POLYPUBLIE

Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates

Kwassi Joseph Dzahini, Michael Kokkolaras, Sébastien Le Digabel

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

Date	2022
Type	Article de revue / Article
Version	Version finale avant publication / Accepted version
URL de PolyPublie PolyPublie URL	https://publications.polymtl.ca/50833/
Évaluation par les pairs Peer reviewed	Révisé par les pairs / Refereed
Conditions d'utilisation Terms of use	© Tous droits réservés / All rights reserved
Référence Citation	Dzahini, K. J., Kokkolaras, M., & Le Digabel, S. (2022). Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. Mathematical Programming, 198, 675-732. https://doi.org/10.1007/s10107-022-01787-7

Publication officielle
Official publication

Titre de la revue Journal title	Mathematical Programming (vol. 198)
Maison d'édition Publisher	Springer Nature
URL officielle Official URL	https://doi.org/10.1007/s10107-022-01787-7
Mention légale Legal notice	The version of record of this article, first published in Mathematical Programming (vol. 198), is available online at Publisher's website: https://doi.org/10.1007/s10107-022-01787-7

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal
https://publications.polymtl.ca

Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates

Kwassi Joseph Dzahini* Michael Kokkolaras† Sébastien Le Digabel*

February 8, 2022

Abstract: This work introduces the StoMADS-PB algorithm for constrained stochastic blackbox optimization, which is an extension of the mesh adaptive direct-search (MADS) method originally developed for deterministic blackbox optimization under general constraints. The values of the objective and constraint functions are provided by a noisy blackbox, i.e., they can only be computed with random noise whose distribution is unknown. As in MADS, constraint violations are aggregated into a single constraint violation function. Since all function values are numerically unavailable, StoMADS-PB uses estimates and introduces probabilistic bounds for the violation. Such estimates and bounds obtained from stochastic observations are required to be accurate and reliable with high, but fixed, probabilities. The proposed method, which allows intermediate infeasible solutions, accepts new points using sufficient decrease conditions and imposing a threshold on the probabilistic bounds. Using Clarke nonsmooth calculus and martingale theory, Clarke stationarity convergence results for the objective and the violation function are derived with probability one.

*GERAD and Département de Mathématiques et de Génie Industriel, Polytechnique Montréal, C.P. 6079, Succ. Centre-ville, Montréal, Québec H3C 3A7, Canada (Kwassi-Joseph-Dzahini-LinkedIn-Profile, www.gerad.ca/Sebastien.Le.Digabel).

†GERAD and McGill University, Mechanical Engineering Department, 845 Rue Sherbrooke Ouest, Montréal, Québec H3A 0G4, Canada (www.mcgill.ca/mecheng/michael-kokkolaras).

1 Introduction

Blackbox optimization (BBO) is concerned with optimization problems in which the functions defining the objective and the constraints are given by a process called a blackbox which returns an output when provided an input but whose inner workings are analytically unavailable [12]. Mesh adaptive direct-search (MADS) [8, 9] with progressive barrier (PB) is an algorithm for deterministic BBO. The present work considers the following constrained stochastic BBO problem

$$\min_{x \in \mathcal{D}} f(x) \tag{1}$$

where $\mathcal{D} = \{x \in \mathcal{X} : c(x) \leq 0\} \subset \mathbb{R}^n$ is the feasible region, $c = (c_1, c_2, \ldots, c_m)^\top$, $\mathcal{X}$ is a subset of $\mathbb{R}^n$, $f(x) = \mathbb{E}_{\Theta_0}\left[f_{\Theta_0}(x)\right]$ with $f \colon \mathcal{X} \mapsto \mathbb{R}$, and $c_j(x) = \mathbb{E}_{\Theta_j}\left[c_{\Theta_j}(x)\right]$ with $c_j \colon \mathcal{X} \mapsto \mathbb{R}$ for all $j \in J := \{1, 2, \ldots, m\}$. $\mathbb{E}_{\Theta_j}$ denotes the expectation with respect to the random variable Θ_j for all $j \in J \cup \{0\}$, which are supposed to be independent with unknown, possibly different, distributions. $f_{\Theta_0}(\cdot)$ denotes the noisy computable version of the numerically unavailable objective function $f(\cdot)$, while for all $j \in J$, $c_{\Theta_j}(\cdot)$ denotes the noisy computable version of the numerically unavailable constraint $c_j(\cdot)$. Note that the noisy objective function f_{Θ_0} and the constraints $c_{\Theta_j}, j \in J$, are the outputs of a blackbox. By means of some useful terminology, constraints that must always be satisfied, such as those defining $\mathcal{X}$, are differentiated from those that need only to be satisfied at the solution, such as $c(x) \leq 0$. The former will be called *unrelaxable* non-quantifiable constraints and the latter, *relaxable* quantifiable constraints [41].

Solving stochastic blackbox optimization problems such as Problem (1), which often arise in signal processing and machine learning [27], has recently been a topic of intense research. Most methods for solving such problems borrow ideas from the stochastic gradient method [49]. Several works have also attempted to transfer ideas from deterministic DFO methods to the stochastic context. However, most of such proposed methods are restricted to unconstrained optimization. Indeed, after [18] which is among the first to propose a stochastic variant of the deterministic Nelder-Mead (NM) method [47], [3] also considered the optimization of functions whose evaluations are subject to random noise and proposed an algorithm which is shown to have convergence properties, based on Markov chain theory [32]. Another stochastic variant of NM was recently proposed in [22] and was proved to have global convergence properties with probability one. Using elements from [17, 40], [23] proposed STORM, a trust-region algorithm designed for stochastic optimization problems, with almost sure global convergence results. Additional research that extends the traditional deterministic trust-region method to stochastic setting have been conducted in [28, 52]. In [48], a classical backtracking Armijo line search method [5] has been adapted to the stochastic optimization setting and was shown to have first-order complexity bounds. Robust-MADS, a kernel smoothing-based variant of MADS [8], was proposed in [13] to approach the minimizer of an objective function whose values can only be computed with random noise, and was shown to possess zeroth-order [10] convergence properties. Another stochastic variant of MADS was proposed in [2] for BBO, where the noise corrupting the blackbox was supposed to be Gaussian. Convergence results of the proposed method have been derived, making use of statistical inference techniques. The StoMADS algorithm [11] is another stochastic optimization approach using estimates of function values obtained from stochastic observations and an algorithmic framework similar to that of MADS. By assuming that such estimates satisfy a variance condition and are sufficiently accurate with a large but fixed probability conditioned to the past, a Clarke [25] stationarity convergence result of StoMADS has been derived with probability one,

using martingale theory. A general framework for stochastic directional direct-search [26] methods was introduced in [33] with expected complexity analysis.

All the above stochastic optimization methods are restricted to unconstrained problems and most of them use estimated gradient information to seek an optimal solution. When the gradient does not exist or is computationally expensive to estimate, heuristics such as simulated annealing methods, genetic algorithms [39], and tabu/scatter search [38], are also used for problems with noisy constraints but do not present any convergence theory. Surrogate model-based methods for constrained stochastic BBO have also been a topic of intense research, including the response surface methodology with stochastic constraints [4] developed for expensive simulation. In [16], the capabilities of the deterministic constrained trust-region algorithm NOWPAC [15] are generalized to the optimization of blackboxes with inherently noisy evaluations of the objective and constraint functions. To mitigate the noise in function evaluations, the resulting gradient-free method SNOWPAC utilizes a Gaussian process surrogate combined with local fully linear surrogate models. Another surrogate-based approach that has gained in popularity in various research fields is Bayesian optimization [45]. Various Bayesian optimization methods for constrained stochastic BBO have been demonstrated to be efficient in practice [42, 54].

Developing direct-search methods for BBO has received renewed interest since such methods are generally known to be reliable and robust in practice [6], thereby appearing as the most promising approach in most of real applications where the gradient does not exist or is computationally expensive to estimate. However, there is relatively scarce research on developing direct-search methods for constrained stochastic BBO, especially when noise is present in the constraint functions. A pattern search and implicit filtering algorithm (PSIFA) [29, 30] was recently developed for linearly constrained problems with a noisy objective function, and was shown to have global convergence properties. A class of direct-search methods for solving smooth linearly constrained problems was also studied in [34] but even though using a probabilistic feasible descent based approach, this work assumes the objective and constraints function values to be exactly computed without noise.

The present work introduces StoMADS-PB, a stochastic variant of the mesh adaptive direct-search with progressive barrier [9], using elements from [8, 9, 11, 17, 23, 48] and is, to the best of our knowledge, the first to propose a directional direct-search [26] stochastic BBO algorithm, capable of handling general noisy constraints without requiring any feasible initial point. Its main contribution is the analysis of the resulting new framework with fully supported theoretical results. StoMADS-PB uses no (approximate) gradient information to find descent directions or to improve feasibility, compared to prior work. Rather, it uses so-called probabilistic estimates [23] of the objective and constraint function values and also introduces probabilistic bounds on constraint violation function values. The reliability of such bounds is assumed to hold with a high, but fixed, probability. Moreover, although no distributions are assumed on the estimates and no assumption is made about the way the estimates are generated, they are required to be sufficiently accurate with large, but fixed, probabilities and satisfy some variance conditions.

The manuscript is organized as follows. Section 2 presents the general framework of the proposed StoMADS-PB algorithm. Section 3 explains how the proposed method results in a stochastic process and discusses requirements on random estimates to guarantee convergence. Section 4 presents the main convergence results. Section 5 shows how random estimates and bounds can be constructed in practice. Computational results are also reported in Section 5 followed by a discussion and suggestions for future work. Additional results are provided in the appendix.

2 The StoMADS-PB algorithm

StoMADS-PB is based on an algorithmic framework similar to that of MADS with PB [9]. For the convergence analysis of Section 4, deterministic constraint violations are aggregated into a single function h called the constraint violation function, defined using the ℓ_1-norm. This in contrast to [9] where an ℓ_2-norm was employed.

$$h(x) := \begin{cases} \sum_{j=1}^{m} \max\{c_j(x), 0\} & \text{if } x \in \mathcal{X} \\ +\infty & \text{otherwise.} \end{cases}$$

According to this definition, $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $x \in \mathcal{D}$, i.e., x is feasible with respect to the relaxable constraints if and only if $h(x) = 0$. Moreover, if $0 < h(x) < +\infty$, then x is called infeasible and satisfies the unrelaxable constraints but not the relaxable ones. Notice that $h(x) = \infty$ when x does not satisfy the unrelaxable constraints.

In MADS with PB, feasibility improvement is achieved by decreasing h, specifically by comparing its function value at a current point x^k to that of a trial point $x^k + s^k$, where s^k denotes a trial step around x^k. Likewise, to decrease f, MADS with PB uses objective function values since they are available in the deterministic setting.

For the StoMADS-PB algorithm, one must guarantee some form of decrease in both f and h, using only noisy information provided by the noisy blackbox outputs f_{Θ_0} and c_{Θ_j}, $j \in J$. This section shows how this can be achieved, making use of ε-accurate estimates introduced in [23] and then presents the general framework of the proposed method.

2.1 Feasibility and objective function improvements

At iteration k, let x^k and $x^k + s^k$ be two points in $\mathcal{X}$. Since the constraint function values $c_j(x^k)$ and $c_j(x^k + s^k)$, $j \in J = \{1, 2, \ldots, m\}$, are numerically unavailable, their corresponding estimates are respectively constructed using evaluations of the noisy blackbox outputs c_{Θ_j}, $j \in J$. In general for the remainder of the manuscript, unless otherwise stated, given a function $g : \mathcal{X} \to \mathbb{R}$, an estimate of $g(x^k)$ is denoted by $g_0^k(x^k)$ (or simply by g_0^k if there is no ambiguity) while an estimate of $g(x^k+s^k)$ is denoted by $g_s^k(x^k+s^k)$ or g_s^k. In StoMADS-PB, the violations of the estimates $c_{j,0}^k(x^k)$ and $c_{j,s}^k(x^k+s^k)$ of $c_j(x^k)$ and $c_j(x^k + s^k)$, respectively, are aggregated in *estimated violations* $h_0^k(x^k)$ and $h_s^k(x^k + s^k)$ defined as

$$h_0^k(x^k) = \begin{cases} \sum_{j=1}^{m} \max\left\{c_{j,0}^k(x^k), 0\right\} & \text{if } x^k \in \mathcal{X} \\ +\infty & \text{otherwise} \end{cases} \tag{2}$$

$$\text{and} \quad h_s^k(x^k + s^k) = \begin{cases} \sum_{j=1}^{m} \max\left\{c_{j,s}^k(x^k + s^k), 0\right\} & \text{if } x^k + s^k \in \mathcal{X} \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

In order for such estimated constraint violations to be reliable enough to determine whether $h(x^k + s^k) < h(x^k)$, the estimates $c_{j,0}^k(x^k)$ and $c_{j,s}^k(x^k + s^k)$ need to be sufficiently accurate. The following definition, similar to that of [11], is adapted from [23].

Definition 2.1. *Let $\varepsilon > 0$ be a constant and δ_p^k be a nonnegative real number. For a given function $g\colon \mathcal{X} \mapsto \mathbb{R}$ and $y^k \in \mathcal{X}$, let g^k be an estimate of $g(y^k)$. Then g^k is said to be an ε-accurate estimate of $g(y^k)$ for the given δ_p^k, if*

$$\left|g^k - g(y^k)\right| \leq \varepsilon(\delta_p^k)^2.$$

As in [11], the role of δ_p^k will be played by the *poll size* parameter introduced later in Section 2.2. The following result provides bounds on $h(x^k)$ and $h(x^k + s^k)$, respectively, which will allow, later in Proposition 2.4, a decrease in the constraint violation function h by means of a sufficient decrease condition on the estimated violations h_0^k and h_s^k.

Proposition 2.2. *Let $c_{j,0}^k$ and $c_{j,s}^k$ be ε-accurate estimates of $c_j(x^k)$ and $c_j(x^k + s^k)$, respectively, for the given δ_p^k, with x^k and $x^k + s^k \in \mathcal{X}$. Then the following hold:*

$$\ell_0^k(x^k) := \sum_{j=1}^{m} \max\left\{c_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\} \leq h(x^k) \leq \sum_{j=1}^{m} \max\left\{c_{j,0}^k + \varepsilon(\delta_p^k)^2, 0\right\} =: u_0^k(x^k) \tag{4}$$

and

$$\ell_s^k(x^k+s^k) := \sum_{j=1}^{m} \max\left\{c_{j,s}^k - \varepsilon(\delta_p^k)^2, 0\right\} \leq h(x^k+s^k) \leq \sum_{j=1}^{m} \max\left\{c_{j,s}^k + \varepsilon(\delta_p^k)^2, 0\right\} =: u_s^k(x^k+s^k)$$

Proof. The result is shown for $h(x^k)$ but the proof for $h(x^k + s^k)$ is the same. Since $c_{j,0}^k$ is an ε-accurate estimate of $c_j(x^k)$ for all $j \in J$, it follows from Definition 2.1 that

$$c_{j,0}^k - \varepsilon(\delta_p^k)^2 \leq c_j(x^k) \leq c_{j,0}^k + \varepsilon(\delta_p^k)^2, \quad \text{for all } j \in J,$$

which implies that

$$\max\left\{c_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\} \leq \max\left\{c_j(x^k), 0\right\} \leq \max\left\{c_{j,0}^k + \varepsilon(\delta_p^k)^2, 0\right\}. \tag{5}$$

Finally, summing each term of (5) from $j = 1$ to m leads to (4). □

Definition 2.3. *The estimates $\ell_0^k(x^k)$ and $u_0^k(x^k)$ of* Proposition 2.2, *satisfying $\ell_0^k(x^k) \leq h(x^k) \leq u_0^k(x^k)$, are said to be ε-reliable bounds for $h(x^k)$. Similarly, the estimates $\ell_s^k(x^k+s^k)$ and $u_s^k(x^k+s^k)$ satisfying $\ell_s^k(x^k + s^k) \leq h(x^k + s^k) \leq u_s^k(x^k + s^k)$ are said to be ε-reliable bounds for $h(x^k + s^k)$.*

The following result provides sufficient conditions to identify a decrease in h and will be also used to determine an iteration type later in Section 2.2.

Proposition 2.4. *Let h_0^k and h_s^k be the estimated constraint violations at x^k and $x^k + s^k \in \mathcal{X}$, respectively, that are constructed using of ε-accurate estimates $c_{j,0}^k$ and $c_{j,s}^k$. Let $\gamma > 2$ be a constant. Then the following holds:*

$$\text{if } h_s^k - h_0^k \leq -\gamma m\varepsilon(\delta_p^k)^2, \text{ then } h(x^k + s^k) - h(x^k) \leq -(\gamma - 2)m\varepsilon(\delta_p^k)^2 < 0. \tag{6}$$

Proof. It follows from Proposition 2.2 that

$$h(x^k + s^k) - h(x^k) \leq \sum_{j=1}^{m} \max\left\{c_{j,s}^k + \varepsilon(\delta_p^k)^2, 0\right\} - \sum_{j=1}^{m} \max\left\{c_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\}. \tag{7}$$

By noticing that

$$\sum_{j=1}^{m} \max\left\{c_{j,s}^k + \varepsilon(\delta_p^k)^2, 0\right\} \leq \sum_{j=1}^{m} \max\left\{c_{j,s}^k, 0\right\} + m\varepsilon(\delta_p^k)^2 = h_s^k + m\varepsilon(\delta_p^k)^2 \tag{8}$$

and

$$\sum_{j=1}^{m} \max\left\{c_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\} \geq \sum_{j=1}^{m} \max\left\{c_{j,0}^k, 0\right\} - m\varepsilon(\delta_p^k)^2 = h_0^k - m\varepsilon(\delta_p^k)^2, \tag{9}$$

it follows from (7) that

$$h(x^k + s^k) - h(x^k) \leq h_s^k - h_0^k + 2m\varepsilon(\delta_p^k)^2 \leq -(\gamma - 2)m\varepsilon(\delta_p^k)^2,$$

where the last inequality follows from the assumption that $h_s^k - h_0^k \leq -\gamma m\varepsilon(\delta_p^k)^2$. The proof is complete by noticing that $\gamma > 2$. □

Remark 2.5. *The result of Proposition 2.4 is very important since it allows to identify a decrease in h making use of the estimated violations h_0^k and h_s^k. However, it can be observed that deriving easily Inequalities (8) and (9) in order to prove (6) is greatly favored by the use of an ℓ_1-norm in the definitions of h and both h_0^k and h_s^k. Indeed, proving a result similar to (6) when an ℓ_2-norm is used should not be as easy as it is in the latter proof. This observation motivates in fact the definition of the progressive barrier function h (and consequently the definitions of h_0^k and h_s^k) using an ℓ_1-norm unlike [9] where an ℓ_2-norm was preferred for the analysis of* MADS *with* PB.

The ε-reliable upper bound $u_0^k(x^k)$ previously obtained for $h(x^k)$ also allows one to determine the feasibility with respect to the relaxable constraints of a given trial point $x^k \in \mathcal{X}$. Indeed, it obviously follows from (4) that $h(x^k) = 0$ if $u_0^k(x^k) = 0$, which is satisfied provided that $c_{j,0}^k(x^k) \leq -\varepsilon(\delta_p^k)^2$, for all $j \in J$. This means that in order for $h(x^k) = 0$ to hold, all the estimates of constraint function values must be sufficiently negative and not simply zero. By means of the following definition, StoMADS-PB partitions the trial points into *ε-feasible* and *ε-infeasible* points making use of a non-negative barrier threshold $h_{\max}^k$ which is introduced in the present research, inspired by [9]

Definition 2.6. *Let $x^k \in \mathcal{X}$ be any trial point and let $u_0^k(x^k)$ be an ε-reliable upper bound for $h(x^k)$. Then x^k is called ε-feasible if $u_0^k(x^k) = 0$, and it is called ε-infeasible if $0 < u_0^k(x^k) \leq h_{\max}^k$. Similarly, $x^k + s^k \in \mathcal{X}$ is called ε-feasible if $u_s^k(x^k + s^k) = 0$, and it is called ε-infeasible if $0 < u_s^k(x^k + s^k) \leq h_{\max}^k$.*

StoMADS-PB does not require that the starting point be ε-feasible. The algorithm can be applied to any problem satisfying only the following assumption adapted from [9].

Assumption 1. *There exists some point $x^0 \in \mathcal{X}$ such that $f_0^0(x^0)$ and $u_0^0(x^0)$ are both finite, and $u_0^0(x^0) \leq h_{\max}^0$.*

The next result similar to that in [11] provides a sufficient condition to identify a decrease in f and also allows one to determine an iteration type in Section 2.2.

Proposition 2.7. *Let f_0^k and f_s^k be ε-accurate estimates of $f(x^k)$ and $f(x^k + s^k)$, respectively, for x^k and $x^k + s^k \in \mathcal{X}$. Let $\gamma > 2$ be a constant. Then the following holds:*

$$\text{if } f_s^k - f_0^k \leq -\gamma\varepsilon(\delta_p^k)^2, \text{ then } f(x^k + s^k) - f(x^k) \leq -(\gamma - 2)\varepsilon(\delta_p^k)^2 < 0. \tag{10}$$

Proof. The proof follows from Definition 2.1 and the equality

$$f(x^k + s^k) - f(x^k) = f(x^k + s^k) - f_s^k + \left(f_s^k - f_0^k\right) + f_0^k - f(x^k).$$

□

The incumbent solutions x_{inf}^k and x_{feas}^k at the start of a given iteration k are defined in Definition 2.9 after ranking the trial mesh points of $\mathcal{X}$, making use of the following dominance notion inspired by [9].

Definition 2.8. *The ε-feasible point $x^k + s^k$ is said to dominate the ε-feasible point x^k, denoted $x^k + s^k \prec_{f;\varepsilon} x^k$, provided $f_s^k - f_0^k \leq -\gamma\varepsilon(\delta_p^k)^2$ and $u_s^k(x^k + s^k) = 0$.*
The ε-infeasible point $x^k + s^k$ is said to dominate the ε-infeasible point x^k, denoted $x^k + s^k \prec_{h;\varepsilon} x^k$, provided $f_s^k - f_0^k \leq -\gamma\varepsilon(\delta_p^k)^2$, $h_s^k - h_0^k \leq -\gamma m\varepsilon(\delta_p^k)^2$ and $0 < u_s^k(x^k + s^k) \leq h_{\max}^k$.

Definition 2.9. *Let $\mathscr{E}_k$ be the set of points where the objective and constraint functions have been evaluated at a given iteration k. If no ε-feasible point is generated by Algorithm 2, then there is no ε-feasible solution. Otherwise, let $t \geq 1$ be such that $t-1$ is the iteration where a first ε-feasible point is found. Then $x_{\text{feas}}^t \in \{x^{t-1} + s^{t-1} \in \mathscr{E}_{t-1} : u_s^{t-1}(x^{t-1} + s^{t-1}) = 0\}$ is an ε-feasible incumbent solution at the start of iteration t. Define $\mathscr{F}_k(y) = \left\{x^k + s^k \in \mathscr{E}_k : u_s^k(x^k + s^k) = 0 \text{ and } x^k + s^k \prec_{f;\varepsilon} y\right\}$ for all $k \geq t$ with $\mathscr{F}_k(y) = \emptyset$ if $k \leq t-1$. Define the sets $\mathscr{D}_k(x^k) = \left\{x^k + s^k \in \mathscr{E}_k : x^k + s^k \prec_{h;\varepsilon} x^k\right\}$ and $\mathscr{I}_k(x^k) = \left\{x^k + s^k \in \mathscr{E}_k : h_s^k(x^k + s^k) - h_0^k(x^k) \leq -\gamma m\varepsilon(\delta_p^k)^2\right\}$ for all $k \geq 0$. Let $x_{\text{inf}}^0 \in \mathcal{X}$ be a starting point. For all $k \geq t$, an ε-feasible incumbent solution at iteration $k+1$ is defined as:*

$$x_{\text{feas}}^{k+1} \in \begin{cases} \mathscr{F}_k(x_{\text{feas}}^k) & \text{if } \mathscr{F}_k(x_{\text{feas}}^k) \neq \emptyset \\ \left\{x_{\text{feas}}^k\right\} & \text{otherwise.} \end{cases}$$

For all $k \geq 0$, an ε-infeasible incumbent solution at iteration $k+1$ is defined as:

$$x_{\text{inf}}^{k+1} \in \begin{cases} \mathscr{D}_k(x_{\text{inf}}^k) & \text{if } \mathscr{F}_k(x_{\text{feas}}^k) = \emptyset \text{ and } \mathscr{D}_k(x_{\text{inf}}^k) \neq \emptyset \\ \underset{x_{\text{inf}}^k + s^k \in \mathscr{I}_k(x_{\text{inf}}^k)}{\arg\min} u_s^k(x_{\text{inf}}^k + s^k) & \text{if } \mathscr{I}_k(x_{\text{inf}}^k) \neq \emptyset \text{ and } \mathscr{F}_k(x_{\text{feas}}^k) \cup \mathscr{D}_k(x_{\text{inf}}^k) = \emptyset \\ \left\{x_{\text{inf}}^k\right\} & \text{otherwise.} \end{cases}$$

2.2 The StoMADS-PB algorithm and parameter update

Recall first that MADS with PB is an iterative algorithm where every iteration is comprised of two main steps: an optional step called the SEARCH [9, 12], and the POLL. The SEARCH which typically consists of a global exploration may use a plethora of strategies like those based on interpolation models, heuristics and surrogate functions or simplified physics models [9] to explore the variables space. Each iteration of StoMADS-PB can also allow a SEARCH step, but it is not shown here for simplicity. Similarly to MADS with PB, the POLL step of StoMADS-PB is more rigidly defined, unlike the freedom of the SEARCH, and consists of a local exploration. During each of these two steps, a finite number of trial points is generated on an underlying *mesh* $\mathcal{M}^k$. The mesh is a discretization of the variable space, whose coarseness or fineness is controlled by a mesh size parameter δ_m^k, thus deviating from the notation Δ_k^m from [9], since uppercase letters will be used to denote random variables. For the remainder of the manuscript, $s^k = \delta_m^k d^k$ where d^k is a nonzero direction. The POLL

step is governed by the poll size parameter δ_p^k which is linked to δ_m^k by $\delta_m^k = \min\{\delta_p^k, (\delta_p^k)^2\}$ [12]. As specified earlier, $\{\delta_p^k\}_{k\in\mathbb{N}}$ will play the role of the sequence of nonnegative real numbers introduced in Definition 2.1. Let $\hat{z} \in \mathbb{N}$ be a large fixed integer and $\tau \in (0,1) \cap \mathbb{Q}$ be a fixed rational constant. For the needs of Section 4, note also that as in [11], δ_p^k is supposed to be bounded above by the positive and fixed constant $\tau^{-\hat{z}}$ in order for the random poll size parameter Δ_p^k introduced later in Section 3 to be integrable. The notion of a *positive spanning set* introduced in the following definition from [12] is required to define the mesh $\mathcal{M}^k$ and the POLL set $\mathcal{P}^k$.

Definition 2.10. *The positive span of a set $\mathbb{D} \subseteq \mathbb{R}^n$, denoted* pspan$(\mathbb{D})$*, is the set of all nonnegative linear combinations of vectors in $\mathbb{D}$:*

$$\text{pspan}(\mathbb{D}) = \left\{ \sum_{\ell} \lambda_\ell d^\ell : \lambda_\ell \geq 0, d^\ell \in \mathbb{D} \right\} \subseteq \mathbb{R}^n.$$

The set $\mathbb{D}$ is a positive spanning set for $\mathbb{R}^n$ if and only if pspan$(\mathbb{D}) = \mathbb{R}^n$.

The definitions of the mesh $\mathcal{M}^k$ and the POLL set $\mathcal{P}^k$ inspired by [9] are given next.

Definition 2.11. *Let $\mathbf{D} \in \mathbb{R}^{n\times p}$ be a matrix, with columns denoted by the set $\mathbb{D}$, which form a positive spanning set. At the beginning of iteration k, let x_{inf}^k and x_{feas}^k denote respectively the ε-infeasible and the ε-feasible incumbent solutions (there might be only one), and let $\mathcal{V}^k := \{x_{\text{inf}}^k, x_{\text{feas}}^k\}$ be the set of such incumbents. The mesh $\mathcal{M}^k$ and the POLL set $\mathcal{P}^k$ are respectively*

$$\mathcal{M}^k := \{x^k + \delta_m^k d : x^k \in \mathcal{V}^k,\ d = \mathbf{D}y,\ y \in \mathbb{Z}^p\} \quad \textit{and} \quad \mathcal{P}^k := \mathcal{P}^k(x_{\text{inf}}^k) \cup \mathcal{P}^k(x_{\text{feas}}^k),$$

where $\forall x^k \in \mathcal{M}^k \cap \mathcal{X}$, $\mathcal{P}^k(x^k) = \{x^k + \delta_m^k d^k \in \mathcal{M}^k \cap \mathcal{X} : \delta_m^k \left\|d^k\right\|_\infty \leq \delta_p^k b,\ d^k \in \mathbb{D}_p^k(x^k)\}$ is called a frame around x^k, with $b = \max\{\|d'\|_\infty, d' \in \mathbb{D}\}$. $\mathbb{D}_p^k(x^k)$ is a positive spanning set which is said to be a set of frame directions used for polling around x^k. The set $\mathbb{D}_p^k$ of all polling directions at iteration k is defined by $\mathbb{D}_p^k := \mathbb{D}_p^k(x_{\text{inf}}^k) \cup \mathbb{D}_p^k(x_{\text{feas}}^k)$. When there is no incumbent ε-feasible solution x_{feas}^k, then the set $\mathcal{V}^k$ is reduced to $\{x_{\text{inf}}^k\}$, in which case $\mathcal{P}^k = \mathcal{P}^k(x_{\text{inf}}^k)$ and $\mathbb{D}_p^k = \mathbb{D}_p^k(x_{\text{inf}}^k)$.

The set $\mathbb{D}_p^k(x^k)$ of directions used for polling in Algorithm 2 can be created using Algorithm 1 [11, 12]. A definition of the round$(\cdot)$ function used in the latter algorithm can be found in [11].

Algorithm 1: Creating the set $\mathbb{D}_p^k(x^k)$ of directions for polling around x^k

Given $x^k \in \mathcal{M}^k \cap \mathcal{X}, v^k \in \mathbb{R}^n$ with $\left\|v^k\right\| = 1$ 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 $\mathbf{H}^k = [h_1\ h_2 \ldots h_n]$

[2] Create poll set

Define $\mathbb{B}^k = \{b_1, b_2, \ldots, b_n\}$ with $b_j = \text{round}\left(\frac{\delta_p^k}{\delta_m^k} \frac{h_j}{\|h_j\|_\infty}\right) \in \mathbb{Z}^n$, $j = 1, 2, \ldots, n$

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

After the POLL step is completed, StoMADS-PB computes not only estimates f_0^k, f_s^k, h_0^k and h_s^k of $f(x^k)$, $f(x^k + s^k)$, $h(x^k)$ and $h(x^k + s^k)$, respectively at trial points $x^k \in \mathcal{V}^k$ and $x^k + s^k \in \mathcal{P}^k$, but

also upper bounds $u_s^k(x^k+s^k)$ and $u_0^k(x_{\text{inf}}^k)$, respectively for $h(x^k+s^k)$ and $h(x_{\text{inf}}^k)$. The values of such estimates and bounds determine the iteration type of the algorithm and govern the way δ_p^k is updated. Adapting the terminologies from [9] and depending on the values of the aforementioned estimates and bounds, there are four StoMADS-PB iteration types: an iteration can be either f-Dominating, h-Dominating (the former and the latter are referred to as Dominating iterations), Improving, or Unsuccessful. During a Dominating iteration, either the algorithm has evaluated its first ε-feasible point or a trial point that dominates an incumbent is generated. An iteration which is Improving is not Dominating but it aims to improve the feasibility of the ε-infeasible incumbent. Unsuccessful iterations are those that are neither Dominating nor Improving.

- At the beginning of iteration k, if no available ε-feasible solution has been evaluated yet, then the iteration is called f-Dominating if for $x^k \in \mathcal{V}^k$, a trial point $x^k + s^k \in \mathcal{P}^k$ satisfying $u_s^k(x^k + s^k) = 0$ is found, in which case $h(x^k + s^k) = 0$ due to Proposition 2.2, meaning that x^k+s^k is ε-feasible. Otherwise, if an ε-feasible point that dominates the incumbent is generated, i.e., $x^k+s^k \prec_{f;\varepsilon} x_{\text{feas}}^k$ for some $x^k \in \mathcal{V}^k$, then the inequality $f_s^k(x^k+s^k)-f_0^k(x_{\text{feas}}^k) \leq -\gamma\varepsilon(\delta_p^k)^2$ leads to a decrease in f due to Proposition 2.7. In either case, $x_{\text{feas}}^{k+1} := x^k + s^k$ and $\delta_p^{k+1} = \min\{\tau^{-1}\delta_p^k, \tau^{-\hat{z}}\}$. The ε-infeasible incumbent x_{inf}^k is not updated, since there is no feasibility improvement.

- Iteration k is said to be h-Dominating whenever an ε-infeasible point that dominates the incumbent is generated, i.e., $x_{\text{inf}}^k + s^k \prec_{h;\varepsilon} x_{\text{inf}}^k$, which means that both inequalities $f_s^k(x_{\text{inf}}^k + s^k) - f_0^k(x_{\text{inf}}^k) \leq -\gamma\varepsilon(\delta_p^k)^2$ and $h_s^k(x_{\text{inf}}^k + s^k) - h_0^k(x_{\text{inf}}^k) \leq -\gamma m\varepsilon(\delta_p^k)^2$ hold. Consequently, it follows from Propositions 2.4 and 2.7 that decreases occur both in f and h. In this case, $x_{\text{feas}}^{k+1} = x_{\text{feas}}^k$ and since feasibility is improved, x_{inf}^{k+1} is set to equal $x_{\text{inf}}^k + s^k$ while the poll size parameter is updated as in f-Dominating iterations.

- Iteration k is said to be Improving if it is not Dominating but at least one ε-infeasible point $x_{\text{inf}}^k + s^k$ is evaluated satisfying $h_s^k(x_{\text{inf}}^k + s^k) - h_0^k(x_{\text{inf}}^k) \leq -\gamma m\varepsilon(\delta_p^k)^2$. Indeed, this means that $x_{\text{inf}}^k + s^k$ improves the feasibility of the ε-infeasible incumbent x_{inf}^k since the previous inequality leads to a decrease in h due to Proposition 2.4. In this case, δ_p^k is updated as in Dominating iterations, $x_{\text{feas}}^{k+1} = x_{\text{feas}}^k$ while the ε-infeasible incumbent is updated according to

$$x_{\text{inf}}^{k+1} \in \underset{x_{\text{inf}}^k+s^k}{\text{argmin}} \left\{ u_s^k(x_{\text{inf}}^k + s^k) : h_s^k(x_{\text{inf}}^k + s^k) - h_0^k(x_{\text{inf}}^k) \leq -\gamma m\varepsilon(\delta_p^k)^2 \right\}.$$

- Finally, an iteration is called Unsuccessful if it is neither Dominating nor Improving. In this case, $\delta_p^{k+1} = \tau\delta_p^k$ while neither x_{inf}^k nor x_{feas}^k are updated.

While x_{inf}^k is updated at the end of each iteration of StoMADS-PB, the barrier threshold is computed at the beginning of each iteration according to $h_{\max}^k = u_0^k(x_{\text{inf}}^k)$ in order to avoid keeping its possibly inaccurate values from one iteration to another. In fact, estimates in StoMADS-PB are always computed at the beginning of each iteration and their accuracies are improved compared to previous iterations as will be seen in Section 5.1. Consequently, even though the sequence $\{h_{\max}^k\}_{k\in\mathbb{N}}$ has a decreasing tendency, it can possibly increase between successive iterations, unlike in the deterministic setting. The goal of StoMADS-PB is to accept only trial points satisfying $h(x^k) \leq h_{\max}^k$. Any trial point x^k for which the inequality $u_0^k(x^k) \leq h_{\max}^k$ does not hold is discarded since such an inequality implies that $h(x^k) \leq h_{\max}^k$ due to (4). However, this is a sufficient acceptance condition since

$u_0^k(x^k) > h_{\max}^k$ does not necessarily imply that $h(x^k) \leq h_{\max}^k$ does not hold, but rather leads to a situation of uncertainty which is not explicitly distinguished in the present manuscript for the sake of simplicity.

Remark 2.12. *Let $t \geq 1$ be such that $t-1$ is the index of the first f-Dominating iteration of Algorithm 2 and assume that $t < +\infty$. Then in Algorithm 2, $x_{\text{feas}}^k = x_{\text{inf}}^0$ for all $k = 0, 1, \ldots, t-1$ while $x_{\text{feas}}^t = x_{\text{feas}}^{(t-1)+1} \neq x_{\text{inf}}^0$. Moreover, even though estimates $f_0^k(x_{\text{feas}}^k)$, $f_s^k(x_{\text{feas}}^k + s^k)$, $h_0^k(x_{\text{feas}}^k)$ and $h_s^k(x_{\text{feas}}^k + s^k)$ are computed at x_{feas}^k and $x_{\text{feas}}^k + s^k \in \mathcal{P}^k$ respectively for all $k \leq t-1$, they are not used by the algorithm until the end of iteration $t-1$ and one can even notice that for all $k \leq t-1$, x_{feas}^k is not an ε-feasible point in the sense of Definition 2.9. Furthermore, no point in $\mathcal{P}^k$ generated using $\mathbb{D}_p^k(x_{\text{feas}}^k)$ is evaluated until the end of iteration $t-1$. In fact, setting x_{feas}^0 to equal x_{inf}^0 as is done in Algorithm 2 and then computing the latter estimates are not necessary in practice. However, doing so allows simply the aforementioned estimates to be defined for all $k \geq 0$ for theoretical needs, specifically the construction of the σ-algebra $\mathcal{F}_{k-1}^{C \cdot F}$ in Section 3. As emphasized in Definition 2.11, observe that for all $k \leq t-1$, there is only one incumbent (ε-infeasible) solution x_{inf}^k according to Definition 2.9.*

2.3 Frame center selection rule

Before describing the frame center selection rule, recall the set $\mathcal{V}^k$ of incumbent solutions introduced in Definition 2.11 and the fact that POLL trial points are generated inside frames around such incumbents. At a given iteration, there are either one or two frame centers in $\mathcal{V}^k$. When $\mathcal{V}^k$ contains only one point, then using terminologies from [9], that point is called the *primary frame center*. In the event that there are two incumbent solutions x_{inf}^k and x_{feas}^k, one of them is chosen as the primary frame center while the other one is the *secondary frame center*. Because of the unavailability of f function values for StoMADS-PB, a specific frame center selection strategy (inspired by Section 2.5 of [9]) using estimates of such function values is proposed and relies on the following result.

Proposition 2.13. *Let $f_0^k(x_{\text{feas}}^k)$ and $f_0^k(x_{\text{inf}}^k)$ be ε-accurate estimates of $f(x_{\text{feas}}^k)$ and $f(x_{\text{inf}}^k)$ respectively. Let $\rho > 0$ be a scalar.*

$$\text{If } f_0^k(x_{\text{feas}}^k) - \rho > f_0^k(x_{\text{inf}}^k) + 2\varepsilon(\delta_p^k)^2, \text{ then } f(x_{\text{feas}}^k) - \rho > f(x_{\text{inf}}^k). \tag{11}$$

Proof. Assume that $f_0^k(x_{\text{feas}}^k) - \rho > f_0^k(x_{\text{inf}}^k) + 2\varepsilon(\delta_p^k)^2$. Then, it follows from the ε-accuracy of $f_0^k(x_{\text{feas}}^k)$ and $f_0^k(x_{\text{inf}}^k)$ that

$$\begin{aligned} f(x_{\text{inf}}^k) - f(x_{\text{feas}}^k) &= \left[f(x_{\text{inf}}^k) - f_0^k(x_{\text{inf}}^k)\right] + \left[f_0^k(x_{\text{inf}}^k) - f_0^k(x_{\text{feas}}^k)\right] + \left[f_0^k(x_{\text{feas}}^k) - f(x_{\text{feas}}^k)\right] \\ &< 2\varepsilon(\delta_p^k)^2 - (\rho + 2\varepsilon(\delta_p^k)^2) = -\rho. \end{aligned}$$

□

Based on the result of Proposition 2.13, x_{feas}^k is always chosen as the StoMADS-PB primary frame center unless the estimates $f_0^k(x_{\text{feas}}^k)$ and $f_0^k(x_{\text{inf}}^k)$ satisfy the sufficient decrease condition in (11) leading to the inequality $f(x_{\text{feas}}^k) - \rho > f(x_{\text{inf}}^k)$, which as in [9] allows the choice of the infeasible incumbent solution as primary frame center.

Algorithm 2: StoMADS-PB

[0] Initialization

choose $x_{\text{inf}}^0 \in \mathcal{X}$, $\delta_p^0 > 0$, $\tau \in (0,1) \cap \mathbb{Q}$, $\varepsilon > 0$, $\gamma > 2$ and $\hat{z} \in \mathbb{N}^*$

set the feasibility success $flag$ = FALSE, $\mathcal{V}^0 \leftarrow \{x_{\text{inf}}^0\}$ and $x_{\text{feas}}^0 \leftarrow x_{\text{inf}}^0$

set the iteration counter $k \leftarrow 0$

[1] Parameter Update

set $\delta_m^k \leftarrow \min\{\delta_p^k, (\delta_p^k)^2\}$

[2] Poll

generate a finite list $\mathcal{P}^k$ of candidates using the polling directions $\mathbb{D}_p^k(x_{\text{inf}}^k) \cup \mathbb{D}_p^k(x_{\text{feas}}^k)$

obtain estimates f_0^k, f_s^k, h_0^k and h_s^k of $f(x^k), f(x^k+s^k), h(x^k)$ and $h(x^k+s^k)$ respectively, at $x^k \in \mathcal{V}^k \cup \{x_{\text{feas}}^k\}$, $x^k + s^k \in \mathcal{P}^k$, then compute bounds $u_s^k(x^k+s^k)$ and $u_0^k(x_{\text{inf}}^k)$, using blackbox evaluations

set the barrier threshold $h_{\max}^k \leftarrow u_0^k(x_{\text{inf}}^k)$

if $flag$ = FALSE and $u_s^k(x^k+s^k) = 0$ or $flag$ = TRUE and $x^k + s^k \prec_{f;\varepsilon} x_{\text{feas}}^k$ for some $x^k \in \mathcal{V}^k$ and $s^k \in \{\delta_m^k d^k : d^k \in \mathbb{D}_p^k(x^k)\}$ (**f-Dominating**)

set $x_{\text{inf}}^{k+1} \leftarrow x_{\text{inf}}^k$, $x_{\text{feas}}^{k+1} \leftarrow x^k + s^k$ and $\delta_p^{k+1} \leftarrow \min\{\tau^{-1}\delta_p^k, \tau^{-\hat{z}}\}$

reset the feasibility success $flag$ = TRUE, set $\mathcal{V}^{k+1} \leftarrow \{x_{\text{inf}}^{k+1}, x_{\text{feas}}^{k+1}\}$ and go to **[4]**

else if $x_{\text{inf}}^k + s^k \prec_{h;\varepsilon} x_{\text{inf}}^k$ for some $s^k \in \{\delta_m^k d^k : d^k \in \mathbb{D}_p^k(x_{\text{inf}}^k)\}$ (**h-Dominating**)

set $x_{\text{inf}}^{k+1} \leftarrow x_{\text{inf}}^k + s^k$, $x_{\text{feas}}^{k+1} \leftarrow x_{\text{feas}}^k$ and $\delta_p^{k+1} \leftarrow \min\{\tau^{-1}\delta_p^k, \tau^{-\hat{z}}\}$

else if $h_s^k(x_{\text{inf}}^k + s^k) - h_0^k(x_{\text{inf}}^k) \leq -\gamma m\varepsilon(\delta_p^k)^2$ for some previously evaluated $x_{\text{inf}}^k + s^k$ (**Improving**)

set $x_{\text{inf}}^{k+1} \in \operatorname{argmin}_{x_{\text{inf}}^k + s^k}\{u_s^k(x_{\text{inf}}^k + s^k) : h_s^k(x_{\text{inf}}^k + s^k) - h_0^k(x_{\text{inf}}^k) \leq -\gamma m\varepsilon(\delta_p^k)^2\}$

$x_{\text{feas}}^{k+1} \leftarrow x_{\text{feas}}^k$ and $\delta_p^{k+1} \leftarrow \min\{\tau^{-1}\delta_p^k, \tau^{-\hat{z}}\}$

otherwise (**Unsuccessful**), set $x_{\text{inf}}^{k+1} \leftarrow x_{\text{inf}}^k$, $x_{\text{feas}}^{k+1} \leftarrow x_{\text{feas}}^k$ and $\delta_p^{k+1} \leftarrow \tau\delta_p^k$

[3] Feasibility update

if $flag$ = TRUE

set $\mathcal{V}^{k+1} \leftarrow \{x_{\text{inf}}^{k+1}, x_{\text{feas}}^{k+1}\}$

otherwise, $\mathcal{V}^{k+1} \leftarrow \{x_{\text{inf}}^{k+1}\}$

[4] Termination

if no termination criterion is met

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

otherwise stop

Figure 1: StoMADS-PB algorithm for constrained stochastic optimization.

3 Stochastic process generated by StoMADS-PB

The stochastic quantities in the present work are all defined on the same probability space $(\Omega, \mathcal{G}, \mathbb{P})$. The nonempty set Ω is referred to as the *sample space* and its subsets are called *events*. The collection $\mathcal{G}$ of such events is called a *σ-algebra* or *σ-field* and $\mathbb{P}$ is a finite measure satisfying $\mathbb{P}(\Omega) = 1$, referred to as *probability measure* and defined on the measurable space $(\Omega, \mathcal{G})$. Each element $\omega \in \Omega$ is referred to as a *sample point*. Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel σ-algebra of $\mathbb{R}^n$, i.e., the one generated by its open sets. A random variable X is a measurable map defined on $(\Omega, \mathcal{G}, \mathbb{P})$ into the measurable space

$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, where measurability means that each event $\{X \in B\} := X^{-1}(B)$ belongs to $\mathcal{G}$ for all $B \in \mathcal{B}(\mathbb{R}^n)$ [20, 33].

The estimates $f_0^k(x^k)$, $f_s^k(x^k+s^k)$, $c_{j,0}^k(x^k)$ and $c_{j,s}^k(x^k+s^k)$, for $j = 1, 2, \ldots, m$, $x^k \in \{x_{\text{inf}}^k, x_{\text{feas}}^k\}$ and $x^k + s^k \in \mathcal{P}^k$, of function values are computed at every iteration of Algorithm 2 using the noisy blackbox evaluations. Because of the randomness of the blackbox outputs, such estimates can respectively be considered as realizations of random estimates $F_0^k(X^k)$, $F_s^k(X^k + S^k)$, $C_{j,0}^k(X^k)$ and $C_{j,s}^k(X^k + S^k)$, for $j = 1, 2, \ldots, m$. Since each iteration k of Algorithm 2 is influenced by the randomness stemming from such random estimates, Algorithm 2 results in a stochastic process. For the remainder of the manuscript, uppercase letters will be used to denote random quantities while their realizations will be denoted by lowercase letters. Thus, $x^k = X^k(\omega)$, $x_{\text{inf}}^k = X_{\text{inf}}^k(\omega)$, $x_{\text{feas}}^k = X_{\text{feas}}^k(\omega)$, $s^k = S^k(\omega)$, $\delta_p^k = \Delta_p^k(\omega)$ and $\delta_m^k = \Delta_m^k(\omega)$ denote respectively realizations of X^k, X_{inf}^k, X_{feas}^k, S^k, Δ_p^k and Δ_m^k. Similarly, $f_0^k(x^k) = F_0^k(X^k)(\omega)$, $f_s^k(x^k + s^k) = F_s^k(X^k + S^k)(\omega)$, $c_{j,0}^k(x^k) = C_{j,0}^k(X^k)(\omega)$, $c_{j,s}^k(x^k + s^k) = C_{j,s}^k(X^k + S^k)(\omega)$, $h_0^k(x^k) = H_0^k(X^k)(\omega)$, $h_s^k(x^k + s^k) = H_s^k(X^k + S^k)(\omega)$, $\ell_0^k(x^k) = L_0^k(X^k)(\omega)$, $\ell_s^k(x^k + s^k) = L_s^k(X^k + S^k)(\omega)$, $u_0^k(x^k) = U_0^k(X^k)(\omega)$ and $u_s^k(x^k + s^k) = U_s^k(X^k + S^k)(\omega)$. When there is no ambiguity, F_0^k will be used instead of $F_0^k(X^k)$, etc. In general, following the notations in [11, 21, 23, 33, 48], F_0^k, F_s^k, H_0^k and H_s^k are respectively the estimates of $f(X^k)$, $f(X^k + S^k)$, $h(X^k)$ and $h(X^k + S^k)$. Moreover, as highlighted in [11], the notation "$f(X^k)$" is used to denote the random variable with realizations $\{f(X^k(\omega)) : \omega \in \Omega\}$.

The present research aims to show that the stochastic process $\{X_{\text{inf}}^k, X_{\text{feas}}^k, \Delta_p^k, \Delta_m^k, F_0^k, F_s^k, H_0^k, H_s^k, L_0^k, U_0^k, L_s^k, U_s^k\}$ resulting from Algorithm 2 has desirable convergence properties with probability one under some assumptions on the estimates F_0^k, F_s^k, $C_{j,0}^k$, $C_{j,s}^k$, H_0^k, H_s^k and on the bounds L_0^k, U_0^k, L_s^k, U_s^k. In particular, the estimates F_0^k, F_s^k, $C_{j,0}^k$ and $C_{j,s}^k$ will be assumed to be ε-accurate while the bounds will be assumed to be ε-reliable, with sufficiently high, but fixed, probabilities conditioned on the past.

Probabilistic bounds and probabilistic estimates

The notion of conditioning on the past is formalized following [11, 21, 23, 33, 48]. Denote by $\mathcal{F}_{k-1}^{C\cdot F}$ the σ-algebra generated by $F_0^\ell(X^\ell)$, $F_s^\ell(X^\ell + S^\ell)$, $C_{j,0}^\ell(X^\ell)$ and $C_{j,s}^\ell(X^\ell + S^\ell)$, for $j = 1, 2, \ldots, m$, for $X^\ell \in \{X_{\text{inf}}^\ell, X_{\text{feas}}^\ell\}$ and for $\ell = 0, 1, \ldots, k-1$. For completeness, $\mathcal{F}_{-1}^{C\cdot F}$ is set to equal $\sigma(x^0) = \sigma(x_{\text{inf}}^0)$. Thus, $\{\mathcal{F}_k^{C\cdot F}\}_{k\geq -1}$ is a filtration, i.e., a sequence of increasing σ-algebras of $\mathcal{G}$.

Sufficient accuracy of function estimates is measured using the poll size parameter and is formalized, following [11, 21, 23, 33, 48], by means of the definitions below.

Definition 3.1. *A sequence of random estimates $\{F_0^k, F_s^k\}$ is said to be β-probabilistically ε-accurate with respect to the sequence $\{X^k, S^k, \Delta_p^k\}$ if the events*

$$J_k = \{F_0^k, F_s^k, \text{ are } \varepsilon\text{-accurate estimates of } f(x^k) \text{ and } f(x^k + s^k), \text{ respectively for } \Delta_p^k\}$$

satisfy the submartingale-like condition

$$\mathbb{P}\left(J_k \mid \mathcal{F}_{k-1}^{C\cdot F}\right) = \mathbb{E}\left(\mathbb{1}_{J_k} \mid \mathcal{F}_{k-1}^{C\cdot F}\right) \geq \beta,$$

where $\mathbb{1}_{J_k}$ denotes the indicator function of the event J_k, i.e., $\mathbb{1}_{J_k} = 1$ if $\omega \in J_k$ and $\mathbb{1}_{J_k} = 0$ otherwise. The estimates are called "good" if $\mathbb{1}_{J_k} = 1$. Otherwise they are called "bad".

Definition 3.2. *A sequence of random estimates* $\{C_{j,0}^k, C_{j,s}^k\}$ *is said to be* $\alpha^{1/m}$*-probabilistically* ε*-accurate for some* $j = 1, 2, \ldots, m$ *with respect to the corresponding sequence* $\{X^k, S^k, \Delta_p^k\}$ *if the events*

$$I_k^j = \{C_{j,0}^k, C_{j,s}^k, \text{ are } \varepsilon\text{-accurate estimates of } c_j(x^k) \text{ and } c_j(x^k + s^k), \text{ respectively for } \Delta_p^k\}$$

satisfy the submartingale-like condition

$$\mathbb{P}\left(I_k^j \mid \mathcal{F}_{k-1}^{C\cdot F}\right) = \mathbb{E}\left(\mathbb{1}_{I_k^j} \mid \mathcal{F}_{k-1}^{C\cdot F}\right) \geq \alpha^{1/m}.$$

Recall the ε-reliable bounds ℓ_0^k, u_0^k for $h(x^k)$, and ℓ_s^k, u_s^k for $h(x^k + s^k)$ provided by Proposition 2.2 making use of ε-accurate estimates $c_{j,0}^k$ and $c_{j,s}^k$, for all $j \in J$. Since Algorithm 2 results in a stochastic process introducing random bounds L_0^k, U_0^k, L_s^k and U_s^k with realizations $\ell_0^k, u_0^k, \ell_s^k$ and u_s^k respectively, the reliability of such random bounds has to be quantified probabilistically, using Definition 3.2 and inspired by Definition 3.1. Indeed, in order to show later in Section 4 that the stochastic process resulting from Algorithm 2 has desirable convergence properties, the reliability of the random bounds will be required to hold with sufficiently high probability. This notion of sufficient reliability introduced in the present work is formalized next.

Definition 3.3. *A sequence of random bounds* $\{L_0^k, U_0^k, L_s^k, U_s^k\}$ *is said to be* α*-probabilistically* ε*-reliable with respect to the corresponding sequence* $\{X^k, S^k, \Delta_p^k\}$ *if the events*

$$\begin{aligned} I_k \quad = \quad & \{\text{“}L_0^k \text{ and } U_0^k \text{ are } \varepsilon\text{-reliable bounds for } h(x^k)\text{”, and “}L_s^k \text{ and } U_s^k \text{ are } \varepsilon\text{-reliable bounds} \\ & \text{for } h(x^k + s^k)\text{”, respectively for } \Delta_p^k\} \end{aligned}$$

satisfy the submartingale-like condition

$$\mathbb{P}\left(I_k \mid \mathcal{F}_{k-1}^{C\cdot F}\right) = \mathbb{E}\left(\mathbb{1}_{I_k} \mid \mathcal{F}_{k-1}^{C\cdot F}\right) \geq \mathbb{P}\left(\bigcap_{j=1}^{m} I_k^j \mid \mathcal{F}_{k-1}^{C\cdot F}\right) \geq \alpha,$$

The bounds are called “good” if $\mathbb{1}_{I_k} = 1$*. Otherwise,* $\mathbb{1}_{I_k} = 0$ *and they are called “bad”.*

The p-integrability of random variables [11, 20] is defined below and will be useful for the analysis of Algorithm 2.

Definition 3.4. *Let* $(\Omega, \mathcal{G}, \mathbb{P})$ *be a probability space and let* $p \in [1, +\infty)$ *be an integer. Then the space* $\mathbb{L}^p(\Omega, \mathcal{G}, \mathbb{P})$ *of so-called* p*-integrable random variables is the set of all real-valued random variables* X *such that*

$$\|X\|_p := \left(\int_\Omega |X(\omega)|^p \, \mathbb{P}(d\omega)\right)^{\frac{1}{p}} =: \left(\mathbb{E}\left(|X|^p\right)\right)^{\frac{1}{p}} < +\infty.$$

As in [11], the following is assumed in order for the random variables $f(X^k)$, $h(X^k)$ and $c_j(X^k)$, $j \in J$, to be integrable so that the conditional expectations $\mathbb{E}\left(f(X^k)|\mathcal{F}_{k-1}^{C\cdot F}\right)$, $\mathbb{E}\left(c_j(X^k)|\mathcal{F}_{k-1}^{C\cdot F}\right)$, $j \in J$ and $\mathbb{E}\left(h(X^k)|\mathcal{F}_{k-1}^{C\cdot F}\right)$ are well-defined [20].

In the assumption below, f and h are assumed to be locally Lipschitz. In general, a function $g : \mathcal{X} \to \mathbb{R}$ is called locally Lipschitz if it is Lipschitz with a finite constant in some nonempty open neighborhood intersected with $\mathcal{X}$ [9].

Assumption 2. *The objective function f and the constraint violation function h are locally Lipschitz with constants $\lambda^f > 0$ and $\lambda^h > 0$, respectively. The constraint functions c_j, $j \in J$, are continuous on $\mathcal{X}$. The set $\mathcal{U} \subset \mathcal{X}$ containing all incumbents realizations is compact.*

Proposition 3.5. *Under Assumption 2, there exists a finite constant $\kappa^f_{\max}$ satisfying $\left|f(x^k)\right| \leq \kappa^f_{\max}$ for all $x^k \in \mathcal{U}$. Moreover, the random variables $f(X^k)$, $h(X^k)$, $c_j(X^k)$ and Δ^k_p belong to $\mathbb{L}^1(\Omega, \mathcal{G}, \mathbb{P})$, for all $j \in J$ and for all $k \geq 0$.*

Proof. Since f is locally Lipschitz on the compact set $\mathcal{U}$, f is bounded on $\mathcal{U}$. Consequently, there exists a finite constant $\kappa^f_{\max}$ such that $\left|f(x^k)\right| \leq \kappa^f_{\max}$ for all $x^k \in \mathcal{U}$. Similarly, there exist $\kappa^h_{\max}$ satisfying $\left|h(x^k)\right| \leq \kappa^h_{\max}$ and $\kappa^c_{\max}$ such that $\left|c_j(x^k)\right| \leq \kappa^c_{\max}$ for all $j \in J$ and all $x^k \in \mathcal{U}$ since h is locally Lipschitz and c_j is continuous on $\mathcal{U}$. Thus, $\mathbb{E}\left(\left|f(X^k)\right|\right) := \int_\Omega \left|f(X^k(\omega))\right| \mathbb{P}(d\omega) \leq \kappa^f_{\max} < +\infty$. Similarly, $\mathbb{E}\left(\left|h(X^k)\right|\right) \leq \kappa^h_{\max} \leq +\infty$ and for all $j \in J$, $\mathbb{E}\left(\left|c_j(X^k)\right|\right) \leq \kappa^c_{\max} \leq +\infty$. Finally, the integrability of Δ^k_p follows from the fact that $\Delta^k_p(\omega) \leq \tau^{-\hat{z}}$ for all $\omega \in \Omega$, which implies that $\mathbb{E}\left(\left|\Delta^k_p\right|\right) := \int_\Omega \left|\Delta^k_p(\omega)\right| \mathbb{P}(d\omega) \leq \tau^{-\hat{z}} < +\infty$. □

Next are stated some key assumptions on the stochastic variables in Algorithm 2, some of which are made in [11] and will be useful for the convergence analysis in Section 4. Approaches for computing random estimates and bounds satisfying these assumptions in a simple random noise framework are discussed in Section 5.1.

Assumption 3. *For fixed $\alpha, \beta \in (0, 1)$, the following hold for the random quantities generated by Algorithm 2 at iteration k.*

(i) The sequence of estimates $\{F^k_0, F^k_s\}$ generated by Algorithm 2 is β-probabilistically ε-accurate.

(ii) The sequence of estimates $\{F^k_0, F^k_s\}$ generated by Algorithm 2 satisfies the following variance condition for all $k \geq 0$:

$$\begin{aligned} &\mathbb{E}\left(\left|F^k_s - f(X^k + S^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2(1 - \sqrt{\beta})(\Delta^k_p)^4 \\ \text{and} \quad &\mathbb{E}\left(\left|F^k_0 - f(X^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2(1 - \sqrt{\beta})(\Delta^k_p)^4. \end{aligned} \tag{12}$$

(iii) For all $j = 1, 2, \ldots, m$, the sequence of estimates $\{C^k_{j,0}, C^k_{j,s}\}$ is $\alpha^{1/m}$-probabilistically ε-accurate.

(iv) For all $j = 1, 2, \ldots, m$, the sequence of estimates $\{C^k_{j,0}, C^k_{j,s}\}$ satisfies the following variance condition for all $k \geq 0$:

$$\begin{aligned} &\mathbb{E}\left(\left|C^k_{j,s} - c_j(X^k + S^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2\left(1 - \alpha^{1/2m}\right)(\Delta^k_p)^4 \\ \text{and} \quad &\mathbb{E}\left(\left|C^k_{j,0} - c_j(X^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2\left(1 - \alpha^{1/2m}\right)(\Delta^k_p)^4. \end{aligned} \tag{13}$$

(v) The sequence of random bounds $\{L^k_0, U^k_0, L^k_s, U^k_s\}$ is α-probabilistically ε-reliable.

An iteration k for which $\mathbb{1}_{I_k}\mathbb{1}_{J_k} = 1$, i.e., for which the events I_k and J_k both occur, will be called "*true*". Otherwise, k will be called "*false*". Even though the present algorithmic framework does not allow one to determine which iterations are true or false, Theorem 3.6 shows that true iterations occur infinitely often. Theorem 3.6 will also be useful for the convergence analysis of Algorithm 2, more precisely in Subsection 4.3.

Theorem 3.6. *Assume that Assumption 3 holds for $\alpha\beta \in (1/2, 1)$. Then true iterations of Algorithm 2 occur infinitely often.*

Proof. Consider the random walk

$$W_k = \sum_{i=0}^{k} (2 \cdot \mathbb{1}_{I_i} \mathbb{1}_{J_i} - 1). \tag{14}$$

Then, since W_k is a submartingale with bounded increments (and, as such, cannot converge), the result follows from the fact that $\left\{ \limsup_{k \to +\infty} W_k = +\infty \right\}$ occurs almost surely, the proof of which can be derived from that of Theorem 4.16 in [23], where a similar random walk was studied. Indeed, the latter result means that

$$\mathbb{P}\left(\left\{ \omega \in \Omega : \exists K(\omega) \subset \mathbb{N} \text{ such that } \lim_{k \in K(\omega)} W_k(\omega) = +\infty \right\} \right) = 1,$$

which implies that $\mathbb{1}_{I_i} \mathbb{1}_{J_i} = 1$ infinitely often. □

The following lemma will be useful later in the analysis of StoMADS-PB. In fact, for a given realization of Algorithm 2, the inequality $h_s^k - h_0^k \leq -\gamma m \varepsilon (\delta_p^k)^2$ leads to a decrease in h at iteration k as was shown in Proposition 2.4 if the event I_k occurs, i.e., when the bounds are ε-reliable. But when the bounds are not ε-reliable, the algorithm can accept a step which leads to an increase in h. Later in the proof of Theorem 4.2, such an increase will be controlled in expectation by making use of (15).

Lemma 3.7. *Let Assumption 3-(iv) hold for $\alpha \in (0, 1)$. The sequence of random estimated violations $\{H_0^k, H_s^k\}$ satisfies*

$$\begin{aligned} & \mathbb{E}\left(\left| H_s^k - h(X^k + S^k) \right| \,|\, \mathcal{F}_{k-1}^{C \cdot F} \right) \leq m\varepsilon(1-\alpha)^{1/2} (\Delta_p^k)^2 \\ \textit{and} \quad & \mathbb{E}\left(\left| H_0^k - h(X^k) \right| \,|\, \mathcal{F}_{k-1}^{C \cdot F} \right) \leq m\varepsilon(1-\alpha)^{1/2} (\Delta_p^k)^2. \end{aligned} \tag{15}$$

Proof. Before showing (15), observe that

$$\begin{aligned} \left| H_0^k - h(X^k) \right| &= \left| \sum_{j=1}^{m} \max\{C_{j,0}^k, 0\} - \sum_{j=1}^{m} \max\{c_j(X^k), 0\} \right| \\ &\leq \sum_{j=1}^{m} \left| \max\{C_{j,0}^k, 0\} - \max\{c_j(X^k), 0\} \right| \leq \sum_{j=1}^{m} \left| C_{j,0}^k - c_j(X^k) \right|, \end{aligned} \tag{16}$$

where the last inequality in (16) follows from the inequality $|\max\{x, 0\} - \max\{y, 0\}| \leq |x - y|$, for all $x, y \in \mathbb{R}$. Moreover, it follows from the conditional Cauchy-Schwarz inequality [20] that for all $j \in J$,

$$\begin{aligned} \mathbb{E}\left(\left| C_{j,0}^k - c_j(X^k) \right| | \mathcal{F}_{k-1}^{C \cdot F} \right) &\leq \left[\mathbb{E}\left(\left| C_{j,0}^k - c_j(X^k) \right|^2 | \mathcal{F}_{k-1}^{C \cdot F} \right) \right]^{1/2} \times \left[\mathbb{E}\left(1 | \mathcal{F}_{k-1}^{C \cdot F} \right) \right]^{1/2} \\ &\leq \varepsilon \left(1 - \alpha^{1/2m} \right)^{1/2} (\Delta_p^k)^2 \leq \varepsilon \left(1 - \alpha \right)^{1/2} (\Delta_p^k)^2 \end{aligned} \tag{17}$$

where the first inequality in (17) follows from (13). Thus, taking the conditional expectation with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ in (16) and then using (17) yield

$$\mathbb{E}\left(\left|H_0^k - h(X^k)\right| |\mathcal{F}_{k-1}^{C\cdot F}\right) \leq \sum_{j=1}^{m} \mathbb{E}\left(\left|C_{j,0}^k - c_j(X^k)\right| |\mathcal{F}_{k-1}^{C\cdot F}\right) \leq m\varepsilon\,(1-\alpha)^{1/2}\,(\Delta_p^k)^2,$$

$$\text{and similarly} \quad \mathbb{E}\left(\left|H_s^k - h(X^k+S^k)\right| |\mathcal{F}_{k-1}^{C\cdot F}\right) \leq m\varepsilon\,(1-\alpha)^{1/2}\,(\Delta_p^k)^2.$$

□

4 Convergence analysis

Using ideas inspired by [9, 11, 23, 40, 48], this section presents convergence results of StoMADS-PB, most of which are stochastic variants of the convergence results in [9]. It introduces the random time T at which Algorithm 2 generates a first ε-feasible solution. Then assuming that T is either almost surely finite or almost surely infinite, a so-called zeroth-order result [10, 11] is derived showing that there exists a subsequence of Algorithm 2-generated random incumbents with mesh realizations becoming infinitely fine and which converges with probability one to a limit. This is achieved by showing, by means of Theorem 4.2, that the sequence of random poll size parameters converges to zero with probability one. Section 4.2 analyzes the function h and the random ε-infeasible incumbents generated by Algorithm 2. In particular, it gives conditions under which an almost sure limit of a subsequence of such incumbents is shown in Theorem 4.10 to satisfy a first-order necessary optimality condition via the Clarke generalized derivative of h with probability one. Then, a similar result for f and the sequence of ε-feasible incumbents is derived in Theorem 4.14 of Section 4.3. The proofs of the main results of this section are presented in the Appendix. For the sake of clarity in the presentation, the following definition is introduced.

Definition 4.1. *A sequence* $\left\{X^k\right\}_{k\in\mathbb{N}}$ *of* StoMADS-PB *random incumbents is either a sequence* $\left\{X_{\text{feas}}^{k\vee T}\right\}_{k\in\mathbb{N}}$ *of random* ε*-feasible incumbents provided* $T < +\infty$ *almost surely, or a sequence* $\left\{X_{\text{inf}}^k\right\}_{k\in\mathbb{N}}$ *of random* ε*-infeasible incumbents. A similar definition is considered for the sequences of realizations* $\left\{x^k\right\}_{k\in\mathbb{N}}$, $\left\{x_{\text{feas}}^{k\vee t}\right\}_{k\in\mathbb{N}}$ *and* $\left\{x_{\text{inf}}^k\right\}_{k\in\mathbb{N}}$ *of* $\left\{X^k\right\}_{k\in\mathbb{N}}$, $\left\{X_{\text{feas}}^{k\vee T}\right\}_{k\in\mathbb{N}}$ *and* $\left\{X_{\text{inf}}^k\right\}_{k\in\mathbb{N}}$ *respectively, where* t *denotes a realization of* T.

4.1 Zeroth-order convergence

Recall Remark 2.12 and denote by $\mathscr{S}_X^k = \{X_{\text{feas}}^\ell : X_{\text{feas}}^\ell \neq x_{\text{inf}}^0,\ \ell \leq k\}$ the set of all random ε-feasible incumbents generated by Algorithm 2 until the beginning of iteration k. Consider the random time T defined by

$$T := \inf\{k \geq 0 : \mathscr{S}_X^k \neq \emptyset\}. \tag{18}$$

Then, $T \geq 1$ and for all $k \geq 1$, the occurrence of the event $\{T \leq k\}$ is determined by observing the random quantities generated by Algorithm 2 until the iteration $k-1$, which means that T is a stopping time [32] for the stochastic process generated by Algorithm 2. For a given $\omega \in \Omega$, $t = T(\omega)$ is the number of iterations required by Algorithm 2 to find a first point which is ε-feasible in the

sense of Definition 2.6. Just as a BBO method in a deterministic framework is not always guaranteed to find feasible points even though the feasible region $\mathcal{D}$ is nonempty, the algorithmic framework proposed in the present manuscript does not guarantee that t will always be finite for every realization of the stochastic process generated by Algorithm 2 even when $\mathcal{D}$ is nonempty. Thus, T could either be finite almost surely depending on the optimization problem or satisfy $\mathbb{P}(T < +\infty) \leq 1 - \zeta$ for some $\zeta \in (0, 1]$. In the latter case, the algorithm will fail to generate ε-feasible incumbents with a probability of at least ζ, in which case an almost sure convergence result related to such incumbents cannot be derived. The following is therefore assumed for the remainder of the analysis.

Assumption 4. *The stopping time T associated to the stochastic process generated by Algorithm 2 is either almost surely finite or almost surely infinite.*

The next result implies that the sequence $\{\Delta_p^k\}_{k\in\mathbb{N}}$ of random frame size parameters converges to zero with probability one and will be useful for the Clarke stationarity results of Sections 4.2 and 4.3. It holds under the assumption below.

Assumption 5. *The objective function f is bounded from below, i.e., there exists $\kappa_{\min}^f \in \mathbb{R}$ such that $-\infty < \kappa_{\min}^f \leq f(x)$, for all $x \in \mathbb{R}^n$.*

Theorem 4.2. *Let Assumptions 1, 2, 4 and 5 be satisfied. Let $\gamma > 2$ and $\tau \in (0, 1) \cap \mathbb{Q}$. Let $\nu \in (0, 1)$ be chosen such that*

$$\frac{\nu}{1-\nu} \geq \frac{2(\tau^{-2} - 1)}{\gamma - 2}. \tag{19}$$

Assume further that Assumption 3 holds for α and β chosen such that

$$\alpha\beta \geq \frac{4\nu}{(1-\nu)(1-\tau^2)} \left[(1-\alpha)^{1/2} + 2(1-\beta)^{1/2}\right]. \tag{20}$$

Then, the sequence $\{\Delta_p^k\}_{k\in\mathbb{N}}$ of frame size parameters generated by Algorithm 2 satisfies

$$\sum_{k=0}^{+\infty} (\Delta_p^k)^2 < +\infty \quad \textit{almost surely}. \tag{21}$$

The following result is an immediate consequence of Theorem 4.2. It shows that the sequences $\{\Delta_m^k\}_{k\in\mathbb{N}}$ and $\{\Delta_p^k\}_{k\in\mathbb{N}}$ converge to zero almost surely respectively.

Corollary 4.3. *The following hold under all the assumptions made in Theorem 4.2*

$$\lim_{k\to+\infty} \Delta_m^k = 0 \textit{ almost surely} \quad \textit{and} \quad \lim_{k\to+\infty} \Delta_p^k = 0 \textit{ almost surely}.$$

The next result shows that, with probability one, the differences between the estimates and their corresponding true function values converge to zero. This means that Algorithm 2 behaves like an exact deterministic method asymptotically. This result will also be useful in Subsection 4.3 for the proof of Theorem 4.13.

Corollary 4.4. *Let all assumptions that were made in Theorem 4.2 hold. Then,*

$$\lim_{k\to+\infty} \left|H_0^k - h(X^k)\right| = 0 \textit{ almost surely} \quad \textit{and} \quad \lim_{k\to+\infty} \left|F_0^k - f(X^k)\right| = 0 \textit{ almost surely}. \tag{22}$$

Likewise, $\left|H_s^k - h(X^k + S^k)\right|$ and $\left|F_s^k - f(X^k + S^k)\right|$ respectively.

Definition 4.5. *A convergent subsequence $\{x^k\}_{k\in\mathcal{K}}$ of Algorithm 2 incumbents, for some subset of indices $\mathcal{K}$, is called a refining subsequence if and only if the corresponding subsequence $\{\delta_m^k\}_{k\in\mathcal{K}}$ converges to zero. The limit $\hat{x}$ is called a refined point.*

Corollary 4.3, along with the compactness hypothesis of Assumption 2 was shown in the proof of Theorem 2 in [11] to be enough to ensure the existence of refining subsequences. Indeed, conditioned on the event $E_0 = \left\{\omega \in \Omega : \lim_{k\to+\infty} \Delta_m^k(\omega) = 0\right\}$ which is almost sure due to Corollary 4.3, the aforementioned proof applies to the next theorem.

Theorem 4.6. *Let the assumptions that were made in Corollary 4.3 hold. Then there almost surely exists at least one refining subsequence $\{X^k\}_{k\in K}$ (where K is a sequence of random variables) that converges to a refined point $\hat{X}$.*

4.2 Nonsmooth optimality conditions: Results for h

This subsection aims to show that there almost surely exists a refining subsequence $\{X_{\text{inf}}^k\}_{k\in K}$ generated by StoMADS-PB that converges to a refined point $\hat{X}_{\text{inf}}$ which satisfies a necessary optimality condition via the Clarke generalized derivative of h with probability one. As in [11], this optimality result strongly relies on the requirement that the polling directions $d^k \in \mathbb{D}_p^k(x_{\text{inf}}^k)$ of Algorithm 2 are such that $\delta_p^k \left\|d^k\right\|_\infty$ never approaches zero for all k. The way such a requirement can be met is discussed in [11]. Indeed, by choosing the columns of the matrix $\mathbf{D}$ used in the definition of the mesh $\mathcal{M}^k$ to be the $2n$ positive and negative coordinate directions, $\delta_p^0 = 1$ and $\tau = 1/2$, the directions $\delta_p^k d^k$ were shown in [11] to satisfy $\delta_p^k \left\|d^k\right\|_\infty \geq 1$ whenever d^k is constructed by means of Algorithm 1. The latter key result which holds under conditions that can also be found in Theorem 8.5 of [12], is in fact proved in [11], inspired by the proof of the latter theorem, making use of the ℓ_∞ norm $\|\cdot\|_\infty$, for the polling directions. This motivates in particular the use of an ℓ_∞ norm in the analysis of StoMADS-PB and later in Definition 4.8 of *refining directions* unlike [9] where an ℓ_2 norm (i.e., the Euclidean norm) was used for the analysis of MADS with PB. The following assumption is made for the remainder of the analysis.

Assumption 6. *Let $d^k \in \mathbb{D}_p^k$ be any polling direction used by Algorithm 2 at iteration k. Then there exists a constant $d_{\min} > 0$ such that $\delta_p^k \left\|d^k\right\|_\infty \geq d_{\min}$ for all $k \geq 0$.*

The main result of this subsection relies on the properties of the random function Ψ_k^h introduced next, similar to the one used in [11].

Lemma 4.7. *Let the same assumptions that were made in Theorem 4.2 hold and assume in addition to (20) that $\alpha\beta \in (1/2, 1)$. Consider the random function Ψ_k^h with realizations ψ_k^h defined by*

$$\psi_k^h := \frac{h(x_{\text{inf}}^k) - h(x_{\text{inf}}^k + \delta_m^k d^k)}{\delta_p^k} \quad \textit{for all } k \geq 0,$$

where $d^k \in \mathbb{D}_p^k(x_{\text{inf}}^k)$. Then,

$$\liminf_{k\to+\infty} \Psi_k^h \leq 0 \textit{ almost surely.} \tag{23}$$

The following definition of refining directions [8, 12] will be useful in the analysis.

Definition 4.8. *Let $\hat{x}$ be the refined point associated with a convergent refining subsequence $\{x^k\}_{k\in\mathcal{K}}$. A direction v is said to be a refining direction for $\hat{x}$ if and only if there exists an infinite subset $\mathcal{L}\subseteq\mathcal{K}$ with polling directions $d^k\in\mathbb{D}_p^k(x^k)$ such that $v=\lim_{k\in\mathcal{L}}\frac{d^k}{\|d^k\|_\infty}$.*

The analysis in this subsection also relies on the following definitions [9]. The Clarke generalized derivative $h^\circ(\hat{x};v)$ of h at $\hat{x}\in\mathcal{X}$ in the direction $v\in\mathbb{R}^n$ is defined by

$$h^\circ(\hat{x};v):=\limsup_{\substack{y\to\hat{x},\ y\in\mathcal{X}\\ t\searrow 0,\ y+tv\in\mathcal{X}}}\frac{h(y+tv)-h(y)}{t}. \tag{24}$$

As highlighted in [9], this definition from [36] is a generalization of the original one by Clarke [25] to the case where the constraint violation function h is not defined outside $\mathcal{X}$.

The analysis involves a specific cone $T_{\mathcal{X}}^H(\hat{x}_{\text{inf}})$ called the hypertangent cone [50] to $\mathcal{X}$ at $\hat{x}_{\text{inf}}$. The hypertangent cone to a subset $\mathcal{O}\subseteq\mathcal{X}$ at $\hat{x}$ is defined by

$$T_{\mathcal{O}}^H(\hat{x}):=\{v\in\mathbb{R}^n:\exists\bar{\epsilon}>0 \text{ such that } y+tw\in\mathcal{O}\ \forall y\in\mathcal{O}\cap\mathcal{B}_{\bar{\epsilon}}(\hat{x}),w\in\mathcal{B}_{\bar{\epsilon}}(v) \text{ and } 0<t<\bar{\epsilon}\}.$$

The next lemma from elementary analysis [9] will be useful in the present analysis.

Lemma 4.9. *If $\{a_k\}$ is a bounded real sequence and $\{b_k\}$ is a convergent real sequence, then*

$$\limsup_k(a_k+b_k)=\limsup_k a_k+\lim_k b_k.$$

The next result which is a stochastic variant of Theorem 3.5 in [9] presents a necessary optimality condition (see e.g. Theorem 6.10 of [12]) based on the hypertangent cone definition. It states that almost surely, the refined point $\hat{X}_{\text{inf}}$ of a convergent ε-infeasible refining subsequence $\{X_{\text{inf}}^k\}_{k\in K}$ is a hypertangent stationary point of h over $\mathcal{X}$. Since the inequality $h(x_{\text{inf}}^k+\delta_m^k d^k)-h(x_{\text{inf}}^k)\geq 0$, on which relies the proof of Theorem 3.5 in [9] does not hold in the present stochastic setting, the proof of this novel result uses the random function Ψ_k^h and the result of Lemma 4.7.

Theorem 4.10. *Let Assumptions 1, 6 and all the assumptions made in Theorem 4.2 and Lemma 4.7 hold. Then there almost surely exists a convergent ε-infeasible refining subsequence $\{X_{\text{inf}}^k\}_{k\in K}$ generated by Algorithm 2, for some sequence $K\subseteq K'$ of random variables satisfying $\lim_{K'}\Psi_k^h\leq 0$ almost surely, such that if $\hat{x}_{\text{inf}}\in\mathcal{X}$ is a refined point for a realization $\{x_{\text{inf}}^k\}_{k\in\mathcal{K}}$ of $\{X_{\text{inf}}^k\}_{k\in K}$ for which the events $\Delta_p^k\to 0$ and $\lim_{K'}\Psi_k^h\leq 0$ both occur, and if $v\in T_{\mathcal{X}}^H(\hat{x}_{\text{inf}})$ is a refining direction for $\hat{x}_{\text{inf}}$, then $h^\circ(\hat{x}_{\text{inf}};v)\geq 0$. In particular, this means that*

$$\begin{aligned}\mathbb{P}\Big(\Big\{\omega\in\Omega:\exists K(\omega)\subseteq\mathbb{N} \text{ and } \exists\hat{X}_{\text{inf}}(\omega)=\lim_{k\in K(\omega)}X_{\text{inf}}^k(\omega),\hat{X}_{\text{inf}}(\omega)\in\mathcal{X}, \text{ such that}\\ \forall V(\omega)\in T_{\mathcal{X}}^H(\hat{X}_{\text{inf}}(\omega)),\ h^\circ(\hat{X}_{\text{inf}}(\omega);V(\omega))\geq 0\Big\}\Big)=1.\end{aligned} \tag{25}$$

Next is stated a stochastic variant of a result in [9], showing that Clarke stationarity is guaranteed with probability one when the set of refining directions is dense in a nonempty hypertangent cone to $\mathcal{X}$.

Corollary 4.11. *Let all assumptions that were made in Theorem 4.10 hold. Let $\{X_{\text{inf}}^k\}_{k\in K}$ be the ε-infeasible refining subsequence of Theorem 4.10 for some sequence $K\subseteq K'$ satisfying $\lim_{K'}\Psi_k^h\leq 0$ almost surely. If $\hat{x}_{\text{inf}}\in\mathcal{X}$ is a refined point for a realization $\{x_{\text{inf}}^k\}_{k\in\mathcal{K}}$ of $\{X_{\text{inf}}^k\}_{k\in K}$ for which the events $\Delta_p^k\to 0$ and $\lim_{K'}\Psi_k^h\leq 0$ both occur, and if the set of refining directions for $\hat{x}_{\text{inf}}$ is dense in $T_{\mathcal{X}}^H(\hat{x}_{\text{inf}})\neq\emptyset$, then $\hat{x}_{\text{inf}}$ is a Clarke stationary point for the problem $\min_{x\in\mathcal{X}}h(x)$.*

4.3 Nonsmooth optimality conditions: Results for f

The analysis presented in this subsection assumes that Algorithm 2 generates infinitely many ε-feasible points. It aims to show that there almost surely exists a refining subsequence $\{X^k_{\text{feas}}\}_{k\in K}$ generated by StoMADS-PB that converges to a refined point $\hat{X}_{\text{feas}}$ which satisfies a necessary optimality condition via the Clarke derivative of f with probability one. The following lemma will be useful in the analysis.

Lemma 4.12. *Let the same assumptions that were made in Theorem 4.2 hold and assume in addition to (20) that $\alpha\beta \in (1/2, 1)$. Assume that the random time T with realizations t is finite almost surely. Consider the random function $\Psi^{f,T}_k$ with realizations $\psi^{f,t}_k$ defined by*

$$\psi^{f,t}_k := \frac{f(x^{k\vee t}_{\text{feas}}) - f(x^{k\vee t}_{\text{feas}} + \delta^k_m d^k)}{\delta^k_p} \quad \text{for all } k \geq 0,$$

where $k \vee t := \max\{k, t\}$ and $d^k \in \mathbb{D}^k_p(x^{k\vee t}_{\text{feas}})$ denotes any polling direction at iteration k. Then,

$$\liminf_{k\to+\infty} \Psi^{f,T}_k \leq 0 \text{ almost surely.} \tag{26}$$

The next theorem shows that the almost sure limit $\hat{X}_{\text{feas}}$ of any convergent refining subsequence of ε-feasible incumbents which drives the random estimated violations $H^k_0(X^k_{\text{feas}})$ to zero almost surely, satisfies $\mathbb{P}\left(\hat{X}_{\text{feas}} \in \mathcal{D}\right) = 1$. First, the existence of such a refining subsequence can be assumed. Indeed, it is known from Theorem 3.6 that true iterations occur infinitely often provided the estimates and bounds are sufficiently accurate. In addition, every ε-feasible point x^k_{feas} accepted by Algorithm 2 satisfies $u^k_0(x^k_{\text{feas}}) = 0 \leq h^k_0(x^k_{\text{feas}}) + m\varepsilon(\delta^k_p)^2$, which implies $\ell^k_0(x^k_{\text{feas}}) = 0 \geq h^k_0(x^k_{\text{feas}}) - m\varepsilon(\delta^k_p)^2$ (see e.g. (8) and (9)), and consequently $-m\varepsilon(\delta^k_p)^2 \leq h^k_0(x^k_{\text{feas}}) \leq m\varepsilon(\delta^k_p)^2$, thus leading to the overall conclusion that $\liminf_{k\to+\infty} H^k_0(X^k_{\text{feas}}) = 0$ almost surely, which is implicitly assumed in the next theorem.

Theorem 4.13. *Let all the assumptions of Lemma 4.12 hold. Let $\hat{X}_{\text{feas}}$ be the almost sure limit of a convergent ε-feasible refining subsequence $\{X^{k\vee T}_{\text{feas}}\}_{k\in K}$, for which $\lim_{k\in K} H^k_0(X^{k\vee T}_{\text{feas}}) = 0$ almost surely. Then*

$$\mathbb{P}\left(\hat{X}_{\text{feas}} \in \mathcal{D}\right) = 1. \tag{27}$$

The following result which is a stochastic variant of Theorem 3.3 in [9] presents a necessary optimality condition based on the hypertangent cone definition. It states that the limit $\hat{X}_{\text{feas}}$ of an almost surely convergent ε-feasible refining subsequence $\{X^k_{\text{feas}}\}_{k\in K}$ is a hypertangent stationary point of f over the feasible region $\mathcal{D}$, with probability one.

Theorem 4.14. *Let Assumptions 1, 6 and all assumptions that were made in Theorem 4.2 and Lemma 4.12 hold. Let $\{X^{k\vee T}_{\text{feas}}\}_{k\in K}$ be an almost surely convergent ε-feasible refining subsequence, for some sequence K of random variables satisfying $\lim_K \Psi^{f,T}_k \leq 0$ and $\lim_K H^k_0(X^{k\vee T}_{\text{feas}}) = 0$ almost surely. If $\hat{x}_{\text{feas}} \in \mathcal{D}$ is a refined point for a realization $\{x^{k\vee t}_{\text{feas}}\}_{k\in\mathcal{K}}$ of $\{X^{k\vee T}_{\text{feas}}\}_{k\in K}$ for which the events $\Delta^k_p \to 0$, $\lim_K \Psi^{f,T}_k \leq 0$ and $\lim_K H^k_0(X^{k\vee T}_{\text{feas}}) = 0$ occur, and if $v \in T^H_{\mathcal{D}}(\hat{x}_{\text{feas}})$ is a refining direction for $\hat{x}_{\text{feas}}$, then $f^\circ(\hat{x}_{\text{feas}}; v) \geq 0$. In particular, this means that*

$$\mathbb{P}\left(\left\{\omega \in \Omega : \exists K(\omega) \subseteq \mathbb{N} \text{ and } \exists \hat{X}_{\text{feas}}(\omega) = \lim_{k\in K(\omega)} X^{k\vee T}_{\text{feas}}(\omega), \hat{X}_{\text{feas}}(\omega) \in \mathcal{D}, \text{ such that } \forall V(\omega) \in T^H_{\mathcal{D}}(\hat{X}_{\text{feas}}(\omega)),\ f^\circ(\hat{X}_{\text{feas}}(\omega); V(\omega)) \geq 0\right\}\right) = 1. \tag{28}$$

Next is stated a stochastic variant of a result in [9] showing that Clarke stationarity is guaranteed with probability one when the set of refining directions is dense in a nonempty hypertangent cone to $\mathcal{D}$.

Corollary 4.15. *Let all assumptions that were made in Theorem 4.14 hold. Let $\{X_{\text{feas}}^{k\vee T}\}_{k\in K}$ be the ε-feasible refining subsequence of Theorem 4.14 where K is the sequence of random variables satisfying $\lim_K \Psi_k^{f,T} \leq 0$ and $\lim_K H_0^k(X_{\text{feas}}^{k\vee T}) = 0$ almost surely. If $\hat{x}_{\text{feas}} \in \mathcal{D}$ is a refined point for a realization $\{x_{\text{feas}}^{k\vee t}\}_{k\in\mathcal{K}}$ of $\{X_{\text{feas}}^{k\vee T}\}_{k\in K}$ for which the events $\Delta_p^k \to 0$, $\lim_K \Psi_k^{f,T} \leq 0$ and $\lim_K H_0^k(X_{\text{feas}}^{k\vee T}) = 0$ occur, and if the set of refining directions for $\hat{x}_{\text{feas}}$ is dense in $T_{\mathcal{D}}^H(\hat{x}_{\text{feas}}) \neq \emptyset$, then $\hat{x}_{\text{feas}}$ is a Clarke stationary point for (1).*

5 Estimates computation and computational experiments

Section 5.1 discusses approaches for computing ε-accurate random estimates and ε-reliable bounds satisfying Assumption 3 in a simple random noise framework, and hence how corresponding deterministic estimates can be obtained using evaluations of the stochastic blackbox. These approaches strongly rely on the computation of $\alpha^{1/m}$-probabilistically ε-accurate estimates $\{C_{j,0}^k, C_{j,s}^k\}$, using techniques derived in [23]. Computational experiments comparing StoMADS-PB to MADS with PB are then presented in Section 5.2.

5.1 Computation of probabilistically accurate estimates and reliable bounds

Consider the following typical noise assumption often made in the stochastic optimization literature:

$$\begin{array}{rcl} \mathbb{E}_{\Theta_0}\left[f_{\Theta_0}(x)\right] & = & f(x) \quad \text{and} \quad \mathbb{V}_{\Theta_0}\left[f_{\Theta_0}(x)\right] \leq V_0 < +\infty \ \text{ for all } x \in \mathcal{X} \\ \mathbb{E}_{\Theta_j}\left[c_{\Theta_j}(x)\right] & = & c_j(x) \quad \text{and} \quad \mathbb{V}_{\Theta_j}\left[c_{\Theta_j}(x)\right] \leq V_j < +\infty \ \text{ for all } x \in \mathcal{X} \text{ and for all } j \in J, \end{array}$$

where $V_i > 0$ is a constant for all $i = 0, 1, \ldots, m$. Let $V = \max\{V_0, V_1, \ldots, V_m\}$.

For some fixed $j \in J$, let Θ_j^0 and Θ_j^s be two independent random variables following the same distribution as Θ_j. Let $\Theta_{j,\ell}^0$, $\ell = 1, 2, \ldots, p_j^k$ and $\Theta_{j,\ell}^s$, $\ell = 1, 2, \ldots, p_j^k$ be independent random samples of Θ_j^0 and Θ_j^s respectively, where $p_j^k \geq 1$ is an integer denoting the sample size. In order to satisfy Assumption 3-(iii), define $C_{j,0}^k$ and $C_{j,s}^k$ respectively by

$$C_{j,0}^k = \frac{1}{p_j^k} \sum_{\ell=1}^{p_j^k} c_{\Theta_{j,\ell}^0}(x^k) \quad \text{and} \quad C_{j,s}^k = \frac{1}{p_j^k} \sum_{\ell=1}^{p_j^k} c_{\Theta_{j,\ell}^s}(x^k + s^k). \tag{29}$$

Since $\mathbb{E}\left(C_{j,0}^k\right) = c_j(x^k)$ and that $\mathbb{V}\left(C_{j,0}^k\right) \leq \frac{V}{p_j^k}$ for all j, it follows from the Chebyshev inequality that

$$\mathbb{P}\left(\left|C_{j,0}^k - c_j(x^k)\right| > \varepsilon(\delta_p^k)^2\right) = \mathbb{P}\left(\left|C_{j,0}^k - \mathbb{E}\left(C_{j,0}^k\right)\right| > \varepsilon(\delta_p^k)^2\right) \leq \frac{V}{p_j^k \varepsilon^2 (\delta_p^k)^4}. \tag{30}$$

Thus, choosing p_j^k such that

$$p_j^k \geq \frac{V}{\varepsilon^2 \left(1 - \alpha^{1/2m}\right)(\delta_p^k)^4} \tag{31}$$

ensures that $\frac{V}{p_j^k \varepsilon^2 (\delta_p^k)^4} \leq 1 - \alpha^{1/2m}$. Then, combining (30) and (31) yields, for all $j \in J$,

$$\mathbb{P}\left(\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right) \geq \alpha^{1/2m} \tag{32}$$

and similarly, $\mathbb{P}\left(\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right) \geq \alpha^{1/2m}$. It follows from the independence of the random variables Θ_j^0 and Θ_j^s and both previous inequalities that

$$\mathbb{P}\left(\left\{\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\} \cap \left\{\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \alpha^{1/m}, \tag{33}$$

which means that Assumption 3-*(iii)* holds. Estimates $c_{j,0}^k = C_{j,0}^k(\omega)$ and $c_{j,s}^k = C_{j,s}^k(\omega)$, obtained by averaging p_j^k realizations of c_{Θ_j} resulting from the evaluations of the stochastic blackbox, respectively at x^k and $x^k + s^k$, are obviously $\alpha^{1/m}$-probabilistically ε-accurate.

In order to satisfy Assumption 3-*(v)*, notice that the independence of the random variables $\Theta_j, j \in J$ combined with (32) implies

$$\mathbb{P}\left(\bigcap_{j=1}^m \left\{\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) = \prod_{j=1}^m \mathbb{P}\left(\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right) \geq \alpha^{1/2} \tag{34}$$

$$\text{and similarly,} \quad \mathbb{P}\left(\bigcap_{j=1}^m \left\{\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \alpha^{1/2}. \tag{35}$$

Define the random bounds $L_0^k(x^k)$, $L_s^k(x^k + s^k)$, $U_0^k(x^k)$ and $U_s^k(x^k + s^k)$, respectively by

$$\begin{aligned} L_0^k(x^k) &= \sum_{j=1}^m \max\left\{C_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\}, \qquad U_0^k(x^k) = \sum_{j=1}^m \max\left\{C_{j,0}^k + \varepsilon(\delta_p^k)^2, 0\right\}, \\ L_s^k(x^k + s^k) &= \sum_{j=1}^m \max\left\{C_{j,s}^k - \varepsilon(\delta_p^k)^2, 0\right\} \text{ and } U_s^k(x^k + s^k) = \sum_{j=1}^m \max\left\{C_{j,s}^k + \varepsilon(\delta_p^k)^2, 0\right\}. \end{aligned}$$

Define the events E_0^k and E_s^k respectively by

$$E_0^k = \left\{L_0^k(x^k) \leq h(x^k) \leq U_0^k(x^k)\right\} \text{ and } E_s^k = \left\{L_s^k(x^k + s^k) \leq h(x^k + s^k) \leq U_s^k(x^k + s^k)\right\} \tag{36}$$

Because

$$\bigcap_{j=1}^m \left\{\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\} = \bigcap_{j=1}^m \left\{C_{j,0}^k - \varepsilon(\delta_p^k)^2 \leq c_j(x^k) \leq C_{j,0}^k + \varepsilon(\delta_p^k)^2\right\} \subseteq E_0^k \tag{37}$$

and

$$\bigcap_{j=1}^m \left\{\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\} \subseteq E_s^k, \tag{38}$$

and combining respectively (34) and (37), and (35) and (38), lead to

$$\mathbb{P}\left(E_0^k\right) \geq \mathbb{P}\left(\bigcap_{j=1}^m \left\{\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \alpha^{1/2} \tag{39}$$

$$\text{and} \quad \mathbb{P}\left(E_s^k\right) \geq \mathbb{P}\left(\bigcap_{j=1}^{m} \left\{\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \alpha^{1/2}. \tag{40}$$

It follows from the independence of the random variables $\Theta_{j,\ell}^0$ and $\Theta_{j,\ell}^s$, for all $j \in J$ and for all $\ell = 1, 2, \ldots, p_j^k$, that the events E_0^k and E_s^k are also independent. Hence, inequalities (39) and (40) imply that

$$\begin{aligned} \alpha &\leq \mathbb{P}\left(\bigcap_{j=1}^{m} \left\{\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \times \mathbb{P}\left(\bigcap_{j=1}^{m} \left\{\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \\ &\leq \mathbb{P}\left(E_0^k\right) \times \mathbb{P}\left(E_s^k\right) = \mathbb{P}\left(E_0^k \cap E_s^k\right), \end{aligned}$$

which shows that Assumption 3-*(v)* holds.

In order to show that Assumption 3-*(iv)* holds, notice that $\mathbb{E}\left(C_{j,0}^k - c_j(x^k)\right) = 0$ for all $j \in J$, which implies that for all $j \in J$,

$$\mathbb{E}\left(\left|C_{j,0}^k - c_j(x^k)\right|^2\right) = \mathbb{V}\left(C_{j,0}^k - c_j(x^k)\right) \leq \frac{V}{p_j^k} \leq \varepsilon^2\left(1 - \alpha^{1/2m}\right)(\delta_p^k)^4, \tag{41}$$

where the last inequality in (41) follows from (31). Similarly, since $\mathbb{E}\left(C_{j,s}^k - c_j(x^k + s^k)\right) = 0$ for all $j \in J$, then

$$\mathbb{E}\left(\left|C_{j,s}^k - c_j(x^k + s^k)\right|^2\right) \leq \varepsilon^2\left(1 - \alpha^{1/2m}\right)(\delta_p^k)^4, \tag{42}$$

which shows that Assumption 3-*(iv)* holds.

Finally, in order to compute estimates F_0^k and F_s^k that satisfy Assumption 3-*(i)* and *(ii)*, let Θ_0^0 and Θ_0^s be two independent random variables following the same distribution as Θ_0. Let $\Theta_{0,\ell}^0$, $\ell = 1, 2, \ldots, p_0^k$ and $\Theta_{0,\ell}^s$, $\ell = 1, 2, \ldots, p_0^k$ be independent random samples of Θ_0^0 and Θ_0^s respectively, where $p_0^k \geq 1$ denotes the sample size. Define F_0^k and F_s^k respectively by

$$F_0^k = \frac{1}{p_0^k}\sum_{\ell=1}^{p_0^k} f_{\Theta_{0,\ell}^0}(x^k) \quad \text{and} \quad F_s^k = \frac{1}{p_0^k}\sum_{\ell=1}^{p_0^k} f_{\Theta_{0,\ell}^s}(x^k + s^k). \tag{43}$$

Then $\mathbb{E}\left(F_0^k\right) = f(x^k)$, which implies that $\mathbb{V}\left(F_0^k\right) \leq \frac{V}{p_0^k}$. Thus, it is easy to notice that the proof of Assumption 3-*(i)* follows that of Assumption 3-*(iii)*. More precisely, the following inequality holds:

$$\mathbb{P}\left(\left\{\left|F_0^k - f(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\} \cap \left\{\left|F_s^k - f(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \beta, \tag{44}$$

provided that

$$p_0^k \geq \frac{V}{\varepsilon^2\left(1 - \sqrt{\beta}\right)(\delta_p^k)^4} \tag{45}$$

Estimates $f_0^k = F_0^k(\omega)$ and $f_s^k = F_s^k(\omega)$, obtained by averaging p_0^k realizations of f_{Θ_0}, respectively at x^k and $x^k + s^k$, are obviously β-probabilistically ε-accurate. The proof of Assumption 3-*(ii)* follows that of Assumption 3-*(iv)*. Specifically,

$$\mathbb{E}\left(\left|F_0^k - f(x^k)\right|^2\right) \leq \varepsilon^2(1 - \sqrt{\beta})(\delta_p^k)^4 \quad \text{and} \quad \mathbb{E}\left(\left|F_s^k - f(x^k + s^k)\right|^2\right) \leq \varepsilon^2(1 - \sqrt{\beta})(\delta_p^k)^4,$$

provided p_0^k is chosen according to (45).

5.2 Computational experiments

This section illustrates the performance and the efficiency of StoMADS-PB using noisy variants of 42 continuous constrained problems from the optimization literature. The sources and characteristics of these problems are summarized in Table 1. The number of variables ranges from $n = 2$ to $n = 20$, where every problem has at least one constraint ($m > 0$) other than bound constraints. In order to show the capability of StoMADS-PB to cope with noisy constrained problems compared to MADS with PB [9], referred to as MADS-PB, two variants of the latter algorithm are compared to several variants of StoMADS-PB. For all computational investigations of both algorithms, only the POLL step is used, i.e., no SEARCH step is involved. Based on the result of Proposition 2.13, the ε-feasible incumbent solution is always chosen as the StoMADS-PB primary frame center unless the estimates $f_0^k(x_{\text{feas}}^k)$ and $f_0^k(x_{\text{inf}}^k)$ satisfy the sufficient decrease condition in (11), in which case the choice of the infeasible incumbent solution as primary frame center is preferred. As in [9], StoMADS-PB places less effort in polling around the secondary frame center than the primary one. Specifically, as default strategy, a maximal positive basis [12] is used for the primary frame center while only two directions, with one being the negative of the first, are used for the secondary frame center. The OrthoMADS-$2n$ directions [1] are used for the POLL which is ordered by means of an opportunistic strategy [12] , i.e., trial points around the primary frame center are evaluated first. Then, all the points around a given frame center are sorted relatively to the successful direction from the last h-Dominating iteration in StoMADS-PB, while in MADS-PB, they are sorted relatively to the last successful direction both in the noisy objective and constraint violation functions. MADS-PB and all the proposed variants of StoMADS-PB are implemented in MATLAB.

The stochastic variants of the 42 deterministic constrained optimization problems are solved using three different infeasible initial points for a total of 126 problem instances. Inspired by [11], the stochastic variants are constructed by additively perturbing the objective f by a random variable Θ_0 and by additively perturbing each constraint $c_j, j = 1, 2, \ldots, m$ by a random variable Θ_j. That is,

$$f_{\Theta_0}(x) = f(x) + \Theta_0 \quad \text{and} \quad c_{\Theta_j}(x) = c_j(x) + \Theta_j, \text{ for all } j \in J, \tag{46}$$

where Θ_0 is uniformly generated in the interval $I(\sigma, x^0, f) = [-\sigma\,|f(x^0) - f^*|\,, \sigma\,|f(x^0) - f^*|]$ and Θ_j is uniformly generated in $I(\sigma, x^0, c_j) = [-\sigma\,|c_j(x^0)|\,, \sigma\,|c_j(x^0)|]$. The scalar $\sigma > 0$ is used to define different noise levels, x^0 denotes an initial point and f^* is the best known feasible minimum value of f. Thus, the *test set* used during the experiments is representative of real-world applications, since consisting of a collection of more than 20 stochastically noisy *test problems* (see e.g. [12], appendix A.1.) The bounds of $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$ are respectively expressed in terms of $|f(x^0) - f^*|$ and $|c_j(x^0)|$ in order to take into account the effort of a given algorithm when reducing the value of the objective function from $f(x^0)$ to f^*, and when reducing the values of the constraints from $c_j(x^0)$ to zero (whenever $c_j(x^0) > 0$), for a given problem. The random variables $\Theta_0, \Theta_1, \ldots, \Theta_m$ are independent. For the remainder of the study, the process which returns the vector $[f_{\Theta_0}(x), c_{\Theta_1}(x), c_{\Theta_2}(x), \ldots, c_{\Theta_m}(x)]$ when provided the input x will be referred to as the noisy blackbox.

The MADS-PB algorithm [9], of which StoMADS-PB is a stochastic variant, and to which the latter is compared is an iterative direct-search method originally developed for deterministic constrained blackbox optimization. In MADS-PB, feasibility is sought by progressively decreasing a threshold imposed on a constraint violation function into which all the constraint violations are aggregated. Any trial point with a constraint violation value greater than that threshold is rejected. A full description

of MADS-PB iterations and its behavior can also be found in [12].

The relative performance and efficiency of algorithms are assessed by performance profiles [31, 46] and data profiles [46], which require the definition of a convergence test for a given problem instance. For each of the 126 problem instances (defined by the 42 functions f, minimized using three different initial points), denote by x^N the best feasible point found after N evaluations of the noisy blackbox and let x^* be the best feasible point of f obtained by the six compared algorithms (described later) on the three stochastic problem instances involving f. This means that for a given objective function f, the value of x^* is the best among eighteen, which can be considered significant and representative in an expensive-to-evaluate BBO framework. The convergence test from [14] used for the experiments is defined as

$$f(x^N) \leq f(x^*) + \tau(\bar{f}_{feas} - f(x^*)), \tag{47}$$

where $\tau \in [0, 1]$ is the convergence tolerance and $\bar{f}_{feas}$ is a reference value obtained by taking the average of the available first feasible f values over all instances of a given problem over all algorithms. If no feasible point is found, then the convergence test fails. Otherwise, a problem is said to be successfully solved within the tolerance τ if (47) holds. As highlighted in [14], $\bar{f}_{feas} = f(x^0)$ for unconstrained problems, where x^0 denotes the initial point.

Inspired by [31], the number of noisy objective function evaluations is chosen as *performance measure*. The horizontal axis of the performance profiles shows a ratio of the performance on a given problem by a given algorithm to the best performance by any algorithm on that problem, while the fraction of problems solved within the convergence tolerance τ is shown on the vertical axis. On the horizontal axis of the data profiles is shown the number of function calls to the noisy blackbox divided by $(n+1)$[1] while the vertical axis shows the proportion of problems solved by all instances of a given algorithm within a tolerance τ. As emphasized in [12], performance profiles show information concerning speed of convergence (i.e., the quality of a given algorithm's output in terms of the number of objective function evaluations) and robustness (i.e., the fraction of problems solved) in a compact graphical format, while data profiles also examine the robustness and efficiency from a different perspective.

Recall that in StoMADS-PB, according to Section 5.1, specifically (29) and (43), the noisy blackbox needs to be evaluated many times at a given point in order to compute function estimates unlike the MADS-PB method, in which it is evaluated only once at each point. Thus, the limited budget of $1000(n+1)$ noisy blackbox evaluations allocated to all the algorithms during the computational experiments should not be exhausted too quickly by doing replications at a given current point when computing estimates for StoMADS-PB. However, given that such estimates are required to be sufficiently accurate in order for the solutions to be satisfactory, a procedure inspired by [11] aimed at improving the estimates accuracy by making use of available samples at a given current point is proposed. The following computation scheme is described only for $f_0^k(x^k)$ but is the same for $f_s^k(x^k + s^k)$, $c_{j,0}^k(x^k)$ and $c_{j,s}^k(x^k + s^k)$, for all $j \in J$. During the optimization, all trial points x^k used by StoMADS-PB and all corresponding values $f_{\Theta_0}(x^k)$ are stored in a cache. When constructing an estimate of $f(x^k)$ at the iteration $k \geq 1$, denote by $a^k(x^k)$[2] the number of sample values of $f_{\Theta_0}(x^k)$ available in the cache from previous blackbox evaluations until iteration $k-1$. Since all the values of the noisy objective function f_{Θ_0} are always computed independently of each other, the aforementioned sample

[1] $n+1$ is the number of evaluations required to construct a linear interpolant or a simplex gradient [12] in $\mathbb{R}^n$ [14, 46].

[2] It is implicitly assumed without any loss of generality that $a^k(x^k) \geq 1$.

values can be considered as independent realizations $f_{\theta_{0,1}}(x^k), f_{\theta_{0,2}}(x^k), \ldots, f_{\theta_{0,a^k(x^k)}}(x^k)$ of $f_{\Theta_0}(x^k)$, where for all $\ell = 1, 2, \ldots, a^k(x^k)$, $\theta_{0,\ell}$ is a realization of the random variable $\Theta_{0,\ell}$ following the same distribution as Θ_0. Now let $n^k \geq 1$ be the number of blackbox evaluations at x^k and consider the independent realizations $\theta_{0,a^k(x^k)+1}, \theta_{0,a^k(x^k)+2}, \ldots, \theta_{0,a^k(x^k)+n^k}$ of Θ_0. Then using (43), an estimate $f_0^k(x^k)$ of $f(x^k)$ is computed according to

$$f_0^k(x^k) = \frac{1}{p^k} \sum_{\ell=1}^{p^k} f_{\theta_{0,\ell}}(x^k), \tag{48}$$

where $p^k = n^k + a^k(x^k)$ is the sample size. Note that this computation procedure is very efficient in practice as highlighted in [11] even though it is inherently biased.

The same values are used to initialize most of the common parameters to StoMADS-PB and MADS-PB. Specifically, the mesh refining parameter $\tau = 1/2$, the frame center trigger $\rho = 0.1$ and $\delta_m^0 = \delta_p^0 = 1$ are common to both methods. However, in MADS-PB, the initial barrier threshold is set equal its default value, i.e., $h_{\max}^0 = +\infty$ [9] while in StoMADS-PB it equals $u_0^0(x_{\text{inf}}^0)$, with $u_0^k(x^k)$ defined in (4) for all $k \in \mathbb{N}$. The default values of Algorithm 2 parameters $\gamma > 2$ and $\varepsilon > 0$[3] are borrowed from [11], that is, $\gamma = 17$ and $\varepsilon = 0.01$.

Table 1: Description of the set of 42 constrained problems.

No	Name	Source	n	m	Bnds	No	Name	Source	n	m	Bnds
1	ANGUN	[54]	2	1	Yes	22	MAD1	[43]	2	1	No
2	BARNES	[51]	2	3	Yes	23	MAD2	[43]	2	1	No
3	BERTSIMAS	[19]	2	2	No	24	MAD6	[43]	7	7	Yes
4	CHENWANG_F2	[24]	8	6	Yes	25	MEZMONTES	[44]	2	2	Yes
5	CHENWANG_F3	[24]	10	8	Yes	26	NEW-BRANIN	[54]	2	1	Yes
6	CONSTR-BRANIN	[54]	2	1	Yes	27	OPTENG-BENCH4	[37]	2	1	Yes
7	CRESCENT	[9]	10	2	No	28	OPTENG-BENCH5	[37]	2	3	Yes
8	DEMBO5	[43]	8	3	Yes	29	OPTENG-RBF	[37]	3	4	Yes
9	DISK	[9]	10	1	No	30	PENTAGON	[43]	6	15	No
10	G23	[10]	3	2	Yes	31	PRESSURE-VESSEL	[44]	4	4	Yes
11	G210	[10]	10	2	Yes	32	SASENA	[54]	2	1	Yes
12	G220	[10]	20	2	Yes	33	SNAKE	[9]	2	2	No
13	GOMEZ	[54]	2	1	Yes	34	SPEED-REDUCER	[44]	7	11	Yes
14	HS15	[35]	2	2	Yes	35	SPRING	[51]	3	4	Yes
15	HS19	[35]	2	2	Yes	36	TAOWANG_F1	[53]	2	2	Yes
16	HS22	[35]	2	2	No	37	TAOWANG_F2	[53]	7	4	Yes
17	HS23	[35]	2	5	Yes	38	WELDED-BEAM	[44]	4	7	Yes
18	HS29	[35]	3	1	No	39	WONG2	[43]	10	3	No
19	HS43	[35]	4	3	No	40	ZHAOWANG_F5	[55]	13	9	Yes
20	HS108	[35]	9	13	Yes	41	ZILONG_G4	[54]	5	1	Yes
21	HS114	[35]	10	5	Yes	42	ZILONG_G24	[54]	2	1	Yes

[3]The use of ε_f instead of ε is favored in [11].

Table 2: Percentage of problems solved for each noise level σ within a convergence tolerance τ.

	$\tau = 10^{-1}$			$\tau = 10^{-3}$		
Algorithm	$\sigma = 0.01$	$\sigma = 0.03$	$\sigma = 0.05$	$\sigma = 0.01$	$\sigma = 0.03$	$\sigma = 0.05$
StoMADS-PB $n^k = 1$	73.81%	76.98%	73.81%	37.30%	36.51%	37.30%
StoMADS-PB $n^k = 2$	73.81%	75.40%	76.19%	43.65%	42.86%	44.44%
StoMADS-PB $n^k = 3$	76.19%	62.70%	66.67%	45.24%	38.89%	33.33%
StoMADS-PB $n^k = 4$	75.40%	74.60%	73.81%	45.24%	45.24%	38.89%
ℓ_1-MADS-PB	61.90%	59.52%	49.21%	29.37%	33.33%	26.19%
ℓ_2-MADS-PB	65.87%	59.52%	51.59%	37.30%	29.37%	24.60%

Four variants of StoMADS-PB corresponding to $n^k \in \{1, 2, 3, 4\}$ are compared to two variants of MADS-PB : a variant referred to as ℓ_1-MADS-PB using an ℓ_1-norm (as in StoMADS-PB) for the definition of the barrier function h, and a second one, ℓ_2-MADS-PB, which uses the Euclidean norm as in [9]. The data and performance profiles used for the comparisons are depicted in Figures 2, 4 and 6 and Figures 3, 5 and 7. Three levels of noise are used during the experiments, which correspond to $\sigma = 0.01$, $\sigma = 0.03$ and $\sigma = 0.05$. For each of the three levels of noise, since each of the six compared algorithms were applied to the stochastic variants of the 126 problem instances, a total of $3 \times 6 \times 126 = 2268$ algorithm runs were necessary to obtain the data and performance profiles. The execution of a StoMADS-PB variant is random by design but the reported results presented for one run of each variant are representative. For a given algorithm, the percentage of problems solved after $1000(n + 1)$ noisy blackbox evaluations for each noise level within a convergence tolerance τ are reported in Table 2.

The data and performance profiles show that when given sufficient budget, StoMADS-PB generally outperforms MADS-PB. They also show that MADS-PB is very efficient for moderate values of the noise level σ, but its performance degrades quickly as σ increases. As expected, ℓ_2-MADS-PB outperforms ℓ_1-MADS-PB since the latter should introduce some nondifferentiability in the deterministic minimization problem as discussed in [12]. Moreover as in [11], varying the value of the convergence tolerance τ in the data profiles does not significantly alter the conclusions drawn from the performance profiles. Indeed, as expected, it can be easily observed from Table 2 that the higher the tolerance parameter τ, the larger the percentage of problems solved by all algorithms for a fixed noise level σ. Notice that while for a given τ, the fraction of problems solved by each variant of MADS-PB has a decreasing tendency when the noise level increases from $\sigma = 0.01$ to $\sigma = 0.05$, this seems not to be the case for StoMADS-PB variants. Before giving an insight as to why, recall that in the present constrained framework, the success or failure of the convergence test (47) depends not only on the values of the objective function f but also on whether a feasible point is found or not, unlike the framework of [11] where no constraints are involved. In fact, as highlighted in [11] from which the scheme (48) was inspired, even though the robustness and efficiency of each StoMADS-PB variant depends on the number n^k of noisy blackbox evaluations which is constant for all k, the quality of the solutions is influenced by the sample size $p^k = n^k + a^k(x^k)$, which is not constant. On one hand, this is the reason why for $n^k = 1$, StoMADS-PB does not have the same behavior as ℓ_1-MADS-PB. On the other hand, (48) naturally favors StoMADS-PB by improving the accuracy of the estimates of the constraint functions, thus allowing it to find a higher amount of feasible solutions

compared to MADS-PB and consequently possibly solve a larger fraction of problems when the noise level increases for a fixed tolerance parameter τ.

Based on Table 2, observe that for a given convergence tolerance τ, varying σ seems not to have a significant influence on the fraction of problems solved by the StoMADS-PB variants corresponding to $n^k = 1, n^k = 2$ and $n^k = 4$. Moreover, even though for the lowest noise level studied, $\sigma = 0.01$, StoMADS-PB with $n^k = 3$ solved the most problems, the corresponding percentage is not significantly larger than those of StoMADS-PB with $n^k = 2$ and $n^k = 4$. For all these reasons, the variant with $n^k = 2$ seems preferable for constrained stochastic blackbox optimization problems with limited budget while the variant $n^k = 4$ should be preferred for stochastic blackbox optimization problems with higher evaluations budgets. Finally, recall that ℓ_2-MADS-PB outperforms ℓ_1-MADS-PB in the present stochastic framework. Thus, given that even in a deterministic BBO context, squared violations were shown in [7] to perform better (thus motivating the use of an ℓ_2 barrier function h in [9]), one could expect a StoMADS-PB variant using an ℓ_2 barrier function to outperform the one analyzed in the present manuscript even though the convergence of this ℓ_2 variant is not yet demonstrated.

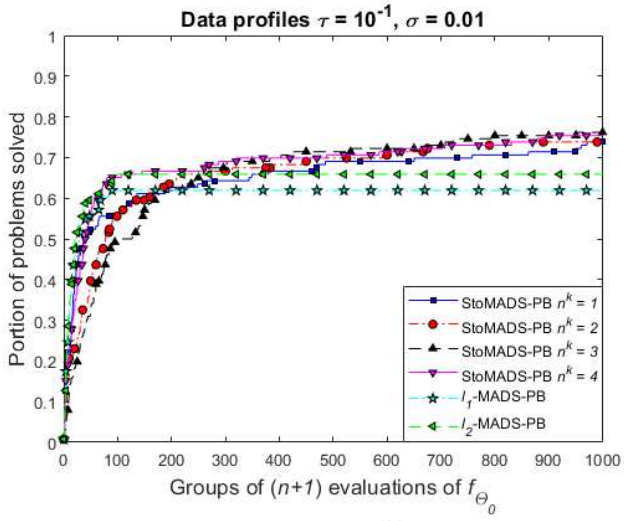


Figure 2: Data profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.01$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.

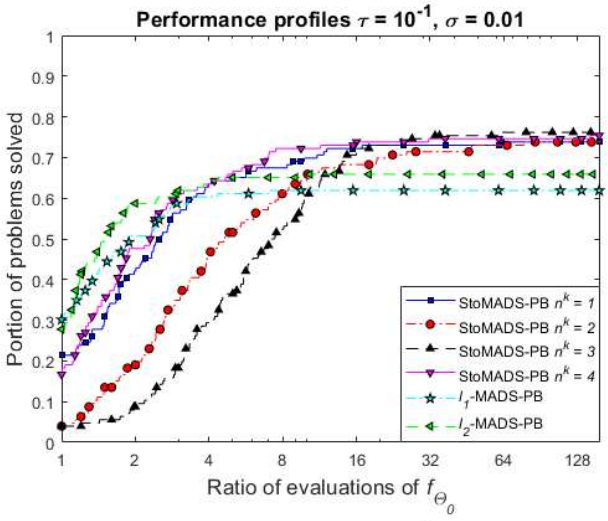
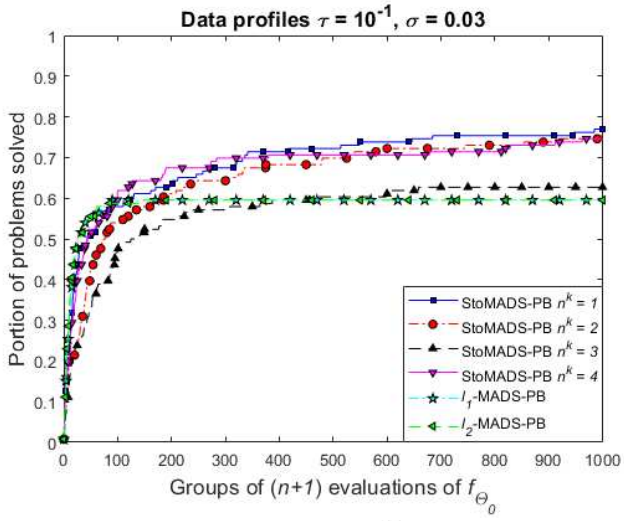


Figure 3: Performance profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.01$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.



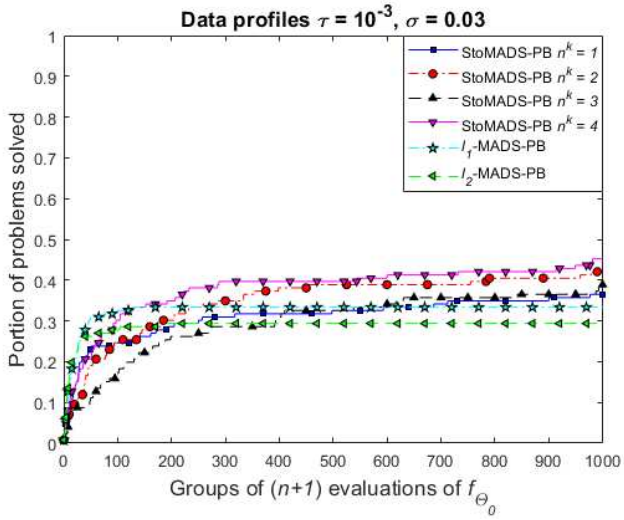
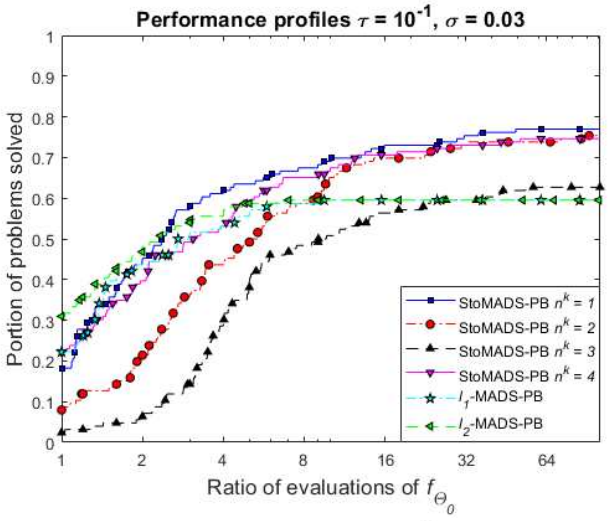


Figure 4: Data profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.03$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.



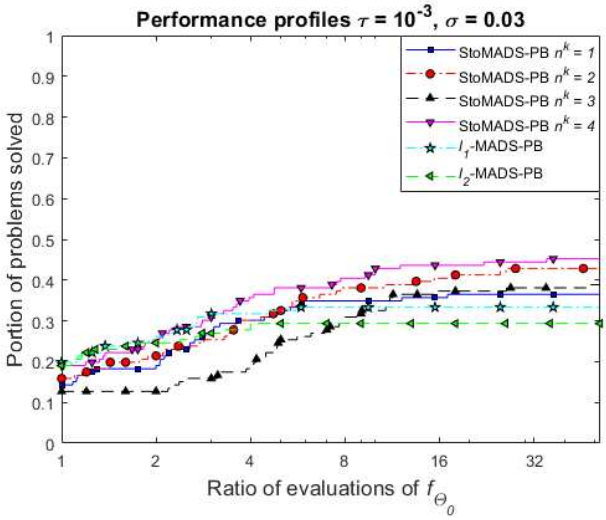


Figure 5: Performance profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.03$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.

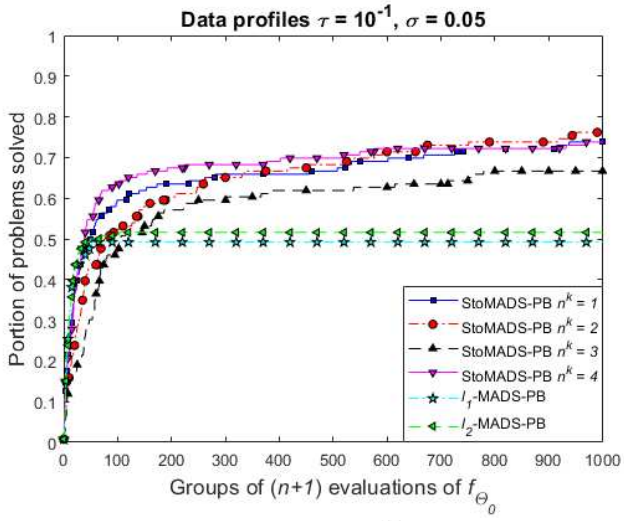


Figure 6: Data profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.05$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.

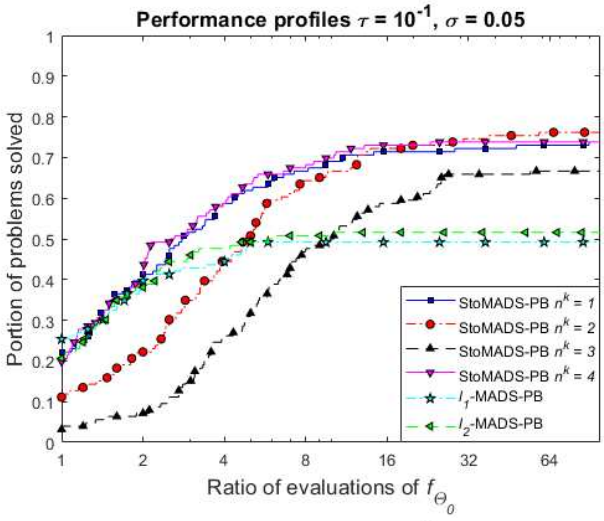


Figure 7: Performance profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.05$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.

Concluding remarks

The StoMADS-PB algorithm introduced in the present work is developed for constrained stochastic blackbox optimization. The proposed method, which uses an algorithmic framework similar to that of MADS, considers the optimization of objective and constraint functions whose values can only be accessed through a stochastically noisy blackbox. It treats constraints using a progressive barrier approach, by aggregating their violations into a single function. It does not use any model or approximate gradient information to find descent directions or improve feasibility unlike prior works. Instead, StoMADS-PB uses function estimates and introduces probabilistic bounds on which sufficient decrease conditions are imposed. By requiring the accuracy of such estimates and bounds to hold with sufficiently high, but fixed, probabilities, convergence results for StoMADS-PB are derived, most of which are stochastic variants of those of MADS.

Computational experiments conducted on several variants of StoMADS-PB on a collection of constrained stochastically noisy problems showed the proposed method eventually outperforms MADS, and show some of its variants to be almost robust to random noise despite the use of very inaccurate estimates.

This research is, to the best of our knowledge, the first to propose a stochastic directional direct-search algorithm for BBO, developed to cope with both a stochastically noisy objective and constraints.

Future research could focus on improving the proposed method to handle large-scale machine learning problems, making use, for example, of parallel space decomposition.

Acknowledgments

The authors are grateful to Charles Audet from Polytechnique Montréal for valuable discussions and constructive suggestions. This work is supported by the NSERC CRD RDCPJ 490744-15 grant and by an InnovÉÉ grant, both in collaboration with Hydro-Québec and Rio Tinto, and by a FRQNT fellowship.

Appendix

This appendix presents the proofs of a series of results stated in Section 4.

Proof of Theorem 4.2

Proof. This theorem is proved using ideas from [11, 21, 23, 33, 40, 48] and conditioning on the disjoint events $\{T = +\infty\}$ and $\{T < +\infty\}$ that are almost sure due to Assumption 4. The proof considers two different parts. Part 1 considers two separate cases conditioned on the event $\{T = +\infty\}$ (i.e., no ε-feasible point is found by Algorithm 2): "good bounds" and "bad bounds", each of which is separated into whether an iteration is h-Dominating, Improving or Unsuccessful. Part 2 considers three separates cases conditioned on the event $\{T < +\infty\}$: "good estimates and good bounds", "bad estimates and good bounds" and "bad bounds", each of which is separated into whether an iteration is f-Dominating, h-Dominating, Improving or Unsuccessful.

In order to show (21), the goal of Part 1 is to show that there exists a constant $\eta > 0$ such that conditioned on the almost sure event $\{T = +\infty\}$, the following holds for all $k \in \mathbb{N}$:

$$\mathbb{E}\left(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2, \tag{49}$$

where Φ_k is the random function defined by

$$\Phi_k := \frac{\nu}{m\varepsilon} h(X_{\text{inf}}^k) + (1-\nu)(\Delta_p^k)^2, \quad \text{for all } k \in \mathbb{N}. \tag{50}$$

Indeed, assume that (49) holds. Since $\Phi_k > 0$ for all $k \in \mathbb{N}$, then summing (49) over $k \in \mathbb{N}$ and taking expectations on both sides lead to

$$\mathbb{E}\left[\sum_{k=0}^{+\infty} (\Delta_p^k)^2\right] \leq \frac{\mathbb{E}\left(\Phi_0\right)}{\eta} = \frac{\Phi_0}{\eta}, \tag{51}$$

That is, (21) holds. Then, Part 2 aims to show that for the same previous constant η, conditioned on the almost sure event $\{T < +\infty\}$ and making use of the following random function

$$\Phi_k^T := \frac{\nu}{\varepsilon}(f(X_{\text{feas}}^{k\vee T}) - \kappa_{\min}^f) + \frac{\nu}{m\varepsilon} h(X_{\text{inf}}^k) + (1-\nu)(\Delta_p^k)^2, \quad \text{for all } k \in \mathbb{N}, \tag{52}$$

where $k \vee T := \max\{k, T\}$, the following holds for all $k \in \mathbb{N}$:

$$\mathbb{E}\left(\Phi_{k+1}^T - \Phi_k^T | \mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2. \tag{53}$$

Indeed, assume that (53) holds. Since $\Phi_k^T > 0$ for all $k \geq 0$, then summing (53) over $k \in \mathbb{N}$ and taking expectations on both sides, yields

$$\begin{aligned} \mathbb{E}\left[\sum_{k=0}^{+\infty} (\Delta_p^k)^2\right] &\leq \frac{\mathbb{E}\left(\Phi_0^T\right)}{\eta} = \frac{1}{\eta}\left[\frac{\nu}{\varepsilon}\left(\mathbb{E}\left[f(X_{\text{feas}}^T)\right] - \kappa_{\min}^f\right) + \frac{\nu}{m\varepsilon} h(x_{\text{inf}}^0) + (1-\nu)(\delta_p^0)^2\right] \\ &\leq \frac{1}{\eta}\left[\frac{\nu}{\varepsilon}\left(\kappa_{\max}^f - \kappa_{\min}^f\right) + \frac{\nu}{m\varepsilon} h(x_{\text{inf}}^0) + (1-\nu)(\delta_p^0)^2\right] =: \mu, \end{aligned} \tag{54}$$

where the last inequality in (54) follows from the inequality $f(X_{\text{feas}}^k) \leq \kappa_{\max}^f$ for all $k \geq 0$, due to Proposition 3.5, and the fact that T is finite almost surely.

The remainder of the proof is devoted to showing that (49) and (53) hold. The following events are introduced for the sake of clarity in the analysis.

$\mathcal{D}_f := \{\text{The iteration is } f\text{-Dominating}\}$, $\mathcal{D}_h := \{\text{The iteration is } h\text{-Dominating}\}$,
$\mathcal{I} := \{\text{The iteration is Improving}\}$, $\mathcal{U} := \{\text{The iteration is Unsuccessful}\}$.

Part 1 ($T = +\infty$ almost surely). The random function Φ_k defined in (50) will be shown to satisfy (49) with $\eta = \frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2)$, regardless of the change in the objective function f on the ε-infeasible incumbents encountered by Algorithm 2. Moreover, since T is infinite almost surely, then no iteration of Algorithm 2 can be f-Dominating. Two separate cases are distinguished and all that follows is conditioned on the almost sure event $\{T = +\infty\}$.

Case 1 (Good bounds, $\mathbb{1}_{I_k} = 1$). No matter the type of iteration which occurs, the random function Φ_k will be shown to decrease and the smallest decrease is shown to happen on Unsuccessful iterations, thus yielding

$$\mathbb{E}\left[\mathbb{1}_{I_k}(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^{C\cdot F}\right] \leq -\alpha(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{55}$$

(i) The iteration is h-Dominating ($\mathbb{1}_{\mathcal{D}_h} = 1$). The iteration is h-Dominating and the bounds are good, so a decrease occurs in h according to (6), i.e.,

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}(h(X_{\text{inf}}^{k+1}) - h(X_{\text{inf}}^{k})) \leq -\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}\nu(\gamma-2)(\Delta_p^k)^2. \tag{56}$$

The frame size parameter is updated according to $\Delta_p^{k+1} = \min\{\tau^{-1}\Delta_p^k, \delta_{\max}\}$, which implies that

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)[(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] \leq \mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \tag{57}$$

Then, by choosing ν according to (19), the right-hand side of (56) dominates that of (57). That is,

$$-\nu(\gamma-2)(\Delta_p^k)^2 + (1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2 \leq -\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{58}$$

Combining (56), (57) and (58) leads to

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{59}$$

(ii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}} = 1$). The iteration is Improving and the bounds are good, so again, a decrease occurs in h according to (6). Moreover, Δ_p^k is updated as in h-Dominating iterations. Thus, the change in Φ_k follows from (59) by replacing $\mathbb{1}_{\mathcal{D}_h}$ by $\mathbb{1}_{\mathcal{I}}$. Specifically,

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{I}}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{60}$$

(iii) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}} = 1$). The value of h is unchanged while the frame size parameter is decreased. Consequently,

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1} - \Phi_k) = -\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{U}}(1-\nu)(1-\tau^2)(\Delta_p^k)^2 \tag{61}$$

Because ν satisfies (19) and because $1-\tau^2 < \tau^{-2}-1$, Unsuccessful iterations, vis a vis (61), provide the worst case decrease when compared to (59) and (60). That is,

$$-\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2 \leq -(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{62}$$

Thus, it follows from (59), (60), (61) and (62) that the change in Φ_k is bounded like

$$\mathbb{1}_{I_k}(\Phi_{k+1} - \Phi_k) = \mathbb{1}_{I_k}(\mathbb{1}_{\mathcal{D}_h} + \mathbb{1}_{\mathcal{I}} + \mathbb{1}_{\mathcal{U}})(\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{I_k}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{63}$$

Since Assumption 3 holds, taking conditional expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of the inequality in (63) leads to (55).

Case 2 (Bad bounds, $\mathbb{1}_{\bar{I}_k} = 1$). Since the bounds are bad, Algorithm 2 can accept a step which leads to an increase in h and Δ_p^k, and hence in Φ_k. Such an increase in Φ_k is controlled by making use of (15). Then, the probability of $\bar{I}_k$ is chosen to be sufficiently small so that Φ_k can be reduced sufficiently in expectation. More precisely, the next result will be proved

$$\mathbb{E}\left[\mathbb{1}_{\bar{I}_k}(\Phi_{k+1} - \Phi_k)|\mathcal{F}_{k-1}^{C\cdot F}\right] \leq 2\nu(1-\alpha)^{1/2}(\Delta_p^k)^2. \tag{64}$$

(i) The iteration is h-Dominating ($\mathbb{1}_{\mathcal{D}_h} = 1$). The change in h is bounded like

$$\begin{aligned}
&\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}(h(X_{\text{inf}}^{k+1}) - h(X_{\text{inf}}^{k})) \\
&\qquad \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}\left[(H_s^k - H_0^k) + \left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right] \\
&\qquad \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\nu\left[-\gamma(\Delta_p^k)^2 + \frac{1}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right)\right], \quad (65)
\end{aligned}$$

where (65) follows from $H_s^k - H_0^k \leq -\gamma m\varepsilon(\Delta_p^k)^2$ which is satisfied in every h-Dominating iteration. Moreover, the change in Δ_p^k can be obtained simply by replacing $\mathbb{1}_{I_k}$ by $\mathbb{1}_{\bar{I}_k}$ in (57). That is,

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)[(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \quad (66)$$

Because ν satisfies (19), $-\nu\gamma(\Delta_p^k)^2 + (1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2 \leq 0$. Combining (65) and (66),

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right). \quad (67)$$

(ii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}} = 1$). Δ_p^k is updated as in h-Dominating iterations. The increase in h is bounded as in (65). Thus, the bound on the change in Φ_k can be obtained by replacing $\mathbb{1}_{\mathcal{D}_h}$ by $\mathbb{1}_{\mathcal{I}}$ in (67). That is,

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right). \quad (68)$$

(iii) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}} = 1$). The value of h is unchanged and Δ_p^k is decreased. Thus, the change in Φ_k follows from (61) by replacing $\mathbb{1}_{I_k}$ by $\mathbb{1}_{\bar{I}_k}$ and is trivially bounded like

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right). \quad (69)$$

It follows from (67), (68), (69) and the inequality $\mathbb{1}_{\bar{I}_k} \leq 1$, that

$$\mathbb{1}_{\bar{I}_k}(\Phi_{k+1} - \Phi_k) \leq \frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right). \quad (70)$$

Taking conditional expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (70) and using the inequalities (15) of Lemma 3.7, leads to (64).

Combining (55) and (64) yields,

$$\begin{aligned}
\mathbb{E}\left(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{C\cdot F}\right) &= \mathbb{E}\left[(\mathbb{1}_{I_k} + \mathbb{1}_{\bar{I}_k})(\Phi_{k+1} - \Phi_k)|\mathcal{F}_{k-1}^{C\cdot F}\right] \\
&\leq \left[-\alpha(1-\nu)(1-\tau^2) + 2\nu(1-\alpha)^{1/2}\right](\Delta_p^k)^2. \quad (71)
\end{aligned}$$

Choosing α according to (20) implies that $\alpha \geq \dfrac{4\nu(1-\alpha)^{1/2}}{(1-\nu)(1-\tau^2)}$, which ensures

$$-\alpha(1-\nu)(1-\tau^2) + 2\nu(1-\alpha)^{1/2} \leq -\frac{1}{2}\alpha(1-\nu)(1-\tau^2) \leq -\frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2). \quad (72)$$

Thus, (49) follows from (71) and (72) with $\eta = \frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2)$.

Part 2 ($T < +\infty$ almost surely). In order to show that the random function Φ_k^T defined by

$$\Phi_k^T = \frac{\nu}{\varepsilon}(f(X_{\text{feas}}^{k\vee T}) - \kappa_{\min}^f) + \frac{\nu}{m\varepsilon}h(X_{\text{inf}}^k) + (1-\nu)(\Delta_p^k)^2$$

satisfies (53) with the same constant η derived in Part 1, notice that whenever the event $\{T > k\}$ occurs, then $f(X_{\text{feas}}^{(k+1)\vee T}) - f(X_{\text{feas}}^{k\vee T}) = 0$ since $\max\{k, T\} := k \vee T = (k+1) \vee T = T$. Thus, on the event $\{T > k\}$, the random function Φ_k used in Part 1 has the same increment as Φ_k^T. Specifically,

$$\mathbb{1}_{\{T<+\infty\}}\mathbb{1}_{\{T>k\}}(\Phi_{k+1}^T - \Phi_k^T) = \mathbb{1}_{\{T<+\infty\}}\mathbb{1}_{\{T>k\}}(\Phi_{k+1} - \Phi_k).$$

Moreover, it follows from the definition of the stopping time T that no iteration can be f-Dominating when the event $\{T > k\}$ occurs. Consequently, it easily follows from the analysis in Part 1 and the fact that the random variable $\mathbb{1}_{\{T>k\}}$ is $\mathcal{F}_{k-1}^{C\cdot F}$-measurable that,

$$\mathbb{1}_{\{T>k\}}\mathbb{E}\left(\Phi_{k+1}^T - \Phi_k^T | \mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2\mathbb{1}_{\{T>k\}}. \tag{73}$$

The remainder of the proof is devoted to showing that

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left(\Phi_{k+1}^T - \Phi_k^T | \mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}, \tag{74}$$

since combining (73) and (74) leads to (53), which is the overall goal. In all that follows, it is assumed that the event $\{T \leq k\}$ occurs.

Case 1 (Good estimates and good bounds, $\mathbb{1}_{I_k}\mathbb{1}_{J_k} = 1$). Regardless of the iteration type, the smallest decrease in Φ_k^T will be shown to happen on Unsuccessful iterations, and it will be shown that

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\mathbb{1}_{I_k}\mathbb{1}_{J_k}(\Phi_{k+1}^T - \Phi_k^T)|\mathcal{F}_{k-1}^{C\cdot F}\right] \leq -\alpha\beta(1-\nu)(1-\tau^2)(\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}. \tag{75}$$

(i) The iteration is f-Dominating ($\mathbb{1}_{\mathcal{D}_f} = 1$). The iteration is f-Dominating and the estimates are good, so a decrease occurs in f according to (10). That is,

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}(f(X_{\text{feas}}^{(k+1)\vee T}) - f(X_{\text{feas}}^{k\vee T})) \\ \leq \quad -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}\nu(\gamma-2)(\Delta_p^k)^2.\end{aligned} \tag{76}$$

Since the ε-infeasible incumbent is not updated, The value of h is unchanged. The frame size parameter is updated according to $\Delta_p^{k+1} = \min\{\tau^{-1}\Delta_p^k, \delta_{\max}\}$, thus implying that

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)[(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \tag{77}$$

Because ν satisfies (19), (58) holds, which implies that the right-hand side of (76) dominates that of (77), leading to the inequality

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}(\Phi_{k+1}^T - \Phi_k^T) \leq -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{78}$$

(ii) The iteration is h-Dominating ($\mathbb{1}_{\mathcal{D}_h} = 1$). The value of f is unchanged since X_{feas}^k is not updated. Thus, the bound on the change in Φ_k^T follows from multiplying both sides of (59) by $\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{J_k}$, and replacing Φ_k by Φ_k^T. That is,

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1}^T - \Phi_k^T) \leq -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_h}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{79}$$

(iii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}} = 1$). Again, the value of f is unchanged. Thus, the bound on the change in Φ_k^T follows from multiplying both sides of (60) by $\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{J_k}$, and replacing Φ_k by Φ_k^T. That is,

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1}^T - \Phi_k^T) \leq -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{I}}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{80}$$

(iv) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}} = 1$). The value of f and h is unchanged since no incumbent is updated, while Δ_p^k is decreased. Consequently, the bound on the change in Φ_k^T follows from multiplying both sides of (61) by $\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{J_k}$, and replacing Φ_k by Φ_k^T. That is,

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1}^T - \Phi_k^T) = -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{U}}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{81}$$

Combining (78), (79), (80), (81) and (62) yields

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}(\Phi_{k+1}^T - \Phi_k^T) \leq -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{82}$$

The following holds under Assumption 3: $\mathbb{E}\left(\mathbb{1}_{I_k}\mathbb{1}_{J_k}|\mathcal{F}_{k-1}^{C\cdot F}\right) \geq \alpha\beta$. Then, taking expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (82) and using the $\mathcal{F}_{k-1}^{C\cdot F}$-measurability of the random variables $\mathbb{1}_{\{T\leq k\}}$ and Δ_p^k leads to (75).

Case 2 (Bad estimates and good bounds, $\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k} = 1$). An increase in the difference of Φ_k^T may occur since good bounds might not provide enough decrease to cancel the increase which occurs in f whenever Algorithm 2 wrongly accepts an incumbent due to bad estimates. Specifically, the f-Dominating case dominates the worst-case increase in the change of Φ_k^T, leading to

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}^T - \Phi_k^T)|\mathcal{F}_{k-1}^{C\cdot F}\right] \leq 2\nu(1-\beta)^{1/2}(\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}. \tag{83}$$

(i) The iteration is f-Dominating ($\mathbb{1}_{\mathcal{D}_f} = 1$). Whenever bad estimates occur and the iteration is f-Dominating, the change in f is bounded like

$$\begin{aligned}
&\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}(f(X_{\text{feas}}^{(k+1)\vee T}) - f(X_{\text{feas}}^{k\vee T})) \\
&\quad\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}\left[(F_s^k - F_0^k) + \left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right] \\
&\quad\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\nu\left[-\gamma(\Delta_p^k)^2 + \frac{1}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right)\right]
\end{aligned} \tag{84}$$

where the last inequality in (84) follows from $F_s^k - F_0^k \leq -\gamma\varepsilon(\Delta_p^k)^2$ which is satisfied for every f-Dominating iteration. While the value of h remains unchanged since X_{inf}^k is not updated, the change in Δ_p^k follows (77) by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$. That is,

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)[(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \tag{85}$$

Then, (84), (85), (19) and the inequality $-\nu\gamma(\Delta_p^k)^2 + (1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2 \leq 0$ yield

$$\begin{aligned}
&\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}(\Phi_{k+1}^T - \Phi_k^T) \\
&\qquad\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right).
\end{aligned} \tag{86}$$

(ii) The iteration is h-Dominating ($\mathbb{1}_{\mathcal{D}_h} = 1$). The bound on the change in Φ_k^T, which can be obtained by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$ in (79), is trivially bounded like

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1}^T - \Phi_k^T) \\ \leq \quad \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right).\end{aligned} \tag{87}$$

(iii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}} = 1$). Again, the change in Φ_k^T which can be obtained by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$ in (80), is trivially bounded like

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1}^T - \Phi_k^T) \\ \leq \quad \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{I}}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right).\end{aligned} \tag{88}$$

(iv) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}} = 1$). Because of the decrease of the frame size parameter and hence the decrease in Φ_k^T, the bound on the change in Φ_k^T follows

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1}^T - \Phi_k^T) \\ \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{U}}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right).\end{aligned} \tag{89}$$

Then, combining (86), (87), (88) and $\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k} \leq 1$, yields

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}^T - \Phi_k^T) \\ \leq \quad \mathbb{1}_{\{T\leq k\}}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right).\end{aligned} \tag{90}$$

Since Assumption 3 holds, it follows from the conditional Cauchy-Schwarz inequality [20] that

$$\begin{aligned}\mathbb{E}\left(\left|f(X_{\text{feas}}^k) - F_0^k\right| |\mathcal{F}_{k-1}^{C\cdot F}\right) &\leq \mathbb{E}\left(1|\mathcal{F}_{k-1}^{C\cdot F}\right)^{1/2}\left[\mathbb{E}\left(\left|f(X_{\text{feas}}^k) - F_0^k\right|^2 |\mathcal{F}_{k-1}^{C\cdot F}\right)\right]^{1/2} \\ &\leq \varepsilon(1-\beta)^{1/2}(\Delta_p^k)^2,\end{aligned} \tag{91}$$

where (91) follows from (12) and the fact that $\mathbb{E}\left(1|\mathcal{F}_{k-1}^{C\cdot F}\right) = 1$. Similarly,

$$\mathbb{E}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| |\mathcal{F}_{k-1}^{C\cdot F}\right) \leq \varepsilon(1-\beta)^{1/2}(\Delta_p^k)^2. \tag{92}$$

Taking expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (90) and then using (91), (92) and the $\mathcal{F}_{k-1}^{C\cdot F}$-measurability of the random variables $\mathbb{1}_{\{T\leq k\}}$ and Δ_p^k, leads to (83).

Case 3 (Bad bounds, $\mathbb{1}_{\bar{I}_k} = 1$). The difference in Φ_k^T may increase since even though good estimates of f values occur, they might not provide enough decrease to cancel the increase in h whenever Algorithm 2 wrongly accepts an incumbent due to bad bounds. It will be shown that

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\mathbb{1}_{\bar{I}_k}(\Phi_{k+1}^T - \Phi_k^T)|\mathcal{F}_{k-1}^{C\cdot F}\right] \leq 2\nu\left[(1-\alpha)^{1/2} + (1-\beta)^{1/2}\right](\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}. \tag{93}$$

(i) The iteration is f-Dominating ($\mathbb{1}_{\mathcal{D}_f} = 1$). The change in Φ_k^T is bounded, taking into account the possible increase in f. Since the value of h is unchanged, the bound on the change in Φ_k^T can be derived from (86) by replacing $\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}$ by $\mathbb{1}_{\bar{I}_k}$. That is,

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_f}(\Phi_{k+1}^T - \Phi_k^T) \\ \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1}) - F_s^k\right| + \left|f(X_{\text{feas}}^k) - F_0^k\right|\right).\end{aligned} \tag{94}$$

(ii) The iteration is h-Dominating ($\mathbb{1}_{\mathcal{D}_h} = 1$). Since the value of f is unchanged, the bound on the change in Φ_k^T is obtained by multiplying both sides of (67) by $\mathbb{1}_{\{T\leq k\}}$ and replacing Φ_k by Φ_k^T. That is,

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1}-\Phi_k) \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right). \quad (95)$$

(iii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}} = 1$). The frame size parameter is updated as in h-Dominating iterations and the value of f is unchanged. Thus, the bound on the change in Φ_k^T follows from (95) by replacing $\mathbb{1}_{\mathcal{D}_h}$ by $\mathbb{1}_{\mathcal{I}}$. That is, follows

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1}-\Phi_k) \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right). \quad (96)$$

(iv) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}} = 1$). Because of the decrease of the frame size parameter and hence the decrease in Φ_k^T, the bound on the change in Φ_k^T is

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}&(\Phi_{k+1}^T-\Phi_k^T)\\ &\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}\nu\left[\frac{1}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right)\right.\\ &\left.+\frac{1}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right)\right]\end{aligned} \quad (97)$$

Since (97) dominates (94), (95) and (96), combining all four cases leads to

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}&(\Phi_{k+1}^T-\Phi_k^T)\\ &\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\nu\left[\frac{1}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right)\right.\\ &\left.+\frac{1}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right)\right]\end{aligned} \quad (98)$$

Taking expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (98) and using (15), (91) and (92) lead to (93). Combining the main results of Case 1, Case 2 and Case 3 of Part 2, specifically (75), (83) and (93),

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\Phi_{k+1}^T-\Phi_k^T|\mathcal{F}_{k-1}^{C\cdot F}\right] \leq \left[-\alpha\beta(1-\nu)(1-\tau^2)+2\nu(1-\alpha)^{1/2}\right.\\ \left.+4\nu(1-\beta)^{1/2}\right](\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}.\end{aligned} \quad (99)$$

Choosing α and β according to (20) ensures that

$$-\alpha\beta(1-\nu)(1-\tau^2)+2\nu(1-\alpha)^{1/2}+4\nu(1-\beta)^{1/2} \leq -\frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2), \quad (100)$$

and (74) follows from (99) and (100) with the same constant $\eta = \frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2)$ as Part 1, which completes the proof. □

Proof of Corollary 4.4

Proof. Only (22) is proved but the proof also applies for $\left|H_s^k - h(X^k + S^k)\right|$ and $\left|F_s^k - f(X^k + S^k)\right|$. According to Assumption 3*(vi)*, $\mathbb{E}\left(\left|H_0^k - h(X^k)\right| \mid \mathcal{F}_{k-1}^{C\cdot F}\right) \leq m\varepsilon(1-\alpha)^{1/2}(\Delta_p^k)^2$, which implies that

$$\mathbb{E}\left(\left|H_0^k - h(X^k)\right|\right) \leq m\varepsilon(1-\alpha)^{1/2}\mathbb{E}\left[(\Delta_p^k)^2\right]. \tag{101}$$

By summing each side of (101) over k from 0 to N, and observing that

$$0 \leq S_N^h := \sum_{k=0}^{N}\left|H_0^k - h(X^k)\right| \nearrow \sum_{k=0}^{+\infty}\left|H_0^k - h(X^k)\right|, \text{ and } 0 \leq S_N^\Delta := \sum_{k=0}^{N}(\Delta_p^k)^2 \nearrow \sum_{k=0}^{+\infty}(\Delta_p^k)^2,$$

it follows from the monotone convergence theorem (see e.g. Theorem 1.6.6 in [32]) that

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=0}^{+\infty}\left|H_0^k - h(X^k)\right|\right) &= \mathbb{E}\left(\lim_{N\to+\infty} S_N^h\right) = \lim_{N\to+\infty}\mathbb{E}\left(S_N^h\right) = \sum_{k=0}^{+\infty}\mathbb{E}\left(\left|H_0^k - h(X^k)\right|\right) \\
&\leq m\varepsilon(1-\alpha)^{1/2}\sum_{k=0}^{+\infty}\mathbb{E}\left[(\Delta_p^k)^2\right] = m\varepsilon(1-\alpha)^{1/2}\lim_{N\to+\infty}\mathbb{E}\left(S_N^\Delta\right) \\
&= m\varepsilon(1-\alpha)^{1/2}\mathbb{E}\left(\lim_{N\to+\infty} S_N^\Delta\right) = m\varepsilon(1-\alpha)^{1/2}\mathbb{E}\left[\sum_{k=0}^{+\infty}(\Delta_p^k)^2\right] \\
&\leq \mu \times m\varepsilon(1-\alpha)^{1/2} < +\infty,
\end{aligned}$$

where μ is from (54). This means that $\sum_{k=0}^{+\infty}\left|H_0^k - h(X^k)\right| < +\infty$ almost surely, which implies the first result of (22). The proof for $\left|F_0^k - f(X^k)\right|$ is similar by observing that (see (91))

$$\mathbb{E}\left(\left|F_0^k - f(X^k)\right| \mid \mathcal{F}_{k-1}^{C\cdot F}\right) \leq \varepsilon(1-\beta)^{1/2}(\Delta_p^k)^2.$$

□

Proof of Lemma 4.7

Proof. The proof uses ideas from [11, 23]. The result is proved by contradiction conditioned on the almost sure event $E_1 = \{\Delta_p^k \to 0\}$. All that follows is conditioned on the event E_1. Assume that with nonzero probability, there exists a random variable $\mathcal{E}' > 0$ such that

$$\Psi_k^h \geq \mathcal{E}', \quad \text{for all } k \in \mathbb{N}, \tag{102}$$

that is,

$$\mathbb{P}\left(\left\{\omega \in \Omega : \exists \mathcal{E}'(\omega) > 0 \text{ such that } \forall k \in \mathbb{N}, \Psi_k^h(\omega) \geq \mathcal{E}'(\omega)\right\}\right) > 0. \tag{103}$$

Let $\{x_{\text{inf}}^k\}_{k\in\mathbb{N}}$, $\{s^k\}_{k\in\mathbb{N}}$, $\{\delta_p^k\}_{k\in\mathbb{N}}$ and $\epsilon' > 0$ be realizations of $\{X_{\text{inf}}^k\}_{k\in\mathbb{N}}$, $\{S^k\}_{k\in\mathbb{N}}$, $\{\Delta_p^k\}_{k\in\mathbb{N}}$ and $\mathcal{E}'$, respectively for which $\psi_k^h \geq \epsilon'$, for all $k \in \mathbb{N}$. Let $\hat{z}$ be the parameter of Algorithm 2 satisfying $\delta_p^k \leq \tau^{-\hat{z}}$ for all $k \geq 0$. Since $\delta_p^k \to 0$ due to the conditioning on E_1, there exists $k_0 \in \mathbb{N}$ such that

$$\delta_p^k < \lambda := \min\left\{\frac{\epsilon'}{m\varepsilon(\gamma+2)}, \tau^{1-\hat{z}}\right\}, \quad \text{for all } k \geq k_0. \tag{104}$$

Consequently and since $\tau < 1$, the random variable R_k with realizations $r_k := -\log_\tau\left(\frac{\delta_p^k}{\lambda}\right)$ satisfies $r_k < 0$ for all $k \geq k_0$. The main idea of the proof is to show that such realizations occur only with probability zero, thus leading to a contradiction. First $\{R_k\}_{k\in\mathbb{N}}$ is shown to be a submartingale. Let $k \geq k_0$ be an iteration for which the events I_k and J_k both occur, which happens with probability at least $\alpha\beta > 1/2$. Then, it follows from the definition of the event I_k (see Definition 3.3) that

$$h(x_{\text{inf}}^k) \;\leq\; u_0^k(x_{\text{inf}}^k) \leq \sum_{j=1}^{m} \max\left\{c_{j,0}^k(x_{\text{inf}}^k), 0\right\} + m\varepsilon(\delta_p^k)^2 = h_0^k(x_{\text{inf}}^k) + m\varepsilon(\delta_p^k)^2, \quad (105)$$

$$\text{and} \quad h(x_{\text{inf}}^k + s^k) \;\geq\; \ell_s^k(x_{\text{inf}}^k + s^k) \geq h_s^k(x_{\text{inf}}^k + s^k) - m\varepsilon(\delta_p^k)^2. \quad (106)$$

$$\begin{aligned}\text{Hence,}\quad h_s^k(x_{\text{inf}}^k + s^k) - h_0^k(x_{\text{inf}}^k) &= [h(x_{\text{inf}}^k + s^k) - h(x_{\text{inf}}^k)] + [h(x_{\text{inf}}^k) - h_0^k(x_{\text{inf}}^k)] \\ &\quad + [h_s^k(x_{\text{inf}}^k + s^k) - h(x_{\text{inf}}^k + s^k)] \\ &\leq 2m\varepsilon(\delta_p^k)^2 - \epsilon'\delta_p^k \leq 2m\varepsilon(\delta_p^k)^2 - m\varepsilon(\gamma+2)(\delta_p^k)^2 = -\gamma m\varepsilon(\delta_p^k)^2 \end{aligned} \quad (107)$$

where the first inequality in (107) follows from (102), (105) and (106) while the last inequality follows from (104). Consequently, iteration k of Algorithm 2 cannot be Unsuccessful. Thus, the frame size parameter is updated according to $\delta_p^{k+1} = \tau^{-1}\delta_p^k$ since $\delta_p^k < \tau^{1-\hat{z}}$. Hence, $r_{k+1} = r_k + 1$.

Let $\mathcal{F}_{k-1}^{I\cdot J} = \sigma(I_0, I_1, \ldots, I_{k-1}) \cap \sigma(J_0, J_1, \ldots, J_{k-1})$. For all other outcomes of I_k and J_k, which will occur with a total probability of at most $1 - \alpha\beta$, the inequality $\delta_p^{k+1} \geq \tau\delta_p^k$ always holds, thus implying that $r_{k+1} \geq r_k - 1$. Hence,

$$\begin{aligned}\mathbb{E}\left(\mathbb{1}_{I_k\cap J_k}(R_{k+1} - R_k)|\mathcal{F}_{k-1}^{I\cdot J}\right) &= \mathbb{P}\left(I_k \cap J_k|\mathcal{F}_{k-1}^{I\cdot J}\right) \geq \alpha\beta \\ \text{and}\quad \mathbb{E}\left(\mathbb{1}_{\overline{I_k\cap J_k}}(R_{k+1} - R_k)|\mathcal{F}_{k-1}^{I\cdot J}\right) &\geq -\mathbb{P}\left(\overline{I_k \cap J_k}|\mathcal{F}_{k-1}^{I\cdot J}\right) \geq \alpha\beta - 1.\end{aligned}$$

Thus, $\mathbb{E}\left(R_{k+1} - R_k|\mathcal{F}_{k-1}^{I\cdot J}\right) \geq 2\alpha\beta - 1 > 0$, implying that $\{R_k\}$ is a submartingale. The remainder of the proof is almost identical to that of the proof of the $\liminf$-type first-order result in [23].

Next is constructed a random walk W_k with realizations w_k on the same probability space as R_k, which will serve as a lower bound on R_k. Define W_k as in (14) by

$$W_k = \sum_{i=0}^{k}(2\cdot\mathbb{1}_{I_i}\mathbb{1}_{J_i} - 1), \quad (108)$$

where the indicator random variables $\mathbb{1}_{I_i}$ and $\mathbb{1}_{J_i}$ are such that $\mathbb{1}_{I_i} = 1$ if I_i occurs, $\mathbb{1}_{I_i} = 0$ otherwise, and similarly, $\mathbb{1}_{J_i} = 1$ if J_i occurs while $\mathbb{1}_{J_i} = 0$ otherwise. Then following the proof of Theorem 3.6, observe that $\{W_k\}$ is a $\mathcal{F}_{k-1}^{I\cdot J}$-submartingale with bounded (nonzero) increments (and, as such, cannot converge to any finite value; see also [23] for the same result), thus leading to the conclusion that the event $\left\{\limsup_{k\to+\infty} W_k = +\infty\right\}$ occurs almost surely. Since by construction

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

then with probability one, R_k is positive infinitely often. Thus, the sequence of realizations r_k such that $r_k < 0$ for all $k \geq k_0$ occurs with probability zero. Thus, the assumption that (103) holds is false. This implies that

$$\mathbb{P}\left(\left\{\omega \in \Omega : \forall \mathcal{E}'(\omega) > 0, \exists k \in \mathbb{N} \text{ such that } \Psi_k^h(\omega) < \mathcal{E}'(\omega)\right\}\right) = 1,$$

which means that (23) holds. □

Proof of Theorem 4.10

Proof. The theorem is proved using ideas from [9, 11]. Define the events E_1 and E_2 by

$$E_1 = \left\{\omega \in \Omega : \Delta_p^k(\omega) \to 0\right\} \quad \text{and} \quad E_2 = \left\{\omega \in \Omega : \exists K'(\omega) \subset \mathbb{N} \text{ such that } \lim_{K'(\omega)} \Psi_k^h(\omega) \leq 0\right\}.$$

Then E_1 and E_2 are almost sure due to Corollary 4.3 and (23) respectively. Let $\omega \in E_1 \cap E_2$ be an arbitrary outcome and note that the event $E_1 \cap E_2$ is also almost sure as a countable intersection of almost sure events. Then $\lim_{K'(\omega)} \Delta_p^k(\omega) = 0$. It follows from the compactness hypothesis of Assumption 2 that there exists $K(\omega) \subseteq K'(\omega)$ for which the subsequence $\{X_{\text{inf}}^k(\omega)\}_{k \in K(\omega)}$ converges to a limit $\hat{X}_{\text{inf}}(\omega)$. Specifically, $\hat{X}_{\text{inf}}(\omega)$ is a refined point for the refining subsequence $\{X_{\text{inf}}^k(\omega)\}_{k \in K(\omega)}$. Let $v \in T_{\mathcal{X}}^H(\hat{X}_{\text{inf}}(\omega))$ be a refining direction for $\hat{X}_{\text{inf}}(\omega)$. Denote by V the random vector with realizations v, i.e., $v = V(\omega)$, and let $\hat{x}_{\text{inf}} = \hat{X}_{\text{inf}}(\omega)$, $x_{\text{inf}}^k = X_{\text{inf}}^k(\omega)$, $\delta_p^k = \Delta_p^k(\omega)$, $\delta_m^k = \Delta_m^k(\omega)$, $\psi_k^h = \Psi_k^h(\omega)$ and $\mathcal{K} = K(\omega)$. Since v is a refining direction, there exists $\mathcal{L} \subseteq \mathcal{K}$ and polling directions $d^k \in \mathbb{D}_p^k(x_{\text{inf}}^k)$ such that $v = \lim_{k \in \mathcal{L}} \frac{d^k}{\|d^k\|_\infty}$. For each $k \in \mathcal{L}$, define

$$t_k = \delta_m^k \|d^k\|_\infty \to 0, \qquad y^k = x_{\text{inf}}^k + t_k \left(\frac{d^k}{\|d^k\|_\infty} - v\right) \to \hat{x}_{\text{inf}},$$
$$a_k = \frac{h(y^k + t_k v) - h(x_{\text{inf}}^k)}{t_k} \quad \text{and} \quad b_k = \frac{h(x_{\text{inf}}^k) - h(y^k)}{t_k},$$

where the fact that $t_k \to 0$ follows from Definition 2.11, specifically the inequality $\delta_m^k \|d^k\|_\infty \leq \delta_p^k b$. Since h is λ^h–locally Lipschitz,

$$|a_k| \leq \frac{\lambda^h}{t_k} \|(y^k + t_k v) - x_{\text{inf}}^k\|_\infty = \lambda^h \quad \text{and} \quad |b_k| \leq \frac{\lambda^h}{t_k} \|x_{\text{inf}}^k - y^k\|_\infty = \lambda^h \left\| \frac{d^k}{\|d^k\|_\infty} - v \right\|_\infty \to 0,$$

which shows that Lemma 4.9 applies to both subsequences $\{a_k\}_{k \in \mathcal{L}}$ and $\{b_k\}_{k \in \mathcal{L}}$. Moreover, combining the inequality $\lim_{\mathcal{L}} \psi_k^h \leq 0$ and Assumption 6 (the fact that $\delta_p^k \|d^k\|_\infty \geq d_{\min} > 0$), yields

$$\begin{aligned} \lim_{k \in \mathcal{L}} \left(\frac{-\psi_k^h}{\delta_p^k \|d^k\|_\infty}\right) &= \lim_{k \in \mathcal{L}} \frac{h(x_{\text{inf}}^k + \delta_m^k d^k) - h(x_{\text{inf}}^k)}{(\delta_p^k)^2 \|d^k\|_\infty} \\ &= \lim_{k \in \mathcal{L}} \frac{h(x_{\text{inf}}^k + \delta_m^k d^k) - h(x_{\text{inf}}^k)}{t_k} \geq -d_{\min}^{-1} \lim_{k \in \mathcal{L}} \psi_k^h \geq 0, \end{aligned} \tag{109}$$

where the equality in (109) follows from $\delta_m^k = (\delta_p^k)^2$ for sufficiently large k. Thus, by adding and subtracting $h(x_{\text{inf}}^k)$ to the numerator of the definition of the Clarke derivative, and using the fact that

$x^k_{\text{inf}} + \delta^k_m d^k \in \mathcal{X}$ for sufficiently large $k \in \mathcal{L}$ since v is a hypertangent direction,

$$\begin{aligned} h^\circ(\hat{x}_{\text{inf}}; v) &\geq \limsup_{k\in\mathcal{L}} \frac{h(y^k + t_k v) - h(x^k_{\text{inf}}) + h(x^k_{\text{inf}}) - h(y^k)}{t_k} = \limsup_{k\in\mathcal{L}} (a_k + b_k) \\ &= \limsup_{k\in\mathcal{L}} a_k + \lim_{k\in\mathcal{L}} b_k = \limsup_{k\in\mathcal{L}} \frac{h(x^k_{\text{inf}} + \delta^k_m d^k) - h(x^k_{\text{inf}})}{t_k} \geq 0, \end{aligned}$$

where the last inequality follows from (109). Every outcome ω arbitrarily chosen in $E_1 \cap E_2$ therefore belongs to the event

$$\begin{aligned} E_3 := \Big\{ \omega \in \Omega : \exists K(\omega) \subseteq \mathbb{N} \text{ and } \exists \hat{X}_{\text{inf}}(\omega) = \lim_{k\in K(\omega)} X^k_{\text{inf}}(\omega), \hat{X}_{\text{inf}}(\omega) \in \mathcal{X}, \text{ such that} \\ \forall V(\omega) \in T^H_{\mathcal{X}}(\hat{X}_{\text{inf}}(\omega)),\ h^\circ(\hat{X}_{\text{inf}}(\omega); V(\omega)) \geq 0 \Big\}, \end{aligned}$$

thus implying that $E_1 \cap E_2 \subseteq E_3$. Then the proof is complete since $\mathbb{P}(E_1 \cap E_2) = 1$. □

Proof of Corollary 4.11

Proof. The proof is almost identical to the proof of a similar result (Corollary 3.6) in [9]. Recall the sequence K' of random variables and the almost sure event $E_1 \cap E_2$ in the proof of Theorem 4.10 and let $\omega \in E_1 \cap E_2$. Following the latter proof, there exists $K(\omega) \subseteq K'(\omega)$ such that $\lim_{K(\omega)} X^k_{\text{inf}}(\omega) = \hat{X}_{\text{inf}}(\omega) = \hat{x}_{\text{inf}}$. Moreover, it follows from Theorem 4.10 that $h^\circ(\hat{x}_{\text{inf}}; v) = h^\circ(\hat{X}_{\text{inf}}(\omega); V(\omega)) \geq 0$ for a set of refining directions v which is dense in the closure $\text{cl}\left(T^H_{\mathcal{X}}(\hat{x}_{\text{inf}})\right)$ of $T^H_{\mathcal{X}}(\hat{x}_{\text{inf}})$. Then the proof is complete by noticing that $\text{cl}\left(T^H_{\mathcal{X}}(\hat{x}_{\text{inf}})\right) = T^{Cl}_{\mathcal{X}}(\hat{x}_{\text{inf}})$ wherever $T^H_{\mathcal{X}}(\hat{x}_{\text{inf}}) \neq \emptyset$ [50], with $T^{Cl}_{\mathcal{X}}(\hat{x}_{\text{inf}})$ denoting the Clarke tangent cone to $\mathcal{X}$ at $\hat{x}_{\text{inf}}$. □

Proof of Lemma 4.12

Proof. The proof is almost identical to those of Lemma 4.7 and a similar result in [11]. Hence, full details are not provided here again. Unless otherwise stated, all the sequences, events and constants considered are defined as in the proof of Lemma 4.7. The result is proved by contradiction and all that follows is conditioned on the almost sure event $E_1 \cap \{T < +\infty\}$. Assume that with nonzero probability there exists a random variable $\mathcal{E}'' > 0$ such that

$$\Psi^{f,T}_k \geq \mathcal{E}'', \quad \text{for all } k \geq 0. \tag{110}$$

Let t, $\{x^{k\vee t}_{\text{feas}}\}_{k\in\mathbb{N}}$, $\{s^k\}_{k\in\mathbb{N}}$, $\{\delta^k_p\}_{k\in\mathbb{N}}$ and $\epsilon'' > 0$ be realizations of T, $\{X^{k\vee T}_{\text{feas}}\}_{k\in\mathbb{N}}$, $\{S^k\}_{k\in\mathbb{N}}$, $\{\Delta^k_p\}_{k\in\mathbb{N}}$ and $\mathcal{E}''$, respectively for which $\psi^{f,t}_k \geq \epsilon''$ for all $k \geq 0$. Let $\bar{k}_0 \in \mathbb{N}^*$ be such that

$$\delta^k_p < \lambda := \min\left\{ \frac{\epsilon''}{\varepsilon(\gamma+2)}, \tau^{1-\hat{z}} \right\} \quad \text{for all } k \geq \bar{k}_0. \tag{111}$$

The key element of the proof is to show that an iteration $k \geq k_0 := \max\{\bar{k}_0, t\}$ for which the events I_k and J_k both occur cannot be Unsuccessful, and hence $\{R_k\}$ is a submartingale.

It follows from (110) and (111) that

$$f(x^k_{\text{feas}} + s^k) - f(x^k_{\text{feas}}) \leq -\epsilon'' \delta^k_p \leq -(\gamma + 2)\varepsilon(\delta^k_p)^2, \quad \text{for all } k \geq k_0.$$

Since J_k occurs,
$$\begin{aligned} f^k_s(x^k_{\text{feas}} + s^k) - f^k_0(x^k_{\text{feas}}) &= [f(x^k_{\text{feas}} + s^k) - f(x^k_{\text{feas}})] + [f(x^k_{\text{feas}}) - f^k_0(x^k_{\text{feas}})] \\ &\quad + [f^k_s(x^k_{\text{feas}} + s^k) - f(x^k_{\text{feas}} + s^k)] \\ &\leq -(\gamma + 2)\varepsilon(\delta^k_p)^2 + 2\varepsilon(\delta^k_p)^2 = -\gamma\varepsilon(\delta^k_p)^2, \end{aligned}$$

which implies that the iteration $k \geq k_0$ of Algorithm 2 cannot be Unsuccessful. The rest of the proof follows that of Lemma 4.7. □

Proof of Theorem 4.13

Proof. The proof follows from Corollary 4.4 and the assumption $\lim_{k\in K} H^k_0(X^{k\vee T}_{\text{feas}}) = 0$ almost surely, by observing that for any outcome ω in the almost sure event

$$E_4 := \Big\{\omega \in \Omega : \forall K(\omega) \subseteq \mathbb{N}, \lim_{k\in K(\omega)} \big|H^k_0(X^{k\vee T}_{\text{feas}})(\omega) - h(X^{k\vee T}_{\text{feas}}(\omega))\big| = 0 \text{ and } \lim_{k\in K(\omega)} H^k_0(X^{k\vee T}_{\text{feas}})(\omega) = 0\Big\} \cap \{T < +\infty\},$$

the inequalities

$$\begin{aligned} h(X^{k\vee T}_{\text{feas}}(\omega)) - \big|H^k_0(X^{k\vee T}_{\text{feas}})(\omega)\big| &\leq \big|h(X^{k\vee T}_{\text{feas}}(\omega)) - \big|H^k_0(X^{k\vee T}_{\text{feas}})(\omega)\big|\big| \\ &\leq \big|h(X^{k\vee T}_{\text{feas}}(\omega)) - H^k_0(X^{k\vee T}_{\text{feas}})(\omega)\big| \end{aligned}$$

and the continuity of $y \mapsto |y|$, yield

$$\begin{aligned} \lim_{k\in K(\omega)} h(X^{k\vee T}_{\text{feas}}(\omega)) &\leq \lim_{k\in K(\omega)} \big(\big|h(X^{k\vee T}_{\text{feas}}(\omega)) - H^k_0(X^{k\vee T}_{\text{feas}})(\omega)\big| + \big|H^k_0(X^{k\vee T}_{\text{feas}})(\omega)\big|\big) \\ &= \lim_{k\in K(\omega)} \big|h(X^{k\vee T}_{\text{feas}}(\omega)) - H^k_0(X^{k\vee T}_{\text{feas}})(\omega)\big| + \Big|\lim_{k\in K(\omega)} H^k_0(X^{k\vee T}_{\text{feas}})(\omega)\Big| = 0. \end{aligned}$$

This means that

$$h(\hat{X}_{\text{feas}}(\omega)) = \lim_{k\in K(\omega)} h(X^{k\vee T}_{\text{feas}}(\omega)) = 0 \tag{112}$$

since h is nonnegative, where the first equality in (112) follows from the continuity of h in $\mathcal{X}$. Consequently,

$$\mathbb{P}\left(h(\hat{X}_{\text{feas}}) = 0\right) = \mathbb{P}\left(\hat{X}_{\text{feas}} \in \mathcal{D}\right) = 1.$$

□

Proof of Theorem 4.14

Proof. First, $\mathbb{P}\left(\hat{X}_{\text{feas}} \in \mathcal{D}\right) = 1$ follows from Theorem 4.13. The proof follows from that of Theorem 4.10, by replacing h by f, $\hat{x}_{\text{inf}} = \hat{X}_{\text{inf}}(\omega)$ by $\hat{x}_{\text{feas}} = \hat{X}_{\text{feas}}(\omega)$, $x^k_{\text{inf}} = X^k_{\text{inf}}(\omega)$ by $x^{k\vee t}_{\text{feas}} = X^{k\vee T}_{\text{feas}}(\omega)$, $\psi^h_k = \Psi^h_k(\omega)$ by $\psi^{f,t}_k = \Psi^{f,T}_k(\omega)$ with $t = T(\omega)$ and $T^H_{\mathcal{X}}(\cdot)$ by $T^H_{\mathcal{D}}(\cdot)$, for ω fixed and arbitrarily chosen in the almost sure event $E_1 \cap E_5 \cap \{T < +\infty\}$, where

$$E_5 = \Big\{\omega \in \Omega : \exists K(\omega) \subseteq \mathbb{N} \text{ such that } \hat{X}_{\text{feas}}(\omega) = \lim_{k\in K(\omega)} X^{k\vee T}_{\text{feas}}(\omega),\ \hat{X}_{\text{feas}}(\omega) \in \mathcal{D},$$
$$\lim_{k\in K(\omega)} \Psi^{f,T}_k(\omega) \leq 0 \ \text{ and } \lim_{k\in K(\omega)} H^k_0(X^{k\vee T}_{\text{feas}})(\omega) = 0\Big\}.$$

□

Proof of Corollary 4.15

Proof. The proof is almost identical to the proof of a similar result (Corollary 3.4) in [9]. Let ω be arbitrarily chosen in the almost sure event $E_1 \cap E_5 \cap \{T < +\infty\}$. It follows from Theorem 4.14 that $f^\circ(\hat{x}_{\text{feas}}; v) = f^\circ(\hat{X}_{\text{feas}}(\omega); V(\omega)) \geq 0$ for a set of refining directions v which is dense in the closure of $T^H_{\mathcal{D}}(\hat{x}_{\text{feas}})$. Then the proof is complete by noticing that the closure of the hypertangent cone coincides with the Clarke tangent cone wherever the hypertangent cone is nonempty [9, 50]. □

References

[1] 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.

[2] S. Alarie, C. Audet, P.-Y. Bouchet, and S. Le Digabel. Optimization of noisy blackboxes with adaptive precision. *SIAM Journal on Optimization*, 31(4):3127–3156, 2021.

[3] 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.

[4] E. Angün, J. Kleijnen, D. den Hertog, and G. Gürkan. Response surface methodology with stochastic constraints for expensive simulation. *Journal of the operational research society*, 60(6):735–746, 2009.

[5] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

[6] C. Audet. A survey on direct search methods for blackbox optimization and their applications. In P.M. Pardalos and T.M. Rassias, editors, *Mathematics without boundaries: Surveys in interdisciplinary research*, chapter 2, pages 31–56. Springer, New York, NY, 2014.

[7] C. Audet and J.E. Dennis, Jr. A pattern search filter method for nonlinear programming without derivatives. *SIAM Journal on Optimization*, 14(4):980–1010, 2004.

[8] C. Audet and J.E. Dennis, Jr. Mesh Adaptive Direct Search Algorithms for Constrained Optimization. *SIAM Journal on Optimization*, 17(1):188–217, 2006.

[9] C. Audet and J.E. Dennis, Jr. A Progressive Barrier for Derivative-Free Nonlinear Programming. *SIAM Journal on Optimization*, 20(1):445–472, 2009.

[10] 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.

[11] C. Audet, K.J. Dzahini, M. Kokkolaras, and S. Le Digabel. Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. *Computational Optimization and Applications*, 79(1):1–34, 2021.

[12] C. Audet and W. Hare. *Derivative-Free and Blackbox Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, Switzerland, 2017.

[13] 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.

[14] C. Audet, S. Le Digabel, and C. Tribes. The Mesh Adaptive Direct Search Algorithm for Granular and Discrete Variables. *SIAM Journal on Optimization*, 29(2):1164–1189, 2019.

[15] F. Augustin and Y.M. Marzouk. NOWPAC: A provably convergent derivative-free nonlinear optimizer with path-augmented constraints. Technical report, arXiv, 2014.

[16] F. Augustin and Y.M. Marzouk. A trust-region method for derivative-free nonlinear constrained stochastic optimization. Technical Report 1703.04156, arXiv, 2017.

[17] 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.

[18] R.R. Barton and J.S. Ivey, Jr. Nelder-Mead simplex modifications for simulation optimization. *Management Science*, 42(7):954–973, 1996.

[19] D. Bertsimas, O. Nohadani, and K.M. Teo. Nonconvex robust optimization for problems with constraints. *INFORMS Journal on Computing*, 22(1):44–58, 2010.

[20] R.N. Bhattacharya and E.C. Waymire. *A basic course in probability theory*, volume 69. Springer, 2007.

[21] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence Rate Analysis of a Stochastic Trust Region Method via Submartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

[22] 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.

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

[24] 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.

[25] F.H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley and Sons, New York, 1983. Reissued in 1990 by SIAM Publications, Philadelphia, as Vol. 5 in the series Classics in Applied Mathematics.

[26] A.R. Conn, K. Scheinberg, and L.N. Vicente. *Introduction to Derivative-Free Optimization*. MOS-SIAM Series on Optimization. SIAM, Philadelphia, 2009.

[27] F.E. Curtis and K. Scheinberg. Adaptive Stochastic Optimization: A Framework for Analyzing Stochastic Optimization Algorithms. *IEEE Signal Processing Magazine*, 37(5):32–42, 2020.

[28] F.E. Curtis, K. Scheinberg, and R. Shi. A Stochastic Trust Region Algorithm Based on Careful Step Normalization. *INFORMS Journal on Optimization*, 1(3):200–220, 2019.

[29] M.A. Diniz-Ehrhardt, D.G. Ferreira, and S.A. Santos. A pattern search and implicit filtering algorithm for solving linearly constrained minimization problems with noisy objective functions. *Optimization Methods and Software*, 34(4):827–852, 2019.

[30] M.A. Diniz-Ehrhardt, D.G. Ferreira, and S.A. Santos. Applying the pattern search implicit filtering algorithm for solving a noisy problem of parameter identification. *Computational Optimization and Applications*, pages 1–32, 2020.

[31] E.D. Dolan and J.J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

[32] R. Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

[33] K.J. Dzahini. Expected complexity analysis of stochastic direct-search. *Computational Optimization and Applications*, 81(1):179–200, 2021.

[34] S. Gratton, C.W. Royer, L.N. Vicente, and Z. Zhang. Direct search based on probabilistic feasible descent for bound and linearly constrained problems. *Computational Optimization and Applications*, 72(3):525–559, 2019.

[35] 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.

[36] J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer, Berlin, 1994.

[37] S. Kitayama, M. Arakawa, and K. Yamazaki. Sequential approximate optimization using radial basis function network for engineering optimization. *Optimization and Engineering*, 12(4):535–557, 2011.

[38] K.J. Klassen and R. Yoogalingam. Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management*, 18(4):447–458, 2009.

[39] T. Lacksonen. Empirical comparison of search algorithms for discrete event simulation. *Computers & Industrial Engineering*, 40(1-2):133–148, 2001.

[40] J. Larson and S.C. Billups. Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and Applications*, 64(3):619–645, 2016.

[41] S. Le Digabel and S.M. Wild. A Taxonomy of Constraints in Simulation-Based Optimization. Technical Report G-2015-57, Les cahiers du GERAD, 2015.

[42] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

[43] L. Lukšan and J. Vlček. Test problems for nonsmooth unconstrained and linearly constrained optimization. Technical Report V-798, ICS AS CR, 2000.

[44] E. Mezura-Montes and C.A. Coello. Useful Infeasible Solutions in Engineering Optimization with Evolutionary Algorithms. In *Proceedings of the 4th Mexican International Conference on Advances in Artificial Intelligence*, MICAI'05, pages 652–662, Berlin, Heidelberg, 2005. Springer-Verlag.

[45] J. Mockus. *Bayesian approach to global optimization: theory and applications*, volume 37 of *Mathematics and Its Applications*. Springer Science & Business Media, 2012.

[46] J.J. Moré and S.M. Wild. Benchmarking Derivative-Free Optimization Algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009.

[47] J.A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[48] C. Paquette and K. Scheinberg. A Stochastic Line Search Method with Expected Complexity Analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[49] H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[50] R.T. Rockafellar. Generalized directional derivatives and subgradients of nonconvex functions. *Canad. J. Math.*, 32(2):257–280, 1980.

[51] 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.

[52] S. Shashaani, F.S. Hashemi, and R. Pasupathy. ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM Journal on Optimization*, 28(4):3145–3176, 2018.

[53] 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.

[54] Z. Wang and M. Ierapetritou. Constrained optimization of black-box stochastic systems using a novel feasibility enhanced Kriging-based method. *Computers & Chemical Engineering*, 118:210–223, 2018.

[55] 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.