal{D}$ and infinity elsewhere. Thus, for all points x that are not in $\mathcal{D}$, both $f_{\eta, \mathcal{D}}(x)$ and $f_{\mathcal{D}}(x)$ are set to $+\infty$.

Such problems are of utmost importance and often arise in modern statistical machine learning, where the random variable η represents a data point drawn according to some unknown distribution and $f_\eta(x)$ measures the fit of some model parameter x to the data point η [13, 23, 31].

The study of these problems and specifically, developing provable algorithms to solve (1), has been a topic of intense research and in the recent years, several methods have been developed, most of which are extensions of existing traditional deterministic derivative-free optimization (DFO) [11, 22] methods to stochastic functions [18, 19, 36]. Such methods can be classified according to Angün and Kleijnen [5] into two categories [18]: white-box methods and blackbox methods. [18] described white-box methods as being those where one has the ability to carry out an estimation of the gradient by mean of a single simulation, perturbation analysis [18] and likelihood ratio function method [27] being some examples among many others. [18] also described blackbox methods as being those who essentially process the simulation model as a blackbox, stochastic approximation method [30] and response surface methodology [3] being some examples, in addition to many metaheuristics [3] such as scatter search, tabu search, genetic and evolutionary algorithm, simulated annealing. Thorough descriptions of stochastic approximation and response surface methodology are provided in [4].

However, since in many real applications, either it can be too complex to know the simulation model [18] or the estimation of the gradient can be computationally expensive, direct search blackbox optimization methods, generally known to be robust and reliable in practice [39], appear to be the only option. It is therefore worthwhile to specify that the analysis in the present work does not assume the existence of derivatives, i.e., first-order informations and consequently, no gradient approximations will be carried out throughout this work. Examples of existing traditional deterministic direct search blackbox optimization methods that have been extended to stochastic functions include the Nelder-Mead (NM) simplex method [35] and MADS [7].

After Barton and Ivey [15] who are among the first authors to propose a variant of the NM algorithm designed to cope with noisy function evaluations, Anderson and Ferris [4] also considered an unconstrained optimization of functions with evaluations subject to a random noise. These last authors used an algorithmic framework similar to that of NM, making use of so-called *structures* instead of simplexes and proposed an algorithm involving reflection, expansion and contraction steps, which is shown to converge to a point with probability one, based on Markov chains theory [26]. Chang [18]

recently proposed a new variant of the classic NM method, the stochastic Nelder-Mead method. After replacing the shrink step of the classic NM by the *adaptive random search*, which is a new local and global search framework in order to avoid a precocious convergence of the new algorithm, this author prove the convergence of the stochastic Nelder-Mead method to the global optima with probability one.

Audet et al. [12] recently proposed Robust-MADS, a kernel smoothing-based variant of the MADS algorithm designed to approach the minimizer of an objective function when only having access to its noisy function values. At each iteration of Robust-MADS, the incumbent solution is determined, based on values of the smoothed version of the noisy available objective constructed from a list of mesh trial points. This list is then eventually updated with the best iterate found before the next iteration of the algorithm. The proposed method is shown to have zeroth-order [9] convergence properties: iterates produced by Robust-MADS converge to a point which is “*the limit of mesh local optimizers on meshes that get infinitely fine*” [7]. Note however that even though this method produces interesting results when applied to problems including those involving granular and discrete variables [10], the corresponding work presented no computational tests to show how the proposed algorithm behaves on problems involving random noise, i.e, in a stochastic framework. Furthermore, Robust-MADS results in a deterministic algorithm in the sense that it uses only deterministic algorithmic objects, i.e, mesh and frame size parameters, smoothed function values, etc. to ensure improvements, in such a way that the resulting convergence of algorithm iterates should be understood from a deterministic and non-stochastic angle.

This work proposes Sto-MADS, a stochastic variant of MADS, designed to cope with the constrained optimization of stochastic blackbox functions and guarantees the convergence of the proposed method to a Clarke-hypertangent stationary point of the objective function f provided that certain conditions are satisfied. The proposed work uses an algorithmic framework similar to that of MADS in addition to assumptions including those taken from [19, 36]. More precisely, on one hand, it has been assumed that function estimates that are used to ensure improvements in the algorithm need to be sufficiently accurate with a sufficiently high but fixed probability which does not have to equal one but simply needs to be above a certain constant [19, 36]; on the other hand, in addition to the fact that such estimates are further assumed to satisfy a variance condition [36] that will be specified later, nothing has been assumed neither about their distribution nor about the way they are generated.

The main novelty of this work is that no model or gradient information is needed to find descent directions, compared to prior works, in particular [19, 36] and [41]. The work simply uses direct search techniques and then exploits the ability of the proposed algorithm to generate an asymptotically dense set of search directions to guarantee convergence. To the best of our knowledge, this research is the first to propose a stochastic variant of the MADS algorithm with full-supported convergence results, obtained with the help of martingale theory.

This work is organized as follows. Section 2 introduces the general framework of the proposed mesh-based stochastic method and discusses the requirements on random estimates to guarantee convergence in addition to how such estimates can be obtained in practice. It is followed by Section 3 which presents the main convergence results. Computational results are presented in Section 4.

2 Sto-MADS algorithm and probabilistic estimates

This section presents the general framework of Sto-MADS, introduces random quantities such as probabilistic estimates that will be useful for its convergence analysis and then shows how such estimates can be constructed.

2.1 The Sto-MADS algorithm

Similarly to the MADS algorithm [7], Sto-MADS is an iterative algorithm where each iteration is characterized by two main steps: an optional SEARCH step which consists of a global search that can use various strategies including the use of surrogate functions and heuristics, to explore the variables

space and a local POLL step which follows stricter rules and performs a local search in the space of variables. During each of these two steps, a finite number of trial points are generated on a discretization of the space of variables called the *mesh*. However, local searches are performed during the POLL step in a subset of the mesh called the *frame*. Note that in order to make use of random variables Δ_m^k and Δ_p^k that will be introduced later, the mesh and frame size parameters are respectively denoted by δ_m^k and δ_p^k in the present work in a deterministic framework, instead of δ^k and Δ^k used in [11].

The following definition is taken from [11] and defines the current mesh at each iteration of the algorithm.

Definition 1 Let $\mathbf{G} \in \mathbb{R}^{n \times n}$ be invertible and $\mathbf{Z} \in \mathbb{Z}^{n \times p}$ be such that the columns of $\mathbf{Z}$ form a positive spanning set [24] for $\mathbb{R}^n$ (i.e, $\mathbb{R}^n$ must be spanned by nonnegative linear combinations of the columns of $\mathbf{Z}$). Define $\mathbf{D} = \mathbf{GZ}$. The mesh of coarseness $\delta_m^k > 0$ generated by $\mathbf{D}$, centered at the incumbent solution $x^k \in \mathbb{R}^n$ is defined by

$$\mathcal{M}^k = \{x^k + \delta_m^k d : d = \mathbf{D}y, y \in \mathbb{N}^p\} \subset \mathbb{R}^n, \quad (2)$$

where δ_m^k is called the mesh size parameter.

At iteration k , given an incumbent feasible solution $x^k \in \mathcal{M}^k \cap \mathcal{D}$, the Sto-MADS algorithm seeks to find a trial “improved mesh point” [7] $y = x^k + \delta_m^k d$ whose objective function value is lower than the current unknown incumbent value $f_{\mathcal{D}}(x^k)$, i.e $f_{\mathcal{D}}(y) < f_{\mathcal{D}}(x^k)$. In the present work, f_0^k and f_s^k denote respectively the estimates of $f_{\mathcal{D}}(x^k)$ and $f_{\mathcal{D}}(x^k + s^k)$, where $s^k = \delta_m^k d$. Since the objective function values of f are not available, the estimates f_0^k and f_s^k are constructed, using evaluations of the available noisy blackbox f_{η} . Such estimates are then compared in a specific way that will be specified below, to determine whether a trial point $x^k + s^k$ may be an improved mesh point or not. Note however that in order to avoid unnecessary evaluations of the blackbox f_{η} at infeasible trial points when constructing f_0^k and f_s^k , estimates construction may be handled as follows. The constraints that define $\mathcal{D}$ are first tested in order to determine whether a trial point $y = x^k + s^k$ is feasible or not. Indeed, as described in [7], one should order the constraints in such a way that the easiest can be tested first, since some of them that define $\mathcal{D}$ might be expensive or difficult to test. If $y \notin \mathcal{D}$, then f_s^k is set to $+\infty$ without evaluating f_{η} and all the constraints that define $\mathcal{D}$, which means that all infeasible trial points are discarded. But if $y \in \mathcal{D}$, which implies that $f(y) = f_{\mathcal{D}}(y)$ and $f_{\eta}(y) = f_{\eta, \mathcal{D}}(y)$, then $f_{\eta}(y)$ is evaluated to construct the estimates f_s^k .

The following definition is taken from [11] and defines the current frame at each iteration of the algorithm.

Definition 2 Let $\mathbf{G} \in \mathbb{R}^{n \times n}$ be invertible and $\mathbf{Z} \in \mathbb{Z}^{n \times p}$ be such that the columns of $\mathbf{Z}$ form a positive spanning set [24] for $\mathbb{R}^n$. Define $\mathbf{D} = \mathbf{GZ}$ and let $\mathbb{D}$ be the columns of $\mathbf{D}$. Select a mesh size parameter $\delta_m^k > 0$ and let $\delta_p^k > 0$ be such that $\delta_m^k = \min\{\delta_p^k, (\delta_p^k)^2\}$. The frame of extent δ_p^k generated by $\mathbf{D}$ centered at the incumbent solution $x^k \in \mathbb{R}^n$ is defined by

$$\mathcal{F}^k = \{x \in \mathcal{M}^k, \|x - x^k\|_{\infty} \leq \delta_p^k b\}, \quad (3)$$

with $b = \max\{\|d'\|_{\infty}, d' \in \mathbb{D}\}$ and where δ_p^k is called the frame size parameter.

In both SEARCH and POLL steps, unlike the MADS algorithm where function values $f_{\mathcal{D}}(x^k)$ and $f_{\mathcal{D}}(x^k + s^k)$ are available, informations provided by the estimates f_0^k and f_s^k are used to determine whether a trial point $x^k + s^k$ may be an improved mesh point or not, i.e whether an iteration is successful or not. Thus, such estimates are needed to be accurate enough. Let state the following definition which is a modified version of that presented in [19].

Definition 3 Let $\varepsilon_f > 0$ be an arbitrary small but fixed constant. Denote by f_x an estimate of $f_{\mathcal{D}}(x)$ at a feasible point x . Then f_x is said to be an ε_f -accurate estimate of $f_{\mathcal{D}}(x)$ for a given δ_p^k , if

$$|f_x - f_{\mathcal{D}}(x)| \leq \varepsilon_f (\delta_p^k)^2. \quad (4)$$

Note that unlike [19, 41], ε_f does not play a crucial role in the convergence analysis of the present work, but allows to adjust the amplitude of the so-called uncertainty interval $\mathcal{I}_{\gamma, \varepsilon_f}(\delta_p^k)$ that will be introduced later.

Let next state and proof a result that will allow us to introduce the definition of successful, certain unsuccessful and uncertain unsuccessful iterations.

Proposition 1 *Let f_0^k and f_s^k be ε_f -accurate estimates of $f_{\mathcal{D}}(x^k)$ and $f_{\mathcal{D}}(x^k + s^k)$, respectively, and let $\gamma \in (2, +\infty)$ be a fixed constant. Then the followings hold:*

$$\begin{aligned} \text{if } f_s^k - f_0^k &\leq -\gamma\varepsilon_f(\delta_p^k)^2, \text{ then } f_{\mathcal{D}}(x^k + s^k) - f_{\mathcal{D}}(x^k) < 0 \\ \text{and if } f_s^k - f_0^k &\geq \gamma\varepsilon_f(\delta_p^k)^2, \text{ then } f_{\mathcal{D}}(x^k + s^k) - f_{\mathcal{D}}(x^k) > 0. \end{aligned}$$

Proof. The proof is immediate, using Definition 3 and noting that

$$f_{\mathcal{D}}(x^k + s^k) - f_{\mathcal{D}}(x^k) = f_{\mathcal{D}}(x^k + s^k) - f_s^k + (f_s^k - f_0^k) + f_0^k - f_{\mathcal{D}}(x^k).$$

□

This latter result motivates the following definition.

Definition 4 *Let f_0^k and f_s^k be ε_f -accurate estimates of $f_{\mathcal{D}}(x^k)$ and $f_{\mathcal{D}}(x^k + s^k)$, respectively, and let $\gamma \in (2, +\infty)$ be a fixed constant. If $f_s^k - f_0^k \leq -\gamma\varepsilon_f(\delta_p^k)^2$, then the iteration is called *successful*. Otherwise if $f_s^k - f_0^k \geq \gamma\varepsilon_f(\delta_p^k)^2$, then the unsuccessful iteration is called *certain*. Otherwise, $|f_s^k - f_0^k| < \gamma\varepsilon_f(\delta_p^k)^2$ and it is called *uncertain*.*

During the SEARCH or POLL step, the condition $f_s^k - f_0^k \leq -\gamma\varepsilon_f(\delta_p^k)^2$ is checked for some directions $s^k = \delta_m^k d$. If it holds, then the iterate $x^k + s^k$ is successful according to Proposition 1. Hence, the current iterate and the frame size parameter are updated respectively according to $x^{k+1} = x^k + s^k$ and $\delta_p^{k+1} = \tau^{-2}\delta_p^k$, with $\tau \in (0, 1) \cap \mathbb{Q}$ being a fixed constant and then a new iteration is initiated with a new mesh size parameter δ_m^{k+1} which satisfies $\delta_m^{k+1} = \min\{\delta_p^{k+1}, (\delta_p^{k+1})^2\}$.

If no improved mesh point is found during the SEARCH step, then the POLL step is invoked and if the condition $f_s^k - f_0^k \leq -\gamma\varepsilon_f(\delta_p^k)^2$ does not hold, the iterate is unsuccessful according to Proposition 1. Note that unlike the MADS Algorithm, Sto-MADS presents two types of unsuccessful iterations: *certain unsuccessful iterations* and *uncertain unsuccessful iterations*. In both *certain* and *uncertain unsuccessful iterations*, the current iterate is not updated, i.e. $x^{k+1} = x^k$ and the corresponding frame $\mathcal{F}^k$ is said to be a *minimal frame* with *minimal frame center* x^k , also called a *mesh local optimizer* [12]. However, one may notice that if the unsuccessful iteration is certain, then the frame size parameter is reduced according to $\delta_p^{k+1} = \tau^2\delta_p^k$ so that the resolution of the mesh can be increased, thus allowing the evaluation of f_{η} and hence estimates computation at trial mesh points that are closer to the current solution. Note that unlike [11], the use of τ^2 instead of τ has been motivated by the need to reduce the frame size parameter less aggressively during uncertain unsuccessful iterations as claimed next. Indeed, in the case of uncertain unsuccessful iterations, i.e., when $f_s^k - f_0^k$ belongs to the so called *uncertainty interval* $\mathcal{I}_{\gamma, \varepsilon_f}(\delta_p^k)$ defined by

$$\mathcal{I}_{\gamma, \varepsilon_f}(\delta_p^k) :=] -\gamma\varepsilon_f(\delta_p^k)^2, \gamma\varepsilon_f(\delta_p^k)^2 [, \quad (5)$$

then the frame size parameter is reduced less aggressively, specifically according to $\delta_p^{k+1} = \tau\delta_p^k$, so that the uncertainty interval is reduced and just like before, a new iteration is initiated with a new mesh size parameter δ_m^{k+1} which satisfies $\delta_m^{k+1} = \min\{\delta_p^{k+1}, (\delta_p^{k+1})^2\}$. The details of Sto-MADS are presented in Algorithm 1.

Algorithm 1 Sto-MADS.**[0] Initialization**

Choose $x^0 \in \mathcal{D}$, $\delta_p^0 > 0$, $\tau \in (0, 1) \cap \mathbb{Q}$, $\varepsilon_f \in (0, 1)$, $\varepsilon_{stop} \in [0, +\infty)$ and $\gamma > 2$.
Set the iteration counter $k \leftarrow 0$.

[1] Parameter Update

Set the mesh size parameter to $\delta_m^k \leftarrow \min\{\delta_p^k, (\delta_p^k)^2\}$.

[2] Poll

Select a positive spanning set $\mathbb{D}_p^k$ such that $x^k + \delta_m^k d \in \mathcal{F}^k$ for all $d \in \mathbb{D}_p^k$.

Estimates computation

Obtain estimates f_0^k and f_s^k of $f_{\mathcal{D}}(x^k)$ and $f_{\mathcal{D}}(x^k + s^k)$ respectively.

Success

if $f_s^k - f_0^k \leq -\gamma\varepsilon_f(\delta_p^k)^2$ for some $s^k = \delta_m^k d \in \{\delta_m^k d : d \in \mathbb{D}_p^k\}$,

set $x^{k+1} \leftarrow x^k + s^k$, and $\delta_p^{k+1} \leftarrow \tau^{-2}\delta_p^k$.

Failure

Certain failure: Otherwise if $f_s^k - f_0^k \geq \gamma\varepsilon_f(\delta_p^k)^2$ for all $s^k \in \{\delta_m^k d : d \in \mathbb{D}_p^k\}$,
set $x^{k+1} \leftarrow x^k$ and $\delta_p^{k+1} \leftarrow \tau^2\delta_p^k$.

Uncertain failure: Otherwise, set $x^{k+1} \leftarrow x^k$ and $\delta_p^{k+1} \leftarrow \tau\delta_p^k$.

[3] Termination

if $\delta_p^k \geq \varepsilon_{stop}$,

set $k \leftarrow k + 1$ and go to [1].

Otherwise stop.

2.2 Probabilistic estimates

The estimates f_0^k and f_s^k of function values are constructed at each iteration of Algorithm 1, using evaluations of the noisy blackbox f_{η} . Because of the randomness of f_{η} , such estimates can be respectively considered as realizations of random estimates F_0^k and F_s^k , obtained based on some random samples of the stochastic function $f_{\eta}(x)$. The behavior of F_0^k and F_s^k then influences each iteration of Algorithm 1 (as it is the case in [19, 36, 41]) in such a way that the k -th iterate, the polling directions, the mesh and frame size parameters are also random quantities. Thus, Algorithm 1 therefore results in a stochastic process $\{D^k, X^k, S^k, \Delta_p^k, \Delta_m^k, F_0^k, F_s^k\}$. All the random variables in this work are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Ω being the sample space, $\mathcal{F}$ being a σ -algebra, that is a collection of all the events (subset of Ω) and $\mathbb{P}$ is a probability measure, that is a function that returns an event's probability. Any single outcome from the sample space Ω will be denoted by ω . In general, random variables will be denoted by uppercase letters within the proposed algorithmic framework, while their realizations will be denoted by lowercase letters. Thus, $d^k = D^k(\omega)$, $x_k = X^k(\omega)$, $s^k = S^k(\omega)$, $\delta_p^k = \Delta_p^k(\omega)$ and $\delta_m^k = \Delta_m^k(\omega)$ denotes respectively realizations of random variables $D^k, X^k, S^k, \Delta_p^k$ and Δ_m^k . Similarly, $f_0^k = F_0^k(\omega)$ and $f_s^k = F_s^k(\omega)$, where $\{F_0^k, F_s^k\}$ denote estimates of $f(X^k)$ and $f(X^k + S^k)$ respectively. The goal of this work is to show that the resulting stochastic process converges with probability one under some assumptions on $\{F_0^k, F_s^k\}$. In particular, such estimates will be assumed to be accurate with sufficiently high but fixed probability, “conditioned on the past” [19].

The notion of conditioning on the past is formalized as follows as proposed in [19, 36]. Let $\mathcal{F}_{k-1}^F$ denote the σ -algebra generated by $F_0^0, F_s^0, F_0^1, F_s^1, \dots, F_0^{k-1}$ and F_s^{k-1} . For completeness, we set $\mathcal{F}_{-1}^F = \sigma(x^0)$. Thus, $\{\mathcal{F}_k^F\}_{k \geq -1}$ is a filtration, that is a subsequence of increasing σ -algebras of $\mathcal{F}$ and one can also notice by construction of the random variables X^k and Δ_p^k from Algorithm 1 that $\mathbb{E}(X^k | \mathcal{F}_{k-1}^F) = X^k$ and $\mathbb{E}(\Delta_p^k | \mathcal{F}_{k-1}^F) = \Delta_p^k$ for all $k \geq 0$.

Closeness or sufficient accuracy of function estimates is measured, using a known quantity: the current frame size parameter. This notion is formalized, using the following definition which is a modified version of those in [16, 19, 36] and which is similar to that in [41].

Definition 5 A sequence of random estimates $\{F_0^k, F_s^k\}$ is said to be β -probabilistically ε_f -accurate with respect to the corresponding sequence $\{X^k, S^k, \Delta_p^k\}$ if the events

$$J_k = \{F_0^k, F_s^k, \text{ are } \varepsilon_f\text{-accurate estimates of } f_{\mathcal{D}}(x^k) \text{ and } f_{\mathcal{D}}(x^k + s^k), \text{ respectively}\} \quad (6)$$

satisfy the following submartingale-like condition

$$\mathbb{P}(J_k | \mathcal{F}_{k-1}^F) = \mathbb{E}(\mathbb{1}_{J_k} | \mathcal{F}_{k-1}^F) \geq \beta, \quad (7)$$

where $\mathbb{1}_{J_k}$ denotes the indicator function of the event J_k , that is $\mathbb{1}_{J_k} = 1$ if $\omega \in J_k$ and 0 otherwise.

The following key assumption (similar to that made in [36]) on the nature of the stochastic information in Algorithm 1 will be useful for the convergence analysis presented in Section 3.

Assumption 1 Let $\varepsilon_f > 0$ be the same arbitrary small but fixed constant of Proposition 1. The following holds for the quantities in the algorithm:

- (i) The sequence of estimates $\{F_0^k, F_s^k\}$ generated by Algorithm 1 at feasible points is β -probabilistically ε_f -accurate for some sufficiently large $\beta \in (1/2, 1)$.
- (ii) There exists $\kappa_F > 0$ such that the sequence of estimates $\{F_0^k, F_s^k\}$ generated by Algorithm 1 at feasible points satisfies the following κ_F -variance condition for all $k \geq 0$,

$$\begin{aligned} \mathbb{E}\left(|F_s^k - f_{\mathcal{D}}(X^k + S^k)|^2 | \mathcal{F}_{k-1}^F\right) &\leq (\kappa_F)^2 (\Delta_p^k)^4 \\ \text{and} \quad \mathbb{E}\left(|F_0^k - f_{\mathcal{D}}(X^k)|^2 | \mathcal{F}_{k-1}^F\right) &\leq (\kappa_F)^2 (\Delta_p^k)^4. \end{aligned} \quad (8)$$

Remark 1 In regard to Assumption 1, note that unlike the MADS algorithm [7], the role of the frame size parameter Δ_p^k in the stochastic framework of this work is twofold. First, it updates the resolution of the mesh (which, as it will be seen, gets infinitely fine) as mentioned earlier, and second, it adaptively controls the variance which again, as it will be seen, will be driven to zero when Algorithm 1 progresses, thus allowing it to reach a desired accuracy. Therefore, no other “control size” parameter has not been required in the analysis of this work in order to control the variance as needed and described for the line search method proposed in [36]. At point (ii) of Assumption 1, the integrability of random quantities $|F_0^k - f_{\mathcal{D}}(X^k)|^2$ and $|F_s^k - f_{\mathcal{D}}(X^k + S^k)|^2$ and hence straightforwardly that of $|F_0^k - f_{\mathcal{D}}(X^k)|$ and $|F_s^k - f_{\mathcal{D}}(X^k + S^k)|$ is implicitly assumed [36] for all k at feasible points.

Using this key assumption on the accuracy of function estimates, a bound on β will be derived, under which convergence of Algorithm 1 holds. Such a bound, as it will be seen, is more general, different from “ $\beta > 1/2$ ” and defined in term of τ , κ_F and ε_f . Nevertheless, β is required to be greater than 1/2 in order to prove the existence of a Clarke-hypertangent stationary point [7] at Section 3. Before delving into the convergence analysis at Section 3 let us next state and prove a useful lemma, a slightly modified version of that presented in [36], showing the relationship between the variance assumption on the function values and the probability of obtaining inaccurate estimates.

Lemma 1 Let Assumption 1 hold. Suppose $\{X^k, F_0^k, F_s^k, \Delta_p^k\}$ is a random process generated by Algorithm 1 and $\{F_0^k, F_s^k\}$ are β -probabilistically ε_f -accurate estimates. Then for every $k \geq 0$,

$$\begin{aligned} \mathbb{E}(\mathbb{1}_{\bar{J}_k} |F_s^k - f_{\mathcal{D}}(X^k + S^k)| | \mathcal{F}_{k-1}^F) &\leq (1 - \beta)^{1/2} \kappa_F (\Delta_p^k)^2 \\ \text{and} \quad \mathbb{E}(\mathbb{1}_{\bar{J}_k} |F_0^k - f_{\mathcal{D}}(X^k)| | \mathcal{F}_{k-1}^F) &\leq (1 - \beta)^{1/2} \kappa_F (\Delta_p^k)^2. \end{aligned} \quad (9)$$

Proof. The result is shown for $F_0^k - f_{\mathcal{D}}(X^k)$ using Hölder inequality for conditional expectations as proposed in [36], but the proof for $F_s^k - f_{\mathcal{D}}(X^k + S^k)$ is the same.

$$\begin{aligned} \mathbb{E}\left(\frac{\mathbb{1}_{\bar{J}_k} |F_0^k - f_{\mathcal{D}}(X^k)|}{\kappa_F (\Delta_p^k)^2} \middle| \mathcal{F}_{k-1}^F\right) &\leq (\mathbb{E}(\mathbb{1}_{\bar{J}_k} | \mathcal{F}_{k-1}^F))^{1/2} \left[\mathbb{E}\left(\frac{|F_0^k - f_{\mathcal{D}}(X^k)|^2}{(\kappa_F)^2 (\Delta_p^k)^4} \middle| \mathcal{F}_{k-1}^F\right) \right]^{1/2} \\ &\leq (1 - \beta)^{1/2} \left[\mathbb{E}\left(\frac{|F_0^k - f_{\mathcal{D}}(X^k)|^2}{(\kappa_F)^2 (\Delta_p^k)^4} \middle| \mathcal{F}_{k-1}^F\right) \right]^{1/2} \end{aligned}$$

and the result follows after noticing by (8) that

$$\left[\mathbb{E} \left(\frac{|F_0^k - f_{\mathcal{D}}(X^k)|^2}{(\kappa_F)^2 (\Delta_p^k)^4} \middle| \mathcal{F}_{k-1}^F \right) \right]^{1/2} \leq 1.$$

□

2.3 Computing probabilistic estimates

This section aims to show how random estimates F_0^k and F_s^k satisfying Assumption 1 can be constructed, that is, the way one can access β -probabilistically ε_f -accurate estimates F_0^k and F_s^k which further satisfy the κ_F -variance condition in [Assumption 1 (ii)] in a *simple stochastic noise* framework. However, note that since full details about such estimates construction are already provided in [19, 36, 41], they are not provided here again.

Now recall that f_η denoted the noisy available blackbox which is the computable version of the objective f and consider the following typical noise assumption often used in stochastic optimization literature, i.e, suppose that the noise η is unbiased for all f , that is,

$$\begin{aligned} \mathbb{E}_\eta[f_\eta(x)] &= f(x), \quad \text{for all } x, \\ \text{and } \text{Var}_\eta[f_\eta(x)] &\leq V < +\infty, \quad \text{for all } x, \end{aligned}$$

where $V > 0$ is a constant. let η_1 and η_2 be two independent copies of the random variable η . Define estimates F_0^k and F_s^k respectively by $F_0^k = \frac{1}{p_k} \sum_{i=1}^{p_k} f_{\eta_{1,i}}(x^k)$ and $F_s^k = \frac{1}{p_k} \sum_{i=1}^{p_k} f_{\eta_{2,i}}(x^k + s^k)$, where $\eta_{1,1}, \eta_{1,2}, \dots, \eta_{1,p_k}$ and $\eta_{2,1}, \eta_{2,2}, \dots, \eta_{2,p_k}$ are independent random samples of η_1 and η_2 respectively. Then, the estimates F_0^k and F_s^k satisfy Assumption 1, provided that $p_k \geq \frac{V}{(\varepsilon_f)^2 (\delta_p^k)^4 (1-\sqrt{\beta})}$.

3 Convergence analysis

Using techniques different from those proposed by Audet and Dennis [7] for the convergence analysis of MADS and treating constraints with the extreme barrier approach, let us present convergence results of Sto-MADS, which are somehow roughly the stochastic variants of those proved in [7]. Let us first prove a *zeroth order* result, i.e, Sto-MADS “generates a limit point $\hat{X}$, which is the almost sure limit of mesh local minimizers on meshes that get infinitely fine” [9]. More formally, Sto-MADS generates a convergent subsequence of feasible iterates $\{X^k\}_{k \in K}$ such that $\lim_{k \in K} X^k = \hat{X}$ almost surely and $\lim_{k \in K} \Delta_m^k = 0$ almost surely. This latter result which will allow us to show that function estimates match their true corresponding function values with probability one, when the algorithm progresses. With regard to this latter convergence result about the random sequence $\{\Delta_m^k\}$, we prove a more general result, i.e, $\lim_{k \rightarrow \infty} \Delta_m^k = 0$ almost surely, which is somehow much stronger than the liminf-type result of [7]. Then, under a feasibility assumption on the initial point $X^0 := x^0$, under a compactness assumption on the set containing all iterates produced by Sto-MADS and under a local Lipschitz assumption on f , we make use of the Clarke calculus [21] of nonsmooth functions and prove that with probability one, each realization $\hat{x}$ of $\hat{X}$ is a *Clarke-hypertangent stationary point* [1, 7].

3.1 Zeroth order convergence

Let us next state a lemma similar to those derived in [19, 36], which guarantees an amount of decrease of the objective function f when true successful iterations occur.

Lemma 2 (*Good estimates $\Rightarrow$ decrease in function*). *Let $\varepsilon_f \in (0, 1)$ and $\gamma \in (2, +\infty)$ be arbitrary fixed constants, with ε_f sufficiently small and suppose $\{f_0^k, f_s^k\}$ are ε_f -accurate estimates. If the iteration is successful, then the improvement in function value is*

$$f_{\mathcal{D}}(x^{k+1}) \leq f_{\mathcal{D}}(x^k) - (\gamma - 2)\varepsilon_f (\delta_p^k)^2. \quad (10)$$

Proof. Since the iterate $x^k + s^k$ is successful and because the estimates are ε_f -accurate,

$$\begin{aligned} f_{\mathcal{D}}(x^k + s^k) - f_{\mathcal{D}}(x^k) &= f_{\mathcal{D}}(x^k + s^k) - f_s^k + (f_s^k - f_0^k) + f_0^k - f_{\mathcal{D}}(x^k) \\ &\leq \varepsilon_f(\delta_p^k)^2 - \gamma\varepsilon_f(\delta_p^k)^2 + \varepsilon_f(\delta_p^k)^2 \\ &\leq -(\gamma - 2)\varepsilon_f(\delta_p^k)^2. \end{aligned}$$

□

We are now ready to state and prove a theorem, similar to those derived in [19, 36] and which represents the corner stone of the convergence results in this work. For that purpose, the following assumptions on f are needed, most of which are made use in [1, 7, 19, 36].

Assumption 2 Assume an initial feasible point $x^0 := X^0$ is given. Let $\mathcal{S}(x^0)$ define the set in $\mathbb{R}^n$ which contains all iterates of Algorithm 1. Assume that the function f is bounded from below on $\mathcal{S}(x^0)$. More precisely, we assume that there exists $f_{\min} \in \mathbb{R}$ such that $-\infty < f_{\min} \leq f_{\mathcal{D}}(x)$, for all $x \in \mathcal{D}$. Assume that all iterates $\{x^k\}$ generated by Algorithm 1 lie in a compact set.

The following theorem states that the sequence of mesh size parameter $\{\Delta_m^k\}$ converges to zero with probability 1.

Theorem 1 Let Assumption 2 be satisfied. Let $\varepsilon_f \in (0, 1)$, $\tau \in (0, 1) \cap \mathbb{Q}$ and $\gamma \in (2, +\infty)$. Let $\nu \in (0, 1)$ be chosen such that

$$\frac{\nu}{1 - \nu} \geq \frac{2(\tau^{-4} - 1)}{\varepsilon_f(\gamma - 2)}, \quad (11)$$

and assume that Assumption 1 holds for $\beta \in (1/2, 1)$ chosen such that

$$\frac{\beta}{\sqrt{1 - \beta}} \geq \frac{4\nu\kappa_F}{(1 - \nu)(1 - \tau^2)}. \quad (12)$$

Then the sequence of mesh size parameter $\{\Delta_m^k\}$, generated by Algorithm 1 satisfies

$$\sum_{k=0}^{+\infty} \Delta_m^k < +\infty \quad \text{almost surely.} \quad (13)$$

Proof. This theorem is proved, using techniques and ideas derived from [19, 36] and making use of properties of the following random function

$$\Phi_k = \nu(f_{\mathcal{D}}(X^k) - f_{\min}) + (1 - \nu)(\Delta_p^k)^2, \quad (14)$$

a similar of which is used in [19], where $\nu \in (0, 1)$ is a fixed constant specified below. Recall that $\Delta_m^k = \min\{\Delta_p^k, (\Delta_p^k)^2\}$. The overall goal is to show that there exists a constant $\sigma > 0$ such that for all k ,

$$\mathbb{E}(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^F) \leq -\sigma(\Delta_p^k)^2 < 0. \quad (15)$$

Indeed, assume (15) holds on every iteration. Since f is bounded from below by $f_{\min}$ and Δ_p^k is positive, then Φ_k is bounded from below for all k . Hence, by summing over $k \in \mathbb{N}$ and taking expectations on both sides of (15), we can conclude that (13) holds with probability 1. Thus, to prove the theorem, we need to prove that on each iteration (15) holds.

The proof of this theorem considers two separate cases: good or accurate estimates and bad estimates, each of which will be broken into whether an iteration is successful, an unsuccessful iteration is certain or uncertain. For sake of clarity of the analysis, let us introduce the following events as suggested in [36]:

$$\begin{aligned}
S &:= \{X_k + S_k \text{ is successful}\}, \\
\bar{S} &:= \{\text{The iteration is unsuccessful}\}, \\
\bar{S}^C &:= \{\text{The unsuccessful iteration is certain}\}, \\
\bar{S}^{\bar{C}} &:= \{\text{The unsuccessful iteration is uncertain}\}.
\end{aligned}$$

Case 1 (Accurate estimates, $\mathbb{1}_{J_k} = 1$). We will show that Φ_k decreases no matter what type of iteration occurs and that the smallest decrease happens on the uncertain unsuccessful iteration. Thus, this case dominates the other two and overall, we conclude that

$$\mathbb{E}(\mathbb{1}_{J_k}(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^F) \leq -\beta(1 - \nu)(1 - \tau^2)(\Delta_p^k)^2. \quad (16)$$

- (i) *Successful iteration* ($\mathbb{1}_S = 1$). The iteration is successful and estimates are accurate so a decrease in the true objective occurs, specifically, (10) from lemma 2 applies:

$$\mathbb{1}_{J_k} \mathbb{1}_S \nu(f_{\mathcal{D}}(X^{k+1}) - f_{\mathcal{D}}(X^k)) \leq -\mathbb{1}_{J_k} \mathbb{1}_S \nu(\gamma - 2)\varepsilon_f(\Delta_p^k)^2 \quad (17)$$

As the iteration is successful, Δ_p^k increases by τ^{-2} . Consequently, we deduce that

$$\mathbb{1}_{J_k} \mathbb{1}_S (1 - \nu) [(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] = \mathbb{1}_{J_k} \mathbb{1}_S (1 - \nu)(\tau^{-4} - 1)(\Delta_p^k)^2. \quad (18)$$

We choose ν large enough so that the right hand side term of (17) dominates the right hand side of (18), specifically,

$$-\nu(\gamma - 2)\varepsilon_f(\Delta_p^k)^2 + (1 - \nu)(\tau^{-4} - 1)(\Delta_p^k)^2 \leq -\frac{1}{2}\nu(\gamma - 2)\varepsilon_f(\Delta_p^k)^2, \quad (19)$$

Then, we combine Equations (17) and (18) to conclude that

$$\mathbb{1}_{J_k} \mathbb{1}_S (\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{J_k} \mathbb{1}_S \frac{1}{2}\nu(\gamma - 2)\varepsilon_f(\Delta_p^k)^2. \quad (20)$$

Note however that condition (19) is equivalent to

$$\frac{\nu}{1 - \nu} \geq \frac{2(\tau^{-4} - 1)}{\varepsilon_f(\gamma - 2)}, \quad (21)$$

but for clarity of presentation, the condition (19) will be used throughout this proof.

- (ii) *Certain unsuccessful iterate* ($\mathbb{1}_{\bar{S}^C} = 1$). The iterate is unsuccessful, so there is a change of 0 in the function values while Δ_p^k decreases. Hence,

$$\mathbb{1}_{J_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^C} (\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{J_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^C} (1 - \nu)(1 - \tau^4)(\Delta_p^k)^2 \quad (22)$$

- (iii) *Uncertain unsuccessful iterate* ($\mathbb{1}_{\bar{S}^{\bar{C}}} = 1$). It's easy to notice that the behavior of Algorithm 1 at *uncertain unsuccessful iteration* is obtained from that at *certain unsuccessful iteration* simply by replacing τ^2 by τ . Thus, the bound in the change of Φ_k follows straightforwardly from (22) by replacing $\mathbb{1}_{\bar{S}^C}$ by $\mathbb{1}_{\bar{S}^{\bar{C}}}$ and τ^4 by τ^2 as follows

$$\mathbb{1}_{J_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^{\bar{C}}} (\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{J_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^{\bar{C}}} (1 - \nu)(1 - \tau^2)(\Delta_p^k)^2. \quad (23)$$

We choose ν large enough so that uncertain unsuccessful iterate (23), provides the worst case decrease when compared to (20) and (22). More precisely, we choose ν according to

$$-\frac{1}{2}\nu(\gamma - 2)\varepsilon_f(\Delta_p^k)^2 \leq -(1 - \nu)(1 - \tau^4)(\Delta_p^k)^2 \leq -(1 - \nu)(1 - \tau^2)(\Delta_p^k)^2, \quad (24)$$

but using inequalities $1 - \tau^2 < 1 - \tau^4 < \tau^{-4} - 1$, we can notice that (24) is satisfied whenever ν is chosen according to (19).

Thus, in the case of accurate estimates, using (20), (22), (23) and (24), we bound the change in Φ_k by

$$\begin{aligned}\mathbb{1}_{J_k}(\Phi_{k+1} - \Phi_k) &= \mathbb{1}_{J_k}(\mathbb{1}_S + \mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C} + \mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C})(\Phi_{k+1} - \Phi_k) \\ &\leq -\mathbb{1}_{J_k}(1 - \nu)(1 - \tau^2)(\Delta_p^k)^2.\end{aligned}\quad (25)$$

We take conditional expectations with respect to $\mathcal{F}_{k-1}^F$ and using Assumption 1, Equation (16) holds.

Case 2 (Bad estimates, $\mathbb{1}_{\bar{J}_k} = 1$). Because of inaccurate estimates, the algorithm can accept an iterate which leads to an increase in f and Δ_p^k , and hence in Φ_k . To control this increase in Φ_k , we bound the variance in the function estimates, as in (8). Then, we adjust the probability of outcome (Case 2) to be sufficiently small in order to ensure that in expectation, Φ_k is sufficiently reduced. More precisely, we will prove that

$$\mathbb{E}(\mathbb{1}_{\bar{J}_k}(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^F) \leq 2\nu(1 - \beta)^{1/2} \kappa_F (\Delta_p^k)^2. \quad (26)$$

On a successful iteration, we have the following bound

$$\begin{aligned}\mathbb{1}_{\bar{J}_k} \mathbb{1}_S \nu(f_{\mathcal{D}}(X^{k+1}) - f_{\mathcal{D}}(X^k)) &\leq \mathbb{1}_{\bar{J}_k} \mathbb{1}_S \nu[(F_s^k - F_0^k) + |f_{\mathcal{D}}(X^{k+1}) - F_s^k| + |F_0^k - f_{\mathcal{D}}(X^k)|] \\ &\leq \mathbb{1}_{\bar{J}_k} \mathbb{1}_S \nu[-\gamma\varepsilon_f(\Delta_p^k)^2 + |f_{\mathcal{D}}(X^{k+1}) - F_s^k| + |F_0^k - f_{\mathcal{D}}(X^k)|]\end{aligned}\quad (27)$$

where the last inequality is due to the decrease condition $F_s^k - F_0^k \leq -\gamma\varepsilon_f(\Delta_p^k)^2$ which holds at every successful iterations. As before, let us consider three separate cases.

- (i) *Successful iteration* ($\mathbb{1}_S = 1$). Since the iteration is successful, then as in Case 1, Δ_p^k increases by τ^{-2} , so,

$$\mathbb{1}_{\bar{J}_k} \mathbb{1}_S (1 - \nu) [(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] = \mathbb{1}_{\bar{J}_k} \mathbb{1}_S (1 - \nu)(\tau^{-4} - 1)(\Delta_p^k)^2. \quad (28)$$

By noticing that choosing ν according to (19) implies

$$-\nu\gamma\varepsilon_f(\Delta_p^k)^2 + (1 - \nu)(\tau^{-4} - 1)(\Delta_p^k)^2 \leq 0, \quad (29)$$

then, combining (27) and (28) leads to

$$\mathbb{1}_{\bar{J}_k} \mathbb{1}_S (\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{J}_k} \mathbb{1}_S (\nu |f_{\mathcal{D}}(X^{k+1}) - F_s^k| + \nu |F_0^k - f_{\mathcal{D}}(X^k)|) \quad (30)$$

- (ii) *Certain unsuccessful iterate* ($\mathbb{1}_{\bar{S}^C} = 1$). As Δ_p^k is decreased and since the change in function values is 0, then the bound in the change of Φ_k follows straightforwardly from that obtained in (22) by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$. Specifically,

$$\begin{aligned}\mathbb{1}_{\bar{J}_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^C} (\Phi_{k+1} - \Phi_k) &\leq -\mathbb{1}_{\bar{J}_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^C} (1 - \nu)(1 - \tau^4)(\Delta_p^k)^2 \\ &\leq -\mathbb{1}_{\bar{J}_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^C} (1 - \nu)(1 - \tau^2)(\Delta_p^k)^2\end{aligned}\quad (31)$$

- (iv) *Uncertain unsuccessful iterate* ($\mathbb{1}_{\bar{S}^C} = 1$). Here again, we get the bound in the change of Φ_k from that obtained in (23), simply by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$. Specifically,

$$\mathbb{1}_{\bar{J}_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^C} (\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{\bar{J}_k} \mathbb{1}_{\bar{S}} \mathbb{1}_{\bar{S}^C} (1 - \nu)(1 - \tau^2)(\Delta_p^k)^2. \quad (32)$$

By noticing that $\bar{S}^C \cup \bar{S}^C = \bar{S}$, then combining (31) and (32) leads to

$$\mathbb{1}_{\bar{J}_k} \mathbb{1}_{\bar{S}} (\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{\bar{J}_k} \mathbb{1}_{\bar{S}} (1 - \nu)(1 - \tau^2)(\Delta_p^k)^2. \quad (33)$$

Finally, since (30) dominates (33), then in all three cases,

$$\mathbb{1}_{\bar{J}_k} (\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{J}_k} (\nu |f_{\mathcal{D}}(X^{k+1}) - F_s^k| + \nu |F_0^k - f_{\mathcal{D}}(X^k)|). \quad (34)$$

Taking expectation of (34) and applying lemma 1 leads to

$$\mathbb{E}(\mathbb{1}_{\bar{J}_k}(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^F) \leq 2\nu(1 - \beta)^{1/2} \kappa_F (\Delta_p^k)^2. \quad (35)$$

Now, we combine expectations (16) and (35) to obtain

$$\begin{aligned}\mathbb{E}(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^F) &= \mathbb{E}((\mathbb{1}_{J_k} + \mathbb{1}_{\bar{J}_k})(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^F) \\ &\leq -\beta(1-\nu)(1-\tau^2)(\Delta_p^k)^2 + 2\nu(1-\beta)^{1/2}\kappa_F(\Delta_p^k)^2 \\ &\leq \left[-\beta(1-\nu)(1-\tau^2) + 2\nu\kappa_F(1-\beta)^{1/2}\right](\Delta_p^k)^2.\end{aligned}\quad (36)$$

Then, choosing β in $(1/2, 1)$ so that

$$\frac{\beta}{\sqrt{1-\beta}} \geq \frac{4\nu\kappa_F}{(1-\nu)(1-\tau^2)}, \quad (37)$$

leads to

$$-\beta(1-\nu)(1-\tau^2) + 2\nu\kappa_F(1-\beta)^{1/2} \leq -\frac{1}{2}\beta(1-\nu)(1-\tau^2). \quad (38)$$

Hence, it follows from (36) and (38) that

$$\mathbb{E}(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^F) \leq -\sigma(\Delta_p^k)^2, \quad (39)$$

where $\sigma = \frac{1}{2}\beta(1-\nu)(1-\tau^2) > 0$, and the proof is complete after noticing that $\Delta_m^k = \min\{\Delta_p^k, (\Delta_p^k)^2\}$. $\square$

Corollary 1 *Let the same assumptions that were made in Theorem 1 hold. Then, almost surely, the mesh $\mathcal{M}^k$ gets infinitely fine.*

Proof. From Theorem 1, we know that $\sum_{k=0}^{+\infty} \Delta_m^k < +\infty$ almost surely. As a consequence, the sequence of mesh size parameters Δ_m^k converges to zero almost surely. $\square$

Remark 2 *Because the sequence $\{\Delta_p^k\}$ converges to zero with probability one according to Theorem 1, then both conditions of Assumption 1 (ii), $\mathbb{E}\left(|F_s^k - f_{\mathcal{D}}(X^k + S^k)|^2 | \mathcal{F}_{k-1}^F\right) \leq (\kappa_F)^2(\Delta_p^k)^4$ and $\mathbb{E}\left(|F_0^k - f_{\mathcal{D}}(X^k)|^2 | \mathcal{F}_{k-1}^F\right) \leq (\kappa_F)^2(\Delta_p^k)^4$, used to adaptively control the variance in function estimates, drive the variance to zero, thus leading Algorithm 1 to reach the desired accuracy where function estimates match their corresponding exact function values with probability one.*

Remark 3 *Recall the definition of mesh local optimizers in Section 2.1 as being centers of minimal frames obtained when unsuccessful iterations occur. Note that some of such minimal frames centers may be “false local optimizers”, that is, iterates generated by Algorithm 1 when bad estimates occur or those which can not be updated since $f_s^k - f_0^k$ belongs to the uncertainty interval $\mathcal{I}_{\gamma, \varepsilon_f}(\delta_p^k)$. However, even though the probability $\beta < 1$ of obtaining good estimates remains the same, it must be the case that Algorithm 1 produces infinitely often “true mesh local optimizers” since on one hand, $\mathcal{I}_{\gamma, \varepsilon_f}(\delta_p^k)$ converges to $\{0\}$ almost surely and on the other hand, function estimates reach naturally their maximum level of accuracy with probability one, as explained in Remark 2. Note that the result in this latter remark is a consequence of a result which is proved more rigorously in Corollary 2.*

Next, motivated by both previous remarks, let us introduce the notions of the so-called *refining* subsequences, points and directions [6, 7] by means of the following definition similar to that in [11].

Definition 6 *A convergent subsequence $\{x^k\}_{k \in K}$ of the Sto-MADS iterates is said to be a refining subsequence, if and only if $\{\delta_m^k\}_{k \in K}$ converges to zero. The limit $\hat{x}$ of $\{x^k\}_{k \in K}$ is called a refined point. Given a refining subsequence $\{x^k\}_{k \in K}$ and its corresponding refined point $\hat{x}$, a direction d is said to be a refining direction if and only if there exists a infinite subset $L \subseteq K$ with poll direction $d^k \in \mathbb{D}_p^k$ such that $x^k + \delta_m^k d^k \in \mathcal{D}$ and $\lim_{k \in L} \frac{d^k}{\|d^k\|} = \frac{d}{\|d\|}$.*

In order to prove the existence of a convergent subsequence $\{x^k\}_{k \in K}$ of the Sto-MADS iterates let us introduce some theorems that will be also useful in the Clarke-type convergence analysis.

The following auxiliary result [14, 19] taken from martingale literature [26] will be useful later in the analysis of this work.

Theorem 2 Let $\{G_k\}$ be a submartingale, i.e, a sequence of random variables which, for every k , satisfy

$$\mathbb{E}(G_k | \mathcal{F}_{k-1}^G) \geq G_{k-1},$$

where $\mathcal{F}_{k-1}^G = \sigma(G_0, G_1, \dots, G_{k-1})$ is the σ -algebra generated by $G_0, G_1, \dots, G_{k-1}$, and $\mathbb{E}(G_k | \mathcal{F}_{k-1}^G)$ denotes the conditional expectation of G_k , given the past history of events $\mathcal{F}_{k-1}^G$.

Assume further that $G_k - G_{k-1} \leq M < +\infty$, for every k . Then,

$$\mathbb{P} \left(\left\{ \lim_{k \rightarrow \infty} G_k < \infty \right\} \cup \left\{ \limsup_{k \rightarrow \infty} G_k = \infty \right\} \right) = 1. \quad (40)$$

Theorem 3 Let the same assumptions that were made in Theorem 1 hold. Define the random function Ψ_k with realizations ψ_k as follows

$$\psi_k = \frac{f_{\mathcal{D}}(x^k) - f_{\mathcal{D}}(x^k + \delta_m^k d^k)}{\delta_p^k}, \quad (41)$$

where d^k is any direction used by Sto-MADS, generated by Algorithm 2 and satisfying $x^k + \delta_m^k d^k \in \mathcal{D}$. Then, almost surely,

$$\liminf_{k \rightarrow +\infty} \Psi_k \leq 0. \quad (42)$$

Algorithm 2 Creating the set $\mathbb{D}_p^k$ of poll directions [11].

Given $v^k \in \mathbb{R}^n$ with $\|v^k\| = 1$ and $\delta_p^k \geq \delta_m^k > 0$

[1] **Create Householder matrix**

use v^k to create its associated Householder matrix $\mathbf{H}^k = I - 2v^k v^{k\top} \in \mathbb{R}^{n \times n}$
and let $\mathbf{H}^k = [h_1 \ h_2 \ \dots \ h_n]$

[2] **Create poll set**

Define $\mathbb{B}^k = \{b_1, b_2, \dots, b_n\}$ with $b_j = \text{round} \left(\frac{\delta_p^k}{\delta_m^k} \frac{h_j}{\|h_j\|_\infty} \right) \in \mathbb{Z}^n$

set $\mathbb{D}_p^k = \mathbb{B}^k \cup (-\mathbb{B}^k)$

Proof. Using ideas in the proof of the liminf-type first-order convergence result in [19] (more precisely, see Theorem 4.16), we prove this result by contradiction conditioned on the almost sure event $V = \{\Delta_p^k \rightarrow 0\}$. All that follows is conditioned on the almost sure event V . Let's assume that with positive probability, there exists fixed numbers $\epsilon > 0$ such that,

$$\psi_k \geq \epsilon(\gamma + 2), \quad \text{for all } k, \quad (43)$$

where $\gamma \in (2, +\infty)$ is the same constant chosen in Algorithm 1 and recall $s^k = \delta_m^k d^k$ for all k . Let $\{x^k\}$, $\{\delta_p^k\}$ and $\{s^k\}$ be realizations of $\{X^k\}$, $\{\Delta_p^k\}$ and $\{S^k\}$, respectively for which $\psi_k \geq \epsilon(\gamma + 2)$, for all k . Since $\lim_{k \rightarrow +\infty} \delta_p^k = 0$ (because we conditioned on V), there exists $k_0 \in \mathbb{N}$ such that

$$\text{for all } k \geq k_0, \quad \delta_p^k < \frac{\epsilon}{\epsilon_f}. \quad (44)$$

Let us define the random variable R_k with realizations $r_k = -\frac{1}{2} \log_\tau \left(\frac{\epsilon_f \delta_p^k}{\epsilon} \right)$. Then, $r_k < 0$ for all $k \geq k_0$. The main idea of the proof is to show that such realizations occur only with probability zero, hence obtaining a contradiction. We first show that R_k is a submartingale. Recall the events J_k in the Definition 5 for some $\epsilon_f \in (0, 1)$ and consider some iterate $k \geq k_0$ for which J_k occurs, which happens with probability at least $\beta > 1/2$. Now, noticing that (43) and (44) imply

$$f_{\mathcal{D}}(x^k + s^k) - f_{\mathcal{D}}(x^k) \leq -\epsilon(\gamma + 2)\delta_p^k \leq -\epsilon_f(\gamma + 2)(\delta_p^k)^2, \quad \text{for all } k \geq k_0, \quad (45)$$

we have, for all $k \geq k_0$,

$$\begin{aligned} f_s^k - f_0^k &= \{f_{\mathcal{D}}(x^k + s^k) - f_{\mathcal{D}}(x^k)\} + \{f_{\mathcal{D}}(x^k) - f_0^k\} + \{f_s^k - f_{\mathcal{D}}(x^k + s^k)\} \\ &\leq -\varepsilon_f(\gamma + 2)(\delta_p^k)^2 + 2\varepsilon_f(\delta_p^k)^2 = -\gamma\varepsilon_f(\delta_p^k)^2. \end{aligned} \quad (46)$$

Hence, the k -th iteration of Algorithm 1 is successful, so the frame size parameter δ_p^k is increased. Consequently, $r_{k+1} = r_k + 1$.

Let $\mathcal{F}_{k-1}^J = \sigma(J_0, J_1, \dots, J_{k-1})$. If $\mathbb{1}_{J_k} = 0$, which occurs with probability at most $1 - \beta$, we (always) have $\delta_p^{k+1} \geq \tau^2 \delta_p^k$ which implies that $r_{k+1} \geq r_k - 1$. Thus,

$$\begin{aligned} \mathbb{E}(\mathbb{1}_{J_k}(R_{k+1} - R_k) | \mathcal{F}_{k-1}^J) &= \mathbb{P}(J_k | \mathcal{F}_{k-1}^J) \geq \beta \\ \text{and } \mathbb{E}(\mathbb{1}_{\bar{J}_k}(R_{k+1} - R_k) | \mathcal{F}_{k-1}^J) &\geq -\mathbb{P}(\bar{J}_k | \mathcal{F}_{k-1}^J) \geq \beta - 1. \end{aligned}$$

Hence, $\mathbb{E}(R_{k+1} - R_k | \mathcal{F}_{k-1}^J) \geq 2\beta - 1 > 0$, implying that R_k is a submartingale.

Now, let us construct the following random walk W_k on the same probability space as R_k , which will serve as a lower bound on R_k and for which $\left\{ \limsup_{k \rightarrow +\infty} W_k = +\infty \right\}$ holds almost surely

$$W_k = \sum_{i=0}^k (2 \cdot \mathbb{1}_{J_i} - 1).$$

From the submartingale-like property enforced in Definition 5, it easily follows that W_k is a submartingale. In fact,

$$\begin{aligned} \mathbb{E}(W_k | \mathcal{F}_{k-1}^J) &= \mathbb{E}(W_{k-1} | \mathcal{F}_{k-1}^J) + \mathbb{E}(2 \cdot \mathbb{1}_{J_k} - 1 | \mathcal{F}_{k-1}^J) \\ &= W_{k-1} + 2\mathbb{E}(\mathbb{1}_{J_k} | \mathcal{F}_{k-1}^J) - 1 \\ &= W_{k-1} + 2\mathbb{P}(J_k | \mathcal{F}_{k-1}^J) - 1 \\ &\geq W_{k-1}. \end{aligned}$$

Since the submartingale W_k has ± 1 (and hence, bounded) increments, it cannot have a finite limit. Thus, it follows from Theorem 2 that the event $\left\{ \limsup_{k \rightarrow +\infty} W_k = +\infty \right\}$ occurs almost surely.

Since R_k and W_k are constructed in such a way that

$$r_k - r_{k_0} = -\frac{1}{2} \log_{\tau} \left(\frac{\delta_p^k}{\delta_p^{k_0}} \right) = k - k_0 \geq w_k - w_{k_0}, \quad (47)$$

(w_k denoting a realization of W_k), then with probability one, R_k has to be positive infinitely often. Consequently, the sequence of realizations r_k such that $r_k < 0$ for all $k \geq k_0$ occurs with probability zero. Thus, the assumption “ $\psi_k \geq \epsilon(\gamma + 2)$ holds for all k with positive probability” is false and

$$\liminf_{k \rightarrow +\infty} \Psi_k \leq 0$$

holds almost surely. □

Corollary 2 *Let the same assumptions that were made in Theorem 3 hold. Then, almost surely, there exists (at least) a subset $K'' \subset \mathbb{N}$ containing an infinite number of unsuccessful iterations, say $k \in \tilde{K}$, such that,*

$$\lim_{k \in K''} \Psi_k \leq 0. \quad (48)$$

Proof. The existence of a subset K'' satisfying (48) follows straightforwardly from Theorem 3. Now, let $K' \subset \mathbb{N}$ denote the subset of all unsuccessful iterations k . To achieve the proof, let us next

prove that there exists an infinite subset $\tilde{K} \subset K'$ such that almost surely, $\lim_{k \in \tilde{K}} \Psi_k \leq 0$. Assume that the following almost sure event $V = \{\Delta_p^k \rightarrow 0\}$ occurs. Consider some iteration $k \in K'$. Then $F_s^k - F_0^k > -\gamma \varepsilon_f (\Delta_p^k)^2$. We consider two separates cases: Good estimates ($\mathbb{1}_{J_k} = 1$) and Bad estimates ($\mathbb{1}_{\bar{J}_k} = 1$).

When good estimates occur (which may happen with probability at least β thanks to Assumption 1), then it follows from Definition 3 that $f(X^k + S^k) - f(X^k) > -(\gamma + 2)\varepsilon_f (\Delta_p^k)^2$, whence

$$\Psi_k \leq (\gamma + 2)\varepsilon_f \Delta_p^k. \quad (49)$$

Hence,

$$\begin{aligned} \mathbb{E}(\mathbb{1}_{J_k} \Psi_k | \mathcal{F}_{k-1}^F) &\leq \mathbb{E}(\mathbb{1}_{J_k} | \mathcal{F}_{k-1}^F) (\gamma + 2)\varepsilon_f \Delta_p^k \\ &= \mathbb{P}(J_k | \mathcal{F}_{k-1}^F) (\gamma + 2)\varepsilon_f \Delta_p^k \leq (\gamma + 2)\varepsilon_f \Delta_p^k. \end{aligned} \quad (50)$$

On the other hand, when bad estimates occur (which may happen with probability at most $1 - \beta$), the change in f is bounded as follows:

$$f(X^k + S^k) - f(X^k) \geq F_s^k - F_0^k - |f(X^k + S^k) - F_s^k| - |f(X^k) - F_0^k|,$$

which implies that

$$\mathbb{1}_{\bar{J}_k} (f(X^k) - f(X^k + S^k)) \leq \mathbb{1}_{\bar{J}_k} (\gamma \varepsilon_f (\Delta_p^k)^2 + |f(X^k + S^k) - F_s^k| + |f(X^k) - F_0^k|). \quad (51)$$

Thus, applying Lemma 1 to (51) leads to

$$\begin{aligned} \mathbb{E}(\mathbb{1}_{\bar{J}_k} \Psi_k | \mathcal{F}_{k-1}^F) &\leq \mathbb{E}(\mathbb{1}_{\bar{J}_k} | \mathcal{F}_{k-1}^F) \gamma \varepsilon_f \Delta_p^k + 2(1 - \beta)^{1/2} \kappa_F \Delta_p^k \\ &\leq \left[(1 - \beta) \gamma \varepsilon_f + 2(1 - \beta)^{1/2} \kappa_F \right] \Delta_p^k. \end{aligned} \quad (52)$$

Then, combining (50) and (52) leads to

$$\mathbb{E}(\Psi_k | \mathcal{F}_{k-1}^F) \leq \mu \Delta_p^k, \quad (53)$$

where $\mu = [((\gamma + 2) + (1 - \beta)\gamma)\varepsilon_f + 2(1 - \beta)^{1/2} \kappa_F] > 0$.

Now, because of (53), it must be the case that almost surely, for all $\epsilon > 0$, there exists an infinite subset $\hat{K} \subset K'$ such that $\Psi_k < \epsilon$ for all $k \in \hat{K}$. Indeed, assume by contradiction that with positive probability, there exists some $\epsilon' > 0$ for which $\Psi_k < \epsilon'$ for only a finite number of iterations $k \in K'$. Then, it must be the case that $\Psi_k \geq \epsilon'$ for infinitely many k , say $k \in \tilde{K}$, thus leading to

$$\epsilon' \leq \mathbb{E}(\Psi_k | \mathcal{F}_{k-1}^F) \leq \mu \Delta_p^k, \quad \text{for all } k \in \tilde{K}, \quad (54)$$

which implies that $\lim_{k \in \tilde{K}} \Delta_p^k \geq \epsilon' > 0$, whence with positive probability, Δ_p^k does not tend to zero, thus leading to a contradiction to the almost sure convergence result of Δ_p^k to zero.

We have thus proved that with probability one, for all $\epsilon > 0$, $\Psi_k < \epsilon$ for all $k \in \hat{K}$. Consequently, with probability one, there exists a subsequence $\{\Psi_k\}_{k \in \hat{K}}$, for some infinite subset $\tilde{K} \subset \hat{K} \subset K'$, such that $\lim_{k \in \tilde{K}} \Psi_k \leq 0$, which achieves the proof. $\square$

Next, let us show the existence of convergent refining subsequences and refining directions. But before delving into the proof, note that in the stochastic setting of this work, iterates are updated based on informations provided by (random) estimates, which are accurate with some probability at least β , unlike the deterministic framework (see e.g [6, 7]) where iterates x^k are updated based on the known corresponding function values $f(x^k)$. Thus, it is the case that iterates generated by Algorithm 1 on unsuccessful iterations might not be accurate with probability one necessarily. Hence, they might not

be “true” *mesh local optimizers* in the sense of [6]. However, since the variance in function estimates goes naturally to zero when the algorithm progresses as explained in Remarks 2 and 3, we exploit the fact that Algorithm 1 may produce infinitely often “true” mesh local optimizers asymptotically, thus providing a refining subsequence, which is formalized in the corollary below.

Corollary 3 *Let the assumptions that were made in Corollary 2 and Assumption 2 hold. Then, there exists at least one almost surely convergent refining subsequence $\{X^k\}_{k \in K}$ and hence one refining direction.*

Proof. The proof of this corollary uses ideas derived in [6] in a deterministic framework. Let us prove this result by conditioning on the almost sure event $V' = \{\lim_{k \rightarrow +\infty} \Delta_m^k\} \cap \{\lim_{k \in \tilde{K}} \Psi_k \leq 0\}$, where the subset $\tilde{K}$ is the one provided by Corollary 2 and assume that V' occurs. Then, using on one hand the fact that $\lim_{k \in \tilde{K}} \delta_m^k = 0$ and on the other hand $\lim_{k \in \tilde{K}} \psi_k \leq 0$, (since we conditioned on V'), which implies in particular that

$$\lim_{k \in \tilde{K}} [f(x^k + s^k) - f(x^k)] \geq 0,$$

it is the case that Algorithm 1 generates asymptotically a subsequence of “true” mesh local optimizers x^k on meshes that get infinitely fine. Therefore, it follows from the compactness hypothesis of Assumption 2 that there exists a subset of indices $K \subset \tilde{K}$ for which the subsequence of iterates $\{x^k\}_{k \in K}$ converges. The existence of a refining direction d is justified by the compactness of the closed ball [11]. $\square$

3.2 Clarke-type first order convergence

The following definition [1, 11, 37] is needed in the main theorem.

Definition 7 (Hypertangent cone). *A vector $d \in \mathbb{R}^n$ is said to be a hypertangent vector to the set $\mathcal{D} \subset \mathbb{R}^n$ at the point $x \in \mathcal{D}$ if there exists a scalar $\epsilon > 0$ such that*

$$y + tw \in \mathcal{D} \text{ for all } y \in \mathcal{D} \cap B_\epsilon(x), \quad w \in B_\epsilon(d), \text{ and } 0 < t < \epsilon. \quad (55)$$

The set of all hypertangent vectors to $\mathcal{D}$ at x is called the hypertangent cone to $\mathcal{D}$ at x and is denoted by $T_{\mathcal{D}}^H(x)$.

In order to establish the main result of this work, let us also make use of the following generalization of the Clarke [21] directional derivative, derived in [29] where only points in the domain $\mathcal{D}$ are evaluated.

Definition 8 (Clarke generalized directional derivative). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be L -Lipschitz near $x \in \mathbb{R}^n$. The Clarke generalized directional derivative of f at x in the direction $d \in \mathbb{R}^n$ is*

$$f^\circ(x; d) = \limsup_{\substack{y \rightarrow x, y \in \mathcal{D} \\ t \searrow 0, y+td \in \mathcal{D}}} \frac{f(y+td) - f(y)}{t}. \quad (56)$$

Let us next state a useful result taken from [11], that provides in particular a lower bound on the Clarke directional derivative.

Lemma 3 *Suppose $f : \mathcal{D} \rightarrow \mathbb{R}$ is L -Lipschitz near $\hat{x} \in \mathcal{D} \subseteq \mathbb{R}^n$, and $d \in T_{\mathcal{D}}^H(\hat{x})$ is an hypertangent direction, then the Clarke directional derivative of f at $\hat{x}$ in the direction d satisfies*

$$f^\circ(\hat{x}; d) = \limsup_{x \rightarrow \hat{x}, v \rightarrow d, t \searrow 0} \frac{f(x+tv) - f(x)}{t}. \quad (57)$$

One of the most important requirements on which the Clarke-based convergence result of this work relies is that the search directions d^k should be chosen in such a way that $\delta_p^k \|d^k\|_\infty$ goes not to zero

whenever δ_p^k does. In fact, note that such expectation is not obvious to be met since $\delta_m^k \|d^k\|_\infty$ always goes to zero [11] whenever δ_m^k does, no matter how the search direction d^k is constructed. Let us next state and prove a lemma which states that if the search directions d^k are chosen according to Algorithm 2, then $\delta_p^k \|d^k\|_\infty$ remains asymptotically lower bounded by a constant that depends only on the dimension of iterates space.

Lemma 4 *Let $d^k \in \mathbb{R}^n$ be any poll direction generated by Algorithm 1 at iteration k . Assume that $\lim_{k \rightarrow +\infty} \delta_p^k = 0$. Then,*

$$\lim_{k \rightarrow +\infty} \delta_m^k \|d^k\|_\infty = 0 \quad \text{while} \quad \lim_{k \rightarrow +\infty} \delta_p^k \|d^k\|_\infty \geq \frac{1}{2\sqrt{n}}. \quad (58)$$

In other words, $\delta_p^k \|d^k\|_\infty$ doesn't go to 0 when k tends to $+\infty$.

Proof. The first result in (58) is taken from [11] and follows straightforwardly from the definition of a frame in Definition 2, specifically from the following inequality which holds at each iteration:

$$\delta_m^k \|d^k\|_\infty \leq \delta_p^k b,$$

where $b = \max\{\|d'\|_\infty, d' \in \mathbb{D}\}$.

Now let $\mathbb{H}^k$ denote the columns of the Householder matrix $\mathbf{H}^k$ generated by Algorithm 2 at iteration k . In order to prove the second result in (58), let's recall that

$$d^k = \text{round} \left(\frac{\delta_p^k}{\delta_m^k} \frac{h}{\|h\|_\infty} \right),$$

where $h := (h_{.1}, h_{.2}, \dots, h_{.n})^\top \in \mathbb{H}^k$. Because $\lim_{k \rightarrow +\infty} \delta_p^k = 0$, there exists $k_0 \in \mathbb{N}$ such that $\delta_p^k \leq \frac{1}{2\sqrt{n}} < 1$ for all $k \geq k_0$. Since $\|h\|_2 = 1$, then using the equivalence of norms [17] $\|\cdot\|_2$ and $\|\cdot\|_\infty$ on $\mathbb{R}^n$, we have that $1 \geq \|h\|_\infty \geq \frac{1}{\sqrt{n}}$. Thus, there exists $j \in \{1, 2, \dots, n\}$ such that $|h_{.j}| \geq \frac{1}{\sqrt{n}}$. Hence, for all $k \geq k_0$,

$$\left| \text{round} \left(\frac{\delta_p^k}{\delta_m^k} \frac{h_{.j}}{\|h\|_\infty} \right) \right| = \left| \text{round} \left(\frac{1}{\delta_m^k} \frac{h_{.j}}{\|h\|_\infty} \right) \right| = \text{round} \left(\frac{1}{\delta_m^k} \frac{|h_{.j}|}{\|h\|_\infty} \right) \geq \frac{1}{\sqrt{n}\delta_p^k} - 1,$$

which implies that

$$\delta_p^k \left| \text{round} \left(\frac{\delta_p^k}{\delta_m^k} \frac{h_{.j}}{\|h\|_\infty} \right) \right| \geq \frac{1}{2\sqrt{n}}, \quad \text{for all } k \geq k_0,$$

whence $\lim_{k \rightarrow +\infty} \delta_p^k \|d^k\|_\infty \geq \frac{1}{2\sqrt{n}}$. $\square$

Next, using properties of the Ψ_k function defined in Theorem 3, let us prove the following result which is a stochastic variant of that in [7] and which states that with probability one, Sto-MADS generates a point $\hat{X}$ at which the Clarke generalized derivative of f in all of the directions in the hypertangent cone to $\mathcal{D}$ at $\hat{X}$ is nonnegative. It is however worthwhile to mention that this proof is quite different from that in [7] because in the analysis of this work, the fact that $f_{\mathcal{D}}(x^k + \delta_m^k d^k) - f_{\mathcal{D}}(x^k) \geq 0$ on every unsuccessful iterations is false since some of such iterations can be uncertain, where $f_{\mathcal{D}}(x^k + \delta_m^k d^k) - f_{\mathcal{D}}(x^k)$ belongs to the uncertainty interval $\mathcal{I}_{\gamma+2, \varepsilon_f}(\delta_p^k)$.

Theorem 4 (Convergence of Sto-MADS). *Let assumptions of Corollary 3 hold. Let $\hat{x} \in \mathcal{D}$ be a refined point and $d \in T_{\mathcal{D}}^H(\hat{x})$ be a refining direction for $\hat{x}$. Denote by $\hat{X}$ and D the random variables with realizations $\hat{x}$ and d respectively. Assume f is Lipschitz near $\hat{x} \in \mathcal{D}$. Then there exists an almost sure event V , such that for all $\omega \in V$,*

$$f^\circ(\hat{x}; d) := f^\circ(\hat{X}(\omega); D(\omega)) \geq 0, \quad (59)$$

i.e., almost surely,

$$f^\circ(\hat{X}; D) \geq 0. \quad (60)$$

Proof. Since the goal is to show that $f^\circ(\hat{X}; D/\|D\|_\infty) \geq 0$ almost surely, it can be shown by conditioning on an almost sure event. All that follows is conditioned on the almost sure event $V = \{\Delta_m^k \rightarrow 0\}$.

Let $\{\delta_p^k\}$ be the realization of $\{\Delta_p^k\}$ and $\{x^k\}$ be the sequence of (feasible) iterates generated by Sto-MADS and therefore for which $\lim_{k \rightarrow +\infty} \delta_p^k = 0$.

In order to show (60), recall the random function Ψ_k with realizations ψ_k defined in Theorem 3

$$\psi_k = \frac{f_{\mathcal{D}}(x^k) - f_{\mathcal{D}}(x^k + \delta_m^k d^k)}{\delta_p^k}, \quad (61)$$

where d^k is any direction that is used by Sto-MADS and generated by Algorithm 2. Recall the subset L in Definition 6 and specify that “if d is an hypertangent direction to $\mathcal{D}$ at x , then Definition 7 guarantees that $x^k + \delta_m^k d^k \in \mathcal{D}$ for all sufficiently large $k \in L$ ” [11]. Also recall that $\delta_m^k = (\delta_p^k)^2$ for all sufficiently large k .

Since $\lim_{k \in L} \delta_p^k = 0$ (because we conditioned on V), it follows from Lemma 4 that $\delta_p^k \|d^k\|_\infty$ doesn't go to 0 when $k \in L$. Then using $\lim_{k \in L} \psi_k \leq 0$ as a consequence of Corollary 2, we have

$$\lim_{k \in L} \left(\frac{-\psi_k}{\delta_p^k \|d^k\|_\infty} \right) = \lim_{k \in L} \frac{f_{\mathcal{D}}(x^k + \delta_m^k d^k) - f_{\mathcal{D}}(x^k)}{\delta_m^k \|d^k\|_\infty} \geq 0. \quad (62)$$

Then, applying Lemmas 3, 4 and Corollary 3 using sequences $x^k \rightarrow \hat{x}$, $d^k/\|d^k\|_\infty \rightarrow d/\|d\|_\infty$ and $\delta_m^k \|d^k\|_\infty \searrow 0$, we find that

$$\begin{aligned} f^\circ\left(\hat{x}; \frac{d}{\|d\|_\infty}\right) &\geq \limsup_{x \rightarrow \hat{x}, v \rightarrow d/\|d\|_\infty, t \searrow 0} \frac{f_{\mathcal{D}}(x + tv) - f_{\mathcal{D}}(x)}{t} \\ &\geq \limsup_{k \in L} \frac{f_{\mathcal{D}}\left(x^k + \delta_m^k \|d^k\|_\infty \frac{d^k}{\|d^k\|_\infty}\right) - f_{\mathcal{D}}(x^k)}{\delta_m^k \|d^k\|_\infty} \\ &\geq \lim_{k \in L} \frac{f_{\mathcal{D}}\left(x^k + \delta_m^k \|d^k\|_\infty \frac{d^k}{\|d^k\|_\infty}\right) - f_{\mathcal{D}}(x^k)}{\delta_m^k \|d^k\|_\infty} \geq 0, \end{aligned} \quad (63)$$

where the last inequality in (63) follows from (62). Thus, we always have $f^\circ(\hat{x}; d/\|d\|_\infty) \geq 0$ conditioned on the almost sure event V , whence $f^\circ(\hat{X}; D/\|D\|_\infty) \geq 0$ almost surely. $\square$

4 Computational study

In this section we analyze the performance of Sto-MADS on a collection of stochastic noisy functions artificially created from deterministically constrained problems in the optimization literature. We omit here to compare our proposed method to algorithms using first or second order informations since they are likely to outperform Sto-MADS significantly. Thus, Several variants of Sto-MADS have been compared to Robust-MADS [12] which is a noisy blackbox optimization algorithm available in the NOMAD [32] Software package (version 3.9.1), making use of the OrthoMADS $2n$ directions [2] of MADS without quadratic models and disabling the anisotropic mesh. Note that all proposed variants of Sto-MADS are implemented in Matlab throughout this section.

The characteristics and sources of the 22 analytical deterministically constrained problems used for the experiments are summarized in Table 1. They are used with different feasible starting points giving a total of 66 constrained problems, 33 of which have bounds constraints in addition and whose dimensions range from 2 to 20. In other words, different starting points define different problems in the present work. Note that the objectives of these problems are given by functions whose function values cannot equal $+\infty$ and some of which are in the form of a sum of squares function, that is,

$$f(x) = \sum_{i=1}^{\ell} (f_i(x))^2, \quad (64)$$

$f_i(x)$ being a smooth function for each $i \in \{1, \dots, \ell\}$.

Table 1: Description of the set of 22 analytical problems.

No	Name	Source	n	m	Bnds	No	Name	Source	n	m	Bnds
1	BARNES	[38]	2	3	Yes	12	HS30	[28]	3	1	Yes
2	CHENWANG_F2	[20]	8	6	Yes	13	HS43	[28]	4	3	No
3	CHENWANG_F3	[20]	10	8	Yes	14	MAD1	[33]	2	1	No
4	CRESCENT	[8]	10	2	No	15	MAD2	[33]	2	1	No
5	DISK	[8]	10	1	No	16	MAD6	[33]	5	7	No
6	G23	[9]	3	2	Yes	17	PENTAGON	[33]	6	15	No
7	G210	[9]	10	2	Yes	18	SNAKE	[8]	2	2	No
8	G220	[9]	20	2	Yes	19	SPRING	[38]	3	4	Yes
9	HS15	[28]	2	2	Yes	20	TAOWANG_F2	[40]	7	4	Yes
10	HS22	[28]	2	2	No	21	WONG2	[33]	10	3	No
11	HS29	[28]	3	1	No	22	ZHAOWANG_F3	[42]	13	9	Yes

The type of noise that is tested is referred to as “*additive*” noise, i.e., for objectives that are in the form (64), each component f_i is additively perturbed by some random variable η_i generated uniformly in $[-u, u]$, with $u > 0$, that is

$$f_\eta(x) = \sum_{i=1}^{\ell} (f_i(x) + \eta_i)^2, \quad (65)$$

while for those that are not in the form (64), f_η is simply given by

$$f_\eta(x) = f(x) + \eta, \quad (66)$$

where η has the same distribution as η_i , $i \in \{1, \dots, \ell\}$. It obviously follows from (66) that $\mathbb{E}_\eta[f_\eta(x)] = f(x)$, while it follows from (65) that $\mathbb{E}_\eta[f_\eta(x)] = f(x) + \sum_{i=1}^{\ell} \mathbb{E}[(\eta_i)^2]$. However, optimization results are not affected by this constant bias term since $\min_x \mathbb{E}_\eta[f_\eta(x)] = \min_x f(x)$.

Constraints are treated using the “*extreme barrier*” approach, which means that the algorithms are not directly applied to the noisy objective $f_\eta(x)$ but to the barrier function $f_{\eta, \mathcal{D}}(x)$ defined to equal $f_\eta(x)$ unless one constraint at least is violated, in which case it equals $+\infty$. A first advantage of this barrier approach is that it transforms all constrained problems into unconstrained problems that can be handled by unconstrained methods such as Robust-MADS [12], while its second and key advantage, as explained in [7], is the fact that it avoids evaluations of the objective $f_\eta(x)$ and thus expensive function calls to f_η at infeasible points. That being said, during all our experiments, infeasible mesh trial points are simply rejected and their corresponding barrier function evaluations are not counted.

In our experiments, Sto-MADS is compared to 2 variants of Robust-MADS, varied in term of the choice of the smoothing parameter β' involved in its kernel smoothing technique. Indeed, Robust-MADS is a smoothing-based algorithm designed to cope with noisy blackbox optimization problems. At each iteration of Robust-MADS, a best mesh local optimizer is determined, based on values of the smoothed version of the noisy available objective constructed from a list of mesh trial points and making use of a Gaussian kernel parameterized by a constant β' that controls the smoothing intensity [12]. This list is then eventually updated with the best iterate found before the next iteration of the algorithm. Although experiments in [12] have been conducted on deterministically noisy problems, the smoothing-based technique and hence the proposed method does not depend at all on the link between the objective function f and its noisy available version, which means that Robust-MADS is supposed to cope with stochastically noisy problems.

A presentation of data profiles and performance profiles [25, 34] in order to assess if algorithms have successfully generated solution values close to the best function f values requires a convergence test. For each of the 66 problems, let x^N denote a best point found by an algorithm after N function calls to the noisy objective f_η , x^0 be a feasible starting point and x^* be the best known solution. We say that the problem is solved within the convergence tolerance $\tau \in [0, 1]$ if

$$\frac{f(x^N) - f(x^0)}{f(x^*) - f(x^0)} \geq 1 - \tau. \quad (67)$$

The horizontal axis of a data profile and a performance profile show respectively the number of noisy function evaluations divided by $n + 1$ and a ratio of the number of function calls to the noisy blackbox while their vertical axis show the portion of problems solved within a given convergence tolerance τ . In all our experiments, we set a budget of $1000(n + 1)$ noisy function evaluations, i.e all algorithms stop as soon as the number of function calls to f_η reaches $1000(n + 1)$. All tests in both Sto-MADS and Robust-MADS use only a Poll step, i.e, where the Search step is disabled, with the OrthoMADS $2n$ directions [2] ordered by means of an opportunistic strategy [11]. For the initialization, the same common parameters to both methods were used: $\delta_m^0 = \delta_p^0 = 1$ and $\tau = 1/2$. Sto-MADS parameters γ and ε_f are chosen so that $\gamma\varepsilon_f = 0.2$. However, for the choice of the sample rate p_k , it is worthwhile to mention that Robust-MADS is not in line with the theory analyzed in this work, especially in term of sample sizes which are not involved at all in its theory. Indeed, the blackbox is evaluated by Robust-MADS at each point only once, while it needs to be evaluated at least p_k times by Sto-MADS at each point in order to construct estimates. This remark motivates the choice of p_k in such a way that it increases when necessary, very slowly from one iteration to another. Thus, two variants “Sto-MADS rate 1” and “Sto-MADS rate 2” of Sto-MADS are tested, respectively with the sample rates $p_k = 1$ for all k and $p_k = \min\{\lfloor 1 + k \rfloor^{1/3}, \lceil 10^{-2}/(\delta_p^k)^4 \rceil\}$, with $p_k = 1$ corresponding to the worst case of estimates accuracy. Note that in order to increase estimates accuracy even though p_k is set to equal 1 whenever unsuccessful iterations occur, the following update procedure has been used. First, recall that $x^{k+1} = x^k$ on unsuccessful iterations k and that $f(x^k) \approx f_0^k = \frac{1}{p_k} \sum_{i=1}^{p_k} f_{\eta_i}(x^k)$. The estimate f_0^{k+1} of $f(x^{k+1})$ is computed using p_k function calls to the blackbox according to

$$f_0^{k+1} = \frac{p_k f_0^k + \sum_{j=1}^{p_{k+1}} f_{\eta_j}(x^{k+1})}{p_k + p_{k+1}}, \quad (68)$$

instead of $p_k + p_{k+1}$ function calls. This procedure improves estimates accuracy by making use of available samples at the current point during estimates computation, thus avoiding additional blackbox evaluations and seems to be very useful for blackboxes that are expensive in term of evaluations.

Note that although these latter 2 choices of sample rates do not meet the theoretical prescription derived in Section 2.3, they seemed to work well enough compared to many various other choices of p_k that have been tested. Two variants of Robust-MADS have been tested, which correspond respectively to the smoothing parameters $\beta' = 1$ and $\beta' = 2$, this latter choice being motivated by the fact that it appeared to be an adequate compromise for deterministically noisy problems in [12].

Figure 1 presents the data profiles which compare the two variants of Sto-MADS with Robust-MADS for various noise levels and convergence tolerances. For both algorithms, we notice that as expected, the lower the noise’s variance, the higher the number of problems solved. Furthermore, even though the estimates are poor in term of accuracy, “Sto-MADS rate 1” outperforms Robust-MADS for the largest value of the variance: $4/3$ for $u = 2$. When $u = 1$, in which case the variance equals $1/3$, Robust-MADS may be able to outperform Sto-MADS for certain convergence tolerances whenever the estimates are poor. For $u = 0.5$, in which case the variance equals $1/12$, we note that Sto-MADS estimates have the necessary quality to find better solutions. It follows from all these remarks that Sto-MADS outperforms Robust-MADS in a stochastic framework. In addition, the behavior of “Sto-MADS rate 2” shows that Sto-MADS could solve the majority of problems when given the time and when the estimates are sufficiently accurate. Figure 2 presents performance profiles comparing variants of Sto-MADS and Robust-MADS for various convergence tolerances and noise levels. In addition to data profiles, performance profiles show that Sto-MADS outperforms Robust-MADS and could solve more problems when given the time.

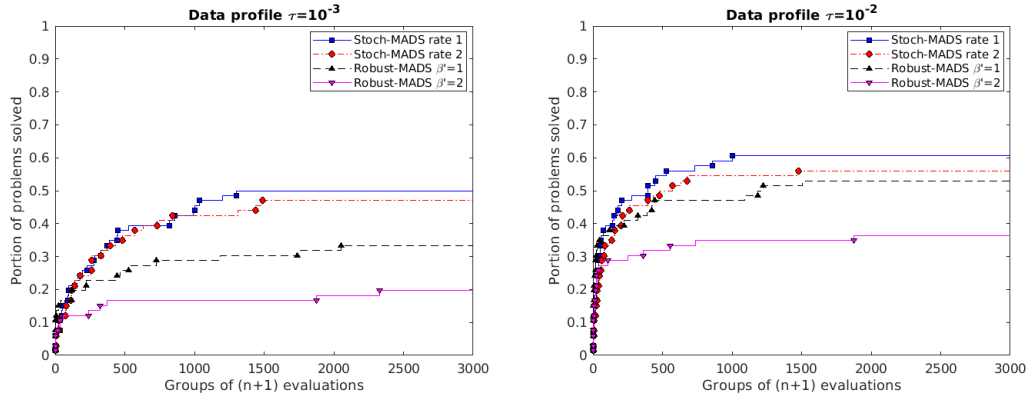
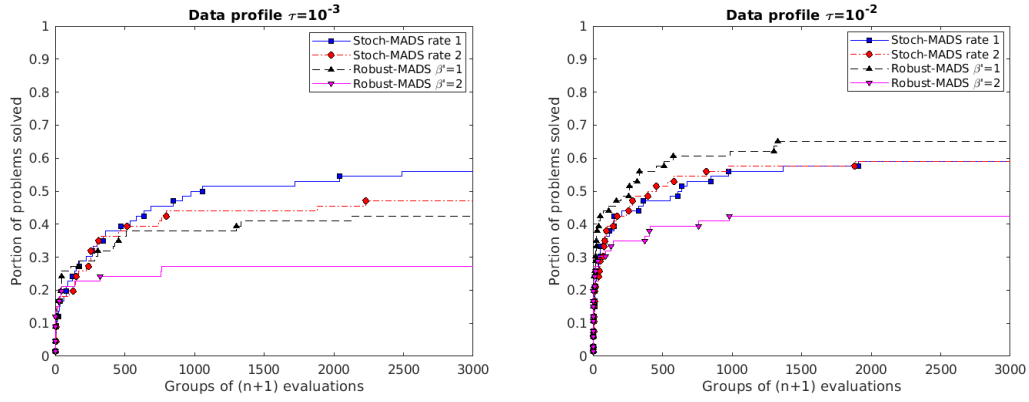
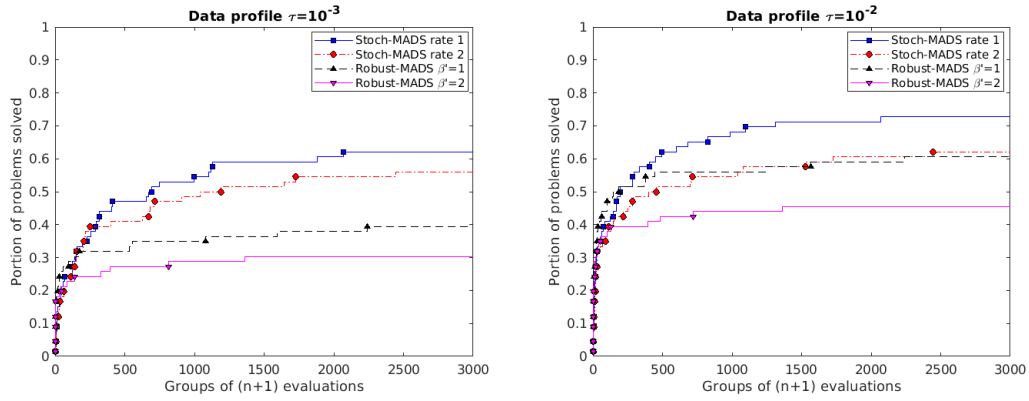
(a) $u = 2$ (b) $u = 1$ (c) $u = 0.5$

Figure 1: Data profiles obtained with Sto-MADS and Robust-MADS for convergence tolerance $\tau = 10^{-3}$ and $\tau = 10^{-2}$ on 66 analytical constrained test problems additively perturbed by a random noise uniformly generated in $[-u, u]$.

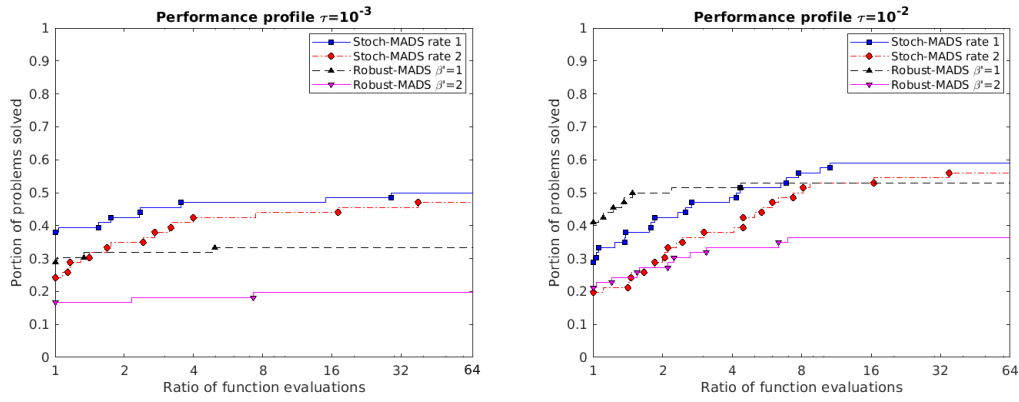
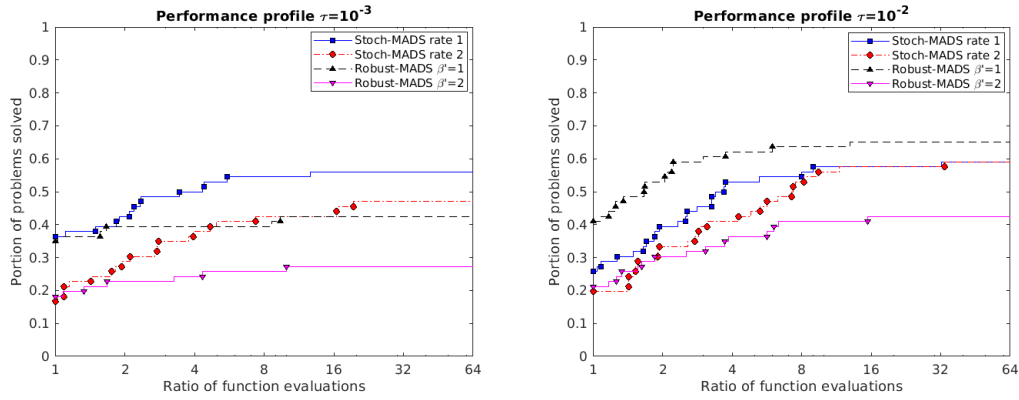
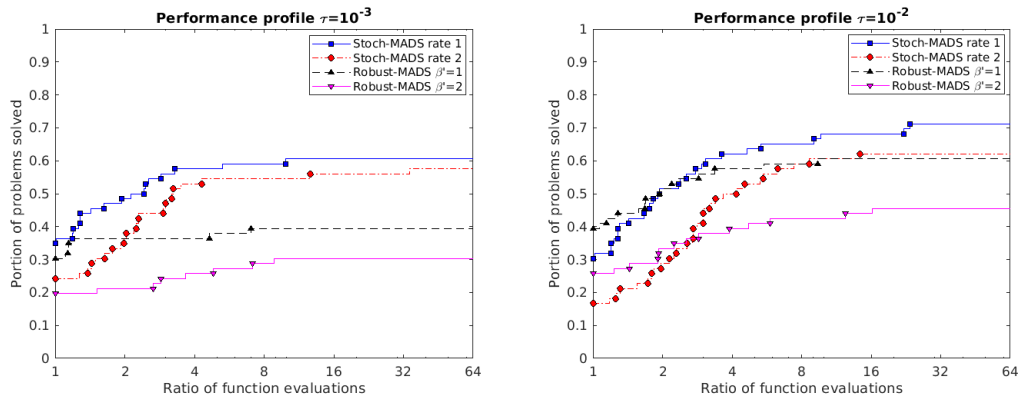
(a) $u = 2$ (b) $u = 1$ (c) $u = 0.5$

Figure 2: Performance profiles obtained with Sto-MADS and Robust-MADS for convergence tolerance $\tau = 10^{-3}$ and $\tau = 10^{-2}$ on 66 analytical constrained test problems additively perturbed by a random noise uniformly generated in $[-u, u]$.

References

- [1] M.A. Abramson and C. Audet. Convergence of mesh adaptive direct search to second-order stationary points. *SIAM Journal on Optimization*, 17(2):606–619, 2006.
- [2] M.A. Abramson, C. Audet, J.E. Dennis, Jr., and S. Le Digabel. OrthoMADS: A Deterministic MADS Instance with Orthogonal Directions. *SIAM Journal on Optimization*, 20(2):948–966, 2009.
- [3] S. Amaran, N.V. Sahinidis, B. Sharda, and S.J. Bury. Simulation optimization: a review of algorithms and applications. *4OR*, 12(4):301–333, 2014.
- [4] E.J. Anderson and M.C. Ferris. A Direct Search Algorithm for Optimization with Noisy Function Evaluations. *SIAM Journal on Optimization*, 11(3):837–857, 2001.
- [5] E. Angiın and J. Kleijnen. An asymptotic test of optimality conditions in multiresponse simulation optimization. *INFORMS Journal on Computing*, 24(1):53–65, 2012.
- [6] C. Audet and J.E. Dennis, Jr. Analysis of generalized pattern searches. *SIAM Journal on Optimization*, 13(3):889–903, 2003.
- [7] C. Audet and J.E. Dennis, Jr. Mesh Adaptive Direct Search Algorithms for Constrained Optimization. *SIAM Journal on Optimization*, 17(1):188–217, 2006.
- [8] C. Audet and J.E. Dennis, Jr. A Progressive Barrier for Derivative-Free Nonlinear Programming. *SIAM Journal on Optimization*, 20(1):445–472, 2009.
- [9] C. Audet, J.E. Dennis, Jr., and S. Le Digabel. Parallel Space Decomposition of the Mesh Adaptive Direct Search Algorithm. *SIAM Journal on Optimization*, 19(3):1150–1170, 2008.
- [10] C. Audet, S. Le Digabel, and C. Tribes. The mesh adaptive direct search algorithm for granular and discrete variables. Technical Report G-2018-16, Les cahiers du GERAD, 2019. To appear in *SIAM Journal on Optimization*.
- [11] C. Audet and W. Hare. *Derivative-Free and Blackbox Optimization*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Berlin, 2017.
- [12] C. Audet, A. Ihaddadene, S. Le Digabel, and C. Tribes. Robust optimization of noisy blackbox problems using the Mesh Adaptive Direct Search algorithm. *Optimization Letters*, 12(4):675–689, 2018.
- [13] K. Balasubramanian and S. Ghadimi. Zeroth-order Nonconvex Stochastic Optimization: Handling Constraints, High-Dimensionality and Saddle-Points. arXiv preprint arXiv:1809.06474, pages 651–676, 2019.
- [14] A.S. Bandeira, K. Scheinberg, and L.N. Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.
- [15] R.R. Barton and J.S. Ivey, Jr. Nelder-Mead simplex modifications for simulation optimization. *Management Science*, 42(7):954–973, 1996.
- [16] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method for nonconvex optimization. arXiv preprint arXiv:1609.07428, 2016.
- [17] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.
- [18] K.H. Chang. Stochastic nelder-mead simplex method - a new globally convergent direct search method for simulation optimization. *European Journal of Operational Research*, 220(3):684–694, 2012.
- [19] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, 2018.
- [20] X. Chen and N. Wang. Optimization of short-time gasoline blending scheduling problem with a DNA based hybrid genetic algorithm. *Chemical Engineering and Processing: Process Intensification*, 49(10):1076–1083, 2010.
- [21] F.H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York, 1983. Reissued in 1990 by SIAM Publications, Philadelphia, as Vol. 5 in the series Classics in Applied Mathematics.
- [22] A.R. Conn, K. Scheinberg, and L.N. Vicente. *Introduction to Derivative-Free Optimization*. MOS-SIAM Series on Optimization. SIAM, Philadelphia, 2009.
- [23] F.E. Curtis, K. Scheinberg, and R. Shi. A Stochastic Trust Region Algorithm Based on Careful Step Normalization. arXiv preprint arXiv:1712.10277, 2017.
- [24] C. Davis. Theory of positive linear dependence. *American Journal of Mathematics*, 76:733–746, 1954.
- [25] E.D. Dolan and J.J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.
- [26] R. Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

- [27] M.C. Fu. Gradient estimation. *Handbooks in operations research and management science*, 13:575–616, 2006.
- [28] W. Hock and K. Schittkowski. *Test Examples for Nonlinear Programming Codes*, volume 187 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, Germany, 1981.
- [29] J. Jahn. *Introduction to the theory of nonlinear optimization*. Springer Science & Business Media, 2007.
- [30] J. Kiefer, J. Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.
- [31] A. Kulunchakov and J. Mairal. Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise. *arXiv preprint arXiv:1901.08788*, 2019.
- [32] S. Le Digabel. Algorithm 909: NOMAD: Nonlinear Optimization with the MADS algorithm. *ACM Transactions on Mathematical Software*, 37(4):44:1–44:15, 2011.
- [33] L. Lukšan and J. Vlček. Test problems for nonsmooth unconstrained and linearly constrained optimization. Technical Report V-798, ICS AS CR, 2000.
- [34] J.J. Moré and S.M. Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009.
- [35] J.A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.
- [36] C. Paquette and K. Scheinberg. A stochastic line search method with convergence rate analysis. *arXiv preprint arXiv:1807.07994*, 2018.
- [37] R.T. Rockafellar. Generalized directional derivatives and subgradients of nonconvex functions. *Canad. J. Math.*, 32(2):257–280, 1980.
- [38] J.F. Rodríguez, J.E. Renaud, and L.T. Watson. Trust Region Augmented Lagrangian Methods for Sequential Response Surface Approximation and Optimization. *Journal of Mechanical Design*, 120(1):58–66, 1998.
- [39] W.H. Swann. Direct Search Methods. In W. Murray, editor, *Numerical Methods for Unconstrained Optimization*, pages 13–28. Academic Press, London and New York, 1972.
- [40] J. Tao and N. Wang. DNA Double Helix Based Hybrid GA for the Gasoline Blending Recipe Optimization Problem. *Chemical Engineering and Technology*, 31(3):440–451, 2008.
- [41] X. Wang and Y. Yuan. Stochastic Trust Region Methods with Trust Region Radius Depending on Probabilistic Models. *arXiv preprint arXiv:1904.03342*, 2019.
- [42] J. Zhao and N. Wang. A bio-inspired algorithm based on membrane computing and its application to gasoline blending scheduling. *Computers and Chemical Engineering*, 35(2):272–283, 2011.