A weak tail-bound probabilistic condition for function estimation in stochastic derivative-free optimization

F. Rinaldi * L. N. Vicente † D. Zeffiro‡

February 22, 2022

Abstract

In this paper, we use tail bounds to define a tailored probabilistic condition for function estimation that eases the theoretical analysis of stochastic derivative-free optimization methods. In particular, we focus on the unconstrained minimization of a potentially non-smooth function, whose values can only be estimated via stochastic observations, and give a simplified convergence proof for both a direct search and a basic trust-region scheme.

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 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., [13] for further details). In simulation-based engineering applications, the function $F(x, \xi)$ is simply related to some noisy computable version of the

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

†Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Avenue, Bethlehem, PA 18015-1582, USA. Support for this author was partially provided by the Centre for Mathematics of the University of Coimbra under grant FCT/MCTES UIDB/MAT/00324/2020.

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

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 (1.1) 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., [14, 17], 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 functions, and only recently adapted to deal with stochastic observations (see, e.g., [7] 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, 9] for further details on these two classes of methods).

In [14], 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 i.i.d. with zero mean and finite variance, that is they aim at solving a smooth version of problem (1.2), when ξ is additive noise. In the same line of research, the authors in [17] 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 i.i.d. noise 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 [7], 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 [5]. The framework analyzed in [5, 7] extends the the trust region DFO method based on probabilistic models described in [4]. It is important to notice that the randomness in the models described in [4] 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 [10] 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 [5]. 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 (see [8]) of the objective function by assuming that stochastic observations are sufficiently accurate and satisfy a variance condition. The analysis adapts to the considered gradient-free framework the theoretical analysis given in [16] for a class of stochastic

gradient-based methods.

The main goal of this paper is to analyze in depth some tail-bound probabilistic conditions for function estimation and show how those theoretical tools allow to achieve a sharp analysis of stochastic derivative-free methods. More specifically, we focus 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 tests of those derivative-free algorithms, and suitably bound the probability that the error in that estimate is large. Those simple tail-bound probabilistic conditions are then compared with the other theoretical tools currently used in the literature. We indeed see how:

- our conditions are implied by the variance conditions considered in [2] and by the probabilistically accurate function estimate assumption used in [2, 7, 16];
- one of our conditions is implied by a tail bound used in [14];
- the finite variance oracle usually considered in the literature (see, e.g., [14, 17]) can be replaced by a finite moment oracle (see Remark 2.5 and Subsection A.2 for further details) when constructing estimates satisfying our conditions.

We finally define two different algorithmic schemes, that is

- a simple stochastic direct-search strategy (see, e.g., [3, 9, 12] for further details on direct-search approaches);
- a stochastic version of the basic trust-region scheme reported in [15].

Those new algorithmic schemes are simply obtained by replacing the function values with their estimates in the acceptance tests of the deterministic counterparts. Both schemes work as follows: they choose a direction over the unit sphere; then generate the new iterate by either moving along the direction, in case of the direct search, or by solving a trust-region subproblem, for the trust-region method; finally they use a suitable acceptance test to decide if the new point can be accepted (successful iteration) or not. Convergence of the methods is then carried out by simply assuming that our tail-bounds hold. 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 limit points.

The paper is organized as follows. In Section 2, we introduce our tail-bound probabilistic conditions and compare them with the existing conditions from the literature. We then analyze the direct-search and trust-region schemes in Section 3 and 4, respectively. We finally draw some conclusions and discuss some possible extensions in Section 5. In order to improve readability and ease the comprehension, we include some technical results into a small appendix.

2 A weak tail-bound probabilistic condition for function estimation

In order to give convergence results for our algorithms, we first need some probabilistic assumptions on the accuracy of the oracle. In this section, we hence describe our tail-bound condition and compare it with other existing conditions from the literature. 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, that is the space of the realizations of the algorithms under analysis. Any single outcome of the sample space Ω will be denoted by w. Our algorithms hence generate a random process whose random quantity realizations are indicated as follows. In our algorithmic schemes a search direction is generated at each iteration and a suitable stepsize is used to move along it. The search direction realization is denoted by g_k as a shorthand for $g_k(w)$, and the stepsize realization is indicated with Δ_k as a shorthand for $\Delta_k(w)$. Realizations related to the estimates of the function values $f(x_k)$ and $f(x_k + \Delta_k g_k)$ are indicated with f_k and f_k^g as a shorthand for $f_k(w)$ and $f_k^g(w)$, respectively. For a random variable X defined in Ω we use the shorthand $\{X \in A\}$ to denote $\{w \mid X(w) \in A\}$. Let $\mathcal{F}_{k-1}$ be the $\sigma-$algebra of events up to the choice of x_k. We assume that g_k is measurable with respect to $\mathcal{F}_{k-1}$. Finally, we use $\mathbb{E}$ to denote expectation and conditional expectation, and $\mathbb{N}_0$ to denote the set of nonnegative integers.

2.1 The weak tail-bound probabilistic condition

We now introduce our tail bound assumptions.

Assumption 2.1. *For every $p \in (0, 1]$ and some $\varepsilon_f > 0$ (independent of p):*

$$\mathbb{P}\left(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \frac{\varepsilon_f}{p}\Delta_k^2 \mid \mathcal{F}_{k-1}\right) \leq p\,. \tag{A1}$$

Assumption 2.2. *For every $p \in (0, 1]$ and some $\varepsilon_q > 0$ (independent of p):*

$$\mathbb{P}\left(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \sqrt{\frac{\varepsilon_q}{p}}\Delta_k^2\right) \leq p\,. \tag{A2}$$

Notice that we are only assuming error bounds 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; furthermore, only the LHS in (A1) is conditioned on $\mathcal{F}_{k-1}$, while the LHS in (A2) is not conditioned at all. 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.

2.2 Comparison with the existing conditions

Our conditions are weaker than the ones imposed in [2]. More precisely, they are implied by [2, Equation (2)], rewritten in our notation as

$$\begin{aligned} \mathbb{E}[|f_k^g - f(x_k + \Delta_k g_k)|^2 \mid \mathcal{F}_{k-1}] &\leq k_f^2 \Delta_k^4 \\ \mathbb{E}[|f_k - f(x_k)|^2 \mid \mathcal{F}_{k-1}] &\leq k_f^2 \Delta_k^4\,, \end{aligned} \tag{2.1}$$

for a constant $k_f > 0$. The k_f-variance condition in (2.1) is a gradient free version of [16, Assumption 2.4, (iii)], and more precisely can be obtained from the latter by removing the gradient related terms in the RHS. However, in [16] as well as in other works on smooth stochastic derivative free optimization (see, e.g., [7, 14, 17] 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 2.3. *Condition (2.1) implies Assumption 2.1 and Assumption 2.2 for $\varepsilon_f = 2k_f$ and $\varepsilon_q = 4k_f^2$ respectively.*

Proof. First, notice that

$$\begin{aligned}&\mathbb{E}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))|^2 \mid \mathcal{F}_{k-1}]\\ &\leq 2(\mathbb{E}[|f_k^g - f(x_k + \Delta_k g_k)|^2 \mid \mathcal{F}_{k-1}] + \mathbb{E}[|f_k - f(x_k)|^2 \mid \mathcal{F}_{k-1}])\\ &\leq 4k_f^2\Delta_k^4\,,\end{aligned} \tag{2.2}$$

where we used the squared triangular inequality in the first inequality, and (2.1) in the second.

We now prove (A1). In order to do so, we only need a bound on the first moment $\mathbb{E}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \mid \mathcal{F}_{k-1}]$, implied by the bound on the second moment (2.2) thanks to conditional Jensen's inequality:

$$\begin{aligned}&\mathbb{E}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \mid \mathcal{F}_{k-1}]\\ &\leq \sqrt{\mathbb{E}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))|^2 \mid \mathcal{F}_{k-1}]} \leq 2k_f\Delta_k^2\,.\end{aligned} \tag{2.3}$$

We can now conclude by noticing

$$\begin{aligned}&\mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \frac{\varepsilon_f}{p}\Delta_k^2 \,|\mathcal{F}_{k-1})\\ &\leq p\frac{\mathbb{E}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \,|\mathcal{F}_{k-1})}{\varepsilon_f\Delta_k^2} \leq p\frac{2k_f}{\varepsilon_f}\,,\end{aligned} \tag{2.4}$$

where we used the conditional Chebyshev's inequality (see, e.g., [6, Corollary 3.1.1]) in the first inequality, and (2.3) in the second inequality. In particular, (2.1) implies (A1) for $\varepsilon_f = 2k_f$.

As for (A2), we have

$$\begin{aligned}&\mathbb{P}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \frac{\sqrt{\varepsilon_q}}{p}\Delta_k^2 \mid \mathcal{F}_{k-1}]\\ &= \mathbb{P}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))|^2 \geq \frac{\varepsilon_q}{p}\Delta_k^4 \mid \mathcal{F}_{k-1}]\\ &\leq p\frac{\mathbb{E}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))|^2 \mid \mathcal{F}_{k-1}]}{\varepsilon_q\Delta_k^4} \leq p\frac{4k_f^2}{\varepsilon_q}\,,\end{aligned}$$

where we used the conditional Chebyshev's inequality in the first inequality, and (2.2) in the second inequality. By setting $\varepsilon_q = 4k_f^2$ in the above equation we obtain

$$\mathbb{P}[|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \frac{\sqrt{\varepsilon_q}}{p}\Delta_k^2 \mid \mathcal{F}_{k-1}] \leq p\,. \tag{2.5}$$

Since (2.5) holds, a fortiori the non conditional version (A2) holds. □

Remark 2.4. The algorithm proposed in [2] uses a strategy to select the search direction at iteration k that depends on the function estimates. This makes the analysis of the algorithm more complicated, requiring in particular the stronger assumptions (2.1). In our algorithm, instead, the search direction g_k only depends on the information $\mathcal{F}_{k-1}$ available before the estimates are computed.

Remark 2.5. As a corollary of Proposition 2.3, our assumptions can always be satisfied if the variance of the oracle is finite (see Appendix A.1 in for details). Furthermore, both Assumption 2.1 and a weaker version of Assumption 2.2, which still suffice to prove convergence of methods like the ones in Sections 3.1 and 4.1, can be satisfied by simply assuming that the r-th moment of the oracle is finite, for any $r \in (1, 2]$ (see Appendix A.2).

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

$$\mathbb{P}(\{|f_k - f(x_k)| \leq \tau_f \Delta_k^2\} \cap \{|f_k^g - f(x_k + \Delta_k g_k)| \leq \tau_f \Delta_k^2|\} \; \mathcal{F}_{k-1}) \geq \beta \,, \tag{2.6}$$

used in [2, 7, 16] in combination with other assumptions. In particular, conditions (2.1) are used in [2] and [16] (as discussed above), and a probabilistic assumption on the accuracy of random models for the objective is considered in [7].

We show that if (2.6) is satisfied for every β in a certain interval, with τ_f depending from an accuracy parameter ε, then also our assumptions are satisfied with $\varepsilon_f, \varepsilon_q$ dependent from ε. Note that the parameter τ_f is upper bounded by a function of β, arbitrarily large for β close to 1, but the result holds for any positive τ_f within the prescribed interval.

Proposition 2.6. *Let $\varepsilon > 0$ and $\bar{p} \in (0, 1)$. Assume that (2.6) holds for every $\beta \in [1 - \bar{p}, 1)$.*

- *If $\tau_f < \frac{\varepsilon}{2(1-\beta)}$, then Assumption 2.1 holds with $\varepsilon_f = \frac{\varepsilon}{\bar{p}}$.*
- *If $\tau_f < \frac{1}{2}\sqrt{\frac{\varepsilon}{1-\beta}}$, then Assumption 2.2 holds with $\varepsilon_q = \sqrt{\frac{\varepsilon}{\bar{p}}}$.*

Proof. First observe that by the triangular inequality

$$|f_k - f(x_k)| + |f_k^g - f(x_k + \Delta_k g_k)| \geq |f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \,.$$

Let $p \in (0, 1]$ be arbitrary. Then

$$\begin{aligned} &\{|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \leq \frac{\varepsilon}{\bar{p}p} \Delta_k^2\} \\ &\supset \{|f_k - f(x_k)| \leq \tau_f \Delta_k^2\} \cap \{|f_k^g - f(x_k + \Delta_k g_k)| < \tau_f \Delta_k^2\} \end{aligned} \tag{2.7}$$

for any $\tau_f < \frac{\varepsilon}{2p\bar{p}}$. Therefore, for $\beta = 1 - p\bar{p}$,

$$\begin{aligned} &\mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \frac{\varepsilon}{\bar{p}p} \Delta_k^2 \; |\mathcal{F}_{k-1}) \\ &= (1 - \mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| < \frac{\varepsilon}{\bar{p}p} \Delta_k^2 \; |\mathcal{F}_{k-1})) \\ &\leq (1 - \mathbb{P}(\{|f_k - f(x_k)| \leq \tau_f(\beta) \Delta_k^2\} \cap \{|f_k^g - f(x_k + \Delta_k g_k)| \leq \tau_f(\beta) \Delta_k^2\} \mid \mathcal{F}_{k-1})) \\ &\leq 1 - \beta = p\bar{p} \leq p \,, \end{aligned}$$

where we were able to apply (2.7) in the first inequality since by assumption $\tau_f(\beta) < \frac{\varepsilon}{2(1-\beta)} = \frac{\varepsilon}{2p\bar{p}}$. Given that $p \in (0,1]$ is arbitrary, this proves the first point of the thesis, and an analogous reasoning holds for the second. □

We now show how the tail bound [14, Condition 2] is stronger than (a slight modification of) Assumption 2.1. We remark that in [14] this tail bound is combined with a probabilistically accurate difference estimate assumption and fully linear local model in order to prove convergence. We first recall the tail bound assumption [14, Condition 2]:

$$\mathbb{P}(f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k)) > (\beta\eta + \varepsilon)\min\{\Delta_k, \Delta_k^2\} \mid \mathcal{F}_{k-1}) \leq \frac{\theta}{\varepsilon}, \tag{2.8}$$

for every $\varepsilon > 0$, $k > \hat{k}$, and some $\beta, \eta, \theta > 0$. We now introduce the following modification of Assumption 2.1, essentially equivalent for our purposes:

$$\mathbb{P}\left(f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k)) > \frac{\varepsilon_f}{p}\Delta_k^2 \mid \mathcal{F}_{k-1}\right) \leq p, \tag{2.9}$$

for every $p \in (0,1]$, $k > \hat{k}$ and some $\varepsilon_f > 0$. It is straightforward to check that all of our results still hold if we replace (A1) with (2.9).

Lemma 2.7. *If (2.8) holds with*

$$\theta + \beta\eta < \varepsilon_f, \tag{2.10}$$

then (2.9) holds.

Proof. We rewrite (2.9) as

$$\mathbb{P}(f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k)) > \alpha\Delta_k^2 \mid \mathcal{F}_{k-1}) \leq \frac{\varepsilon_f}{\alpha},$$

for every $\alpha \geq \varepsilon_f$. First, for every $\alpha \geq \varepsilon_f$ we have

$$\frac{\theta}{\alpha - \eta\beta} \leq \frac{\varepsilon_f}{\alpha} \tag{2.11}$$

under (2.10), since

$$\frac{\theta}{1 - \eta\beta/\alpha} \leq \frac{\theta}{1 - \eta\beta/\varepsilon_f} = \frac{\varepsilon_f\theta}{\varepsilon_f - \eta\beta} \leq \varepsilon_f,$$

where we used $\alpha \geq \varepsilon_f$ in the first inequality and (2.10) in the last inequality.

Now, for every $\alpha \geq \varepsilon_f$:

$$\begin{aligned}
&\mathbb{P}(f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k)) > \alpha\Delta_k^2 \mid \mathcal{F}_{k-1}) \\
&= \mathbb{P}(f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k)) > (\eta\beta + (\alpha - \eta\beta))\Delta_k^2 \mid \mathcal{F}_{k-1}) \\
&\leq \mathbb{P}(f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k)) > (\eta\beta + (\alpha - \eta\beta))\min\{\Delta_k, \Delta_k^2\} \mid \mathcal{F}_{k-1}) \\
&\leq \frac{\theta}{\alpha - \eta\beta} \leq \frac{\varepsilon_f}{\alpha},
\end{aligned} \tag{2.12}$$

where we used (2.11) in the last inequality. □

3 Direct search for stochastic non-smooth functions

In this section, we first describe a simple stochastic direct-search algorithm for the unconstrained minimization problem given in (1.1), where f is possibly non-smooth, and then analyze its convergence.

3.1 A simple 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 the estimate values f_k^g and f_k 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 $\theta\Delta_k^2$. 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`

1 **Initialization.** Choose a point x_0, $\Delta_0, \theta > 0$, $\tau \in (0,1)$, $\bar{\tau} \in [1, 1+\tau]$.
2 **For** $k = 0, 1 \ldots$
3 Select a direction g_k in the unitary sphere.
4 Compute estimates f_k and f_k^g for f in x_k and $x_k + \Delta_k g_k$.
5 **If** $f_k - f_k^g \geq \theta\Delta_k^2$, **Then** set `SUCCESS` = `true`, $x_{k+1} = x_k + \Delta_k g_k$, $\Delta_{k+1} = \bar{\tau}\Delta_k$.
6 **Else** set `SUCCESS` = `false`, $x_{k+1} = x_k$, $\Delta_{k+1} = (1-\tau)\Delta_k$.
7 **End if**
8 **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 on the unit sphere can be generated using a suitable pseudorandom sequence (see, e.g., [11, 15]).

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 (A1) allows us to give a unified argument for unsuccessful and successful steps. 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_f:

$$\theta > \frac{4\varepsilon_f}{2-\tau}\,, \tag{3.1}$$

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 (3.1) holds then*

$$\sum_{k\in\mathbb{N}_0} \Delta_k^2 < \infty \tag{3.2}$$

a.s. in Ω.

Proof. Let $\Phi_k = f(x_k) - f^* + \eta\Delta_k^2$, with $\eta = \frac{\theta}{4\tau}$, and

$$\varepsilon = -\varepsilon_f + \frac{(2-\tau)\theta}{4} > 0\,,$$

where the inequality follows by (3.1).

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

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}] \geq \varepsilon\Delta_k^2\,. \tag{3.3}$$

The thesis then follows as in [2, Theorem 1]: for every k,

$$\varepsilon\mathbb{E}\left[\sum_{i=0}^{k}\Delta_i^2\right] \leq \mathbb{E}\left[\sum_{i=0}^{k}\mathbb{E}[\Phi_i - \Phi_{i+1} \mid \mathcal{F}_{k-1}]\right] = \mathbb{E}\left[\sum_{i=0}^{k}\Phi_i - \Phi_{i+1}\right] = \mathbb{E}[\Phi_0 - \Phi_{k+1}] \leq \mathbb{E}[\Phi_0],$$

where we used $\Phi_{k+1} \geq 0$ in the last inequality, which follows immediately from the definition of Φ_k. Passing to the limit, we obtain

$$\varepsilon\mathbb{E}\left[\sum_{i\in\mathbb{N}_0}\Delta_i^2\right] \leq \mathbb{E}[\Phi_0] < \infty\,,$$

and therefore in particular (3.2) must hold a.s. in Ω.

It remains to prove (3.3). Let $\delta_k \in \mathbb{R}$ be such that $f(x_k) - f(x_k + \Delta_k g_k) = (\theta - \delta_k)\Delta_k^2$, and let J_k be the event that the step k is successful. We have

$$\begin{aligned}
&\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] = \mathbb{E}[(\Phi_k - \Phi_{k+1})(\mathbb{1}_{J_k} \ + (1 - \mathbb{1}_{J_k}))|\mathcal{F}_{k-1}] \\
&= (f(x_k) - f(x_{k+1}) + \eta(\Delta_k^2 - \Delta_{k+1}^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad + (f(x_k) - f(x_{k+1}) + \eta(\Delta_k^2 - \Delta_{k+1}^2))\mathbb{E}[\ 1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&= (f(x_k) - f(x_k + \Delta_k g_k) + \eta(\Delta_k^2 - \Delta_{k+1}^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad + \eta(\Delta_k^2 - \Delta_{k+1}^2)\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\geq (((\theta - \delta_k) - \eta(2\tau + \tau^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta(2\tau - \tau^2)\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}])\Delta_k^2,
\end{aligned} \tag{3.4}$$

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,

$$\begin{aligned}
&(((\theta - \delta_k) - \eta(2\tau + \tau^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta(2\tau - \tau^2)\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}])\Delta_k^2 \\
&= ((\theta - \delta_k - 4\eta\tau)\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta(2\tau - \tau^2))\Delta_k^2 \\
&= -\delta_k\Delta_k^2\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta(2\tau - \tau^2)\Delta_k^2\,,
\end{aligned} \tag{3.5}$$

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 $\theta = 4\eta\tau$ in the second one. By combining (3.4) and (3.5) we can therefore conclude

$$\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] \geq -\delta_k\Delta_k^2\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta(2\tau - \tau^2)\Delta_k^2\,. \tag{3.6}$$

Notice that if the step is successful then $f_k - f_k^g \geq \theta\Delta_k^2$, which implies

$$f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k)) \geq \theta\Delta_k^2 - (\theta - \delta_k)\Delta_k^2 = \delta_k\Delta_k^2\,.$$

In particular

$$J_k \subset \{|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \delta_k\Delta_k^2\}\,,$$

and we can write

$$\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = \mathbb{P}(J_k|\mathcal{F}_{k-1}) \leq \mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \delta_k\Delta_k^2|\mathcal{F}_{k-1})\,. \tag{3.7}$$

Thanks to (3.7) we always have

$$-\delta_k\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \geq -\delta_k\mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| > \delta_k\Delta_k^2|\mathcal{F}_{k-1}) \geq -\varepsilon_f\,, \tag{3.8}$$

with the last inequality trivial if $\delta_k \leq \varepsilon_f$, and a direct consequence of (A1) for $p = \varepsilon_f/\delta_k$ otherwise. Hence,

$$-\delta_k\Delta_k^2\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta(2\tau - \tau^2)\Delta_k^2 \geq (-\varepsilon_f + \eta(2\tau - \tau^2))\Delta_k^2 = \varepsilon\Delta_k^2\,, \tag{3.9}$$

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 2.2 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 Assumptions 2.1–2.2 and (3.1) we have a.s. in Ω*

$$\liminf_{k\in K,\, k\to\infty} \frac{f(x_k + \Delta_k g_k) - f(x_k)}{\Delta_k} \geq 0\,. \tag{3.10}$$

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 (A2) with $p = \varepsilon_q m^2 \Delta_k^2$, we have

$$\mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \frac{\Delta_k}{m}) \leq m^2 \Delta_k^2 \varepsilon_q\,.$$

By Theorem 3.1 we have that for some $V \subset \Omega$ with $\mathbb{P}(V) = 1$ (3.2) holds. On this subset, we have

$$\sum_{k \in \mathbb{N}_0} \mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \frac{\Delta_k}{m}) \leq \sum_{k \in \mathbb{N}_0} m^2 \Delta_k^2 \varepsilon_q < \infty\,,$$

where we used (3.2) in the last inequality. In particular, by the Borel-Cantelli theorem

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

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, for every $k \in K$ large enough and a.s. (in particular on V'),

$$\frac{f(x_k + \Delta_k g_k) - f(x_k)}{\Delta_k} \geq \frac{f_k^g - f_k - |f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))|}{\Delta_k} \geq -\theta \Delta_k - \frac{1}{m}\,, \tag{3.13}$$

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.

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. Under Assumptions 2.1–2.2, 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 for $w \in \Omega$ let $S(w) \subset L(w)$ be such that

$$\lim_{k \in S(w),\, k \to \infty} g_k = d\,.$$

By definition of Clarke stationarity, we just need to prove that a.s. (for an event w independent from d)

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

By Lemma 3.2 we have $V'\subset\Omega$ with $\mathbb{P}(V')=1$ such that (3.10) holds for every $w\in V'$. Then on V' we can write

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

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 $k\in S(w)\subset L(w)$ we get

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

for every $w\in V'$, where we used $\|g_k-d\|\to 0$ by construction in the first inequality and (3.14) in the second. □

4 Trust-region methods 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 [15], and analyze its convergence properties under tail-bound probabilistic conditions like the ones used in Section 3. The main difference between this new version and the original Basic DFO-TRNS algorithm is in the presence of an upper bound on the trust-region radius. This is a fundamental tool in the stochastic context. Indeed, in successful iterations a better control on the amplitude of the step must be exercised in order to prevent errors to accumulate too much. This allows to prove convergence of the squared trust-region radii series (i.e., the analogous of Lemma 2.1 in [15]), stating convergence to zero of the trust-region radius. Using this property, we manage to show that the proposed modification of the original Basic DFO-TRNS algorithm enjoys almost surely convergence to Clarke stationary points.

4.1 A simple stochastic trust-region scheme

As already said, the simple trust-region algorithm that we report here is a minor modification of the Basic DFO-TRNS algorithm proposed in [15]. Indeed, there are two differences between the Basic DFO-TRNS algorithm and its stochastic counterpart.

The first difference consists in the exponent of the acceptance ration ρ_k which we fix to 2 whereas in [15] is $1+p$ with $p>0$.

The second, more relevant difference, is in the updating rule related to the trust-region radius. In our modification, we choose $\tau \in (0,1)$ and then define $\gamma_1 = 1-\tau$ and $\gamma_2 \in [1, 1+\tau]$. Furthermore, in case of a successful iteration, the new trust-region is possibly expanded provided that the new radius does not exceed a maximum length $\Delta_{\max}$. This guarantees to perform a more stringent control with respect to larger steps that could possibly give rise to higher uncertainty in the objective function values.

The detailed scheme of the algorithm (see Algorithm 2) is reported below. At every iteration a symmetric matrix B_k is built by interpolation or regression on a sample set of points and a direction g_k is generated on the unit sphere (independently of the objective function estimates generated so far). 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 onto the spherical trust-region constraint. Once the current step has been computed, we evaluate the estimate of the true objective function on the tentative point $x_k + s_k$ and compute the acceptance ratio ρ_k. A new estimate f_k of f at x_k is recomputed at every iteration (for use in Step 5 and especially 6). Note that, as in [15], the non-standard acceptance ratio is motivated by convergence requirements. In the 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 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. We point out that such an assumption is weaker than what is usually done in trust-region methods, where an upper bound on the norm of B_k is traditionally imposed. Our theory allows B_k to be unbounded as long as it is bounded by a negative power of the trust-region radius (we shall prove convergence of the series of squared trust-region radii).

Assumption 4.1. *There exist $q \in (0,1)$, $m, M > 0$, such that: The maximal eigenvalue of B_k satisfies*

$$\lambda_{\max}(B_k) \;\leq\; M\Delta_k^{-q}.$$

When B_k has negative eigenvalues, its minimal eigenvalue satisfies

$$-\lambda_{\min}(B_k) \;\leq\; m\Delta_k^{-q}.$$

4.2 Convergence analysis under the tail-bound probabilistic condition

In order to analyze the method introduced above, we adapt Assumptions 2.1–2.2, using $g_k = s_k/\|s_k\|$ and $\Delta_k = \|s_k\|$. Now Δ_k stands for the trust-region radius. Hence, we obtain the following assumptions.

Algorithm 2 `Stochastic DFO Trust-Region Algorithm`

1 **Initialization.** Select $x_0 \in \mathbb{R}^n$, $\theta > 0$, $\tau \in (0,1)$, $\bar{\tau} \in [1, 1+\tau]$, $\Delta_0 > 0$.

2 **For** $k = 0, 1 \ldots$

3 Select a direction g_k on the unit sphere and build a symmetric matrix B_k.

4 Compute an estimate f_k of $f(x_k)$.

5 Compute
$$s_k \in \operatorname*{argmin}_{\|s\| \leq \Delta_k} f_k + g_k^\top s + \frac{1}{2} s^\top B_k s. \tag{4.1}$$

6 Compute an estimate f_k^s of $f(x_k + s_k)$ and let
$$\rho_k = \frac{f_k - f_k^s}{\theta \|s_k\|^2}.$$

7 **If** $\rho_k \geq 1$ **Then** set `SUCCESS` = `true`, $x_{k+1} = x_k + s_k$, $\Delta_{k+1} = \min(\Delta_{\max}, \bar{\tau}\Delta_k)$,

8 **Else** set `SUCCESS` = `false`, $x_{k+1} = x_k$, $\Delta_{k+1} = (1-\tau)\Delta_k$.

9 **End If**

10 **End For**

Assumption 4.2. *For some $\varepsilon_f > 0$ and every $p \in (0,1]$:*

$$\mathbb{P}\left(|f_k - f_k^s - (f(x_k) - f(x_k + s_k))| \geq \frac{\varepsilon_f}{p} \|s_k\|^2 \ | \mathcal{F}_{k-1} \right) \leq p. \tag{A1'}$$

Assumption 4.3. *For some $\varepsilon_q > 0$ and every $p \in (0,1]$:*

$$\mathbb{P}\left(|f_k - f_k^s - (f(x_k) - f(x_k + s_k))| \geq \sqrt{\frac{\varepsilon_q}{p}} \|s_k\|^2 \right) \leq p\,. \tag{A2'}$$

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

Theorem 4.4. *Let $\bar{M} = \frac{1}{M^2 \Delta_{\max}^{2-2q}}$. Under Assumption 4.2, if*

$$\theta > \frac{4\varepsilon_f}{\bar{M}(2-\tau)}, \tag{4.2}$$

then

$$\sum_{k \in \mathbb{N}_0} \Delta_k^2 < \infty$$

a.s. in Ω.

Proof. Let $\Phi_k = f(x_k) - f^* + \eta \Delta_k^2$, with

$$\eta = \frac{\theta \bar{M}}{4\tau}. \tag{4.3}$$

By definition we have either that s_k is on the boundary of the trust region, that is $\|s_k\| = \Delta_k$, or $s_k = B_k^{-1} g_k$. In the latter case

$$\|s_k\|^2 = \|B_k^{-1} g_k\|^2 \geq \frac{1}{M^2}\Delta_k^{2q} \geq \frac{1}{M^2 \Delta_{\max}^{2-2q}}\Delta_k^2 = \bar{M}\Delta_k^2 \,. \tag{4.4}$$

We can assume $\bar{M} \leq 1$, otherwise we must always have $\|s_k\| = \Delta_k$ and the result follows exactly as in Theorem 3.1. Let $\delta_k \in \mathbb{R}$ be such that $f(x_k) - f(x_k + s_k) = (\theta - \delta_k)\|s_k\|^2$ and let J_k represent the event that the step k is successful.

In the first part of the proof we will show that an $\varepsilon > 0$ exists such that

$$\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] \geq \varepsilon \Delta_k^2. \tag{4.5}$$

Following the same argument as in Theorem 3.1, we have

$$\begin{aligned}
&\mathbb{E}[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}] \\
&\quad = (f(x_k) - f(x_k + s_k) + \eta(\Delta_k^2 - \Delta_{k+1}^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta(\Delta_k^2 - \Delta_{k+1}^2)\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad \geq ((\theta - \delta_k)\|s_k\|^2 - \eta\Delta_k^2(2\tau + \tau^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\Delta_k^2(2\tau - \tau^2)\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \,.
\end{aligned} \tag{4.6}$$

where the inequality in the above relation comes from the fact that $\Delta_{k+1} \leq (1 + \tau)\Delta_k$.

Notice that if the step is successful then $f_k - f_k^s \geq \theta\|s_k\|^2$, so that

$$f_k - f_k^s - (f(x_k) - f(x_k + s_k)) \geq \theta\|s_k\|^2 - (\theta - \delta_k)\|s_k\|^2 = \delta_k\|s_k\|^2 \,.$$

In particular

$$J_k \subset \{|f_k - f_k^s - (f(x_k) - f(x_k + s_k))| \geq \delta_k\|s_k\|^2\} \,,$$

and we can write

$$\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = \mathbb{P}(J_k|\mathcal{F}_{k-1}) \leq \mathbb{P}(|f_k - f_k^s - (f(x_k) - f(x_k + s_k))| \geq \delta_k\|s_k\|^2|\mathcal{F}_{k-1}) \,. \tag{4.7}$$

Let us now consider the following three cases: $\delta_k < 0$; $0 \leq \delta_k \leq \varepsilon_f$; $\delta_k > \varepsilon_f$.

$\underline{\delta_k < 0}$ Then $\delta_k\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \leq \varepsilon_f$ so that $-\delta_k\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \geq -\varepsilon_f$

$\underline{0 \leq \delta_k \leq \varepsilon_f}$ Then $-\delta_k \geq -\varepsilon_f$ so that, again, $-\delta_k\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \geq -\varepsilon_f$

$\underline{\delta_k > \varepsilon_f}$ Then, let $p = \varepsilon_f/\delta_k$. By (A1') we have that

$$\delta_k\mathbb{P}(|f_k - f_k^s - (f(x_k) - f(x_k + s_k))| \geq \frac{\varepsilon_f}{p}\|s_k\|^2|\mathcal{F}_{k-1}) \leq \varepsilon_f$$

Now, by recalling (4.7), we can write

$$\delta_k\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \leq \delta_k\mathbb{P}(|f_k - f_k^s - (f(x_k) - f(x_k + s_k))| \geq \frac{\varepsilon_f}{p}\|s_k\|^2|\mathcal{F}_{k-1}) \leq \varepsilon_f$$

so that, we have

$$-\delta_k\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \geq -\varepsilon_f.$$

Considering the three cases all together, we can always write

$$-\delta_k \mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \geq -\varepsilon_f\,. \tag{4.8}$$

We can now proceed as follows

$$\begin{aligned}
&((\theta-\delta_k)\|s_k\|^2 - \eta\Delta_k^2(2\tau+\tau^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\Delta_k^2(2\tau-\tau^2)\mathbb{E}[1-\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] \\
&\quad = ((\theta-\delta_k)\|s_k\|^2 - \eta\Delta_k^2(2\tau+\tau^2) - \eta\Delta_k^2(2\tau-\tau^2))\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] + \eta\Delta_k^2(2\tau-\tau^2) \\
&\quad \geq (\theta\|s_k\|^2 - 4\eta\tau\Delta_k^2)\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] - \varepsilon_f\|s_k\|^2 + \eta\Delta_k^2(2\tau-\tau^2) \\
&\quad \geq (\bar{M}\theta - 4\eta\tau)\Delta_k^2\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] - \varepsilon_f\Delta_k^2 + \eta\Delta_k^2(2\tau-\tau^2) \\
&\quad = (-\varepsilon_f + \eta(2\tau-\tau^2))\Delta_k^2 = \varepsilon\Delta_k^2\,,
\end{aligned} \tag{4.9}$$

where we used (4.8) in the first inequality, (4.4) together with $\|s_k\| \leq \Delta_k$ in the second, and $\bar{M}\theta - 4\eta\tau = 0$ in the final equality. From (4.2), we have that

$$4\varepsilon_f < \bar{M}\theta(2-\tau) = 4\eta(2\tau-\tau^2)$$

from which we assure that

$$\varepsilon = -\varepsilon_f + \eta(2\tau-\tau^2) > 0.$$

By concatenating (4.6) and (4.9) the inequality (4.5) follows, and we can conclude following the same reasoning as in Theorem 3.1. □

The property below basically says that the solutions of our trust-region subproblems tend to a step along the negative gradient almost surely, when the radius converges to zero (which is the probabilistic version of [15, Proposition 2.1]).

Property 4.5. Any sequence $\{(x_k, s_k, \Delta_k)\}$ is such that, for k sufficiently large

$$s_k \;=\; -\Delta_k D_k g_k,$$

with $D_k \in \mathbb{R}^{n\times n}$ satisfying

$$\lim_{k\to\infty} D_k \;=\; I.$$

In the next proposition, we show that our simple algorithm exhibits such a property.

Proposition 4.6. *Let Assumption 4.1 hold. Assume also that all trust-region subproblems (4.1) are solved up to optimality. Then Algorithm 2 generates sequences $\{(x_k, s_k, \Delta_k)\}$ satisfying Property 4.5 (for k sufficiently large and a.s.).*

Proof. The proof trivially follows from the proof of [15, Proposition 2.1] by considering that, by Theorem 4.4, $\sum_{k=0}^{\infty}\Delta_k^2 < \infty$ a.s. in Ω, and thus $\Delta_k \to 0$ a.s.. □

Following the same lines 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 4.7. *Let K be the set of indices of unsuccessful iterations (notice that K is random). Then under Assumption 4.2, Assumption 4.3 and (4.2) 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. Clearly it suffices to show, that for any given $m \in \mathbb{N}$ and a.s.

$$\liminf_{k\in K} \frac{f(x_k+s_k)-f(x_k)}{\|s_k\|} \geq -\frac{1}{m}\,. \tag{4.10}$$

To start with, by applying (A2') with $p = \varepsilon_q m^2 \|s_k\|^2$, we have

$$\mathbb{P}(|f_k - f_k^s - (f(x_k)-f(x_k+s_k))| \geq \frac{\|s_k\|}{m}) \leq m^2\|s_k\|^2\varepsilon_q\,.$$

From $\|s_k\| \leq \Delta_k$ and Theorem 4.4 we have that, for some $V \subset \Omega$ with $\mathbb{P}(V) = 1$, it holds

$$\sum_{k\in\mathbb{N}_0} \|s_k\|^2 < \infty\,. \tag{4.11}$$

On this subset, we have

$$\sum_{k\in\mathbb{N}_0} \mathbb{P}(|f_k - f_k^s - (f(x_k)-f(x_k+s_k))| \geq \frac{\|s_k\|}{m}) \leq \sum_{k\in\mathbb{N}_0} m^2\|s_k\|^2\varepsilon_q < \infty\,,$$

where we used (4.11) in the last inequality. In particular, reasoning as in the proof of Lemma 3.2 we have that a subset $V' \subset V$ exists such that $V' \to V$ and for every $w \in V'$

$$|f_k - f_k^s - (f(x_k)-f(x_k+s_k))| \leq \frac{\|s_k\|}{m}, \quad \text{for } k \text{ large enough.} \tag{4.12}$$

Then, for $k \in K$ large enough and a.s. (in particular on V'),

$$\frac{f(x_k+s_k)-f(x_k)}{\|s_k\|} \geq \frac{f_k^s - f_k - |f_k - f_k^s - (f(x_k)-f(x_k+s_k))|}{\|s_k\|} \geq -\theta\|s_k\| - \frac{1}{m}\,, \tag{4.13}$$

where we used (4.12) combined with the unsuccessful step condition of Algorithm 2 in the second inequality. Finally, (4.10) follows passing to the liminf for $k \to \infty$ in (4.13). □

We now prove that the Clarke generalized derivative is nonnegative along some limiting normalized trust-region step.

Lemma 4.8. *Assume that f is Lipschitz continuous with constant L_f^* around any limit point of the sequence of iterates $\{x_k\}$. Let Assumption 4.1 hold. Let $\{(x_k, s_k, \Delta_k)\}$ Let K be the index of unsuccessful iterations and $L \subseteq K$ be an index set such that*

$$\lim_{k\in L,\, k\to\infty} x_k = x^* \quad \textit{and} \tag{4.14}$$

$$\lim_{k\in L,\, k\to\infty} \frac{s_k}{\|s_k\|} = s^*. \tag{4.15}$$

Then, $f^\circ(x^*; s^*) \geq 0$.

Proof. For every $k \in L$ we have that

$$\frac{f(x_k + s_k) - f(x_k)}{\|s_k\|} = \frac{f(x_k + \|s_k\|[s_k/\|s_k\|]) - f(x_k)}{\|s_k\|} =$$
$$= \frac{f(x_k + \|s_k\|s^*) - f(x_k)}{\|s_k\|} - \frac{f(x_k + \|s_k\|s^*) - f(x_k + \|s_k\|[s_k/\|s_k\|])}{\|s_k\|}$$

By Lipschitz continuity, we have, for k large enough

$$\frac{|f(x_k + \|s_k\|s^*) - f(x_k + \|s_k\|[s_k/\|s_k\|])|}{\|s_k\|} \leq L_f^* \|[s_k/\|s_k\|] - s^*\|,$$

The proof is now completed as in the proof of Theorem 3.3 replacing $\Delta_k, g_k, \Delta_k d, \Delta_k g_k$ by respectively $\|s_k\|, s_k/\|s_k\|, \|s_k\|s^*, s_k$. □

Using Lemma 4.8, we finally prove the convergence result related to our trust-region method.

Theorem 4.9. *Assume that f is Lipschitz continuous with constant Ł$_f^*$ around any limit point of the sequence of iterates $\{x_k\}$. Let K be the set of indices of unsuccessful iterations. Let Assumptions 4.2 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 $\{g_k\}_{k\in L}$ 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. Recalling Proposition 4.6, we know that a.s. in Ω,

$$\lim_{k\to\infty} \left(\frac{s_k}{\|s_k\|} + g_k \right) = 0.$$

Let now d be a direction in the unitary sphere, and $S(w) \subset L(w)$ an index set such that

$$\lim_{k\in S(w),\, k\to\infty} g_k = -d\,,$$

so that

$$\lim_{k\in S(w),\, k\to\infty} \frac{s_k}{\|s_k\|} = d\,.$$

Then, by using Lemma 4.8, we have that a.s. $f^\circ(x^*; d) \geq 0$ concluding the proof. □

5 Concluding remarks and future work

This paper proposed new tail-bound probabilistic conditions for function estimation in stochastic derivative-free optimization. Those conditions, which are weaker than the usual assumptions for the potentially non-smooth case, allowed us to obtain convergence of both a direct-search and a trust-region method. Surprisingly, once proved that the series of squared step-sizes/radii is almost surely finite, a single tail-bound is sufficient to prove 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 within global optimization schemes, and numerical test on real world problems with noisy objectives.

A Appendix

A.1 Finite variance oracle

A common assumption in stochastic derivative-free optimization is that the stochastic oracle is exact in expected value and with bounded variance [14, 17]:

$$\begin{aligned} f(x) &= \mathbb{E}_\xi[F(x,\xi)]\,, \\ \mathrm{Var}_\xi[F(x,\xi)] &\leq V < +\infty\,. \end{aligned} \tag{A.1}$$

In other words, the objective is assumed to be the expected value of a random variable $F(x,\xi)$ parametrized by x, with the black-box oracle given by sampling on the ξ space. The random estimate f_k can then be computed by averaging on p_k i.i.d. samples $\{\xi_{k,i}\}_{i=1}^{p_k}$:

$$f_k = \frac{1}{p_k}\sum_{i=1}^{p_k} F(x_k,\xi_{k,i})\,,$$

and analogously f_k^g can be computed by averaging on p_k^g samples $\{\xi_{k,i}^g\}_{i=1}^{p_k^g}$.

We have that $\lceil V/(k_f^2\Delta_k^4)\rceil$ samples are enough to satisfy (2.1) and therefore in particular our conditions for $\varepsilon_f = 2k_f$ and $\varepsilon_q = 4k_f^2$, thanks to Proposition 2.3. Indeed for $p_k \geq \lceil V/(k_f^2\Delta_k^4)\rceil$ we have

$$\begin{aligned} \mathbb{E}[|f_k - f(x_k)|^2 \mid \mathcal{F}_{k-1}] =& \mathbb{E}\left[\left(\frac{1}{p_k}\sum_{i=1}^{p_k} F(x_k,\xi_{k,i}) - f(x_k)\right)^2 \mid \mathcal{F}_{k-1}\right] \\ =& \frac{1}{p_k}\mathbb{E}\left[\frac{1}{p_k}\sum_{i=1}^{p_k}(F(x_k,\xi_{k,i}) - f(x_k))^2 \mid \mathcal{F}_{k-1}\right] \\ =& \frac{1}{p_k}\mathrm{Var}[F(x_k,\xi)] \leq \frac{V}{p_k} \leq k_f^2\Delta_k^4\,, \end{aligned}$$

where we used the $\mathcal{F}_{k-1}$ measurability of p_k in the second equality, that $\{\xi_{k,i}\}_{i=1}^{p_k}$ are i.i.d. and also independent from $\mathcal{F}_{k-1}$ in the third equality, and the assumption (A.1). The inequality for f_k^g can be proved analogously.

A.2 Finite moment oracle

We now describe a weaker version of condition (A2) given in Assumption 2.2. This condition can be satisfied, together with (A1), when a finite r-th moment oracle, for some $r \in (1, 2]$, is used to build up function estimates. It is not difficult to see that Theorem 3.3 still holds under this version of the assumptions, with an analogous proof.

For convenience, we now rewrite (A1) by setting $\alpha = \frac{\varepsilon_f}{p}$, so that the assumption becomes

$$\mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \alpha \Delta_k^2 \,|\mathcal{F}_{k-1}) \leq \frac{\varepsilon_f}{\alpha}\,,$$

for every $\alpha \geq \varepsilon_f$. Analogously, we rewrite (A2) by setting $\alpha = \sqrt{\frac{\varepsilon_q}{p}}$:

$$\mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \alpha \Delta_k^2) \leq \frac{\varepsilon_q}{\alpha^2}\,,$$

for every $\alpha \geq \varepsilon_q$.

Consider now this generalization of (A2):

$$\mathbb{P}(|f_k - f_k^g - (f(x_k) - f(x_k + \Delta_k g_k))| \geq \alpha \Delta_k^h) \leq \frac{\varepsilon_q}{\alpha^{r(h)}}\,, \tag{A2'}$$

for $r(h) = \frac{2}{h-1}$, with a fixed $h \geq 2$. By classic results from probability theory (see, e.g., [18, Section 5]), when the stochastic oracle is exact in expected value and with finite $r(h)$-th moment, then both (A1) and (A2') can be satisfied with a large enough number of samples. Hence, the finite variance assumption discussed in Section A.1 is no longer needed.

A.3 Conditional Chebycheff's inequality

Thanks to the properties of conditional expectations, the conditional Chebycheff's inequality that we use in (2.4) can be proved in the same way as the standard Chebycheff's inequality. We include here the proof for completeness.

Proposition A.1. *Given random variables X, ϵ defined on $\mathbb{R}^n$ with $\epsilon > 0$ measurable with respect to a sub σ-field $\mathcal{F}$, we have*

$$\mathbb{P}(|X| \geq \epsilon \mid \mathcal{F}) \leq \frac{\mathbb{E}[|X| \mid \mathcal{F}]}{\epsilon}\,.$$

Proof. We have

$$\begin{aligned}\epsilon \mathbb{P}(|X| \geq \epsilon \mid \mathcal{F}) &= \epsilon \mathbb{E}[\mathbb{1}_{|X| \geq \epsilon} \mid \mathcal{F}] \\ &= \mathbb{E}[\epsilon \mathbb{1}_{|X| \geq \epsilon} \mid \mathcal{F}] \leq \mathbb{E}[|X| \mid \mathcal{F}]\,,\end{aligned}$$

where we used that ϵ is $\mathcal{F}$ measurable in the second equality and the monotonicity of the conditional expectation together with $\epsilon \mathbb{1}_{|X| \geq \epsilon} \leq |X|$ in the inequality. □

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*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, Switzerland, 2017.

[4] 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.

[5] 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.

[6] V. S. Borkar. *Probability Theory: An Advanced Course*. Springer Science & Business Media, New York, 2012.

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

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

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

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

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

[12] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Rev.*, 45:385–482, 2003.

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

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

[15] 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.

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

[17] 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.

[18] B. von Bahr and C.-G. Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36:299–303, 1965.