Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 22–43

A WEAK TAIL-BOUND PROBABILISTIC CONDITION FOR FUNCTION ESTIMATION IN STOCHASTIC DERIVATIVE-FREE OPTIMIZATION

F. RINALDI, L. N. VICENTE AND D. ZEFFIRO

ABSTRACT: Using tail bounds, we introduce a new probabilistic condition for function estimation in stochastic derivative-free optimization which leads to a reduction in the number of samples and eases algorithmic analyses. Moreover, we develop simple stochastic direct-search and trust-region methods for the optimization of a potentially non-smooth function whose values can only be estimated via stochastic observations. For trial points to be accepted, these algorithms require the estimated function values to yield a sufficient decrease measured in terms of a power larger than 1 of the algoritmic stepsize.

Our new tail-bound condition is precisely imposed on the reduction estimate used to achieve such a sufficient decrease. This condition allows us to select the stepsize power used for sufficient decrease in such a way to reduce the number of samples needed per iteration. In previous works, the number of samples necessary for global convergence at every iteration k of this type of algorithms was $O(\Delta_k^{-4})$, where Δ_k is the stepsize or trust region radius. However, using the new tail-bound condition, and under mild assumptions on the noise, one can prove that such a number of samples is only $O(\Delta_k^{-2-\varepsilon})$, where $\varepsilon > 0$ can be made arbitrarily small by selecting the power of the stepsize in the sufficient decrease test arbitrarily close to 1. The global convergence properties of the stochastic direct-search and trust-region algorithms are established under the new tail bound condition.

KEYWORDS: Derivative-free optimization, Direct search, Nonsmooth optimization, Trust-region methods.
MATH. SUBJECT CLASSIFICATION (2010): 90C15, 90C30, 90C56.

1. Introduction

We consider the following unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1.1}$$

with f locally Lipschitz continuous and possibly non-smooth function with $\inf f = f^* \in \mathbb{R}$. We assume that the original function $f(x)$ is not computable, and the only information available on f is given by a stochastic oracle producing an estimate $\tilde{f}(x)$ for any $x \in \mathbb{R}^n$. In some contexts, we can assume

Received December 6, 2022.

Support for LNV was partially provided by the Centre for Mathematics of the University of Coimbra under grant FCT/MCTES UIDB/MAT/00324/2020.

that the estimate is a random variable parameterized by x, that is

$$\tilde{f}(x) = F(x, \xi),$$

with the black-box oracle given by sampling on the ξ space. When dealing with, e.g., statistical learning problems, the function $F(x, \xi)$ evaluates the loss of the decision rule parametrized by x on a data point ξ (see, e.g., [17] for further details). In simulation-based engineering applications, the function $F(x, \xi)$ is simply related to some noisy computable version of the original function. In this case ξ represents the random variable that induces the noise (a classic example is given by Monte Carlo simulations). A detailed overview is given in, e.g., [1].

When this random variable is exact in expected value, problem (1.3) turns out to be the expected loss formulation

$$\min_{x \in \mathbb{R}^n} \mathbb{E}_\xi[F(x, \xi)], \tag{1.2}$$

a case addressed in recent literature, see, e.g., [18, 26], for further details.

Although the role of derivative-free optimization is particularly important when the black-box representing the function is somehow noisy or, in general, of a stochastic type, traditional DFO methods have been developed primarily for deterministic We consider the following unconstrained optimization problem

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

where f is locally Lipschitz continuous and possibly non-smooth with $\inf f = f^* \in \mathbb{R}$. We assume that the original function f is not computable and that the only information available about f is given by a stochastic oracle producing an estimate $\tilde{f}(x)$ for any $x \in \mathbb{R}^n$. In some contexts, we can assume that the estimate is a random variable parameterized by x, that is

$$\tilde{f}(x) = F(x, \xi),$$

with the black-box oracle given by sampling on the ξ space. When dealing with statistical learning problems [17], the function $F(x, \xi)$ evaluates the loss of the decision rule parametrized by x on a data point ξ. In simulation-based engineering applications [1], the function $F(x, \xi)$ is simply related to some noisy computable version of the original function. In this case, ξ represents the random variable that induces the noise (a classic example is given by Monte Carlo simulations). When this random variable is exact in expected value, problem (1.3) turns out to be the expected loss formulation

$$\min_{x \in \mathbb{R}^n} \mathbb{E}_\xi[F(x, \xi)], \tag{1.4}$$

a case addressed in recent literature, see, e.g., [18, 26] for further details.

1.1. A short review of stochastic derivative-free optimization. Although the role of derivative-free optimization is particularly important when the black-box representing the function is somehow noisy or, in general, of a stochastic type, traditional DFO methods have been developed primarily for deterministic functions, and only recently adapted to deal with stochastic observations (see, e.g., [8] for a detailed discussion on this matter). We give here a brief overview of the main results available in the literature by first focusing on *model-based* strategies and then moving to *direct-search* approaches (see, e.g., [3, 10] for further details on these two classes of methods).

In [18], the authors describe a trust-region algorithm to handle noisy objectives and prove convergence when f is sufficiently smooth (i.e., with Lipschitz continuous gradient) and the noise is drawn independently from a distribution with zero mean and finite variance, that is they aim at solving a smooth version of problem (1.4), when ξ is additive noise. In the same line of research, the authors in [26] developed a class of derivative-free trust-region algorithms, called ASTRO-DF, for unconstrained optimization problems whose objective function has Lipschitz continuous gradient and can only be implicitly expressed via a Monte Carlo oracle. The authors consider again an objective with noise drawn independently from a distribution with zero mean, finite variance and a bound on the $4v$-th moment (with $v \geq 2$), and prove the almost sure convergence of their method when using stochastic polynomial interpolation models. Another relevant reference in this context is given by [8], where the authors analyze a trust-region model-based algorithm for solving unconstrained stochastic optimization problems. They consider random models of a smooth objective function, obtained from stochastic observations of the function or its gradient. Convergence rates for this class of methods are reported in [6]. The frameworks analyzed in [6, 7, 8] extend the trust-region DFO method based on probabilistic models described in [5]. It is important to notice that the randomness in the models described in [5] comes from the way sample points are chosen, rather than from noise in the function evaluations. All the above-mentioned model-based approaches consider functions with a certain degree of smoothness (e.g., with Lipschitz continuous gradient) and assume that a probabilistically accurate gradient estimate (e.g., some kind of probabilistically fully-linear model) can be generated, while of course such an estimate is not available when dealing with non-smooth functions.

A detailed convergence rate analysis of stochastic direct search variants is reported in [12] for the smooth case, i.e., for an objective function with Lipschitz continuous gradient. The main theoretical results are obtained by suitably adapting the supermartingale-based framework proposed in [6]. A stochastic mesh adaptive direct search for black-box nonsmooth optimization is proposed in [2]. The authors prove convergence with probability one to a Clarke stationary point [9] of the objective function by assuming that stochastic observations are sufficiently accurate and satisfy a variance condition. The considered analysis adapts to the direct-search gradient-free framework the theoretical analysis given in [23] for a class of stochastic gradient-based methods. It was extended in [13] to the constrained case.

1.2. The contributions of this paper. The main goal of this paper is to introduce a tail-bound probabilistic condition needed to deal with a stochastic black-box function in general direct-search and trust-region schemes. This probabilistic condition focuses on the reduction estimate (i.e., the estimate of the difference between the function at the current iterate and at a potential next iterate) used in the acceptance test of those derivative-free algorithms. It expresses a bound on the probability that the reduction estimate error is greater than a fraction of a stepsize power characterizing the sufficient decrease needed for trial-point acceptance, and can therefore be easily adapted to different choices of the power defining such a sufficient decrease.

Our condition enables us to define a trade-off between noise, algorithm parameters, and number of samples per iteration needed to achieve global convergence. One of our results is that if all the noise moments are finite, like in the case of Gaussian noise, we only need $O(\Delta_k^{-2-\varepsilon})$ samples, where Δ_k is the stepsize at iteration k (see Corollary 2.2.1). Here, $\varepsilon > 0$ can be made arbitrarily small by selecting the sufficient decrease power arbitrarily close to 1. This result compares to the $O(\Delta_k^{-4})$ number of samples required in previous works on stochastic trust-region methods (see, e.g., [6, 8, 26]) and stochastic direct-search methods (see [2, 12, 13]). We further show that the number of samples needed can be lowered to $O(\Delta_k^{-\varepsilon})$ when the sampling errors are suitably correlated and the random number generator is known (see Corollary 2.8).

We introduce two different algorithmic schemes, namely a simple stochastic direct-search strategy and a stochastic version of the basic deterministic trust-region scheme reported in [19]. Both schemes work as follows: they randomly generate a direction (direct search) or a linear term (trust region); then generate the new iterate by either moving along the direction (direct search) or by solving a trust-region subproblem (trust region); finally they use a sufficient decrease acceptance test to decide if the new point can be accepted (successful iteration) or not. In this work, we use stochastic function estimates in the acceptance tests rather than exact values. Our tail bound condition applies to the function reduction estimates of both schemes, and it allows us to deduce global convergence and to take advantage of the improvement in the number of samples per iteration. We remark that the convergence analysis of our trust-region scheme is developed under a new bound on the Hessian of the quadratic model which allow us to generate non-unit linear terms, and thus generalizing the deterministic version given in [19].

Lastly, we show that, for suitable choices of the algorithmic parameters, our tail-bound condition is implied by the variance conditions considered in [2] and by the probabilistically accurate function estimate assumption used in [2, 8, 23]. It is also interesting to notice that the finite variance oracle usually considered in the literature (see, e.g., [18, 26]) can be replaced by a more general finite moment oracle (see Subsection 2.2.1 for further details) when constructing estimates satisfying our conditions.

1.3. Outline of the paper. In Section 2, we introduce our tail-bound probabilistic condition, prove the new bounds on the number of samples needed per iteration to satisfy the condition, and compare it to existing conditions from the literature. We then analyze the direct-search and trust-region schemes in Sections 3 and 4, respectively. In both cases, the analysis has two main steps. In the first one, we show a result that implies convergence of the stepsize/trust-region radius to zero almost surely. In the second one, we focus on the random sequence of the unsuccessful iterations and prove, by exploiting the first result, Clarke stationarity at certain limit points. Numerical results for the direct-search scheme on a standard set of problems are reported in Section 5. Finally, we draw some conclusions and discuss some possible extensions in Section 6. In order to improve readability and ease the comprehension, we leave some proofs and additional numerical results to an appendix.

2. A weak tail-bound probabilistic condition for function estimation

In order to give convergence results for our algorithms, we need to introduce a tail-bound probabilistic condition on the accuracy of the function oracle. The stochastic quantities defined hereafter lie in a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, with probability measure $\mathbb{P}$ and σ-algebra $\mathcal{F}$ containing subsets of Ω, which is the space of the realizations of the algorithms under analysis. Any single outcome of the sample space Ω will be denoted by w. For a random variable X defined in Ω we use the shorthand $\{X \in A\}$ to denote $\{w \mid X(w) \in A\}$.

Our algorithms take a step along a certain direction, which can be a direct-search direction or a trust-region step, and in both cases there is a suitable stepsize quantifying the displacement. The algorithms generate a random process whose random quantity realizations are indicated as follows. The direction, the stepsize, and the current point are denoted by G_k, Δ_k, and X_k, with realizations g_k, δ_k, and x_k respectively. The function values $f(X_k)$ and $f(X_k + \Delta_k G_k)$ are denoted by F_k and F_k^g, with realizations f_k and f_k^g respectively. In the direct-search case, the acceptance criterion will be defined as

$$f_k - f_k^g \geq \theta \delta_k^q \,, \tag{2.1}$$

for some $\theta > 0$ and $q > 1$, with δ_k replaced by the norm of the step $\|s_k\|$ in the trust-region case. We define $\mathcal{F}_{k-1}$ as the $\sigma-$algebra of events up to the choice of G_k (so that in particular G_k is measurable with respect to $\mathcal{F}_{k-1}$). More explicitly, we define $\mathcal{F}_{k-1}$ as the σ-algebra generated by $(F_j, F_j^g)_{j=0}^{k-1}$ and $(G_j)_{j=0}^k$. Finally, we use $\mathbb{E}$ to denote expectation and conditional expectation, $\hat{x}$ as a shorthand for $x/\|x\|$, with $\hat{x} = 0$ for $x = 0$, and a.s. as a shorthand for *almost surely*.

2.1. The weak tail-bound probabilistic condition.

We now introduce our tail bound assumption, related to the acceptance criterion (2.1).

Assumption 2.1. *For some $\varepsilon_q > 0$ (independent of k):*

$$\mathbb{P}\left(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha \Delta_k^q \;|\mathcal{F}_{k-1}\right) \leq \frac{\varepsilon_q}{\alpha^{q/(q-1)}} \tag{AS1}$$

a.s. for every $\alpha > 0$.

The above assumption is in particular a power law tail bound with exponent $q/(q-1)$. Notice that we are only assuming an error bound for the estimate of the difference $f(X_k) - f(X_k + \Delta_k G_k)$ and not for the estimates of $f(X_k)$ and $f(X_k + \Delta_k G_k)$ taken individually; we basically want to bound the probability that the error in that estimate is large, as such an estimation plays a crucial role in the acceptance tests of our algorithms. We will see in Sections 3.2 and 4.2 that knowledge of an upper bound ε_q is needed in order to ensure convergence in the proposed algorithms.

Remark 2.2. As we will see in Section 2.2.1, Assumption 2.1 can be posed for any q provided that the r-th moment of the evaluation noise is finite, for $r = q/(q-1)$. Furthermore, for $q \in (1,2)$, the number of samples needed to satisfy Assumption 2.1 is just $O(\Delta_k^{-2q})$ rather than the standard $O(\Delta_k^{-4})$ required under finite variance assumptions [2]. This improvement is possible thanks to the strict relation between the tail bound (AS1) and the acceptance criterion (2.1), together with classic results from probability theory on the convergence rate for the law of large numbers. More precisely, we will use the fact that, for A average of m i.i.d. samples with finite r-th finite moment, there is a tail bound of the form

$$\mathbb{P}(A \geq \alpha) \leq \frac{K_{m,r}}{\alpha^r},$$

with $K_{m,r} \propto m^{-\frac{r}{2}}$ (as a consequence of Rosenthal's inequality [15]).

In our convergence arguments we will need Assumption 2.1 with a $\mathcal{F}_{k-1}$ measurable random variable A rather than a real number α. This is justified by the following lemma.

Lemma 2.3. *Let* A *be a positive* $\mathcal{F}_{k-1}$ *measurable random variable. If* (AS1) *holds, then it holds also with* A *instead of* α.

Proof: We prove the result in the case where A is a discrete random variable with a countable set of possible realizations $\{a_i\}_{i\in\mathbb{N}}$, which is sufficient since the general case then follows by approximation. Let $Y = |F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))|/\Delta_k^q$, and $r = \frac{q}{q-1}$. By the definition of conditional probability, (AS1) holds with A instead of α if and only if, for every $F \in \mathcal{F}_{k-1}$:

$$\mathbb{E}[\mathbb{1}_F \mathbb{1}_{\{Y \geq \mathrm{A}\}}] \leq \mathbb{E}[\mathbb{1}_F \frac{\varepsilon_q}{\mathrm{A}^r}]. \tag{2.2}$$

Indeed we have

$$\begin{aligned}\mathbb{E}[\mathbb{1}_F\mathbb{1}_{\{Y\geq \mathrm{A}\}}] &= \sum_{i\in\mathbb{N}}\mathbb{E}[\mathbb{1}_F\mathbb{1}_{\{Y\geq \mathrm{A}\}}\mathbb{1}_{\{\mathrm{A}=a_i\}}] = \sum_{i\in\mathbb{N}}\mathbb{E}[\mathbb{1}_{F\cap\{\mathrm{A}=a_i\}}\mathbb{1}_{\{Y\geq a_i\}}]\\ &\leq \sum_{i\in\mathbb{N}}\mathbb{E}[\mathbb{1}_{F\cap\{\mathrm{A}=a_i\}}\frac{\varepsilon_q}{a_i^r}] = \sum_{i\in\mathbb{N}}\mathbb{E}[\mathbb{1}_F\mathbb{1}_{\{\mathrm{A}=a_i\}}\frac{\varepsilon_q}{a_i^r}] = \mathbb{E}[\mathbb{1}_F\frac{\varepsilon_q}{\mathrm{A}^r}]\end{aligned} \tag{2.3}$$

as desired, where we used that $F\cap\{\mathrm{A}=a_i\}$ is measurable w.r.t. $\mathcal{F}_{k-1}$ together with (AS1) for $\alpha = a_i$ in the inequality. ■

In the remaining of this section, we will report the bounds on the number of samples needed to satisfy Assumption 2.1, as well as a comparison with existing conditions. The proofs are rather technical and can be found in the appendix.

2.2. Sampling improvement under the new condition. We will show that our tail-bound condition can be satisfied under a reduced number of function samples.

2.2.1. *Finite moment oracle.* We deal first with the case where the error of the oracle has finite r-th moment, for some $r>1$:

$$\begin{aligned} f(x) &= \mathbb{E}_\xi[F(x,\xi)]\,,\\ \mathrm{E}_\xi[|F(x,\xi)-f(x)|^r] &\leq M_r < +\infty\,. \end{aligned} \tag{2.4}$$

Recall that finite r-th moment implies finite r'-th moment for any $r'\in(1,r]$. Thus for $r<2$ assumption (2.4) is weaker than assuming finite variance, while for $r>2$ (2.4) is stronger than assuming finite variance. The next result describes the number of samples needed asymptotically to satisfy our tail bound conditions as a function of r.

Theorem 2.4. *Assume that* (2.4) *holds with* $r=\frac{q}{q-1}$. *If* $q\geq 2$, *then Assumption* 2.1 *can be satisfied using*

$$O\left(\Delta_k^{-q^2}\right) \tag{2.5}$$

i.i.d. samples, while if $q\in(1,2)$, *it can be satisfied using*

$$O\left(\Delta_k^{-2q}\right) \tag{2.6}$$

i.i.d. samples.

We thus have the following corollary illustrating an improvement on the number of samples per iteration with respect to the finite variance case.

Corollary 2.5. *Let $\varepsilon \in (0,1]$. Then, for $q = 1+\varepsilon$, $O(\Delta_k^{-2-\varepsilon})$ samples are sufficient to satisfy Assumption 2.1, under the finite moment assumption (2.4) for $r = \frac{q}{q-1}$.*

2.2.2. *Correlated errors.* We assume now that the objective is given in the form (1.4), and that we have access to the random number generator, that is we can sample different x with fixed ξ. Let now

$$\bar{F}(x,\xi) = F(x,\xi) - f(x)$$

be the sampling error. We assume that the sampling errors of close points are correlated in the following way:

$$\mathbb{E}_\xi[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^r] \leq D_r\|x-y\|^r \tag{2.7}$$

for some $D_r > 0$. The use of the term correlation to describe (2.7) can be better understood with the following lemma.

Proposition 2.6. *Assume that, for some $V, \bar{l} > 0$, we have*

$$\mathrm{Var}_\xi(F(x,\xi)) = V\,,$$

and

$$\mathrm{Cov}_\xi(F(x,\xi), F(y,\xi)) \geq V(1 - \bar{l}\|x-y\|^2)$$

for every $x, y \in \mathbb{R}^n$. Then (2.7) is satisfied for $D_r = 2V\bar{l}$ and $r = 2$.

We now show how the bound given in Theorem 2.4 improves under (2.7), for $r \geq 2$.

Theorem 2.7. *If the random number generator is known and (2.7) holds with $r = \frac{q}{q-1}$, then Assumption 2.1 can be obtained for $q \in (1,2]$ using*

$$O\left(\Delta_k^{2-2q}\right) \tag{2.8}$$

i.i.d. samples.

As a corollary we can state a further improvement in samples per iteration with respect to Corollary 2.5.

Corollary 2.8. *If $q = 1 + \frac{\varepsilon}{2}$ then $O(\Delta_k^{-\varepsilon})$ samples are sufficient to get Assumption 2.1 under (2.7) for $r = \frac{q}{q-1}$.*

We conclude with an example where (2.7) is satisfied, with the noise modelled as a Gaussian process, as it is common practice in Bayesian optimization (see, e.g., [25]).

Proposition 2.9. *Assume that $F(x, \xi)$ is a Gaussian process with expectation $f(x)$ and exponentiated quadratic kernel, so that in particular*

$$\mathrm{Cov}_{\xi}(F(x,\xi), F(y,\xi)) = \sigma^2 \exp\left(-\frac{\|x-y\|^2}{2l^2}\right) \tag{2.9}$$

for some $\sigma, l > 0$ and every $x, y \in \mathbb{R}^n$. Then assumption (2.7) is satisfied, for every $r \geq 2$ (with D_r depending on r).

2.3. Comparison with existing conditions. In this subsection, we compare our condition with others found in the literature. We will start by showing that our condition is weaker than the ones imposed in [2]. More precisely, it is implied by [2, Equation (2)], rewritten in our notation as

$$\begin{aligned} \mathbb{E}[|F_k^g - f(X_k + \Delta_k G_k)|^2 \mid \mathcal{F}_{k-1}] \leq k_f^2 \Delta_k^4 \\ \mathbb{E}[|F_k - f(X_k)|^2 \mid \mathcal{F}_{k-1}] \leq k_f^2 \Delta_k^4 \,, \end{aligned} \tag{2.10}$$

for a constant $k_f > 0$. The k_f-variance condition in (2.10) is a gradient free version of [23, Assumption 2.4, (iii)], and more precisely can be obtained from the latter by removing the gradient related terms in the right-hand side. It is important to note here that in [23] as well as in other works on smooth stochastic derivative free optimization (see, e.g., [8, 18, 26] and references therein), a probabilistically accurate gradient estimate is also used, while of course such an estimate is not available in a possibly non-smooth setting.

Proposition 2.10. *Condition (2.10) implies Assumption 2.1 for $\varepsilon_q = 4k_f^2$ and $q = 2$.*

The proof of the above result relies on the conditional Chebyshev's inequality (see the proof in the appendix for details).

Remark 2.11. In the algorithm proposed in [2] the direct-search direction at iteration k is chosen before the function estimates to be used in the acceptance test are computed. Thus our analysis can be extended also to that algorithm.

We now describe the relation between our assumption and the β-probabilistically accurate function estimate assumption

$$\mathbb{P}(\{|F_k - f(X_k)| \leq \tau_f \Delta_k^2\} \cap \{|F_k^g - f(X_k + \Delta_k G_k)| \leq \tau_f \Delta_k^2|\} \mid \mathcal{F}_{k-1}) \geq \beta \,, \tag{2.11}$$

used in [2, 8, 23] in combination with other assumptions. In particular, conditions (2.10) are used in [2] and [23] (as discussed above), and a probabilistic assumption on the accuracy of random models for the objective is considered in [8].

We show that if (2.11) is satisfied for every β in a certain interval, with τ_f depending on β and an accuracy parameter ε, then also our assumption is satisfied with ε_q dependent on ε.

Proposition 2.12. *Let $\varepsilon > 0$ and $\bar{p} \in (0, 1)$. Assume that (2.11) holds for every $\beta \in [1 - \bar{p}, 1)$, with $\tau_f = \tau_f(\beta) < \frac{1}{2}\sqrt{\frac{\varepsilon}{1-\beta}}$. Then Assumption 2.1 holds with $\varepsilon_q = \frac{\varepsilon}{\bar{p}}$ and $q = 2$.*

The proposition above follows from the inclusion

$$\begin{aligned} &\{|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| < \alpha \Delta_k^2\} \\ &\supset \{|F_k - f(X_k)| \leq \tau_f(\beta)\Delta_k^2\} \cap \{|F_k^g - f(X_k + \Delta_k G_k)| \leq \tau_f(\beta)\Delta_k^2\}, \end{aligned} \tag{2.12}$$

whenever $\tau_f(\beta) < \frac{\alpha}{2}$ (see the proof in the appendix for details).

3. A simple direct-search method for stochastic non-smooth functions

In this section, we first describe a simple stochastic direct-search algorithm for the unconstrained minimization problem given in (1.3), where f is possibly non-smooth, and then analyze its convergence.

3.1. The stochastic direct-search scheme. A detailed description of our stochastic direct-search method is given in Algorithm 1. At each iteration, we generate a direction g_k in the unitary sphere (independently of the estimates of the objective function generated so far; see Step 3), and perform a step along the direction g_k with stepsize δ_k. Then, at Step 4, we compute f_k^g and f_k, that is the estimate values of the function at the resulting trial point $x_k + \delta_k g_k$ and also at x_k. We then accept or reject the trial point based on a sufficient decrease condition, imposing that the improvement on the objective estimate at the trial point is at least $\theta \delta_k^q$. If the sufficient decrease condition is satisfied, we have a successful iteration. We hence update our iterate x_{k+1} by setting it equal to the trial point and expand or keep the same stepsize at Step 5. Otherwise, the iteration is unsuccessful, so we do not move (i.e., $x_{k+1} = x_k$) and shrink the stepsize (see Step 6).

Algorithm 1 Stochastic direct search

Initialization. Choose a point x_0, $\delta_0, \theta > 0$, $\tau \in (0,1)$, $\bar{\tau} \in [1, 1+\tau]$.
For $k = 0, 1 \dots$
    Select a direction g_k in the unitary sphere.
    Compute estimates f_k and f_k^g for f in x_k and $x_k + \delta_k g_k$.
    If $f_k - f_k^g \geq \theta \delta_k^q$, **Then** set `SUCCESS` $=$ `true`, $x_{k+1} = x_k + \delta_k g_k$, $\delta_{k+1} = \bar{\tau}\delta_k$.
    Else set `SUCCESS` $=$ `false`, $x_{k+1} = x_k$, $\delta_{k+1} = (1-\tau)\delta_k$.
    End if
End for

In order for the method to convergence to Clarke stationary points, the sequence $\{g_k\}$ must be dense in the unit sphere on certain subsequences (see Theorem 3.3). We remark that a dense sequence in the unit sphere can be generated using a suitable quasirandom sequence (see, e.g., [14, 19]).

3.2. Convergence analysis under the tail-bound probabilistic condition. The following theorem, which implies that the stepsize sequence $\{\Delta_k\}$ converges to zero almost surely, is a key result in the convergence analysis. By taking a look at the proof, we can see how the use of the tail-bound probabilistic condition (AS1) allows us to give a unified argument for unsuccessful and successful steps.

We define now for convenience the positive constants $\tau_q^+ = (1+\tau)^q - 1$, $\tau_q^- = 1 - (1-\tau)^q$, and $\bar{\tau}_q = \tau_q^+ + \tau_q^-$. To obtain our result we need the following lower bound on the parameter θ defining the sufficient decrease condition, dependent on the stepsize update parameter τ and the tail bound parameter ε_q:

$$\theta > \frac{\sqrt[r(q)]{\varepsilon_q}\,\bar{\tau}_q}{\tau_q^-}\,, \tag{3.1}$$

with $r(q) = \frac{q}{q-1}$. Notice that since $\tau \in (0,1)$ we must always have $\theta > 0$. The bound (3.1) allows us to relate stepsize expansions to improvements of the objective.

Theorem 3.1. *Under Assumption 2.1, if Inequality (3.1) holds then*

$$\sum_{k \in \mathbb{N}_0} \mathbb{E}[\Delta_k^q] < \infty \tag{3.2}$$

a.s. in Ω.

Proof: Let $\varepsilon_f = \sqrt[r(q)]{\varepsilon_q}$, $\Phi_k = f(X_k) - f^* + \eta\Delta_k^q$, with $\eta = \frac{\theta}{\bar{\tau}_q}$, and

$$\varepsilon = -\varepsilon_f + \frac{\tau_q^- \theta}{\bar{\tau}_q} > 0\,,$$

where the inequality follows by (3.1).

We will prove, for every $k \geq 0$, that

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}] \geq \varepsilon \Delta_k^q\,. \tag{3.3}$$

The thesis then follows as in [2, Theorem 1] (or directly by Robbins-Siegmund Theorem [24]).

Let ρ_k be the random variable such that $f(X_k) - f(X_k + \Delta_k G_k) = (\theta - \rho_k)\Delta_k^q$, and let J_k be the event that the step k is successful. We have

$$\begin{aligned}
&\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] = \mathbb{E}[(\Phi_k - \Phi_{k+1})(\mathbb{1}_{J_k} + (1 - \mathbb{1}_{J_k}))|\mathcal{F}_{k-1}] \\
&= (f(X_k) - f(X_{k+1}) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad + (f(X_k) - f(X_{k+1}) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}[\,1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&= (f(X_k) - f(X_k + \Delta_k G_k) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad + \eta(\Delta_k^q - \Delta_{k+1}^q)\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\geq (((\theta - \rho_k) - \eta\tau_q^+)\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^- \mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}])\Delta_k^q,
\end{aligned} \tag{3.4}$$

where we used $X_k = X_{k+1}$ for unsuccessful steps in the second equality, and $\Delta_{k+1} = \bar{\tau}\Delta_k \leq (1+\tau)\Delta_k$ for successful steps in the inequality. In turn,

$$\begin{aligned}
&(((\theta - \rho_k) - \eta\tau_q^+)\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^- \mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}])\Delta_k^q \\
&= ((\theta - \rho_k - \eta\bar{\tau}_q)\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-)\Delta_k^q \\
&= -\rho_k\Delta_k^q\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-\Delta_k^q\,,
\end{aligned} \tag{3.5}$$

where we used $\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = 1 - \mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}]$ in the first equality, and $\theta = \eta\bar{\tau}_q$ in the second one. By combining (3.4) and (3.5) we can therefore conclude

$$\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] \geq -\rho_k\Delta_k^q\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-\Delta_k^q\,. \tag{3.6}$$

Notice that if the step is successful then $f_k - f_k^g \geq \theta\delta_k^q$, which implies

$$f_k - f_k^g - (f(x_k) - f(x_k + \delta_k g_k)) \geq \theta\delta_k^q - (\theta - \rho_k(w))\delta_k^q = \rho_k(w)\delta_k^q\,.$$

In particular

$$J_k \subset \{|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \rho_k\Delta_k^q\}\,,$$

and we can write, for $\rho_k^+ = \rho_k \mathbb{1}_{\rho_k>0}$,

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = \mathbb{E}[\mathbb{1}_{J_k}\mathbb{1}_{\{\rho_k>0\}} + \mathbb{1}_{J_k}\mathbb{1}_{\{\rho_k\le 0\}}|\mathcal{F}_{k-1}] \\
&= \mathbb{E}[\mathbb{1}_{J_k\cap\{\rho_k>0\}}|\mathcal{F}_{k-1}] + \mathbb{1}_{\{\rho_k\le 0\}}\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\le \mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \ge \rho_k^+\Delta_k^q|\mathcal{F}_{k-1}) \\
&\quad + \mathbb{1}_{\{\rho_k\le 0\}}\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}]\,,
\end{aligned} \tag{3.7}$$

where we used the measurability of ρ_k w.r.t. $\mathcal{F}_{k-1}$ in the second equality. We now have

$$\begin{aligned}
&-\rho_k\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \ge -\rho_k^+\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\ge -\rho_k^+(\mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \ge \rho_k^+\Delta_k^q|\mathcal{F}_{k-1}) \\
&\quad + \mathbb{1}_{\{\rho_k\le 0\}}\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}]) \\
&= -\rho_k^+\mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \ge \rho_k^+\Delta_k^q|\mathcal{F}_{k-1}) \\
&\ge -\rho_k^+\min\left(1, \frac{\varepsilon_q}{(\rho_k^+)^{r(q)}}\right) = -\rho_k^+\min\left(1, \frac{\varepsilon_f^{r(q)}}{(\rho_k^+)^{r(q)}}\right) \ge -\rho_k^+\min\left(1, \frac{\varepsilon_f}{\rho_k^+}\right) \\
&\ge -\varepsilon_f\,,
\end{aligned} \tag{3.8}$$

where we applied (3.7) in the first inequality, and the second inequality is a direct consequence of (AS1) for $\alpha = \rho_k^+$. Hence,

$$-\rho_k\Delta_k^q\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\tau_q^-\Delta_k^q \ge (-\varepsilon_f + \eta\tau_q^-)\Delta_k^q = \varepsilon\Delta_k^q\,, \tag{3.9}$$

where we used (3.8) in the inequality.

Claim (3.3) can finally be obtained by concatenating (3.6) and (3.9). ■

The lemma we now state will be useful for the proof of the optimality result of Theorem 3.3 which is based on the Clarke generalized directional derivative. We notice that Assumption 2.1 plays a key role in this result, allowing us to upper bound the error of the reduction estimate by a quantity that depends on the stepsize Δ_k.

Lemma 3.2. *Let K be the set of indices of unsuccessful iterations (notice that K is random). Then under Assumption 2.1 and (3.1) we have a.s. in Ω*

$$\liminf_{k\in K,\, k\to\infty} \frac{f(X_k+\Delta_k G_k) - f(X_k)}{\Delta_k} \ge 0\,. \tag{3.10}$$

Proof: Clearly it suffices to show that, for any given $m \in \mathbb{N}$ and a.s.,

$$\liminf_{k\in K,\, k\to\infty} \frac{f(X_k+\Delta_k G_k)-f(X_k)}{\Delta_k} \geq -\frac{1}{m}\,. \tag{3.11}$$

To start with, by applying (AS1) with $\alpha = \frac{\Delta_k^{1-q}}{m}$ we have

$$\mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \geq \frac{\Delta_k}{m} \mid \mathcal{F}_{k-1}) \leq m^{r(q)}\Delta_k^q \varepsilon_q\,,$$

and therefore taking expectations on both sides

$$\mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \geq \frac{\Delta_k}{m}) \leq m^{r(q)}\mathbb{E}[\Delta_k^q]\varepsilon_q\,.$$

We can now deduce

$$\sum_{k\in\mathbb{N}_0} \mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \geq \frac{\Delta_k}{m}) \leq \sum_{k\in\mathbb{N}_0} m^{r(q)}\mathbb{E}[\Delta_k^q]\varepsilon_q < \infty\,,$$

where we applied Theorem 3.1 in the last inequality. In particular, by the Borel-Cantelli's first lemma

$$\mathbb{P}\left(\left\{|F_k - f_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \geq \frac{\Delta_k}{m}\right\} \text{ i.o.}\right) = 0\,,$$

where "i.o." stands for *infinitely often*. Hence, we have a.s.

$$|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \leq \frac{\Delta_k}{m} \quad \text{for } k \text{ large enough.} \tag{3.12}$$

From this we can infer that a.s., for every $k \in K$ large enough

$$\frac{f(X_k+\Delta_k G_k)-f(X_k)}{\Delta_k} \geq \frac{F_k^g - F_k - |F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))|}{\Delta_k}$$
$$\geq -\theta\Delta_k - \frac{1}{m}\,, \tag{3.13}$$

where we used (3.12) combined with the unsuccessful step condition of Algorithm 1 in the second inequality. Finally, (3.11) follows passing to the liminf for $k \to \infty$ in (3.13). ■

We now report the main convergence result for our stochastic direct-search scheme. We refer to $\mathcal{V}$ as the event with probability one that (3.10) holds.

Theorem 3.3. *Assume that f is Lipschitz continuous with constant L_f^* around any limit point of the sequence of iterates $\{X_k\}$. Let K be the set of indices of unsuccessful iterations. Let Assumptions 2.1 and (3.1) hold.*

Then, the following property holds a.s. in Ω: if $L \subset K$ (notice that L, K are random) is such that $\{G_k\}_{k\in L}$ is dense in the unit sphere and

$$\lim_{k\in L,\, k\to\infty} X_k = X^*\,,$$

then the point X^* is Clarke stationary.

Proof: Let d be a direction in the unitary sphere, and let $S \subset L$ be such that

$$\lim_{k\in S,\, k\to\infty} G_k = d\,.$$

By definition of Clarke stationarity, we just need to prove that on $\mathcal{V}$ (and therefore a.s.)

$$\limsup_{k\in S,\, k\to\infty} \frac{f(X_k+\Delta_k d)-f(X_k)}{\Delta_k} \geq 0\,.$$

For $w \in \mathcal{V}$ we can write

$$\limsup_{k\in S,\, k\to\infty} \frac{f(X_k+\Delta_k G_k)-f(X_k)}{\Delta_k} \geq \liminf_{k\in K,\, k\to\infty} \frac{f(X_k+\Delta_k G_k)-f(X_k)}{\Delta_k} \geq 0\,, \tag{3.14}$$

where the last inequality follows by (3.10).

Now using the Lipschitz property of f we can write, for $k \in S(w)$ large enough,

$$\begin{aligned}\frac{f(X_k+\Delta_k d)-f(X_k)}{\Delta_k} =& \frac{f(X_k+\Delta_k G_k)-f(X_k)}{\Delta_k} \\ &+ \frac{f(X_k+\Delta_k d)-f(X_k+\Delta_k G_k)}{\Delta_k} \\ \geq& \frac{f(X_k+\Delta_k G_k)-f(X_k)}{\Delta_k} - L_f^*\|G_k-d\|.\end{aligned}$$

Passing to the limsup for $k \in S \subset L$ we get

$$\limsup_{k\in S,\ k\to\infty} \frac{f(X_k+\Delta_k d)-f(X_k)}{\Delta_k} \geq \limsup_{k\in S,\ k\to\infty} \frac{f(X_k+\Delta_k G_k)-f(X_k)}{\Delta_k} \geq 0\,,$$

for every $w \in \mathcal{V}$, where we used $\|G_k - d\| \to 0$ by construction in the first inequality and (3.14) in the second. ∎

4. A simple trust-region method for stochastic non-smooth functions

After having analyzed a simple stochastic direct-search method, we focus on a stochastic version of the Basic DFO-TRNS presented in [19], and analyze its convergence properties under tail-bound probabilistic conditions like the ones used in Section 3. Some minor changes in notation are convenient and will be introduced with a clear reference to the corresponding elements of Algorithm 1.

4.1. The stochastic trust-region scheme. As already mentioned, the simple trust-region algorithm that we report here is a minor modification of the Basic DFO-TRNS algorithm proposed in [19]. Indeed, there are two differences between the Basic DFO-TRNS algorithm and its stochastic counterpart.

The first difference is in the updating rule related to the trust-region radius. In our modification, we choose $\tau \in (0, 1)$ and then choose $1-\tau$ as contraction factor and $\bar{\tau} \in [1, 1+\tau]$ as expansion factor.

The second, more relevant difference is the fact that the linear term g_k is not constrained to the unit sphere. This makes more sense when modeling cases where g_k resembles an approximation of the gradient.

The detailed scheme is reported in Algorithm 2. At every iteration k, a symmetric matrix B_k is built from interpolation or regression on a sample set of points. The linear term g_k needs to randomly cover the unit sphere when normalized. By using these quantities, a quadratic model of the objective function around x_k is built. The step s_k is obtained by solving the trust-region subproblem, i.e., by minimizing the quadratic model within the spherical trust-region constraint. Once the current step has been computed, we estimate the true objective function f at the trial point x_k+s_k and recompute a new estimate at x_k, after which we compute the acceptance ratio $\bar{\rho}_k$. Note that, as in [19], the non-standard acceptance ratio is motivated by convergence requirements. In this scheme, realizations related to the estimate of the function value at the current iterate $f(x_k)$ and at the potential next iterate $f(x_k+s_k)$ are indicated with f_k and f_k^s (thus replacing f_k^g used in the direct-search scheme) as a shorthand for $F_k(w)$ and $F_k^s(w)$, respectively.

For convergence purposes, we require the Hessian model to satisfy the assumption below.

Algorithm 2 `Stochastic DFO Trust-Region Algorithm`

Initialization. Select $x_0 \in \mathbb{R}^n$, $\theta > 0$, $\tau \in (0,1)$, $\bar{\tau} \in [1, 1+\tau]$, $\delta_0 > 0$, $q > 1$.
For $k = 0, 1 \dots$
  Select a direction $g_k \neq 0$ and build a symmetric matrix B_k.
  Compute
$$s_k \in \arg\min_{\|s\| \leq \delta_k} g_k^\top s + \frac{1}{2} s^\top B_k s. \tag{4.1}$$
  Compute estimates f_k, f_k^s for f at x_k, $x_k + \delta_k$, respectively, and let
$$\bar{\rho}_k = \frac{f_k - f_k^s}{\theta \|s_k\|^q}.$$
  If $\bar{\rho}_k \geq 1$ **Then** set `SUCCESS` $=$ `true`, $x_{k+1} = x_k + s_k$, $\delta_{k+1} = \bar{\tau}\delta_k$.
  Else set `SUCCESS` $=$ `false`, $x_{k+1} = x_k$, $\delta_{k+1} = (1-\tau)\delta_k$.
  End If
End For

Assumption 4.1. *There exist $\rho \in (0,1]$ such that, for every $k \in \mathbb{N}_0$,*
$$\|B_k\| \leq \frac{1}{\rho} \frac{\|G_k\|}{\Delta_k}.$$

When $\|G_k\| = 1$, the above assumption is essentially saying that B_k can be unbounded as long at it does not go to infinity faster than $1/\Delta_k$.

We now show, under Assumption 4.1, that every trust-region subproblem solution S_k has norm equal to Δ_k, up to a constant. This will allow us to deduce convergence to 0 of the trust-region radius from convergence to 0 of the solution norm.

Lemma 4.2. *Under Assumption 4.1 we have*
$$\|S_k\| \geq \rho \Delta_k. \tag{4.2}$$

Proof: The thesis is clear if S_k is on the boundary of the trust region, which includes the case $B_k = 0$ since $G_k \neq 0$ by assumption. Otherwise, if S_k is in the interior we must have
$$B_k S_k = -G_k,$$
and therefore
$$\|B_k\| \|S_k\| \geq \|G_k\| \geq \rho \Delta_k \|B_k\|,$$
where we used (4.1) in the second inequality, and the proof is completed. ∎

4.2. Convergence analysis under the tail-bound probabilistic condition. In order to analyze the method introduced above, we adapt Assumption 2.1, replacing G_k with $\hat{S}_k$ and Δ_k with $\|S_k\|$. Now Δ_k stands for the trust-region radius. Hence, we obtain the following tail-bound condition.

Assumption 4.3. *For some $\varepsilon_q > 0$ (independent of k):*

$$\mathbb{P}\left(|F_k - F_k^g - (f(X_k) - f(X_k + S_k))| \geq \alpha\|S_k\|^q \;|\mathcal{F}_{k-1}\right) \leq \frac{\varepsilon_q}{\alpha^{q/(q-1)}}\,, \qquad \text{(A2')}$$

a.s. for every $\alpha > 0$.

The next theorem states convergence of the series of trust-region radii elevated to the q almost surely. This obviously implies that the trust-region radius converges to zero almost surely.

Theorem 4.4. *Under Assumptions 4.1 and 4.3, if*

$$\theta > \frac{(\rho^q \tau_q^- + \tau_q^+)\sqrt[r(q)]{\varepsilon_q}}{\rho^q \tau_q^-}\,, \tag{4.3}$$

then

$$\sum_{k\in\mathbb{N}_0} \mathbb{E}[\Delta_k^q] < \infty$$

a.s. in Ω.

Proof: Reasoning along the lines of Theorem 3.1, using the merit function $\Phi_k = f(X_k) - F^* + \eta\|S_k\|^q$, with $\eta = \frac{\theta\rho^q}{\tau_q^+ + \rho^q\tau_q^-}$, we obtain

$$\sum_{k\in\mathbb{N}_0} \mathbb{E}[\|S_k\|^q] < +\infty\,,$$

and therefore

$$\sum_{k\in\mathbb{N}_0} \mathbb{E}[\Delta_k^q] \leq \frac{1}{\rho^q}\sum_{k\in\mathbb{N}_0} \mathbb{E}[\|S_k\|^q] < +\infty\,,$$

where we used (4.2) in the inequality. ■

As for the analysis of our direct-search scheme in Section 3, we now state a lemma that will be useful for the proof of the optimality result based on the Clarke generalized derivative.

Lemma 4.5. *Let K be the set of indices of unsuccessful iterations (notice that K is random). Then under Assumptions 4.1, 4.3, and (4.3) we have a.s.*

$$\liminf_{k\in K,\, k\to\infty} \frac{f(X_k + S_k) - f(X_k)}{\|S_k\|} \geq 0\,.$$

Proof: Follows analogously to Lemma 3.2. ■

We now state a convergence result generalizing Theorem 3.3 to our trust-region method.

Theorem 4.6. *Assume that f is Lipschitz continuous with constant L_f^* around any limit point of the sequence of iterates $\{X_k\}$. Let K be the set of indices of unsuccessful iterations. Let Assumptions 4.1, 4.3, and (4.3) hold. Then, the following property holds a.s. in Ω: if $L \subset K$ (notice that L, K are random) is such that $\{\hat{S}_k\}_{k\in L}$ is dense in the unit sphere and*

$$\lim_{k\in L,\, k\to\infty} X_k = X^*\,,$$

then the point X^* is Clarke stationary.

Proof: The proof follows the lines of Theorem 3.3's proof, replacing Δ_k, G_k, $\Delta_k G_k$ by respectively $\|S_k\|, \hat{S}_k, S_k$. ■

We now introduce a stronger version of Assumption 4.1, and show that under this stronger assumption the trust-region scheme becomes at the limit a search along a direction G_k with stepsize Δ_k.

Assumption 4.7. *For some positive sequence $\{a_k\}$ such that $a_k \to 0$*

$$\|B_k\| \leq a_k \frac{\|G_k\|}{\Delta_k}\,. \tag{4.4}$$

Trivially, Assumption 4.7 implies Assumption 4.1, with $\rho = \frac{1}{\max(\{a_k\})}$.

Proposition 4.8. *Let Assumptions 4.3, 4.7, and (4.3) hold. Then we have a.s.*

$$\lim_{k\to\infty} \hat{G}_k + \hat{S}_k = 0\,.$$

Proof: First, notice that $\|\hat{G}_k\| = 1$, as well as $\|\hat{S}_k\| = 1$ since G_k must be always different from 0 and therefore S_k as well. Now define F_k^m as the local model:

$$F_k^m(s) = G_k^\top s + \frac{1}{2} s^\top B_k s\,,$$

and let $\gamma_k = \hat{G}_k^\top \hat{S}_k$ be the cosine of the angle between $\hat{G}_k$ and $\hat{S}_k$. We need to prove $\gamma_k \to -1$ (a.s.).

We have on the one hand

$$\begin{aligned} F_k^m(S_k) &= S_k^\top G_k + \frac{1}{2} S_k^\top B_k S_k = \gamma_k \|S_k\| \|G_k\| + \frac{1}{2} S_k^\top B_k S_k \\ &\geq \min(0, \gamma_k) \Delta_k \|G_k\| - \frac{1}{2} \|B_k\| \Delta_k^2 \,, \end{aligned} \tag{4.5}$$

where we used $\|S_k\| \leq \Delta_k$ in the inequality. On the other hand

$$F_k^m(-\Delta_k \hat{G}_k) = -\Delta_k \|G_k\| + \frac{\Delta_k^2}{2} \hat{G}_k^\top B_k \hat{G}_k \leq -\Delta_k \|G_k\| + \frac{1}{2} \Delta_k^2 \|B_k\| \,. \tag{4.6}$$

Putting (4.5) and (4.6) together we obtain

$$\begin{aligned} -\Delta_k \|G_k\| + \frac{1}{2} \Delta_k^2 \|B_k\| \geq F_k^m(-\Delta_k \hat{G}_k) &\geq F_k^m(S_k) \\ &\geq \min(0, \gamma_k) \Delta_k \|G_k\| - \frac{1}{2} \|B_k\| \Delta_k^2 \,, \end{aligned} \tag{4.7}$$

where in the second inequality we used that S_k is a solution of the trust-region subproblem. Then rearranging (4.7) and dividing by $\Delta_k \|G_k\|$ we get

$$(1 + \min(0, \gamma_k)) \leq \frac{\|B_k\| \Delta_k}{\|G_k\|} \,. \tag{4.8}$$

Since the right-hand side of (4.8) converges to 0 a.s. by Theorem 4.4, we get $1 + \min(0, \gamma_k) \to 0$ a.s., and we can conclude $\gamma_k \to -1$ a.s. as desired. ∎

Under the conditions of Proposition 4.8, we just need to ensure that $\hat{G}_k$ is dense in the unit sphere on subsequences to obtain convergence to Clarke stationary points, as expressed in the following corollary.

Corollary 4.9. *Assume that f is Lipschitz continuous with constant L_f^* around any limit point of the sequence of iterates $\{X_k\}$. Let K be the set of indices of unsuccessful iterations. Let Assumption 4.3, 4.7 and (4.3) hold. Then, the following property holds a.s. in Ω: if $L \subset K$ (notice that L, K are random) is such that $\{\hat{G}_k\}_{k \in L}$ is dense in the unit sphere and*

$$\lim_{k \in L,\, k \to \infty} X_k = X^* \,,$$

then the point X^* is Clarke stationary.

Proof: Thanks to Proposition 4.8 since $\{\hat{G}_k\}_{k \in L}$ is dense in the unit sphere $\{\hat{S}_k\}_{k \in L}$ also is, and we can therefore apply Theorem 4.6. ∎

5. Numerical results

We describe in this section some numerical results comparing two instances of Algorithm 1 for different choices of the sufficient decrease parameter and sampling strategies, corresponding to different values of q and r in the algorithmic scheme and in the assumptions. Our main goal is to show that the theoretical improvement from $O(\Delta_k^{-4})$ samples per iteration to $O(\Delta_k^{-2-\varepsilon})$, proved for a suitable choice of q and r, also leads to an experimental improvement on the total number of samples needed to find a good solution. We then show that an analogous experimental improvement can be observed in the case of correlated errors discussed in Section 2.2.2. Focusing on Algorithm 1 rather than Algorithm 2 allows us to only count function samples necessary to satisfy the weak tail bound conditions introduced in this paper, rather than having to take into account the samples used to build the trust-region model as well.

Remark 5.1. It is of course not always the case that an improvement in number of samples per iteration leads to an improvement in the solution found with a fixed budget of samples, since using lower values of q might increase the iteration complexity. For instance, for smooth objectives with deterministic oracles a complexity of $O(\epsilon^{-\frac{q}{q-1}})$ was proved in [27] for a scheme analogous to Algorithm 1, with $q \in (1, 2]$. Then in this case the lower number of samples per iteration for q approaching 1 comes at the price of a potentially much higher iteration complexity. However, it is important to note that the estimates from [27] heavily rely on the Lipschitz continuity of the gradient, so that this trade-off does not necessarily generalize to potentially non smooth objectives.

To compare the performance of the two algorithms, we will use data and performance profiles as defined in [22]. We briefly recall here their definitions. Given a set S of algorithms and a set P of problems, for $s \in S$ and $p \in P$, let $t_{p,s}$ be the number of function evaluations required by algorithm s on problem p to satisfy the condition

$$f(x_k) \leq f_L + \gamma_p(f(x_0) - f_L)\,, \tag{5.1}$$

where $\gamma_p \in (0, 1)$ and f_L is the best objective function value achieved by any solver on problem p. Then, the performance and data profiles of solver s are defined by

$$\rho_s(\alpha) = \frac{1}{|P|}\left|\left\{p \in P : \frac{t_{p,s}}{\min\{t_{p,s'} : s' \in S\}} \leq \alpha\right\}\right|,$$
$$d_s(\kappa) = \frac{1}{|P|}\left|\{p \in P : t_{p,s} \leq \kappa(n_p+1)\}\right|,$$

where n_p is the dimension of problem p. We used a budget of $10000(n_p+1)$ sample evaluations for both algorithms and two different accuracies for (5.1), that is $\gamma_p \in \{10^{-2}, 10^{-4}\}$. All the profiles are built with the true function values rather than the noisy estimates used in the algorithms. The set P includes 96 well known instances of derivative-free unconstrained nonsmooth optimization problems reported in Table 1 with dimensions and references. Each of the instances is used 10 times, so that both algorithms perform 10 runs on every instance, thus getting $|P| = 960$.

We tuned the parameters with a basic grid search to obtain good performances for both instances of Algorithm 1 to $\tau = 0.001$, $\bar{\tau} = 1.001$, $\theta = 0.5$ and $\delta_0 = 2$ (see Section A.2). We set $q = 2$ and $q = 1.5$ for the first and the second instance, thus comparing a standard choice (see, e.g., [2, 12, 13]) to one that allows us to use a lower number of samples per iteration as proved in Theorems 2.4 and 2.7. We simulate the process of sampling p_k noisy function estimates by adding Gaussian noise with 0 mean and $1/p_k$ variance.

Remark 5.2. It is not difficult to check that the bound (3.1) translates to $\theta > c$ with $c = 4$ and $c \approx 9$ for $q = 2$ and $q = 1.5$ respectively. However, both algorithms show bad relative performance for θ greater than 1 (see Figure 3). We conjecture that lower values of θ and therefore a more tolerant acceptance test might still lead to convergence in practice in most cases, with a lower number of samples needed to find a good solution due to the resulting more aggressive exploration. Finding weaker versions of (3.1) that still guarantee convergence under reasonable assumptions remains of course an open problem to be studied more in depth in future works.

We first deal with the general case of finite r-th moment and uncorrelated errors (Figure 1). We refer in this case to the two instances of Algorithm 1 as SDSq for $q \in \{2, 1.5\}$. By Theorem 2.4, we have that $O\left(\Delta_k^{-2q}\right)$ samples are needed to satisfy the weak tail bound assumptions. Given that the Gaussian noise has finite r-th moment for every r, we can apply Theorem 2.4 with $r = q/(q-1)$. The number of samples needed per iteration is then $O(\Delta_k^{-4})$

and $O(\Delta_k^{-3})$ respectively for SDS2 and SDS1.5. We thus set $p_k = \lceil \delta_k^{-4} \rceil$ and $p_k = \lceil \delta_k^{-3} \rceil$ in our sampling simulation.

We then consider the case of correlated errors (Figure 2). We refer in this case to the two instances of Algorithm 1 as SDSqc for $q \in \{2, 1.5\}$. By using Theorem 2.7 and reasoning as for the general case, we obtain that $O(\Delta_k^{-2})$ and $O(\Delta_k^{-1})$ samples are needed per iteration to satisfy the weak tail bound assumptions. In the implementation we thus set $p_k = \lceil \delta_k^{-2} \rceil$ and $p_k = \lceil \delta_k^{-1} \rceil$.

By taking a look at the profiles, we can easily see that SDS1.5 and SDS1.5c outperform SDS2 and SDS2c, respectively. We can then conclude for both the general and the correlated error case, using $q = 1.5$ (and consequently fewer samples per iteration) gives better performances in the end.

TABLE 1. Problems used in numerical experiments.

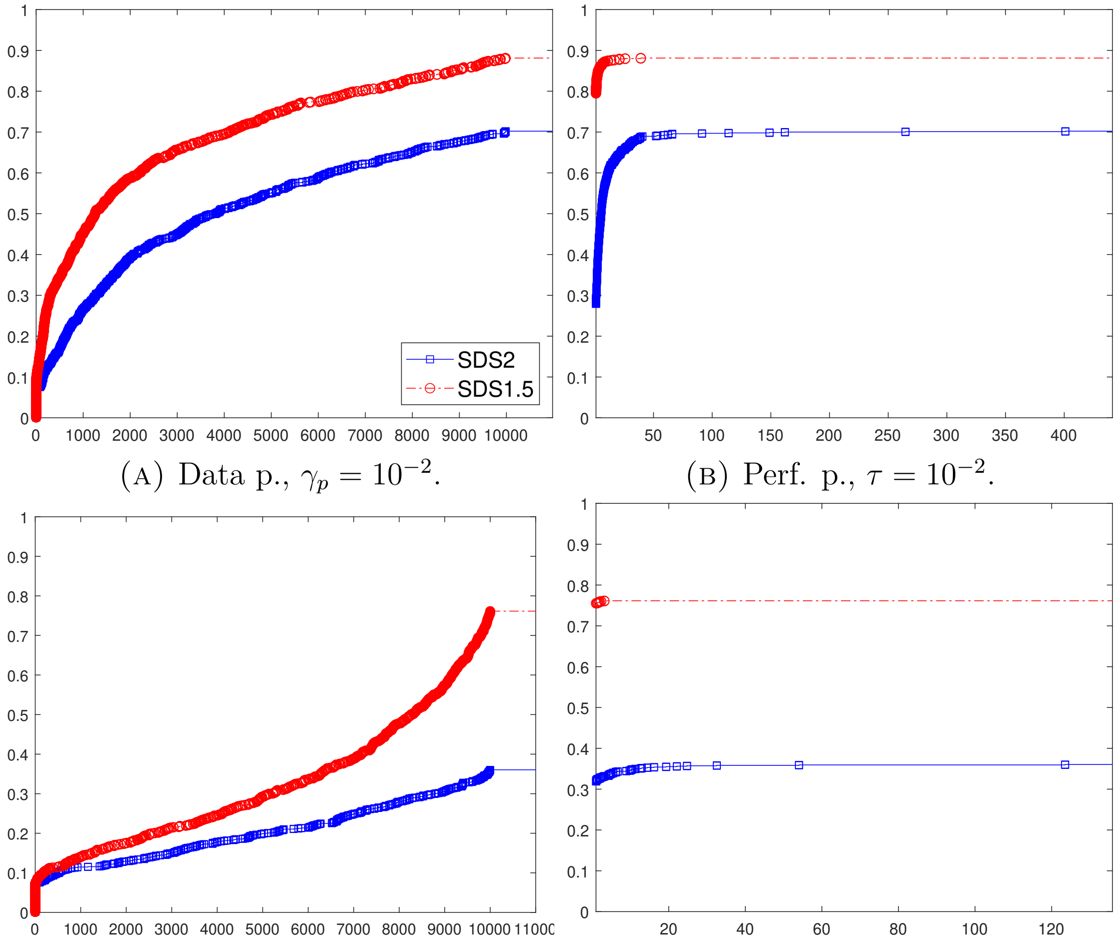
name	dimension	reference
crescent	2	[22]
cb2	2	[21]
charconn1	2	[20]
charconn2	2	[20]
demyanov-malozemov	2	[20]
dennis-woods	2	[11]
wong1	7	[21]
wong2	10	[21]
wong3	20	[21]
elattar	6	[21]
goffin	50	[21]
hald-madsen 1	2	[20]
lq	2	[21]
ql	2	[21]
maxl	20	[21]
maxq	20	[22]
mifflin 1	2	[16]
mifflin 2	2	[16]
rosen-suzuki	4	[21]
wf	2	[21]
spiral	2	[21]
evd 52	3	[21]
kowalik-osborne	4	[21]
oet 5	4	[21]
oet 6	4	[21]
gamma	4	[21]
exp	5	[21]
pbc1	5	[21]
evd61	6	[21]
filter	9	[21]
polak 2	10	[21]
polak 3	11	[21]
polak 6	4	[21]
watson	20	[21]
osborne 2	11	[21]

shor	5	[21]
colville 1	5	[21]
hs 78	5	[21]
maxquad	10	[21]
gill	10	[21]
mxhilb	50	[16]
l1hilb	50	[21]
davidon 2	4	[21]
shelldual	15	[21]
steiner 2	12	[21]
transformer	6	[21]
polak 6.10	1	[21]
wild1	20	[16]
wild2	20	[16]
wild3	20	[16]
wild19	20	[16]
wild11	20	[16]
wild16	20	[16]
wild20	20	[16]
wild15	20	[16]
wild21	20	[16]
maxq	{10, 20, 30, 40}	[16]
l1hilb	{10, 20, 30, 40}	[22]
lq	{10, 20, 30, 40}	[22]
cb3	{10, 20, 30, 40}	[22]
cb32	{10, 20, 30, 40}	[22]
af	{10, 20, 30, 40}	[22]
brown	{10, 20, 30, 40}	[22]
mifflin2	{10, 20, 30, 40}	[22]
crescent	{10, 20, 30, 40}	[22]
crescent2	{10, 20, 30, 40}	[22]

6. Concluding remarks and future work

This paper proposed a new tail-bound condition for function estimation in stochastic derivative-free optimization, provably weaker than probabilistic conditions appearing in previous works. We showed how this condition can be obtained under a finite moment assumption on the black box noise, generalizing finite variance. This naturally led to defining a trade-off between noise moment and number of samples per iteration, generalizing the classic $O(\Delta_k^{-4})$ sample bound of the finite variance case, with improvements for higher moments.

Our tail bound assumption allowed us to obtain convergence of both a direct-search and a trust-region method. Surprisingly, unlike in previous



(A) Data p., $\gamma_p = 10^{-2}$.

(B) Perf. p., $\tau = 10^{-2}$.

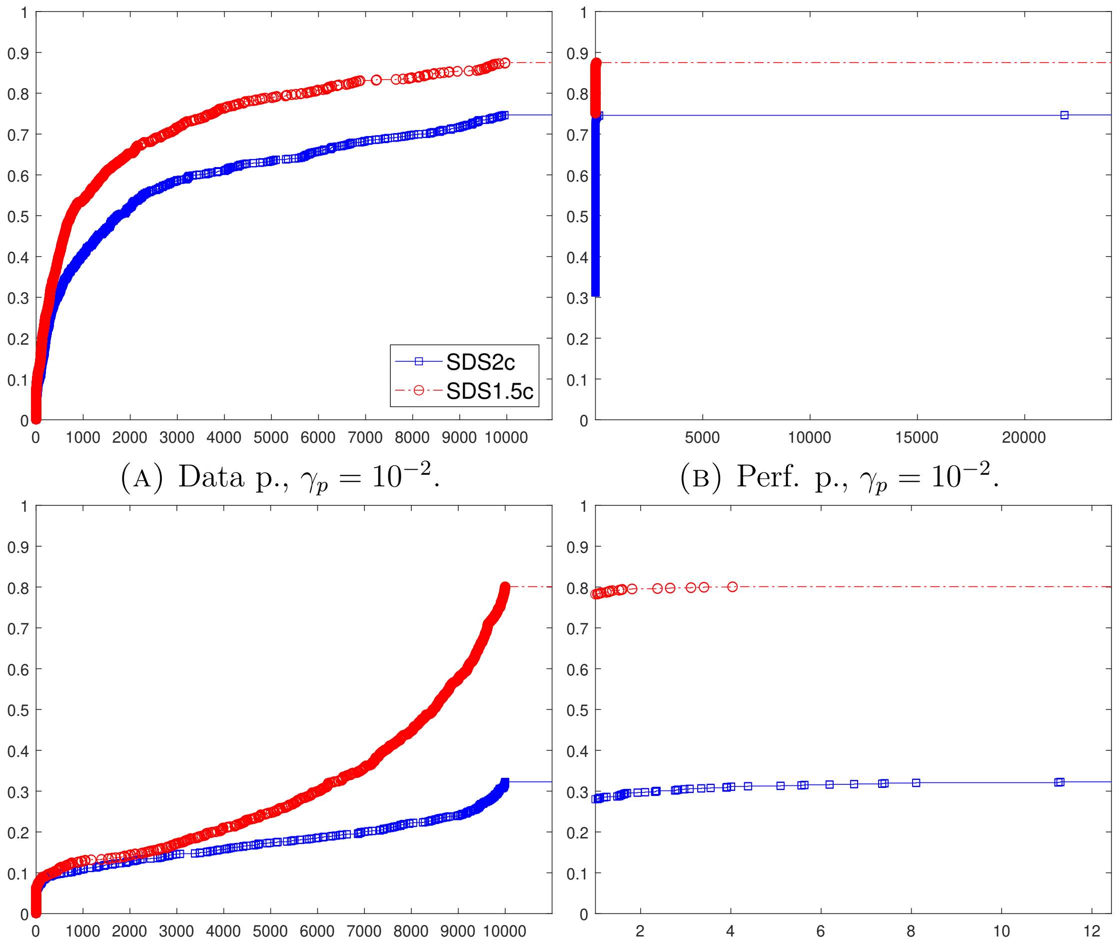
(C) Data p., $\gamma_p = 10^{-4}$.

(D) Perf. p., $\gamma_p = 10^{-4}$.

FIGURE 1. Profiles for SDS2 and SDS1.5.

works on stochastic DFO requiring multiple probabilistic conditions for convergence, in this work a single tail-bound is sufficient to prove that the sequence of stepsizes/radii converges to 0, and to conclude convergence to Clarke stationary points.

There are a few future research developments. A first one is the analysis of trust-region algorithms based on non-smooth random local models under the new conditions. Possible choices of the model include piecewise linear models and random smooth functions like the ones used in Bayesian optimization. Studying tailored models for special cases where the objective is the non smooth composition of smooth functions (like for instance the maximum of smooth functions) is a related challenge. Other possible research topics include the extension of our analysis to the constrained case, its integration



(A) Data p., $\gamma_p = 10^{-2}$. (B) Perf. p., $\gamma_p = 10^{-2}$.

(C) Data p., $\gamma_p = 10^{-4}$. (D) Perf. p., $\gamma_p = 10^{-4}$.

FIGURE 2. Profiles for SDS2c and SDS1.5c.

within global optimization schemes, and numerical tests for the trust-region scheme.

Appendix A. Appendix

A.1. Proofs. In this appendix, we report the missing proofs. We first recall Rosenthal inequality (see, e.g., [15]), together with a corollary useful for several results. This inequality states that, for $\{Z_i\}_{i\in[1:p]}$ independent and with 0 mean and finite r-th moment, $r \geq 2$, and $S = \frac{1}{p}\sum_{i=1}^{p} Z_i$, one has

$$\mathbb{E}[|S|^r] \leq p^{-r} C(r) \max\left(\sum_{i=1}^{p} \mathbb{E}[|Z_i|^r], \left(\sum_{i=1}^{p} \mathbb{E}\left[|Z_i|^2\right]\right)^{\frac{r}{2}}\right). \tag{A.1}$$

Note that when the $\{Z_i\}$ are i.i.d. and having the same distribution of a certain random variable Z, (A.1) reduces to

$$\mathbb{E}[|S|^r] \leq p^{-r}C(r)\max\left(pE[|Z|^r], p^{\frac{r}{2}}\mathbb{E}\left[|Z|^2\right]^{\frac{r}{2}}\right). \tag{A.2}$$

Now, we have

$$\begin{aligned} &p^{-r}C(r)\max\left(pE[|Z|^r], p^{\frac{r}{2}}\mathbb{E}\left[|Z|^2\right]^{\frac{r}{2}}\right) \\ &\leq C(r)p^{-r}\max\left(p\mathbb{E}[|Z|^r], p^{\frac{r}{2}}\mathbb{E}[|Z|^r]\right) \leq C(r)p^{-\frac{r}{2}}\mathbb{E}[|Z|^r], \end{aligned} \tag{A.3}$$

where we used Jensen's inequality on the second argument of the max operator in the first inequality and $r \geq 2$ and $p \geq 1$ in the second inequality. By concatenating (A.2) and (A.3), we obtain the following corollary of Rosenthal's inequality

$$\mathbb{E}[|S|^r] \leq C(r)p^{-\frac{r}{2}}\mathbb{E}[|Z|^r]. \tag{A.4}$$

Proof of Theorem 2.4. Let $\bar{F}_k = F_k - f(X_k)$ and $\bar{F}_k^g = F_k^g - f(X_k + \Delta_k G_k)$, for F_k and F_k^g average of p_k samples:

$$\begin{aligned} F_k &= \frac{1}{p_k}\sum_{i=1}^{p_k} F(X_k, \xi_{k,i}) \\ F_k^g &= \frac{1}{p_k}\sum_{i=1}^{p_k} F(X_k + \Delta_k G_k, \xi_{k,i}^g). \end{aligned}$$

We start with the case $q \geq 2$, implying $r \in (1, 2]$. By the conditional version of [4, Theorem 2], we have

$$\mathbb{E}[|\bar{A}_k|^r \mid \mathcal{F}_{k-1}] \leq 2M_r p_k^{1-r} \tag{A.5}$$

for $\bar{A}_k = \bar{F}_k, \bar{F}_k^g$. Let now $A_k = \bar{F}_k - \bar{F}_k^g$. We can then prove

$$\mathbb{E}[|A_k|^r \mid \mathcal{F}_{k-1}] \leq 2^{r-1}\mathbb{E}[|\bar{F}_k|^r + |\bar{F}_k^g|^r \mid \mathcal{F}_{k-1}] \leq 2^{r+1}M_r p_k^{1-r}, \tag{A.6}$$

where we used $||a| + |b||^r \leq 2^{r-1}(|a|^r + |b|^r)$ for $a, b \in \mathbb{R}$ in the first inequality, and (A.5) in the second. Applying (A.6) we obtain

$$\begin{aligned} &\mathbb{P}(|A_k| \geq \alpha\Delta_k^{\frac{r}{r-1}} \mid \mathcal{F}_{k-1}) = \mathbb{P}(|A_k|^r \geq \alpha\Delta_k^{r^2/r-1} \mid \mathcal{F}_{k-1}) \\ &\leq \frac{\mathbb{E}[|A_k|^r \mid \mathcal{F}_{k-1}]}{\alpha^r\Delta_k^{r^2/(r-1)}} \leq 2^{r+1}M_r\frac{p_k^{1-r}}{\alpha^r\Delta_k^{r^2/(r-1)}}, \end{aligned} \tag{A.7}$$

where for $p_k = O(\Delta_k^{-\frac{r^2}{(r-1)^2}}) = O(\Delta_k^{-q^2})$ the right-hand side of (A.7) is $O(1/\alpha^r)$ and Assumption 2.1 follows.

We now deal with the case $q \in (1,2)$, corresponding to $r \in (2,+\infty)$. We will apply the conditional version of (A.4) with $S = \bar{F}_k$ and $Z = F(X_k,\xi) - f(X_k)$, and write

$$\mathbb{E}[|\bar{F}_k|^r \mid \mathcal{F}_{k-1}] \le C(r) p_k^{-\frac{r}{2}} \mathbb{E}[|Z|^r \mid \mathcal{F}_{k-1}] \le C(r) M_r p_k^{-\frac{r}{2}}\,, \tag{A.8}$$

where we used (2.4) in the second inequality. Of course (A.8) holds with $\bar{F}_k^g$ instead of $\bar{F}_k$ as well. Then, reasoning as in (A.5), we get

$$\mathbb{E}[|A_k|^r \mid \mathcal{F}_{k-1}] \le 2^r C(r) M_r p_k^{-\frac{r}{2}}\,,$$

and analogously to (A.7):

$$\begin{aligned} &\mathbb{P}(|A_k| \ge \alpha \Delta_k^{\frac{r}{r-1}} \mid \mathcal{F}_{k-1}) = \mathbb{P}(|A_k|^r \ge \alpha^r \Delta_k^{\frac{r^2}{r-1}} \mid \mathcal{F}_{k-1}) \\ &\le \frac{\mathbb{E}[|A_k|^r \mid \mathcal{F}_{k-1}]}{\alpha^r \Delta_k^{r2/(r-1)}} \le \frac{2^r C(r) M_r p_k^{-\frac{r}{2}}}{\alpha^r \Delta_k^{r2/(r-1)}}\,. \end{aligned} \tag{A.9}$$

In particular, for $p_k = O(\Delta_k^{\frac{-2r}{r-1}}) = O(\Delta_k^{-2q})$, we retrieve Assumption 2.1. ∎

Proof of Proposition 2.6. We have

$$\begin{aligned} &\mathbb{E}_\xi[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^2] = \mathbb{E}_\xi[\bar{F}(x,\xi)^2] + \mathbb{E}_\xi[\bar{F}(y,\xi)^2] - 2\mathbb{E}_\xi[\bar{F}(x,\xi)\bar{F}(y,\xi)] \\ &= \mathrm{Var}_\xi(F(x,\xi)) + \mathrm{Var}_\xi(F(y,\xi)) - 2\mathrm{Cov}_\xi(F(x,\xi), F(y,\xi)) \\ &\le 2V - 2V(1 - \bar{l}\|x-y\|^2) = 2V\bar{l}\|x-y\|^2\,. \end{aligned}$$

∎

Proof of Theorem 2.7. Let A_k be an estimate of the difference between the errors in the current and the tentative point obtained with p_k samples:

$$A_k = \frac{1}{p_k}\sum_{i=1}^{p_k} (\bar{F}(X_k,\xi_{k,i}) - \bar{F}(X_k + \Delta_k G_k, \xi_{k,i}))\,.$$

Then, for $Z = \bar{F}(X_k,\xi) - \bar{F}(X_k + \Delta_k G_k, \xi)$, and $C(r)$ constant depending only on r

$$\begin{aligned} \mathbb{E}[|A_k|^r \mid \mathcal{F}_{k-1}] &\le C(r) p_k^{-\frac{r}{2}} \mathbb{E}[|Z|^r \mid \mathcal{F}_{k-1}] \le D_r C(r) p_k^{-\frac{r}{2}} \|\Delta_k G_k\|^r \\ &= D_r C(r) p_k^{-\frac{r}{2}} \Delta_k^r\,, \end{aligned} \tag{A.10}$$

where we used the conditional version of (A.4) in the first inequality, (2.7) in the second inequality, and $\|G_k\| = 1$ in the last equality.

We thus have

$$\begin{aligned}
&\mathbb{P}(|A_k| \geq \alpha \Delta_k^{\frac{r}{r-1}} \mid \mathcal{F}_{k-1}) = \mathbb{P}(|A_k|^r \geq \alpha^r \Delta_k^{\frac{r^2}{r-1}} \mid \mathcal{F}_{k-1}) \\
&\leq \frac{\mathbb{E}[|A_k|^r \mid |F_{k-1}]}{\alpha^r \Delta_k^{r^2/(r-1)}} \leq \frac{D_r C(r) \Delta_k^{-\frac{r}{r-1}}}{p_k^{\frac{r}{2}} \alpha^r},
\end{aligned}$$

where we used the conditional Chebyshev's inequality in the first inequality, and (A.10) in the last inequality. Hence we obtain Assumption 2.1 for $p_k = O(\Delta_k^{-\frac{2}{r-1}}) = O(\Delta_k^{2-2q})$ as desired. ∎

Proof of Proposition 2.9: By setting $x = y$ in (2.9) we get

$$\mathrm{Var}_\xi[F(x,\xi)] = \sigma^2 .$$

Moreover, we have

$$\mathrm{Cov}_\xi(F(x,\xi), F(y,\xi)) = \sigma^2 \exp\left(-\frac{\|x-y\|^2}{2l^2}\right) \geq \sigma^2(1 - \frac{\|x-y\|^2}{2l^2}),$$

where we used (2.9) in the equality and $e^x > 1 + x$ in the inequality. We can then apply Proposition 2.6 with $V = \sigma^2$ and $\bar{l} = \frac{1}{2l^2}$ to obtain

$$\mathbb{E}_\xi[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^2] \leq \frac{\sigma^2}{l^2}\|x-y\|^2 . \tag{A.11}$$

Let now $V_{x,y} = \mathbb{E}_\xi[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^2] = \mathrm{Var}_\varepsilon[\bar{F}(x,\xi) - \bar{F}(y,\xi)]$. Since any linear combination of jointly Gaussian variables is still Gaussian, $\bar{F}(x,\xi) - \bar{F}(y,\xi)$ is Gaussian with 0 mean and variance $V_{x,y}$. In particular, we can write $\bar{F}(x,\xi) - \bar{F}(y,\xi) = \sqrt{V_{x,y}}N$, with N having standard normal distribution. We conclude by noticing, for $r \geq 2$,

$$\mathbb{E}_\xi[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^r] = \mathbb{E}[V_{x,y}^{\frac{r}{2}}|N|^r] = V_{x,y}^{\frac{r}{2}}\mathbb{E}[|N|^r] = V_{x,y}^{\frac{r}{2}}\bar{M}_r \leq \frac{\sigma^r}{l^r}\|x-y\|^r \bar{M}_r ,$$

where we used (A.11) in the second inequality, and $\bar{M}_r$ is the r-th moment of a folded normal distribution of mean 0 and variance 1. ∎

Proof of Proposition 2.10. First, notice that

$$\begin{aligned}
&\mathbb{E}[|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))|^2 \mid \mathcal{F}_{k-1}] \\
&\leq 2(\mathbb{E}[|F_k^g - f(X_k + \Delta_k G_k)|^2 \mid \mathcal{F}_{k-1}] + \mathbb{E}[|F_k - f(X_k)|^2 \mid \mathcal{F}_{k-1}]) \qquad \text{(A.12)} \\
&\leq 4k_f^2 \Delta_k^4 ,
\end{aligned}$$

where we used $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$ in the first inequality, and (2.10) in the second. We now have

$$\begin{aligned}
&\mathbb{P}[|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha\Delta_k^2 \mid \mathcal{F}_{k-1}] \\
&= \mathbb{P}[|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))|^2 \geq \alpha^2\Delta_k^4 \mid \mathcal{F}_{k-1}] \\
&\leq \frac{\mathbb{E}[|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))|^2 \mid \mathcal{F}_{k-1}]}{\alpha^2\Delta_k^4} \leq \frac{4k_f^2}{\alpha^2},
\end{aligned}$$

where we used the conditional Chebyshev's inequality in the first inequality, and (A.12) in the second inequality. By setting $\varepsilon_q = 4k_f^2$ in the above equation we obtain

$$\mathbb{P}[|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha\Delta_k^2 \mid \mathcal{F}_{k-1}] \leq \frac{\varepsilon_q}{\alpha^2} \tag{A.13}$$

as desired. ∎

Proof of Proposition 2.12. Notice that (AS1) is trivially satisfied for $\alpha < \sqrt{\varepsilon_q}$. We then just need to deal with the case $\alpha \geq \sqrt{\varepsilon_q}$. First observe that by the triangular inequality

$$|F_k - f(X_k)| + |F_k^g - f(X_k + \Delta_k G_k)| \geq |F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))|,$$

which proves in particular (2.12). Let $\alpha \geq \sqrt{\varepsilon_q}$ be arbitrary. For $\beta = 1 - \frac{\varepsilon_q}{\alpha^2}\bar{p} \in (1, 1-\bar{p}]$,

$$\begin{aligned}
&\mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha\Delta_k^2 \, |\mathcal{F}_{k-1}) \\
&= 1 - \mathbb{P}(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| < \alpha\Delta_k^2 \, |\mathcal{F}_{k-1}) \\
&\leq 1 - \\
&\mathbb{P}(\{|F_k - f(X_k)| \leq \tau_f(\beta)\Delta_k^2\} \cap \{|F_k^g - f(X_k + \Delta_k G_k)| \leq \tau_f(\beta)\Delta_k^2\} \mid \mathcal{F}_{k-1}) \\
&\leq 1 - \beta = \frac{\varepsilon_q}{\alpha^2}\bar{p} \leq \frac{\varepsilon_q}{\alpha^2},
\end{aligned}$$

where we were able to apply (2.12) in the first inequality since by assumption

$$\tau_f(\beta) < \frac{1}{2}\sqrt{\frac{\varepsilon}{1-\beta}} = \frac{\alpha}{2},$$

and the second inequality follows from (2.11). Given that $\alpha > \varepsilon_f$ is arbitrary, this proves the first point of the thesis, and an analogous reasoning holds for the second. ∎

A.2. Additional numerical results. We report here details about the grid search used to fine tune Algorithm 1 parameters in the tests of Section 5. We fixed $\tau = 0.001$ (using $\tau = 0.01$ lead to similar results) and defined $\bar{\tau} = 1 + \tau$. We then considered, after some preliminary tests to determine a range of competitive parameters, $\theta \in \bar{\theta} = (0, 0.5, 1, 2)$ and $\delta_0 \in \bar{\delta}_0 = (0.5, 1, 2, 4)$. We defined $\bar{H} \in \mathbb{R}^{4\times4\times4\times4}$, with $\bar{H}(i_1, j_1, i_2, j_2) \in [0, 1]$ equal to the fraction of instances from the test set where SDS1.5 with $(\delta_0, \theta) = (\bar{\delta}_0(i_1), \bar{\theta}(j_1))$ outperformed SDS2 with $(\delta_0, \theta) = (\bar{\delta}_0(i_2), \bar{\theta}(j_2))$. Finally, we defined $H_{1.5} \in \mathbb{R}^{4\times4}$ as

$$H_{1.5}(i_1, j_1) = \min_{i_2, j_2 \in (1,2,3,4)^2} \bar{H}(i_1, j_1, i_2, j_2)\,,$$

i.e., the worst case performance of SDS1.5 for a fixed choice of (δ_0, θ). The matrix $H_2 \in \mathbb{R}^{4\times4}$ was defined analogously, switching the roles of SDS2 and SDS1.5. The results, in Figure 3, clearly show that the performance of SDS1.5 is robust with respect to (δ_0, θ), in most cases outperforming SDS2 in at least half the instances.

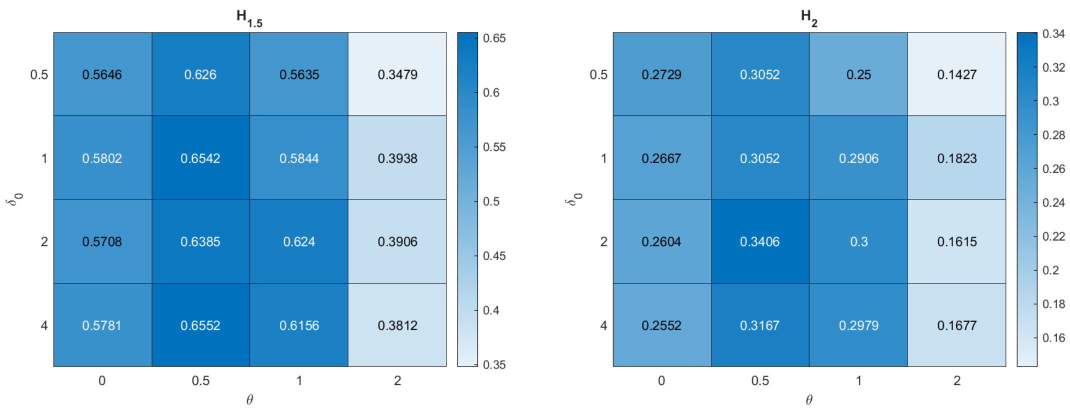


FIGURE 3. From left to right, heatmaps representing worst case performance of SDS1.5 against SDS2 and conversely, for various choices of the optimization parameters.

References

[1] S. Amaran, N. V. Sahinidis, B. Sharda, and S. J. Bury. Simulation optimization: A review of algorithms and applications. *Ann. Oper. Res.*, 240:351–380, 2016.

[2] C. Audet, K. J. Dzahini, M. Kokkolaras, and S. Le Digabel. Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. *Comput. Optim. Appl.*, 79:1–34, 2021.

[3] C. Audet and W. Hare. *Derivative-Free and Blackbox Optimization*, volume 2 of *Springer Series in Operations Research and Financial Engineering*. Springer, Cham, Switzerland, 2017.

[4] B. Von Bahr and C.-G. Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Stat.*, 36:299–303, 1965.

[5] A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM J. Optim.*, 24:1238–1264, 2014.

[6] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS J. Optim.*, 1:92–119, 2019.

[7] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Math. Program.*, 169:337–375, 2018.

[8] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Math. Program.*, 169:447–487, 2018.

[9] F. H. Clarke. *Optimization and Nonsmooth Analysis.* John Wiley & Sons, New York, 1983. Reissued by SIAM, Philadelphia, 1990.

[10] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization.* MPS/SIAM Ser. Optim. SIAM, Philadelphia, 2009.

[11] J.E. Dennis and D. J. Woods. Optimization on microcomputers: The Nelder-Mead simplex algorithm. *New Computing Environments: Microcomputers in Large-Scale Computing*, 11:6–122, 1987.

[12] K. J. Dzahini. Expected complexity analysis of stochastic direct-search. *Comput. Optim. Appl.*, 81:179–200, 2022.

[13] K. J. Dzahini, M. Kokkolaras, and S. Le Digabel. Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. *Math. Program.*, 2022, to appear.

[14] J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.*, 2:84–90, 1960.

[15] R. Ibragimov and S. Sharakhmetov. The exact constant in the Rosenthal inequality for random variables with mean zero. *Theory Probab. Appl.*, 46:127–132, 2002.

[16] N. Karmitsa. Test problems for large-scale nonsmooth minimization. *Reports of the Department of Mathematical Information Technology. Series B, Scientific computing*, 2007.

[17] G. Lan. *First-Order and Stochastic Optimization Methods for Machine Learning.* Data Sciences. Springer, Switzerland, 2020.

[18] J. Larson and S. C. Billups. Stochastic derivative-free optimization using a trust region framework. *Comput. Optim. Appl.*, 64:619–645, 2016.

[19] G. Liuzzi, S. Lucidi, F. Rinaldi, and L. N. Vicente. Trust-region methods for the derivative-free optimization of nonsmooth black-box functions. *SIAM J. Optim.*, 29:3012–3035, 2019.

[20] G. Liuzzi, S. Lucidi, and M. Sciandrone. A derivative-free algorithm for linearly constrained finite minimax problems. *SIAM J. Optim.*, 16:1054–1075, 2006.

[21] L. Lukšan and J. Vlcek. Test problems for nonsmooth unconstrained and linearly constrained optimization. Technical report, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2000.

[22] J. J. Moré and S. M. Wild. Benchmarking derivative-free optimization algorithms. *SIAM J. Optim.*, 20:172–191, 2009.

[23] C. Paquette and K. Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM J. Optim.*, 30:349–376, 2020.

[24] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

[25] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[26] S. Shashaani, F. S. Hashemi, and R. Pasupathy. ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM J. Optim.*, 28:3145–3176, 2018.

[27] L. N. Vicente. Worst case complexity of direct search. *EURO J. Comput. Optim.*, 1:143–153, 2013.

F. Rinaldi
Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy (rinaldi@math.unipd.it).

L. N. Vicente
Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Avenue, Bethlehem, PA 18015-1582, USA and Centre for Mathematics of the University of Coimbra (CMUC), Portugal (lnv@mat.uc.pt).

D. Zeffiro
Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy (zeffiro@math.unipd.it).