Titre: Title:	Méthodes de recherche directe pour l'optimisation stochastique de boîtes noires
Auteur: Author:	Kwassi Joseph Dzahini
Date:	2020
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Dzahini, K. J. (2020). Méthodes de recherche directe pour l'optimisation stochastique de boîtes noires [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/5524/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/5524/
Directeurs de recherche: Advisors:	Sébastien Le Digabel, & Michael Kokkolaras
Programme: Program:	Doctorat en mathématiques

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Méthodes de recherche directe pour l'optimisation stochastique de boîtes noires

KWASSI JOSEPH DZAHINI
Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Mathématiques

Décembre 2020

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Méthodes de recherche directe pour l'optimisation stochastique de boîtes noires

présentée par **Kwassi Joseph DZAHINI**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Michel GENDREAU, président
Sébastien LE DIGABEL, membre et directeur de recherche
Michael KOKKOLARAS, membre et codirecteur de recherche
Richard LABIB, membre
Matt MENICKELLY, membre externe

DÉDICACE

À mon père Jean-Sébastien-Bach
et à ma mère Ama

REMERCIEMENTS

La réalisation de cette thèse n'aurait pas été possible sans le concours de plusieurs personnes à qui je voudrais témoigner ma profonde gratitude.

Je tiens d'abord à remercier les professeurs Sébastien Le Digabel et Michael Kokkolaras pour avoir accepté de m'encadrer tout au long de cette thèse. Qu'ils soient également remerciés pour leur gentillesse, leur disponibilité permanente, leurs connaissances et expériences personnelles qu'ils ont partagées avec moi, leur patience ainsi que leurs précieuses suggestions qui ont largement facilité l'aboutissement de cette longue aventure. Je voudrais particulièrement vous dire un grand merci pour votre confiance. Ce fut un immense plaisir de faire des recherches sous votre supervision.

Je voudrais exprimer ma reconnaissance envers le professeur Charles Audet pour la qualité de ses cours qui ont su nourrir mes réflexions et qui m'ont donné des bases nécessaires pour la réalisation des différents projets de cette thèse. Je voudrais sincèrement le remercier pour sa patience et sa disponibilité à répondre à mes innombrables questions. Merci également pour ses suggestions pertinentes et sa précieuse collaboration, qui ont permis de mener à bien le premier projet de cette thèse.

Merci également à Hydro-Québec, à Rio-Tinto, au FRQNT et au GERAD pour leur support financier tout au long de la thèse. Merci à Christophe et Viviane pour avoir toujours répondu à toutes mes questions. Merci à tous mes collègues et amis du GERAD, en particulier Dounia, Khalil, Ludovic, Pierre-Yves, Vilmar, Alexis, et surtout Mathieu Tanneau et Christian Bingane, à qui j'ai posé tellement de questions. Un grand merci à mes amis Kwassi Holali Degue, Josette Kuagbenu, Mawunyo K. Dagnon, Joseph Agbavor et surtout Francis Hounkpe pour les multiples discussions et collaborations.

Je voudrais également que les professeurs Mylène Maïda et David Dereudre de l'université Lille 1 trouvent ici les sentiments de ma plus profonde gratitude pour leurs encouragements, leurs recommandations et pour m'avoir donné à travers leurs cours de solides bases en probabilité. Un spécial merci à Mme Maïda chez qui j'avais appris tellement de choses durant mon stage de Master 2.

Je tiens spécialement à remercier le professeur Yaogan Mensah de l'université de Lomé, pour les solides bases acquises au travers de ses cours. Merci également pour ses recommandations et ses grands efforts avec les professeurs Midzodzi K. Pekpe et Djidula K. Motchon, grâce auxquels j'avais pu intégrer l'université Lille 1. Je tiens également à témoigner toute ma gratitude aux professeurs Edi K. Gagnassi pour ses nombreux soutiens, et Elias Apetogbo grâce à qui j'avais compris pour la toute première fois au Lycée ce que précision et rigueur voulaient dire en mathématiques.

Merci à mes chers parents pour leur soutien inconditionnel et pour avoir cru en moi, et un spécial merci à Estelle et Aurore pour m'avoir épaulé chaque jour et poussé vers l'avant.

RÉSUMÉ

L'optimisation informatique est omniprésente dans de nombreuses communautés et a suscité un intérêt sans précédent au cours des dernières décennies en raison de l'utilisation croissante des simulations informatiques pour la résolution de problèmes complexes survenant dans une multitude de domaines et plus particulièrement en ingénierie. "L'optimisation de boîtes noires", thème majeur de cette thèse, n'émet aucune hypothèse sur l'existence de dérivées des fonctions ciblées et se concentre sur des problèmes où la fonction objectif ainsi que les contraintes sont données par une boîte noire dont les simulations informatiques constituent un exemple typique. Les valeurs exactes de ces fonctions déterministes au moyen desquelles sont modélisés les problèmes couramment rencontrés dans le domaine de l'apprentissage automatique, du traitement du signal, de la médecine ainsi que la biologie etc., ne sont en général pas toujours accessibles numériquement. En effet, elles ne peuvent parfois être obtenues que par le biais d'une boîte noire dont les évaluations sont corrompues par un bruit aléatoire, engendrant ainsi des problèmes d'optimisation stochastique de boîtes noires. Par ailleurs, même si la conception d'algorithmes performants d'optimisation stochastique a suscité un regain d'intérêt ces dernières années, il est à noter que la plupart des méthodes d'optimisation dites *sans dérivée* utilisent des approximations de gradient, se limitent à des problèmes sans contraintes ou sinon, supposent que ces dernières sont déterministes. Ainsi, les méthodes de recherche directe connues pour leur fiabilité et leur robustesse en pratique s'avèrent plus prometteuses pour les problèmes d'optimisation de boîtes noires. Cependant, force est de constater qu'assez rares sont les travaux portant sur la conception d'algorithmes de recherche directe pour l'optimisation stochastique de boîtes noires, dotés d'une preuve de convergence et surtout en présence de contraintes bruitées aléatoirement. Cette thèse se concentre alors sur la conception ainsi que l'analyse de convergence d'algorithmes de recherche directe pour l'optimisation stochastique de boîtes noires.

Le premier projet de la présente thèse introduit une variante stochastique de l'algorithme de recherche directe par treillis adaptatif (MADS) initialement conçu pour l'optimisation de boîtes noires déterministes. La méthode proposée nommée StoMADS considère l'optimisation sans contrainte d'une fonction objectif dont les valeurs ne peuvent être obtenues que via une boîte noire corrompue par un bruit aléatoire de distribution inconnue. Étant donné que les valeurs déterministes exactes de ladite fonction sous-jacente ne sont pas accessibles numériquement, StoMADS fait usage de leurs estimés obtenus à partir d'observations aléatoires. En supposant ces estimés précis avec une probabilité suffisamment grande mais fixe, et de variance satisfaisant une certaine condition, un résultat de convergence de StoMADS a été obtenu grâce au calcul non lisse de Clarke et la théorie des martingales. Des études numériques comparant StoMADS à certains algorithmes disponibles dans le logiciel NOMAD d'op-

timisation de boîtes noires, ont révélé la méthode proposée très prometteuse pour des applications concrètes.

Le second projet généralise le cadre algorithmique de StoMADS en introduisant une large classe d'algorithmes de recherche directe dits de type directionnel, pour l'optimisation stochastique de boîtes noires (SDDS), et qui acceptent de nouveaux itérés en imposant une condition de décroissance suffisante à des estimés satisfaisant les mêmes hypothèses du premier projet. Sous une hypothèse supplémentaire de différentiabilité de la fonction objectif déterministe numériquement inaccessible, le taux de convergence de la méthode proposée a été étudié à l'aide des supermartingales. Quoique ne faisant usage d'aucune information de gradient contrairement à des méthodes existantes telles que la recherche linéaire stochastique ainsi que les algorithmes de région de confiance stochastiques, le taux de convergence de SDDS est similaire à celui des deux dernières méthodes et à celui de la large classe d'algorithmes de recherche directe déterministes de type directionnel qui acceptent également de nouveaux itérés en imposant une condition de décroissance suffisante.

Le troisième projet de cette thèse introduit StoMADS-PB, un algorithme d'optimisation stochastique de boîtes noires sous contraintes. StoMADS-PB est une variante stochastique de l'algorithme MADS utilisant une approche de barrière progressive (PB) pour la gestion des contraintes, initialement conçu pour l'optimisation de boîtes noires déterministes sous contraintes générales. Similairement à StoMADS, les valeurs de la fonction objectif et des contraintes sont accessibles uniquement via une boîte noire bruitée aléatoirement. Ainsi, StoMADS-PB utilise également des estimés et par ailleurs, traite les contraintes en regroupant leurs *violations* en une seule *fonction de violation de contrainte*. Sous l'hypothèse que ces estimés sont précis et fiables avec des probabilités suffisamment grandes mais fixes, des résultats de convergence de la méthode vers des points stationnaires au sens de la dérivée généralisée de Clarke des fonctions objectif et de violation, ont été obtenus grâce à la théorie des martingales. Enfin, des expériences numériques menées sur 126 variantes stochastiques de 42 problèmes sous contraintes tirés de la littérature ont démontré l'efficacité de StoMADS-PB par rapport à MADS avec PB pour l'optimisation stochastique de boîtes noires.

ABSTRACT

Computational optimization is ubiquitous in many communities and has attracted an unparalleled interest during the last decades due to the growing use of computer simulations when solving complex problems arising in a plethora of fields, especially in engineering. Blackbox optimization (BBO), a major theme in the present thesis, does not assume the existence of derivatives and focuses on problems where the objective and constraint functions are given by a blackbox, typical examples of which are computer simulations. For many of such problems arising in machine learning, signal processing, medicine and biology to name a few, the target deterministic objective function and constraints can only be accessed through a blackbox corrupted by some random noise, thus resulting in stochastic BBO problems. Even though developing provable algorithms for stochastic optimization has recently received renewed interest, most of derivative-free optimization (DFO) methods either use estimated gradient informations, are restricted to unconstrained problems, or assume that the constraints are deterministic. Thus, direct-search methods known to be reliable and robust in practice appear to be the most promising approach for BBO problems. However in stochastic BBO, there is relatively scarce research on developing direct-search methods with full-supported convergence analysis especially when constraints function evaluations are also corrupted by some random noise. This thesis therefore focuses on the development and convergence analysis of direct-search methods for stochastic BBO.

The first project of this thesis introduces a stochastic extension of the mesh adaptive direct-search (MADS) algorithm originally designed for deterministic BBO. The proposed method called StoMADS considers the unconstrained optimization of an objective function when only having access to its noisy evaluations available through a blackbox corrupted by some random noise following an unknown distribution. Thus, since the exact deterministic values of the underlying objective function are not available, their estimates obtained from stochastic observations are used. By requiring the accuracy of such estimates to hold with a large but fixed probability and by assuming them to satisfy a variance condition, a Clarke stationarity convergence result of StoMADS is proved by means of martingale theory. Computational studies comparing StoMADS to algorithms available in the NOMAD BBO software revealed that the proposed method is very promising for real-life applications.

The second project generalizes the algorithmic framework of StoMADS by introducing a broad class of stochastic directional direct-search (SDDS) algorithms which accept new iterates by imposing a sufficient decrease condition on function estimates required to satisfy the same StoMADS assumptions. By assuming in addition the objective function to be differentiable, the expected complexity of SDDS is analyzed making use of an existing supermartingale-based framework. Despite using no gradient information unlike prior methods such as stochastic line-search and stochastic trust-region

methods, the convergence rate of SDDS is shown to be similar to that of both latter methods and to that of the broad class of deterministic directional direct-search methods which accept new iterates using a sufficient decrease condition.

By introducing the StoMADS-PB algorithm, the third project of the thesis focuses on the constrained stochastic BBO. The proposed method is a stochastic extension of the MADS method using a progressive barrier (PB) approach for constraints handling, originally developed for deterministic BBO under general constraints. Similarly to the framework of StoMADS, the objective and constraints function values can only be computed through a noisy blackbox. Constraints are treated by StoMADS-PB by aggregating their corresponding violations into a single *constraint violation function*. Since all the underlying deterministic functions values are not available, estimates are used and so-called *probabilistic bounds* are introduced for the violation function. By requiring such estimates and bounds to be accurate and reliable with sufficiently high but fixed probabilities, Clarke stationarity results for the objective and the violation function are derived with probability one, making use of the Clarke nonsmooth calculus and martingale theory. Numerical experiments conducted on 126 stochastic variants of 42 constrained problems from literature demonstrated StoMADS-PB to outperform MADS with PB.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

BBO	Blackbox Optimization
DFO	Derivative-Free Optimization
MADS	Mesh Adaptive Direct Search
SDDS	Stochastic Directional Direct Search
PB	Progressive Barrier
MADS-PB	Algorithme MADS avec barrière progressive
EB	Extreme Barrier
CS	Coordinate Search
NM	Nelder Mead
SNM	Stochastic Nelder Mead
STORM	Stochastic Optimization with Random Models
STARS	Stepsize Approximation in Random Search
SPSA	Simultaneous Perturbation Stochastic Approximation
RSM	Response Surface Methodology
$\mathbb{E}_{\Theta}(\cdot)$	Espérance par rapport à Θ
$\|\cdot\|_{\infty}$	Opérateur norme infini
$\|\cdot\|$	Opérateur valeur absolue
$\bar{(\cdot)}$	Opérateur complémentaire
$\mathcal{B}_r(x)$	Boule centrée en x et de rayon r
$a \vee b$	Maximum de a et b
$\lfloor \cdot \rfloor$	Opérateur partie entière

CHAPITRE 1 INTRODUCTION

L'optimisation est une branche des mathématiques qui étudie la modélisation, l'analyse ainsi que la résolution analytique ou numérique de problèmes de minimisation ou de maximisation d'une fonction éventuellement soumise à des contraintes. Après Euclide qui aurait formulé les premiers problèmes d'optimisation au III^{e} siècle avant notre ère, l'optimisation connu plus tard d'importantes avancées. Même si ces dernières ne sauraient être résumées en quelques lignes, on peut tout de même évoquer la mise au point d'une méthode itérative par Newton au XVII^{e} siècle, faisant intervenir la notion de *dérivée* issue des travaux de Leibnitz ainsi que la création de la théorie de l'optimisation linéaire en 1939 par les mathématiciens Leonid Kantorovitch et Charles Koopmans. Par ailleurs, des progrès notables ayant émergé pendant la seconde guerre mondiale qui marqua le début de l'optimisation moderne, ont donné naissance à de nouvelles méthodes à l'instar de celle du simplexe pour la programmation linéaire.

Au cours des dernières décennies, l'optimisation, devenue un pilier des mathématiques, a eu d'importantes applications dans des domaines très diversifiés, notamment dans la résolution de problèmes complexes survenant en médecine, en biologie, en traitement du signal, en apprentissage automatique et en ingénierie. Pour la plupart de ces applications, la résolution du problème d'intérêt passe souvent par la construction d'un modèle mathématique identifiant les variables de décision ainsi que les paramètres du problème. Cependant, ce modèle consistant formellement en une fonction objectif soumise à des contraintes n'a toujours pas une expression analytique mais se présente également sous la forme d'une simulation informatique appelée boîte noire. En aéronautique par exemple, la conception de forme de profils aérodynamiques ou d'ailes d'avions commerciaux [82] visant à minimiser la traînée et la consommation en carburant, constitue un problème d'optimisation complexe modélisé par de coûteuses simulations ou boîtes noires. La modélisation par boîtes noires intervient également en cardiologie pédiatrique lors de l'identification de la forme d'une greffe chirurgicale dans le but de retarder autant que possible une prochaine opération [98]. Cette modélisation intervient aussi en océanographie lors de l'identification de meilleurs emplacements de bouées de détection de tsunami [16].

Dépendamment de la nature du problème, plusieurs méthodes d'optimisation peuvent être utilisées, faisant usage ou non d'informations issues des calculs de gradients, de matrices hessiennes, ou sinon des approximations de ceux-ci. Toutefois, pour la plupart des modèles décrivant des problèmes concrets comme celui résultant du tsunami de l'océan indien de 2004 [16], il est peu probable que l'utilisation d'évaluations de fonction pour l'estimation des dérivées soit utile en raison de la complexité de la simulation numérique, ou sinon ces dérivées sont simplement inaccessibles car les fonctions sont bruitées. Ainsi, comme souligné dans [16], les méthodes de recherche directe pour l'opti-

misation de boîtes noires (*blackbox optimization*, BBO) constituent un choix naturel pour la résolution du problème survenant dans des situations comme cette dernière. Plus précisément, en présence d'un bruit aléatoire et sans aucune information sur l'existence de dérivées, il faut avoir recours à des méthodes de recherche directe pour BBO stochastique, dont la conception ainsi que l'analyse de convergence font l'objet de cette thèse.

1.1 Définitions et concepts de base

1.1.1 Optimisation sans dérivée et optimisation de boîte noire

Lorsque l'information du gradient est disponible, fiable et peut être obtenue à un coût raisonnable pour un problème donné, les méthodes d'optimisation sans dérivée (*derivative-free optimization*, DFO) et de BBO ne pourront presque jamais être meilleures que les méthodes modernes à base de gradient qui devront alors être privilégiées [16].

La DFO est l'étude mathématique des algorithmes d'optimisation n'utilisant pas de dérivées [16]. Elle traite des problèmes pour lesquels les gradients existent mais sont inaccessibles. Elle se distingue de la BBO qui étudie la conception ainsi que l'analyse d'algorithmes supposant que la fonction objectif et/ou les contraintes n'admettent pas de dérivées et sont accessibles via une boîte noire. En optimisation, une boîte noire est un processus dont les mécanismes internes ne sont pas analytiquement disponibles mais qui renvoie une sortie lorsqu'on lui fournit une entrée [16]. Quoique la forme la plus courante de boîte noire soit une simulation informatique, notons que certaines expériences de laboratoires en constituent également une autre.

En BBO, lorsque les valeurs de la fonction objectif et/ou des contraintes sont accessibles uniquement via une boîte noire corrompue d'un bruit stochastique, on parle alors d'optimisation stochastique de boîtes noires.

1.1.2 Recherche directe en BBO

En raison de leur robustesse et de leur fiabilité, cette thèse se focalise particulièrement sur les méthodes de recherche directe pour l'optimisation stochastique de boîtes noires. En BBO, les méthodes de recherche directe constituent une large classe de méthodes de BBO où, à chaque itération, l'algorithme de BBO évalue la fonction objectif à un ensemble de points et agit uniquement sur la base des valeurs obtenues sans aucune construction de modèle ni approximation de dérivée [16, 42].

1.2 Éléments de la problématique

Étant donné l'utilisation croissante des outils informatiques durant ces dernières années, nombreux sont les problèmes d'optimisation comportant une fonction objectif f ainsi que des contraintes (lorsque ces dernières existent) de la forme $c_j(x) \leq 0$, où pour tout $j = 1, 2, \ldots, m$, les fonctions c_j et f sont toutes issues d'une simulation informatique ou boîte noire. Toutefois, pour certains de ces problèmes fréquemment rencontrés par exemple lors de l'optimisation d'hyperparamètres algorithmiques ou en biologie structurale lors d'une recherche de l'alignement volumétrique optimal de structures protéiques [34], les valeurs exactes de f et/ou des fonctions c_j sont impossibles d'accès et ne peuvent être obtenues que via des évaluations bruitées de la boîte noire. En d'autres termes, l'algorithme d'optimisation n'a pas accès aux valeurs de f et des c_j, mais plutôt à celles de leurs versions bruitées données par les sorties de la boîte noire, et couramment modélisées respectivement par $f_{\Theta_0}(x) = f(x) + \Theta_0$ et $c_{\Theta_j}(x) = c_j(x) + \Theta_j$, où pour tout $j = 0, 1, \ldots, m$, Θ_j est une variable aléatoire de distribution inconnue.

Les méthodes actuelles d'optimisation sans dérivées de résolution de pareils problèmes sont soit sans contraintes, ou heuristiques (sans preuve rigoureuse de convergence) ou sinon, supposent que la fonction f ainsi que les contraintes sont différentiables, puis utilisent par conséquent des approximations de gradients. Ainsi, ces méthodes sont inefficaces en contexte de boîtes noires en cas de non-différentiabilité de l'objectif ou des contraintes. De plus, il n'existe au meilleur de nos connaissances, aucune méthode de recherche directe d'optimisation stochastique de boîtes noires avec des contraintes bruitées aléatoirement disposant de preuve de convergence. De même, dans le cadre de l'optimisation sans contrainte, aucune recherche ne s'est jusqu'à présent intéressée à l'analyse de complexité de pareilles méthodes. Notons enfin que même le logiciel académique québécois NOMAD [1] [69] d'optimisation de boîte noire, présente malheureusement de grandes limitations en optimisation (stochastique) de pareilles boîtes noires bruitées.

1.3 Objectifs de recherche

Le but visé par les recherches dans cette thèse est donc le développement de méthodes de recherche directe justifiées par des analyses de convergence, pour l'optimisation stochastique de boîte noire, aussi bien en optimisation sans contrainte qu'en présence de contraintes bruitées. Plus particulièrement, la théorie de la complexité étant à l'origine d'une grande partie de l'optimisation moderne, permettant une comparaison équitable entre des méthodes numériques concurrentes, cette thèse aborde égale-

1. NOMAD est librement distribué et utilisé par des chercheurs universitaires et des entreprises de toutes tailles dont Airbus, Boeing, Exxon-Mobil, Rio Tinto et Hydro-Québec ; depuis la sortie de la troisième version en 2008, NOMAD compte plus de 4000 téléchargements vérifiés dans le monde entier.

ment, pour une première fois dans le domaine de la BBO, la question du taux de convergence de méthodes de recherche directe pour l'optimisation stochastique sans contrainte de boîte noire. Son objectif principal est traité par les trois sous-objectifs suivants :

1. Développer et analyser un algorithme de recherche directe pour l'optimisation stochastique et sans contrainte de boîtes noires.
2. Étudier le taux de convergence des méthodes de recherche directe de type directionnel pour l'optimisation stochastique sans contrainte de boîte noire.
3. Concevoir et analyser un algorithme de recherche directe pour l'optimisation stochastique sous contraintes bruitées de boîtes noires.

CHAPITRE 2 REVUE DE LITTÉRATURE

Ce chapitre résume la revue de littérature nécessaire à la réalisation de ce doctorat. Il est subdivisé en trois parties :

— Méthodes d'optimisation de boîtes noires déterministes
— Méthodes d'optimisation de boîtes noires bruitées
— Processus stochastiques à temps discret et utilité en optimisation stochastique sans dérivée

2.1 Méthodes d'optimisation de boîtes noires déterministes

On distingue deux grandes familles de méthodes en DFO : les méthodes basées sur des modèles [41] qui opèrent par construction d'approximations de la fonction objectif et des contraintes, ainsi que les méthodes de recherche directe particulièrement intéressantes en BBO, qui se basent uniquement sur la comparaison des valeurs de l'objectif et l'évaluation des contraintes pour rejeter ou retenir un point courant.

2.1.1 Méthodes de recherche directe en BBO

Dans la grande famille de ces méthodes, on distingue celles basées sur les simplexes comme l'algorithme classique de Nelder-Mead [80] qui utilise des opérations répétées de réflexions, d'expansions et de contractions appliquées à un simplexe de $n + 1$ points dans $\mathbb{R}^n$, ainsi que les algorithmes de recherche directe de type directionnel où une amélioration de la fonction objectif est garantie en se déplaçant le long d'une direction définie par un meilleur point [95]. Dans cette dernière classe de méthodes, on note la recherche par coordonnées (CS) [16], la recherche généralisée par motifs (GPS) [16] et l'algorithme de recherche directe par treillis adaptatifs (MADS) [12, 16]. Notons que MADS est l'algorithme principal dont la conception ainsi que l'analyse de variantes pour l'optimisation stochastique de boîtes noires font objet de cette thèse.

2.1.2 Algorithme MADS pour la BBO

MADS est un algorithme qui a été développé pour pallier aux difficultés rencontrées dans les problèmes d'optimisation de boîtes noires non lisses ou qui présentent des contraintes cachées. Les problèmes d'optimisation ciblés par MADS sont de la forme suivante :

$$\min_{x \in \mathcal{D}} f(x) \tag{2.1}$$

où $\mathcal{D} = \{x \in \mathcal{X} : c(x) \leq 0\} \subseteq \mathbb{R}^n$ est le domaine réalisable, $f : \mathcal{X} \mapsto \mathbb{R} \cup \{+\infty\}$ et $c : \mathcal{X} \mapsto (\mathbb{R} \cup \{+\infty\})^m$ avec $c = (c_1, c_2, \ldots, c_m)^\top$, sont des fonctions sur lesquelles aucune hypothèse de régularité n'est émise. Les contraintes, lorsqu'elles existent, sont traitées soit par l'approche de la barrière extrême (EB) [12] ou celle de la barrière progressive (PB) [13]. Cette dernière approche plus sophistiquée, et particulièrement d'intérêt dans cette thèse, utilise la fonction de violation de contraintes [16], également connue sous le nom de fonction d'agrégation, définie par :

$$h(x) = \begin{cases} \sum_{j \in J} (\max\{c_j(x), 0\})^2 & \text{si } x \in \mathcal{X}, \\ +\infty & \text{sinon,} \end{cases}$$

où $J = \{1, 2, \ldots, m\}$ représente l'ensemble des indices des contraintes. La fonction de violation des contraintes est non négative et vaut zéro si et seulement si x appartient à $\mathcal{D}$.

MADS est un algorithme de recherche itérative dans laquelle chaque itération est constituée de deux principales étapes : La "*recherche globale*" ou étape de SEARCH est celle qui peut utiliser des stratégies variées incluant l'utilisation de fonctions "*surrogates*" et des heuristiques pour explorer l'espace des variables, et la "*sonde locale*" ou étape de POLL est celle qui suit des règles plus strictes et qui effectue une exploration locale dans l'espace des variables dans une région délimitée par le "*vecteur de taille de sonde*" $\boldsymbol{\delta_p^k} \in \mathbb{R}_+^n$. En pratique, une étape de SEARCH bien conçue peut permettre à l'algorithme d'échapper à des solutions optimales locales [10] tandis que l'étape de POLL assure la convergence théorique et pratique [12].

Ces deux étapes génèrent leurs points candidats sur une discrétisation de l'espace des variables appelée maillage ou "*mesh*". La finesse de cette discrétisation est contrôlée par le "*vecteur de taille de maillage*" $\boldsymbol{\delta_m^k} \in \mathbb{R}_+^n$. Formellement, à l'itération k, le maillage est défini comme suit :

$$\mathcal{M}^k = \mathcal{V}^k + \left\{ \text{diag}(\boldsymbol{\delta_m^k}) z : z \in \mathbb{Z}^n \right\}, \tag{2.2}$$

où la "*cache*" $\mathcal{V}^k$ est l'ensemble des points visités par l'algorithme au début de l'itération k. Le premier ensemble $\mathcal{V}^0$ est l'ensemble d'un ou plusieurs points initiaux fournis par l'utilisateur. Les vecteurs des tailles de maillage et de sonde sont mis à jour à la fin de chaque itération. Les composantes des deux vecteurs sont réduites lorsque l'itération n'améliore pas la solution courante, et sinon, elles sont soit augmentées ou gardent les mêmes valeurs. L'algorithme 1 fournit une description de haut niveau de MADS.

Le résultat fondamental de convergence de MADS stipule que si toutes les suites de points tests appartiennent à un ensemble borné, et si l'ensemble des soi-disant "*directions raffinantes*" [16] est suffisamment riche, alors il existe un point d'accumulation $\hat{x}$ également appelé "*point raffinant*" tel que la dérivée généralisée $f^\circ(\hat{x}, d)$ de Clarke [38] est non négative pour toute direction d "*hypertan-*

Algorithm 1: MADS

Un ensemble de points initiaux étant donné par l'utilisateur : $\mathcal{V}^0 \in \mathbb{R}^n$,
et un maillage initial ainsi qu'un vecteur de tailles de sonde : typiquement $\boldsymbol{\delta^0_{m,i}} = \boldsymbol{\delta^0_{p,i}} = 1$ pour $i = 1, 2, \ldots, n$.
Initialiser le compteur d'itération : $k \leftarrow 0$

[1]-SEARCH (optionnel)

Lancer la boîte noire sur un ensemble fini $\mathcal{S}^k$ de points du maillage
Si échec, aller à l'étape **[3]**

[2]-POLL

Lancer la boîte noire sur un sous-ensemble fini de points de sonde

[3]-Mises à jour

Mettre à jour la cache $\mathcal{V}^{k+1}$ avec le point courant x^{k+1}
et les vecteurs des tailles de maillage et de sonde $\boldsymbol{\delta^{k+1}_m}$ et $\boldsymbol{\delta^{k+1}_p}$.
Augmenter le compteur d'itération : $k \leftarrow k + 1$ et aller à **[1]**.

Figure 2.1 Algorithme MADS

gente" [16] au domaine Ω en $\hat{x}$ pourvu que $\hat{x}$ soit réalisable. On a un résultat similaire pour h dans les situations où les itérés ne s'approchent jamais du domaine réalisable : si le point raffinant $\hat{x}$ est non réalisable, alors la dérivée généralisée de Clarke $h^\circ(\hat{x}, d)$ est non négative dans toutes les directions d hypertangentes à l'ensemble $\mathcal{X}$ en $\hat{x}$.

Ces résultats de convergence ont lieu indépendamment de l'étape SEARCH de l'algorithme 1, tant que l'ensemble $\mathcal{S}^k$ est fini et appartient au maillage $\mathcal{M}^k$.

2.2 Méthodes d'optimisation de boîtes noires bruitées

Diverses méthodes ont été conçues pour optimiser les problèmes de type boîte noire. On distingue : les méthodes de recherche directe, les méthodes de recherche aléatoire (*sample path optimization methods*), les méthodes à gradient, les méthodes utilisant des modèles, les méthodes d'optimisation Lipschitzienne, etc.

2.2.1 Méthodes de recherche directe

La recherche directe peut également être définie comme l'examen séquentiel de solutions expérimentales générées par une certaine stratégie [60].

Méthodes de recherche déterministes

Au cours des dernières années, beaucoup de recherches se sont de plus en plus intéressées aux méthodes de recherche directe pour l'optimisation sans contrainte. La plupart de ces méthodes ne font pas d'estimations de gradient et font relativement usage de peu d'évaluations de fonctions à chaque itération. La méthode NM est l'une des plus populaires [4, 25] et couramment utilisées dans cette classe. Cependant, les méthodes basées sur une approche par simplexe ont l'inconvénient de possiblement ne pas converger pour certains types de problèmes. Il existe d'ailleurs des exemples tels que celui très célèbre dû à McKinnon [16, 74] où NM converge vers un point non stationnaire. Bien que réussies dans des contextes déterministes, les applications directes de NM à l'optimisation de fonctions bruitées ont de sérieuses limitations : premièrement, NM manque d'un mécanisme efficace de sélection de taille d'échantillon pour contrôler le bruit ; Deuxièmement, c'est un algorithme heuristique [90] dont la qualité de la solution optimale estimée produite ne peut pas être quantifiée. Afin d'amener NM à optimiser des fonctions bruitées, [25] en avait considéré des variantes. Cinq années plus tard, [5] a proposé une autre variante de NM utilisant des *structures* presque similaires aux simplexes, pour l'optimisation de fonctions bruitées. Tout récemment, [31] a proposé une nouvelle variante de NM désignée sous le nom de *Stochastic Nelder-Mead* (SNM) susceptible d'optimiser les fonctions bruitées en utilisant uniquement leurs évaluations, et en a prouvé la convergence vers des optima globaux avec probabilité un. Il est cependant très important de souligner que même si ces méthodes sont fiables et robustes en pratique comme toute méthode de recherche directe [9], elles se limitent à l'optimisation sans contrainte et ne disposent d'aucun résultat de complexité. Par ailleurs, nulle d'entre elles ne dispose d'une analyse de convergence beaucoup plus adaptée au contexte de la BBO, i.e., vers des points Clarke-stationnaires.

Méthodes de recherche aléatoire

Elles constituent l'une des classes de méthodes d'optimisation de fonctions bruitées datant d'un demi-siècle [36]. Contrairement aux méthodes classiques de recherche directe déterministe [1, 2, 12, 72, 93, 94], les méthodes de recherche aléatoire ont l'avantage d'accélérer l'optimisation en utilisant des vecteurs aléatoires comme directions de recherche. Elles partagent un cadre de base simple et se sont avérées prometteuses pour résoudre des problèmes sans dérivées à grande échelle [54, 91]. Par exemple, en s'inspirant des récents travaux de Nesterov [81], [36] a proposé une approche qui établit des limites de complexité pour la convergence de méthodes sans dérivées aléatoires pour des fonctions convexes et non convexes. Par ailleurs, incorporant la technique de lissage gaussien de Nesterov [81], [54] propose une méthode de recherche aléatoire sans dérivées pour l'optimisation stochastique et montre que la complexité de l'itération de l'algorithme proposé améliore le résultat de Nesterov par un facteur

d'ordre n dans le cas lisse et convexe. En s'inspirant des travaux de Nesterov [81], [36] propose un nouvel algorithme STARS (*STepsize Approximation in Random Search*) où le choix de la longueur de pas a été grandement motivé par des travaux récents de Moré et Wild [78, 79]. Une analyse du taux de convergence de STARS pour la résolution de problèmes convexes avec des bruits stochastiques additif et multiplicatif a été effectuée et il a également été démontré que sous des hypothèses non restrictives sur le bruit, la méthode admet un taux de convergence pour les fonctions convexes bruitées, qui est identique à celui de la méthode de recherche aléatoire de Nesterov pour les fonctions convexes lisses. Malheureusement, contrairement aux méthodes de recherche directe, toutes ces méthodes de recherche aléatoire ne sont pas parfaitement adaptées au contexte de BBO car faisant usage de directions de recherche obtenues par des approximations de gradient stochastique (indisponible en BBO), qui en cas d'imprécision, peut pointer dans de très mauvaises directions.

2.2.2 Méthodes RSM (*Response Surface Methodology*)

RSM est typiquement utile dans le contexte de problèmes d'optimisation continue et se concentre sur l'apprentissage des relations *input-output* pour approximer la boîte noire bruitée sous-jacente par une surface également appelée métamodèle ou substitut, pour laquelle l'on définit une forme fonctionnelle. Cette dernière peut ensuite être utilisée dans des techniques d'optimisation basées sur l'utilisation de dérivées. Les approches RSM peuvent soit construire des modèles *surrogates* efficaces dans des régions locales, et les utiliser de façon séquentielle pour guider la recherche, ou sinon construire des *surrogates* pour l'ensemble de l'espace de recherche, puis les utiliser dans le but de trouver de meilleures solutions dans des zones d'intérêt. Ces méthodes ont pour inconvénient d'être fortement dépendantes de la qualité des surrogates utilisés et dont la construction peut parfois s'avérer difficile dans le cas de boîtes noires très coûteuses en évaluation. Par ailleurs, il n'existe au meilleur de nos connaissances aucune méthode RSM d'optimisation de boîtes noires sous contraintes bruitées, dotée d'une analyse de convergence.

Méthodes de région de confiance

Les méthodes de région de confiance [39, 42] peuvent être utilisées pour implémenter des RSM séquentiels. Les régions de confiance fournissent un moyen de contrôler la région d'approximation, fournissant des critères de mise à jour pour les modèles *surrogates*, et sont utiles dans l'analyse des propriétés de convergence. Une fois qu'un métamodèle ou surface de réponse g est construit autour d'un centre x_i de région de confiance, les algorithmes de région de confiance introduisent le sous-problème $(\min_s\ g(x_i + s) :\ s \in \mathcal{B}(x_i, \Delta))$, où $\mathcal{B}(x_i, \Delta)$ désigne la boule centrée en x_i et de rayon Δ. Il existe des critères bien définis pour mettre à jour le centre et le rayon de la région de confiance [39]

qui définiront la région d'approximation subséquente. L'utilisation de régions de confiance dans l'optimisation de fonctions issues de simulations est relativement récente et a été étudiée dans une certaine mesure [32,45]. Tout récemment, en s'inspirant de [46], [35] a utilisé une méthode de région de confiance pour proposer l'Algorithme 2, STORM (*STochastic Optimisation with Random Models*), basée sur la construction de modèles aléatoires dans un cadre d'optimisation sans dérivée et sans contrainte pour l'optimisation de boîtes noires bruitées, où la taille de la région de confiance peut être augmentée ou diminuée en fonction de la diminution empirique de la fonction observée et de la taille des gradients approximatifs observés. Cependant, ces méthodes sont beaucoup plus avantageuses en contexte de DFO qu'en BBO car faisant usage d'approximations de gradient et parfois même de matrices hessiennes [35] (non disponibles en BBO) lors de la construction de modèles. Il suffit alors que ces approximations soient médiocres pour que l'algorithme de région de confiance pointe dans une très mauvaise direction.

2.2.3 Méthodes d'approximation stochastique

Les méthodes d'approximation stochastique ou d'approche par gradients sont celles visant à chercher une direction de descente en faisant usage d'informations de "gradient estimé". Les techniques d'approximation stochastique sont l'une des plus anciennes méthodes d'optimisation d'une boîte noire bruitée. Robbins et Monro [85] ainsi que Kiefer et Wolfowitz [62] ont été les premiers à élaborer des techniques d'approximation stochastique. Dans un contexte d'optimisation sans dérivée, les gradients sont estimés à l'aide de schémas de différences finies. Typiquement, une estimation de différence directe impliquerait l'échantillonnage d'au moins $n+1$ points distincts, mais des performances supérieures ont été observées par des estimations de perturbation simultanées nécessitant des échantillons à deux points seulement [90], une méthode appelée *Simultaneous Perturbation Stochastic Approximation* (SPSA). Dans le but de trouver des directions de descente fiables, ces méthodes requièrent en général un grand nombre d'évaluations de la boîte noire pour approximer le gradient de la fonction objectif, et ne sont clairement pas avantageuses en contexte de BBO où les dérivées ne sont pas disponibles.

2.2.4 Approximation par moyenne échantillonnale

En raison de l'inaccessibilité de la fonction inconnue sous-jacente f, l'approximation par moyenne échantillonnale utilise plutôt un estimateur généralement cohérent tel que la moyenne échantillonnale d'évaluations indépendantes en un point donné de la fonction objectif bruitée accessible $\tilde{f}$. À titre d'exemple, on peut durant l'optimisation, utiliser $F_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{f}(x;\eta_i)$ au lieu de la fonction $\mathbb{E}_\eta[\tilde{f}(x;\eta)] = f(x)$ elle-même, où $\eta_1, \eta_2, \ldots, \eta_n$ forment un échantillon aléatoire de taille n de η.

Certains des premiers travaux utilisant cette technique sont ceux de [58, 88]. Plusieurs autres travaux, comme ceux dans [45, 84], discutent des résultats de convergence des algorithmes dans ce contexte. Cependant, notons que l'analyse de ces méthodes reposent fortement sur la façon dont les estimés utilisés sont générés (e.g., la taille n de l'échantillon définissant F_n, l'indépendance des variables aléatoires $\eta_i, i = 1, 2, \ldots, n$), contrairement aux travaux présentés dans la présente thèse où aucune hypothèse de ce genre n'a été émise pour la convergence des diverses méthodes proposées.

2.3 Processus stochastiques à temps discret et utilité en optimisation stochastique sans dérivée

Nombreux sont les domaines utilisant des observations en fonction du temps et qui se traduisent par une courbe bien définie dans les cas les plus simples. L'interprétation de ces observations se présentant dans certaines situations de façon plus ou moins erratique, est donc soumise à une certaine incertitude pouvant être traduite par l'usage de processus stochastiques ou aléatoires pour les représenter. L'utilisation de ces processus a suscité un regain d'intérêt durant la dernière décennie, notamment dans la modélisation du comportement ainsi que l'analyse d'algorithmes de DFO utilisant un cadre aléatoire [24, 29, 35, 68, 83]. En effet, l'Algorithme 2, STORM, fut proposé dans [34, 35] dans le but de résoudre le problème sans contrainte

$$\min_{x \in \mathbb{R}^n} f(x) \tag{2.3}$$

où les valeurs de la fonction f supposée continue ne sont accessibles numériquement que via la version bruitée

$$\tilde{f}(x) = f(x, \varepsilon), \tag{2.4}$$

ε étant une variable aléatoire de distribution inconnue. Plus précisément, étant donné que les valeurs de f en un point courant x_k et en un point voisin $x_k + s_k$ sont inconnues, STORM utilise alors des estimés $f_k^0 \approx f(x_k)$ et $f_k^s \approx f(x_k + s_k)$ ainsi que des modèles m_k, construits à base d'observations aléatoires de l'objectif bruité $\tilde{f}$. Ces estimés et modèles sont en effet des réalisations d'estimés aléatoires F_k^0 et F_k^s, et de modèles M_k aléatoires, dont les comportements influent sur chaque itération k de l'algorithme qui résulte par conséquent en un processus stochastique $\{M_k, X_k, S_k, \Delta_k, F_k^0, F_k^s\}$. En attendant la présentation de la notion de variable aléatoire à la Section 2.3.1, notons que dans le processus précédent, X_k, S_k et Δ_k désignent respectivement des variables aléatoires de réalisations respectives x_k, s_k et du rayon de confiance δ_k. Notons enfin que dans l'Algorithme 2, $m_k(x)$ désigne plus précisément un modèle servant d'approximation de $f(x)$ dans la boule $\mathcal{B}(x_k, \delta_k)$. Par ailleurs, en s'inspirant des travaux de [34, 35], Paquette et Scheinberg ont proposé dans [83] l'Algorithme 3 dans le but de résoudre des problèmes d'optimisation stochastique comme ceux donnés par (2.3). La méthode proposée est une variante stochastique d'une méthode de recherche linéaire classique due à Armijo [8], et qui a été analysée au moyen de la théorie des processus stochastiques en s'inspirant

Algorithm 2: STORM

0-Initialisation

Choisir un point initial x_0 et un rayon de confiance initial $\delta_0 \in (0, \delta_{\max})$ où $\delta_{\max} > 0$

Choisir des constantes $\gamma > 1, \eta_1 \in (0,1), \eta_2 > 0$ et $\kappa_{fcd} \in (0,1]$

$k \leftarrow 0$

1-Construction de modèle

Construire un modèle $m_k(x^k + s) = f_k + g_k^\top s + s^\top H_k s$ qui approxime $f(x)$ sur la boule $\mathcal{B}(x_k, \delta_k)$, où $s = x - x_k$

2-Calcul de longueur de pas

Calculer approximativement $s_k = \underset{s:\|s\| \leq \delta_k}{\arg\min}\ m_k(s)$ de sorte que

$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{fcd}}{2} \|g_k\| \min\left\{\frac{\|g_k\|}{\|H_k\|}, \delta_k\right\}$

3-Calcul d'estimés

Calculer les estimés respectifs f_k^0 et f_k^s de $f(x_k)$ et $f(x_k + s_k)$

4-Acceptation du point test

Calculer $\rho_k = \frac{f_k^0 - f_k^s}{m_k(x_k) - m_k(x_k + s_k)}$

Si $\rho_k \geq \eta_1$ et $\|g_k\| \geq \eta_2 \delta_k$, alors $x_{k+1} = x_k + s_k$

Sinon $x_{k+1} = x_k$

5-Mise à jour du rayon de confiance

Si $\rho_k \geq \eta_1$ et $\|g_k\| \geq \eta_2 \delta_k$, alors $\delta_{k+1} = \min\ \{\gamma \delta_k, \delta_{\max}\}$

Sinon $\delta_{k+1} = \gamma^{-1} \delta_k$, $k \leftarrow k + 1$ et aller à **1**

Figure 2.2 Algorithme STORM pour l'optimisation stochastique sans dérivée

notamment de l'analyse de convergence de STORM et des travaux présentés dans [24, 29, 30]. En effet, une étude de complexité de l'Algorithme 3 a été effectuée après avoir démontré que ce dernier génère également un processus stochastique $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$ où G_k désigne un estimé aléatoire de gradient, Δ_k un paramètre de contrôle, et $\mathcal{A}_k$ une longueur de pas. Il est enfin très important d'insister sur le fait que les cadres algorithmiques de toutes les méthodes proposées dans cette thèse ainsi que leur analyse de convergence se sont fortement inspirés de ceux de MADS, de STORM, de l'Algorithme 3 et des recherches menées dans [24, 29, 68]. De plus, toutes les méthodes proposées n'utilisent ni modèle ni information de gradient.

2.3.1 Notions de base de probabilité générale

L'étude des propriétés de convergence des processus stochastiques $\{M_k, X_k, S_k, \Delta_k, F_k^0, F_k^s\}$ et $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$ par le biais de la théorie des martingales a précisément permis d'obtenir celles de STORM et de l'Algorithme 3. Les différentes notions suivantes sont donc introduites dans le but de présenter aux prochaines sections les martingales également utilisées pour les analyses de

Algorithm 3: Méthode de recherche linéaire stochastique

0-Initialisation

Choisir les constantes $\gamma > 1$, $\theta \in (0,1)$ et $\alpha_{\max} > 0$. Choisir un point initial x_0, $\alpha_0 = \gamma^{j_0}\alpha_{\max}$ pour un certain $j_0 \leq 0$ et $\delta_0 > 0$.

Répéter pour $k = 0, 1, \ldots$

1-Calcul d'estimé du gradient

En se basant au point x_k, construire un estimé g_k du gradient satisfaisant l'Hypothèse 1. Poser $s_k = -\alpha_k g_k$.

2-Calcul d'estimés de fonction

Basé sur δ_k, g_k et x_k, obtenir des estimés f_k^0 et f_k^s de $f(x_k)$ et $f(x_k + s_k)$, respectivement, satisfaisant l'Hypothèse 1.

3-Succès

Si $f_k^s \leq f_k^0 - \alpha_k\theta \left\|g_k\right\|^2$, alors $x_{k+1} = x_k + s_k$ et $\alpha_{k+1} = \min\left\{\alpha_{\max}, \gamma\alpha_k\right\}$

Fiabilité du pas : Si $\alpha_k \left\|g_k\right\|^2 \geq \delta_k^2$, alors $\delta_{k+1}^2 = \gamma\delta_k^2$.

Non fiabilité du pas : Si $\alpha_k \left\|g_k\right\|^2 < \delta_k^2$, alors $\delta_{k+1}^2 = \gamma^{-1}\delta_k^2$.

4-Échec

Sinon, $x_{k+1} = x_k$, $\alpha_{k+1} = \gamma^{-1}\alpha_k$ et $\delta_{k+1}^2 = \gamma^{-1}\delta_k^2$.

convergence des diverses méthodes proposées dans cette thèse.

Notons tout d'abord qu'une expérience aléatoire est une épreuve que l'on peut en principe répéter autant de fois que l'on veut et telle que l'ensemble de tous les résultats possibles est connu, mais telle qu'on ne peut pas prévoir avec certitude lequel de ces résultats sera obtenu. Il peut par exemple s'agir de tirages d'urnes, de lancers de pièces de monnaies ou de dés, d'une durée de vie, etc. L'univers ou l'espace échantillon ou l'espace des observables Ω, est l'ensemble de tous les résultats possibles d'une expérience aléatoire. Tout sous-ensemble de Ω est appelé événement. Lorsque ce dernier se réduit à un singleton $\{\omega\} \subset \Omega$, alors il est qualifié d'élémentaire ou simple.

Toutes les recherches menées dans [26,29,30,35,68,83,96] se sont en effet fortement inspirée de [24] où quoique utilisant des modèles aléatoires pour la résolution de problèmes sans contrainte comme celui du (2.3), il a été supposé que les valeurs de l'objectif peuvent être calculées exactement, i.e., sans bruit aléatoire. L'analyse de la méthode de région de confiance proposée qui génère également des modèles aléatoires M_k, repose cependant sur le fait que même si la précision de ces derniers peut dépendre de l'historique, $(M_1, \ldots, M_{k-1})$, elle est suffisamment bonne avec une certaine probabilité d'au moins p, indépendamment de cet historique. En d'autres termes, ces modèles ont été supposés précis avec une probabilité p assez grande mais fixe. Dans [24], ainsi que [26, 29, 30, 35, 68, 83, 96], un pareil historique des différents algorithmes proposés à une itération k donnée a été modélisé par l'introduction d'une *tribu* qui contient tous les événements antérieurs au début de l'itération k. En mots simples, cette tribu correspond à tout ce qu'on sait du parcours de l'algorithme concerné jusqu'à

l'itération k. La notion de tribu également utile aux diverses analyses proposées dans la présente thèse est introduite dans la définition ci-après tirée de [52].

Définition 1. *On dit qu'une famille $\mathcal{G}$ de parties de Ω est une* tribu *ou σ-algèbre si elle est stable par réunion dénombrable et par passage au complémentaire, et contient l'ensemble vide $\emptyset$. Le couple $(\Omega, \mathcal{G})$ est appelé* espace probabilisable *(ou* mesurable*).*

Il importe à présent de souligner qu'un calcul de probabilité ne peut avoir lieu que dans le cadre d'un espace probabilisé plus ou moins bien précisé [52]. Ces notions de probabilité et d'espace probabilisé peuvent maintenant rigoureusement être définies comme suit :

Définition 2. *Une* probabilité *sur l'espace probabilisable $(\Omega, \mathcal{G})$ est une fonction $\mathbb{P}$ définie de $\mathcal{G}$ à valeurs dans l'intervalle $[0, 1]$ telle que :*

1. $\mathbb{P}(\Omega) = 1$
2. $\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} B_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(B_n)$ *pour toute suite $\{B_n\}_{n\in\mathbb{N}}$ d'événements deux à deux disjoints : c'est la propriété d'additivité dénombrable.*

Le triplet $(\Omega, \mathcal{G}, \mathbb{P})$ est appelé espace probabilisé *ou* espace de probabilité. *Les événements de probabilité 1 sont dits* presque sûrs *tandis que ceux de probabilité nulle sont dits* négligeables.

Proposition 1. *Soit $(\Omega, \mathcal{G}, \mathbb{P})$ un espace probabilisé. On a les propriétés suivantes :*

1. *$\mathbb{P}$ est croissante au sens de l'inclusion : $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$, pour tous $A, B \in \mathcal{G}$.*
2. *Toute intersection dénombrable d'événements presque sûrs est presque sûre.*

La notion d'indépendance étant au cœur des analyses présentées dans ce manuscrit, elle se définit de la façon suivante [52] dans le cas des événements en attendant le cas général des variables aléatoires qui seront présentées ensuite.

Définition 3. *Soit $(\Omega, \mathcal{G}, \mathbb{P})$ un espace probabilisé. Des événements $A_i \in \mathcal{G}$ avec $i = 1, 2, \ldots, n$ sont dits indépendants lorsque*

$$\mathbb{P}\left(\bigcap_{\ell=1}^{n} A_{i_\ell}\right) = \prod_{\ell=1}^{n} \mathbb{P}(A_{i_\ell}) \quad \textit{pour tout } 1 \leq i_1 < \cdots < i_k \leq n.$$

Variables aléatoires

La définition suivante d'application mesurable est utile pour l'introduction de celle d'une variable aléatoire.

Définition 4. *Soient $(\Omega, \mathcal{G})$ et $(E, \mathcal{C})$ deux espaces mesurables. Une application f de Ω à valeurs dans E est dite $\mathcal{C}$-mesurable (ou mesurable lorsqu'il n'y a pas d'ambigüité) si*

$$\forall C \in \mathcal{C}, f^{-1}(C) := \{\omega \in \Omega : f(\omega) \in C\} \in \mathcal{G}.$$

Définition 5. *Soit $(\Omega, \mathcal{G}, \mathbb{P})$ un espace probabilisé et $(E, \mathcal{C})$ un espace mesurable. Une* variable aléatoire *X est une application mesurable de $(\Omega, \mathcal{G}, \mathbb{P})$ vers $(E, \mathcal{C})$. Lorsque $(E, \mathcal{C}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ où $\mathcal{B}(\mathbb{R}^d)$ désigne la tribu borélienne de $\mathbb{R}^d$, i.e., celle engendrée par les ouverts, alors l'application X est qualifiée de* vecteur aléatoire *multidimensionnel (ou de variable aléatoire réelle lorsque $d = 1$).*

Définition 6. *Soit $p \in [1, +\infty)$ un entier et $(\Omega, \mathcal{G}, \mathbb{P})$ un espace probabilisé. L'espace $\mathbb{L}^p(\Omega, \mathcal{G}, \mathbb{P})$ de variables aléatoires dites p-intégrables est l'ensemble de toutes les variables aléatoires réelles X telles que*

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

Définition 7. *Soit X une variable aléatoire réelle positive ou intégrable définie sur un espace probabilisé $(\Omega, \mathcal{G}, \mathbb{P})$. L'intégrale*

$$\int X d\mathbb{P} := \int_\Omega X(\omega) \mathbb{P}(d\omega)$$

est appelée espérance mathématique *de X et est notée $\mathbb{E}(X)$.*

Convergence de suites de variables aléatoires

Bon nombre de résultats de convergence d'algorithmes d'optimisation stochastique sans dérivée ont été établis au sens *presque sûr*. Par exemple, [31] a démontré la convergence presque sûre de la méthode de NM stochastique vers des optima globaux. Plusieurs résultats de convergence presque sûre peuvent également être trouvés dans [5,24,35,83]. Notons en particulier que l'analyse de convergence de STORM qui suppose la différentiabilité de l'objectif f a permis d'établir l'existence d'une suite d'itérés $\{X_k\}_{k\in\mathbb{N}}$, pour laquelle la convergence de la suite $\|\nabla f(X_k)\|_{k\in\mathbb{N}}$ a lieu presque sûrement. Avant de présenter cette notion de convergence presque sûre qui a été utilisée dans toutes les analyses du présent manuscrit, précisons qu'il s'agit d'une convergence très forte en ce sens qu'elle implique aussi bien la convergence en probabilité que celle en loi.

Définition 8. *Soient $\{X_n\}_{n\in\mathbb{N}}$ une suite de variables aléatoires définies sur un même espace probabilisé $(\Omega, \mathcal{G}, \mathbb{P})$. On dit que la suite $\{X_n\}_{n\in\mathbb{N}}$ converge presque sûrement vers la variable aléatoire X définie sur $(\Omega, \mathcal{G}, \mathbb{P})$, et on note $\lim_{n\to+\infty} X_n = X \ \ p.s.$, lorsque*

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to+\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

Le théorème classique suivant, dit de convergence monotone [28, 50], permet d'intervertir la limite et l'intégrale et a été d'une grande utilité dans l'analyse de convergence de STOMADS-PB. Il est en particulier utilisé pour démontrer le lemme de Fatou [28, 50] ainsi que le théorème classique de la convergence dominée de Lebesgue [28, 50].

Theorème 1. *(Convergence monotone)*
Soit $\{X_n\}_{n\in\mathbb{N}}$ *une suite presque sûrement croissante de variables aléatoires réelles positives. Alors,*

$$\lim_{n\to+\infty}\mathbb{E}\left(X_n\right)=\mathbb{E}\left(\lim_{n\to+\infty}X_n\right).$$

2.3.2 Espérance conditionnelle et processus stochastiques à temps discret

Définition 9. *Soit* X *une variable aléatoire intégrable définie sur un espace probabilisé* $(\Omega,\mathcal{G},\mathbb{P})$ *et soit* $\mathcal{F}$ *une sous-tribu de* $\mathcal{G}$*.* X *n'est à priori pas* $\mathcal{F}$*-mesurable. On appelle* espérance conditionnelle *de* X *sachant* $\mathcal{F}$*, l'unique variable aléatoire* $\mathcal{F}$*-mesurable* $Z=\mathbb{E}\left(X|\mathcal{F}\right)$ *telle que pour toute variable aléatoire* $\mathcal{F}$*-mesurable bornée* Y*, on a*

$$\mathbb{E}\left(XY\right)=\mathbb{E}\left(ZY\right).$$

L'un des travaux dont s'est fortement inspirée l'analyse du taux de convergence des algorithmes SDDS de recherche directe de type directionnel présentés au Chapitre 5 est [83], portant sur l'analyse de taux de convergence de l'Algorithme 3. L'analyse de ce dernier repose notamment sur "l'hypothèse clé" [83] suivante, sur la nature de l'information stochastique dans la méthode proposée, et dont des similaires ont été utilisées dans la présente thèse.

Hypothèse 1. *On a les propriétés suivantes pour les quantités aléatoires de l'Algorithme 3 :*

(i) Il existe des constantes $p_g\in(0,1]$ *suffisamment grand et* $\kappa_g>0$ *telles que la suite* $\{G_k\}_{k\in\mathbb{N}}$ *de directions aléatoires satisfait*

$$\mathbb{P}\left(\left\{\|G_k-\nabla f(X_k)\|\leq\kappa_g\mathcal{A}_k\|G_k\|\right\}|\mathcal{F}_{k-1}^{G\cdot F}\right)\geq p_g$$

(ii) Il existe des constantes $p_f\in(0,1]$ *suffisamment grand et* $\varepsilon_f\leq\frac{\theta}{4\alpha_{\max}}$ *telles que la suite* $\{F_k^0,F_k^s\}_{k\in\mathbb{N}}$ *d'estimés satisfait*

$$\mathbb{P}\left(\left\{\left|F_k^0-f(x_k)\right|\leq\varepsilon_f\mathcal{A}_k^2\|G_k\|^2\right\}\cap\left\{|F_k^s-f(x_k+s_k)|\leq\varepsilon_f\mathcal{A}_k^2\|G_k\|^2\right\}|\mathcal{F}_{k-1/2}^{G\cdot F}\right)\geq p_f$$

(iii) La suite $\{F_k^0,F_k^s\}_{k\in\mathbb{N}}$ *d'estimés aléatoires satisfait les conditions de variance suivantes pour*

une certaine constante $\kappa_f > 0$ *:*

$$\mathbb{E}\left(\left|F_k^s - f(X_k + S_k)\right|^2 |\mathcal{F}_{k-1/2}^{G\cdot F}\right) \leq \max\left\{\kappa_f^2 \mathcal{A}_k^2 \left\|\nabla f(X_k)\right\|^4, \theta^2 \Delta_k^4\right\}$$

et $$\mathbb{E}\left(\left|F_k^0 - f(X_k)\right|^2 |\mathcal{F}_{k-1/2}^{G\cdot F}\right) \leq \max\left\{\kappa_f^2 \mathcal{A}_k^2 \left\|\nabla f(X_k)\right\|^4, \theta^2 \Delta_k^4\right\}$$

Les propriétés suivantes tirées de [28] sont très utiles pour les analyses de convergence proposées.

Proposition 2. *Soit* $(\Omega, \mathcal{G}, \mathbb{P})$ *un espace probabilisé,* $\mathcal{F}$ *et* $\mathcal{H}$ *des sous-tribus de* $\mathcal{G}$*, et* $X, Y \in \mathbb{L}^1(\Omega, \mathcal{G}, \mathbb{P})$*. On a les propriétés suivantes* $\mathbb{P}$*-presque sûrement :*

1. $\mathbb{E}\left[\mathbb{E}(X|\mathcal{F})\right] = \mathbb{E}(X)$.
2. *Si* X *est* $\mathcal{F}$*-mesurable, alors* $\mathbb{E}(X|\mathcal{F}) = X$.
3. $\mathbb{E}(aX + bY|\mathcal{F}) = a\mathbb{E}(X|\mathcal{F}) + b\mathbb{E}(Y|\mathcal{F})$*, pour tout* $a, b \in \mathbb{R}$.
4. *Si* $X \leq Y$ $p.s.$*, alors* $\mathbb{E}(X|\mathcal{F}) \leq \mathbb{E}(Y|\mathcal{F})$.

Proposition 3. *(Inégalité de Cauchy-Schwarz conditionnelle)*
Soit X *et* Y *des variables aléatoires de carrés intégrables définies sur un espace probabilisé* $(\Omega, \mathcal{G}, \mathbb{P})$*, et* $\mathcal{F}$ *une sous-tribu de* $\mathcal{G}$*. Alors,*

$$\left|\mathbb{E}(XY|\mathcal{F})\right|^2 \leq \mathbb{E}\left(X^2|\mathcal{F}\right)\mathbb{E}\left(Y^2|\mathcal{F}\right).$$

Processus stochastiques à temps discret

La théorie des processus stochastiques à temps discret est au cœur des analyses faites dans cette thèse. Son utilisation en optimisation stochastique sans dérivée ne date cependant pas d'aujourd'hui. En effet, Anderson et Ferris avaient utilisé la théorie des Chaînes de Markov [50] dans le but d'étudier la convergence d'un algorithme de recherche directe proposé dans [5], et qui utilise un cadre algorithmique similaire à celui de NM. L'analyse de convergence de STORM et de l'Algorithme 3 repose par contre sur la théorie des martingales. Des outils introduits dans [29] basés sur les surmartingales ont particulièrement permis d'étudier l'analyse du taux de convergence de STORM, et par la suite celle de l'Algorithme 3 dans [83]. Dans ces deux dernières recherches, les différents taux de complexité ont précisément été obtenus en majorant les temps d'arrêts associés aux processus stochastiques générés par les divers algorithmes.

La théorie élémentaire des martingales, fournissant sans grands efforts des résultats très profonds dont les implications sont nombreuses dans toute la théorie de probabilité, est due à J.L. Doob qui l'a élaborée vers le milieu du vingtième siècle. Dans les définitions ci-après tirées de [52] (voir également [50] pour la Définition 11), l'espace probabilisé $(\Omega, \mathcal{G}, \mathbb{P})$ est muni d'une suite croissante (au sens de l'inclusion) $\{\mathcal{G}_n\}_{n\in\mathbb{N}}$ de sous-tribus de $\mathcal{G}$, appelée *filtration*.

Définition 10. *Un processus stochastique à temps discret est une famille* $\{X_k\}_{k\in\mathbb{N}}$ *de variables aléatoires définies sur le même espace probabilisé.*

Définition 11. *Une suite* $\{X_n\}_{n\in\mathbb{N}}$ *de variables aléatoires réelles intégrables est une* surmartingale *si :*

1. *Elle est adaptée, i.e.,* X_n *est* $\mathcal{G}_n$*-mesurable pour tout* $n \in \mathbb{N}$.
2. $\mathbb{E}\left(X_{n+1}|\mathcal{G}_n\right) \leq X_n$ *p.s., pour tout* $n \in \mathbb{N}$.

La suite $\{X_n\}_{n\in\mathbb{N}}$ *est une* sousmartingale *si la suite* $\{-X_n\}_{n\in\mathbb{N}}$ *est une surmartingale, i.e. qu'elle est adaptée et satisfait* $\mathbb{E}\left(X_{n+1}|\mathcal{G}_n\right) \geq X_n$ *p.s., pour tout* $n \in \mathbb{N}$. $\{X_n\}_{n\in\mathbb{N}}$ *est une* martingale *lorsqu'elle est à la fois une surmartingale et une sousmartingale, i.e. qu'elle est adaptée et* $\mathbb{E}\left(X_{n+1}|\mathcal{G}_n\right) = X_n$ *p.s., pour tout* $n \in \mathbb{N}$.

Définition 12. *Un* temps d'arrêt T *est une variable aléatoire à valeurs dans* $\mathbb{N} \cup \{+\infty\}$ *telle que* $\{T \leq n\} \in \mathcal{G}_n$ *pour tout* $n \in \mathbb{N}$.

Notons enfin que dans les analyses de complexité de STORM présentée dans [29] et de l'Algorithme 3 [83], en notant par T_ε le temps d'arrêt défini par

$$T_\varepsilon = \inf\left\{k \geq 0 : \left\|\nabla f(X^k)\right\| \leq \varepsilon\right\},$$

il a été démontré que

$$\mathbb{E}\left(T_\varepsilon\right) \leq \mathcal{O}(1) \times \frac{p}{2p-1} \times \frac{1}{\varepsilon^2} + 1,$$

où $\varepsilon > 0$ et $p \in (0,1)$ désigne une probabilité.

CHAPITRE 3 ORGANISATION DE LA THÈSE

Cette thèse est rédigée par articles. Elle incorpore trois articles qui sont respectivement reportés dans les chapitres 4, 5 et 6.

La rédaction de la thèse est organisée comme suit. Après l'introduction présentée au Chapitre 1, une revue critique de littérature situant les contributions de la thèse a été présentée au Chapitre 2. Le Chapitre 4 présente les premières contributions de la thèse, notamment la méthode de recherche directe d'optimisation stochastique sans contrainte StoMADS ainsi que son analyse de convergence. Ces travaux ont été acceptés pour publication dans la revue *Computational Optimization and Applications (COAP)* [15]. Le cadre algorithmique de StoMADS a été généralisé au Chapitre 5 par l'introduction d'une large classe de méthodes de recherche directe dites de type directionnel (SDDS) pour l'optimisation stochastique et sans contrainte de boîtes noires. Le taux de convergence de SDDS a ensuite été présenté et les travaux de ce chapitre ont été soumis pour publication dans la revue *Computational Optimization and Applications (COAP)* [51]. Le Chapitre 6 se concentre sur l'optimisation stochastique sous contraintes bruitées aléatoirement, en introduisant l'algorithme de recherche directe StoMADS-PB qui traite les contraintes par l'approche de la barrière progressive. Une analyse de convergence de StoMADS-PB également été proposée et tous les travaux de ce chapitre ont été soumis pour publication dans la revue *Mathematical Programming*. Le Chapitre 7 présente une discussion générale sur les contributions des trois articles de la thèse, suivie d'une conclusion générale au Chapitre 8.

CHAPITRE 4 ARTICLE 1: STOCHASTIC MESH ADAPTIVE DIRECT SEARCH FOR BLACKBOX OPTIMIZATION USING PROBABILISTIC ESTIMATES

Charles Audet, Kwassi Joseph Dzahini, Michael Kokkolaras and Sébastien Le Digabel. Stochastic Mesh Adaptive Direct Search for Blackbox Optimization Using Probabilistic Estimates. Accepted for publication in Computational Optimization and Applications (COAP).

Abstract: We present a stochastic extension of the mesh adaptive direct search (MADS) algorithm originally developed for deterministic blackbox optimization. The algorithm, called StoMADS, considers the unconstrained optimization of an objective function f whose values can be computed only through a blackbox corrupted by some random noise following an unknown distribution. The proposed method is based on an algorithmic framework similar to that of MADS and uses random estimates of function values obtained from stochastic observations since the exact deterministic computable version of f is not available. Such estimates are required to be accurate with a sufficiently large but fixed probability and to satisfy a variance condition. The ability of the proposed algorithm to generate an asymptotically dense set of search directions is then exploited using martingale theory to prove convergence to a Clarke stationary point of f with probability one.

Keywords: Blackbox optimization, derivative-free optimization, stochastic optimization, mesh adaptive direct search, probabilistic estimates.

4.1 Introduction

Blackbox optimization (BBO) considers the development and analysis of algorithms under the assumption that the objective and/or constraint functions are provided by blackboxes, i.e., "*any computational process whose inner workings are analytically unavailable and which returns an output when provided an input*" [16]. Mesh adaptive direct search (MADS) is an algorithm for deterministic BBO [12].

We consider the unconstrained stochastic blackbox optimization problem

$$\min_{x\in\mathbb{R}^n} f(x), \quad \text{with} \quad f(x) = \mathbb{E}_\Theta\left[f_\Theta(x)\right] \tag{4.1}$$

where Θ is a random variable obeying some unknown distribution, $\mathbb{E}_\Theta$ denotes the expectation with respect to Θ, and $f_\Theta(x)$ [1] denotes a stochastic blackbox, i.e., the noisy computable version of a numerically unavailable objective function $f\colon \mathbb{R}^n \to \mathbb{R}$. We assume that the deterministic objective function is locally Lipschitz continuous and bounded from below. Such problems arise in modern statistical machine learning, where the random variable Θ represents a data point drawn according to some unknown distribution and $f_\Theta(x)$ measures the fit of some model parameter x to the data point Θ [23, 44, 66].

Developing provable algorithms to solve the problem in Eq. (4.1) has been the goal of intense research efforts. Several methods have been proposed in recent years, most of which are extensions of existing deterministic derivative-free optimization (DFO) methods [16, 42] to stochastic functions [22,31,35,68,83,89]. Such methods are classified according to Angün and Kleijnen as whitebox and blackbox methods [6]. Whitebox methods rely on the ability to compute the gradient of f with a single simulation [53]. On the contrary, blackbox methods, such as the stochastic approximation method [5, 62], response surface methodology [4], and many heuristics [4], need several function evaluations to approximate the gradient of the objective function.

Direct search methods are promising candidates for blackbox optimization as they are considered to be robust and reliable in practice [9]. It is important to emphasize that their analysis does not assume the existence of derivatives; consequently, no gradient approximations are attempted. Existing deterministic direct search blackbox optimization methods that have been extended to stochastic functions include the Nelder-Mead (NM) method [80]. Barton and Ivey were among the first to propose a variant of the NM algorithm considering noisy function evaluations [25]. Anderson and Ferris also considered the unconstrained optimization of functions subject to random noise [5]. They used an

1. The notations $\tilde{f}(x,\xi)$ [29], $\tilde{f}(x;\xi)$ [83] and $f(x;\varepsilon)$ [35] are often used for the noisy computable versions of f, where ξ and ε are random variables. We use the more compact notation $f_\Theta(x)$.

algorithmic framework similar to that of NM, making use of so-called *structures* instead of simplices and proposed an algorithm involving reflection, expansion and contraction steps, which generates a sequence of iterates converging to a cluster point with probability one based on the theory of Markov chains [50]. Chang proposed another stochastic variant of the classic NM method which replaces the shrink step by an *adaptive random search* and proved convergence to global optima with probability one [31].

Audet et al. recently proposed Robust-MADS, a kernel smoothing-based variant of the MADS algorithm, designed to approach the minimizer of an objective function when only having access to noisy function values [18]. At each iteration of Robust-MADS, the incumbent solution is determined based on values of a smoothed version of the noisy objective function constructed from a list of trial points. This list is eventually updated with the best iterate found before the next iteration of the algorithm, which is shown to possess zero-order convergence properties [14]. However, while this method produces good results when applied to a variety of problems, including problems with granular and discrete variables [20], problems involving random noise have not been considered. Moreover, Robust-MADS is a deterministic algorithm in the sense that it uses only deterministic algorithmic parameters and information (e.g., mesh and frame size parameters and smoothed function values).

[3] considered the optimization of functions that are numerically unavailable and whose values can only be computed through a blackbox corrupted by Gaussian random noise. Using an algorithmic framework similar to that of MADS, the algorithm proposed in [3] aims at minimizing such unknown functions by adaptively driving to zero the standard deviation of the estimators of the unavailable function values, making use of statistical inference techniques. However, even though this algorithm is shown to have desirable convergence properties, it needs to be improved since obtaining satisfactory solutions in practice requires a large number of blackbox evaluations, thus making the method computationally expensive.

In this paper, we present a stochastic extension of MADS using elements from [12, 24, 35, 83]. This algorithm, which we name StoMADS, can guarantee convergence to a Clarke stationary point provided that certain conditions are satisfied. More precisely, we assume that the function estimates used to ensure improvement are sufficiently accurate with a fixed probability that does not have to be equal to one but needs to be larger than a certain constant [35, 83]. While we assume that the estimates satisfy a condition on the variance [83] that will be specified later, no other assumption is made about the way they are generated.

The main novelty of the present work is that no model or gradient information is needed to find descent directions, as opposed to [35, 83], and [96]. We use direct search techniques and exploit the ability of the proposed algorithm to generate an asymptotically dense set of search directions to

guarantee convergence. To the best of our knowledge, this research is the first to propose a stochastic variant of MADS with fully supported convergence results obtained using martingale theory.

This manuscript is organized as follows. Section 4.2 introduces the general framework of the proposed stochastic algorithm and discusses the requirements on random estimates to guarantee convergence in addition to how such estimates can be obtained in practice. Section 4.3 presents the main convergence results. Computational results are reported in Section 4.4, followed by a discussion and suggestions for future work.

4.2 The StoMADS algorithm and probabilistic estimates

This section presents the general framework of StoMADS, introduces probabilistic estimates, and shows how to construct these estimates.

4.2.1 The StoMADS algorithm

Similarly to MADS, StoMADS is an iterative algorithm where each iteration comprises two steps: an optional SEARCH step, which typically consists of a global exploration based on various strategies including surrogates and heuristics, and a local POLL step, which performs a local exploration in a subset of the space of variables called the *frame*. During each of these two steps, a finite number of trial points is generated on a discretization of the space of variables called the *mesh*. The discretizations of the mesh and the frame are controlled by the mesh and frame size parameters δ_m^k and δ_p^k, respectively. Note that we are deviating from the notation δ^k and Δ^k used in [16] because Δ_m^k and Δ_p^k will be used here to denote random variables.

Let $\mathbf{D} \in \mathbb{R}^{n\times p}$ be a matrix whose columns form a positive spanning set $\mathbb{D}$. The mesh $\mathcal{M}^k$ and the frame $\mathcal{F}^k$ are defined as

$$\mathcal{M}^k := \{x^k + \delta_m^k d : d = \mathbf{D}y,\ y \in \mathbb{Z}^p\} \quad \text{and} \quad \mathcal{F}^k := \{x \in \mathcal{M}^k : \left\|x - x^k\right\|_\infty \leq \delta_p^k b\},$$

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

At iteration k, given an incumbent solution $x^k \in \mathcal{M}^k$, the StoMADS algorithm seeks to find a better mesh point $y = x^k + \delta_m^k d$ whose objective function value is less than the current unknown incumbent value $f(x^k)$, i.e., $f(y) < f(x^k)$. We use f_0^k and f_s^k to denote the estimates of $f(x^k)$ and $f(x^k + s^k)$ (with $s^k = \delta_m^k d$), respectively, obtained using evaluations of the noisy blackbox f_Θ. These estimates are then compared as specified below to determine whether a trial point $x^k + s^k$ is a better mesh point; therefore, they need to be sufficiently accurate. The following definition is adapted from [35].

Definition 1. *Let $\varepsilon_f > 0$ be a fixed constant and f_x be an estimate of $f(x)$. Then f_x is said to be an ε_f-accurate estimate of $f(x)$ for a given δ_p^k, if*

$$|f_x - f(x)| \leq \varepsilon_f (\delta_p^k)^2.$$

Note that, unlike in [35, 96], ε_f does not play a crucial role in the convergence analysis but allows to adjust the initial amplitude of the so-called uncertainty interval $\mathcal{I}_{\gamma,\varepsilon_f}(\delta_p^k)$ that will be introduced later. The next result provides information on how to determine the iteration type.

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

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

Proof. The proof is immediate using Definition 1 and observing that

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

□

The following definition distinguishes three types of iterations.

Definition 2. *Let f_0^k and f_s^k be ε_f-accurate estimates of $f(x^k)$ and $f(x^k + s^k)$, respectively, and let $\gamma \in (2, +\infty)$ be a fixed constant. Then the iteration is called:*

$$\begin{cases} \textit{successful} & \textit{if} \quad f_s^k - f_0^k \leq -\gamma \varepsilon_f (\delta_p^k)^2, \\ \textit{certain unsuccessful} & \textit{if} \quad f_s^k - f_0^k \geq \gamma \varepsilon_f (\delta_p^k)^2, \\ \textit{uncertain unsuccessful} & \textit{if} \quad f_s^k - f_0^k \in \mathcal{I}_{\gamma,\varepsilon_f}(\delta_p^k) := \left(-\gamma \varepsilon_f (\delta_p^k)^2, \gamma \varepsilon_f (\delta_p^k)^2\right), \end{cases}$$

where $\mathcal{I}_{\gamma,\varepsilon_f}(\delta_p^k)$ is the so-called uncertainty interval that is reduced during uncertain unsuccessful iterations.

Let $\tau \in (0, 1) \cap \mathbb{Q}$ be a fixed rational constant and $\hat{z} \in \mathbb{N}$ be a large fixed integer. Note that for the needs of the convergence analysis of Section 4.3, unlike MADS, the frame size parameter of StoMADS is assumed to be bounded above by a positive fixed constant $\tau^{-\hat{z}}$ in order for the random frame size parameter Δ_p^k that will be introduced in the next subsection to be integrable.

During the SEARCH or POLL step, if the *sufficient decrease condition* $f_s^k - f_0^k \leq -\gamma \varepsilon_f (\delta_p^k)^2$ is satisfied for some direction $s^k = \delta_m^k d$, then the iterate $x^k + s^k$ is successful according to Proposition 1.

Hence, the current iterate and the frame size parameter are updated according to $x^{k+1} = x^k + s^k$ and $\delta_p^{k+1} = \min\{\tau^{-2}\delta_p^k, \tau^{-\hat{z}}\}$, respectively. A new iteration is then initiated with a new mesh size parameter $\delta_m^{k+1} = \min\{\delta_p^{k+1}, (\delta_p^{k+1})^2\}$.

If no better mesh point is found during the SEARCH step, then the POLL step is invoked. If the condition $f_s^k - f_0^k \leq -\gamma\varepsilon_f(\delta_p^k)^2$ does not hold, then the iterate is unsuccessful according to Proposition 1. In the case of a certain or uncertain unsuccessful iteration, the current iterate is not updated, i.e., $x^{k+1} = x^k$, and the corresponding frame $\mathcal{F}^k$ is said to be a *minimal frame* with *minimal frame center* x^k, also called a *mesh local optimizer* [18]. However, in the case of a certain unsuccessful iteration, the frame size parameter is reduced according to $\delta_p^{k+1} = \tau^2\delta_p^k$ so that the resolution of the mesh can be increased, thus allowing the evaluation of f_Θ and hence the computation of estimates at trial mesh points that are closer to the current solution. Note that the use of τ^2 instead of τ (used in [16]) has been motivated by the need to reduce the frame size parameter less aggressively during uncertain unsuccessful iterations. Indeed, in the case of uncertain unsuccessful iterations, i.e., when $f_s^k - f_0^k$ belongs to the uncertainty interval $\mathcal{I}_{\gamma,\varepsilon_f}(\delta_p^k)$, the frame size parameter is reduced less aggressively (using $\delta_p^{k+1} = \tau\delta_p^k$), so that the uncertainty interval is reduced. A new iteration is then initiated with a new mesh size parameter δ_m^{k+1}. An overview of the algorithm and its details are presented in Figure 4.1 and Algorithm 4, respectively.

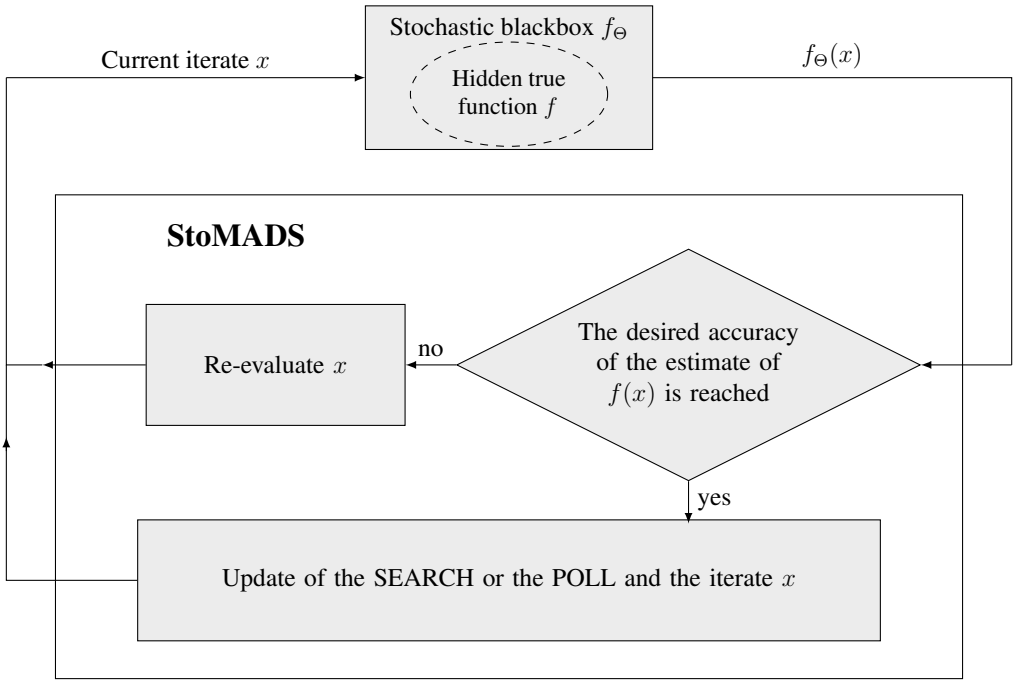


Figure 4.1 Overview of the StoMADS algorithm

Algorithm 4: StoMADS

[0] Initialization

Choose $x^0 \in \mathbb{R}^n$, $\delta_p^0 = 1$, $\tau = \frac{1}{2}$, $\varepsilon_f > 0$, $\epsilon_{stop} \geq 0$, $\gamma > 2$ and $\hat{z} \in \mathbb{N}^*$.

Set the iteration counter $k \leftarrow 0$.

[1] Parameter Update

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

[2] Search

Select a finite subset $\mathcal{S}^k$ of $\mathcal{M}^k$.

Obtain estimates f_0^k and f_s^k of f respectively at x^k and $x^k + s^k \in \mathcal{S}^k$, using blackbox evaluations.

If $f_s^k - f_0^k \leq -\gamma\varepsilon_f(\delta_p^k)^2$ for some $x^k + s^k \in \mathcal{S}^k$, then
set $x^{k+1} \leftarrow x^k + s^k$ and $\delta_p^{k+1} \leftarrow \min\{\tau^{-2}\delta_p^k, \tau^{-\hat{z}}\}$ and go to **[4]**.

Go to **[3]**.

[3] Poll

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

Obtain estimates f_0^k and f_s^k of $f(x^k)$ and $f(x^k + s^k)$, respectively, using blackbox evaluations.

Success

If $f_s^k - f_0^k \leq -\gamma\varepsilon_f(\delta_p^k)^2$ for some $s^k \in \{\delta_m^k d : d \in \mathbb{D}_p^k\}$, then
set $x^{k+1} \leftarrow x^k + s^k$, and $\delta_p^{k+1} \leftarrow \min\{\tau^{-2}\delta_p^k, \tau^{-\hat{z}}\}$.

Failure

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

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

[4] Termination

If $\delta_p^k \geq \epsilon_{stop}$, then
set $k \leftarrow k + 1$ and go to **[1]**.

Otherwise, stop.

Figure 4.2 The StoMADS algorithm

4.2.2 Probabilistic estimates

All random variables considered here are defined on the same probability space $(\Omega, \mathcal{G}, \mathbb{P})$, with Ω being the sample space, $\mathcal{G}$ being a σ-algebra, and $\mathbb{P}$ being a probability measure. Any single outcome of the sample space Ω will be denoted by ω. In general, random variables will be denoted by uppercase letters, while their realizations will be denoted by lowercase letters.

The estimates f_0^k and f_s^k of function values are constructed at each iteration of Algorithm 4 using evaluations of the noisy blackbox f_Θ. Because of the randomness of f_Θ, such estimates can be respectively considered as realizations of random estimates denoted by F_0^k and F_s^k, obtained based on some random samples of the stochastic function $f_\Theta(x)$. The behavior of F_0^k and F_s^k influences the outcome of each iteration of Algorithm 4 (as it is the case in [35, 83, 96]) in such a way that the iterates X^k, the polling directions D^k, the mesh size parameter Δ_m^k and the frame size parameter Δ_p^k are also random quantities. As mentioned above, $d^k = D^k(\omega)$, $x_k = X^k(\omega)$, $\delta_p^k = \Delta_p^k(\omega)$ and $\delta_m^k = \Delta_m^k(\omega)$ denote realizations of the random variables D^k, X^k, Δ_p^k, and Δ_m^k, respectively. Similarly, S^k denotes the random variable with realizations s^k; $f_0^k = F_0^k(\omega)$ and $f_s^k = F_s^k(\omega)$ denote estimates of $f(X^k)$ and $f(X^k + S^k)$, respectively. In other words, Algorithm 4 results in a stochastic process $\{X^k, S^k, \Delta_p^k, \Delta_m^k, F_0^k, F_s^k\}$. Note that $f(X^k)$ is used to denote the random variable with realizations $f(X^k(\omega))$.

One of the objectives of this work is to show that the abovementioned stochastic process converges with probability one under some assumptions on $\{F_0^k, F_s^k\}$. The probabilistic estimates will be assumed to be accurate with a sufficiently large, but fixed, probability *"conditioned on the past"* [29,35]. The notion of conditioning on the past is formalized as proposed in [35, 83]. Let $\mathcal{F}_{k-1}^F$ denote the σ-algebra generated by $F_0^0, F_s^0, F_0^1, F_s^1, \ldots, F_0^{k-1}$, and F_s^{k-1}. For completeness, $\mathcal{F}_{-1}^F$ is set to equal $\sigma(x^0)$. Thus, $\{\mathcal{F}_k^F\}_{k\geq -1}$ is a filtration, i.e., an increasing subsequence of σ-algebras of $\mathcal{G}$. Closeness or sufficient accuracy of function estimates is measured using the current frame size parameter. This notion is formalized using the following definition, which is a modified version of those in [29, 30, 35, 83] and similar to that in [96].

Definition 3. *A sequence of random estimates $\{F_0^k, F_s^k\}$ is said to be β-probabilistically ε_f-accurate with respect to the corresponding sequence $\{X^k, S^k, \Delta_p^k\}$ if the events*

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

satisfy the following submartingale-like condition

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

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

An iteration k is called "true" and an estimate is called "good" if $\mathbb{1}_{J_k} = 1$. Otherwise the iteration is called "false" and the estimate is called "bad".

The following definition of p-integrable random variables [28] is useful for the analysis of Algorithm 4.

Definition 4. *Let $p \in [1, +\infty)$ be an integer and $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space. 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*

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

In order for $f(X^k)$ to be integrable so that the conditional expectation $\mathbb{E}\left(f(X^k) \mid \mathcal{F}_{k-1}^F \right)$ can be well defined [28] for the needs of the analysis of StoMADS, the following is assumed.

Assumption 1. *The objective function f is locally Lipschitz continuous everywhere and all iterates x^k generated by Algorithm 4 lie in a compact set $\mathcal{X}$.*

The following result shows that $f(X^k)$ is integrable if Assumption 1 holds.

Proposition 2. *If Assumption 1 holds, then both Δ_p^k and $f(X^k) \in \mathbb{L}^1(\Omega, \mathcal{G}, \mathbb{P})$ for all k.*

Proof. The function f is bounded on $\mathcal{X}$ since f is locally Lipschitz and $\mathcal{X}$ is compact. Consequently, there exists a finite constant $F_{\max}$ such that all the iterates x^k lying in $\mathcal{X}$ satisfy $\left| f(x^k) \right| \leq F_{\max}$. In other words, all realizations $f(X^k(\omega))$ of the random variable $f(X^k)$ satisfy $\left| f(X^k(\omega)) \right| \leq F_{\max}$. Therefore, $\mathbb{E}\left(\left| f(X^k) \right| \right) := \int_\Omega \left| f(X^k(\omega)) \right| \mathbb{P}(d\omega) \leq F_{\max} < +\infty$.

Moreover, the integrability of Δ_p^k and hence, that of Δ_m^k follows straightforwardly from the fact that for all $\omega \in \Omega$, $\Delta_p^k(\omega) \leq \tau^{-\hat{z}}$. Indeed, $\mathbb{E}\left(\left| \Delta_p^k \right| \right) := \int_\Omega \left| \Delta_p^k(\omega) \right| \mathbb{P}(d\omega) \leq \tau^{-\hat{z}} < +\infty$. □

The following key assumption similar to that made in [83] on the nature of the stochastic information in Algorithm 4 will be useful for the convergence analysis presented in Section 4.3.

Assumption 2. *Let $\varepsilon_f > 0$ be the constant of Proposition 1. The following holds for the random quantities derived from Algorithm 4:*

(i) The sequence of estimates $\{F_0^k, F_s^k\}$ generated by Algorithm 4 is β-probabilistically ε_f-accurate for some $\beta \in (0, 1)$.

(ii) There exists $\kappa_F > 0$ such that the sequence of estimates $\{F_0^k, F_s^k\}$ generated by Algorithm 4 satisfies the following κ_F-variance conditions for all $k \geq 0$

$$\begin{aligned} & \mathbb{E}\left(\left|F_s^k - f(X^k + S^k)\right|^2 \mid \mathcal{F}_{k-1}^F\right) \leq (\kappa_F)^2(\Delta_p^k)^4 \\ \textit{and} \qquad & \mathbb{E}\left(\left|F_0^k - f(X^k)\right|^2 \mid \mathcal{F}_{k-1}^F\right) \leq (\kappa_F)^2(\Delta_p^k)^4. \end{aligned} \tag{4.2}$$

Remark 1. *The role of the frame size parameter Δ_p^k in the stochastic framework of this work is twofold. First, it updates the resolution of the mesh (which, as it will be seen, gets infinitely fine), and second, it adaptively controls the variance which again, as it will be seen, will be driven to zero as Algorithm 4 progresses, thus allowing it to reach a desired accuracy. Therefore, no other "control size" parameter is required for the analysis in order to control the variance as needed and described for the line search method proposed in [83]. As in [83], note that at point (ii) of Assumption 2, the integrability of random quantities $\left|F_0^k - f(X^k)\right|^2$ and $\left|F_s^k - f(X^k + S^k)\right|^2$ and hence straightforwardly that of $\left|F_0^k - f(X^k)\right|$ and $\left|F_s^k - f(X^k + S^k)\right|$ is implicitly assumed for all k.*

Using this key assumption on the accuracy of function estimates, a lower bound on β, defined in terms of τ, κ_F, and ε_f will be derived, necessary for the convergence property of Algorithm 4. Before delving into the convergence analysis, we state and prove a useful lemma (slightly modified from [83]), that demonstrates the relationship between the variance assumption on the function values and the probability of obtaining bad estimates.

Lemma 1. *Let Assumption 2 hold. Suppose that $\{X^k, F_0^k, F_s^k, \Delta_p^k\}$ is a random process generated by Algorithm 4. Then for every $k \geq 0$,*

$$\begin{aligned} & \mathbb{E}\left(\mathbb{1}_{\bar{J}_k}\left|F_s^k - f(X^k + S^k)\right| \mid \mathcal{F}_{k-1}^F\right) \leq (1-\beta)^{1/2}\kappa_F(\Delta_p^k)^2 \\ \textit{and} \qquad & \mathbb{E}\left(\mathbb{1}_{\bar{J}_k}\left|F_0^k - f(X^k)\right| \mid \mathcal{F}_{k-1}^F\right) \leq (1-\beta)^{1/2}\kappa_F(\Delta_p^k)^2. \end{aligned}$$

Proof. The result is shown for $F_0^k - f(X^k)$ using ideas derived from [83] (specifically, by making use of the conditional Cauchy-Schwarz inequality [28], the proof for $F_s^k - f(X^k + S^k)$ is the same). However, the proof here is slightly modified to emphasize the integrability of the random variables that define the conditional expectations.

Since it follows trivially from Assumption 2 that $\left|F_0^k - f(X^k)\right| \in \mathbb{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ and that $\mathbb{1}_{\bar{J}_k} \in \mathbb{L}^2(\Omega, \mathcal{G}, \mathbb{P})$, then $\mathbb{1}_{\bar{J}_k}\left|F_0^k - f(X^k)\right| \in \mathbb{L}^1(\Omega, \mathcal{G}, \mathbb{P})$ due to the Cauchy-Schwarz inequality [28]. Thus,

it follows from the conditional Cauchy-Schwarz inequality that

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

where the last inequality follows from (4.2) and the fact that $\mathbb{E}\left(\mathbb{1}_{\bar{J}_k} \mid \mathcal{F}_{k-1}^F\right) = \mathbb{P}\left(\bar{J}_k \mid \mathcal{F}_{k-1}^F\right) \leq 1-\beta$ due to point *(i)* of Assumption 2. □

4.2.3 Computation of probabilistic estimates

This section demonstrates how random estimates F_0^k and F_s^k satisfying Assumption 2 can be constructed in a simple random noise framework and hence how deterministic estimates f_0^k and f_s^k can be obtained through evaluating the blackbox f_Θ using techniques derived from [35, 83, 96].

Recall that f_Θ denotes the noisy available blackbox which is the computable version of the numerically unavailable objective f and consider the following typical noise assumption often used in the stochastic optimization literature [35], i.e., suppose that the noise Θ is unbiased for all f:

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

where $V > 0$ is a constant. Let Θ_0 and Θ_s be two independent random variables following the same distribution as Θ. Define estimates F_0^k and F_s^k respectively by $F_0^k = \frac{1}{p^k}\sum_{i=1}^{p^k} f_{\Theta_{0,i}}(x^k)$ and $F_s^k = \frac{1}{p^k}\sum_{i=1}^{p^k} f_{\Theta_{s,i}}(x^k + s^k)$, where p^k denotes the sample size and $\Theta_{0,1}, \Theta_{0,2}, \ldots, \Theta_{0,p^k}$ and $\Theta_{s,1}, \Theta_{s,2}, \ldots, \Theta_{s,p^k}$ are independent random samples of Θ_0 and Θ_s, respectively. By noticing that $\mathbb{E}\left(F_0^k\right) = f(x^k)$ and that $\text{Var}(F_0^k) \leq \frac{V}{p^k}$, it follows from the Chebyshev inequality that

$$\mathbb{P}\left(\left|F_0^k - f(x^k)\right| > \varepsilon_f(\delta_p^k)^2\right) \leq \frac{\text{Var}(F_0^k)}{(\varepsilon_f)^2(\delta_p^k)^4} \leq \frac{V}{p^k(\varepsilon_f)^2(\delta_p^k)^4}.$$

Thus, choosing p^k according to

$$p^k \geq \frac{V}{(\varepsilon_f)^2(1-\sqrt{\beta})(\delta_p^k)^4}, \tag{4.3}$$

leads to

$$\mathbb{P}\left(\left|F_0^k - f(x^k)\right| \leq \varepsilon_f(\delta_p^k)^2\right) \geq \sqrt{\beta} \text{ and, similarly, } \mathbb{P}\left(\left|F_s^k - f(x^k + s^k)\right| \leq \varepsilon_f(\delta_p^k)^2\right) \geq \sqrt{\beta}. \tag{4.4}$$

Since F_0^k and F_s^k are independent random variables, it follows from (4.4) that

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

which shows that point *(i)* of Assumption 2 holds.

Now, in order to prove point *(ii)*, notice that $\mathbb{E}\left(F_0^k - f(x^k)\right) = 0$, which implies that

$$\mathbb{E}\left(\left|F_0^k - f(x^k)\right|^2\right) = \text{Var}\left(F_0^k\right) \leq \frac{V}{p^k} \leq (\varepsilon_f)^2(1-\sqrt{\beta})(\delta_p^k)^4, \tag{4.5}$$

where the last inequality follows from (4.3). Similarly, since $\mathbb{E}\left(F_s^k - f(x^k+s^k)\right) = 0$, then

$$\mathbb{E}\left(\left|F_s^k - f(x^k+s^k)\right|^2\right) \leq (\varepsilon_f)^2(1-\sqrt{\beta})(\delta_p^k)^4. \tag{4.6}$$

It follows from (4.5) and (4.6) that the point *(ii)* of Assumption 2 holds for κ_F chosen according to

$$(\kappa_F)^2 \geq (\varepsilon_f)^2(1-\sqrt{\beta}).$$

By using the fact that the deterministic estimates f_0^k and f_s^k are realizations of F_0^k and F_s^k, respectively, their values can be estimated by averaging p^k realizations of f_Θ obtained from the evaluations of the stochastic blackbox at x^k and $x^k + s^k$.

We propose the following technique to reduce the required number of evaluations, especially for computationally expensive blackboxes. Recall that $x^{k+1} = x^k + s^k$ after a successful iteration while $x^{k+1} = x^k$ after an unsuccessful one. Denote by $n^k \leq p^k$ the number of blackbox evaluations at a given point when constructing an estimate at the iteration k with $n^0 = p^0$. If iteration k is successful, then the estimate f_0^{k+1} of $f(x^{k+1})$ is computed by

$$f_0^{k+1} = \frac{n^k f_s^k + \sum_{j=n^k+1}^{p^{k+1}} f_{\theta_{s,j}}(x^{k+1})}{p^{k+1}} \tag{4.7}$$

(using $f_s^k = \frac{1}{n^k}\sum_{i=1}^{n^k} f_{\theta_{s,i}}(x^k+s^k)$), where $p^{k+1} = n^k + n^{k+1}$. If iteration k is unsuccessful, then f_0^{k+1} is computed by

$$f_0^{k+1} = \frac{p^k f_0^k + \sum_{j=p^k+1}^{p^{k+1}} f_{\theta_{0,j}}(x^{k+1})}{p^{k+1}}, \tag{4.8}$$

where $p^{k+1} = p^k + n^{k+1}$. Indeed, this procedure is used in Section 4.4 to improve the accuracy of the estimates by making use of available samples at the current iterate, avoiding thus additional blackbox evaluations. Despite the fact that this computation scheme is inherently biased, it seems to be very

efficient in practice, especially for computationally expensive blackboxes.

4.3 Convergence analysis

This section presents convergence results for StoMADS using ideas inspired by [35, 68, 83]. The first result is a *zeroth-order* result [14], showing that there exists a subsequence of StoMADS-generated random iterates with mesh realizations becoming infinitely fine and which converges to a limit with probability one. More formally, StoMADS generates a convergent subsequence $\{X^k\}_{k\in K}$ of random iterates such that $\lim_{k\in K} X^k = \hat{X}$ almost surely, provided that $\lim_{k\to+\infty} \Delta_m^k = 0$ with probability one. This result is stronger than the liminf-type result in [12] about the convergence of the sequence of mesh size parameters. Then, under the assumptions of compactness of the set containing all iterates and local Lipschitz continuity of f, a stochastic variant of the *first-order* necessary optimality condition [12, 16] via the Clarke derivative [38] is proved.

4.3.1 Zeroth-order convergence

In order to prove the existence of an almost surely convergent subsequence of StoMADS random iterates with mesh realizations becoming infinitely fine, it is first proved that, with probability one, the sequence of random mesh size parameters converges to zero almost surely. The following lemma, similar to those derived in [35, 83], guarantees decrease in the objective function f at successful iterations.

Lemma 2. *Let $\varepsilon_f > 0$ and $\gamma > 2$ be fixed constants and suppose $\{f_0^k, f_s^k\}$ are ε_f-accurate estimates. If the iteration is successful, then the improvement in f is bounded as follows*

$$f(x^{k+1}) - f(x^k) \leq -(\gamma - 2)\varepsilon_f(\delta_p^k)^2. \tag{4.9}$$

Proof. Since the iteration is successful and because the estimates are ε_f-accurate,

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

□

Before proving the following theorem that provides a result which is similar to that obtained in [35] and which represents the cornerstone of the convergence results in the present work, the following

assumption on f is needed.

Assumption 3. *The objective function f is bounded from below, i.e., there exists $f_{\min} \in \mathbb{R}$ such that $-\infty < f_{\min} \leq f(x)$, for all $x \in \mathbb{R}^n$.*

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

Theorem 1. *Let Assumptions 1 and 3 be satisfied. Let $\varepsilon_f > 0$, $\tau \in (0,1) \cap \mathbb{Q}$ and $\gamma > 2$. Let $\nu \in (0,1)$ be chosen so that*

$$\frac{\nu}{1-\nu} \geq \frac{2(\tau^{-4}-1)}{\varepsilon_f(\gamma-2)}, \tag{4.10}$$

and $\beta \in (1/2, 1)$ be chosen so that Assumption 2 holds with

$$\frac{\beta}{\sqrt{1-\beta}} \geq \frac{4\nu\kappa_F}{(1-\nu)(1-\tau^2)}. \tag{4.11}$$

Then the sequence $\{\Delta_m^k\}$ of mesh size parameters generated by Algorithm 4 satisfies

$$\sum_{k=0}^{+\infty} \Delta_m^k < +\infty \quad \textit{almost surely}. \tag{4.12}$$

Proof. This theorem is proved, using techniques and ideas derived from [35, 68, 83] and making use of properties of the random function

$$\Phi_k = \nu(f(X^k) - f_{\min}) + (1-\nu)(\Delta_p^k)^2.$$

A similar random function is used in [35, 68], where $\nu \in (0,1)$ is a fixed constant specified below. Recall that $\Delta_m^k = \min\{\Delta_p^k, (\Delta_p^k)^2\}$ and note that $\Phi_k \in \mathbb{L}^1(\Omega, \mathcal{G}, \mathbb{P})$ according to Proposition 2, which implies that the conditional expectation $\mathbb{E}\left(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^F\right)$ is well defined for all k.

The overall goal is to show that there exists a constant $\eta > 0$ such that for all k

$$\mathbb{E}\left(\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}^F\right) \leq -\eta(\Delta_p^k)^2 < 0. \tag{4.13}$$

Let us assume that (4.13) holds at every iteration. Since f is bounded from below by $f_{\min}$ and Δ_p^k is positive, we have that Φ_k is bounded from below for all k. Hence, summing up over $k \in \mathbb{N}$ and taking the expectation of both sides of (4.13), leads to the conclusion that (4.12) holds with probability one. We can then prove the theorem by showing that (4.13) holds at each iteration.

The proof of this theorem considers two separate cases: good estimates and bad estimates, each of which will be broken into three subcases depending on whether an iteration is successful, certain

unsuccessful, or uncertain unsuccessful. We introduce the following events as suggested in [83]:

S := {The iteration is successful}, $\bar{S}$:= {The iteration is unsuccessful},

$\bar{S}^C$:= {The iteration is certain unsuccessful}, $\bar{S}^{\bar{C}}$:= {The iteration is uncertain unsuccessful}.

Case 1 (Good estimates, $\mathbb{1}_{J_k} = 1$). It will be shown that Φ_k decreases no matter what type of iteration occurs and that the smallest decrease occurs at uncertain unsuccessful iterations. Thus, this case dominates the other two, leading to the conclusion that

$$\mathbb{E}\left(\mathbb{1}_{J_k}(\Phi_{k+1} - \Phi_k) \mid \mathcal{F}_{k-1}^F\right) \leq -\beta(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{4.14}$$

(i) *Successful iteration* ($\mathbb{1}_S = 1$). If the iteration is successful and the estimates are good, a decrease in the objective f occurs; specifically, Lemma 2 applies:

$$\mathbb{1}_{J_k}\mathbb{1}_S\, \nu(f(X^{k+1}) - f(X^k)) \leq -\mathbb{1}_{J_k}\mathbb{1}_S \nu(\gamma - 2)\varepsilon_f(\Delta_p^k)^2. \tag{4.15}$$

As the iteration is successful, $\Delta_p^{k+1} = \min\{\tau^{-2}\Delta_p^k, \tau^{-\hat{z}}\}$, and consequently,

$$\mathbb{1}_{J_k}\mathbb{1}_S(1-\nu)\left[(\Delta_p^{k+1})^2 - (\Delta_p^k)^2\right] \leq \mathbb{1}_{J_k}\mathbb{1}_S(1-\nu)(\tau^{-4}-1)(\Delta_p^k)^2. \tag{4.16}$$

Let ν be large enough so that the right-hand side of (4.15) dominates that of (4.16), i.e.,

$$-\nu(\gamma-2)\varepsilon_f(\Delta_p^k)^2 + (1-\nu)(\tau^{-4}-1)(\Delta_p^k)^2 \leq -\frac{1}{2}\nu(\gamma-2)\varepsilon_f(\Delta_p^k)^2, \tag{4.17}$$

which is equivalent to equation (4.10). The combination of (4.15) and (4.16) leads to

$$\mathbb{1}_{J_k}\mathbb{1}_S(\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{J_k}\mathbb{1}_S\frac{1}{2}\nu(\gamma-2)\varepsilon_f(\Delta_p^k)^2. \tag{4.18}$$

(ii) *Certain unsuccessful iteration* ($\mathbb{1}_{\bar{S}^C} = 1$). The iteration is unsuccessful; therefore, there is no change in the function values while Δ_p^k decreases. Hence,

$$\mathbb{1}_{J_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C}(\Phi_{k+1} - \Phi_k) \;\leq\; -\mathbb{1}_{J_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C}(1-\nu)(1-\tau^4)(\Delta_p^k)^2. \tag{4.19}$$

(iii) *Uncertain unsuccessful iteration* ($\mathbb{1}_{\bar{S}^{\bar{C}}} = 1$). The behavior of Algorithm 4 at *uncertain unsuccessful iterations* can be derived from its behavior at *certain unsuccessful iterations* by simply replacing τ^2 with τ. Thus, the bound in the change of Φ_k follows straightforwardly from (4.19) by replacing $\mathbb{1}_{\bar{S}^C}$ with $\mathbb{1}_{\bar{S}^{\bar{C}}}$ and τ^4 with τ^2:

$$\mathbb{1}_{J_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^{\bar{C}}}(\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{J_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^{\bar{C}}}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{4.20}$$

Let ν be large enough so that uncertain unsuccessful iterations (4.20) provide the worst case decrease when compared to (4.18) and (4.19). More precisely, ν is chosen according to

$$-\frac{1}{2}\nu(\gamma-2)\varepsilon_f(\Delta_p^k)^2 \;\leq\; -(1-\nu)(1-\tau^4)(\Delta_p^k)^2 \leq -(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \quad (4.21)$$

The inequalities $1-\tau^2 < 1-\tau^4 < \tau^{-4}-1$ ensure that (4.21) is satisfied whenever ν is chosen according to (4.17).

In the case of accurate estimates, using (4.18), (4.19), (4.20), and (4.21), the change in Φ_k is bounded by

$$\begin{aligned} \mathbb{1}_{J_k}(\Phi_{k+1}-\Phi_k) &= \mathbb{1}_{J_k}(\mathbb{1}_S + \mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C} + \mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}\bar{C}})(\Phi_{k+1}-\Phi_k) \\ &\leq -\mathbb{1}_{J_k}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \end{aligned} \quad (4.22)$$

Taking the conditional expectation with respect to $\mathcal{F}_{k-1}^F$ at both sides of (4.22) and using assumption 2 leads to (4.14).

Case 2 (Bad estimates, $\mathbb{1}_{\bar{J}_k}=1$). When the estimates are bad, the algorithm may accept an iterate that leads to an increase in f and Δ_p^k, and hence in Φ_k. To control such an increase in Φ_k, the variance in the function estimates is bounded making use of (4.2). The probability of outcome (Case 2) is then adjusted to be sufficiently small in order to ensure that the expectation of Φ_k is reduced sufficiently. More precisely, it will be proved that

$$\mathbb{E}\left(\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}-\Phi_k) \mid \mathcal{F}_{k-1}^F\right) \leq 2\nu(1-\beta)^{1/2}\kappa_F(\Delta_p^k)^2. \quad (4.23)$$

As before, let us consider three separate cases.

(i) *Successful iteration* $(\mathbb{1}_S=1)$. Whenever bad estimates occur, a successful iteration leads to the following bound

$$\begin{aligned} &\mathbb{1}_{\bar{J}_k}\mathbb{1}_S\,\nu(f(X^{k+1})-f(X^k)) \\ &\qquad \leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S\,\nu\left[(F_s^k-F_0^k)+\left|f(X^{k+1})-F_s^k\right|+\left|F_0^k-f(X^k)\right|\right] \\ &\qquad \leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S\,\nu\left[-\gamma\varepsilon_f(\Delta_p^k)^2+\left|f(X^{k+1})-F_s^k\right|+\left|F_0^k-f(X^k)\right|\right], \end{aligned} \quad (4.24)$$

where the last inequality is due to the decrease condition $F_s^k-F_0^k \leq -\gamma\varepsilon_f(\Delta_p^k)^2$, which holds at every successful iteration. Since the iteration is successful, $\Delta_p^{k+1} = \min\{\tau^{-2}\Delta_p^k, \tau^{-\hat{z}}\}$. Therefore,

$$\mathbb{1}_{\bar{J}_k}\mathbb{1}_S(1-\nu)\left[(\Delta_p^{k+1})^2-(\Delta_p^k)^2\right] \leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S(1-\nu)(\tau^{-4}-1)(\Delta_p^k)^2. \quad (4.25)$$

Choosing ν according to (4.17) implies

$$-\nu\gamma\varepsilon_f(\Delta_p^k)^2 + (1-\nu)(\tau^{-4}-1)(\Delta_p^k)^2 \leq 0. \tag{4.26}$$

Combining then (4.24) and (4.25) leads to

$$\mathbb{1}_{\bar{J}_k}\mathbb{1}_S(\Phi_{k+1}-\Phi_k) \leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S(\nu\left|f(X^{k+1})-F_s^k\right| + \nu\left|F_0^k - f(X^k)\right|). \tag{4.27}$$

(ii) *Certain unsuccessful iteration* $(\mathbb{1}_{\bar{S}^C}=1)$. Since Δ_p^k is decreased and there is no change in the function values, the bound in the change of Φ_k follows straightforwardly from that obtained in (4.19) by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$. Specifically,

$$\begin{aligned}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C}(\Phi_{k+1}-\Phi_k) &\leq -\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C}(1-\nu)(1-\tau^4)(\Delta_p^k)^2 \\ &\leq -\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^C}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \end{aligned}\tag{4.28}$$

(iii) *Uncertain unsuccessful iteration* $(\mathbb{1}_{\bar{S}^{\bar{C}}}=1)$. Here again, the bound in the change of Φ_k is derived from that obtained in (4.20), simply by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$. Specifically,

$$\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^{\bar{C}}}(\Phi_{k+1}-\Phi_k) \leq -\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}\mathbb{1}_{\bar{S}^{\bar{C}}}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{4.29}$$

By noticing that $\bar{S}^{\bar{C}} \cup \bar{S}^C = \bar{S}$ and combining (4.28) and (4.29) leads to

$$\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}(\Phi_{k+1}-\Phi_k) \leq -\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{4.30}$$

Finally, since (4.27) dominates (4.30), then in all three cases

$$\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}-\Phi_k) \leq \mathbb{1}_{\bar{J}_k}(\nu\left|f(X^{k+1})-F_s^k\right| + \nu\left|F_0^k - f(X^k)\right|). \tag{4.31}$$

Taking the expectation with respect to $\mathcal{F}_{k-1}^F$ on both sides of (4.31) and applying Lemma 1 leads to (4.23).

Now, combining (4.14) and (4.23) leads to

$$\begin{aligned}\mathbb{E}\left(\Phi_{k+1}-\Phi_k \mid \mathcal{F}_{k-1}^F\right) &= \mathbb{E}\left((\mathbb{1}_{J_k}+\mathbb{1}_{\bar{J}_k})(\Phi_{k+1}-\Phi_k) \mid \mathcal{F}_{k-1}^F\right) \\ &\leq -\beta(1-\nu)(1-\tau^2)(\Delta_p^k)^2 + 2\nu(1-\beta)^{1/2}\kappa_F(\Delta_p^k)^2 \\ &\leq \left[-\beta(1-\nu)(1-\tau^2) + 2\nu\kappa_F(1-\beta)^{1/2}\right](\Delta_p^k)^2. \end{aligned}\tag{4.32}$$

Then, choosing β in $(1/2, 1)$ according to (4.11) ensures that

$$-\beta(1-\nu)(1-\tau^2) + 2\nu\kappa_F(1-\beta)^{1/2} \leq -\frac{1}{2}\beta(1-\nu)(1-\tau^2). \tag{4.33}$$

Hence, (4.13) follows from (4.32) and (4.33) with $\eta = \frac{1}{2}\beta(1-\nu)(1-\tau^2) > 0$, and the proof is complete by noticing that $\Delta_m^k = \min\{\Delta_p^k, (\Delta_p^k)^2\}$. □

The following result shows that all the realizations of random iterates X^k generated by StoMADS lie on meshes that become infinitely fine with probability one.

Corollary 1. *Let the assumptions made in Theorem 1 hold. Then, almost surely*

$$\lim_{k\to+\infty} \Delta_m^k = 0. \tag{4.34}$$

Proof. It follows from Theorem 1 that $\sum_{k=0}^{+\infty} \Delta_m^k < +\infty$ almost surely. As a consequence, the sequence $\{\Delta_m^k\}_{k\in\mathbb{N}}$ of mesh size parameters converges to zero almost surely. □

Remark 2. *We emphasize that the result in* (4.34) *is stronger than the one obtained in the deterministic framework of the MADS algorithm, where it was shown that* $\liminf_{k\to+\infty} \delta_m^k = 0$. *Indeed, unlike the deterministic framework of the MADS algorithm where available outputs of the objective function f are directly compared in order to ensure improvement, such behavior of the random sequence of mesh size parameters in the present stochastic framework is due to the use of a sufficient decrease condition in the definition of iteration types (see Proposition 1 and Definition 2). A similar remark about the convergence to zero of a whole sequence of step size parameters is made in [42] when a sufficient decrease condition had been imposed on the "directional direct search method."*

We now introduce the following definition (similar to that in [16]) to show the existence of convergent subsequences of StoMADS iterates.

Definition 5. *A convergent subsequence $\{x^k\}_{k\in\mathcal{K}}$ of* StoMADS *iterates (for some subset of indices $\mathcal{K}$), is said to be a* refining subsequence*, if and only if $\{\delta_m^k\}_{k\in\mathcal{K}}$ converges to zero. The limit $\hat{x}$ of $\{x^k\}_{k\in\mathcal{K}}$ is called a* refined point*.*

The existence of convergent refining subsequences was first proved in the deterministic framework of the Generalized Pattern Search (GPS) algorithm under assumptions including that all the iterates generated by GPS belong to a compact set [11]. This proof was then generalized to the MADS algorithm in [12], but with the latter assumption replaced by one where all the iterates generated by MADS belong to the level set $\mathscr{L}(f(x^0)) := \{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$, which is assumed

to be bounded. For both algorithms, the refining subsequences were shown to be subsequences of mesh local optimizers on meshes becoming infinitely fine. Note that it is impossible for the objective function values to increase in a deterministic framework. The challenge in the present stochastic framework, as in those of related works [29,35,83,96], lies in the fact that the iterates may lie outside the initial level set $\mathscr{L}(f(x^0))$ since the function f values may possibly increase between successive iterations. In other words, StoMADS "*can venture outside the initial level set*" [35]. Motivated by these remarks, the following theorem is proved under Assumption 1 (also used in [11]) to simplify the analysis.

Theorem 2. *Let the assumptions made in Theorem 1 and Assumption 1 hold. Then, there exists at least one almost surely convergent refining subsequence* $\{X^k\}_{k\in K}$.

Proof. The proof uses ideas derived from [11]. The result is shown by making use of the event $V = \{\omega \in \Omega : \lim_{k\to+\infty} \Delta_m^k(\omega) = 0\}$ that is almost sure due to Corollary 1.

For all $\omega \in V$, $\{X^k(\omega)\}_{k\in\mathbb{N}}$ is a sequence of iterates on meshes becoming infinitely fine. It therefore follows from the compactness hypothesis of Assumption 1 that there exists a subset of indices $K(\omega)$ for which the subsequence $\{X^k(\omega)\}_{k\in K(\omega)}$ converges. Denote by $\hat{X}(\omega)$ the limit of $\{X^k(\omega)\}_{k\in K(\omega)}$. The proof follows by noticing that $V \subseteq \{\omega \in \Omega : \lim_{k\in K(\omega)} X^k(\omega) = \hat{X}(\omega)\}$. □

4.3.2 Nonsmooth optimality conditions

The main goal of this section is to show with probability one that any refined point $\hat{X}$ derived in Theorem 2 satisfies a stochastic variant of the first-order necessary optimality condition based on the Clarke derivative stated as Theorem 6.9 in [16].

The Clarke optimality result requires that the polling directions d^k are chosen in such a way that the sequence $\{\delta_m^k \left\|d^k\right\|_\infty\}_{k\in\mathbb{N}}$ converges to zero while $\{\delta_p^k \left\|d^k\right\|_\infty\}_{k\in\mathbb{N}}$ does not converge to zero even though both sequences of mesh and frame size parameters converge to zero. To meet this requirement, we require that i) the columns of the matrix $\mathbf{D}$ used in the definition of the mesh $\mathcal{M}^k$ are the $2n$ positive and negative coordinate directions, ii) the initial frame size parameter δ_p^0 equals 1, iii) the mesh refining parameter τ equals $1/2$, and iv) all search directions used in Algorithm 4 during the POLL step are generated by Algorithm 5 in [16]. Note that under the previously made assumptions, the sequence $\{\delta_m^k \left\|d^k\right\|_\infty\}_{k\in\mathbb{N}}$ is shown in [16] to converge to zero. However, $\delta_p^k \left\|d^k\right\|_\infty \geq 1$ for large values of k. Indeed, consider $d^k = \text{round}\left(\frac{\delta_p^k}{\delta_m^k}\frac{h}{\|h\|_\infty}\right)$, where $h = (h^1, h^2, \ldots, h^n)^\top$ is a column of the

Householder matrix $\mathbf{H}^k$. The function round$(\cdot)$ is defined as

$$\text{round}(t) = \begin{cases} \text{sgn}(t)\,\lfloor |t| + 0.5 \rfloor & \text{if } |t| - \lfloor |t| \rfloor = 0.5 \\ \text{sgn}(t)\,\lfloor |t| \rfloor & \text{if } |t| - \lfloor |t| \rfloor < 0.5 \\ -\text{sgn}(t)\,\lfloor -|t| \rfloor & \text{if } |t| - \lfloor |t| \rfloor > 0.5 \end{cases} \quad \text{for all } t \in \mathbb{R},$$

where $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t and $\text{sgn}(t)$ is the sign of t. For example, $\text{round}(2.4) = \text{round}(2) = 2$, $\text{round}(2.5) = \text{round}(2.8) = 3$, $\text{round}(-2.4) = \text{round}(-2) = -2$, $\text{round}(-2.5) = \text{round}(-2.8) = -3$, and $\text{round}(0) = 0$. Consider the indices j and k_0 such that $|h^j| = \|h\|_\infty$ and $\delta_p^k \leq 1$ for all $k \geq k_0$, respectively. Then, for all $k \geq k_0$, $\delta_p^k \left\| d^k \right\|_\infty \geq 1$ since $1/\delta_p^k$ is an integer and

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

Algorithm 5: Creating the set $\mathbb{D}_p^k$ of poll directions

---

Given $v^k \in \mathbb{R}^n$ 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 $\mathbb{D}_p^k = \mathbb{B}^k \cup (-\mathbb{B}^k)$

We will use the following auxiliary result [24, 35], which is based on martingale theory [50].

Theorem 3. *Let $\{G_k\}_{k \in \mathbb{N}}$ be a submartingale, i.e, a sequence of random variables which, for every $k \in \mathbb{N}$, satisfy*

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

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

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

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

The properties of the random function Ψ_k introduced next will be useful for the proof of the optimality result based on the Clarke derivative in Theorem 5.

Theorem 4. *Let the assumptions made in Theorem 1 hold. Define the random function Ψ_k with realizations ψ_k as*

$$\psi_k = \frac{f(x^k) - f(x^k + \delta_m^k d)}{\delta_p^k},$$

where $d \in \mathbb{D}_p^k$ is any direction used by StoMADS *and that is generated by Algorithm 5. Then, almost surely,*

$$\liminf_{k \to +\infty} \Psi_k \leq 0. \tag{4.35}$$

Proof. Using ideas from the liminf-type first-order convergence proof in [35], we will show this result by contradiction conditioned on the event $V' = \left\{\lim_{k\to+\infty} \Delta_p^k = 0\right\}$ that is almost sure due to Corollary 1. All of the following is conditioned on V'. Assume that with non-zero probability, there exists a random variable $\mathcal{E}$ with realizations $\epsilon > 0$ such that

$$\Psi_k \geq \mathcal{E}(\gamma + 2), \quad \text{for all } k \in \mathbb{N}. \tag{4.36}$$

That is, assume that

$$\mathbb{P}\left(\{\omega \in \Omega : \exists \mathcal{E}(\omega) > 0 \text{ such that } \forall k \in \mathbb{N}, \Psi_k(\omega) \geq \mathcal{E}(\omega)(\gamma + 2)\}\right) > 0, \tag{4.37}$$

where $\gamma \in (2, +\infty)$ is the same constant in Algorithm 4 and recall that $s^k = \delta_m^k d$ for all k. Let $\{x^k\}_{k\in\mathbb{N}}$, $\{\delta_p^k\}_{k\in\mathbb{N}}$, $\{s^k\}_{k\in\mathbb{N}}$, and ϵ be realizations of $\{X^k\}_{k\in\mathbb{N}}$, $\{\Delta_p^k\}_{k\in\mathbb{N}}$, $\{S^k\}_{k\in\mathbb{N}}$ and $\mathcal{E}$, respectively for which $\psi_k \geq \epsilon(\gamma + 2),$ for all $k \in \mathbb{N}$. Since $\lim_{k\to+\infty}\delta_p^k = 0$ because of the conditioning on V', there exists $k_0 \in \mathbb{N}$ such that

$$\delta_p^k < \lambda := \min\left\{\frac{\epsilon}{\varepsilon_f}, \tau^{2-\hat{z}}\right\}, \text{ for all } \ k \geq k_0, \tag{4.38}$$

where $\hat{z}$ is the same parameter of Algorithm 4 satisfying $\delta_p^k \leq \tau^{-\hat{z}}$ for all $k \geq 0$. Define the random variable R_k with realizations $r_k = -\frac{1}{2}\log_\tau\left(\frac{\delta_p^k}{\lambda}\right)$. Since $\tau < 1$ then it is obvious that $r_k < 0$ for all $k \geq k_0$. The main idea of the proof is to show that such realizations occur only with probability zero, hence obtaining a contradiction. Note that such a contradiction will be obtained later by constructing a random variable W_k with realizations w_k satisfying $r_k - r_{k_0} \geq w_k - w_{k_0}$ and for which $\left\{\limsup_{k\to+\infty} W_k = +\infty\right\}$ almost surely, thus implying that with probability one, R_k has to be positive infinitely often.

In order to show that R_k is a submartingale, recall the events J_k in the Definition 3 for some $\varepsilon_f > 0$ and consider some iteration $k \geq k_0$ for which J_k occurs, which happens with probability at least

$\beta > 1/2$. Now, noticing that (4.36) and (4.38) imply

$$f(x^k + s^k) - f(x^k) \leq -\epsilon(\gamma + 2)\delta_p^k \leq -\varepsilon_f(\gamma + 2)(\delta_p^k)^2, \text{ for all } k \geq k_0,$$

then for all $k \geq k_0$,

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

Hence, the k-th iteration of Algorithm 4 is successful, so the frame size parameter δ_p^k is updated according to $\delta_p^{k+1} = \tau^{-2}\delta_p^k$ since $\delta_p^k < \tau^{2-\hat{z}}$. Consequently, $r_{k+1} = r_k + 1$. Define the σ-algebra $\mathcal{F}_{k-1}^J$ by $\mathcal{F}_{k-1}^J = \sigma(J_0, J_1, \ldots, J_{k-1})$. If $\mathbb{1}_{J_k} = 0$, which occurs with probability at most $1 - \beta$, then the inequality $\delta_p^{k+1} \geq \tau^2\delta_p^k$ always holds, which implies that $r_{k+1} \geq r_k - 1$. Thus,

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

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

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

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

From the submartingale-like property enforced in Definition 3, it easily follows that W_k is a submartingale. Indeed,

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

Notice that $W_k \in \{\pm 1\}$ for all k, and hence, W_k has bounded increments, whence cannot have a finite limit. Thus, it follows from Theorem 3 that the event $\left\{\limsup_{k\to+\infty} W_k = +\infty\right\}$ occurs almost surely.

Since R_k and W_k are constructed in such a way that

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

with w_k denoting a realization of W_k, then with probability one, R_k has to be positive infinitely often. Consequently, the sequence of realizations r_k such that $r_k < 0$ for all $k \geq k_0$ occurs with probability zero. Thus, the assumption that (4.37) holds is false. This implies that

$$\mathbb{P}\left(\{\omega \in \Omega : \forall \mathcal{E}(\omega) > 0,\ \exists k \in \mathbb{N} \text{ such that } \Psi_k(\omega) < \mathcal{E}(\omega)(\gamma + 2)\}\right) = 1, \tag{4.39}$$

which means that (4.35) holds almost surely. □

The following definition of refining directions [12, 16] will be useful to our analysis.

Definition 6. *Given a convergent refining subsequence $\{x^k\}_{k\in\mathcal{K}}$ and its corresponding refined point $\hat{x}$, a direction d is said to be a* refining direction *$\hat{x}$ if and only if there exists an infinite subset $\mathcal{L} \subseteq \mathcal{K}$ with poll directions $d^k \in \mathbb{D}_p^k$ such that $\lim_{k\in\mathcal{L}} \frac{d^k}{\|d^k\|_\infty} = \frac{d}{\|d\|_\infty}$.*

Note that for all realizations of StoMADS, the existence of a refining direction d for a given refining subsequence $\{x^k\}_{k\in\mathcal{K}}$ and its corresponding refined point $\hat{x}$ is justified by the compactness of the unit closed ball.

We now state a useful result from [12] that provides a lower bound on the Clarke directional derivative.

Lemma 3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz near $\hat{x} \in \mathbb{R}^n$. Then the Clarke* generalized directional derivative *of f at $\hat{x}$ in the direction $d \in \mathbb{R}^n$ satisfies*

$$f^\circ(\hat{x}; d) \ := \ \limsup_{\substack{y\to\hat{x}\\ t\searrow 0}} \frac{f(y+td) - f(y)}{t} \ = \ \limsup_{\substack{x\to\hat{x}, v\to d\\ t\searrow 0}} \frac{f(x+tv) - f(x)}{t}.$$

The convergence property of the StoMADS algorithm will be shown using properties of the random function Ψ_k defined in Theorem 4; it is a stochastic variant of the result in [12]. It states that the Clarke generalized derivative of f at a refined point in any corresponding refining direction is nonnegative with probability one. We would like to mention that while the proof in [12] relies on the fact that the inequality $f(x^k + \delta_m^k d^k) - f(x^k) \geq 0$ always holds at every unsuccessful iteration, the idea of the proof in the present analysis is different since some unsuccessful iterations can be uncertain, in which case $f(x^k + \delta_m^k d^k) - f(x^k)$ belongs to the uncertainty interval $\mathcal{I}_{\gamma+2,\varepsilon_f}(\delta_p^k)$.

Theorem 5. *(Convergence of* StoMADS*). Let the assumptions of Theorem 2 hold. Then, there exists an almost sure event V'' such that for all $\omega \in V''$, for all refined points $\hat{X}(\omega) \in \mathbb{R}^n$, and for all*

refining directions $D(\omega) \in \mathbb{R}^n$ for $\hat{X}(\omega)$, the generalized directional derivative of f at $\hat{X}(\omega)$ in the direction $D(\omega)$ is nonnegative, i.e.,

$$f^\circ\left(\hat{X}(\omega); D(\omega)\right) \geq 0. \tag{4.40}$$

Proof. It follows from Corollary 1 and Theorem 4 that the event

$$V'' := \left\{\omega \in \Omega : \lim_{k\to+\infty} \Delta_m^k(\omega) = 0\right\} \bigcap \left\{\omega \in \Omega : \exists K'(\omega) \subset \mathbb{N}, \lim_{k\in K'(\omega)} \Psi_k(\omega) \leq 0\right\}$$

is almost sure as countable intersection of almost sure events. Consider some arbitrary outcome $\omega \in V''$. Denote by $\mathcal{K}' = K'(\omega)$ and recall that $\delta_m^k = \Delta_m^k(\omega)$, $\delta_p^k = \Delta_p^k(\omega)$, $x^k = X^k(\omega)$, and $\psi_k = \Psi_k(\omega)$. Since $\lim_{k\in\mathcal{K}'} \delta_m^k = 0$, using similar arguments as in the proof of Theorem 2, there exists a subset $\mathcal{K} \subset \mathcal{K}'$ such that $\lim_{k\in\mathcal{K}} x^k = \hat{x}$. It then follows from the compactness of the closed unit ball that there exists a subset $\mathcal{L} \subset \mathcal{K}$ such that the normalized subsequence $\left\{d^k/\left\|d^k\right\|_\infty\right\}_{k\in\mathcal{L}}$ of polling directions used by StoMADS converges to a limit $d/\|d\|_\infty = D(\omega)/\|D(\omega)\|_\infty$ while $\lim_{k\in\mathcal{L}} \psi_k \leq 0$.

Since $\delta_p^k\left\|d^k\right\|_\infty$ does not approach zero even though $\lim_{k\in\mathcal{L}} \delta_p^k = 0$, the following holds

$$\lim_{k\in\mathcal{L}} \left(\frac{-\psi_k}{\delta_p^k \|d^k\|_\infty}\right) = \lim_{k\in\mathcal{L}} \frac{f(x^k + \delta_m^k d^k) - f(x^k)}{\delta_m^k \|d^k\|_\infty} \geq 0. \tag{4.41}$$

Then, by applying Lemma 3 using the sequences $x^k \to \hat{x}$, $d^k/\left\|d^k\right\|_\infty \to d/\|d\|_\infty$, and $\delta_m^k\left\|d^k\right\|_\infty \searrow 0$, the following holds for the generalized derivative of f:

$$\begin{aligned} f^\circ\left(\hat{X}(\omega); \frac{D(\omega)}{\|D(\omega)\|_\infty}\right) = f^\circ\left(\hat{x}; \frac{d}{\|d\|_\infty}\right) &= \limsup_{x\to\hat{x}, v\to d/\|d\|_\infty, t\searrow 0} \frac{f(x+tv) - f(x)}{t} \\ &\geq \limsup_{k\in\mathcal{L}} \frac{f\left(x^k + \delta_m^k\left\|d^k\right\|_\infty \frac{d^k}{\left\|d^k\right\|_\infty}\right) - f(x^k)}{\delta_m^k \|d^k\|_\infty} \\ &\geq \lim_{k\in\mathcal{L}} \frac{f\left(x^k + \delta_m^k\left\|d^k\right\|_\infty \frac{d^k}{\left\|d^k\right\|_\infty}\right) - f(x^k)}{\delta_m^k \|d^k\|_\infty} \geq 0, \end{aligned} \tag{4.42}$$

where the last inequality in (4.42) follows from (4.41).

□

4.4 Computational study

The performance of StoMADS is investigated numerically using noisy variants of the objective functions of 22 deterministic unconstrained optimization problems from the CUTEst collection [55]. All objective functions are given by analytical expressions. Several variants of StoMADS are compared to Robust-MADS [18], which is currently the only noisy blackbox optimization algorithm available in the NOMAD [69] software package (version 3.9.1). We will refer to Robust-Mads as NOMAD-robust in the remainder of the text. We use only the POLL step for all numerical investigations of StoMADS and NOMAD-robust, i.e., we do not make use of the optional SEARCH step. We use the OrthoMADS $2n$ directions [2] option for the POLL, which is ordered by means of an opportunistic strategy [16] while disabling the quadratic models [40] and the anisotropic mesh [19]. The MADS algorithm [12] with the SEARCH step disabled is referred to as NOMAD-basic. The default algorithm in NOMAD is referred to as NOMAD-default. The compared algorithms are summarized in Table 4.1. Note that all proposed variants of StoMADS are implemented in MATLAB.

Table 4.1 Summary of the compared algorithms

Algorithm	Direction option	Anisotropic mesh	Opportunistic strategy	Quadratic models
StoMADS	OrthoMADS $2n$	No	Yes	No
NOMAD-robust	OrthoMADS $2n$	No	Yes	No
NOMAD-basic	OrthoMADS $2n$	No	Yes	No
NOMAD-default	OrthoMADS $n+1$ [17]	Yes	Yes	Yes

The stochastic variants of the 22 CUTEst unconstrained optimization problems used in [77] are solved using three different initial points for a total of 66 problem instances whose size ranges from $n = 2$ to $n = 12$. Their objective functions are sums of squares, i.e.,

$$f(x) = \sum_{i=1}^{m} (f_i(x))^2$$

with $f_i(x)$ being a smooth function for all $i \in \{1, 2, \ldots, m\}$.

The type of noise that is tested is referred to as "*additive*" noise, i.e., each f_i is additively perturbed by some random variable Θ_i generated uniformly in the interval $I(\sigma, x^0, f^*)$ defined by $I(\sigma, x^0, f^*) = [-\sigma\,|f(x^0) - f^*|\,, \sigma\,|f(x^0) - f^*|]$, i.e.,

$$f_\Theta(x) = \sum_{i=1}^{m} (f_i(x) + \Theta_i)^2, \tag{4.43}$$

where $\sigma > 0$ is a constant used to define different noise levels in the blackbox f_Θ. The random

variables Θ_i, $i \in \{1, 2, \ldots, m\}$, are assumed to be independent; x^0 denotes the initial point and f^* is the best known minimum value of f. Note that although $\mathbb{E}_\Theta[f_\Theta(x)] = f(x) + \sum_{i=1}^m \mathbb{E}[(\Theta_i)^2]$, optimization results are not affected by this constant bias term since $\min_x \mathbb{E}_\Theta[f_\Theta(x)] = \min_x f(x)$.

The NOMAD-robust algorithm to which StoMADS is compared is a smoothing-based algorithm designed to handle noisy blackbox optimization problems. At each iteration of NOMAD-robust, a best mesh local optimizer is determined based on values of the smoothed version of the noisy available objective constructed from a list of trial points and making use of a Gaussian kernel [18]. This list is then updated with the best iterate found before the next iteration of the algorithm in order for the smoothed function quality to increase [18], since the blackbox is evaluated by NOMAD-robust at each point only once. Although experiments in [18] have been conducted on deterministically noisy problems, the smoothing-based technique does not depend on the link between the objective function f and its noisy available version, which means that NOMAD-robust is supposed to be able to handle stochastically noisy problems.

To compare the results and the performance of the algorithms, data profiles [77] and performance profiles [49, 77] are presented using the following convergence test:

$$f(x^N) \leq f(x^*) + \tau(f(x^0) - f(x^*)), \tag{4.44}$$

where, for each of the 66 problems, x^N denotes the best point found by an algorithm after N function calls to the noisy objective f_Θ, x^* is the best known solution and $\tau \in [0, 1]$ is the convergence tolerance. Thus, a problem is said to be solved within the convergence tolerance τ if (4.44) holds.

The horizontal axis of the data profiles shows the number of noisy function evaluations divided by $n + 1$ while the vertical axis shows the portion of problems solved within a given convergence tolerance τ. The horizontal axis of the performance profiles shows the ratio of the number of function calls to the noisy blackbox while the vertical axis shows the portion of problems solved within the tolerance τ. In all the experiments, a budget of $1000(n + 1)$ noisy function evaluations is set, i.e all algorithms stop as soon as the number of function calls to f_Θ reaches $1000(n + 1)$.

For the initialization, the same common parameter values to both methods are used: $\delta_m^0 = \delta_p^0 = 1$ and the mesh refining parameter $\tau = 1/2$. The StoMADS parameters γ and ε_f are set equal to 17 and 0.01, respectively, to satisfy $\gamma > 2$ and $\varepsilon_f > 0$. Note that NOMAD-robust is not in line with the theory presented in this work regarding the choice of the sample size p^k, especially in terms of sample sizes. In fact as mentioned earlier, the blackbox is not evaluated more than once by NOMAD-robust at each point, while it needs to be evaluated at least p^k times by StoMADS at each point in order to construct the estimates $f_0^k = \frac{1}{p^k} \sum_{i=1}^{p^k} f_{\theta_{0,i}}(x^k) \approx f(x^k)$ and $f_s^k = \frac{1}{p^k} \sum_{i=1}^{p^k} f_{\theta_{s,i}}(x^k + s^k) \approx f(x^k + s^k)$, where $\theta_{0,i}$ and $\theta_{s,i}$, $i \in \{1, 2, \ldots, p^k\}$, are the realizations of the random variables $\Theta_{0,i}$ and $\Theta_{s,i}$, respectively

(introduced in Section 4.2.3).

This latter remark and the need for p^k to be large in order for the estimates to be sufficiently accurate present a challenge in obtaining satisfactory solutions within an allocated budget. Recall that n^k denotes the number of blackbox evaluations at a given point when constructing an estimate at iteration k. Five variants of StoMADS corresponding to $n^k = 1$, $n^k = 2$, $n^k = 3$, $n^k = 4$, and $n^k = 5$, respectively, for all k are therefore compared to NOMAD-robust, NOMAD-basic, and NOMAD-default. Despite the fact that the resulting values of p^k do not meet the theoretical prescription derived in Section 4.2.3, they seemed to work well enough compared to many various other choices of n^k that have been tested. However, in order to increase accuracy of function value estimates using limited blackbox evaluations, the procedure described in Section 4.2.3 is used. Recall that it improves the accuracy of the estimates by making use of previously computed estimates at the current iterate. When the iteration k is successful, the estimate f_0^{k+1} of $f(x^{k+1})$ is computed according to (4.7), while after an unsuccessful iteration k, f_0^{k+1} is given by (4.8).

Three levels of noise are considered in the numerical experiments: $\sigma = 0.01$, $\sigma = 0.03$, and $\sigma = 0.05$. These values were chosen arbitrarily to study if and how the portion of problems solved successfully by StoMADS varies. Consider, for example, the two-dimensional version of the Rosenbrock function given by

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \tag{4.45}$$

with the initial point $x^0 = (-1.2, 1)$ and the minimum value $f^* = 0$. Then, $|f(x^0) - f^*| = 24.2$ and the corresponding noisy function is given by

$$f_\Theta(x) = \left[10(x_2 - x_1^2) + \Theta_1\right]^2 + \left[(1 - x_1) + \Theta_2\right]^2, \tag{4.46}$$

where Θ_1 and Θ_2 are independent random variables uniformly generated in the interval $I(\sigma, x^0, f^*) = [-24.2\sigma, 24.2\sigma]$. Figure 4.3 depicts $f(x)$ and realizations of $f_\Theta(x)$. Figures 4.4, 4.6, and 4.8 and Figures 4.5, 4.7, and 4.9 present the data and performance profiles, respectively, to compare the five variants of StoMADS with NOMAD-robust, NOMAD-basic, and NOMAD-default at various noise levels and convergence tolerances.

These data and performance profiles show that, in general, StoMADS outperforms NOMAD-robust and both deterministic blackbox optimization algorithms NOMAD-basic and NOMAD-default (which are obviously not appropriate for stochastic optimization). Varying the value of the tolerance parameter τ in the performance profiles does not significantly alter the conclusions drawn from the data profiles. It can be observed that for a given τ, the higher the noise level, the lower the fraction of problems solved for most variants of StoMADS (as expected). Since the variance of the blackbox increases with the noise level, it follows from Section 4.2.3 that the estimates need to be sufficiently

accurate to generate satisfactory solutions and consequently allow the solution of a larger fraction of problems. Similarly, for a fixed noise level, the higher the convergence tolerance, the larger the fraction of problems solved by most algorithms.

Even though the number n^k of blackbox evaluations is constant from one iteration to another for a given variant of StoMADS, this is not the case for the sample size p^k involved in the computation of the estimates. Indeed, it follows respectively from (4.7) and (4.8) that $p^{k+1} = 2n^k$ when the iteration k is successful while $p^{k+1} = p^k + n^{k+1}$ when it is unsuccessful. Thus, even though the efficiency of each StoMADS variant depends on its corresponding evaluation parameter n^k, the quality of the solutions that are generated is influenced by the sample rate p^k, which is not constant. This explains why varying the blackbox evaluation parameter n^k from one to five does not necessarily improve the performance of the corresponding StoMADS variants. Note that this also explains why the behavior of the StoMADS variant corresponding to $n^k = 1$ is not similar to that of MADS. Indeed, no computation of estimates is carried out in MADS and moreover, MADS is unable to show how an improvement in a noisy blackbox can lead to a decrease in an unavailable objective function unlike StoMADS.

We can conclude, especially based on the profiles corresponding to the tolerance $\tau = 10^{-3}$, that StoMADS can handle the optimization of stochastically noisy blackboxes that are expensive in term of blackbox evaluations, since its variants corresponding to $n^k = 1$ and $n^k = 2$ are able to generate satisfactory solutions with few blackbox evaluations. However, the choice $n^k = 4$ seems to be preferable for stochastic blackbox optimization problems with higher evaluations budgets.

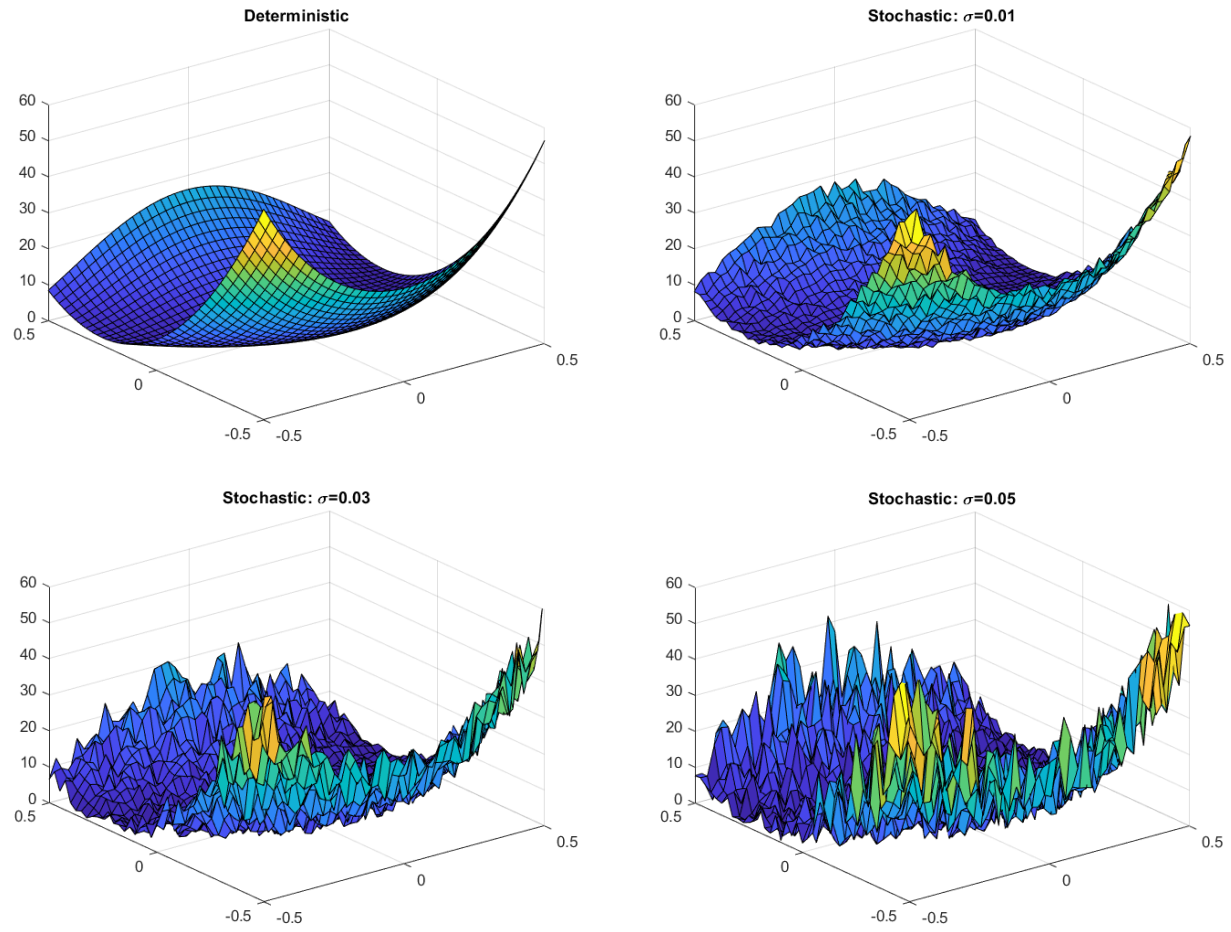
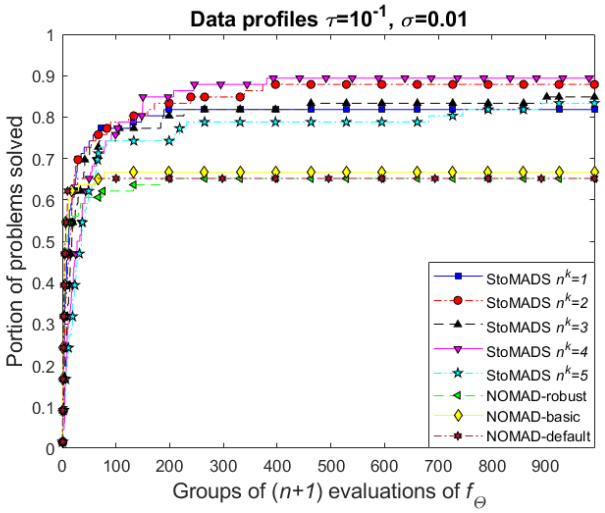


Figure 4.3 Plots of the deterministic Rosenbrock function (4.45) and corresponding realizations of $f_\Theta(x)$ (4.46) on the box $[-0.5, 0.5] \times [-0.5, 0.5]$. The random variables defining the noisy functions f_Θ are uniformly generated in $[-24.2\sigma, 24.2\sigma]$.



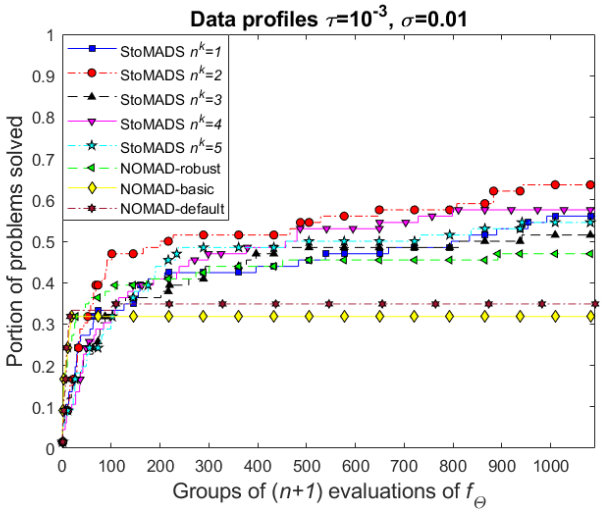
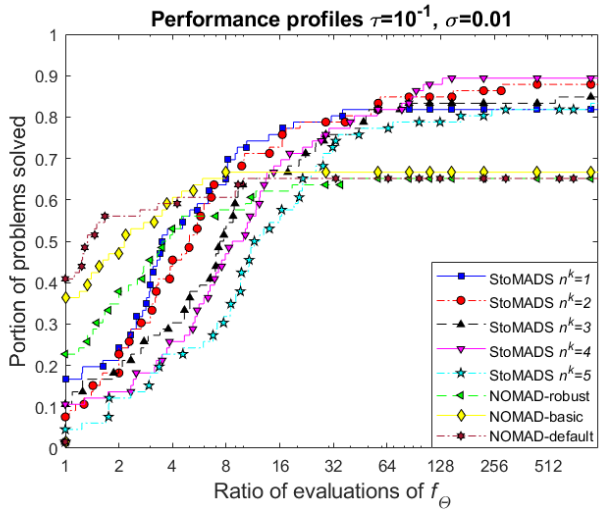


Figure 4.4 Data profiles for noise level $\sigma = 0.01$ and convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$ on 66 analytical unconstrained test problems additively perturbed in the interval $I(\sigma, x^0, f^*)$.



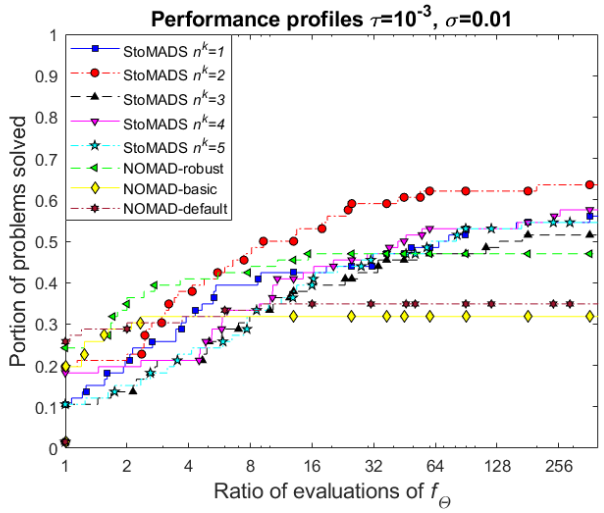
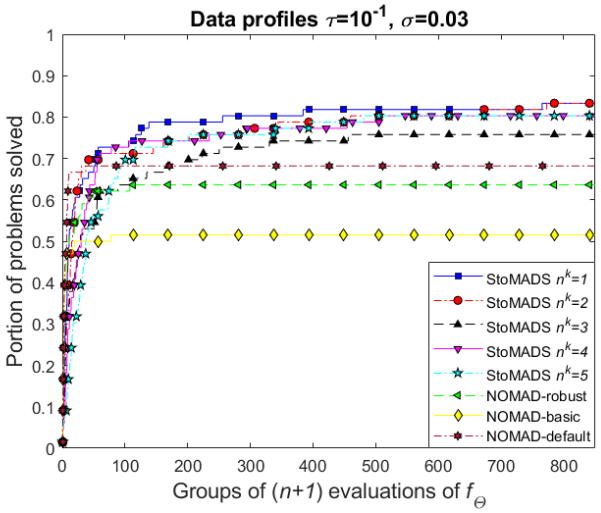
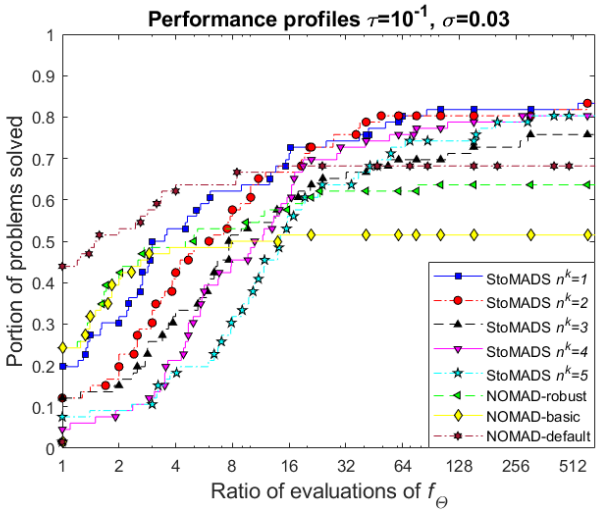
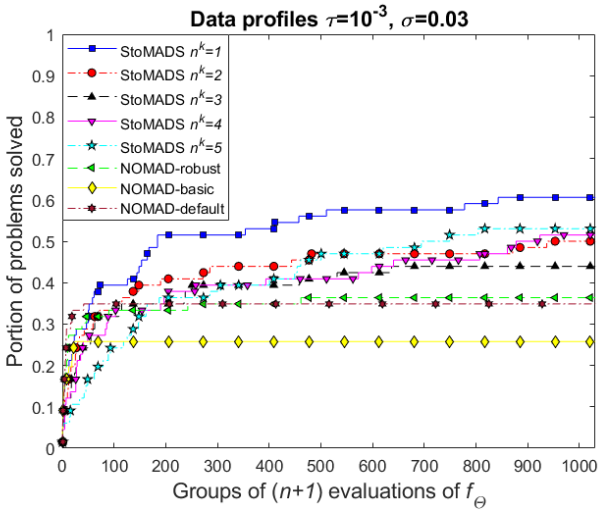


Figure 4.5 Performance profiles for noise level $\sigma = 0.01$ and convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$ on 66 analytical unconstrained test problems additively perturbed in the interval $I(\sigma, x^0, f^*)$.







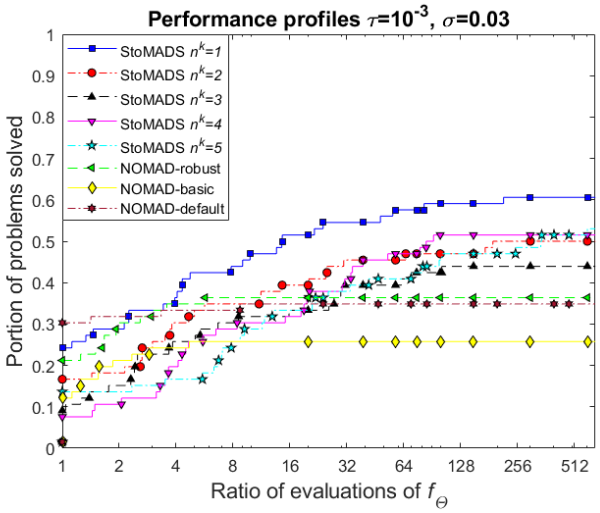
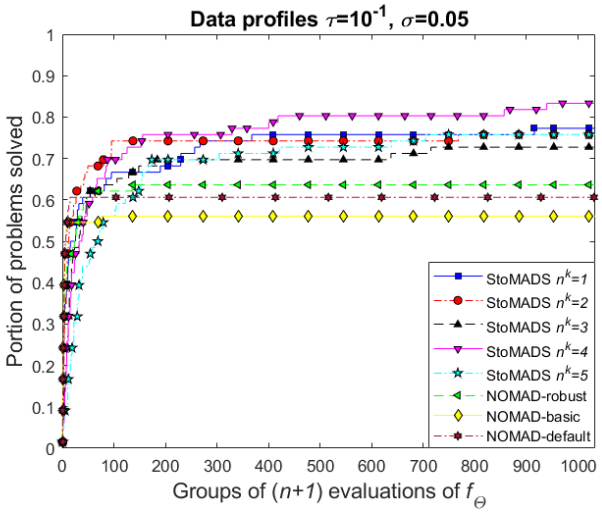
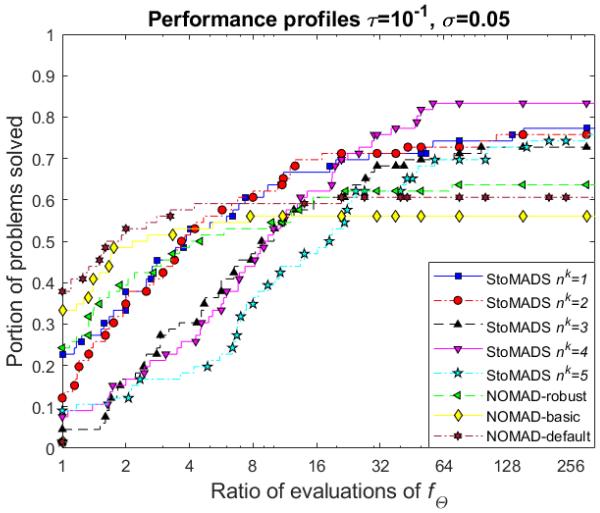
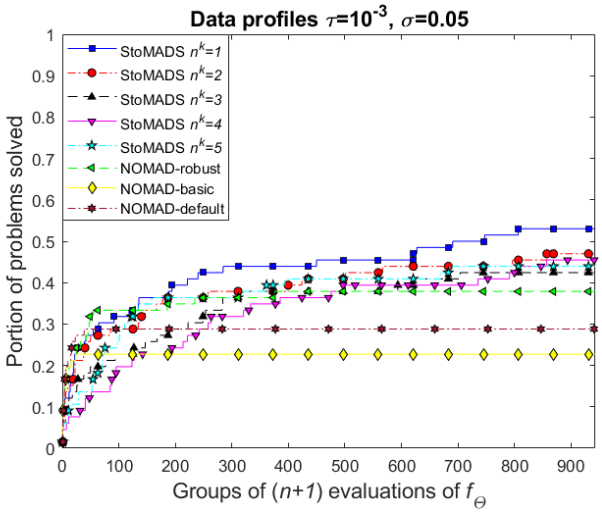


Figure 4.6 Data profiles for noise level $\sigma = 0.03$ and convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$ on 66 analytical unconstrained test problems additively perturbed in the interval $I(\sigma, x^0, f^*)$.

Figure 4.7 Performance profiles for noise level $\sigma = 0.03$ and convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$ on 66 analytical unconstrained test problems additively perturbed in the interval $I(\sigma, x^0, f^*)$.







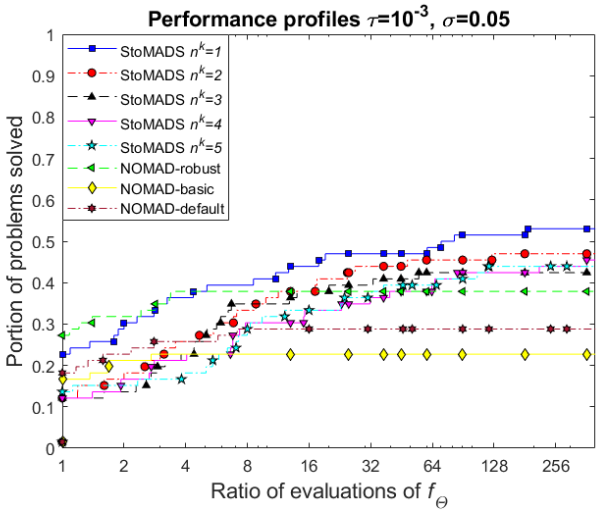


Figure 4.8 Data profiles for noise level $\sigma = 0.05$ and convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$ on 66 analytical unconstrained test problems additively perturbed in the interval $I(\sigma, x^0, f^*)$.

Figure 4.9 Performance profiles for noise level $\sigma = 0.05$ and convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$ on 66 analytical unconstrained test problems additively perturbed in the interval $I(\sigma, x^0, f^*)$.

4.5 Concluding remarks

While MADS is a valuable blackbox optimization algorithm supported by convergence results, it was developed for solving deterministic problems. Even though Robust-MADS, the first variant of MADS modified to solve noisy blackbox optimization, was shown to have *zeroth-order* convergence properties, the corresponding work [18] did not show whether an improvement in the smoothed version of the noisy available blackbox, used to update the iterates, would result in a decrease in the unknown objective.

Thus, unlike Robust-MADS, the StoMADS algorithm presented in this paper shows how an improvement in the estimates of the unavailable objective function values may cause a decrease in the unavailable objective function. This is achieved by defining new iteration types by means of a sufficient decrease condition on these estimates that are required to be probabilistically sufficiently accurate.

Although the convergence analysis of StoMADS uses ideas derived from that of MADS, the analysis itself is different and based on stochastic processes. In addition to the convergence result of the whole sequence of random mesh size parameters, which is stronger than the $\liminf$-type result of MADS, a more general existence proof of *refining* subsequences consisting of StoMADS iterates that are not necessarily mesh local optimizers has been proposed, followed by a stochastic variant of the Clarke optimality result of MADS.

An extensive computational study of several variants of StoMADS on a collection of unconstrained stochastically noisy problems shows that the proposed method outperforms Robust-MADS and also highlights the fact that MADS is not appropriate for stochastic blackbox optimization, even though the accuracy of StoMADS estimates does not meet the prescription derived theoretically.

While all previous works used theoretical principles that are similar to the ones used in this paper, the present research is, to the best of our knowledge, the first that does not require model or gradient information to find descent directions.

Future research will focus on extending this approach to blackbox optimization problems involving stochastically noisy constraints.

Acknowledgments

The authors are grateful to Erick Delage from HEC Montréal and Richard Labib 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.

CHAPITRE 5 ARTICLE 2: EXPECTED COMPLEXITY ANALYSIS OF STOCHASTIC DIRECT-SEARCH

Kwassi Joseph Dzahini. Expected complexity analysis of stochastic direct-search. Submitted for publication to Computational Optimization and Applications (COAP).

Abstract: This work presents the convergence rate analysis of stochastic variants of the broad class of direct-search methods of directional type. It introduces an algorithm designed to optimize differentiable objective functions f whose values can only be computed through a stochastically noisy blackbox. The proposed stochastic directional direct-search (SDDS) algorithm accepts new iterates by imposing a sufficient decrease condition on so called probabilistic estimates of the corresponding unavailable objective function values. The accuracy of such estimates is required to hold with a sufficiently large but fixed probability β. The analysis of this method utilizes an existing supermartingale-based framework proposed for the convergence rates analysis of stochastic optimization methods that use adaptive step sizes. It aims to show that the expected number of iterations required to drive the norm of the gradient of f below a given threshold ϵ is bounded in $\mathcal{O}\left(\epsilon^{\frac{-p}{\min(p-1,1)}}/(2\beta-1)\right)$ with $p > 1$. Unlike prior analysis using the same aforementioned framework such as those of stochastic trust-region methods and stochastic line search methods, SDDS does not use any gradient information to find descent directions. However, its convergence rate is similar to those of both latter methods with a dependence on ϵ that also matches that of the broad class of deterministic directional direct-search methods which accept new iterates by imposing a sufficient decrease condition.

Keywords: Blackbox optimization, Derivative-free optimization, Stochastic optimization, Convergence rate, Direct-search, Stochastic processes.

5.1 Introduction

Direct-search methods constitute a broad class of derivative-free optimization (DFO) methods where at each iteration, the DFO algorithm evaluates the objective function at a collection of points and acts solely based on those function values without any model building or derivative approximation [16,42]. Such methods include as well those based on simplices like the classical Nelder-Mead method and its numerous variants, as those of directional type where an improvement in the objective function is guaranteed by moving along a direction defined by a better point [95].

This work focuses on the convergence rate analysis of stochastic variants of the broad class of directional direct-search methods analyzed in [95], using a supermartingale-based framework proposed in [29] and elements from [15, 24, 33, 65, 83]. It introduces a stochastic directional direct-search (SDDS) algorithm designed for stochastic blackbox optimization (BBO) and aims to solve the following unconstrained stochastic blackbox optimization problem which often arises in modern statistical machine learning:

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{with} \quad f(x) = \mathbb{E}_\Theta[f_\Theta(x)] \tag{5.1}$$

where Θ is a real-valued random variable following some unknown distribution, f_Θ denotes the blackbox, the stochastically noisy computable version of the objective function $f : \mathbb{R}^n \to \mathbb{R}$ which is numerically unavailable and $\mathbb{E}_\Theta$ denotes the expectation with respect to Θ. Note that Θ is considered as a data point for many machine learning problems [83].

Significant theoretical and algorithmic advances have been made in the field of stochastic DFO in the recent years with the aim of solving Problem (5.1). Thus, numerous algorithms have been developed, most of which carry out either an estimation of the gradient of f using a single simulation, or a processing of the simulation model as a blackbox. However, since the simulation model can be inaccessible in many real applications, or the gradient can be too expensive to estimate computationally, direct-search optimization methods "appear to be the most promising option" [15].

Several recent works have proposed directional direct-search algorithms with full supported convergence rates analysis. Vicente [95] proved that to drive the gradient of an objective function below a threshold $\epsilon \in (0,1)$, the number of iterations required by the broad class of directional direct-search methods that use a sufficient decrease condition when accepting new iterates, is bounded in $\mathcal{O}\left(\epsilon^{\frac{-p}{\min(p-1,1)}}\right)$, with $p > 1$. Directional direct-search methods based on *probabilistic descent*, that incorporate *random gradient*, was recently proposed and analyzed by Gratton et al. [57] with worst-case complexity, and global rates results. However, both latter works assume that the objective function is deterministic, i.e., function values are exactly computed.

Audet et al. [15] recently proposed StoMADS, a stochastic variant of the mesh adaptive direct-search

(MADS) algorithm [12], with convergence analysis based on Clarke calculus and martingale theory. Alarie et al. [3] also proposed another variant of MADS capable to optimize noisy blackboxes corrupted with Gaussian noise, and proved convergence results using statistical inference techniques. Nevertheless, no convergence rates analysis have been carried out for both methods.

The main novelty of the present work is that unlike many prior research on convergence rate analysis of stochastic DFO methods (see for example [26, 29, 83, 96] and references therein), especially those on stochastic trust-region [29] and line search [83] methods, SDDS does not use any first-order information to find descent directions. Rather, such directions are provided by a positive spanning set and are chosen so that they never become close to losing the positive spanning property. However, as emphasized in [95], "*it is not unreasonable*" to expect that SDDS shares a similar worst case complexity bound of the latter methods in term of the expected number of iterations. Indeed, one of the directions of any positive spanning set makes an acute angle with the negative gradient, provided that the objective function is continuously differentiable [65, 95]. This latter remark is in fact the cornerstone of the analysis in the present manuscript. Moreover, unlike the deterministic framework [95], SDDS accepts new iterates by imposing a sufficient decrease condition on so called *probabilistic estimates* of the corresponding unavailable objective function values, which accuracy is required to hold with a sufficiently large but fixed probability $\beta > 1/2$. However, even though β is not required to equal one, SDDS is shown to have desirable convergence properties. Specifically, as main theoretical result of the present work, the expected number of iterations required by SDDS to drive the gradient of f below a threshold ϵ is shown to be bounded in $\mathcal{O}\left(\epsilon^{\frac{-p}{\min(p-1,1)}}/(2\beta-1)\right)$, using a supermartingale-based framework proposed in [29]. Moreover, a subsequence of random iterates generated by SDDS is shown to drive the norm of the gradient of f to zero with probability one. Finally, note that the analysis in the present manuscript is not limited to $p = 2$ as is the case for several similar works, but rather extends to $p > 1$. To the best of our knowledge, this research is the first to propose a convergence rate analysis of a stochastic direct-search algorithm of directional type.

This manuscript is organized as follows. Section 5.2 introduces an outline of the proposed stochastic algorithm and requirements on so-called probabilistic estimates that guarantee convergence at an appropriate rate. Section 5.3 presents a general framework of a stochastic process that is required for the convergence rate analysis in Section 5.4. The latter section presents also a $\liminf$-type first-order convergence result for SDDS, followed by a discussion and suggestions for future work.

5.2 The SDDS method and probabilistic estimates

This section introduces the general framework of SDDS, probabilistic estimates and then discusses the requirements on such estimates that guarantee the convergence of the algorithm.

5.2.1 The SDDS algorithm

SDDS uses an algorithmic framework similar to that of the broad class of methods analyzed in [95], i.e., a framework that can describe the main features of generating set search (GSS) [65], pattern search and generalized pattern search (GPS) [11].

Each iteration of a directional direct-search method is composed of two main steps: the SEARCH step which is optional and the POLL step on which relies the convergence analysis. For simplicity of presentation, Algorithm 6 does not show any SEARCH step. During the POLL, trial points are generated in a subset $\mathcal{P}^k = \{x^k + \delta^k d : d \in \mathbb{D}^k\}$ of the space of variables, where x^k denotes the incumbent solution, δ^k the step size and $\mathbb{D}^k$ is a positive spanning set [16, 42]. Thus, the POLL step which follows stricter rules, consists of a local exploration of the variables space, unlike the SEARCH step which consists of a global exploration.

In Algorithm 6, since objective function values $f(x)$ are unavailable, f_0^k and f_s^k denote respectively the estimates of $f(x^k)$ and $f(x^k + s^k)$ (with $s^k = \delta^k d$), constructed making use of evaluations of the noisy blackbox f_Θ. In order for the information provided by f_0^k and f_s^k to determine the iteration type, i.e., successful or unsuccessful, both estimates are required to be $\varepsilon_{f,p}$-accurate, with $\varepsilon_{f,p} > 0$, according to the following definition similar to those in [15, 29, 35, 83].

Definition 7. *Let $\rho_p : (0, +\infty) \to (0, +\infty)$ be a continuous and non-decreasing function satisfying $\rho_p(t)/t \to 0$ when $t \searrow 0$. f^k is called $\varepsilon_{f,p}$-accurate estimate of $f(x^k)$ for a given δ^k if*

$$\left| f^k - f(x^k) \right| \leq \varepsilon_{f,p} \rho_p(\delta^k).$$

Following the terminology in [65], the function ρ_p in Definition 7 represents the "*forcing function*". Sufficient information to determine the iteration type is provided next.

Proposition 3. *Let f_0^k and f_s^k be $\varepsilon_{f,p}$-accurate estimates of $f(x^k)$ and $f(x^k + s^k)$ respectively, and let $\gamma > 2$ be a fixed constant. Then the followings hold:*

$$\textit{if} \quad f_s^k - f_0^k \leq -\gamma \varepsilon_{f,p} \rho_p(\delta^k), \ \textit{then} \ f(x^k + s^k) - f(x^k) \leq -(\gamma - 2)\varepsilon_{f,p} \rho_p(\delta^k) := u_{\text{sto}}^k \quad (5.2)$$

$$\textit{if} \quad f_s^k - f_0^k > -\gamma \varepsilon_{f,p} \rho_p(\delta^k), \ \textit{then} \ f(x^k + s^k) - f(x^k) > -(\gamma + 2)\varepsilon_{f,p} \rho_p(\delta^k) := \ell_{\text{sto}}^k \quad (5.3)$$

Proof. The proof straightforwardly follows from Definition 7 and the equality

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

□

In addition to the results in Proposition 3, the definition of the iteration type in Algorithm 6 is motivated by the following remarks. First, recall the framework of StoMADS where $\rho_p(t) = t^2$ and its so-called *frame* size parameter δ_p^k, and the fact that while its *sufficient decrease condition* $f_s^k - f_0^k \leq -\gamma\varepsilon_{f,p}\rho_p(\delta_p^k)$ leads to a decrease in f, the inequality $f_s^k - f_0^k > -\gamma\varepsilon_{f,p}\rho_p(\delta_p^k)$ on the contrary does not necessarily lead to an increase just like in (5.3). Indeed, two situations can be distinguished: on the one hand, the inequality $f_s^k - f_0^k \geq \gamma\varepsilon_{f,p}\rho_p(\delta_p^k)$ implies that $f(x^k + s^k) - f(x^k) \geq (\gamma - 2)\varepsilon_{f,p}\rho_p(\delta_p^k) > 0$ while on the other hand, the fact that $f_s^k - f_0^k \in (-\gamma\varepsilon_{f,p}\rho_p(\delta_p^k), \gamma\varepsilon_{f,p}\rho_p(\delta_p^k))$ leads to $-(\gamma + 2)\varepsilon_{f,p}\rho_p(\delta^k) < f(x^k + s^k) - f(x^k) < (\gamma + 2)\varepsilon_{f,p}\rho_p(\delta^k)$. Thus, there are two unsuccessful iterations for StoMADS. In the former situation, the unsuccessful iteration is called *certain* while it is called *uncertain* in the latter. Then, even though updating the frame size parameter according to $\delta_p^{k+1} = \tau\delta_p^k$ on *uncertain* unsuccessful iterations (where $\tau \in (0, 1)$ is a rational number), and $\delta_p^{k+1} = \tau^2\delta_p^k$ whenever the unsuccessful iteration is *certain*, the corresponding sequence $\{\delta_p^k\}_{k\in\mathbb{N}}$ was shown in [15] to converge to zero. Note also that this kind of update is the only one that differentiates *certain* iterations from those that are *uncertain*. In the present work, the step size parameter δ^k is updated on unsuccessful iterations according to $\delta^{k+1} = \tau\delta^k$, where $\tau \in (0, 1)$ is a real number. As a consequence, *certain* unsuccessful iterations will not be differentiated from *uncertain* ones. In other words, every iteration such that $f_s^k - f_0^k > -\gamma\varepsilon_{f,p}\rho_p(\delta^k)$ will be called unsuccessful.

However, let put an emphasis on the specific choice of τ by means of the following additional remarks. Note that in the general deterministic framework described in [95], the amount of decrease in the objective function on successful iterations is such that $f(x^k + s^k) - f(x^k) \leq -\rho_p(\delta^k) := u_{\text{det}}^k$ while unsuccessful iterations are characterized by $f(x^k + s^k) - f(x^k) > -\rho_p(\delta^k) := \ell_{\text{det}}^k$. Thus, the equality $\ell_{\text{det}}^k = u_{\text{det}}^k$ always holds, which is not the case in stochastic settings where $\ell_{\text{sto}}^k < u_{\text{sto}}^k$. Moreover, since $\delta^{k+1} < \delta^k$ whenever the iteration k is unsuccessful, then $\ell_{\text{det}}^{k+1} > u_{\text{det}}^k$. Likewise, since $\delta^{k+1} > \delta^k$ on successful iterations, then $u_{\text{det}}^{k+1} < \ell_{\text{det}}^k$. Given that the equality $\ell_{\text{sto}}^k = u_{\text{sto}}^k$ can not hold in the present stochastic settings, then τ must be chosen in such a way that at least, both inequalities $\ell_{\text{sto}}^{k+1} > u_{\text{sto}}^k$ and $u_{\text{sto}}^{k+1} < \ell_{\text{sto}}^k$ hold respectively on unsuccessful and successful iterations, analogously to the deterministic framework. This means using (5.2) and (5.3) that τ must be chosen according to

$$\rho_p(\tau\delta^k) < \frac{\gamma - 2}{\gamma + 2}\rho_p(\delta^k) \quad \text{and} \quad \rho_p(\tau^{-1}\delta^k) > \frac{\gamma + 2}{\gamma - 2}\rho_p(\delta^k). \tag{5.4}$$

It follows from (5.4) that depending on the expression of the forcing function ρ_p, the choice of τ could depend on δ^k and hence should be made at each iteration. Thus, in order to make the present analysis simpler, the following assumption is made.

Assumption 4. *The forcing function $\rho_p\colon (0, +\infty) \to (0, +\infty)$ is such that $\rho_p(t) = ct^p$, where $c > 0$ and $p > 1$ are fixed constants.*

Under Assumption 4, the choice of τ does not depend on δ^k. More precisely, it follows from (5.4) that τ must be chosen according to $0 < \tau^p < \frac{\gamma-2}{\gamma+2}$, for all $k \in \mathbb{N}$, as specified in Algorithm 6.

Algorithm 6: SDDS

[0]-Initialization

Choose $x^0 \in \mathbb{R}^n$, $\delta^0 > 0$, $\varepsilon_{f,p} > 0$, $\gamma > 2$, $c > 0$, $p > 1$, $0 < \tau < \left(\frac{\gamma-2}{\gamma+2}\right)^{1/p}$, $j_{\max} \in \mathbb{N}$ and $\delta_{\max} = \tau^{-j_{\max}}\delta^0$.
Set the iteration counter $k \leftarrow 0$.

[1]-Poll

Select a positive spanning set $\mathbb{D}^k$.
Generate a set $\mathcal{P}^k$ of Poll points such that $\mathcal{P}^k = \{x^k + \delta^k d : d \in \mathbb{D}^k\}$.
Obtain estimates f_0^k and f_s^k of $f(x^k)$ and $f(x^k + s^k)$, respectively at x^k and $x^k + s^k \in \mathcal{P}^k$ using objective function evaluations.

Success

If $f_s^k - f_0^k \leq -\gamma c \varepsilon_{f,p} (\delta^k)^p$ for some $s^k \in \{\delta^k d : d \in \mathbb{D}^k\}$,
Set $x^{k+1} \leftarrow x^k + s^k$, and $\delta^{k+1} \leftarrow \min\{\tau^{-1}\delta^k, \delta_{\max}\}$.

Failure

Otherwise set $x^{k+1} \leftarrow x^k$ and $\delta^{k+1} \leftarrow \tau\delta^k$.

[2]-Termination

If no termination criterion is met,
Set $k \leftarrow k + 1$ and go to **[1]**.
Otherwise stop.

Figure 5.1 The SDDS algorithm

5.2.2 Probabilistic estimates

Following the notation in [28], all stochastic quantities in the present manuscript live on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a nonempty set referred to as the sample space, $\mathcal{F}$ is a collection of events (subsets of Ω) called a σ-field and $\mathbb{P}$ is a finite measure on the measurable space $(\Omega, \mathcal{F})$ satisfying $\mathbb{P}(\Omega) = 1$ and referred to as probability measure. The elements $\omega \in \Omega$ are referred to as possible outcomes or sample points. When $\mathbb{R}^n$ is given its Borel σ-field $\mathcal{B}(\mathbb{R}^n)$, i.e., the one generated by the open sets, a random variable or random map X is a measurable map on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into the measurable space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Measurability meaning that each event $\{X \in B\} := X^{-1}(B)$ belongs to $\mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R}^n)$ [28].

The estimates f_s^k and f_0^k constructed at iteration k of Algorithm 6, based on random information provided by the noisy objective f_Θ, can be considered as realizations of random estimates F_s^k and F_0^k respectively. Thus, because of the randomness stemming from such random estimates whose behavior influences each iteration k, Algorithm 6 results in a stochastic process $\{X^k, S^k, \Delta^k, F_s^k, F_0^k\}$. In

general, uppercase letters will be used to denote random variables while lowercase letters will be used for their realizations. For example, $x^k = X^k(\omega)$, $s^k = S^k(\omega)$ and $\delta^k = \Delta^k(\omega)$ denote respectively realizations of the random variables X^k, S^k and Δ^k. Similarly, following the notations in [15, 29, 35, 83], $f_0^k = F_0^k(\omega)$ and $f_s^k = F_s^k(\omega)$ where F_0^k and F_s^k are respectively estimates of $f(X^k)$ and $f(X^k + S^k)$.

The goal of this work is to show that the stochastic process resulting from Algorithm 6 converges at an appropriate rate with probability one, provided that the sequence $\{(F_0^k, F_s^k)\}$ is sufficiently accurate with sufficiently high but fixed probability, *conditioned on the past* [29, 35, 83].

To formalize the notion of conditioning on the past, let $\mathcal{F}_{k-1}^F$ denotes the σ-field generated by F_0^0, F_s^0, $F_0^1, F_s^1, \ldots, F_0^{k-1}$ and F_s^{k-1}, and set $\mathcal{F}_{-1}^F$ to equal $\sigma(x^0)$ for completeness. As a result, $\left\{\mathcal{F}_k^F\right\}_{k\geq -1}$ is a filtration, i.e., an increasing subsequence of σ-fields of $\mathcal{F}$. Moreover, one can notice that $\mathbb{E}\left(\Delta^k | \mathcal{F}_{k-1}^F\right) = \Delta^k$ and $\mathbb{E}\left(X^k | \mathcal{F}_{k-1}^F\right) = X^k$ for all $k \geq 0$, since the random variables Δ^k and X^k are $\mathcal{F}_{k-1}^F$-measurable by construction in Algorithm 6.

The following definition similar to those in [15,29,35,83,96] is used to measure closeness or sufficient accuracy.

Definition 8. *A sequence of random estimates $\{(F_0^k, F_s^k)\}$ is said to be β-probabilistically $\varepsilon_{f,p}$-accurate with respect to the corresponding sequence $\{X^k, S^k, \Delta^k\}$ if the events*

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

satisfy the following submartingale-like condition

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

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

An estimate is called "good" if $\mathbb{1}_{J_k} = 1$. Otherwise it is called "bad" [15].

Global convergence properties of deterministic directional direct-search methods strongly rely on having the step size parameters approaching zero [95] and the fact that the function f values never increase between successive iterations. The main challenge of the analysis in the present stochastic framework lies in the fact that this monotonicity is not always guaranteed. The key to the analysis of Algorithm 6 therefore relies on the assumption that accuracy in function estimates "*improves in coordination with the perceived progress of the algorithm*" [29]. The analysis is based on properties of supermartingales whose increments have a decreasing tendency and depend on the change in objective

function values between iterations.

In order to show that the sequence $\{\Delta^k\}_{k\in\mathbb{N}}$ of random step size parameters converges to zero with probability one, the following key assumption similar to those in [15, 83] is made.

Assumption 5. *For some fixed $\beta \in (0,1)$, and $\varepsilon_{f,p} > 0$, the following holds for the random quantities derived from Algorithm 6.*

(i) The sequence $\{(F_0^k, F_s^k)\}$ of estimates is β-probabilistically $\varepsilon_{f,p}$-accurate.

(ii) The sequence $\{(F_0^k, F_s^k)\}$ satisfies the following variance condition

$$\begin{aligned} & \mathbb{E}\left(\left|F_0^k - f(X^k)\right|^2 \mid \mathcal{F}_{k-1}^F\right) \leq {\varepsilon_{f,p}}^2(1-\beta)[\rho_p(\Delta^k)]^2 \\ \textit{and} \quad & \mathbb{E}\left(\left|F_s^k - f(X^k+S^k)\right|^2 \mid \mathcal{F}_{k-1}^F\right) \leq {\varepsilon_{f,p}}^2(1-\beta)[\rho_p(\Delta^k)]^2 \end{aligned} \tag{5.5}$$

By means of Assumption 5-*(ii)*, the variance in function estimates is adaptively controlled. Showing therefore that the sequence of random step size parameters converges to zero with probability one, ensures that this variance is driven to zero even though the probability β of encountering good estimates remains fixed, thus allowing Algorithm 6 to behave like an exact deterministic method asymptotically.

Moreover, since the estimates satisfying Assumption 5 can easily be constructed using techniques proposed in [29, 35, 83], then thorough details about their computations are not provided here again. Note however that if both Θ_0 and Θ_s are independent random variables following the same distribution as Θ defined in (5.1), and if $\Theta_{0,i}$ and $\Theta_{s,i}$, $i = 1, 2, \ldots, p^k$ are independent random samples of Θ_0 and Θ_s respectively, then the estimates

$$F_0^k = \frac{1}{p^k}\sum_{i=1}^{p^k} f_{\Theta_{0,i}}(x^k) \quad \text{and} \quad F_s^k = \frac{1}{p^k}\sum_{i=1}^{p^k} f_{\Theta_{s,i}}(x^k + s^k)$$

satisfy Assumption 5 provided that the sample size p^k satisfies

$$p^k \geq \frac{V}{{\varepsilon_{f,p}}^2(1-\sqrt{\beta})[\rho_p(\delta^k)]^2},$$

where the constant $V > 0$ is such that the variance of $f_\Theta(x)$ satisfies $\mathbb{V}\left[f_\Theta(x)\right] \leq V < +\infty$, for all $x \in \mathbb{R}^n$.

Next is stated a useful lemma similar to those in [15, 83], linking the probability of obtaining bad estimates to the variance assumption on function values.

Lemma 4. *Let Assumption 5 holds. Then for all $k \geq 0$, the followings hold for the random process*

$\{X^k, F_0^k, F_s^k, \Delta^k\}$ *generated by Algorithm 6*

$$\begin{aligned} & \mathbb{E}\left(\mathbb{1}_{\bar{J}_k}\left|F_0^k - f(X^k)\right| \mid \mathcal{F}_{k-1}^F\right) \leq \varepsilon_{f,p}(1-\beta)[\rho_p(\Delta^k)] \\ \textit{and} \quad & \mathbb{E}\left(\mathbb{1}_{\bar{J}_k}\left|F_s^k - f(X^k+S^k)\right| \mid \mathcal{F}_{k-1}^F\right) \leq \varepsilon_{f,p}(1-\beta)[\rho_p(\Delta^k)] \end{aligned} \tag{5.6}$$

Proof. The result is proved using ideas derived from [15, 83]. The proof follows straightforwardly from the conditional Cauchy-Schwarz inequality [28] as follows

$$\begin{aligned} \mathbb{E}\left(\mathbb{1}_{\bar{J}_k}\left|F_s^k - f(X^k+S^k)\right| \mid \mathcal{F}_{k-1}^F\right) &\leq \left[\mathbb{E}\left(\mathbb{1}_{\bar{J}_k} \mid \mathcal{F}_{k-1}^F\right)\right]^{1/2}\left[\mathbb{E}\left(\left|F_s^k - f(X^k+S^k)\right|^2 \mid \mathcal{F}_{k-1}^F\right)\right]^{1/2} \\ &\leq (1-\beta)^{1/2}\varepsilon_{f,p}(1-\beta)^{1/2}[\rho_p(\Delta^k)], \end{aligned}$$

where the last inequality follows from (5.5) and the fact that $\mathbb{E}\left(\mathbb{1}_{\bar{J}_k} \mid \mathcal{F}_{k-1}^F\right) = \mathbb{P}\left(\bar{J}_k \mid \mathcal{F}_{k-1}^F\right) \leq 1-\beta$ thanks to Assumption 5-*(i)*. The proof for $F_0^k - f(X^k)$ is the same. □

5.3 A renewal-reward martingale process

This section presents a general stochastic process and its associated stopping time T introduced in [29] for the convergence rate analysis of a stochastic trust-region method. It introduces some relevant definition, assumptions and theorem derived in the analysis of a renewal-reward process in [29], that will be useful for the convergence rate analysis presented in Section 5.4. Specifically, by considering the stopping time consisting of the time required by SDDS to reach a desired accuracy, Section 5.4 will aim to show how the properties of this general stochastic process are satisfied for Algorithm 6. Note that some results derived in analyzing this stochastic process in [29] are also used in [83] for the convergence rate analysis of a stochastic line search method.

Definition 9. *A random variable T is said to be a stopping time with respect to a given discrete time stochastic process $\{X_k\}_{k\in\mathbb{N}}$ if, for each $k \in \mathbb{N}$, the event $\{T = k\}$ belongs to the σ-field $\sigma(X_1, X_2, \ldots, X_k)$ generated by $X_1, X_2, \ldots, X_k$.*

Consider a stochastic process $\{(\Phi_k, \Delta^k)\}_{k\in\mathbb{N}}$ satisfying $\Phi_k \in [0, +\infty)$ and $\Delta^k \in [0, +\infty)$ for all $k \in \mathbb{N}$. Define on the same probability space as $\{(\Phi_k, \Delta^k)\}_{k\in\mathbb{N}}$, a sequence of biased random walk process $\{W_k\}_{k\in\mathbb{N}}$ such that $W_0 = 1$,

$$\mathbb{P}\left(W_{k+1} = 1 \mid \mathcal{F}_k\right) = q \quad \text{and} \quad \mathbb{P}\left(W_{k+1} = -1 \mid \mathcal{F}_k\right) = 1-q, \tag{5.7}$$

where $q \in (1/2, 1)$ and $\mathcal{F}_k$ denotes the σ-field generated by $\{(\Phi_0, \Delta^0, W_0), (\Phi_1, \Delta^1, W_1), \ldots, (\Phi_k, \Delta^k, W_k)\}$.

Define the following family $\{T_{\epsilon'}\}_{\epsilon'>0}$ of stopping times parameterized by $\epsilon' > 0$, with respect to $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$. The following assumptions are made in [29, 83] in order to derive a bound on $\mathbb{E}(T_{\epsilon'})$.

Assumption 6. *The following hold for the stochastic process* $\{(\Phi_k, \Delta^k, W_k)\}_{k\in\mathbb{N}}$.

(i) There exist constant $\lambda \in (0, +\infty)$ *and* $\delta_{\max} = \delta^0 e^{\lambda j_{\max}}$*, for some integer* $j_{\max} \in \mathbb{Z}$*, such that* $\Delta^k \leq \delta_{\max}$ *for all* $k \in \mathbb{N}$.

(ii) There exists a constant $\delta_{\epsilon'} = \delta^0 e^{\lambda j_{\epsilon'}}$*, for some* $j_{\epsilon'} \in \mathbb{Z}, j_{\epsilon'} \leq 0$*, such that the following holds for all* $k \in \mathbb{N}$*,*

$$\mathbb{1}_{\{T_{\epsilon'}>k\}}\Delta^{k+1} \geq \mathbb{1}_{\{T_{\epsilon'}>k\}} \min\left(\Delta^k e^{\lambda W_{k+1}}, \delta_{\epsilon'}\right), \tag{5.8}$$

where W_{k+1} *satisfies* (5.7) *with* $q > \frac{1}{2}$.

(iii) There exists a nondecreasing function $h : [0, +\infty) \to (0, +\infty)$ *and a constant* $\eta > 0$ *such that*

$$\mathbb{E}(\Phi_{k+1} - \Phi_k \mid \mathcal{F}_k)\,\mathbb{1}_{\{T_{\epsilon'}>k\}} \leq -\eta h(\Delta^k)\mathbb{1}_{\{T_{\epsilon'}>k\}}. \tag{5.9}$$

Note that as highlighted in [29, 83], Assumption 6 states that conditioned on the past, the nonnegative random sequence $\{\Phi_k\}_{k\in\mathbb{N}}$ decreases by at least $\eta h(\Delta^k)$ at each iteration provided that $T_{\epsilon'} > k$ and moreover, the sequence $\{\Delta^k\}_{k\in\mathbb{N}}$ has a tendency to increase whenever it is below some fixed threshold $\delta_{\epsilon'}$.

The following theorem from [29] providing a bound on $\mathbb{E}(T_{\epsilon'})$ is proved by observing that the upward drift in the random walk $\{W_k\}_{k\in\mathbb{N}}$ makes the event $\{\Delta^k \geq \delta_{\epsilon'}\}$ occur sufficiently frequently on average [29, 83]. Hence, $\mathbb{E}(\Phi_{k+1} - \Phi_k)$ can frequently be bounded by some negative fixed constant, thus leading to a bound on the expected stopping time $\mathbb{E}(T_{\epsilon'})$. A proof of this theorem is presented in the Appendix.

Theorem 6. *Let Assumption 6 hold. Then,*

$$\mathbb{E}(T_{\epsilon'}) \leq \frac{q}{2q-1} \times \frac{\Phi_0}{\eta h(\delta_{\epsilon'})} + 1$$

5.4 Convergence rate analysis

It follows from Section 5.3 that Theorem 6 holds for any stopping time $T_{\epsilon'}$ defined with respect to the filtration $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$, provided that Assumption 6 hold for the stochastic process $\{(\Phi_k, \Delta^k, W_k)\}_{k\in\mathbb{N}}$. Thus, the goal of the present section is to show how such a stochastic process satisfying Assumption 6 can be constructed in order to bound the expected number of iterations required by Algorithm 6 to achieve $\left\|\nabla f(X^k)\right\|_\infty \leq \epsilon$, for some arbitrary fixed $\epsilon \in (0, 1)$, where $\|\cdot\|_\infty$ denotes the Euclidean norm of $\mathbb{R}^n$ as in the remainder of the manuscript.

5.4.1 Analysis of the stochastic process generated by SDDS

In order to show that Assumption 6 holds, let impose the following standard assumption on the objective function f.

Assumption 7. *The function f is bounded from below, i.e., there exists $f_{\min} \in \mathbb{R}$ such that $-\infty < f_{\min} \leq f(x)$, for all $x \in \mathbb{R}^n$.*

The following result extending that in [15] from $p = 2$ to any $p > 1$ provides a bound on the expected decrease in the random function

$$\Phi_k := \frac{\nu}{c\varepsilon_{f,p}}(f(X^k) - f_{\min}) + (1-\nu)(\Delta^k)^p. \tag{5.10}$$

Theorem 7. *Let Assumption 4, 5 and 7 hold. Let $\gamma > 2, p > 1$ and $\tau \in (0,1)$. Let $\nu \in (0,1)$ and $\beta \in (1/2, 1)$ be chosen so that*

$$\frac{\nu}{1-\nu} \geq \frac{2(\tau^{-p}-1)}{\gamma - 2} \quad \textit{and} \quad \frac{\beta}{1-\beta} \geq \frac{\nu}{1-\nu} \times \frac{4}{(1-\tau^p)}, \tag{5.11}$$

Then the expected decrease in the random function Φ_k defined in (5.10) *satisfies*

$$\mathbb{E}\left(\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}^F\right) \leq -\frac{1}{2}\beta(1-\nu)(1-\tau^p)(\Delta^k)^p. \tag{5.12}$$

Proof. The proof is almost identical to that in [15], using ideas derived in [35, 68, 83] and making use of properties of the random function Φ_k defined in (5.10). It considers two separate cases: good estimates and bad estimates, each of which will be broken into whether an iteration is successful or unsuccessful. Define the event S by

$$S := \{\text{The iteration is successful}\},$$

and let $\bar{S}$ denote the complement of S.

Case 1 (Good estimates, $\mathbb{1}_{J_k} = 1$) The overall goal is to show that Φ_k decreases no matter what type of iteration occurs thus yielding the following bound

$$\mathbb{E}\left(\mathbb{1}_{J_k}(\Phi_{k+1} - \Phi_k) \mid \mathcal{F}_{k-1}^F\right) \leq -\beta(1-\nu)(1-\tau^p)(\Delta^k)^p. \tag{5.13}$$

(i) *Successful iteration* ($\mathbb{1}_S = 1$). A decrease occurs in f according to (5.2) since estimates are

good and the iteration is successful. Thus,

$$\mathbb{1}_{J_k}\mathbb{1}_S \frac{\nu}{c\varepsilon_{f,p}}(f(X^{k+1}) - f(X^k)) \leq -\mathbb{1}_{J_k}\mathbb{1}_S\nu(\gamma-2)(\Delta^k)^p. \tag{5.14}$$

The step size parameter is updated according to $\Delta^{k+1} = \min\{\tau^{-1}\Delta^k, \delta_{\max}\}$. Hence,

$$\mathbb{1}_{J_k}\mathbb{1}_S(1-\nu)\left[(\Delta^{k+1})^p - (\Delta^k)^p\right] \leq \mathbb{1}_{J_k}\mathbb{1}_S(1-\nu)(\tau^{-p}-1)(\Delta^k)^p. \tag{5.15}$$

Then, choosing ν according to (5.11) ensures that the right-hand side term of (5.14) dominates that of (5.15), i.e.,

$$-\nu(\gamma-2)(\Delta^k)^p + (1-\nu)(\tau^{-p}-1)(\Delta^k)^p \leq -\frac{1}{2}\nu(\gamma-2)(\Delta^k)^p. \tag{5.16}$$

Thus, combining (5.14), (5.15) and (5.16) yields

$$\mathbb{1}_{J_k}\mathbb{1}_S(\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{J_k}\mathbb{1}_S\frac{1}{2}\nu(\gamma-2)(\Delta^k)^p. \tag{5.17}$$

(ii) *Unsuccessful iteration* ($\mathbb{1}_{\bar{S}} = 1$). The step size parameter is decreased while there is a change of zero in function values since the iteration is unsuccessful. Thus,

$$\mathbb{1}_{J_k}\mathbb{1}_{\bar{S}}(\Phi_{k+1} - \Phi_k) = -\mathbb{1}_{J_k}\mathbb{1}_{\bar{S}}(1-\nu)(1-\tau^p)(\Delta^k)^p. \tag{5.18}$$

Then, the inequality $1-\tau^p < \tau^{-p} - 1$ ensures that unsuccessful iterations, specifically (5.18), provide the worst case decrease when compared to (5.17), whenever ν is chosen according to (5.11). Specifically, the following holds

$$-\frac{1}{2}\nu(\gamma-2)(\Delta^k)^p \leq -(1-\nu)(1-\tau^p)(\Delta^k)^p. \tag{5.19}$$

Thus, combining (5.17), (5.18), and (5.19), leads to the following bound on the change in Φ_k

$$\mathbb{1}_{J_k}(\Phi_{k+1} - \Phi_k) = \mathbb{1}_{J_k}(\mathbb{1}_S + \mathbb{1}_{\bar{S}})(\Phi_{k+1} - \Phi_k) \leq -\mathbb{1}_{J_k}(1-\nu)(1-\tau^p)(\Delta^k)^p. \tag{5.20}$$

Since Assumption 5 holds, then taking conditional expectations with respect to $\mathcal{F}_{k-1}^F$ in both sides of (5.20) leads to (5.13).

Case 2 (Bad estimates, $\mathbb{1}_{\bar{J}_k} = 1$). Since the estimates are bad, an iterate leading to an increase in f and Δ^k, and hence in Φ_k, can be accepted by Algorithm 6. Such an increase in Φ_k is controlled by bounding the variance in function estimates, using (5.5). Then, in order to guarantee that Φ_k is

sufficiently reduced in expectation, the probability of outcome is adjusted to be sufficiently small. The overall goal is to show that

$$\mathbb{E}\left(\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}-\Phi_k) \mid \mathcal{F}_{k-1}^F\right) \leq 2\nu(1-\beta)(\Delta^k)^p. \tag{5.21}$$

(i) *Successful iteration* ($\mathbb{1}_S = 1$). The change in f is bounded as follows

$$\begin{aligned}
\mathbb{1}_{\bar{J}_k}\mathbb{1}_S\frac{\nu}{c\varepsilon_{f,p}}&(f(X^{k+1})-f(X^k)) \\
&\leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S\frac{\nu}{c\varepsilon_{f,p}}\left[(F_s^k-F_0^k)+\left|f(X^{k+1})-F_s^k\right|+\left|F_0^k-f(X^k)\right|\right] \\
&\leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S\nu\left[-\gamma(\Delta^k)^p+\frac{1}{c\varepsilon_{f,p}}\left(\left|f(X^{k+1})-F_s^k\right|+\left|F_0^k-f(X^k)\right|\right)\right]
\end{aligned} \tag{5.22}$$

where the last inequality in (5.22) follows from the fact that $F_s^k - F_0^k \leq -\gamma c\varepsilon_{f,p}(\Delta^k)^p$ for successful iterations. Moreover, as in Case 1, $\Delta^{k+1} = \min\{\tau^{-1}\Delta^k, \delta_{\max}\}$ since the iteration is successful. Thus,

$$\mathbb{1}_{\bar{J}_k}\mathbb{1}_S(1-\nu)\left[(\Delta^{k+1})^p-(\Delta^k)^p\right] \leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S(1-\nu)(\tau^{-p}-1)(\Delta^k)^p. \tag{5.23}$$

Then, choosing ν according to (5.11) yields

$$-\nu\gamma(\Delta^k)^p+(1-\nu)(\tau^{-p}-1)(\Delta^k)^p \leq 0. \tag{5.24}$$

Thus, combining (5.22), (5.23) and (5.24) leads to

$$\mathbb{1}_{\bar{J}_k}\mathbb{1}_S(\Phi_{k+1}-\Phi_k) \leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_S\frac{\nu}{c\varepsilon_{f,p}}\left(\left|f(X^{k+1})-F_s^k\right|+\left|F_0^k-f(X^k)\right|\right) \tag{5.25}$$

(ii) *Unsuccessful iteration* ($\mathbb{1}_{\bar{S}} = 1$). Δ^k is decreased and the change in function values is zero. Thus, the bound in the change of Φ_k follows straightforwardly from (5.18) by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$. More precisely, the following holds,

$$\begin{aligned}
\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}(\Phi_{k+1}-\Phi_k) &= -\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}(1-\nu)(1-\tau^p)(\Delta^k)^p. \\
&\leq \mathbb{1}_{\bar{J}_k}\mathbb{1}_{\bar{S}}\frac{\nu}{c\varepsilon_{f,p}}\left(\left|f(X^{k+1})-F_s^k\right|+\left|F_0^k-f(X^k)\right|\right)
\end{aligned} \tag{5.26}$$

Then, combining (5.25) and (5.26), yields

$$\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}-\Phi_k) \leq \mathbb{1}_{\bar{J}_k}\frac{\nu}{c\varepsilon_{f,p}}\left(\left|f(X^{k+1})-F_s^k\right|+\left|F_0^k-f(X^k)\right|\right). \tag{5.27}$$

Taking conditional expectations with respect to $\mathcal{F}_{k-1}^F$ in both sides of (5.27) and applying Lemma 4 leads to (5.21).

Now, combining (5.13) and (5.21) leads to

$$\begin{aligned}\mathbb{E}\left(\Phi_{k+1}-\Phi_k \mid \mathcal{F}_{k-1}^F\right) &= \mathbb{E}\left((\mathbb{1}_{J_k}+\mathbb{1}_{\bar{J}_k})(\Phi_{k+1}-\Phi_k) \mid \mathcal{F}_{k-1}^F\right)\\ &\leq [-\beta(1-\nu)(1-\tau^p)+2\nu(1-\beta)]\,(\Delta^k)^p. \end{aligned} \tag{5.28}$$

Then, choosing β according to (5.11) ensures that

$$-\beta(1-\nu)(1-\tau^p)+2\nu(1-\beta) \leq -\frac{1}{2}\beta(1-\nu)(1-\tau^p). \tag{5.29}$$

Hence, (5.12) follows from (5.28) and (5.29), which achieves the proof. □

Remark 3. *$\nu \in (0,1)$ and $\beta \in (1/2,1)$ can always be chosen according to* (5.11)*. Indeed, first observe that τ, p and γ are fixed, which implies that $\dfrac{2(\tau^{-p}-1)}{\gamma-2}$ is a constant. Since $\lim\limits_{\nu\nearrow 1}\dfrac{\nu}{1-\nu}=+\infty$, then ν can always be chosen sufficiently close to one, so that*

$$\frac{\nu}{1-\nu} \geq \frac{2(\tau^{-p}-1)}{\gamma-2}. \tag{5.30}$$

Now, assume that ν is fixed and chosen according to (5.30)*. Then $\dfrac{\nu}{1-\nu}\times\dfrac{4}{(1-\tau^p)}$ is a constant. By observing that $\lim\limits_{\beta\nearrow 1}\dfrac{\beta}{1-\beta}=+\infty$, then β can always be chosen sufficiently close to one so that*

$$\frac{\beta}{1-\beta} \geq \frac{\nu}{1-\nu}\times\frac{4}{(1-\tau^p)}.$$

Finally, summing (5.12) over $k \in \mathbb{N}$ and taking expectations on both sides lead to the following result extending that in [15] from $p=2$ to any $p>1$, and which shows in particular that the sequence $\{\Delta^k\}_{k\in\mathbb{N}}$ of step size parameters converges to zero with probability one.

Theorem 8. *Let all assumptions that were made in Theorem 7 hold. Then for all $p>1$, the sequence $\{\Delta^k\}_{k\in\mathbb{N}}$ of step size parameters generated by Algorithm 6 satisfies*

$$\sum_{k=0}^{+\infty}(\Delta^k)^p < +\infty \quad \textit{almost surely.}$$

Consider the stochastic process $\{(\Phi_k,\Delta^k,W_k)\}_{k\in\mathbb{N}}$, where Φ_k is the same random function in Theo-

rem 7, Δ^k is the random step size parameter and $W_k = 2(\mathbb{1}_{J_k} - \frac{1}{2})$. Define $\hat{p} = \min(p-1, 1)$ for some fixed $p > 1$. For some arbitrary fixed $\epsilon' \in (0, 1)$, consider the following random time $T_{\epsilon'}$ defined by

$$T_{\epsilon'} = \inf \left\{ k \in \mathbb{N} : \left\| \nabla f(X^k) \right\|_\infty^{1/\hat{p}} \leq \epsilon' \right\} \tag{5.31}$$

Then, $T_{\epsilon'}$ is a stopping time for the stochastic process generated by Algorithm 6 and is consequently a stopping time for $\{(\Phi_k, \Delta^k, W_k)\}_{k\in\mathbb{N}}$ [29, 83]. Moreover, $T_{\epsilon^{1/\hat{p}}}$ is the number of iterations required by Algorithm 6 to drive the norm of the gradient of f below $\epsilon \in (0, 1)$. This latter remark will help to derive the main result of the present work in Theorem 9.

In order to apply Theorem 6 to $T_{\epsilon'}$, the remainder of this section is devoted to showing that Assumption 6 holds for the previous stochastic process. Let show that Assumption 6-*(iii)* holds. First, multiplying both sides of (5.20) by $\mathbb{1}_{\{T_{\epsilon'}>k\}}$ and taking conditional expectation with respect to $\mathcal{F}_{k-1}^F$ lead to

$$\begin{aligned}
\mathbb{E}\left(\mathbb{1}_{\{T_{\epsilon'}>k\}}\mathbb{1}_{J_k}(\Phi_{k+1} - \Phi_k)|\mathcal{F}_{k-1}^F\right) &= \mathbb{1}_{\{T_{\epsilon'}>k\}}\mathbb{E}\left(\mathbb{1}_{J_k}(\Phi_{k+1} - \Phi_k)|\mathcal{F}_{k-1}^F\right) \\
&\leq -\mathbb{E}\left(\mathbb{1}_{\{T_{\epsilon'}>k\}}\mathbb{1}_{J_k}(1-\nu)(1-\tau^p)(\Delta^k)^p|\mathcal{F}_{k-1}^F\right) \\
&= -\mathbb{1}_{\{T_{\epsilon'}>k\}}\mathbb{E}\left(\mathbb{1}_{J_k}(1-\nu)(1-\tau^p)(\Delta^k)^p|\mathcal{F}_{k-1}^F\right) \\
&\leq -\mathbb{1}_{\{T_{\epsilon'}>k\}}\beta(1-\nu)(1-\tau^p)(\Delta^k)^p,
\end{aligned} \tag{5.32}$$

where the equalities in (5.32) follows from the fact that $\mathbb{1}_{\{T_{\epsilon'}>k\}}$ is $\mathcal{F}_{k-1}^F$-measurable since $T_{\epsilon'}$ is a stopping time with respect to the filtration $\left\{\mathcal{F}_{k-1}^F\right\}_{k\geq 0}$. Similarly, it easily follows from (5.27) that

$$\mathbb{1}_{\{T_{\epsilon'}>k\}}\mathbb{E}\left(\mathbb{1}_{\bar{J}_k}(\Phi_{k+1} - \Phi_k) \mid \mathcal{F}_{k-1}^F\right) \leq \mathbb{1}_{\{T_{\epsilon'}>k\}} 2\nu(1-\beta)(\Delta^k)^p. \tag{5.33}$$

Thus, combining (5.32) and (5.33) (just like (5.13) and (5.21) in the proof of Theorem 7) implies that

$$\mathbb{1}_{\{T_{\epsilon'}>k\}}\mathbb{E}\left(\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}^F\right) \leq -\mathbb{1}_{\{T_{\epsilon'}>k\}}\frac{1}{2}\beta(1-\nu)(1-\tau^p)(\Delta^k)^p,$$

which means that Assumption 6-*(iii)* holds with $\eta = \frac{1}{2}\beta(1-\nu)(1-\tau^p)$ and $h(x) = x^p$. Moreover, by choosing λ such that $e^\lambda = \tau^{-1}$ and noticing that $\Delta^k \leq \delta_{\max} = \delta^0 e^{\lambda j_{\max}}$ in Algorithm 6 for all $k \in \mathbb{N}$, then Assumption 6-*(i)* holds.

Now, before showing that Assumption 6-*(ii)* holds, let emphasize that as in the deterministic framework, polling directions in Algorithm 6 are chosen in such a way that their significant deterioration can be avoided asymptotically, i.e., in such a way to ensure that they never become close to loosing the positive spanning property [95]. For this purpose, let recall the following definition of the *cosine*

measure [42, 65] of a positive spanning set $\mathbb{D}^k$ with non-zero vectors

$$\kappa(\mathbb{D}^k) := \min_{v\in\mathbb{R}^n} \max_{d\in\mathbb{D}^k} \frac{v^\top d}{\|v\|_\infty \|d\|_\infty}$$

In order to avoid the abovementioned deterioration of polling directions, the positive spanning sets are required to satisfy the following assumption [65,95] where the size of the directions does not tend to infinite or approaches zero, and the cosine measure always stays positive.

Assumption 8. *The followings hold for all positive spanning sets $\mathbb{D}^k$ used for polling in Algorithm 6. There exists a constant $\kappa_{\min} > 0$ such that $\kappa(\mathbb{D}^k) > \kappa_{\min}$ for all k. There exist constants $d_{\min} > 0$ and $d_{\max} > 0$ such that $d_{\min} \leq \|d\|_\infty \leq d_{\max}$ for all $d \in \mathbb{D}^k$.*

The following result from [65] will be useful for the remaining of the analysis, and specifically the proof of the key result in Lemma 5. It shows by means of the cosine measure $\kappa(\mathbb{D}^k)$, how far can be in the worst case, the steepest descent direction from the vector in $\mathbb{D}^k$ which makes the smallest angle with $v = -\nabla f(x^k)$. This means in term of descent that, there exists $d_*^k \in \mathbb{D}^k$ such that

$$\kappa(\mathbb{D}^k)\left\|\nabla f(x^k)\right\|_\infty \left\|d_*^k\right\|_\infty \leq -\nabla f(x^k)^\top d_*^k. \tag{5.34}$$

For the remaining of the analysis, the following standard assumption is also imposed on the gradient of f.

Assumption 9. *The gradient ∇f of the objective function f is L-Lipschitz continuous everywhere.*

Define the constant $\delta_{\epsilon'}$ as follows

$$\delta_{\epsilon'} = \frac{\epsilon'}{\zeta} \quad \text{with} \quad \zeta > \left[\kappa_{\min}^{-1}\left(Ld_{\max} + (\gamma+2)c\varepsilon_{f,p}d_{\min}^{-1}\right)\right]^{1/\hat{p}}, \tag{5.35}$$

where without loss of generality, $Ld_{\max} > \kappa_{\min}$ so that $\delta_{\epsilon'} < 1$ for the needs of the analysis and specifically, the proof of Lemma 5. Then following [29], it can be assumed without any loss of generality that $\delta_{\epsilon'} = \tau^{-i}\delta^0$, for some integer $i \leq 0$. Hence, for any k, $\Delta^k = \tau^{i_k}\delta_{\epsilon'}$, for some integer i_k. Thus, what remains to be proved in Assumption 6 in order to apply Theorem 6 is the dynamics (5.8). Note however that the proof of the latter dynamics which will be achieved in Lemma 6, needs the following intermediate key result. Indeed, in the stochastic trust-region framework of [29], the proof of a similar dynamics strongly relies on the fact that any iteration k, where $\|\nabla f(x_k)\| > \epsilon$ and for which "*good*" model and estimates occur, is successful provided that the *trust-region radius* δ_k is bellow a threshold Δ_ϵ. Nevertheless, unlike the latter framework where informations can easily be

derived on the true gradient $\nabla f(x^k)$ using those provided by the *gradient estimate* g_k, the algorithmic framework of the present work does not use any gradient information. Thus, the main challenge in proving that Assumption 6-*(ii)* holds, lies in linking the event $\left\{\left\|\nabla f(X^k)\right\|_\infty^{1/\hat{p}} > \epsilon'\right\}$ to a successful iteration of Algorithm 6, which is done next.

Lemma 5. *Assume that Assumptions 8 and 9 hold, and that $\delta^k \leq \delta_{\epsilon'}$. Let f_0^k and f_s^k be $\varepsilon_{f,p}$-accurate estimates of $f(x^k)$ and $f(x^k + s^k)$ respectively. If $\left\|\nabla f(x^k)\right\|_\infty^{1/\hat{p}} > \epsilon'$, then*

$$f_s^k - f_0^k \leq -\gamma c \varepsilon_{f,p} (\delta^k)^p.$$

In particular, this means that the iteration k of Algorithm 6 is successful.

Proof. The proof uses elements derived in [65]. Suppose that $\delta^k \leq \delta_{\epsilon'}$ and assume in contradiction that $f_s^k - f_0^k > -\gamma c \varepsilon_{f,p} (\delta^k)^p$. Since the estimates f_0^k and f_s^k are $\varepsilon_{f,p}$-accurate, then it follows from the following equality

$$f(x^k + s^k) - f(x^k) = f(x^k + s^k) - f_s^k + (f_s^k - f_0^k) + f_0^k - f(x^k)$$

that

$$f(x^k + s^k) - f(x^k) + (\gamma + 2) c \varepsilon_{f,p} (\delta^k)^p \geq 0. \tag{5.36}$$

Recall that $s^k = \delta^k d$ where $d \in \mathbb{D}^k$ denotes any direction used by Algorithm 6 at iteration k. It follows from the mean value theorem, combined with (5.36), that there exists a constant $\mu_k \in [0, 1]$ such that

$$0 \leq \delta^k \nabla f(x^k + \mu_k \delta^k d_*^k)^\top d_*^k + (\gamma + 2) c \varepsilon_{f,p} (\delta^k)^p, \tag{5.37}$$

where d_*^k is the direction satisfying (5.34). Dividing both sides of (5.37) by δ^k and subtracting $\nabla f(x^k)^\top d_*^k$, yields

$$-\nabla f(x^k)^\top d_*^k \leq \left[\nabla f(x^k + \mu_k \delta^k d_*^k) - \nabla f(x^k)\right]^\top d_*^k + (\gamma + 2) c \varepsilon_{f,p} (\delta^k)^{p-1}. \tag{5.38}$$

Putting (5.34) and (5.38) together, yields

$$\kappa(\mathbb{D}^k) \left\|\nabla f(x^k)\right\|_\infty \left\|d_*^k\right\|_\infty \leq \left[\nabla f(x^k + \mu_k \delta^k d_*^k) - \nabla f(x^k)\right]^\top d_*^k + (\gamma + 2) c \varepsilon_{f,p} (\delta^k)^{p-1}. \tag{5.39}$$

Then, dividing both sides of (5.39) by $\kappa(\mathbb{D}^k) \left\|d_*^k\right\|_\infty$ and using Assumption 9 and 8, lead to

$$\begin{aligned} \left\|\nabla f(x^k)\right\|_\infty &\leq \kappa_{\min}^{-1} \left[L d_{\max} \delta^k + (\gamma + 2) c \varepsilon_{f,p} d_{\min}^{-1} (\delta^k)^{p-1}\right] \\ &\leq \kappa_{\min}^{-1} \left(L d_{\max} + (\gamma + 2) c \varepsilon_{f,p} d_{\min}^{-1}\right) (\delta^k)^{\min(p-1,1)}, \end{aligned} \tag{5.40}$$

where the inequality (5.40) follows from the fact that $\delta^k \leq \delta_{\epsilon'} < 1$. Now, recall that $\hat{p} = \min(p-1, 1)$ and let $L_1 := \kappa_{\min}^{-1}\left(Ld_{\max} + (\gamma+2)c\varepsilon_{f,p}d_{\min}^{-1}\right)$. Then, it follows from (5.40) that

$$\left\|\nabla f(x^k)\right\|_\infty^{1/\hat{p}} \leq L_1^{1/\hat{p}}\delta^k \leq L_1^{1/\hat{p}}\delta_{\epsilon'} = L_1^{1/\hat{p}}\frac{\epsilon'}{\zeta} \leq \epsilon', \tag{5.41}$$

where the last inequality in (5.41) follows from (5.35), which achieves the proof. □

Finally, the following result shows that the dynamics (5.8) of Assumption 6-*(ii)* holds.

Lemma 6. *Let Assumption 9 and all assumptions that were made in Theorem 7 hold. Then Assumption 6-(ii) is satisfied for the random variable* $W_k = 2(\mathbb{1}_{J_k} - \frac{1}{2})$, $\lambda = -\ln(\tau)$ *and* $q = \beta$.

Proof. The result is proved by adapting the proof of a similar Lemma from [29]. First, notice that (5.8) trivially holds when $\mathbb{1}_{\{T_{\epsilon'}>k\}} = 0$. Thus, the remaining of the proof is devoted to showing that conditioned on the event $\{T_{\epsilon'} > k\}$, i.e., when $\mathbb{1}_{\{T_{\epsilon'}>k\}} = 1$, then the following holds

$$\Delta^{k+1} \geq \min\left\{\delta_{\epsilon'}, \min\left\{\tau^{-1}\Delta^k, \delta_{\max}\right\}\mathbb{1}_{J_k} + \tau\Delta^k\mathbb{1}_{\bar{J}_k}\right\}. \tag{5.42}$$

Notice that every realization such that $\delta^k > \delta_{\epsilon'}$ also satisfies $\delta^k \geq \tau^{-1}\delta_{\epsilon'}$ whence $\delta^{k+1} \geq \tau\delta^k \geq \delta_{\epsilon'}$. Now, assume that $\delta^k \leq \delta_{\epsilon'}$. Since $T_{\epsilon'} > k$, then it is the case that $\left\|\nabla f(x^k)\right\|_\infty^{1/\hat{p}} > \epsilon'$. If $\mathbb{1}_{J_k} = 1$, then the estimates are good and are specifically $\varepsilon_{f,p}$-accurate. Hence, it follows from Lemma 5 that the kth iteration is successful. Thus, $x^{k+1} = x^k + s^k$ and $\delta^{k+1} = \min\left\{\tau^{-1}\delta^k, \delta_{\max}\right\}$. But if $\mathbb{1}_{J_k} = 0$, i.e., $\mathbb{1}_{\bar{J}_k} = 1$, then the inequality $\delta^{k+1} \geq \tau\delta^k$ always holds by the dynamics of Algorithm 6. The proof is complete by noticing finally that $\mathbb{P}\left(J_k|\mathcal{F}_{k-1}^F\right) \geq q = \beta$. □

5.4.2 Complexity result and first-order optimality conditions

The following result provides a bound on the expected number of iterations taken by Algorithm 6 before $\left\{\left\|\nabla f(X^k)\right\|_\infty \leq \epsilon\right\}$ occurs. It is the main result of the present work.

Theorem 9. *Let Assumption 9 and all assumptions that were made in Theorem 7 hold with* $\beta \in (1/2, 1)$ *and* $\nu \in (0, 1)$ *satisfying* (5.11)*. Consider Algorithm 6 and the corresponding stochastic process. For some arbitrary fixed* $\epsilon \in (0, 1)$*, consider the random time* $T_\epsilon^\star$ *defined by*

$$T_\epsilon^\star = \inf\left\{k \in \mathbb{N} : \left\|\nabla f(X^k)\right\|_\infty \leq \epsilon\right\}. \tag{5.43}$$

Then,

$$\mathbb{E}\left(T_\epsilon^\star\right) \leq \frac{2\Phi_0 L_2}{(2\beta-1)(1-\nu)(1-\tau^p)}\epsilon^{\frac{-p}{\min(p-1,1)}} + 1, \tag{5.44}$$

where $L_2 := \left[1 + \kappa_{\min}^{-1}\left(Ld_{\max} + (\gamma+2)c\varepsilon_{f,p}d_{\min}^{-1}\right)\right]^{\frac{p}{\min(p-1,1)}}$. *i.e., the expected number of iterations taken by Algorithm 6 to reduce the gradient below* $\epsilon \in (0,1)$ *is bounded in* $\mathcal{O}\left(\epsilon^{\frac{-p}{\min(p-1,1)}}/(2\beta-1)\right)$.

Proof. As shown previously, since Assumption 6 holds for the stochastic process $\{(\Phi_k, \Delta^k, W_k)\}_{k\in\mathbb{N}}$ generated by Algorithm 6, with $q = \beta$, $h(x) = x^p$, $\eta = \frac{1}{2}\beta(1-\nu)(1-\tau^p)$ and $\delta_{\epsilon'} = \epsilon'/\zeta$, then Theorem 6 applies for the stopping time $T_{\epsilon'}$ defined in (5.31). Thus, the following inequality holds for all $\epsilon' \in (0,1)$

$$\mathbb{E}\left(T_{\epsilon'}\right) \leq \frac{\beta}{2\beta-1} \times \frac{\Phi_0 \zeta^p}{\eta \epsilon'^p} + 1, \tag{5.45}$$

where $\zeta^p > \left[\kappa_{\min}^{-1}\left(Ld_{\max} + (\gamma+2)c\varepsilon_{f,p}d_{\min}^{-1}\right)\right]^{p/\hat{p}}$ due to (5.35), with $\hat{p} = \min(p-1, 1)$. Now, let $\epsilon \in (0,1)$ be arbitrary fixed. Then, $\epsilon^{1/\hat{p}} \in (0,1)$, which means that (5.45) holds in particular for $\epsilon' = \epsilon^{1/\hat{p}}$. By noticing moreover that $T_{\epsilon^{1/\hat{p}}} = T_\epsilon^\star$, then it follows from (5.45) that

$$\mathbb{E}\left(T_\epsilon^\star\right) \leq \frac{\beta}{2\beta-1} \times \frac{2\Phi_0\zeta^p}{\beta(1-\nu)(1-\tau^p)}\epsilon^{-p/\hat{p}} + 1. \tag{5.46}$$

Finally, (5.44) results from (5.46) for $\zeta^p = \left[1 + \kappa_{\min}^{-1}\left(Ld_{\max} + (\gamma+2)c\varepsilon_{f,p}d_{\min}^{-1}\right)\right]^{p/\hat{p}}$, which achieves the proof. □

Remark 4. *The analysis carried out in the present stochastic directional direct-search framework and especially the complexity result of Theorem 9, favors discussions about the consequences and tradeoffs of allowing for more general* p *in the forcing function. First, observe that* $\frac{p}{\min(p-1,1)} \geq 2$ *for all* $p > 1$ *where the equality holds only for* $p = 2$*, in which case the corresponding expected complexity bound of* $\mathcal{O}\left(\epsilon^{-2}\right)$ *is that of standard nonconvex optimization. When* $p > 2$ *in which case the expected complexity bound is a* $\mathcal{O}\left(\epsilon^{-p}\right)$*, the corresponding convergence rate is slower than that of standard nonconvex optimization since* $\epsilon^{-p} > \epsilon^{-2}$*, but in this case, a less decrease from the forcing function is demanded. At the same time, however, e.g. for* $p = 3$*, an* $\varepsilon_{f,p}$*-accurate solution requires* $\mathcal{O}\left((\delta^k)^3\right)$ *accuracy (and* $\mathcal{O}\left((\delta^k)^6\right)$ *variance bounds per Assumption 5) which is a steep price to pay. On the flip side for* $1 < p < 2$*, especially as* $p \searrow 1$*, a less accuracy in estimates is demanded, which is nice, but in this case the corresponding convergence rate is very slow unfortunately since* $\lim_{p\searrow 1} \frac{p}{\min(p-1,1)} = +\infty$*, and a more decrease per iteration (approaching* $\mathcal{O}(\delta^k)$*) is demanded.*

The following $\liminf$-type first-order necessary optimality condition is a simple consequence of the complexity result of Theorem 9. It shows the existence of a subsequence of random iterates generated by Algorithm 6 which drives the norm of the gradient of f to zero with probability one. Note that a similar corollary has been derived in [83].

Theorem 10. *Let Assumption 9 and all assumptions that were made in Theorem 7 hold. Then the sequence* $\{X^k\}_{k\in\mathbb{N}}$ *of random iterates generated by Algorithm 6 satisfies*

$$\liminf_{k\to+\infty} \left\| \nabla f\left(X^k\right) \right\|_\infty = 0 \quad \textit{almost surely}. \tag{5.47}$$

5.5 Concluding remarks

This manuscript presents the first convergence rate analysis of a broad class of stochastic directional direct-search (SDDS) algorithms designed for the unconstrained optimization of noisy blackboxes, and based on imposing a sufficient decrease condition when accepting new iterates. Using an existing supermartingale-based framework for the analysis, the methodology for deriving the worst case complexity of SDDS algorithms heavily relies on bounding an expected stopping time associated to the stochastic process generated by the algorithms. The analysis demonstrates that SDDS algorithms have the same worst case complexity as any other first-order optimization method in a nonconvex setting. In particular, this complexity bound matches in some sense its deterministic counterparts despite the fact that function estimates are sometimes allowed to be arbitrarily inaccurate. The main novelty of the present research compared to many others on the worst case complexity analysis of stochastic DFO methods, lies in the fact that SDDS algorithms do not need any gradient information to find descent directions.

The analysis in the present manuscript strongly relies on the assumption that function estimates must satisfy a *variance condition*. Obtaining worst case complexity results without the latter condition when such estimates are possibly biased is therefore a topic for future research.

Acknowledgments

The author is grateful to Sébastien Le Digabel and Charles Audet from Polytechnique Montréal and Michael Kokkolaras from McGill university for valuable discussions and constructive suggestions that improve the quality of the presentation. 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

Before presenting the proof of Theorem 6 from [29], which requires the following intermediate results also from [29], recall the stochastic process $\{W_k\}_{k\in\mathbb{N}}$ introduced in (5.7), the constant $j_{\epsilon'}, \delta_{\epsilon'}, \eta$ from Assumption 6, the σ-field $\mathcal{F}_k$ generated by $\{(\Phi_0, \Delta^0, W_0), (\Phi_1, \Delta^1, W_1), \ldots, (\Phi_k, \Delta^k, W_k)\}$ and the family $\{T_{\epsilon'}\}_{\epsilon'>0}$ of stopping times with respect to $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$.

Lemma 7. *Let $\{Z_k\}_{k\in\mathbb{N}}$ be a stochastic process defined by $Z_0 = j_{\epsilon'}$ and $Z_{k+1} = \min(Z_k + W_{k+1}, j_{\epsilon'})$. Define the process $\{A_n\}_{n\in\mathbb{N}}$ by letting $A_0 = 0$ and setting $A_n = \inf\{m > A_{n-1} : Z_m = j_{\epsilon'}\}$. For all $n \geq 1$, let $\tau_n = A_n - A_{n-1}$. Then, for all n,*

$$\mathbb{E}(\tau_n) = \frac{p}{2p-1}.$$

Proof. Consider the simple random walk defined by $\bar{Z}_0 = -1$, $\bar{Z}_{k+1} = \bar{Z}_k + W_{k+1}$ for all $k \geq 1$, and define $\bar{\tau} = \inf\left\{n \geq 0 : \bar{Z}_n = 0\right\}$. It is well known from Wald's identity that

$$\mathbb{E}(\bar{\tau}) = \frac{1}{2p-1}.$$

Then, conditioning on W_1 and noticing that the distribution of $\bar{\tau}$ is the same as that of τ_1 conditioned on $Z_1 = j_{\epsilon'} - 1$, lead to

$$\mathbb{E}(\tau_1) = \mathbb{P}(W_1 = 1) + (1 + \mathbb{E}(\bar{\tau}))\mathbb{P}(W_1 = -1) = p + \left(1 + \frac{1}{2p-1}\right)(1-p).$$

□

Lemma 8. *Assume that Assumption 6-(iii) holds. Let $\{N(k)\}_{k\in\mathbb{N}}$ be a stochastic process defined by $N(k) = \max\{n : A_n \leq k\}$. Then*

$$\mathbb{E}[N(T_{\epsilon'} - 1) + 1] \leq \frac{\Phi_0}{\eta h(\delta_{\epsilon'})}.$$

Proof. Consider the stochastic process $\{R_k\}_{k\in\mathbb{N}}$ defined by $R_0 = \Phi_0$ and

$$R_k = \Phi_{k\wedge T_{\epsilon'}} + \eta \sum_{j=0}^{(k\wedge T_{\epsilon'})-1} h(\Delta^j), \quad \text{where } k \wedge T_{\epsilon'} := \min(k, T_{\epsilon'}).$$

In order to see that $\{R_k\}_{k\in\mathbb{N}}$ is a nonnegative supermartingale with respect to $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$, observe first that

$$\mathbb{E}(R_{k+1}|\mathcal{F}_k) = \mathbb{E}\left(R_{k+1}\mathbb{1}_{\{T_{\epsilon'}>k\}}|\mathcal{F}_k\right) + \mathbb{E}\left(R_{k+1}\mathbb{1}_{\{T_{\epsilon'}\leq k\}}|\mathcal{F}_k\right).$$

Then,

$$\begin{aligned}\mathbb{E}\left(R_{k+1}\mathbb{1}_{\{T_{\epsilon'}\leq k\}}|\mathcal{F}_k\right) &= \mathbb{E}\left[\left(\Phi_{T_{\epsilon'}}+\eta\sum_{j=0}^{T_{\epsilon'}-1}h(\Delta^j)\right)\mathbb{1}_{\{T_{\epsilon'}\leq k\}}|\mathcal{F}_k\right]\\ &= \Phi_{T_{\epsilon'}}\mathbb{1}_{\{T_{\epsilon'}\leq k\}}+\eta\sum_{j=0}^{T_{\epsilon'}-1}h(\Delta^j)\mathbb{1}_{\{T_{\epsilon'}\leq k\}},\end{aligned} \tag{5.48}$$

where the last equality follows from the fact $T_{\epsilon'}$ is $\mathcal{F}_k$-measurable since it is a stopping time. Moreover, noting that $\{T_{\epsilon'}\geq k+1\}=\{T_{\epsilon'}>k\}=\{T_{\epsilon'}\leq k\}^c\in\mathcal{F}_k$ and using (5.9), lead to

$$\begin{aligned}\mathbb{E}\left(R_{k+1}\mathbb{1}_{\{T_{\epsilon'}>k\}}|\mathcal{F}_k\right) &= \mathbb{E}\left(R_{k+1}|\mathcal{F}_k\right)\mathbb{1}_{\{T_{\epsilon'}>k\}}\\ &= \mathbb{E}\left(\Phi_{k+1}|\mathcal{F}_k\right)\mathbb{1}_{\{T_{\epsilon'}>k\}}+\mathbb{E}\left[\eta\sum_{j=0}^{k}h(\Delta^j)|\mathcal{F}_k\right]\mathbb{1}_{\{T_{\epsilon'}>k\}}\\ &\leq \left(\Phi_k-\eta h(\Delta^k)+\eta\sum_{j=0}^{k}h(\Delta^j)\right)\mathbb{1}_{\{T_{\epsilon'}>k\}}\\ &= \left(\Phi_k+\eta\sum_{j=0}^{k-1}h(\Delta^j)\right)\mathbb{1}_{\{T_{\epsilon'}>k\}}.\end{aligned} \tag{5.49}$$

Combining finally (5.48) and (5.49) implies that $\mathbb{E}\left(R_{k+1}|\mathcal{F}_k\right)\leq R_k$ as claimed. Then, because $\Phi_k\geq 0$ for all $k\geq 0$,

$$\eta\mathbb{E}\left[\sum_{j=0}^{(k\wedge T_{\epsilon'})-1}h(\Delta^j)\right]=\mathbb{E}\left(R_k\right)\leq\mathbb{E}\left(R_0\right)=\Phi_0.$$

Now, since $h(\cdot)\geq 0$, observe that

$$0\leq\sum_{j=0}^{(k\wedge T_{\epsilon'})-1}h(\Delta^j)\nearrow\sum_{j=0}^{T_{\epsilon'}-1}h(\Delta^j)\quad\text{as } k\to+\infty,$$

which holds even on the event $\{T_{\epsilon'}=+\infty\}$. Therefore, it follows from the monotone convergence theorem [50] that

$$\eta\mathbb{E}\left[\sum_{j=0}^{T_{\epsilon'}-1}h(\Delta^j)\right]=\lim_{k\to+\infty}\eta\mathbb{E}\left[\sum_{j=0}^{(k\wedge T_{\epsilon'})-1}h(\Delta^j)\right]\leq\mathbb{E}\left(R_0\right)=\Phi_0. \tag{5.50}$$

By definition of $N(\cdot)$, because $\left\{A_n:\Delta^{A_n}\geq\delta_{\epsilon'}\right\}\subseteq\{0,1,\ldots,T_{\epsilon'}\}$ and since $h(\cdot)$ is nondecreasing,

$$\eta\sum_{j=0}^{T_{\epsilon'}-1}h(\Delta^j)\geq\eta\sum_{j=0}^{T_{\epsilon'}-1}h(\Delta^j)\mathbb{1}_{\left\{j\in\{A_i\}_{i=1}^{\infty}\right\}}\geq\eta\left[N(T_{\epsilon'}-1)+1\right]h(\delta_{\epsilon'}),$$

where one was added to $N(T_{\epsilon'}-1)$ in the last inequality because $A_0=0$. Inserting this in (5.50),

$$\mathbb{E}\left[N(T_{\epsilon'}-1)+1\right] \leq \frac{\Phi_0}{\eta h(\delta_{\epsilon'})},$$

which achieves the proof. □

The following well-known theorem concerning expected stopping time, known as Wald's identity, typically proved in the literature under the assumption that the stopping time is finite almost surely, was proved in [29] by dropping the latter assumption since it is equivalent to assuming the optimization algorithm generating the stochastic process to converge.

Theorem 11. (Wald's identity). *Let $\{Y_i\}_{i=1}^n$ be a sequence of independent random variables such that $\mathbb{P}\left(Y_i \in [0,+\infty]\right)=1$. Define $\mathbb{E}\left(Y_i\right)=\mu_i \in [0,+\infty]$ and let $N \in [0,+\infty]$ be a stopping time with respect to the filtration generated by the Y_ns. Define $S_n = Y_1+Y_2+\cdots+Y_n, S_0=0, s_n = \mu_1+\mu_2+\cdots+\mu_n$ and $s_0=0$. Then*

$$\mathbb{E}\left(S_N\right)=\mathbb{E}\left(s_N\right).$$

Now, Theorem 6 is proved by applying Wald's identity to $S_n=A_n=\sum_{i=0}^{n}\tau_i$.

Proof of Theorem 6

Proof. Define $\mathscr{G}_n=\mathcal{F}_{A_n}:=\{A\in\sigma\left(\cup_{m=0}^{\infty}\mathcal{F}_m\right): A\cap\{A_n\leq k\}\in\mathcal{F}_k \text{ for all } k\in\mathbb{N}\}$. Note that $\mathscr{G}_n$ is well defined since A_n is a stopping time with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$.
The random variable $N(T_{\epsilon'}-1)+1$ is a stopping time with respect to $\{\mathscr{G}_n\}_{n\in\mathbb{N}}$. Indeed, since $N(k)\leq k$,

$$\begin{aligned}\{N(T_{\epsilon'}-1)+1\leq n\} &= \cup_{k=0}^{n-1}\{N(k)\leq n-1, T_{\epsilon'}-1=k\}\\ &= \cup_{k=0}^{n-1}\{N(k)+1\leq n, T_{\epsilon'}=k+1\}\subseteq\mathcal{F}_{A_n},\end{aligned}$$

where the inclusion follows from the fact that $N(k)+1$ is a stopping time with respect to $\{\mathcal{F}_{A_n}\}_{n\in\mathbb{N}}$ ($A_n\geq n$ implies $\mathcal{F}_n\subseteq\mathcal{F}_{A_n}$), and $T_{\epsilon'}$ is a stopping time with respect to $\{\mathscr{G}_n\}_{n\in\mathbb{N}}$ by construction.
Now, because of the independence assumption implied by (5.7),

$$\mathbb{E}\left(\tau_{n+1}|\mathscr{G}_n\right)=\mathbb{E}\left(\tau_{n+1}\right)=\frac{p}{2p-1}.$$

Recalling that $A_{N(T_{\epsilon'}-1)+1} = \sum_{k=1}^{N(T_{\epsilon'}-1)+1} \tau_k$, the Wald's identity from Theorem 11 is invoked to conclude that

$$\mathbb{E}\left[A_{N(T_{\epsilon'}-1)+1}\right] = \frac{p}{2p-1}\mathbb{E}\left[N(T_{\epsilon'}-1)+1\right].$$

Finally, since $A_{N(T_{\epsilon'}-1)+1} \geq T_{\epsilon'}-1$, it follows from Lemmas 7 and 8 that

$$\mathbb{E}\left(T_{\epsilon'}-1\right) \leq \mathbb{E}\left(\tau_1\right)\mathbb{E}\left[N(T_{\epsilon'}-1)+1\right] \leq \frac{p}{2p-1}\left(\frac{\Phi_0}{\eta h(\delta_{\epsilon'})}\right),$$

which achieves the proof. □

CHAPITRE 6 ARTICLE 3: CONSTRAINED STOCHASTIC BLACKBOX OPTIMIZATION USING A PROGRESSIVE BARRIER AND PROBABILISTIC ESTIMATES

Kwassi Joseph Dzahini, Michael Kokkolaras and Sébastien Le Digabel. Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. Submitted for publication to Mathematical Programming.

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 functions values are numerically unavailable, StoMADS-PB uses estimates and introduces so-called 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 iterates, 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.

6.1 Introduction

Blackbox optimization (BBO) considers the development and analysis of algorithms designed for objectives and constraints functions that are given by a process called a blackbox which returns an output when provided an input but whose inner workings are analytically unavailable [16]. Mesh adaptive direct-search (MADS) [12,13] with progressive barrier (PB) is an algorithm for deterministic BBO. This work considers the following constrained stochastic BBO problem

$$\min_{x \in \mathcal{D}} f(x) \tag{6.1}$$

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 typically 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 [70].

Solving stochastic blackbox optimization problems such as Problem (6.1), which often arise in signal processing and machine learning [43], has recently been a topic of intense research. Most methods for solving such problems borrow ideas from the stochastic gradient method [85]. 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 [25] which is among the first to propose a stochastic variant of the deterministic Nelder-Mead (NM) method [80], [5] 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 [50]. Another stochastic variant of NM was recently proposed in [31] and was proved to have global convergence properties with probability one. Using elements from [24,68], [35] proposed STORM, a trust-region algorithm designed for stochastic optimization problems, with almost sure global convergence results. Many other researches that extend the traditional deterministic trust-region method to stochastic setting have been conducted in [44,89,96]. In [83], a classical backtracking Armijo line search method [8] has been adapted to the stochastic optimization setting and was shown to have first-order complexity bounds. Robust-MADS, a kernel smoothing-based vari-

ant of MADS [12], was proposed in [18] to approach the minimizer of an objective function whose values can only be computed with a random noise. It was shown to possess zeroth-order [14] convergence properties. Another stochastic variant of MADS was proposed in [3] 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. [15] proposed another stochastic optimization approach using an algorithmic framework similar to that of MADS. StoMADS uses estimates of function values obtained from stochastic observations. 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 [38] stationarity convergence result of StoMADS has been derived with probability one, using martingale theory. A general framework for stochastic directional direct-search [42] methods was introduced in [51] with expected complexity analysis.

All the above stochastic optimization methods are restricted to unconstrained problems and most of them use estimated gradient information when seeking for an optimal solution. When the gradient does not exist or is computationally expensive to estimate, heuristics such as simulated annealing methods, genetic algorithms [67], and tabu/scatter search [64], 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 [7] developed for expensive simulation. In [22], the capabilities of the deterministic constrained trust-region algorithm NOWPAC [21] are generalized for the optimization of blackboxes with inherently noisy evaluations of the objective and constraint functions. To mitigate the noise in the latter functions evaluations, the resulting gradient-free method SNOWPAC utilizes Gaussian process surrogate combined with local fully linear surrogate models. Another surrogate-based approach that has gained in increasing popularity in various research fields is Kriging, also known as Bayesian optimization [76]. Various Bayesian optimization methods for constrained stochastic BBO have been demonstrated to be efficient in practice [71, 97].

Developing direct-search methods for BBO has received renewed interest since such methods generally known to be reliable and robust in practice [9], appear to be 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) [47,48] 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 [56] 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 [13], using elements from [12, 13, 15, 24, 35, 83] and is, to the best of our knowledge, the first to propose a directional direct-search [42] stochastic BBO algorithm, capable to handle 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 gradient information to find descent directions or improve feasibility compared to prior work. Rather, it uses so-called probabilistic estimates [35] of the objective and constraint function values and also introduces probabilistic bounds on a 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 for the estimates and no assumption is made about the way they 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 6.2 presents the general framework of the proposed StoMADS-PB algorithm. Section 6.3 explains how the proposed method results in a stochastic process and discusses requirements on random estimates to guarantee convergence. It also shows how such estimates can be constructed in practice. Section 6.4 presents the main convergence results. Computational results are reported in Section 6.5 followed by a discussion and suggestions for future work. Additional results are provided as an annex.

6.2 The StoMADS-PB algorithm

StoMADS-PB is based on an algorithmic framework similar to that of MADS with PB [13]. For the needs of the convergence analysis of Section 6.4, deterministic constraint violations are aggregated into a single function h called the constraint violation function, defined using the ℓ_1-norm for needs of convergence studies as opposed to [13] where an ℓ_2-norm has been favored

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

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.

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 direction around x^k. Likewise, to decrease f, MADS with PB uses objective function values since they are

available in the deterministic setting.

The main challenge here is to guarantee for StoMADS-PB such decreases as well in f as in h whereas their function values are unavailable numerically, using only 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 so called ε-accurate estimates introduced in [35] and then presents the general framework of the proposed method.

6.2.1 Feasibility and objective function improvements

At iteration k, let x^k and $x^k + s^k$ be two points of $\mathcal{X}$. 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 that 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 so-called *estimated violations* $h_0^k(x^k)$ and $h_s^k(x^k + s^k)$ defined as follows

$$h_0^k(x^k) = \begin{cases} \sum_{j=1}^{m} \max\left\{c_{j,0}^k(x^k), 0\right\} & \text{if } x^k \in \mathcal{X} \\ +\infty & \text{otherwise} \end{cases} \tag{6.2}$$

$$\text{and} \quad h_s^k(x^k + s^k) = \begin{cases} \sum_{j=1}^{m} \max\left\{c_{j,s}^k(x^k + s^k), 0\right\} & \text{if } x^k + s^k \in \mathcal{X} \\ +\infty & \text{otherwise.} \end{cases} \tag{6.3}$$

In order for such estimated constraint violations to be reliable enough to determine whether $h(x^k + s^k) < h(x^k)$ or not, 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 [15] is adapted from [35].

Definition 10. *Let $\varepsilon > 0$ be a fixed constant and $\{\delta_p^k\}_{k\in\mathbb{N}}$ be a sequence of nonnegative real numbers. 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 [15], the role of δ_p^k will be played by the so-called *poll size* parameter introduced later in Section 6.2.2. The following result provides bounds on $h(x^k)$ and $h(x^k + s^k)$, respectively, which

will allow, later in Proposition 5, to guarantee 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 4. *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, with x^k and $x^k + s^k \in \mathcal{X}$. Then the followings hold:*

$$\ell_0^k(x^k) := \sum_{j=1}^{m} \max\left\{c_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\} \leq h(x^k) \leq \sum_{j=1}^{m} \max\left\{c_{j,0}^k + \varepsilon(\delta_p^k)^2, 0\right\} =: u_0^k(x^k) \tag{6.4}$$

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 $j \in J$, then it follows from Definition 10 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

$$\max\left\{c_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\} \leq \max\left\{c_j(x^k), 0\right\} \leq \max\left\{c_{j,0}^k + \varepsilon(\delta_p^k)^2, 0\right\}. \tag{6.5}$$

Finally, summing each term of (6.5) from $j = 1$ to m leads to (6.4). □

Definition 11. *The estimates $\ell_0^k(x^k)$ and $u_0^k(x^k)$ of* Proposition 4*, 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 information to identify a decrease in h and will be also useful to determine an iteration type later in Section 6.2.2.

Proposition 5. *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. Let $\gamma > 2$ be a constant. Then the following holds:*

$$\textit{if } h_s^k - h_0^k \leq -\gamma m\varepsilon(\delta_p^k)^2, \textit{ then } h(x^k + s^k) - h(x^k) \leq -(\gamma - 2)m\varepsilon(\delta_p^k)^2 < 0. \tag{6.6}$$

Proof. It follows from Proposition 4 that

$$h(x^k + s^k) - h(x^k) \leq \sum_{j=1}^{m} \max\left\{c_{j,s}^k + \varepsilon(\delta_p^k)^2, 0\right\} - \sum_{j=1}^{m} \max\left\{c_{j,0}^k - \varepsilon(\delta_p^k)^2, 0\right\}. \tag{6.7}$$

By noticing that

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

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

then it follows from (6.7) 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 $\gamma > 2$. □

The ε-reliable upper bound $u_0^k(x^k)$ previously obtained for $h(x^k)$ also allows 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 (6.4) 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 so-called *ε-feasible* and *ε-infeasible* points, making use of a nonnegative barrier threshold $h_{\max}^k$ which is introduced in the present research as in [13].

Definition 12. *Let $x^k \in \mathcal{X}$ be any trial point and $u_0^k(x^k)$ 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 is ε-feasible. The algorithm can be applied to any problem satisfying only the following assumption adapted from [13].

Assumption 10. *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 [15] provides a sufficient information to identify a decrease in f and also allows to determine an iteration type in Section 6.2.2.

Proposition 6. *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:*

$$\text{if } f_s^k - f_0^k \leq -\gamma\varepsilon(\delta_p^k)^2, \text{ then } f(x^k + s^k) - f(x^k) \leq -(\gamma - 2)\varepsilon(\delta_p^k)^2 < 0. \tag{6.8}$$

Proof. The proof follows from Definition 10 and the next 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 by ranking trial mesh points of $\mathcal{X}$, making use of the following dominance notion inspired from [13].

Definition 13. *The ε-feasible point $x^k + s^k$ is said to dominate the ε-feasible point x^k, denoted $x^k + s^k \prec_{f;\varepsilon} x^k$, when $f_s^k - f_0^k \leq -\gamma\varepsilon(\delta_p^k)^2$, with $u_s^k(x^k + s^k) = 0$.*
The ε-infeasible point $x^k + s^k$ is said to dominate the ε-infeasible point x^k, denoted $x^k + s^k \prec_{h;\varepsilon} x^k$, when $f_s^k - f_0^k \leq -\gamma\varepsilon(\delta_p^k)^2$ and $h_s^k - h_0^k \leq -\gamma m\varepsilon(\delta_p^k)^2$, with $0 < u_s^k(x^k + s^k) \leq h_{\max}^k$.

Definition 14. *Let $\mathscr{E}_k$ be the set of points where the objective and constraint functions have been evaluated at a given iteration k. Let $t \geq 0$ be the iteration where a first ε-feasible point is found. Then $x_{\text{feas}}^{t+1} \in \{x^t + s^t \in \mathscr{E}_t : u_s^t(x^t + s^t) = 0\}$ is an ε-feasible incumbent solution at iteration $t+1$. Define the set $\mathscr{F}_k(y^k) = \left\{x^k + s^k \in \mathscr{E}_k : u_s^k(x^k + s^k) = 0 \text{ and } x^k + s^k \prec_{f;\varepsilon} y^k\right\}$ for all $k \geq t+1$ with $\mathscr{F}_k(x^k) = \emptyset$ if $k \leq t$. For all $k \geq 0$, 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 $\mathscr{I}_k(x^k) = \left\{x^k + s^k \in \mathscr{E}_k : h_s^k(x^k + s^k) - h_0^k(x^k) \leq -\gamma m\varepsilon(\delta_p^k)^2\right\}$. Let $x_{\text{inf}}^0 \in \mathcal{X}$ be a starting point. For all $k \geq t+1$, an ε-feasible incumbent solution at iteration $k+1$ is defined as follows:*

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

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

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

6.2.2 The StoMADS-PB algorithm and parameter update

Recall first that MADS with PB is an iterative algorithm where every iteration comprises two main steps: an optional step called the SEARCH, and the POLL. The SEARCH which typically consists of a global exploration may use a plethora of strategies like those based on interpolatory models, heuristics and surrogate functions or simplified physics models [13] 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 variables space, whose coarseness or fineness is controlled by a mesh size parameter δ_m^k thus deviating from the notation Δ_k^m from [13], 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 around $x^k \in \mathcal{M}^k$. The POLL step is governed by the poll size parameter δ_p^k which is linked to δ_m^k by $\delta_m^k = \min\{\delta_p^k, (\delta_p^k)^2\}$ [16]. As specified earlier, $\{\delta_p^k\}_{k\in\mathbb{N}}$ will play the role of the sequence of nonnegative real numbers introduced in Definition 10. 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 6.4, note also that as in [15], 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 6.3.1 to be integrable. The definitions of the mesh $\mathcal{M}^k$ and the POLL set $\mathcal{P}^k$ inspired from [13] are given next.

Definition 15. *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 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 $\mathbb{D}_p^k = \mathbb{D}_p^k(x_{\text{inf}}^k)$.

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 $x^k \in \mathcal{V}^k$ and $x^k+s^k \in \mathcal{P}^k$, but also the 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 also the way δ_p^k is updated. Adapting the terminologies from [13] and depending on the values of the aforementioned estimates and bounds, there are four StoMADS-PB iterations 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 found a first ε-feasible iterate or a trial point that dominates an incumbent is generated. An iteration which is Improving is not domi-

nating but aims to improve the feasibility of the ε-infeasible incumbent. Unsuccessful iterations are neither dominating nor improving.

— At the beginning of iteration k, if there is no available ε-feasible solution, then the iteration is called f-Dominating if for $x^k \in \mathcal{V}^k$, a first 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 4, meaning that $x^k + s^k$ is ε-feasible. Otherwise, if an ε-feasible point that dominates the incumbent is generated, i.e., $x^k + s^k \prec_{f;\varepsilon} x^k_{\text{feas}}$ for some $x^k \in \mathcal{V}^k$, then the inequality $f_s^k(x^k + s^k) - f_0^k(x^k_{\text{feas}}) \leq -\gamma\varepsilon(\delta_p^k)^2$ leads to a decrease in f due to Proposition 6. In either case, $x^{k+1}_{\text{feas}} := x^k + s^k$ and $\delta_p^{k+1} = \min\{\tau^{-1}\delta_p^k, \tau^{-\hat{z}}\}$. The ε-infeasible incumbent x^k_{inf} 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., $x^k_{\text{inf}} + s^k \prec_{h;\varepsilon} x^k_{\text{inf}}$, which means that both inequalities $f_s^k(x^k_{\text{inf}} + s^k) - f_0^k(x^k_{\text{inf}}) \leq -\gamma\varepsilon(\delta_p^k)^2$ and $h_s^k(x^k_{\text{inf}} + s^k) - h_0^k(x^k_{\text{inf}}) \leq -\gamma m\varepsilon(\delta_p^k)^2$ hold. Consequently, it follows from Propositions 5 and 6 that decreases occur both in f and h. In this case, $x^{k+1}_{\text{feas}} = x^k_{\text{feas}}$ and since feasibility is improved, x^{k+1}_{inf} is set to equal $x^k_{\text{inf}} + s^k$ while the poll size parameter is updated as at f-Dominating iterations.

— Iteration k is said to be Improving if it is not dominating but there is at least one ε-infeasible point $x^k_{\text{inf}} + s^k$ satisfying $h_s^k(x^k_{\text{inf}} + s^k) - h_0^k(x^k_{\text{inf}}) \leq -\gamma m\varepsilon(\delta_p^k)^2$. Indeed, this means that $x^k_{\text{inf}} + s^k$ improves the feasibility of the ε-infeasible incumbent x^k_{inf} since the previous inequality leads to a decrease in h due to Proposition 5. In this case, δ_p^k is updated as in dominating iterations, $x^{k+1}_{\text{feas}} = x^k_{\text{feas}}$ while the ε-infeasible incumbent is updated according to

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

— 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^k_{inf} nor x^k_{feas} are updated.

Now, note that while x^k_{inf} is updated at the end of each iteration of StoMADS-PB, the barrier threshold is rather computed at the beginning of every iteration according to $h^k_{\max} = u_0^k(x^k_{\text{inf}})$ to prevent it from 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 accuracy is improved compared to previous iterations as discussed later in Section 6.3.2. Consequently, even though the sequence $\{h^k_{\max}\}_{k\in\mathbb{N}}$ has a globally decreasing tendency, it is not nonincreasing as in MADS with PB, but can possibly increase between successive iterations. The goal of StoMADS-PB is to accept only the trial points satisfying $h(x^k) \leq h^k_{\max}$, and any trial point x^k for which the inequality $u_0^k(x^k) \leq h^k_{\max}$ does not hold is discarded from consideration since such an inequality implies that $h(x^k) \leq h^k_{\max}$ due

to (6.4). 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 5. *Denote by $t \geq 0$ the number of the first f-Dominating iteration of Algorithm 7 and assume that $t < +\infty$. Then it is easy to notice that $x_{\text{feas}}^k = x_{\text{inf}}^0$ for all $k = 0, 1, \ldots, t$ while $x_{\text{feas}}^{t+1} \neq x_{\text{inf}}^0$. 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 $k \leq t$, they are not used by the algorithm until the end of iteration t and it can also be noticed that no point in $\mathcal{P}^k$ that is generated using $\mathbb{D}_p^k(x_{\text{feas}}^k)$ is evaluated until the end of iteration t. In fact, setting the initial ε-feasible guess to equal x_{inf}^0 as it is in Algorithm 7 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 6.3.*

6.2.3 Frame center selection rule

Before describing the frame center selection rule, recall the set $\mathcal{V}^k$ of incumbent solutions introduced in Definition 15 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 [13], 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. The primary frame center in [13] is chosen to be the infeasible incumbent solution while the secondary frame center is the feasible incumbent whenever $f_k^F - \rho > f_k^I$, where the positive scalar ρ is the so called frame center trigger, f_k^F and f_k^I are respectively the incumbent feasible and infeasible f-values at iteration k. Otherwise if the previous inequality does not hold, the primary and secondary frame centers are the feasible and infeasible incumbent solutions. Because of the unavailability of f function values for StoMADS-PB, a specific frame center selection strategy using estimates of such function values is proposed and relies on the following result.

Proposition 7. *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.*

$$\textit{If } f_0^k(x_{\text{feas}}^k) - \rho > f_0^k(x_{\text{inf}}^k) + 2\varepsilon(\delta_p^k)^2, \textit{ then } f(x_{\text{feas}}^k) - \rho > f(x_{\text{inf}}^k). \tag{6.9}$$

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

$$\begin{aligned} f(x_{\text{inf}}^k) - f(x_{\text{feas}}^k) &= \left[f(x_{\text{inf}}^k) - f_0^k(x_{\text{inf}}^k)\right] + \left[f_0^k(x_{\text{inf}}^k) - f_0^k(x_{\text{feas}}^k)\right] + \left[f_0^k(x_{\text{feas}}^k) - f(x_{\text{feas}}^k)\right] \\ &< 2\varepsilon(\delta_p^k)^2 - (\rho + 2\varepsilon(\delta_p^k)^2) = -\rho. \end{aligned} \tag{6.10}$$

□

Thus motivated by Proposition 7, 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 a sufficient decrease condition leading to the inequality $f(x_{\text{feas}}^k) - \rho > f(x_{\text{inf}}^k)$, which as in [13] allows the choice of the infeasible incumbent solution as primary frame center.

As in [13], StoMADS-PB as implemented for the computational study in Section 6.5 places less effort in polling around the secondary frame center than the primary one. Specifically, the default strategy is to use a maximal positive basis [16] for the primary frame center and only two directions with one being the negative of the first for the secondary frame center.

6.3 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 or a possible outcome. 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 $B \in \mathcal{B}(\mathbb{R}^n)$ [28, 51].

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 7 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 7 is influenced by the randomness stemming from such random estimates, Algorithm 7 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

Algorithm 7: 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 $h_{\max}^k \leftarrow u_0^k(x_{\text{inf}}^k)$

f-Dominating

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 $s^k \in \{\delta_m^k d^k : d^k \in \mathbb{D}_p^k(x^k)\}$
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]**

h-Dominating

else if $x_{\text{inf}}^k + s^k \prec_{h;\varepsilon} x_{\text{inf}}^k$ for some $s^k \in \{\delta_m^k d^k : d^k \in \mathbb{D}_p^k(x_{\text{inf}}^k)\}$
set $x_{\text{inf}}^{k+1} \leftarrow x_{\text{inf}}^k + s^k$, $x_{\text{feas}}^{k+1} \leftarrow x_{\text{feas}}^k$ and $\delta_p^{k+1} \leftarrow \min\{\tau^{-1}\delta_p^k, \tau^{-\hat{z}}\}$

Improving

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

Unsuccessful

otherwise, 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 6.1 StoMADS-PB algorithm for constrained stochastic optimization.

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 [15, 29, 35, 51, 83], 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 [15], the notation "$f(X^k)$" is used to denote the random variable with realizations $f(X^k(\omega))$.

The present research aims to show that the stochastic process $\left\{X_{\text{inf}}^k, X_{\text{feas}}^k, \Delta_p^k, \Delta_m^k, F_0^k, F_s^k, H_0^k, H_s^k, L_0^k, U_0^k, L_s^k, U_s^k\right\}$ resulting from Algorithm 7 converges 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 accurate while the bounds will be assumed to be reliable, with sufficiently high but fixed probabilities, conditioned on the past.

6.3.1 Probabilistic bounds and probabilistic estimates

The previously mentioned notion of conditioning on the past is formalized following [15, 29, 35, 51, 83]. 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 $X^\ell \in \left\{X_{\text{inf}}^\ell, X_{\text{feas}}^\ell\right\}$ 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 subsequence of increasing σ-algebras of $\mathcal{G}$.

Sufficient accuracy of functions estimates is measured using the poll size parameter and is formalized, following [15, 29, 35, 51, 83] by means of the definitions bellow.

Definition 16. *A sequence of random estimates $\{F_0^k, F_s^k\}$ is said to be β-probabilistically ε-accurate with respect to the corresponding sequence $\{X^k, S^k, \Delta_p^k\}$ if the events*

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

satisfy the following 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 17. *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 following submartingale-like condition

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

To formalize the sufficient reliability of random bounds in the present work, the following definition is introduced.

Definition 18. *A sequence of random bounds $\{L_0^k, U_0^k, L_s^k, U_s^k\}$ is said to be α-probabilistically ε-accurate with respect to the corresponding sequence $\{X^k, S^k, \Delta_p^k\}$ if the events*

$$\begin{aligned} I_k \;=\; & \Big\{\text{“}L_0^k \text{ and } U_0^k \text{ are } \varepsilon\text{-reliable bounds for } h(x^k)\text{”, and “}L_s^k \text{ and } U_s^k \text{ are } \varepsilon\text{-reliable bounds} \\ & \text{for } h(x^k + s^k)\text{”, respectively for } \Delta_p^k\Big\} \end{aligned} \tag{6.11}$$

satisfy the following 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 [15,28] is defined below and will be useful for the analysis of Algorithm 7.

Definition 19. *Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space and $p \in [1, +\infty)$ 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 [15], 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 $\mathbb{E}\left(h(X^k)|\mathcal{F}_{k-1}^{C\cdot F}\right)$ can be well defined [28].

Assumption 11. *The objective function f and the constraints 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 iterates realizations is compact.*

Local Lipschitz in the above assumption means, Lipschitz with a finite constant in some nonempty neighborhood intersected with $\mathcal{X}$ [13].

Proposition 8. *Under Assumption 11, 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. The proof is inspired from [15]. Since f is locally Lipschitz on the compact set $\mathcal{U}$, the it 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 nature of the stochastic information in Algorithm 7, some of which are made in [15] and which will be useful for the convergence analysis of Section 6.4.

Assumption 12. *For fixed α and $\beta \in (0,1)$, the followings hold for the random quantities generated by Algorithm 7.*

(i) The sequence of estimates $\{F^k_0, F^k_s\}$ generated by Algorithm 7 is β-probabilistically ε-accurate.

(ii) The sequence of estimates $\{F^k_0, F^k_s\}$ generated by Algorithm 7 satisfies the following variance condition for all $k \geq 0$,

$$\begin{aligned} &\mathbb{E}\left(\left|F^k_s - f(X^k + S^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2(1-\sqrt{\beta})(\Delta^k_p)^4 \\ \text{and} \quad &\mathbb{E}\left(\left|F^k_0 - f(X^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2(1-\sqrt{\beta})(\Delta^k_p)^4. \end{aligned} \tag{6.12}$$

(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 $k \geq 0$,

$$\begin{aligned} &\mathbb{E}\left(\left|C^k_{j,s} - c_j(X^k + S^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2\left(1-\alpha^{1/2m}\right)(\Delta^k_p)^4 \\ \text{and} \quad &\mathbb{E}\left(\left|C^k_{j,0} - c_j(X^k)\right|^2 \mid \mathcal{F}^{C\cdot F}_{k-1}\right) \leq \varepsilon^2\left(1-\alpha^{1/2m}\right)(\Delta^k_p)^4. \end{aligned} \tag{6.13}$$

(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, it will be called "*false*". Even though the present algorithmic framework does

not allow to determine which iterations are true or false, Theorem 12 shows that true iterations occur infinitely often for convergence to hold, provided that estimates and bounds are sufficiently accurate. Theorem 12 will also be useful for the convergence analysis of Algorithm 7, more precisely in Subsection 6.4.3.

Theorem 12. *Assume that Assumption 12 holds for $\alpha\beta \in (1/2, 1)$. Then true iterations of Algorithm 7 occur infinitely often.*

Proof. Consider the following random walk

$$W_k = \sum_{i=0}^{k} (2 \cdot \mathbb{1}_{I_i} \mathbb{1}_{J_i} - 1). \tag{6.14}$$

Then, the result easily follows from the fact that $\left\{ \limsup_{k \to +\infty} W_k = +\infty \right\}$ almost surely, the proof of which can be derived from that of Theorem 4.16 in [35] (using $\mathcal{F}_{k-1}^{C \cdot F}$ instead of $\mathcal{F}_{k-1}^{I \cdot J}$), 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.

Lemma 9. *Let Assumption 12-(iv) hold for a fixed $\alpha \in (0, 1)$. Then the sequence of random estimated violations $\{H_0^k, H_s^k\}$ satisfies*

$$\begin{aligned} & \mathbb{E}\left(\left| H_s^k - h(X^k + S^k) \right| \mid \mathcal{F}_{k-1}^{C \cdot F} \right) \leq m\varepsilon(1-\alpha)^{1/2} (\Delta_p^k)^2 \\ \textit{and} \quad & \mathbb{E}\left(\left| H_0^k - h(X^k) \right| \mid \mathcal{F}_{k-1}^{C \cdot F} \right) \leq m\varepsilon(1-\alpha)^{1/2} (\Delta_p^k)^2. \end{aligned} \tag{6.15}$$

Proof. Before proving (6.15), let first notice that

$$\begin{aligned} \left| H_0^k - h(X^k) \right| &= \left| \sum_{j=1}^{m} \max\{C_{j,0}^k, 0\} - \sum_{j=1}^{m} \max\{c_j(X^k), 0\} \right| \\ &\leq \sum_{j=1}^{m} \left| \max\{C_{j,0}^k, 0\} - \max\{c_j(X^k), 0\} \right| \leq \sum_{j=1}^{m} \left| C_{j,0}^k - c_j(X^k) \right|, \end{aligned} \tag{6.16}$$

where the last inequality in (6.16) 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 [28] that for

all $j \in J$,

$$\begin{aligned}\mathbb{E}\left(\left|C_{j,0}^k - c_j(X^k)\right| |\mathcal{F}_{k-1}^{C\cdot F}\right) &\leq \left[\mathbb{E}\left(\left|C_{j,0}^k - c_j(X^k)\right|^2 |\mathcal{F}_{k-1}^{C\cdot F}\right)\right]^{1/2} \times \left[\mathbb{E}\left(1|\mathcal{F}_{k-1}^{C\cdot F}\right)\right]^{1/2} \\ &\leq \varepsilon\left(1-\alpha^{1/2m}\right)^{1/2}(\Delta_p^k)^2 \leq \varepsilon\,(1-\alpha)^{1/2}(\Delta_p^k)^2 \qquad (6.17)\end{aligned}$$

where the first inequality in (6.17) follows from (6.13). Thus, taking the conditional expectation with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ in (6.16) and then using (6.17) yield

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

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

□

6.3.2 Computation of probabilistically accurate estimates and reliable bounds

This section discusses approaches for computing accurate random estimates and reliable bounds satisfying Assumption 12 in a simple random noise framework, and hence how corresponding deterministic estimates can be obtained using evaluations of the stochastic blackbox. Such 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 [35].

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

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

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 12-(iii), define $C_{j,0}^k$ and $C_{j,s}^k$ respectively by

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

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

$$\mathbb{P}\left(\left|C_{j,0}^k - c_j(x^k)\right| > \varepsilon(\delta_p^k)^2\right) = \mathbb{P}\left(\left|C_{j,0}^k - \mathbb{E}\left(C_{j,0}^k\right)\right| > \varepsilon(\delta_p^k)^2\right) \leq \frac{V}{p_j^k \varepsilon^2 (\delta_p^k)^4}. \tag{6.18}$$

Thus, choosing p_j^k such that

$$p_j^k \geq \frac{V}{\varepsilon^2 \left(1 - \alpha^{1/2m}\right) (\delta_p^k)^4} \tag{6.19}$$

ensures that $\frac{V}{p_j^k \varepsilon^2 (\delta_p^k)^4} \leq 1 - \alpha^{1/2m}$. Then, combining (6.18) and (6.19) yields for all $j \in J$,

$$\mathbb{P}\left(\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right) \geq \alpha^{1/2m} \tag{6.20}$$

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

$$\mathbb{P}\left(\left\{\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\} \cap \left\{\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \alpha^{1/m}, \tag{6.21}$$

which means that Assumption 12-*(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 ε-accurate.

In order to satisfy Assumption 12-*(v)*, notice that the independence of the random variables $\Theta_j, j \in J$ combined with (6.20) implies

$$\mathbb{P}\left(\bigcap_{j=1}^m \left\{\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) = \prod_{j=1}^m \mathbb{P}\left(\left|C_{j,0}^k - c_j(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right) \geq \alpha^{1/2} \tag{6.22}$$

$$\text{and similarly,} \quad \mathbb{P}\left(\bigcap_{j=1}^m \left\{\left|C_{j,s}^k - c_j(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \alpha^{1/2}. \tag{6.23}$$

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

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

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

$$E_0^k = \left\{ L_0^k(x^k) \leq h(x^k) \leq U_0^k(x^k) \right\} \text{ and } E_s^k = \left\{ L_s^k(x^k + s^k) \leq h(x^k + s^k) \leq U_s^k(x^k + s^k) \right\} \tag{6.24}$$

By noticing that

$$\bigcap_{j=1}^{m} \left\{ \left| C_{j,0}^k - c_j(x^k) \right| \leq \varepsilon(\delta_p^k)^2 \right\} = \bigcap_{j=1}^{m} \left\{ C_{j,0}^k - \varepsilon(\delta_p^k)^2 \leq c_j(x^k) \leq C_{j,0}^k + \varepsilon(\delta_p^k)^2 \right\} \subseteq E_0^k \tag{6.25}$$

$$\bigcap_{j=1}^{m} \left\{ \left| C_{j,s}^k - c_j(x^k + s^k) \right| \leq \varepsilon(\delta_p^k)^2 \right\} \subseteq E_s^k, \tag{6.26}$$

then combining respectively (6.22) and (6.25), and (6.23) and (6.26), yields

$$\mathbb{P}\left(E_0^k\right) \geq \mathbb{P}\left(\bigcap_{j=1}^{m} \left\{ \left| C_{j,0}^k - c_j(x^k) \right| \leq \varepsilon(\delta_p^k)^2 \right\} \right) \geq \alpha^{1/2} \tag{6.27}$$

$$\mathbb{P}\left(E_s^k\right) \geq \mathbb{P}\left(\bigcap_{j=1}^{m} \left\{ \left| C_{j,s}^k - c_j(x^k + s^k) \right| \leq \varepsilon(\delta_p^k)^2 \right\} \right) \geq \alpha^{1/2}. \tag{6.28}$$

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, both inequalities (6.27) and (6.28) imply that

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

which shows that Assumption 12-*(v)* holds.

In order to show that Assumption 12-*(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 $j \in J$,

$$\mathbb{E}\left(\left| C_{j,0}^k - c_j(x^k) \right|^2 \right) = \mathbb{V}\left(C_{j,0}^k - c_j(x^k)\right) = \mathbb{V}\left(C_{j,0}^k\right) \leq \frac{V}{p_j^k} \leq \varepsilon^2 \left(1 - \alpha^{1/2m}\right) (\delta_p^k)^4, \tag{6.29}$$

where the last inequality in (6.29) follows from (6.19). Similarly, since $\mathbb{E}\left(C_{j,s}^k - c_j(x^k + s^k)\right) = 0$

for all $j \in J$, then

$$\mathbb{E}\left(\left|C_{j,s}^k - c_j(x^k + s^k)\right|^2\right) \leq \varepsilon^2 \left(1 - \alpha^{1/2m}\right) (\delta_p^k)^4, \tag{6.30}$$

which shows that Assumption 12-*(iv)* holds.

Finally, let compute estimates F_0^k and F_s^k that satisfy Assumption 12-*(i)* and *(ii)*. For that purpose, 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

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

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 12-*(i)* follows that of Assumption 12-*(iii)*. More precisely, the following inequality holds:

$$\mathbb{P}\left(\left\{\left|F_0^k - f(x^k)\right| \leq \varepsilon(\delta_p^k)^2\right\} \cap \left\{\left|F_s^k - f(x^k + s^k)\right| \leq \varepsilon(\delta_p^k)^2\right\}\right) \geq \beta, \tag{6.31}$$

provided that

$$p_0^k \geq \frac{V}{\varepsilon^2 \left(1 - \sqrt{\beta}\right) (\delta_p^k)^4} \tag{6.32}$$

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}, resulting from the evaluations of the stochastic blackbox, respectively at x^k and $x^k + s^k$, are obviously ε-accurate. It is also easy to notice that the proof of Assumption 12-*(ii)* follows that of Assumption 12-*(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 that p_0^k is chosen according to (6.32).

6.4 Convergence analysis

Using ideas inspired by [13, 15, 35, 68, 83] this section presents convergence results of StoMADS-PB, most of which are stochastic variants of those in [13]. It introduces the random time T at which Algorithm 7 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 [14, 15] is derived showing that there exists a subsequence of Algorithm 7-generated random iterates with mesh realizations becoming infinitely fine and which converges with probability one to a limit. This is achieved by showing by means of

Theorem 13 that the sequence of random poll size parameters converges to zero with probability one. Section 6.4.2 analyzes the function h and the random ε-infeasible iterates generated by Algorithm 7. In particular, it gives conditions under which an almost sure limit of a subsequence of such iterates is shown in Theorem 15 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 iterates is derived in Theorem 17 of Section 6.4.3. Note finally that the proofs of the main results of this section are presented in the Appendix.

6.4.1 Zeroth-order convergence

Recall Remark 5 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 iterates generated by Algorithm 7 until the beginning of iteration k. Consider the following random time T defined by

$$T := \inf\{k \geq 0 : \mathscr{S}_X^k \neq \emptyset\}. \tag{6.33}$$

Then it is easy to notice that $T \geq 1$ and that for all $k \geq 1$, the occurrence of the event $\{T \leq k\}$ is determined by observing the random quantities generated by Algorithm 7 until the iteration $k-1$, which means that T is a stopping time [50] for the stochastic process generated by Algorithm 7. The following is assumed for the remainder of the analysis.

Assumption 13. *The stopping time T associated to the stochastic process generated by Algorithm 7 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 poll size parameters converges to zero with probability one and will be useful for the Clarke stationarity results of Sections 6.4.2 and 6.4.3. It holds under the assumption below.

Assumption 14. *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 13. *Let Assumptions 11, 13 and 14 be satisfied. Let $\gamma > 2$ and $\tau \in (0,1)\cap\mathbb{Q}$. Let $\nu \in (0,1)$ be chosen such that*

$$\frac{\nu}{1-\nu} \geq \frac{2(\tau^{-2}-1)}{\gamma-2} \tag{6.34}$$

and assume that Assumption 12 holds for α and β chosen such that

$$\alpha\beta \geq \frac{4\nu}{(1-\nu)(1-\tau^2)}\left[(1-\alpha)^{1/2} + 2(1-\beta)^{1/2}\right]. \tag{6.35}$$

Then, the sequence $\{\Delta_p^k\}_{k\in\mathbb{N}}$ *of frame size parameters generated by Algorithm 7 satisfies*

$$\sum_{k=0}^{+\infty}(\Delta_p^k)^2 < +\infty \quad \textit{almost surely}. \tag{6.36}$$

The following result is a simple consequence of Theorem 13. 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 2. *The followings hold under all the assumptions made in Theorem 13*

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

The next result shows that with probability one, the difference between the estimates and their corresponding true function values converge to zero. This means that Algorithm 7 behaves like an exact deterministic method asymptotically. This result will be also useful in Subsection 6.4.3 for the proof of Theorem 16.

Corollary 3. *Let all assumptions that were made in Theorem 13 hold. Then,*

$$\lim_{k\to+\infty}\left|H_0^k - h(X^k)\right| = 0 \textit{ almost surely} \quad \textit{and} \quad \lim_{k\to+\infty}\left|F_0^k - f(X^k)\right| = 0 \textit{ almost surely}, \tag{6.37}$$

and the same result holds for $\left|H_s^k - h(X^k+S^k)\right|$ *and* $\left|F_s^k - f(X^k+S^k)\right|$ *respectively.*

Definition 20. *A convergent subsequence* $\{x^k\}_{k\in\mathcal{K}}$ *of Algorithm 7 iterates, 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.*

Combining the results of Corollary 2 and the compactness hypothesis of Assumption 11 was shown in [15] to be enough to ensure the existence of refining subsequences. Specifically the following holds.

Theorem 14. *Let the assumptions that were made in Corollary 2 hold. Then there exists at least one refining subsequence* $\{X^k\}_{k\in K}$ *(where* K *is a sequence of random variables) which converges almost surely to a refined point* $\hat{X}$*.*

6.4.2 Nonsmooth optimality conditions: Results for h

This subsection aims to show with probability one that Algorithm 7 generates a refining subsequence $\{X_{\text{inf}}^k\}_{k\in K}$ with refined point $\hat{X}_{\text{inf}}$ which satisfies a first-order necessary optimality condition via the

Clarke generalized derivative of h. As in [15], 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 7 are such that $\delta_p^k \left\| d^k \right\|_\infty$ never approaches zero for all k. The way such an expectation can be met is discussed in [15]. 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 [15] to satisfy $\delta_p^k \left\| d^k \right\|_\infty \geq 1$ whenever d^k is constructed by means of the so-called Householder matrix [16]. Thus, the following assumption is made for the remainder of the analysis.

Assumption 15. *Let $d^k \in \mathbb{D}_p^k$ be any polling direction used by Algorithm 7 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, a similar of which was used in [15].

Lemma 10. *Let the same assumptions that were made in Theorem 13 hold and assume in addition to* (6.35) *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 $d^k \in \mathbb{D}_p^k(x_{\text{inf}}^k)$ denotes any available polling direction around x_{inf}^k at iteration k. Then the following holds,

$$\liminf_{k \to +\infty} \Psi_k^h \leq 0 \textit{ almost surely.} \tag{6.38}$$

The following definition of refining directions [12, 16] will be useful in the analysis.

Definition 21. *Let $\hat{x}$ be the refined point associated to 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 $v = \lim_{k \in \mathcal{L}} \frac{d^k}{\left\| d^k \right\|_\infty}$.*

The analysis in this subsection also relies on the following definitions [13]. 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

$$h^\circ(\hat{x}; v) := \limsup_{\substack{y \to \hat{x},\ y \in \mathcal{X} \\ t \searrow 0,\ y + tv \in \mathcal{X}}} \frac{h(y + tv) - h(y)}{t}. \tag{6.39}$$

As highlighted in [13], this definition from [61] is a generalization of the original one by Clarke [38] to the case where the constraints violation function h is not defined outside $\mathcal{X}$.

The analysis involves a specific cone $T_{\mathcal{X}}^{H}(\hat{x}_{\inf})$ called the hypertangent cone [86] to $\mathcal{X}$ at $\hat{x}_{\inf}$. The hypertangent cone to a subset $\mathcal{O} \subseteq \mathcal{X}$ at $\hat{x}$ is defined by

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

Next is stated a lemma [13] from elementary analysis, that will be useful latter in the present analysis.

Lemma 11. *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 is a stochastic variant of Theorem 3.5 in [13]. Since the inequality $h(x_{\text{inf}}^k + \delta_m^k d^k) - h(x_{\text{inf}}^k) \geq 0$ on which relies the latter theorem does not hold in the present stochastic setting, then the proof of the result below is based on the random function Ψ_k^h $\liminf$-type result of Lemma 10.

Theorem 15. *Let Assumptions 10, 15 and all the assumptions made in Theorem 13 and Lemma 10 hold. Then Algorithm 7 generates a convergent ε-infeasible refining subsequence $\{X_{\text{inf}}^k\}_{k\in K}$, for some sequence $K \subseteq K'$ of random variables satisfying $\lim_{K'} \Psi_k^h \leq 0$ almost surely, such that if $\hat{x}_{\inf} \in \mathcal{X}$ 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 $v \in T_{\mathcal{X}}^{H}(\hat{x}_{\inf})$ is a refining direction for $\hat{x}_{\inf}$, then $h^{\circ}(\hat{x}_{\inf}; v) \geq 0$. In particular, this means that*

$$\begin{aligned}\mathbb{P}\Big(\Big\{\omega \in \Omega : \exists K(\omega) \subseteq \mathbb{N} \textit{ and } \exists \hat{X}_{\inf}(\omega) = \lim_{k\in K(\omega)} X_{\text{inf}}^k(\omega), \hat{X}_{\inf}(\omega) \in \mathcal{X}, \textit{ such that} \\ \forall V(\omega) \in T_{\mathcal{X}}^{H}(\hat{X}_{\inf}(\omega)),\ h^{\circ}(\hat{X}_{\inf}(\omega); V(\omega)) \geq 0\Big\}\Big) = 1.\end{aligned} \tag{6.40}$$

Next is stated a stochastic variant of a result in [13], showing that Clarke stationarity is ensured when the set of refining directions is dense in a nonempty hypertangent cone to $\mathcal{X}$.

Corollary 4. *Let all assumptions that were made in Theorem 15 hold. Let $\{X_{\text{inf}}^k\}_{k\in K}$ be the ε-infeasible refining subsequence of Theorem 15, with realizations $\{x_{\text{inf}}^k\}_{k\in\mathcal{K}}$ which converges to a refined point $\hat{x}_{\inf} \in \mathcal{X}$. If the set of refining directions for $\hat{x}_{\inf}$ is dense in $T_{\mathcal{X}}^{H}(\hat{x}_{\inf}) \neq \emptyset$, then $\hat{x}_{\inf}$ is a Clarke stationary point for the problem $\min_{x\in\mathcal{X}} h(x)$.*

Proof. The proof of this result is almost identical to the proof of a similar result (Corollary 3.6) in [13] and hence will not be presented here again. □

6.4.3 Nonsmooth optimality conditions: Results for f

The analysis presented in this subsection assumes that Algorithm 7 generates infinitely many ε-feasible points. It aims to show with probability one that StoMADS-PB generates a refining subsequence $\{X^k_{\text{feas}}\}_{k\in K}$ with refined point $\hat{X}_{\text{feas}}$, which satisfies a first-order necessary optimality condition based on the Clarke derivative of f. The following lemma will be useful latter in the analysis.

Lemma 12. *Let the same assumptions that were made in Theorem 13 hold and assume in addition to* (6.35) *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*

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

where $k \vee t := \max\{k, t\}$ *and* d^k *denotes any available polling direction around* $x^{k\vee t}_{\text{feas}}$ *at iteration* k. *Then the following holds,*

$$\liminf_{k\to+\infty} \Psi^{f,T}_k \leq 0 \textit{ almost surely.} \tag{6.41}$$

Now let prove that the almost sure limit $\hat{X}_{\text{feas}}$ of any convergent refining subsequence of ε-feasible iterates 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, notice that the existence of such a refining subsequence can be assumed. Indeed, it is known from Theorem 12 that true iterations occur infinitely often provided that estimates and bounds are sufficiently accurate. In addition, every ε-feasible point x^k_{feas} newly accepted by Algorithm 7 satisfies $u^k_0(x^k_{\text{feas}}) = 0$, which implies that $h^k_0(x^k_{\text{feas}}) = 0$, 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 next.

Theorem 16. *Let all the assumptions of Lemma 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*

$$\mathbb{P}\left(\hat{X}_{\text{feas}} \in \mathcal{D}\right) = 1. \tag{6.42}$$

The following result is a stochastic variant of Theorem 3.3 in [13].

Theorem 17. *Let Assumptions 10, 15 and all assumptions that were made in Theorem 13 and Lemma 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. Then, 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*

$$\mathbb{P}\left(\Big\{\omega \in \Omega : \exists K(\omega) \subseteq \mathbb{N} \text{ and } \exists \hat{X}_{\text{feas}}(\omega) = \lim_{k \in K(\omega)} X_{\text{feas}}^{k \vee T}(\omega), \hat{X}_{\text{feas}}(\omega) \in \mathcal{D}, \text{ such that } \forall V(\omega) \in T_{\mathcal{D}}^{H}(\hat{X}_{\text{feas}}(\omega)),\ f^\circ(\hat{X}_{\text{feas}}(\omega); V(\omega)) \geq 0\Big\}\right) = 1. \tag{6.43}$$

Corollary 5. *Let all assumptions that were made in Theorem 17 hold. Let* $\{X_{\text{feas}}^{k \vee T}\}_{k \in K}$ *be the* ε*-feasible refining subsequence of Theorem 17, with realizations* $\{x_{\text{feas}}^{k \vee t}\}_{k \in \mathcal{K}}$ *which converges to a refined point* $\hat{x}_{\text{feas}} \in \mathcal{D}$*. 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* (6.1)*.*

Proof. The proof of this result is almost identical to the proof of a similar result (Corollary 3.4) in [13] and hence will not be presented here again. □

6.5 Computational study

This section illustrates the performance and the efficiency of StoMADS-PB using noisy variants of 42 continuous analytical computational constrained problems from the optimization literature. The sources and characteristics of these problems are summarized in Table 6.1. 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 [13] referred to as MADS-PB, the latter algorithm is compared to several variants of StoMADS-PB. For all numerical investigations of both algorithms, only the POLL step is used, i.e., no SEARCH step is involved. Recall the frame center selection rule of Section 6.2.3 and that of MADS-PB [13]. The OrthoMADS-$2n$ directions [2] are used for the POLL which is ordered by means of an opportunistic strategy [16]. Indeed, 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 abovementioned deterministic constrained optimization problems are solved using three different infeasible initial points for a total of 126 problem instances. Inspired from [15], such stochastic variants are constructed by additively perturbing the objective f by a random variable Θ_0 and each constraint $c_j, j = 1, 2, \ldots, m$ by a random variable Θ_j as follows

$$f_{\Theta_0}(x) = f(x) + \Theta_0 \quad \text{and} \quad c_{\Theta_j}(x) = c_j(x) + \Theta_j, \text{ for all } j \in J, \tag{6.44}$$

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. 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 while reducing the value of the objective function from $f(x^0)$ to f^*, as well as those 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 noisy blackbox. Consider for example the SNAKE problem [13] in Table 6.1 whose objective and constraint functions are given by

$$f(x) = \sqrt{(x_1 - 20)^2 + (x_2 - 1)^2}, \quad c_1(x) = \sin(x_1) - x_2 - \frac{1}{10} \text{ and } c_2(x) = x_2 - \sin(x_1).$$

Consider the initial point $x^0 = (2, 2)$ and the feasible minimum value $f^* = 0.08$ [13]. Then, $|f(x^0) - f^*| = 17.95$, $|c_1(x^0)| = 1.19$ and $|c_2(x^0)| = 1.09$. Hence, $I(\sigma, x^0, f) = [-17.95\sigma, 17.95\sigma]$ while $I(\sigma, x^0, c_1) = [-1.19\sigma, 1.19\sigma]$, $I(\sigma, x^0, c_2) = [-1.09\sigma, 1.09\sigma]$ and the corresponding noisy functions are given by

$$\begin{aligned} f_{\Theta_0}(x) &= \sqrt{(x_1 - 20)^2 + (x_2 - 1)^2} + \Theta_0, \\ c_{\Theta_1}(x) &= \sin(x_1) - x_2 - \frac{1}{10} + \Theta_1 \quad \text{and} \quad c_{\Theta_2}(x) = x_2 - \sin(x_1) + \Theta_2, \end{aligned}$$

where the independent random variables Θ_0, Θ_1 and Θ_2 are uniformly generated in the intervals $I(\sigma, x^0, f)$, $I(\sigma, x^0, c_1)$ and $I(\sigma, x^0, c_2)$ respectively. Figures 6.2 and 6.3 depict $f(x)$ and realizations of $f_{\Theta_0}(x)$, and illustrate in $\mathbb{R}^2$ feasible domains with respect to the deterministic constraints $c_1(x) \leq 0$ and $c_2(x) \leq 0$, and realizations of the noisy constraints $c_{\Theta_1}(x) \leq 0$ and $c_{\Theta_2}(x) \leq 0$.

The MADS-PB algorithm [13] 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 in an adaptive manner 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 out of hand. Full description of MADS-PB iterations and useful information for better understanding of the algorithm behavior can also be found in [16].

The relative performance and efficiency of algorithms are assessed by performance profiles [49, 77] and data profiles [77], which require to define for a given computational problem a convergence test. For each of the 126 problems, denote by x^N the best feasible iterate found after N evaluations of

the noisy blackbox and let x^* be the best feasible point obtained by all tested algorithms on all run instances. Then, the convergence test from [20] used for the experiments is defined as follows:

$$f(x^N) \leq f(x^*) + \tau(\bar{f}_{\text{feas}} - f(x^*)), \tag{6.45}$$

where, $\tau \in [0,1]$ is the convergence tolerance and $\bar{f}_{\text{feas}}$ is a reference value obtained by taking the average of the first feasible f function values over all run instances of a given computational problem for 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 (6.45) holds. As highlighted in [20], $\bar{f}_{\text{feas}} = f(x^0)$ for unconstrained computational problems, where x^0 denotes the initial point.

The horizontal axis of the performance profiles shows the ratio of the number of noisy objective function evaluations while the fraction of computational 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 computational problems solved by all run instances of a given algorithm within a tolerance τ. As emphasized in [16], performance profiles capture information on speed of convergence (i.e., the quality of a given algorithm's output in terms of the objective function evaluations) and robustness (i.e., the fraction of computational problems solved) in a compact graphical format, while data profiles also examine the robustness and efficiency from a different perspective.

Now recall that in StoMADS-PB, according to Section 6.3.2, the noisy blackbox needs to be evaluated many times at a given point in order to compute function estimates unlike the MADS-PB method where it is evaluated only once at each point. But since a limited budget of $1000(n+1)$ noisy blackbox evaluations is set in all the experiments, that is, since MADS-PB and all variants of StoMADS-PB stop as soon as the number of noisy blackbox evaluations reaches $1000(n+1)$, only few calls to the blackbox need to be used when computing StoMADS-PB function estimates. However, given that such estimates are required to be sufficiently accurate in order for the solutions to be satisfactory, a procedure inspired from [15] aiming at improving the estimates accuracy by making use of available samples at a given current point is proposed. Note in passing that the proposed computation procedure is very efficient in practice as highlighted in [15] even though it is inherently biased. 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 $j \in J$. First, let mention that 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

1. $n+1$ is the number of evaluations required to construct a linear interpolant or a simplex gradient [16] in $\mathbb{R}^n$ [20,77].
2. It is implicitly assumed without any loss of generality that $a^k(x^k) \geq 1$.

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 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 following 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, an estimate $f_0^k(x^k)$ of $f(x^k)$ is computed according to,

$$f_0^k(x^k) = \frac{1}{p^k} \sum_{\ell=1}^{p^k} f_{\theta_{0,\ell}}(x^k), \tag{6.46}$$

where $p^k = n^k + a^k(x^k)$ is the sample size.

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 $\delta_m^0 = \delta_p^0 = 1$. Nevertheless in MADS-PB, the initial barrier threshold is set equal its default value, i.e., $h_{\max}^0 = +\infty$ [13] while in StoMADS-PB it equals $u_0^0(x_{\text{inf}}^0)$, with $u_0^k(x^k)$ defined in (6.4) for all $k \in \mathbb{N}$. The default values of Algorithm 7 parameters $\gamma > 2$ and $\varepsilon > 0$ [3] are borrowed from [15] in which StoMADS, an unconstrained stochastic variant of MADS [12] is introduced. Specifically, $\gamma = 17$ and $\varepsilon = 0.01$.

Three variants of StoMADS-PB corresponding to $n^k = 1, n^k = 2$ and $n^k = 3$ are compared to MADS-PB. The data and performance profiles used for the comparisons are depicted on Figures 6.4, 6.6 and 6.8 and Figures 6.5, 6.7 and 6.9. Three levels of noise are used during the experiments, which correspond to $\sigma = 0.01$, $\sigma = 0.03$ and $\sigma = 0.05$. For a given algorithm, the estimated percentages of problems solved after $1000(n+1)$ noisy blackbox evaluations for each noise level within a convergence tolerance τ are reported in Table 6.2. They are obtained based on the profiles graphs using MATLAB tools.

The data and performance profiles show that when given the time, StoMADS-PB eventually outperforms MADS-PB in general. Moreover as in [15], 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 6.2 that the higher the tolerance parameter τ, the larger the percentage of problems solved by all algorithms for a fixed noise level σ. Now notice that while for a given τ, the fraction of problems solved by MADS-PB decreases 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

3. The use of ε_f instead of ε is favored in [15].

Table 6.1 Description of the set of 42 analytical problems.

No	Name	Source	n	m	Bnds	No	Name	Source	n	m	Bnds
1	ANGUN	[97]	2	1	Yes	22	MAD1	[73]	2	1	No
2	BARNES	[87]	2	3	Yes	23	MAD2	[73]	2	1	No
3	BERTSIMAS	[27]	2	2	No	24	MAD6	[73]	7	7	Yes
4	CHENWANG_F2	[37]	8	6	Yes	25	MEZMONTES	[75]	2	2	Yes
5	CHENWANG_F3	[37]	10	8	Yes	26	NEW-BRANIN	[97]	2	1	Yes
6	CONSTR-BRANIN	[97]	2	1	Yes	27	OPTENG-BENCH4	[63]	2	1	Yes
7	CRESCENT	[13]	10	2	No	28	OPTENG-BENCH5	[63]	2	3	Yes
8	DEMBO5	[73]	8	3	Yes	29	OPTENG-RBF	[63]	3	4	Yes
9	DISK	[13]	10	1	No	30	PENTAGON	[73]	6	15	No
10	G23	[14]	3	2	Yes	31	PRESSURE-VESSEL	[75]	4	4	Yes
11	G210	[14]	10	2	Yes	32	SASENA	[97]	2	1	Yes
12	G220	[14]	20	2	Yes	33	SNAKE	[13]	2	2	No
13	GOMEZ	[97]	2	1	Yes	34	SPEED-REDUCER	[75]	7	11	Yes
14	HS15	[59]	2	2	Yes	35	SPRING	[87]	3	4	Yes
15	HS19	[59]	2	2	Yes	36	TAOWANG_F1	[92]	2	2	Yes
16	HS22	[59]	2	2	No	37	TAOWANG_F2	[92]	7	4	Yes
17	HS23	[59]	2	5	Yes	38	WELDED-BEAM	[75]	4	7	Yes
18	HS29	[59]	3	1	No	39	WONG2	[73]	10	3	No
19	HS43	[59]	4	3	No	40	ZHAOWANG_F5	[99]	13	9	Yes
20	HS108	[59]	9	13	Yes	41	ZILONG_G4	[97]	5	1	Yes
21	HS114	[59]	10	5	Yes	42	ZILONG_G24	[97]	2	1	Yes

Table 6.2 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 $n^k = 1$	74.6%	78.57%	73.02%	44.44%	45.24%	45.24%
StoMADS-PB $n^k = 2$	74.6%	76.98%	76.19%	47.62%	47.62%	50.79%
StoMADS-PB $n^k = 3$	76.19%	65.08%	66.67%	48.41%	41.27%	38.10%
MADS-PB	69.5%	64.29%	54.76%	41.27%	36.51%	29.37%

of the convergence test (6.45) does not depend only on the values of the objective function f but also on whether a feasible point is found or not, unlike the framework of [15] where no constraints are involved. In fact, as highlighted in [15] from which is inspired the computation scheme (6.46), even though the robustness and efficiency of each StoMADS-PB variants 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 MADS-PB. On the other hand, such computation scheme naturally favors StoMADS-PB by improving the accuracy of the estimates of its constraints function values, thus allowing it to find more feasible solutions than MADS-PB and consequently possibly solve larger fraction of problems when the noise level increases for a fixed tolerance parameter τ.

Finally, based on Table 6.2, it can be noticed that for a given convergence tolerance τ, varying σ seems not to have significant influences on the fractions of problems solved by StoMADS-PB variants corresponding to $n^k = 1$ and $n^k = 2$. 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 that of StoMADS-PB with $n^k = 2$. For all these reasons, the latter variant seems preferable for constrained stochastic blackbox optimization problems.

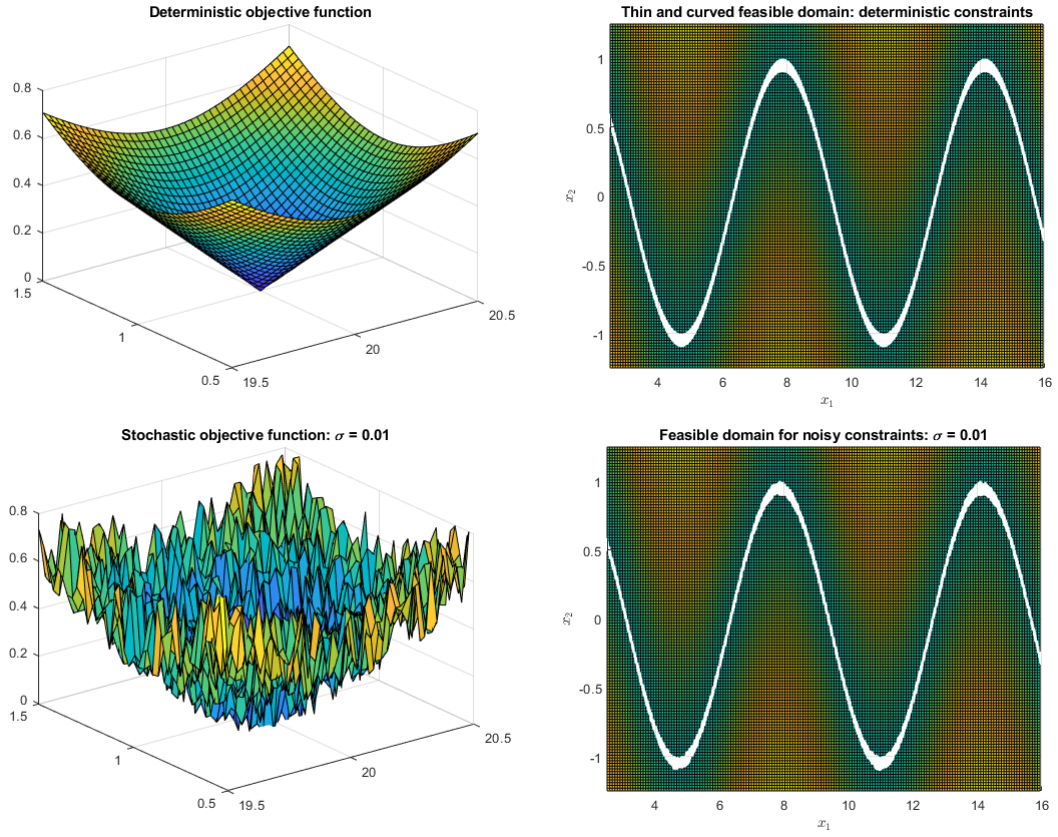


Figure 6.2 Plots of the deterministic objective function, a corresponding realization of f_{Θ_0} and two dimensional illustrations of feasible domains with respect to the deterministic constraints and corresponding realizations of the noisy constraints $c_{\Theta_1}(x) \leq 0$ and $c_{\Theta_2}(x) \leq 0$, for the SNAKE problem. The random variables Θ_0, Θ_1 and Θ_2 defining the noisy functions f_{Θ_0}, c_{Θ_1} and c_{Θ_2} are uniformly generated in $[-17.95\sigma, 17.95\sigma]$, $[-1.19\sigma, 1.19\sigma]$ and $[-1.09\sigma, 1.09\sigma]$, respectively.

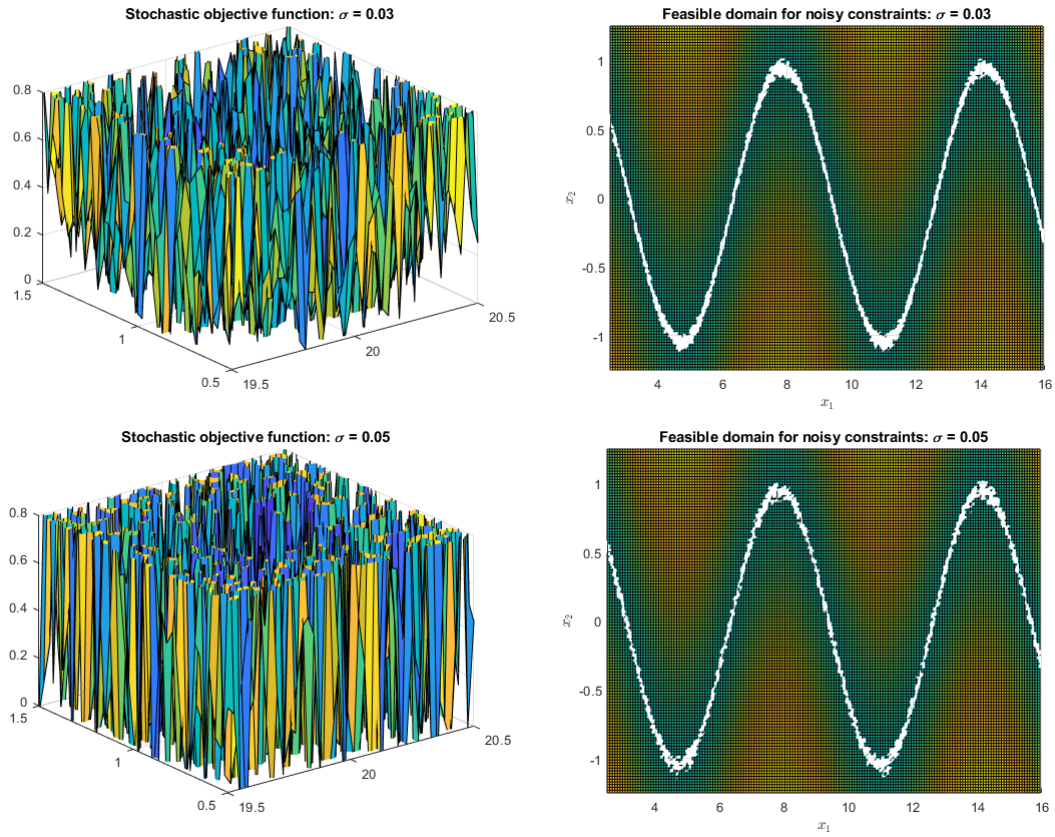


Figure 6.3 Plots of realizations of the noisy objective function f_{Θ_0} and two dimensional illustrations of feasible domains with respect to realizations of the noisy constraints $c_{\Theta_1}(x) \leq 0$ and $c_{\Theta_2}(x) \leq 0$, for the SNAKE problem. The random variables Θ_0, Θ_1 and Θ_2 defining the noisy functions f_{Θ_0}, c_{Θ_1} and c_{Θ_2} are uniformly generated in $[-17.95\sigma, 17.95\sigma]$, $[-1.19\sigma, 1.19\sigma]$ and $[-1.09\sigma, 1.09\sigma]$, respectively.

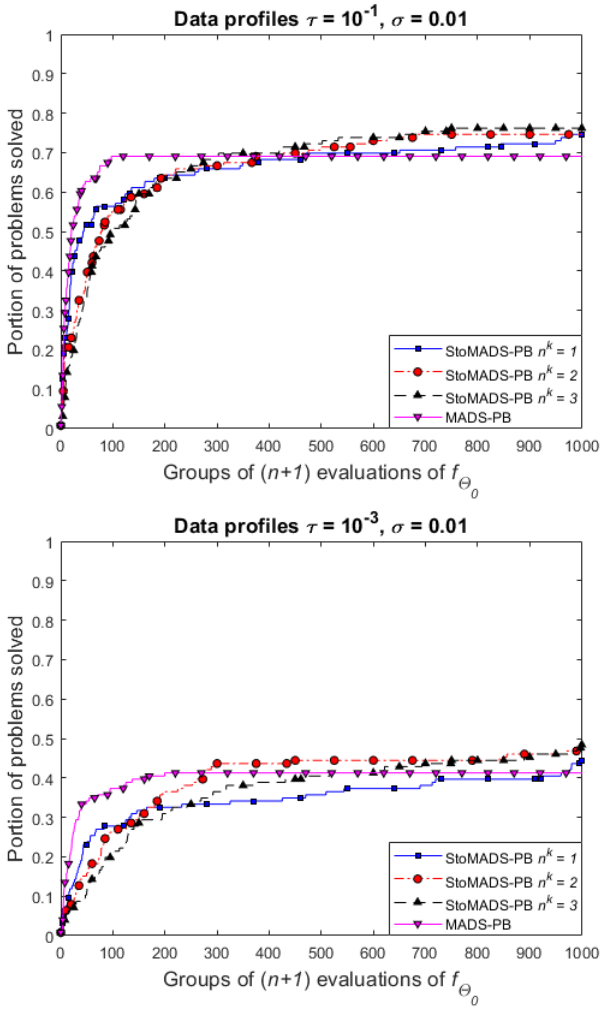


Figure 6.4 Data profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.01$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.

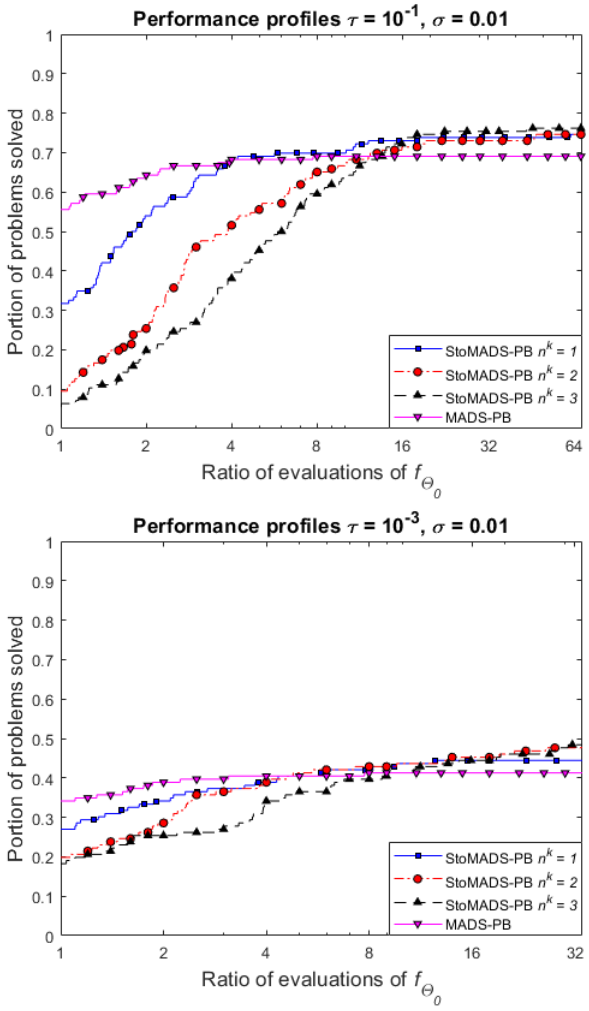
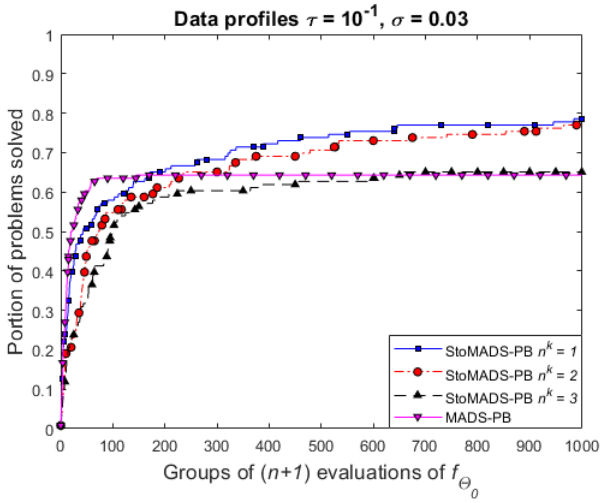


Figure 6.5 Performance profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.01$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.



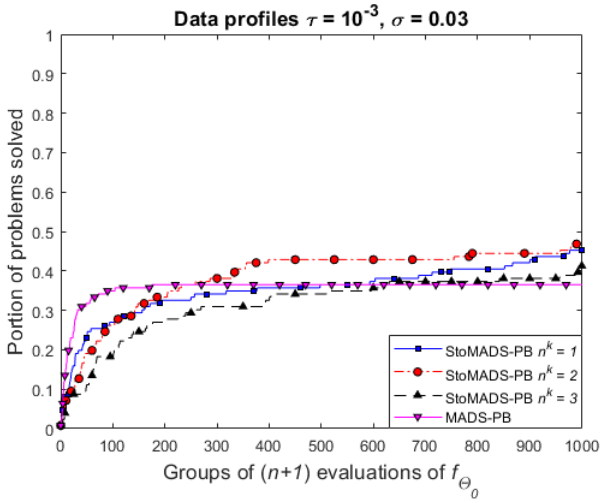
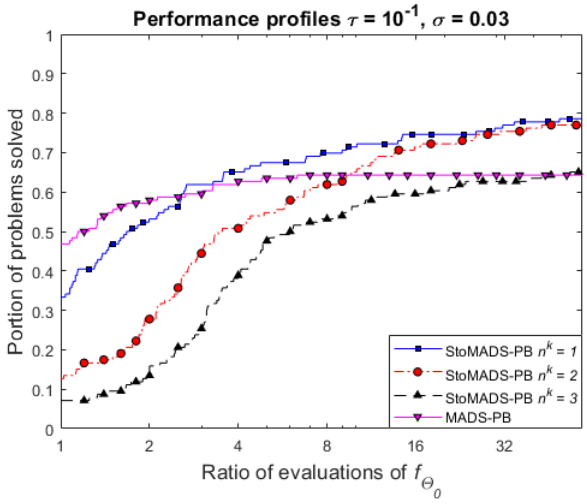


Figure 6.6 Data profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.03$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.



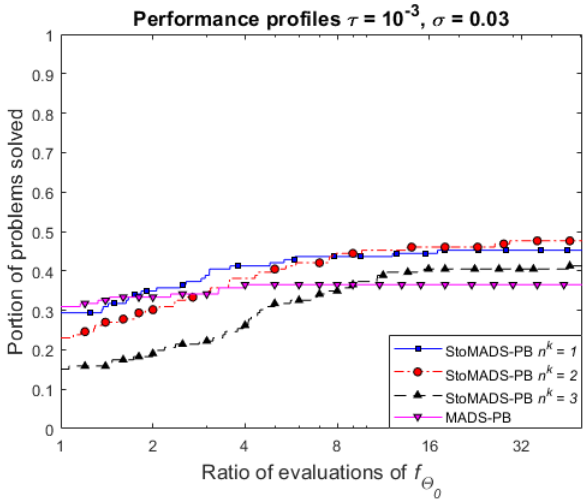
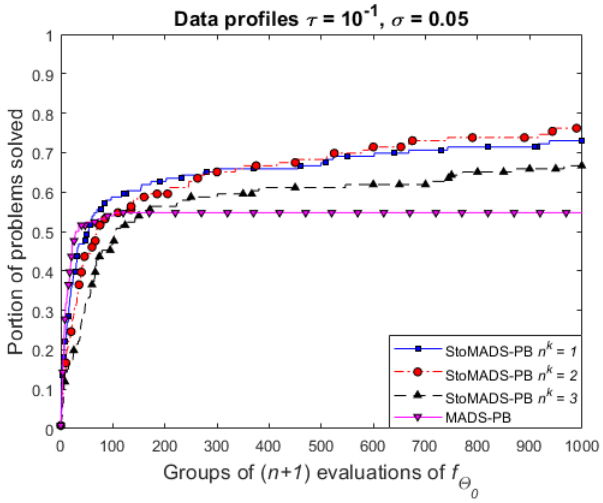


Figure 6.7 Performance profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.03$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.



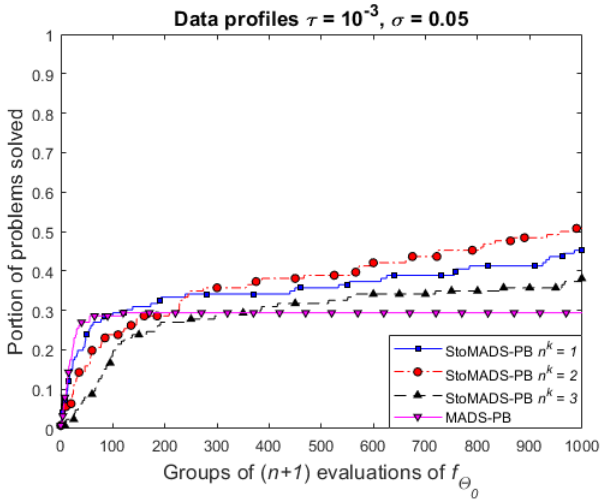
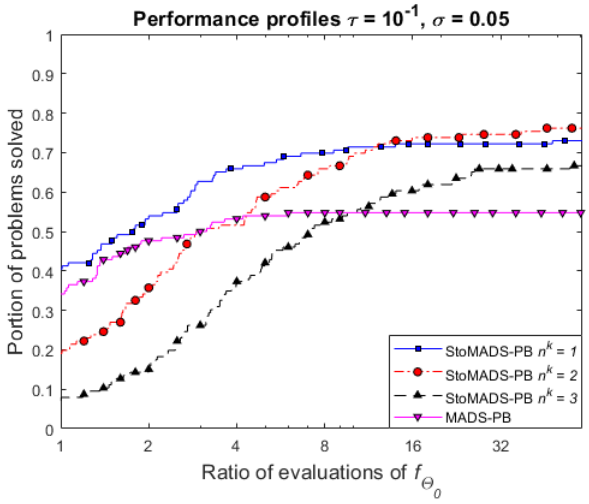


Figure 6.8 Data profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.05$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.



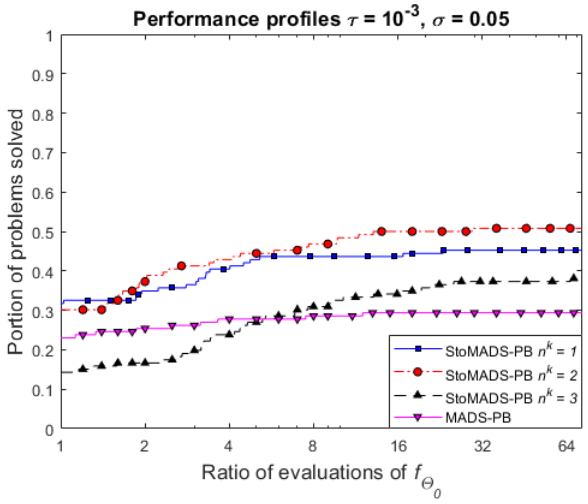


Figure 6.9 Performance profiles for convergence tolerances $\tau = 10^{-1}$ and $\tau = 10^{-3}$, and noise level $\sigma = 0.05$ on 126 analytical constrained test problems additively perturbed in the intervals $I(\sigma, x^0, f)$ and $I(\sigma, x^0, c_j)$.

6.6 Concluding remarks

This research proposes the StoMADS-PB algorithm for constrained stochastic blackbox optimization. The proposed method which uses an algorithmic framework similar to that of MADS considers the optimization of objective and constraints 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 gradient information to find descent directions or improve feasibility unlike prior works, but instead, 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 of 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 to eventually outperform MADS, and also showed 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 a noisy objective and constraints that are also stochastically noisy.

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

Now we prove a sequence of convergence results of Section 6.4.

Proof of Theorem 13

Proof. This theorem is proved using ideas from [15, 29, 35, 51, 68, 83, 96]. According to Assumptions 13, the proof considers two different parts: Part 1 assumes that $T = +\infty$ almost surely, i.e., no ε-feasible iterate is found by Algorithm 7, while Part 2 considers that $T < +\infty$ almost surely. Part 1 considers two separate cases: "good bounds" and "bad bounds", each of which is broken into whether an iteration is h-Dominating, Improving or Unsuccessful. Part 2 considers three separates cases: "good estimates and good bounds", "bad estimates and good bounds" and "bad bounds", each of which is broken into whether an iteration is f-Dominating, h-Dominating, Improving or Unsuccessful.

In order to show (6.36), 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 $k \in \mathbb{N}$

$$\mathbb{E}\left(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2, \tag{6.47}$$

where Φ_k is the random function defined by

$$\Phi_k := \frac{\nu}{m\varepsilon} h(X_{\text{inf}}^k) + (1-\nu)(\Delta_p^k)^2, \quad \text{for all } k \in \mathbb{N}. \tag{6.48}$$

Indeed, assume that (6.47) holds. Since $\Phi_k > 0$ for all $k \in \mathbb{N}$, then summing (6.47) over $k \in \mathbb{N}$ and taking expectations on both sides lead to

$$\mathbb{E}\left[\sum_{k=0}^{+\infty} (\Delta_p^k)^2\right] \leq \frac{\mathbb{E}\left(\Phi_0\right)}{\eta} = \frac{\Phi_0}{\eta}, \tag{6.49}$$

That is, (6.36) holds. Then, making use of the following random function

$$\Phi_k^T := \frac{\nu}{\varepsilon}(f(X_{\text{feas}}^{k\vee T}) - \kappa_{\min}^f) + \frac{\nu}{m\varepsilon} h(X_{\text{inf}}^k) + (1-\nu)(\Delta_p^k)^2, \quad \text{for all } k \in \mathbb{N}, \tag{6.50}$$

where $k \vee T := \max\{k, T\}$, Part 2 aims to show that for the same previous constant $\eta > 0$, then conditioned on the almost sure event $\{T < +\infty\}$, the following holds for all $k \in \mathbb{N}$

$$\mathbb{E}\left(\Phi_{k+1}^T - \Phi_k^T | \mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2. \tag{6.51}$$

Indeed, assume that (6.51) holds. Since $\Phi_k^T > 0$ for all $k \geq 0$, then summing (6.51) over $k \in \mathbb{N}$ and taking expectations on both sides, yield

$$\begin{aligned}\mathbb{E}\left[\sum_{k=0}^{+\infty}(\Delta_p^k)^2\right] \leq \frac{\mathbb{E}\left(\Phi_0^T\right)}{\eta} = \frac{1}{\eta}\left[\frac{\nu}{\varepsilon}\left(\mathbb{E}\left[f(X_{\text{feas}}^T)\right] - \kappa_{\min}^f\right) + \frac{\nu}{m\varepsilon}h(x_{\text{inf}}^0) + (1-\nu)(\delta_p^0)^2\right] \\ \leq \frac{1}{\eta}\left[\frac{\nu}{\varepsilon}\left(\kappa_{\max}^f - \kappa_{\min}^f\right) + \frac{\nu}{m\varepsilon}h(x_{\text{inf}}^0) + (1-\nu)(\delta_p^0)^2\right] =: \mu,\end{aligned} \tag{6.52}$$

where the last inequality in (6.52) follows from the inequality $f(X_{\text{feas}}^k) \leq \kappa_{\max}^f$ for all $k \geq 0$, due to Proposition 8, and the fact that T is finite almost surely.

The remainder of the proof is devoted to showing that (6.47) and (6.51) 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 (6.48) will be shown to satisfy (6.47) with $\eta = \frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2)$, no matter the change led in the objective function f by the ε-infeasible iterates encountered by Algorithm 7. Moreover, since T is infinite almost surely, then no iteration of Algorithm 7 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 is shown to decrease and the smallest decrease is shown to happen on unsuccessful iterations, thus yielding the following conclusion

$$\mathbb{E}\left[\mathbb{1}_{I_k}(\Phi_{k+1} - \Phi_k)|\mathcal{F}_{k-1}^{C\cdot F}\right] \leq -\alpha(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{6.53}$$

(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 (6.6) as follows

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}(h(X_{\text{inf}}^{k+1}) - h(X_{\text{inf}}^k)) \leq -\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}\nu(\gamma-2)(\Delta_p^k)^2 \tag{6.54}$$

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

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)[(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] \leq \mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \tag{6.55}$$

Then, by choosing ν according to (6.34), the right-hand side term of (6.54) dominates that

of (6.55). Specifically, the following holds

$$-\nu(\gamma-2)(\Delta_p^k)^2+(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2\leq-\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{6.56}$$

Then combining (6.54), (6.55) and (6.56) leads to

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1}-\Phi_k)\leq-\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{D}_h}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{6.57}$$

(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 (6.6). Moreover, Δ_p^k is updated as at h-Dominating iterations. Thus, the change in Φ_k follows from (6.57) by replacing $\mathbb{1}_{\mathcal{D}_h}$ by $\mathbb{1}_{\mathcal{I}}$. Specifically,

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1}-\Phi_k)\leq-\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{I}}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{6.58}$$

(iii) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}}=1$). There is a change of zero in h function values while the frame size parameter is decreased. Consequently,

$$\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1}-\Phi_k)=-\mathbb{1}_{I_k}\mathbb{1}_{\mathcal{U}}(1-\nu)(1-\tau^2)(\Delta_p^k)^2 \tag{6.59}$$

Then, the choice of ν according to (6.34) and the fact that $1-\tau^2<\tau^{-2}-1$ ensures that unsuccessful iterations, more precisely (6.59), provide the worst case decrease when compared to (6.57) and (6.58). Specifically, the following holds

$$-\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2\leq-(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{6.60}$$

Thus, it follows from (6.57), (6.58), (6.59) and (6.60) that the change in Φ_k is bounded as follows

$$\mathbb{1}_{I_k}(\Phi_{k+1}-\Phi_k)=\mathbb{1}_{I_k}(\mathbb{1}_{\mathcal{D}_h}+\mathbb{1}_{\mathcal{I}}+\mathbb{1}_{\mathcal{U}})(\Phi_{k+1}-\Phi_k)\leq-\mathbb{1}_{I_k}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{6.61}$$

Since Assumption 12 holds, then taking conditional expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of the inequality in (6.61) leads to (6.53).

Case 2 (Bad bounds, $\mathbb{1}_{\bar{I}_k}=1$). Since the bounds are bad, Algorithm 7 can accept an iterate which leads to an increase in h and Δ_p^k, and hence in Φ_k. Such an increase in Φ_k is controlled making use of (6.15). Then, the probability of outcome (Part 1, Case 2) is adjusted to be sufficiently small so that Φ_k can be reduced sufficiently in expectation. More precisely, the following will be proved

$$\mathbb{E}\left[\mathbb{1}_{\bar{I}_k}(\Phi_{k+1}-\Phi_k)|\mathcal{F}_{k-1}^{C\cdot F}\right]\leq 2\nu(1-\alpha)^{1/2}(\Delta_p^k)^2. \tag{6.62}$$

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

$$\begin{aligned}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}(h(X_{\text{inf}}^{k+1}) - h(X_{\text{inf}}^{k})) & \\ \leq\ & \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}\left[(H_s^k - H_0^k) + \left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right] \\ \leq\ & \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\nu\left[-\gamma(\Delta_p^k)^2 + \frac{1}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right)\right] \quad (6.63)\end{aligned}$$

where (6.63) follows from $H_s^k - H_0^k \leq -\gamma m\varepsilon(\Delta_p^k)^2$ which is satisfied for every h-Dominating iteration. Moreover, the change in Δ_p^k can be obtained simply by replacing in (6.55) $\mathbb{1}_{I_k}$ by $\mathbb{1}_{\bar{I}_k}$ as follows

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)[(\Delta_p^{k+1})^2 - (\Delta_p^k)^2] \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \quad (6.64)$$

Since choosing ν according to (6.34) ensures that $-\nu\gamma(\Delta_p^k)^2 + (1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2 \leq 0$, then combining (6.63) and (6.64), yields

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right). \quad (6.65)$$

(ii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}} = 1$). Δ_p^k is updated as at h-Dominating iterations and because of bad bounds, the increase in h is bounded following (6.63). 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 (6.65) as follows

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right). \quad (6.66)$$

(iii) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}} = 1$). The change in h is zero and Δ_p^k is decreased. Thus, the change in Φ_k follows from (6.59) by replacing $\mathbb{1}_{I_k}$ by $\mathbb{1}_{\bar{I}_k}$ and is trivially bounded as follows

$$\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1} - \Phi_k) \leq \mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right). \quad (6.67)$$

Finally, it follows from (6.65), (6.66), (6.67) and the inequality $\mathbb{1}_{\bar{I}_k} \leq 1$, that

$$\mathbb{1}_{\bar{I}_k}(\Phi_{k+1} - \Phi_k) \leq \frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1}) - H_s^k\right| + \left|h(X_{\text{inf}}^{k}) - H_0^k\right|\right), \quad (6.68)$$

Then, taking conditional expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (6.68) and using the inequalities (6.15) of Lemma 9, lead to (6.62).

Now, combining (6.53) and (6.62) yields,

$$\begin{aligned}\mathbb{E}\left(\Phi_{k+1}-\Phi_k|\mathcal{F}_{k-1}^{C\cdot F}\right) &= \mathbb{E}\left[(\mathbb{1}_{I_k}+\mathbb{1}_{\bar{I}_k})(\Phi_{k+1}-\Phi_k)|\mathcal{F}_{k-1}^{C\cdot F}\right]\\ &\leq \left[-\alpha(1-\nu)(1-\tau^2)+2\nu(1-\alpha)^{1/2}\right](\Delta_p^k)^2. \end{aligned} \tag{6.69}$$

Then, choosing α according to (6.35) implies that $\alpha \geq \dfrac{4\nu(1-\alpha)^{1/2}}{(1-\nu)(1-\tau^2)}$, which ensures

$$-\alpha(1-\nu)(1-\tau^2)+2\nu(1-\alpha)^{1/2} \leq -\frac{1}{2}\alpha(1-\nu)(1-\tau^2) \leq -\frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2). \tag{6.70}$$

Thus, (6.47) follows from (6.69) and (6.70) 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 (6.51) 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 increments 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 as in Part 1 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,

$$\mathbb{1}_{\{T>k\}}\mathbb{E}\left(\Phi_{k+1}^T-\Phi_k^T|\mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2\mathbb{1}_{\{T>k\}}. \tag{6.71}$$

The remainder of the proof is devoted to showing that the following holds

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left(\Phi_{k+1}^T-\Phi_k^T|\mathcal{F}_{k-1}^{C\cdot F}\right) \leq -\eta(\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}, \tag{6.72}$$

since combining (6.71) and (6.72) leads to (6.51), which is the remaining 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 small-

est decrease in Φ_k^T is shown to happen on unsuccessful iterations, thus implying that

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\mathbb{1}_{I_k}\mathbb{1}_{J_k}(\Phi_{k+1}^T-\Phi_k^T)|\mathcal{F}_{k-1}^{C\cdot F}\right]\leq -\alpha\beta(1-\nu)(1-\tau^2)(\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}. \tag{6.73}$$

(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 (6.8) as follows

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}(f(X_{\text{feas}}^{(k+1)\vee T})-f(X_{\text{feas}}^{k\vee T}))&\\ \leq\ -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}\nu(\gamma-2)(\Delta_p^k)^2.&\end{aligned} \tag{6.74}$$

Since the ε-infeasible iterate is not updated, then there is a change of zero in h. The frame size parameter is updated according to $\Delta_p^{k+1}=\min\{\tau^{-1}\Delta_p^k,\delta_{\max}\}$, thus implying that

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)[(\Delta_p^{k+1})^2-(\Delta_p^k)^2]\leq\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \tag{6.75}$$

Then, choosing ν according to (6.34) ensures that (6.56) holds, which implies that the right-hand side term of (6.74) dominates that of (6.75), thus leading to the inequality below

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}(\Phi_{k+1}^T-\Phi_k^T)\leq-\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_f}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{6.76}$$

(ii) The iteration is h-Dominating ($\mathbb{1}_{\mathcal{D}_h}=1$). There is a change of zero in f since X_{feas}^k is not updated. Thus, the bound on the change in Φ_k^T follows from multiplying both sides of (6.57) by $\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{J_k}$, and replacing Φ_k by Φ_k^T as follows

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1}^T-\Phi_k^T)\leq-\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{D}_h}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{6.77}$$

(iii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}}=1$). Again, there is a change of zero in f. Thus, the bound on the change in Φ_k^T easily follows from multiplying both sides of (6.58) by $\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{J_k}$, and replacing Φ_k by Φ_k^T as follows

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1}^T-\Phi_k^T)\leq-\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{I}}\frac{1}{2}\nu(\gamma-2)(\Delta_p^k)^2. \tag{6.78}$$

(iv) The iteration is Unsuccessful ($\mathbb{1}_{\mathcal{U}}=1$). There is a change of zero in f and in h since no iterate is updated, while Δ_p^k is decreased. Consequently, the bound on the change in Φ_k^T follows from multiplying both sides of (6.59) by $\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{J_k}$, and replacing Φ_k by Φ_k^T as follows

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1}^T-\Phi_k^T)=-\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}\mathbb{1}_{\mathcal{U}}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{6.79}$$

Then combining (6.76), (6.77), (6.78), (6.79) and using (6.60), yields

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}(\Phi_{k+1}^T-\Phi_k^T)\leq -\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{J_k}(1-\nu)(1-\tau^2)(\Delta_p^k)^2. \tag{6.80}$$

Now, notice that under Assumption 12, simple calculations lead to $\mathbb{E}\left(\mathbb{1}_{I_k}\mathbb{1}_{J_k}|\mathcal{F}_{k-1}^{C\cdot F}\right)\geq \alpha\beta$. Then, taking expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (6.80) and using the $\mathcal{F}_{k-1}^{C\cdot F}$-measurability of the random variables $\mathbb{1}_{\{T\leq k\}}$ and Δ_p^k, lead to (6.73).

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 occurs since good bounds might not provide enough decrease to cancel the increase which occurs in f whenever Algorithm 7 wrongly accepts an iterate because of bad estimates. Specifically, the f-Dominating case dominates the worst-case increase in the change of Φ_k^T, thus leading to

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}^T-\Phi_k^T)|\mathcal{F}_{k-1}^{C\cdot F}\right]\leq 2\nu(1-\beta)^{1/2}(\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}. \tag{6.81}$$

(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 as follows

$$\begin{aligned}
&\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}(f(X_{\text{feas}}^{(k+1)\vee T})-f(X_{\text{feas}}^{k\vee T}))\\
&\quad\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}\left[(F_s^k-F_0^k)+\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right]\\
&\quad\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\nu\left[-\gamma(\Delta_p^k)^2+\frac{1}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right)\right]
\end{aligned} \tag{6.82}$$

where the last inequality in (6.82) 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 change in h is zero since X_{inf}^k is not updated, that in Δ_p^k follows (6.75) by replacing $\mathbb{1}_{J_k}$ by $\mathbb{1}_{\bar{J}_k}$ as follows

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)[(\Delta_p^{k+1})^2-(\Delta_p^k)^2]\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2. \tag{6.83}$$

Then, (6.82), (6.83) and the inequality $-\nu\gamma(\Delta_p^k)^2+(1-\nu)(\tau^{-2}-1)(\Delta_p^k)^2\leq 0$ due to (6.34) yield

$$\begin{aligned}
&\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}(\Phi_{k+1}^T-\Phi_k^T)\\
&\qquad\leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right).
\end{aligned} \tag{6.84}$$

(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 (6.77) is trivially bounded as follows

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1}^T-\Phi_k^T)&\\ \leq\ \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right)&\end{aligned} \tag{6.85}$$

(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 (6.78) is trivially bounded as follows

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1}^T-\Phi_k^T)&\\ \leq\ \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{I}}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right).&\end{aligned} \tag{6.86}$$

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

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1}^T-\Phi_k^T)&\\ \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\mathbb{1}_{\mathcal{U}}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right).&\end{aligned} \tag{6.87}$$

Then, combining (6.84), (6.85), (6.86) and $\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}\leq 1$, yields

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}(\Phi_{k+1}^T-\Phi_k^T)&\\ \leq\ \mathbb{1}_{\{T\leq k\}}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right).&\end{aligned} \tag{6.88}$$

Since Assumption 12 holds, it follows from the conditional Cauchy-Schwarz inequality [28] that

$$\begin{aligned}\mathbb{E}\left(\left|f(X_{\text{feas}}^k)-F_0^k\right|\,|\mathcal{F}_{k-1}^{C\cdot F}\right)\ &\leq\ \mathbb{E}\left(1|\mathcal{F}_{k-1}^{C\cdot F}\right)^{1/2}\left[\mathbb{E}\left(\left|f(X_{\text{feas}}^k)-F_0^k\right|^2|\mathcal{F}_{k-1}^{C\cdot F}\right)\right]^{1/2}\\ &\leq\ \varepsilon(1-\beta)^{1/2}(\Delta_p^k)^2,\end{aligned} \tag{6.89}$$

where (6.89) follows from (6.12) and the fact that $\mathbb{E}\left(1|\mathcal{F}_{k-1}^{C\cdot F}\right)=1$. Similarly, the following holds

$$\mathbb{E}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|\,|\mathcal{F}_{k-1}^{C\cdot F}\right)\leq\varepsilon(1-\beta)^{1/2}(\Delta_p^k)^2. \tag{6.90}$$

Thus, taking expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (6.88) and then using (6.89), (6.90) and the $\mathcal{F}_{k-1}^{C\cdot F}$-measurability of the random variables $\mathbb{1}_{\{T\leq k\}}$ and Δ_p^k, lead to (6.81).

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 7 wrongly accepts an iterate because of bad bounds. The following will be shown

$$\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\mathbb{1}_{\bar{I}_k}(\Phi_{k+1}^T-\Phi_k^T)|\mathcal{F}_{k-1}^{C\cdot F}\right]\leq 2\nu\left[(1-\alpha)^{1/2}+(1-\beta)^{1/2}\right](\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}. \tag{6.91}$$

(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 aforementioned increase in f. Since the change in h is zero, then it is easy to notice that the bound on the change in Φ_k^T can be derived from (6.84) by replacing $\mathbb{1}_{I_k}\mathbb{1}_{\bar{J}_k}$ by $\mathbb{1}_{\bar{I}_k}$ as follows

$$\begin{aligned}&\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_f}(\Phi_{k+1}^T-\Phi_k^T)\\&\qquad\leq\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_f}\frac{\nu}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right).\end{aligned} \tag{6.92}$$

(ii) The iteration is h-Dominating ($\mathbb{1}_{\mathcal{D}_h}=1$). Since the change in f is zero, the bound on the change in Φ_k^T is obtained by multiplying both sides of (6.65) by $\mathbb{1}_{\{T\leq k\}}$ and replacing Φ_k by Φ_k^T

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}(\Phi_{k+1}-\Phi_k)\leq\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{D}_h}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right). \tag{6.93}$$

(iii) The iteration is Improving ($\mathbb{1}_{\mathcal{I}}=1$). The frame size parameter is updated as at h-Dominating iterations and the change in f is zero. Thus, the bound on the change in Φ_k^T follows from (6.93) by replacing $\mathbb{1}_{\mathcal{D}_h}$ by $\mathbb{1}_{\mathcal{I}}$ as follows

$$\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}(\Phi_{k+1}-\Phi_k)\leq\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{I}}\frac{\nu}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right). \tag{6.94}$$

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

$$\begin{aligned}&\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}(\Phi_{k+1}^T-\Phi_k^T)\\&\qquad\leq\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\mathbb{1}_{\mathcal{U}}\nu\left[\frac{1}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right)\right.\\&\qquad\left.+\frac{1}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right)\right]\end{aligned} \tag{6.95}$$

Since (6.95) dominates (6.92), (6.93) and (6.94), then combining all four cases lead to

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}(\Phi_{k+1}^T-\Phi_k^T)&\\ \leq \mathbb{1}_{\{T\leq k\}}\mathbb{1}_{\bar{I}_k}\nu\Big[&\frac{1}{\varepsilon}\left(\left|f(X_{\text{feas}}^{k+1})-F_s^k\right|+\left|f(X_{\text{feas}}^k)-F_0^k\right|\right)\\ +&\frac{1}{m\varepsilon}\left(\left|h(X_{\text{inf}}^{k+1})-H_s^k\right|+\left|h(X_{\text{inf}}^k)-H_0^k\right|\right)\Big]\end{aligned} \tag{6.96}$$

Now, taking expectations with respect to $\mathcal{F}_{k-1}^{C\cdot F}$ on both sides of (6.96) and using (6.15), (6.89) and (6.90) lead to (6.91). Then, by combining the main results of Case 1, Case 2 and Case 3 of Part 2, specifically (6.73), (6.81) and (6.91), the following holds

$$\begin{aligned}\mathbb{1}_{\{T\leq k\}}\mathbb{E}\left[\Phi_{k+1}^T-\Phi_k^T|\mathcal{F}_{k-1}^{C\cdot F}\right]\leq\Big[&-\alpha\beta(1-\nu)(1-\tau^2)+2\nu(1-\alpha)^{1/2}\\ &+4\nu(1-\beta)^{1/2}\Big](\Delta_p^k)^2\mathbb{1}_{\{T\leq k\}}.\end{aligned} \tag{6.97}$$

Finally, choosing α and β according to (6.35) ensures that

$$-\alpha\beta(1-\nu)(1-\tau^2)+2\nu(1-\alpha)^{1/2}+4\nu(1-\beta)^{1/2}\leq-\frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2), \tag{6.98}$$

and (6.72) obviously follows from (6.97) and (6.98) with the same constant $\eta=\frac{1}{2}\alpha\beta(1-\nu)(1-\tau^2)$ as Part 1, which achieves the proof. □

Proof of Corollary 3

Proof. Only (6.37) 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 12*(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

$$\mathbb{E}\left(\left|H_0^k-h(X^k)\right|\right)\leq m\varepsilon(1-\alpha)^{1/2}\mathbb{E}\left[(\Delta_p^k)^2\right]. \tag{6.99}$$

By summing each side of (6.99) over k from 0 to N, and observing that

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

then, it follows from the monotone convergence theorem [50] 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 the constant of (6.52). This means that $\sum_{k=0}^{+\infty}\left|H_0^k - h(X^k)\right| < +\infty$ almost surely, which implies the first result of (6.37). The proof for $\left|F_0^k - f(X^k)\right|$ is similar by observing that (see (6.89))

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

□

Proof of Lemma 10

Proof. The proof uses ideas derived in [15, 35]. 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

$$\Psi_k^h \geq \mathcal{E}', \quad \text{for all } k \in \mathbb{N}. \tag{6.100}$$

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 (6.100) holds. Let $\hat{z}$ be the same parameter of Algorithm 7 satisfying $\delta_p^k \leq \tau^{-\hat{z}}$ for all $k \geq 0$. Since $\delta_p^k \to 0$ because of the conditioning on E_1, there exists $k_0 \in \mathbb{N}$ such that

$$\delta_p^k < \lambda := \min\left\{\frac{\epsilon'}{m\varepsilon(\gamma+2)}, \tau^{1-\hat{z}}\right\}, \quad \text{for all } k \geq k_0. \tag{6.101}$$

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. Let first show that $\{R_k\}_{k\in\mathbb{N}}$ is 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 of

at least $\alpha\beta > 1/2$. Then, it follows from the definition of the event I_k (see Definition 18) that

$$h(x^k_{\text{inf}}) \leq u^k_0(x^k_{\text{inf}}) \leq \sum_{j=1}^{m} \max\left\{c^k_{j,0}(x^k_{\text{inf}}), 0\right\} + m\varepsilon(\delta^k_p)^2 = h^k_0(x^k_{\text{inf}}) + m\varepsilon(\delta^k_p)^2, \quad (6.102)$$

$$\text{and} \quad h(x^k_{\text{inf}} + s^k) \geq \ell^k_s(x^k_{\text{inf}} + s^k) \geq h^k_s(x^k_{\text{inf}} + s^k) - m\varepsilon(\delta^k_p)^2. \quad (6.103)$$

$$\begin{aligned} \text{Hence,} \quad h^k_s(x^k_{\text{inf}} + s^k) - h^k_0(x^k_{\text{inf}}) &= [h(x^k_{\text{inf}} + s^k) - h(x^k_{\text{inf}})] + [h(x^k_{\text{inf}}) - h^k_0(x^k_{\text{inf}})] \\ &\quad + [h^k_s(x^k_{\text{inf}} + s^k) - h(x^k_{\text{inf}} + s^k)] \\ &\leq 2m\varepsilon(\delta^k_p)^2 - \epsilon'\delta^k_p \leq 2m\varepsilon(\delta^k_p)^2 - m\varepsilon(\gamma+2)(\delta^k_p)^2 = -\gamma m\varepsilon(\delta^k_p)^2 \end{aligned} \quad (6.104)$$

where the first inequality in (6.104) follows from (6.100), (6.102) and (6.103) while the last one follows from (6.101). Consequently, the iteration k of Algorithm 7 can not be unsuccessful. Thus, the frame size parameter is updated according to $\delta^{k+1}_p = \tau^{-1}\delta^k_p$ since $\delta^k_p < \tau^{1-\hat{z}}$. Hence, $r_{k+1} = r_k + 1$.

Let $\mathcal{F}^{I\cdot J}_{k-1} = \sigma(I_0, I_1, \ldots, I_{k-1}) \cap \sigma(J_0, J_1, \ldots, J_{k-1})$. 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 $\delta^{k+1}_p \geq \tau\delta^k_p$ always holds, thus implying that $r_{k+1} \geq r_k - 1$. Hence,

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

Thus, $\mathbb{E}\left(R_{k+1} - R_k|\mathcal{F}^{I\cdot J}_{k-1}\right) \geq 2\alpha\beta - 1 > 0$, 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 [35].

Now, let construct 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 (6.14) by

$$W_k = \sum_{i=0}^{k} (2 \cdot \mathbb{1}_{I_i}\mathbb{1}_{J_i} - 1), \quad (6.105)$$

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 12, it is easy to notice that $\{W_k\}$ is a $\mathcal{F}^{I\cdot J}_{k-1}$-submartingale (see also [35] for the same result), thus leading

to the conclusion that $\left\{\limsup_{k\to+\infty} W_k = +\infty\right\}$ 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 has to be 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. Consequently, the assumption that $\Psi_k^h \geq \mathcal{E}'$ holds for all $k \in \mathbb{N}$ with a positive probability is false, which implies that (6.38) holds. □

Proof of Theorem 15

Proof. The theorem is proved using ideas derived in [13, 15]. 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 2 and (6.38) 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 countable intersection of almost sure events. Then $\lim_{K'(\omega)} \Delta_p^k(\omega) = 0$. It follows from the compactness hypothesis of Assumption 11 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, then 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

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

where the fact that $t_k \to 0$ follows from Definition 15, specifically the inequality $\delta_m^k \left\|d^k\right\|_\infty \leq \delta_p^k b$. Since h is λ^h–locally Lipschitz, then

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

which shows that Lemma 11 applies for 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 15 (the fact that $\delta_p^k \left\| d^k \right\|_\infty \geq d_{\min} > 0$), yields

$$\lim_{k\in\mathcal{L}} \left(\frac{-\psi_k^h}{\delta_p^k \|d^k\|_\infty} \right) = \lim_{k\in\mathcal{L}} \frac{h(x_{\text{inf}}^k + \delta_m^k d^k) - h(x_{\text{inf}}^k)}{t_k} \geq -d_{\min}^{-1} \lim_{k\in\mathcal{L}} \psi_k^h \geq 0. \tag{6.106}$$

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

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

where the last inequality follows from (6.106). Now, notice that it has been showed that every outcome ω arbitrarily chosen in $E_1 \cap E_2$, belongs to the event

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

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

Proof of Lemma 12

Proof. The proof is almost identical to those of Lemma 10 and a similar result in [15]. 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 10. 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

$$\Psi_k^{f,T} \geq \mathcal{E}'', \quad \text{for all } k \geq 0. \tag{6.107}$$

Let $\{x_{\text{feas}}^{k\vee t}\}_{k\in\mathbb{N}}$, $\{s^k\}_{k\in\mathbb{N}}$, $\{\delta_p^k\}_{k\in\mathbb{N}}$ and $\epsilon'' > 0$ be realizations of $\{X_{\text{feas}}^{k\vee T}\}_{k\in\mathbb{N}}$, $\{S^k\}_{k\in\mathbb{N}}$, $\{\Delta_p^k\}_{k\in\mathbb{N}}$ and $\mathcal{E}''$, respectively for which (6.107) holds. Let $\bar{k}_0 \in \mathbb{N}^*$ be such that

$$\delta_p^k < \lambda := \min \left\{ \frac{\epsilon''}{\varepsilon(\gamma+2)}, \tau^{1-\hat{z}} \right\} \quad \text{for all } k \geq \bar{k}_0. \tag{6.108}$$

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 can not be unsuccessful, thus leading to the fact that $\{R_k\}$ is a submartingale.

It follows from (6.107) and (6.108) 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 7 can not be unsuccessful. □

Proof of Theorem 16

Proof. The proof results from Corollary 3 by observing that for all outcome ω in the almost sure event

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

$$\lim_{k \in K(\omega)} \left| H^k_0(X^{k \vee T}_{\text{feas}})(\omega) - h(X^{k \vee T}_{\text{feas}}(\omega)) \right| = \lim_{k \in K(\omega)} h(X^{k \vee T}_{\text{feas}}(\omega)) = h(\hat{X}_{\text{feas}}(\omega)) = 0,$$

where the penultimate equality follows from the continuity of h in $\mathcal{X}$. This means that

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

□

Proof of Theorem 17

Proof. First, notice that the fact that $\mathbb{P}\left(\hat{X}_{\text{feas}} \in \mathcal{D}\right) = 1$ follows from Theorem 16. Then the proof easily follows from that of Theorem 15, 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

$T_{\mathcal{D}}^{H}(\cdot)$, for ω fixed and arbitrarily chosen in the almost sure event $E_1 \cap E_5 \cap \{T < +\infty\}$, where

$$E_5 = \Big\{\omega \in \Omega : \exists K(\omega) \subseteq \mathbb{N} \text{ such that } \hat{X}_{\text{feas}}(\omega) = \lim_{k \in K(\omega)} X_{\text{feas}}^{k \vee T}(\omega),\ \hat{X}_{\text{feas}}(\omega) \in \mathcal{D}, \\ \lim_{k \in K(\omega)} \Psi_k^{f,T}(\omega) \leq 0 \ \text{ and } \lim_{k \in K(\omega)} H_0^k(X_{\text{feas}}^{k \vee T})(\omega) = 0\Big\}. \tag{6.109}$$

□

CHAPITRE 7 DISCUSSION GÉNÉRALE

7.1 Synthèse des travaux

Les contributions de cette thèse sont multiples. Elles étendent plusieurs résultats existants du contexte de la BBO déterministe au cadre stochastique via des analyses de convergence ou de complexité d'algorithmes originaux de recherche directe de type directionnel destinés aussi bien à l'optimisation stochastique sans contrainte qu'à celle sous contraintes bruitées de boîtes noires. Il est d'ailleurs très utile de souligner que les divers résultats principaux issus de ces analyses susmentionnées, sont au meilleur de nos connaissances les toutes premières en BBO stochastique.

Dans le Chapitre 4, l'algorithme StoMADS a été introduit pour l'optimisation stochastique de boîtes noires bruitées sans contraintes. Il n'utilise ni modèle, ni information de gradient pour trouver des directions de descente. En redéfinissant de nouveaux types d'itérations au moyen d'une condition de décroissance suffisante sur des estimés, StoMADS a su faire le lien entre cette décroissance et celle de l'objectif inconnu, démontrant ainsi son efficacité vis-à-vis de Robust-MADS. Notons en effet que même si ce dernier dispose d'une propriété de convergence *d'ordre zéro*, son mécanisme de mise à jour des itérés au moyen d'une version lisse de l'objectif inconnu ne montre pas clairement si une amélioration dans cette version lisse résulterait en une décroissance dans l'objectif inconnu. La convergence de StoMADS a été étudiée sous certaines hypothèses incluant le fait que les estimés sont précis avec une probabilité suffisamment grande mais fixe. Une preuve d'existence plus générale de soi-disant sous-suites raffinantes d'itérés de la méthode, qui ne sont pas nécessairement des optima locaux de treillis, a notamment favorisé le résultat principal d'optimalité obtenu. Il importe enfin de préciser que ces derniers résultats découlent de la convergence de toute la suite de paramètres de treillis vers zéro ; un résultat plus fort que celui de MADS au sens de la $\liminf$.

Les méthodes de recherche directe de type directionnel sont déjà dotées d'une analyse de taux de convergence dans le cas déterministe. Malheureusement, la question n'avait auparavant pas été abordée dans le cas stochastique. Ainsi, en introduisant au Chapitre 5 la large classe de méthodes SDDS généralisant le cadre algorithmique de StoMADS, une analyse de complexité a ensuite été effectuée, faisant usage d'un cadre existant de surmartingales. Cette analyse a révélé que les méthodes SDDS, quoique n'utilisant aucune information de gradient durant l'optimisation, ont toutefois un taux de convergence similaire à celui de toutes les méthodes d'optimisation non convexe du premier ordre. En particulier, le résultat principal obtenu est une extension au cadre stochastique de celui déjà existant dans le contexte déterministe pour les méthodes de recherche directe de type directionnel qui acceptent de nouveaux itérés en imposant une condition de décroissance suffisante.

Le Chapitre 6 étend au cadre stochastique l'algorithme MADS-PB d'optimisation de boîtes noires sous contraintes générales conçu pour le contexte déterministe, en introduisant StoMADS-PB. La fonction objectif qu'est supposé optimiser ce dernier ainsi que les contraintes n'étant accessibles que via une boîte noire corrompue d'un bruit aléatoire, StoMADS-PB utilise également des estimés sur lesquels sont imposées des conditions de décroissance suffisante. Il n'utilise donc ni modèle, ni information de gradient pour trouver des directions de descente. L'introduction de la notion de *probabilistically reliable bounds* déduite de celle d'existants soi-disant *probabilistically accurate estimates*, a notamment favorisé l'amélioration de la réalisabilité des itérés, gérée via la fonction de violation des contraintes. Une analyse de la méthode a permis d'étendre les résultats de convergence de MADS-PB au cadre stochastique, faisant usage du calcul non lisse de Clarke et de la théorie des processus stochastiques.

7.2 Limitations de la solution proposée

Toutes les méthodes proposées dans cette thèse sont souvent qualifiées d'*inexactes* en ce sens qu'elles utilisent des estimés plutôt que les vraies valeurs des fonctions qui sont impossibles d'accès. Ainsi, même si les résultats issus des analyses numériques ont la particularité d'être intéressants malgré la médiocre qualité des estimés utilisés, il découle toutefois des analyses théoriques qu'il est nécessaire d'avoir des estimés suffisamment précis pour espérer des résultats de très grande qualité. Plus précisément, l'obtention d'un estimé très précis nécessite un très grand nombre d'évaluations. Or la plupart des boîtes noires souvent rencontrées en ingénierie ayant la réputation d'être très coûteuses en terme de temps par évaluation, les méthodes proposées peuvent donc parfois s'avérer peu efficaces pour l'obtention de solutions d'une grande précision.

Par ailleurs, telles que présentés sans étape de SEARCH, les algorithmes proposés peuvent être peu performants devant des problèmes de très grande taille, ou simplement converger vers des minima locaux même pour des problèmes de taille relativement petite. Enfin, il faut noter que même dans les situations où la fonction objectif est différentiable, il n'est pas possible de détecter le nombre d'itérations au bout duquel le gradient de ce dernier est réduit en dessous d'un seuil donné. En d'autres termes, les méthodes ne disposent d'aucun critère d'arrêt sophistiqué comme c'est d'ailleurs le cas pour la majorité des algorithmes d'optimisation stochastique sans dérivées.

CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS

Cette thèse propose de nouvelles méthodes de recherche directe pour l'optimisation stochastique de boîtes noires. Elle a la particularité d'aborder pour une première fois en BBO au meilleur de nos connaissances, les questions d'analyse de taux de convergence de ces méthodes lorsqu'elles sont de type directionnel, et de s'intéresser à leur analyse de convergence en présence de contraintes générales également bruitées aléatoirement. En plus des résultats théoriques obtenus, des études numériques portant sur des problèmes tirés de la littérature ont révélé ces méthodes prometteuses pour la résolution de problèmes concrets.

Toutefois, à la lumière des limitations énumérées au Chapitre 7, les améliorations ci-après pourraient s'avérer intéressantes :

1. Proposer aux divers algorithmes une étape de SEARCH utilisant par exemple des modèles aléatoires précis (les soi-disant *probabilistically accurate models* existant dans la littérature) dont la probabilité de précision est contrôlable, dans le but d'éviter le maximum possible des convergences prématurés vers des minima locaux. Étendre au cadre stochastique la technique du NM-Search qui existe en contexte déterministe dans le logiciel NOMAD de BBO.
2. Utiliser une technique de décomposition parallèle de l'espace de recherche afin d'explorer de manière plus efficace l'espace des solutions, et ainsi éviter non seulement certaines évaluations inutiles de la boîte noire, mais aussi afin de pouvoir gérer des problèmes de très grande dimension.
3. Utiliser une base positive minimale pour la génération de points tests durant l'étape de POLL afin d'avoir moins de points à visiter.
4. Envisager l'utilisation d'autres stratégies d'ordonnancement plus sophistiquées que l'opportunisme durant l'étape de POLL.

RÉFÉRENCES

[1] M.A. Abramson and C. Audet. Convergence of Mesh Adaptive Direct Search to Second-Order Stationary Points. *SIAM Journal on Optimization*, 17(2) :606–619, 2006.

[2] M.A. Abramson, C. Audet, J.E. Dennis, Jr., and S. Le Digabel. OrthoMADS : A Deterministic MADS Instance with Orthogonal Directions. *SIAM Journal on Optimization*, 20(2) :948–966, 2009.

[3] S. Alarie, C. Audet, P.-Y. Bouchet, and S. Le Digabel. Optimization of noisy blackboxes with adaptive precision. Technical Report G-2019-84, Les cahiers du GERAD, 2019.

[4] S. Amaran, N.V. Sahinidis, B. Sharda, and S.J. Bury. Simulation optimization : a review of algorithms and applications. *4OR*, 12(4) :301–333, 2014.

[5] 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.

[6] E. Angün and J. Kleijnen. An asymptotic test of optimality conditions in multiresponse simulation optimization. *INFORMS Journal on Computing*, 24(1) :53–65, 2012.

[7] 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.

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

[9] 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, 2014.

[10] C. Audet, V. Béchard, and S. Le Digabel. Nonsmooth optimization through Mesh Adaptive Direct Search and Variable Neighborhood Search. *Journal of Global Optimization*, 41(2) :299–318, 2008.

[11] C. Audet and J.E. Dennis, Jr. Analysis of generalized pattern searches. *SIAM Journal on Optimization*, 13(3) :889–903, 2003.

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

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

[14] 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.

[15] C. Audet, K. J. Dzahini, M. Kokkolaras, and S. Le Digabel. StoMADS : Stochastic blackbox optimization using probabilistic estimates. Technical Report G-2019-30, Les cahiers du GERAD, 2019.

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

[17] C. Audet, A. Ianni, S. Le Digabel, and C. Tribes. Reducing the Number of Function Evaluations in Mesh Adaptive Direct Search Algorithms. *SIAM Journal on Optimization*, 24(2) :621–642, 2014.

[18] 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.

[19] C. Audet, S. Le Digabel, and C. Tribes. Dynamic scaling in the mesh adaptive direct search algorithm for blackbox optimization. *Optimization and Engineering*, 17(2) :333–358, 2016.

[20] 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.

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

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

[23] K. Balasubramanian and S. Ghadimi. Zeroth-order Nonconvex Stochastic Optimization : Handling Constraints, High-Dimensionality and Saddle-Points. *arXiv*, 2019.

[24] 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.

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

[26] A. S. Berahas, L. Cao, and K. Scheinberg. Global Convergence Rate Analysis of a Generic Line Search Algorithm with Noise. *arXiv*, 2019.

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

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

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

[30] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2) :337–375, 2018.

[31] 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.

[32] K.H. Chang, L. J. Hong, and H. Wan. Stochastic trust-region response-surface method (STRONG) – a new response-surface framework for simulation optimization. *INFORMS Journal on Computing*, 25(2) :230–243, 2013.

[33] L. Chen, H. Qiu, C. Jiang, X. Cai, and L. Gao. Ensemble of surrogates with hybrid method using global and local measures for engineering design. *Structural and Multidisciplinary Optimization*, 57(4) :1711–1729, 2018.

[34] R. Chen. *Stochastic derivative-free optimization of noisy functions*. PhD thesis, Lehigh University, 2015.

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

[36] R. Chen and S. Wild. Randomized derivative-free optimization of noisy convex functions. *arXiv*, 2015.

[37] 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.

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

[39] A.R. Conn, N.I.M. Gould, and Ph.L. Toint. *Trust region methods*. SIAM, 2000.

[40] A.R. Conn and S. Le Digabel. Use of quadratic models with mesh-adaptive direct search for constrained black box optimization. *Optimization Methods and Software*, 28(1) :139–158, 2013.

[41] A.R. Conn, K. Scheinberg, and L.N. Vicente. Geometry of sample sets in derivative free optimization : Polynomial regression and underdetermined interpolation. *IMA Journal of Numerical Analysis*, 28(4) :721–749, 2008.

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

[43] F. E. Curtis and K. Scheinberg. Adaptive Stochastic Optimization. *arXiv*, 2020.

[44] F.E. Curtis, K. Scheinberg, and R. Shi. A Stochastic Trust Region Algorithm Based on Careful Step Normalization. *arXiv*, 2017.

[45] G. Deng and M.C. Ferris. Adaptation of the UOBYQA Algorithm for Noisy Functions. In *Proceedings of the 38th Conference on Winter Simulation*, WSC '06, pages 312–319. Winter Simulation Conference, 2006.

[46] G. Deng and M.C. Ferris. Variable-number sample-path optimization. *Mathematical Programming*, 117(2) :1–2, 2009.

[47] 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.

[48] 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.

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

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

[51] K. J. Dzahini. Expected complexity analysis of stochastic direct-search. Technical Report G-2020-18, Les cahiers du GERAD, 2020.

[52] J. Franchi. *Processus aléatoires à temps discret : Cours, exercices et problèmes corrigés*. Ellipse, 2013.

[53] M.C. Fu. Gradient estimation. *Handbooks in operations research and management science*, 13 :575–616, 2006.

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

[55] N.I.M. Gould, D. Orban, and Ph.L. Toint. CUTEst : a Constrained and Unconstrained Testing Environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60(3) :545–557, 2015. Code available at `https://ccpforge.cse.rl.ac.uk/gf/project/cutest/wiki`.

[56] 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.

[57] S. Gratton, C.W. Royer, L.N. Vicente, and Z. Zhang. Direct search based on probabilistic descent. *SIAM Journal on Optimization*, 25(3) :1515–1541, 2015.

[58] K. Healy and L. W. Schruben. Retrospective simulation response optimization. Technical report, 1991.

[59] 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.

[60] R. Hooke and T.A. Jeeves. “Direct Search" Solution of Numerical and Statistical Problems. *Journal of the Association for Computing Machinery*, 8(2) :212–229, 1961.

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

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

[63] 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.

[64] 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.

[65] T.G. Kolda, R.M. Lewis, and V. Torczon. Optimization by direct search : New perspectives on some classical and modern methods. *SIAM Review*, 45(3) :385–482, 2003.

[66] A. Kulunchakov and J. Mairal. Estimate Sequences for Stochastic Composite Optimization : Variance Reduction, Acceleration, and Robustness to Noise. *arXiv*, 2019.

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

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

[69] S. Le Digabel. Algorithm 909 : NOMAD : Nonlinear Optimization with the MADS algorithm. *ACM Transactions on Mathematical Software*, 37(4) :44 :1–44 :15, 2011.

[70] 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.

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

[72] R.M. Lewis, V. Torczon, and M.W. Trosset. Direct search methods : Then and now. *Journal of Computational and Applied Mathematics*, 124(1–2) :191–207, 2000.

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

[74] K.I.M. McKinnon. Convergence of the Nelder-Mead simplex method to a nonstationary point. *SIAM Journal on Optimization*, 9(1) :148–158, 1998.

[75] 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.

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

[77] J.J. Moré and S.M. Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1) :172–191, 2009.

[78] J.J. Moré and S.M. Wild. Estimating Computational Noise. *SIAM Journal on Scientific Computing*, 33(3) :1292–1314, 2011.

[79] J.J. Moré and S.M. Wild. Estimating Derivatives of Noisy Simulations. *ACM Transactions on Mathematical Software*, 38(3) :19 :1–19 :21, 2012.

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

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

[82] S. Painchaud-Ouellet, C. Tribes J. Y. Trépanier, and D. Pelletier. Airfoil Shape Optimization Using a Nonuniform Rational B-Splines Parametrization Under Thickness Constraint. *AIAA Journal*, 44(10) :2170–2178, 2006.

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

[84] E. L. Plambeck, B. R. Fu, S. M. Robinson, and R. Suri. Sample-path optimization of convex stochastic performance functions. *Mathematical Programming*, 75(2) :137–176, 1996.

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

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

[87] 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.

[88] A. Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1) :169–186, 1991.

[89] 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.

[90] J. C. Spall. *Introduction to stochastic search and optimization : estimation, simulation, and control*. John Wiley & Sons, 2003.

[91] S. U. Stich, C.L. Muller, and B. Gartner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2) :1284–1309, 2013.

[92] 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.

[93] V. Torczon. On the convergence of the multidirectional search algorithm. *SIAM Journal on Optimization*, 1 :123–145, 1991.

[94] V. Torczon. On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, 7(1) :1–25, 1997.

[95] L. N. Vicente. Worst case complexity of direct search. *EURO Journal on Computational Optimization*, 1(1) :143–153, 2013.

[96] X. Wang and Y. Yuan. Stochastic Trust Region Methods with Trust Region Radius Depending on Probabilistic Models. *arXiv*, 2019.

[97] 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.

[98] W. Yang, J.A. Feinstein, and A.L. Marsden. Constrained optimization of an idealized y-shaped baffle for the fontan surgery at rest and exercise. *Computer Methods in Applied Mechanics and Engineering*, 199(33-36) :2135–2149, July 2010.

[99] 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.