Expected complexity analysis of stochastic direct-search

Kwassi Joseph Dzahini

Received: 15 June 2020 / Accepted: 1 November 2021 / Published online: 27 November 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

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 \( \beta \). 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 \( \epsilon \) is bounded in \( \mathcal{O}(\epsilon^{\frac{2}{p(2p-1)}} / (2\beta - 1)) \) 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 \( \epsilon \) 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

Kwassi Joseph Dzahini
kwassi-joseph.dzahini@polymtl.ca

GERAD and Département de Mathématiques et de Génie Industriel, Polytechnique Montréal, Succ. Centre-ville, C.P. 6079, Montreal, QC H3C 3A7, Canada
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 \[5, 11\]. Such methods include those based on simplices, like the classical Nelder–Mead method and its numerous variants, as well as those of directional type where an improvement in the objective function is guaranteed by moving along a direction defined by a better point \[17\].

This work focuses on the convergence rate analysis of stochastic variants of the broad class of directional direct-search methods analyzed in \[17\], using a supermartingale-based framework proposed in \[9\] and elements from \[4, 6, 10, 14, 16\]. 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:

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

where \(\Theta\) is a real-valued random variable following some unknown distribution, \(f_\Theta\) 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 \(\Theta\). Indeed, \(\Theta\) is considered as a data point for many machine learning problems \[16\].

Significant theoretical and algorithmic advances have been made in the field of stochastic DFO in recent years with the aim of solving Problem (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” \[4\].

Several recent works have proposed directional direct-search algorithms with convergence rates analysis. Vicente \[17\] 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 \(O\left(\epsilon^{\frac{-p}{\min(p-1,1)}}\right)\), with \(p > 1\).

Directional direct-search methods based on probabilistic descent, that incorporate a random gradient, was recently proposed and analyzed by Gratton et al. \[13\] with worst-case complexity, and global rates results. However, these latter works both assume that the objective function is deterministic, i.e., the function values are exactly computed.

Audet et al. \[4\] recently proposed StoMADS, a stochastic variant of the mesh adaptive direct-search (MADS) algorithm \[3\], with convergence analysis based on Clarke calculus and martingale theory. Alarie et al. \[1\] also proposed another variant of MADS to optimize noisy blackboxes corrupted with Gaussian noise, and proved
convergence results using statistical inference techniques. Nevertheless, no convergence rates analyses have been carried out for either method.

The main novelty of the present work is that unlike many prior research on convergence rate analysis of stochastic DFO methods (see for example [7, 9, 16] and references therein), especially those on stochastic trust-region [9] and line search [16] 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 [17], “it is not unreasonable” to expect that SDDS shares a similar worst case complexity bound of the latter methods in terms 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 [14, 17]. This latter remark is in fact the cornerstone of the analysis in the present manuscript. Moreover, unlike the deterministic framework [17], SDDS accepts new iterates by imposing a sufficient decrease condition on so called probabilistic estimates of the corresponding unavailable objective function values, whose accuracy is required to hold with a sufficiently large but fixed probability \( \beta > 1/2 \). However, even though \( \beta \) is not required to equal one, SDDS is shown to have desirable convergence properties. Specifically, as the main theoretical result of the present work, the expected number of iterations required by SDDS to drive the gradient of \( f \) below a threshold \( \epsilon \) 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 [9]. 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, let us mention that the analysis in the present manuscript is not limited to \( p = 2 \) as in several similar works, but rather extends to any \( p > 1 \). 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 2 introduces an outline of the proposed stochastic algorithm and requirements (including a variance condition) on so-called probabilistic estimates that guarantee convergence at an appropriate rate. Section 3 presents a general framework of a stochastic process that is required for the convergence rate analysis in Sect. 4. The latter section presents also a \( \lim \inf \)-type first-order convergence result for SDDS. Section 5 discusses why the variance assumption on the estimates cannot be dropped from the analysis in Sect. 4, followed by concluding remarks and suggestions for future work.

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.
2.1 The SDDS algorithm

SDDS uses an algorithmic framework similar to that of StoMADS [4] and the broad class of methods analyzed in [17], i.e., a framework that can describe the main features of generating set search (GSS) [14], pattern search and generalized pattern search (GPS) [2].

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 1 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 current solution, $\delta^k$ the step size and $\mathbb{D}^k$ is a positive spanning set [5, 11]. 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 1, since the objective function values $f(x)$ are numerically unavailable, $f_0^k$ and $f^k(s)$ denote respectively the estimates of $f(x^k)$ and $f(x^k + s)$ (with $s \in \mathcal{S}^k := \{\delta^k d : d \in \mathbb{D}^k\}$), constructed making use of evaluations of the noisy blackbox $f_\Theta$. In order for the information provided by $f_0^k$ and $f^k(s)$ to determine the iteration type, i.e., successful or unsuccessful, both estimates are required to be $\varepsilon_f$-accurate, with $\varepsilon_f > 0$, according to the following definition similar to those in [4, 9, 10, 16].

---

Algorithm 1: SDDS

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

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

3. **Success**
   1. If $f^k(s) - f_0^k \leq -\gamma \varepsilon_f (\delta^k)^p$ for some $s \in \mathcal{S}^k$,
   2. set $x^{k+1} \leftarrow x^k + s$, and $\delta^{k+1} \leftarrow \min\{\tau^{-1} \delta^k, \delta_{\text{max}}\}$.

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

5. **Termination**
   1. If no termination criterion is met,
   2. set $k \leftarrow k + 1$ and go to [1].
   3. Otherwise stop.
Definition 1 Let \( \rho : (0, +\infty) \rightarrow (0, +\infty) \) be a continuous and non-decreasing function satisfying \( \rho(t)/t \rightarrow 0 \) when \( t \searrow 0 \). \( f^k \) is called an \( \epsilon_f \)-accurate estimate of \( f(x^k) \) for a given \( \delta^k \) if

\[
|f^k - f(x^k)| \leq \epsilon_f \rho(\delta^k).
\]

Following the terminology in [14], the function \( \rho \) in Definition 1 represents the “forcing function”. In order to make the present analysis simpler, the following assumption is made.

Assumption 1 The forcing function \( \rho : (0, +\infty) \rightarrow (0, +\infty) \) is such that \( \rho(t) = ct^p \), where \( c > 0 \) and \( p > 1 \) are constants.

Sufficient information to determine the iteration type is provided below.

Proposition 1 Let \( f^k_0 \) and \( f^k(s) \) be \( \epsilon_f \)-accurate estimates of \( f(x^k) \) and \( f(x^k + s) \) respectively, and let \( \gamma > 2 \) be a fixed constant. Then the following hold:

\[
\text{if } f^k(s) - f^k_0 \leq -\gamma \epsilon_f \rho(\delta^k), \quad \text{then } f(x^k + s) - f(x^k) \leq -(\gamma - 2)\epsilon_f \rho(\delta^k) =: \epsilon_{sto}^k  
\]  

(2)

\[
\text{if } f^k(s) - f^k_0 > -\gamma \epsilon_f \rho(\delta^k), \quad \text{then } f(x^k + s) - f(x^k) > -(\gamma + 2)\epsilon_f \rho(\delta^k) =: \epsilon_{sto}^k  
\]  

(3)

Proof The proof straightforwardly follows from Definition 1 and the fact that for any \( s \in \mathcal{S}^k \)

\[
f(x^k + s) - f(x^k) = f(x^k + s) - f^k(s) + (f^k(s) - f^k_0) + f^k_0 - f(x^k).
\]

□

An important distinction between StoMADS and our work is that we do not distinguish between different types of unsuccessful iterations. In particular, Algorithm 1 defines an iteration to be unsuccessful if

\[
f^k(s) - f^k_0 > -\gamma \epsilon_f \rho(\delta^k) \quad \text{for all } s \in \mathcal{S}^k = \{ \delta^k d : d \in \mathbb{D}^k \},
\]  

(4)

while in StoMADS, an iteration is called unsuccessful if (4) holds with \( \rho(t) \) := \( t^2 \) and \( \delta^k \) replaced with the StoMADS so-called frame size parameter \( \delta_p^k \). In addition, there are two types of unsuccessful iterations in StoMADS: certain ones which occur whenever the inequality \( f^k(s) - f^k_0 \geq \gamma \epsilon_f \rho(\delta_p^k) \) holds for all \( s \in \mathcal{S}^k \), implying that \( f(x^k + s) - f(x^k) \geq (\gamma - 2)\epsilon_f \rho(\delta_p^k) > 0 \) (see [4], Proposition 1), and those called uncertain when \( f^k(s) - f^k_0 \in (-\gamma \epsilon_f \rho(\delta_p^k), \gamma \epsilon_f \rho(\delta_p^k)) \) for at least one \( s \in \mathcal{S}^k \) since they lead to \( -(\gamma + 2)\epsilon_f \rho(\delta_p^k) < f(x^k + s) - f(x^k) < (\gamma + 2)\epsilon_f \rho(\delta_p^k) \). By updating the frame size parameter according to \( \delta_p^{k+1} - \tau \delta_p^k \) on uncertain unsuccessful iterations (where

\& Springer
\( \tau \in (0, 1) \), and \( \delta^{k+1}_p = \tau^2 \delta^k_p \) whenever the unsuccessful iteration is \textit{certain}, the corresponding sequence \( \{ \delta^k \}_{k \in \mathbb{N}} \) was shown in \([4]\) to converge to zero. Note also that this kind of update is the only one that differentiates \textit{certain} iterations from those that are \textit{uncertain}. In the present work, the step size parameter \( \delta^k \) is updated on unsuccessful iterations according to \( \delta^{k+1} = \tau \delta^k \). As a consequence, \textit{certain} unsuccessful iterations will not be differentiated from \textit{uncertain} ones, i.e., any iteration where \( (4) \) holds will be called unsuccessful.

However, let us put an emphasis on the specific choice of \( \tau \) by means of the following additional remarks. In the general deterministic framework described in \([17]\), the amount of decrease in the objective function on successful iterations is such that \( f(x^k + s) - f(x^k) \leq -\rho(\delta^k) := u^{\text{det}} \) while unsuccessful iterations are characterized by \( f(x^k + s) - f(x^k) > -\rho(\delta^k) := \ell^{\text{det}} \). Thus, the equality \( \ell^{\text{det}} = u^{\text{det}} \) always holds, which is not the case in stochastic settings where \( \ell^{\text{sto}} < u^{\text{sto}} \). Moreover, since \( \delta^{k+1} < \delta^k \) when the iteration \( k \) is unsuccessful, then \( \ell^{\text{det}} > u^{\text{det}} \). Likewise, since \( \delta^{k+1} > \delta^k \) on successful iterations, then \( u^{\text{det}} < \ell^{\text{det}} \). Given that the equality \( \ell^{\text{sto}} = u^{\text{sto}} \) cannot hold in the present stochastic settings, \( \tau \) must be chosen in such a way that at least, both inequalities \( \ell^{\text{sto}} > u^{\text{sto}} \) and \( u^{\text{sto}} < \ell^{\text{sto}} \) hold respectively on unsuccessful and successful iterations, analogously to the deterministic framework. This means using \((2)\) and \((3)\) that \( \tau \) must be chosen according to

\[
\rho(\tau \delta^k) < \frac{\gamma - 2}{\gamma + 2} \rho(\delta^k) \quad \text{and} \quad \rho(\tau^{-1} \delta^k) > \frac{\gamma + 2}{\gamma - 2} \rho(\delta^k),
\]

which means that under Assumption 1, \( \tau \) must be chosen according to \( 0 < \tau^p < \frac{\gamma - 2}{\gamma + 2} \), for all \( k \in \mathbb{N} \), as specified in Algorithm 1.

### 2.2 Probabilistic estimates

Following the notation in \([8]\), all stochastic quantities in the present manuscript live on the same probability space \((\Omega, \mathcal{F}, \mathbb{P})\), where \( \Omega \) is a nonempty set referred to as the sample space, \( \mathcal{F} \) is a collection of events (subsets of \( \Omega \)) called a \( \sigma \)-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 \( \sigma \)-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 \( B \in \mathcal{B}(\mathbb{R}^n) \) \([8]\).

The estimates \( f^k(s) \) and \( f^k_0 \) constructed at iteration \( k \) of Algorithm 1 based on random information provided by the noisy objective \( f_0 \), can be considered as realizations of random estimates \( F^k(S) \) and \( F^k_0 \) respectively \([4]\). Thus, because of the randomness stemming from such random estimates whose behavior influences each iteration \( k \), Algorithm 1 results in a stochastic process \( \{ X^k, \Delta^k, F^k(S), F^k_0 \} \). In general, uppercase letters will be used to denote random variables while lowercase letters will be used for their realizations. For example, \( x^k = X^k(\omega) \) and \( \delta^k = \Delta^k(\omega) \) denote respectively realizations of the random variables \( X^k \) and \( \Delta^k \). Similarly, inspired by
the notations in [4, 9, 10, 16], $f^k_0 = F^k_0(\omega)$ and $f^k(s) = F^k(S)(\omega)$ where $F^k_0$ and $F^k(S)$ are respectively estimates of $f(X^k)$ and $f(X^k + S)$.

The goal of this work is to show that the stochastic process resulting from Algorithm 1 converges at an appropriate rate with probability one, provided that the sequence $\{(F^k_0, F^k(S))\}$ is sufficiently accurate with sufficiently high but fixed probability, conditioned on the past [9, 10, 16].

To formalize the notion of conditioning on the past, let $\mathcal{F}_{k-1}$ denote the $\sigma$-field generated by $F^0_0, F^1_0, F^1_1, F^1(S), \ldots, F^{k-1}_0$ and $F^{k-1}(S)$, and set $\mathcal{F}^F_k$ to equal $\sigma(x^0)$ for completeness [16]. As a result, $\{\mathcal{F}^F_k\}_{k \geq 1}$ is a filtration, i.e., an increasing sequence of $\sigma$-fields of $\mathcal{F}$. Moreover, one can notice that $\mathbb{E}(\Delta^k | \mathcal{F}^F_{k-1}) = \Delta^k$ and $\mathbb{E}(X^k | \mathcal{F}^F_{k-1}) = X^k$ for all $k \geq 0$, since the random variables $\Delta^k$ and $X^k$ are $\mathcal{F}^F_{k-1}$-measurable by construction in Algorithm 1.

The following definition similar to those in [4, 9, 10, 16] is used to measure closeness or sufficient accuracy.

**Definition 2** A sequence of random estimates $\{(F^k_0, F^k(S))\}$ is said to be $\beta$-probability $\epsilon_f$-accurate with respect to the corresponding sequence $\{X^k, \Delta^k\}$ if the events

$$J_k = \{F^k_0, F^k(S), \text{ are } \epsilon_f\text{-accurate estimates of } f(x^k) \text{ and } f(x^k + s), \text{ respectively}\}$$

satisfy the following submartingale-like condition

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

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” [4].

Global convergence properties of deterministic directional direct-search methods strongly rely on having the step size parameters approaching zero [17] 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 1 therefore relies on the assumption that accuracy in function estimates “improves in coordination with the perceived progress of the algorithm” [9]. 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, let us make the following key assumption similar to those in [4, 16].

**Assumption 2** For some fixed $\beta \in (0, 1)$, and $\epsilon_f > 0$, the following hold for the random quantities derived from Algorithm 1.

(i) The sequence $\{(F^k_0, F^k(S))\}$ of estimates is $\beta$-probability $\epsilon_f$-accurate.
(ii) The sequence \(\{ (F^k_0, F^k(S)) \}\) satisfies the following variance condition

\[
\mathbb{E}\left( \left| F^k_0 - f(X^k) \right|^2 \mid \mathcal{F}^F_{k-1} \right) \leq \varepsilon_f^2 (1 - \beta) \rho(\Delta^k)^2
\]

and

\[
\mathbb{E}\left( \left| F^k(S) - f(X^k + S) \right|^2 \mid \mathcal{F}^F_{k-1} \right) \leq \varepsilon_f^2 (1 - \beta) \rho(\Delta^k)^2
\]

Hereafter are provided further details on Assumption 2. First, (i) assumes that the estimates are good (increasingly better, as \(\Delta^k\) gets smaller) with some fixed sufficiently large probability. Thus the estimates may be overall biased, and can be arbitrarily inaccurate with small probability. This allows for outliers in estimating \(f(x)\). On the other hand, by means of (ii), the variance of the function estimates is adaptively controlled as it is also decreasing with \(\Delta^k\). Thus, the fact that the sequence of step sizes converges to zero (which will later result from Theorem 3), forces the variance to go to zero.

Before giving some details on the computation of estimates satisfying Assumption 2, let us specify that, since such estimates can easily be constructed using techniques proposed in [4, 9, 10, 16], then thorough details about their computations are not provided here again. Note however that if both \(\Theta_0\) and \(\Theta_s\) are independent random variables following the same distribution as \(\Theta\) defined in (1), and if \(\Theta_{0,i}\) and \(\Theta_{s,i}\), \(i = 1, 2, \ldots, p^k\) are independent random samples of \(\Theta_0\) and \(\Theta_s\) respectively, then the estimates

\[
F^k_0 = \frac{1}{p^k} \sum_{i=1}^{p^k} f_{\Theta_{0,i}}(x^k) \quad \text{and} \quad F^k(S) = \frac{1}{p^k} \sum_{i=1}^{p^k} f_{\Theta_{s,i}}(x^k + S)
\]

satisfy Assumption 2 provided that the sample size \(p^k\) satisfies

\[
p^k \geq \frac{V}{\varepsilon_f^2 (1 - \sqrt{\beta}) \rho(\Delta^k)^2},
\]

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

Below is stated a useful lemma similar to those in [4, 16], linking the probability of obtaining bad estimates to the variance assumption on function values.

**Lemma 1** Let Assumption 2 hold. Then for all \(k \geq 0\), the following hold for the random process \(\{X^k, F^k_0, F^k(S), \Delta^k\}\) generated by Algorithm 1

\[
\mathbb{E}\left( 1_{I_k} \left| F^k_0 - f(X^k) \right| \mathcal{F}^F_{k-1} \right) \leq \varepsilon_f (1 - \beta) \rho(\Delta^k)
\]

and

\[
\mathbb{E}\left( 1_{I_k} \left| F^k(S) - f(X^k + S) \right| \mathcal{F}^F_{k-1} \right) \leq \varepsilon_f (1 - \beta) \rho(\Delta^k)
\]

**Proof** The result is proved using ideas derived from [4, 16]. The proof follows straightforwardly from the conditional Cauchy–Schwarz inequality [8] as follows
\[
\mathbb{E}\left( I_{J_k} \left| F_k(S) - f(X^k + S) \right| \mathcal{F}_{k-1} \right) \\
\leq \left[ \mathbb{E}\left( I_{J_k} \left| \mathcal{F}_{k-1} \right) \right] \right)^{1/2} \left[ \mathbb{E}\left( \left| F_k(S) - f(X^k + S) \right|^2 \left| \mathcal{F}_{k-1} \right) \right] \right)^{1/2} \\
\leq (1 - \beta)^{1/2} \varepsilon_f (1 - \beta)^{1/2} [\rho(\Delta^k)],
\]
where the last inequality follows from (5) and the fact that \( \mathbb{E}\left( I_{J_k} \left| \mathcal{F}_{k-1} \right) \right] = \mathbb{P}(J_k \left| \mathcal{F}_{k-1} \right) \leq 1 - \beta \) thanks to Assumption 2-(i). The proof for \( F_0^k - f(X^k) \) is the same. \( \square \)

### 3 A renewal-reward martingale process

This section presents a general stochastic process and its associated stopping time \( T \) introduced in [9] 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 [9], that will be useful for the convergence rate analysis presented in Sect. 4. Specifically, by considering the stopping time consisting of the time required by SDDS to reach a desired accuracy, Sect. 4 will aim to show how the properties of this general stochastic process are satisfied for Algorithm 1. Let us specify that some results derived in analyzing this stochastic process in [9] are also used in [16] for the convergence rate analysis of a stochastic line search method.

**Definition 3** 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 \( \sigma \)-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 biased random walk process \( \{W_k\}_{k \in \mathbb{N}} \) such that \( W_0 = 1 \),

\[
\mathbb{P}(W_{k+1} = 1 \mid \mathcal{F}_k) = q \quad \text{and} \quad \mathbb{P}(W_{k+1} = -1 \mid \mathcal{F}_k) = 1 - q,
\]
where \( q \in (1/2, 1) \) and \( \mathcal{F}_k \) denotes the \( \sigma \)-field generated by \( \{\Phi_0, \Delta^0, W_0, (\Phi_1, \Delta^1, W_1), \ldots, (\Phi_k, \Delta^k, W_k)\} \). Let \( \{T_{\epsilon'}\}_{\epsilon' > 0} \) be a family of stopping times parameterized by \( \epsilon' > 0 \), with respect to \( \{\mathcal{F}_k\}_{k \in \mathbb{N}} \). The following assumptions are imposed on \( \{(\Phi_k, \Delta^k, W_k)\}_{k \in \mathbb{N}} \) and \( T_{\epsilon'} \) in [9] in order to derive a bound on \( \mathbb{E}(T_{\epsilon'}) \).

**Assumption 3** The following hold for the stochastic process \( \{(\Phi_k, \Delta^k, W_k)\}_{k \in \mathbb{N}} \):

(i) There exist constants \( \lambda \in (0, +\infty) \) and \( \delta_{\text{max}} = \delta^0 e^j_{\text{max}} \), for some integer \( j_{\text{max}} \in \mathbb{Z} \), such that \( \Delta^k \leq \delta_{\text{max}} \) for all \( k \in \mathbb{N} \).
(ii) There exists a constant $\delta_e = \delta^0 e^{\beta e}$, for some $j_e \in \mathbb{Z}, j_e \leq 0$, such that the following holds for all $k \in \mathbb{N}$,

$$\mathbb{I}_{\{T_e > k\}} \Delta^{k+1} \geq \mathbb{I}_{\{T_e > k\}} \min \left( \Delta^k e^{\beta W_{k+1}}, \delta_e \right),$$

where $W_{k+1}$ satisfies (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

$$\mathbb{E}(\Phi_{k+1} - \Phi_k | \mathcal{F}_k) \mathbb{I}_{\{T_e > k\}} \leq -\eta h(\Delta^k) \mathbb{I}_{\{T_e > k\}}.$$

As highlighted in [9, 16], Assumption 3 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 $T_e > k$ and moreover, the sequence $\{\Delta^k\}_{k \in \mathbb{N}}$ has a tendency to increase whenever it is below some fixed threshold $\delta_e$.

The following theorem providing a bound on $\mathbb{E}(T_e)$ is proved in [9] by observing that the upward drift in the random walk $\{W_k\}_{k \in \mathbb{N}}$ makes the event $\{\Delta_k \geq \delta_e\}$ occur sufficiently frequently on average [9, 16]. 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 $\mathbb{E}(T_e)$.

**Theorem 1** Let Assumption 3 hold. Then,

$$\mathbb{E}(T_e) \leq \frac{q}{2q - 1} \times \frac{\Phi_0}{\eta h(\delta_e)} + 1.$$

## 4 Convergence rate analysis

It follows from Sect. 3 that Theorem 1 holds for any stopping time $T_e$, defined with respect to the filtration $\{\mathcal{F}_k\}_{k \in \mathbb{N}}$, provided that Assumption 3 holds 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 3 can be constructed in order to bound the expected number of iterations required by Algorithm 1 to achieve $\|\nabla f(X^k)\| \leq \varepsilon$, for some arbitrary fixed $\varepsilon \in (0, 1)$, where $\| \cdot \|$ denotes the Euclidean norm of $\mathbb{R}^n$ as in the remainder of the manuscript.

### 4.1 Analysis of the stochastic process generated by SDDS

In order to show that Assumption 3 holds, let us impose the following standard assumption on the objective function $f$. 

 Springer
Assumption 4 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 [4] from \( p = 2 \) to any \( p > 1 \) provides a bound on the expected decrease in the random function

\[
\Phi_k := \frac{\nu}{c \varepsilon_f} (f(X^k) - f_{\min}) + (1 - \nu)(\Delta^k)^p,
\]

where the constants \( p \) and \( c \) pertain to the definition of the forcing function \( \rho \) given in Assumption 1.

Theorem 2 Let Assumptions 1, 2 and 4 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

\[
\frac{\nu}{1 - \nu} \geq \frac{2(\tau^{-p} - 1)}{\gamma - 2} \quad \text{and} \quad \frac{\beta}{1 - \beta} \geq \frac{\nu}{1 - \nu} \times \frac{4}{(1 - \tau^p)}.
\]

Then the expected decrease in the random function \( \Phi_k \) defined in (10) satisfies

\[
\mathbb{E}(\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}^k) \leq -\frac{1}{2} \beta(1 - \nu)(1 - \tau^p)(\Delta^k)^p.
\]

Proof The proof is almost identical to that in [4], using ideas derived in [10, 15, 16] and making use of properties of the random function \( \Phi_k \) defined in (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 \( \mathcal{E} \) by

\( \mathcal{E} := \{\text{The iteration is successful}\} \),

and let \( \mathcal{E}^c \) denote the complement of \( \mathcal{E} \).

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

\[
\mathbb{E}(1_{\mathcal{E}}(\Phi_{k+1} - \Phi_k) \mid \mathcal{F}_{k-1}^k) \leq -\beta(1 - \nu)(1 - \tau^p)(\Delta^k)^p.
\]

(i) Successful iteration (\( 1_{\mathcal{E}} = 1 \)). A decrease occurs in \( f \) according to (2) since estimates are good and the iteration is successful. Thus,

\[
1_{\mathcal{E}}1_{\mathcal{E}} \frac{\nu}{c \varepsilon_f} (f(X^{k+1}) - f(X^k)) \leq -1_{\mathcal{E}}1_{\mathcal{E}} \nu(\gamma - 2)(\Delta^k)^p.
\]

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

\[
1_{\mathcal{E}}1_{\mathcal{E}}(1 - \nu) [ (\Delta^{k+1})^p - (\Delta^k)^p ] \leq 1_{\mathcal{E}}1_{\mathcal{E}} (1 - \nu)(\tau^{-p} - 1)(\Delta^k)^p.
\]

Then, choosing \( \nu \) according to (11) ensures that the right-hand side term of (14) dominates that of (15), i.e.,
Thus, combining (14), (15) and (16) yields

\[ \mathbb{I}_{J_k}(\Phi_{k+1} - \Phi_k) \leq -\mathbb{I}_{J_k}(1 - \nu)(1 - \tau^p)(\Delta^k)^p. \]

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

\[ \mathbb{I}_{J_k}(\Phi_{k+1} - \Phi_k) = -\mathbb{I}_{J_k}(1 - \nu)(1 - \tau^p)(\Delta^k)^p. \]

Then, the inequality \(1 - \tau^p < \tau^{-p} - 1\) ensures that unsuccessful iterations, specifically (18), provide the worst case decrease when compared to (17), whenever \(\nu\) is chosen according to (11). Specifically, the following holds

\[ -\frac{1}{2} \nu(\gamma - 2)(\Delta^k)^p \leq - (1 - \nu)(1 - \tau^p)(\Delta^k)^p. \]

Thus, combining (17), (18), and (19), leads to the following bound on the change in \(\Phi_k\)

\[ \mathbb{I}_{J_k}(\Phi_{k+1} - \Phi_k) = 1 - \frac{1}{2} (\mathbb{I}_{\phi} + \mathbb{I}_{\bar{\phi}})(\Phi_{k+1} - \Phi_k) \leq - \mathbb{I}_{J_k}(1 - \nu)(1 - \tau^p)(\Delta^k)^p. \]

Since Assumption 2 holds, then taking conditional expectations with respect to \(\mathcal{F}_{k-1}\) in both sides of (20) leads to (13).

**Case 2 (Bad estimates, \(\mathbb{I}_{\bar{J}_k} = 1\)).** Since the estimates are bad, an iterate leading to an increase in \(f\) and \(\Delta^k\), and hence in \(\Phi_k\), can be accepted by Algorithm 1. Such an increase in \(\Phi_k\) is controlled by bounding the variance in function estimates, using (5). Then, in order to guarantee that \(\Phi_k\) is sufficiently reduced in expectation, the probability of outcome is adjusted to be sufficiently small. The overall goal is to show that

\[ \mathbb{E}\left( \mathbb{I}_{J_k}(\Phi_{k+1} - \Phi_k) \mid \mathcal{F}_{k-1} \right) \leq 2\nu(1 - \beta)(\Delta^k)^p. \]
where the last inequality in (22) follows from the fact that 
\( F^k(S) - F^k_0 \leq -\gamma c f_0 (\Delta^k)^p \) for successful iterations. Moreover, as in Case 1, 
\( \Delta^{k+1} = \min \{ \tau^{-1} \Delta^k, \delta_{\text{max}} \} \) since the iteration is successful. Thus,
\[
\mathbb{I}_{J_k} \mathbb{I}_{\phi} (1 - \nu) \left[ (\Delta^{k+1})^p - (\Delta^k)^p \right] \leq \mathbb{I}_{J_k} \mathbb{I}_{\phi} (1 - \nu) (\tau^{-p} - 1)(\Delta^k)^p.
\]  
(23)

Then, choosing \( \nu \) according to (11) yields
\[-\nu \gamma (\Delta^k)^p + (1 - \nu)(\tau^{-p} - 1)(\Delta^k)^p \leq 0.\]
(24)

Thus, combining (22), (23) and (24) leads to
\[
\mathbb{I}_{J_k} \mathbb{I}_{\phi} (\Phi_{k+1} - \Phi_k) \leq \mathbb{I}_{J_k} \mathbb{I}_{\phi} \frac{\nu}{c \epsilon f} \left( \left| f(X^{k+1}) - F^k(S) \right| + \left| F^k_0 - f(X^k) \right| \right).
\]  
(25)

(ii) *Unsuccessful iteration* (\( \mathbb{I}_{\phi} = 1 \)). \( \Delta^k \) is decreased and the change in function values is zero. Thus, the bound in the change of \( \Phi_k \) follows straightforwardly from (18) by replacing \( \mathbb{I}_{J_k} \) by \( \mathbb{I}_{J_k} \). More precisely, the following holds,
\[
\mathbb{I}_{J_k} \mathbb{I}_{\phi} (\Phi_{k+1} - \Phi_k) = - \mathbb{I}_{J_k} \mathbb{I}_{\phi} (1 - \nu)(1 - \tau^p)(\Delta^k)^p.
\]
(26)

Then, combining (25) and (26), yields
\[
\mathbb{I}_{J_k} (\Phi_{k+1} - \Phi_k) \leq \mathbb{I}_{J_k} \frac{\nu}{c \epsilon f} \left( \left| f(X^{k+1}) - F^k(S) \right| + \left| F^k_0 - f(X^k) \right| \right).
\]
(27)

Taking conditional expectations with respect to \( \mathcal{F}^{k-1}_{k-1} \) in both sides of (27) and applying Lemma 1 leads to (21).

Now, combining (13) and (21) leads to
\[
\mathbb{E} \left( \Phi_{k+1} - \Phi_k \mid \mathcal{F}^{k-1}_{k-1} \right) = \mathbb{E} \left( (\mathbb{I}_{J_k} + \mathbb{I}_{J_k}) (\Phi_{k+1} - \Phi_k) \mid \mathcal{F}^{k-1}_{k-1} \right)
\]
(28)

Then, choosing \( \beta \) according to (11) ensures that
\[-\beta (1 - \nu)(1 - \tau^p) + 2\nu (1 - \beta) \leq -\frac{1}{2} \beta (1 - \nu)(1 - \tau^p).\]
(29)

Hence, (12) follows from (28) and (29), which achieves the proof. \( \square \)

Note that (11) can always be satisfied. A demonstration of this satisfiability is provided in the following remark.

**Remark 1** \( \nu \in (0, 1) \) and \( \beta \in (1/2, 1) \) can always be chosen according to (11). Indeed, first observe that \( \tau, p \) and \( \gamma \) are fixed, which implies that \( \frac{2(\tau^{-p} - 1)}{\gamma - 2} \) is a con-
stant. Since \( \lim_{\nu \to 1} \frac{\nu}{1-\nu} = +\infty \), then \( \nu \) can always be chosen sufficiently close to one, so that
\[
\frac{\nu}{1-\nu} \geq \frac{2(\tau^{-p} - 1)}{\gamma - 2}.
\] (30)

Finally, assume that \( \nu \) is fixed and chosen according to (30). Consequently \( \frac{\nu}{1-\nu} \times \frac{4}{(1-\tau^p)} \) is a constant. By observing that \( \lim_{\beta \to 1} \frac{\beta}{1-\beta} = +\infty \), then \( \beta \) can always be chosen sufficiently close to one so that
\[
\frac{\beta}{1-\beta} \geq \frac{\nu}{1-\nu} \times \frac{4}{(1-\tau^p)}.
\]

The following result extending that in [4] from \( p = 2 \) to any \( p > 1 \), shows in particular that the sequence \( \{\Delta^k\}_{k \in \mathbb{N}} \) of step size parameters converges to zero with probability one.

**Theorem 3** Let all assumptions that were made in Theorem 2 hold. Then for all \( p > 1 \), the sequence \( \{\Delta^k\}_{k \in \mathbb{N}} \) of step size parameters generated by Algorithm 1 satisfies

\[
\sum_{k=0}^{+\infty} (\Delta^k)^p < +\infty \quad \text{almost surely.}
\]

**Proof** Let \( N \geq 1 \) be an integer and \( \phi = \frac{1}{2} \beta(1 - \nu)(1 - \tau^p) \). Summing (12) over \( k \in \{0, 1, \ldots, N\} \) leads to

\[
\phi \sum_{k=0}^{N} (\Delta^k)^p \leq \sum_{k=0}^{N} \mathbb{E}(\Phi_k - \Phi_{k+1}|\mathcal{F}_{k-1}^F).
\] (31)

Since \( \mathbb{E}[\mathbb{E}(\Phi_k - \Phi_{k+1}|\mathcal{F}_{k-1}^F)] = \mathbb{E}(\Phi_k - \Phi_{k+1}) \), then taking expectations on both sides of (31), yields

\[
\phi \sum_{k=0}^{N} \mathbb{E}[(\Delta^k)^p] \leq \sum_{k=0}^{N} \mathbb{E}(\Phi_k - \Phi_{k+1}) = \mathbb{E}(\Phi_0) - \mathbb{E}(\Phi_{N+1}) \leq \mathbb{E}(\Phi_0) \quad \text{for all } N \geq 1,
\] (32)

which implies that

\[
\sum_{k=0}^{+\infty} \mathbb{E}[(\Delta^k)^p] \leq \frac{\mathbb{E}(\Phi_0)}{\phi},
\] (33)

where the second inequality in (32) follows from the fact that \( \Phi_{N+1} > 0 \). Denote by \( \mathcal{P}(\mathbb{N}) \) the power set of \( \mathbb{N} \), i.e., the set of all subsets of \( \mathbb{N} \) and let us mention that the sum in (33) is an integral with respect to counting measure on the measurable space \( (\mathbb{N}, \mathcal{P}(\mathbb{N})) \). Since the expectation is also an integral with respect to the probability

\[ Springer \]
measure on \((\Omega, \mathcal{F})\) and \((\Delta^k)^p > 0\) for all \(k \in \mathbb{N}\), the sum and the expectation in (33) can be interchanged due to Fubini’s theorem (see e.g., Theorem 1.7.2 in [12]). Thus,

\[
\mathbb{E}
\left[
\sum_{k=0}^{+\infty} (\Delta^k)^p
\right] = \sum_{k=0}^{+\infty} \mathbb{E}[(\Delta^k)^p] \leq \frac{\mathbb{E}(\Phi_0)}{\rho} < +\infty,
\]

whence

\[
\sum_{k=0}^{+\infty} (\Delta^k)^p < +\infty \quad \text{almost surely.}
\]

\(\square\)

Consider the stochastic process \(\{(\Phi_k, \Delta^k, W_k)\}_{k \in \mathbb{N}}\), where \(\Phi_k\) is the same random function of Theorem 2, \(\Delta^k\) is the random step size parameter and \(W_k = 2(\mathbb{1}_{J_k} - \frac{1}{2})\). Define \(\hat{\rho} = \min(\rho - 1, 1)\) for some fixed \(\rho > 1\). For some arbitrary fixed \(\epsilon' \in (0, 1)\), consider the following random time \(T_{\epsilon'}\) defined by

\[
T_{\epsilon'} = \inf \left\{ k \in \mathbb{N} : \left\| \nabla f(X^k) \right\|^{1/\hat{\rho}} \leq \epsilon' \right\}
\]

Then, \(T_{\epsilon'}\) is a stopping time for the stochastic process generated by Algorithm 1 and is consequently a stopping time for \(\{(\Phi_k, \Delta^k, W_k)\}_{k \in \mathbb{N}}\) [9, 16]. Moreover, \(T_{\epsilon'/\hat{\rho}}\) is the number of iterations required by Algorithm 1 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 4.

In order to apply Theorem 1 to \(T_{\epsilon'}\), the remainder of this section is devoted to showing that Assumption 3 holds for the stochastic process \(\{(\Phi_k, \Delta^k, W_k)\}_{k \in \mathbb{N}}\). Let us show that Assumption 3-(iii) holds. First, multiplying both sides of (20) by \(\mathbb{1}_{\{T_{\epsilon'} > k\}}\) and taking conditional expectation with respect to \(\mathcal{F}_{k-1}\) lead to

\[
\mathbb{E}(\mathbb{1}_{\{T_{\epsilon'} > k\}} \mathbb{1}_{J_k} (\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}) \leq -\mathbb{E}(\mathbb{1}_{\{T_{\epsilon'} > k\}} \mathbb{1}_{J_k} (1 - \nu)(1 - \tau^p)(\Delta^k)^p | \mathcal{F}_{k-1})
\]

\[
= -\mathbb{1}_{\{T_{\epsilon'} > k\}} \mathbb{E}(\mathbb{1}_{J_k} (1 - \nu)(1 - \tau^p)(\Delta^k)^p | \mathcal{F}_{k-1})
\]

\[
\leq -\mathbb{1}_{\{T_{\epsilon'} > k\}} \beta(1 - \nu)(1 - \tau^p)(\Delta^k)^p,
\]

where the equalities in (35) follows from the fact that \(\mathbb{1}_{\{T_{\epsilon'} > k\}}\) is \(\mathcal{F}_{k-1}\)-measurable since \(T_{\epsilon'}\) is a stopping time with respect to the filtration \(\{\mathcal{F}_{k-1}\}_{k \geq 0}\). Similarly, it easily follows from (21) that

\[
\mathbb{1}_{\{T_{\epsilon'} > k\}} \mathbb{E}(\mathbb{1}_{J_k} (\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}) \leq \mathbb{1}_{\{T_{\epsilon'} > k\}} 2\nu(1 - \beta)(\Delta^k)^p.
\]

Thus, combining (35) and (36) (as we did for (13) and (21) in the proof of Theorem 2) implies that
\[\mathbb{1}_{\{T_r > k\}} E(\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}) \leq -\mathbb{1}_{\{T_r > k\}} \frac{1}{2} \beta (1 - \nu) (1 - \tau^p) (\Delta^k)^p,\]

which means that Assumption 3-(iii) holds with \(\eta = \frac{1}{2} \beta (1 - \nu) (1 - \tau^p)\) and \(h(x) = x^p\). Moreover, by choosing \(\lambda\) such that \(e^\lambda = \tau^{-1}\) and noticing that \(\Delta^k \leq \delta_{\max} = \delta^0 e^{h_{\max}}\) in Algorithm 1 for all \(k \in \mathbb{N}\), then Assumption 3-(i) holds.

Now, before showing that Assumption 3-(ii) holds, let us emphasize that as in the deterministic framework, polling directions in Algorithm 1 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 losing the positive spanning property \([17]\). For this purpose, let us recall the following definition of the cosine measure \([11, 14]\) of a positive spanning set \(\mathcal{D}^k\) with non-zero vectors

\[\kappa(\mathcal{D}^k) := \min_{v \in \mathbb{R}^n} \max_{d \in \mathcal{D}^k} \frac{v^T d}{||v|| \cdot ||d||},\]

In order to avoid the aforementioned deterioration of the polling directions, the positive spanning sets are required to satisfy the following assumption \([14, 17]\) where the size of the directions does not tend to infinity or approach zero, and the cosine measure always stays positive.

**Assumption 5** The following hold for all positive spanning sets \(\mathcal{D}^k\) used for polling in Algorithm 1. There exists a constant \(\kappa_{\min} > 0\) such that \(\kappa(\mathcal{D}^k) > \kappa_{\min}\) for all \(k\). There exist constants \(d_{\min} > 0\) and \(d_{\max} > 0\) such that \(d_{\min} \leq ||d|| \leq d_{\max}\) for all \(d \in \mathcal{D}^k\).

The following result from \([14]\) will be useful for the remainder of the analysis, and specifically the proof of the key result in Lemma 2. It shows by means of the cosine measure \(\kappa(\mathcal{D}^k)\), how far the steepest descent direction can be in the worst case from the vector in \(\mathcal{D}^k\). This means in terms of descent that, there exists \(d^*_k \in \mathcal{D}^k\) such that

\[\kappa(\mathcal{D}^k) \left\| \nabla f(x^k) \right\| \left\| d^*_k \right\| \leq -\nabla f(x^k)^T d^*_k.\]  

(37)

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

**Assumption 6** The gradient \(\nabla f\) of the objective function \(f\) is \(L\)-Lipschitz continuous everywhere.

Define the constant \(\delta_{\epsilon}\), as follows

\[\delta_{\epsilon} = \frac{e^\epsilon}{\zeta}\]  

with  \(\zeta > \left[ \kappa_{\min}^{-1} (L d_{\max} + (\gamma + 2) c e_f d_{\min}^{-1}) \right]^{1/p},\]  

(38)
where without loss of generality, \( Ld_{\max} > \kappa_{\min} \) so that \( \delta_{\varepsilon} < 1 \) for the needs of the analysis and specifically, the proof of Lemma 2. Then following [9], it can be assumed without any loss of generality that \( \delta_{\varepsilon} = \tau^{-i}\delta_{0}^{i} \), for some integer \( i \leq 0 \). Hence, for any \( k \), \( \Delta^{k} = \tau^{i}\delta_{\varepsilon} \), for some integer \( i_{k} \). Thus, what remains to be proved in Assumption 3 in order to apply Theorem 1 is the dynamics (8). Note however that the proof of the latter dynamics which will be achieved in Lemma 3, needs the following intermediate key result. Indeed, in the stochastic trust-region framework of [9], the proof of a similar dynamics strongly relies on the fact that any iteration \( k \), where \( \|\nabla f(x_{k})\| > \epsilon \) and for which a “good” model and estimates occur, is successful provided that the trust-region radius \( \delta_{k} \) is below a threshold \( \Delta_{\varepsilon} \). Nevertheless, unlike the latter framework where information 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 3-(ii) holds, lies in linking the event \( \{ \|\nabla f(X^{k})\|^{1/\beta} > \epsilon' \} \) to a successful iteration of Algorithm 1, which is done next.

**Lemma 2** Assume that Assumptions 5 and 6 hold, and that \( \delta^{k} \leq \delta_{\varepsilon} \). Let \( f_{0}^{k} \) and \( f^{k}(s) \) be \( \epsilon_{f} \)-accurate estimates of \( f(x^{k}) \) and \( f(x^{k} + s) \) respectively. If \( \|\nabla f(x^{k})\|^{1/\beta} > \epsilon' \), then there exists \( s \in \mathcal{S}^{k} = \{ \delta^{k}d \mid d \in \mathbb{D} \} \) such that

\[
f^{k}(s) - f_{0}^{k} \leq -\gamma c\epsilon_{f} (\delta^{k})^p.
\]

In particular, this means that the iteration \( k \) of Algorithm 1 is successful.

**Proof** The proof uses elements derived in [14]. Suppose that \( \delta^{k} \leq \delta_{\varepsilon} \) and assume in contradiction that \( f^{k}(s) - f_{0}^{k} > -\gamma c\epsilon_{f} (\delta^{k})^p \) for all \( s \in \mathcal{S}^{k} \). In particular, this inequality holds for \( s = \delta^{k}d_{s}^{k} \), where \( d_{s}^{k} \) is the direction satisfying (37). Since the estimates \( f_{0}^{k} \) and \( f^{k}(s) \) are \( \epsilon_{f} \)-accurate, then it follows from the following equality

\[
f(x^{k} + s) - f(x^{k}) = f(x^{k} + s) - f^{k}(s) + (f^{k}(s) - f_{0}^{k}) + f^{k}_{0} - f(x^{k})
\]

that

\[
f(x^{k} + s) - f(x^{k}) + (\gamma + 2)c\epsilon_{f} (\delta^{k})^p \geq 0. \tag{39}
\]

It follows from the mean value theorem, combined with (39), that there exists a constant \( \mu_{k} \in [0, 1] \) such that

\[0 \leq \delta^{k}\nabla f(x^{k} + \mu_{k}\delta^{k}d_{s}^{k})^\top d_{s}^{k} + (\gamma + 2)c\epsilon_{f} (\delta^{k})^p. \tag{40}\]

Dividing both sides of (40) by \( \delta^{k} \) and subtracting \( \nabla f(x^{k})^\top d_{s}^{k} \), yields

\[-\nabla f(x^{k})^\top d_{s}^{k} \leq [\nabla f(x^{k} + \mu_{k}\delta^{k}d_{s}^{k}) - \nabla f(x^{k})]^\top d_{s}^{k} + (\gamma + 2)c\epsilon_{f} (\delta^{k})^{p-1}. \tag{41}\]

Putting (37) and (41) together, yields
\[ \kappa(D^k) \| \nabla f(x^k) \| \| d^k_s \| \leq [\nabla f(x^k + \mu_k \delta^k d^k_s) - \nabla f(x^k)]^T d^k_s + (\gamma + 2) c \varepsilon f(\delta^k)^{p-1}. \]  

\[ \text{(42)} \]

Then, dividing both sides of (42) by \( \kappa(D^k) \| d^k_s \| \) and using Assumptions 6 and 5, lead to

\[ \| \nabla f(x^k) \| \leq \kappa^{-1}_{\text{min}} \left[ L_{\text{max}} \delta^k + (\gamma + 2) c \varepsilon f d^{-1}_{\text{min}}(\delta^k)^{p-1} \right] \leq \kappa^{-1}_{\text{min}} \left[ L_{\text{max}} + (\gamma + 2) c \varepsilon f d^{-1}_{\text{min}}(\delta^k)^{\min(p-1,1)} \right], \]

\[ \text{(43)} \]

where the inequality (43) follows from the fact that \( \delta^k \leq \delta_{\varepsilon'} < 1 \). Now, recall that \( \hat{p} = \min(p - 1, 1) \) and let \( L_1 := \kappa^{-1}_{\text{min}} \left( L_{\text{max}} + (\gamma + 2) c \varepsilon f d^{-1}_{\text{min}} \right) \). Then, it follows from (43) that

\[ \| \nabla f(x^k) \|^{1/\hat{p}} \leq L_1^{1/\hat{p}} \delta^k \leq L_1^{1/\hat{p}} \delta_{\varepsilon'} = L_1^{1/\hat{p}} \frac{\delta_{\varepsilon'}}{\varepsilon} \leq \varepsilon', \]

\[ \text{(44)} \]

where the last inequality in (44) follows from (38), which achieves the proof. \qed

Finally, the following result shows that the dynamics (8) of Assumption 3-(ii) holds.

**Lemma 3** Let Assumptions 5, 6 and all assumptions that were made in Theorem 2 hold. Then Assumption 3-(ii) is satisfied for the random variable \( W_k = 2(\mathbb{1}_{J_k} - \frac{1}{2}) \), with \( \lambda = -\ln(\tau) \) and \( q = \beta \).

**Proof** The result is proved by adapting the proof of a similar Lemma from [9]. First, notice that (8) trivially holds when \( \mathbb{1}_{\{T_{\varepsilon'} > k\}} = 0 \). Thus, the remainder of the proof is devoted to showing that conditioned on the event \( \{T_{\varepsilon'} > k\} \), i.e., when \( \mathbb{1}_{\{T_{\varepsilon'} > k\}} = 1 \), the following holds

\[ \Delta^{k+1} \geq \min \{ \delta_{\varepsilon'}, \min \{ \tau^{-1} \Delta^k, \delta_{\text{max}} \} \mathbb{1}_{J_k} + \tau \Delta^k \mathbb{1}_{J_k} \}. \]

\[ \text{(45)} \]

Recall that \( \delta^k = \tau^i \delta_{\varepsilon'} \) for some integer \( i_k \), so if \( \delta^k > \delta_{\varepsilon'} \) then \( \delta^k \geq \tau^{-1} \delta_{\varepsilon'} \), whence \( \delta^{k+1} \geq \tau \delta^k \geq \delta_{\varepsilon'} \). Now, assume that \( \delta^k \leq \delta_{\varepsilon'} \). Since \( T_{\varepsilon'} > k \), then it is the case that \( \| \nabla f(x^k) \|^{1/\hat{p}} > \varepsilon' \). If \( \mathbb{1}_{J_k} = 1 \), then the estimates are good and are specifically \( \varepsilon f \) -accurate. Hence, it follows from Lemma 2 that the \( k \)th iteration is successful. Thus, \( x^{k+1} = x^k + s \) for some \( s \in A_{\delta^k} \) and \( \delta^{k+1} = \min \{ \tau^{-1} \delta^k, \delta_{\text{max}} \} \). But if \( \mathbb{1}_{J_k} = 0 \), i.e., \( \mathbb{1}_{J_k} = 1 \), then the inequality \( \delta^{k+1} \geq \tau \delta^k \) always holds by the dynamics of Algorithm 1. The proof is complete by noticing finally that \( \mathbb{P}(J_k \mid J_{k-1}) \geq q = \beta \). \qed

### 4.2 Complexity result and first-order optimality conditions

The following result provides a bound on the expected number of iterations taken by Algorithm 1 before \( \{ \| \nabla f(X^k) \| \leq \varepsilon \} \) occurs. It is the main result of the present work.
\textbf{Theorem 4} Let Assumptions 5, 6 and all assumptions that were made in Theorem 2 hold with $\beta \in (1/2, 1)$ and $\nu \in (0, 1)$ satisfying (11). Consider Algorithm 1 and the corresponding stochastic process. For some arbitrary fixed $\epsilon \in (0, 1)$, consider the random time $T_\epsilon^*$ defined by

$$T_\epsilon^* = \inf \left\{ k \in \mathbb{N} : \| \nabla f(X_k) \| \leq \epsilon \right\}. \quad (46)$$

Then,

$$\mathbb{E}(T_\epsilon^*) \leq \frac{2 \Phi_0 L_2}{(2 \beta - 1)(1 - \nu)(1 - \tau^p)} \epsilon^{-\beta p/(\min(p - 1, 1))} + 1, \quad (47)$$

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

\textbf{Proof} In order for the inequality in (38) to be satisfied, let $\zeta$ be chosen according to

$$\zeta = \left[ 1 + \kappa^{-1}_{\min} (L \max + (\gamma + 2)c \epsilon f d^{-1}_{\min}) \right]^{1/\hat{p}}, \quad \text{where } \hat{p} = \min(p - 1, 1). \quad (48)$$

As shown previously, since Assumption 3 holds for the stochastic process $\{(\Phi_k, \Delta_k, W_k)\}_{k \in \mathbb{N}}$ generated by Algorithm 1, with $q = \beta$, $h(x) = x^{\beta}$, $\eta = \frac{1}{2} \beta (1 - \nu)(1 - \tau^p)$ and $\delta_{\sup} = \epsilon' / \zeta$, then Theorem 1 applies for the stopping time $T_{\epsilon'}$ defined in (34). Thus, the following inequality holds for all $\epsilon' \in (0, 1)$

$$\mathbb{E}(T_{\epsilon'}) \leq \frac{\beta}{2 \beta - 1} \times \frac{\Phi_0 \zeta^p}{\eta \epsilon'^p} + 1, \quad (49)$$

Now, let $\epsilon \in (0, 1)$ be arbitrary fixed. Then $\epsilon^{1/\hat{p}} \in (0, 1)$, which means that (49) holds in particular for $\epsilon' = \epsilon^{1/\hat{p}}$. By noticing moreover that $T_{\epsilon^{1/\hat{p}}} = T_\epsilon^*$, then it follows from (49) that

$$\mathbb{E}(T_\epsilon^*) \leq \frac{\beta}{2 \beta - 1} \times \frac{2 \Phi_0 \zeta^p}{\beta (1 - \nu)(1 - \tau^p)} \epsilon^{-\beta p/\hat{p}} + 1. \quad (50)$$

Finally, (47) follows from (50) with the specific choice of $\zeta$ given by (48), which achieves the proof. \qed

The complexity result of Theorem 4 motivates the following discussion about some consequences and tradeoffs of allowing for more general $p$ in the forcing function of the present work. First, observe that the expected complexity bound of (47) can never approach $O(\epsilon^{-1})$ which is what one would expect in a more structured (e.g., convex) optimization. This is because $\frac{p}{\min(p - 1, 1)} \geq 2$ for all $p > 1$, where the equality holds only for $p = 2$, in which case the corresponding expected complexity bound of $O(\epsilon^{-2})$ is that of standard nonconvex optimization. When $p > 2$ in which case the expected complexity bound is $O(\epsilon^{-\hat{p}})$, the corresponding convergence rate is slower than that of standard nonconvex optimization since
of $\varepsilon \in (0, 1)$ and hence $\varepsilon^{-p} > \varepsilon^{-2}$, but in this case, less decrease from the forcing function is demanded. At the same time, however, e.g. for $p = 3$, an $\varepsilon_f$-accurate solution requires $O((\delta_k)^3)$ accuracy (and $O((\delta_k)^6)$ variance bounds per Assumption 2) which is a steep price to pay. On the flip side for $1 < p < 2$, especially as $p \searrow 1$, less accuracy in the estimates is demanded, which is nice, but in this case the corresponding convergence rate is very slow unfortunately since
\[
\lim_{p \searrow 1} \frac{\varepsilon}{\min(p - 1, 1)} = +\infty,
\]
and more decrease per iteration (approaching $O(\delta_k)$) is demanded.

The following lim inf-type first-order necessary optimality condition is a simple consequence of the complexity result of Theorem 4. It shows the existence of a subsequence of random iterates generated by Algorithm 1 which drives the norm of the gradient of $f$ to zero with probability one. Note that a similar corollary has been derived in [16].

**Theorem 5** Let Assumptions 5, 6 and all assumptions that were made in Theorem 2 hold. Then the sequence $\{X^k\}_{k \in \mathbb{N}}$ of random iterates generated by Algorithm 1 satisfies
\[
\liminf_{k \to +\infty} \left\| \nabla f(X^k) \right\| = 0 \quad \text{almost surely.} \tag{51}
\]

**5 Discussion**

This section provides some insight as to why Assumption 2-(ii) cannot be dropped from the analysis of Algorithm 1 in favor of the Taylor’s theorem based approach of [10]. We noticed in the proof of Theorem 2 that by bounding the variance in the function estimates, the possible increase occurring in $f$ and hence in $\Phi_k$ due to bad estimates is bounded in expectation by means of Lemma 1. Indeed a similar variance assumption was used for the analysis of a stochastic line search method in [16], but was not needed in [9, 10] to analyze STORM, a stochastic trust-region algorithm using random models and function estimates. In fact, it has been possible in [9, 10] to bound the increase of the objective function and connect it to the decrease that one hopes to obtain if both the model and the estimates are good. More precisely by making use of a Taylor expansion, the possible amount of increase in $f$ was shown in [9, 10] to be proportional to $\| \nabla f(x_k) \| \delta_k$, i.e.,
\[
f(x_k + s_k) - f(x_k) \leq C_3 \| \nabla f(x_k) \| \delta_k, \tag{52}
\]
while a decrease in $f$ (see e.g. [10], Lemma 4.6) is such that
\[
f(x_k + s_k) - f(x_k) \leq -C_1 \| \nabla f(x_k) \| \delta_k, \tag{53}
\]
where $C_1 > 0$ and $C_3 > 0$ are constants, and $x_k, s_k$ and $\delta_k$ denote respectively a current iterate, the step length and the trust-region radius. Looking closely at the proof of Theorem 3 in [9] (or that of Theorem 4.11 in [10]), one can easily notice that for
the analysis of STORM to be carried out making use of (52) instead of a variance assumption such as Assumption 2-(ii), the decrease in $f$ had to be proportional to $\|\nabla f(x_k)\|\delta_k$ as in (53). Indeed, even though STORM uses a gradient approximation instead of a true gradient information, expressing however the decrease in $f$ in terms of $\nabla f(x_k)$ is achieved by means of the definition of a $k$-fully linear model which is a key tool of the proof of the aforementioned Lemma 4.6 [10]. Moreover, the framework analyzed in [9] making use of (52) does not allow for more general $p > 1$ in the complexity result, i.e., the derived complexity bound is $O(e^{-2})$, which corresponds to $O\left(e^{\frac{p-2}{p-1}}\right)$ as in [17], but only for $p = 2$. Thus, Assumption 2-(ii) cannot be dropped from the analysis of SDDS in favor of the use of a Taylor expansion for two reasons. First, the latter approach prevents from allowing e.g. for more general $p > 1$ in the forcing function and consequently in the complexity result, and second, the present framework does not offer any possibility of showing the decrease in $f$ to be proportional to $\|\nabla f(x^k)\|\delta^k$.

6 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.

Acknowledgements The author is grateful to Sébastien Le Digabel and Charles Audet from Polytechnique Montréal, Michael Kokkolaras from McGill university and Matt Menickelly from the Argonne national laboratory for valuable discussions and constructive suggestions that improve the quality of the presentation. He would also like to thank an anonymous referee for his/her careful reading and helpful suggestions that contributed to improve this work. 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.
References

1. Alarie, S., Audet, C., Bouchet, P.Y., Digabel, S.L.: Optimization of noisy blackboxes with adaptive precision. Technical Report G-2019-84, Les cahiers du GERAD (2019)
2. Audet, C., Dennis, J.E., Jr.: Analysis of generalized pattern searches. SIAM J. Optim. 13(3), 889–903 (2003)
3. Audet, C., Dennis, J.E., Jr.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. 17(1), 188–217 (2006)
4. Audet, C., Dzahini, K.J., Kokkolaras, M., Le Digabel, S.: Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. Comput. Optim. Appl. 79(1), 1–34 (2021)
5. Audet, C., Hare, W.: Derivative-Free and Blackbox Optimization. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham (2017)
6. Bandeira, A.S., Scheinberg, K., Vicente, L.N.: Convergence of trust-region methods based on probabilistic models. SIAM J. Optim. 24(3), 1238–1264 (2014)
7. Berahas, A.S., Cao, L., Scheinberg, K.: Global convergence rate analysis of a generic line search algorithm with noise. SIAM J. Optim. 31(2), 1489–1518 (2021)
8. Bhattacharya, R.N., Waymire, E.C.: A Basic Course in Probability Theory, vol. 69. Springer, Berlin (2007)
9. Blanchet, J., Cartis, C., Menickelly, M., Scheinberg, K.: Convergence rate analysis of a stochastic trust-region method via supermartingales. INFORMS J. Optim. 1(2), 92–119 (2019)
10. Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models. Math. Program. 169(2), 447–487 (2018)
11. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. MOS-SIAM Series on Optimization. SIAM, Philadelphia (2009)
12. Durrett, R.: Probability: Theory and Examples. Cambridge University Press, Cambridge (2010)
13. Gratton, S., Royer, C.W., Vicente, L.N., Zhang, Z.: Direct search based on probabilistic descent. SIAM J. Optim. 25(3), 1515–1541 (2015)
14. Kolda, T.G., Lewis, R.M., Torczon, V.: Optimization by direct search: new perspectives on some classical and modern methods. SIAM Rev. 45(3), 385–482 (2003)
15. Larson, J., Billups, S.C.: Stochastic derivative-free optimization using a trust region framework. Comput. Optim. Appl. 64(3), 619–645 (2016)
16. Paquette, C., Scheinberg, K.: A stochastic line search method with expected complexity analysis. SIAM J. Optim. 30(1), 349–376 (2020)
17. Vicente, L.N.: Worst case complexity of direct search. EURO J. Comput. Optim. 1(1), 143–153 (2013)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.