Noname manuscript No.
(will be inserted by the editor)

Extrapolation-based Direct Search for Nonsmooth Stochastic Zeroth-Order Optimization

A. Palmieri · F. Rinaldi · S. Shashaani

Received: date / Accepted: date

Abstract We propose and analyze a stochastic direct-search method for unconstrained zeroth-order minimization of locally Lipschitz, possibly nonsmooth, objectives. The method combines random polling directions with a stochastic extrapolating line search based on a sufficient-decrease test of order p. Under conditional accuracy assumptions on the stochastic estimates, we prove almost-sure convergence to Clarke stationary points. We further establish an expected iteration complexity bound. Specifically, using a supermartingale stopping-time argument, we prove that $\mathcal{O}\left(\max\left\{r^{-p}, \varepsilon^{-p/(p-1)}\right\}\right)$ iterations are sufficient in expectation to reach an (r, ε)-Goldstein stationary point. Moreover, we derive a corresponding expected tested-point complexity bound of order $\mathcal{O}\big(\varepsilon^{1-n}\max\{r^{-p}, \varepsilon^{-p/(p-1)}\}\big)$. To the best of our knowledge, this is the first convergence and expected-complexity analysis for an *extrapolation-based direct-search* method in a *nonsmooth stochastic* setting. Numerical experiments on a DFO benchmark suite highlight competitive performance against well-established stochastic direct-search methods.

Mathematics Subject Classification (2020) 90C56 · 90C15 · 49J52

A. Palmieri
Dipartimento di Matematica, Università di Padova, Italy
E-mail: anthony.palmieri@studenti.unipd.it

F. Rinaldi
Dipartimento di Matematica, Università di Padova, Italy
E-mail: rinaldi@math.unipd.it

S. Shashaani
Edward P. Fitts Department of Industrial and System Engineering, North Carolina State University, Raleigh, NC, USA
E-mail: sshasha2@ncsu.edu

1 Introduction

We study the unconstrained problem

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a *locally* Lipschitz (possibly nonsmooth) function that cannot be evaluated exactly. Instead, we only have access to a *stochastic zeroth-order oracle* that returns a random estimate $\tilde{f}(x)$ of $f(x)$ at any query point x. For example, the stochastic estimate can be modeled as a random variable parameterized by x, namely

$$\tilde{f}(x) \;=\; F(x,\zeta), \qquad \zeta \sim \mathbb{P},$$

i.e., the oracle draws a random seed ζ from an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and outputs $F(x,\zeta)$. Throughout the paper, all algorithmic quantities are random variables adapted to the natural filtration induced by the oracle and the algorithmic randomness (precise assumptions on the estimator, e.g., probabilistic accuracy, tails, and independence are stated in Section 3). Our goal is to design and analyze a *derivative-free* method that combines random dense directions with a stochastic line search, proving convergence to Clarke stationary points and establishing expected complexity bounds in terms of oracle calls.

1.1 Related Work

Derivative-free optimization (DFO) focuses on problems whose derivative information is unavailable, unreliable, or simply too expensive to obtain, requiring algorithms to work only with (possibly noisy) function values. Two main approaches are common in this context, which are the model-based methods and the direct-search methods. In model-based methods, one builds at each iteration a local surrogate intended to approximate f on a neighborhood of the current point, proposes a step by (approximately) minimizing the model inside that neighborhood, and accepts or rejects it by comparing predicted and observed decrease. Direct-search methods, by contrast, do not form an explicit model, but rather evaluate f on a structured set of trial points and update the current point based on a tailored acceptance test (e.g., sufficient-decrease of the objective function). The deterministic DFO theory is well established; see, e.g., the monograph of Conn, Scheinberg, and Vicente [6], and the survey of Larson, Menickelly, and Wild [17] for further details. Both model-based and direct-search schemes extend to stochastic zeroth-order oracles, but the analysis must then keep track of how estimation error affects acceptance decisions and step-size updates.

On the model-based side, trust-region methods based on probabilistic accuracy conditions [2] demonstrate that much of the classical trust-region mechanism

remains valid when local models and function estimates are sufficiently accurate with high probability, and martingale-based arguments play a central role in the theoretical analysis. This line of work ultimately led to the development of STORM-type frameworks, which systematically integrate randomized local models with stochastic function evaluations within a trust-region mechanism [4]. Expected-complexity guarantees for STORM-type methods can be derived through stopping-time and supermartingale techniques [3], and this analysis has been adapted both to random-subspace trust-region schemes for large-scale problems [10] and to stochastic line-search mechanisms [22]. In a separate framework for derivative-free stochastic trust-region schemes, adaptive sampling allows to drive down the estimation error according to a measure of stationarity, ensuring that the local models and function estimates are eventually sufficiently accurate almost surely; ASTRO-DF is a representative example of an algorithm built on this principle [25]. Subsequent work establishes almost-sure iteration and sample complexity results for ASTRO-DF and quantifies how variance-reduction devices, such as common random numbers (CRN), can improve sample complexity of these methods by replacing stringent per-point accuracy requirements with the difference accuracy requirements [13]. A related development uses similar difference error tail bounds within the supermartingle complexity analysis to obtain reduced sample sizes (with further benefits under CRN) while preserving convergence guarantees for both trust-region and direct-search methods [23].

On the direct-search side, StoMADS extends MADS to stochastic black-box objectives through probabilistic estimates and variance control, and uses a mesh-based strategy to obtain almost-sure convergence to Clarke stationary points [1], although with no complexity analysis. Constrained variants incorporate progressive barriers and probabilistic feasibility control [8]. For smooth objectives, the work by Dzahini [7] provides an expected-complexity analysis for stochastic direct search under a power-type sufficient-decrease condition, again via a supermartingale-based argument, and a recent survey summarizes current direct-search paradigms and guarantees, including stochastic settings and line-search variants [9]. Linesearch-based DFO with extrapolation under noisy oracles has also been studied recently, with convergence and complexity results in smooth regimes [24]. Another zeroth-order paradigm is randomized smoothing, where one replaces the objective by a smoothed surrogate and suitably build random estimators of the surrogate gradient. This idea is central to the random gradient-free framework of Nesterov and Spokoiny [21]. For nonsmooth nonconvex objectives, [18] was among the first works to connect randomized smoothing with Goldstein stationarity and to analyze gradient-free methods for reaching (r, ε)-Goldstein stationary points. This line was later sharpened by Kornowski and Shamir [15], who obtained optimal dimension-dependence for stochastic zero-order nonsmooth nonconvex optimization. Our work fits into the stochastic direct-search/line-search thread, but targets a fully nonsmooth, derivative-free setting: we couple dense random directions with a run-until-failure line-search policy. We prove almost-sure convergence in the spirit of [23], by relying on Clarke [5] and Goldstein/Lebourg nonsmooth

tools [11,14]. Moreover, we derive an expected-complexity result by leveraging the supermartingale framework proposed in [3], while working under minimal oracle assumptions.

1.2 Contributions

The proposed approach combines a line search strategy with random directions uniformly generated on the unit sphere. Under conditional tail-accuracy assumptions (motivated by the weak tail-bound framework introduced in [23]), we prove almost-sure convergence to Clarke stationary points and, by suitably adapting the analyses in [3,22], we derive an expected iteration complexity bound of order $\mathcal{O}(\max\{r^{-p}, \varepsilon^{-p/(p-1)}\})$ for reaching an (r,ε)-Goldstein stationary point. We also obtain the corresponding expected tested-point complexity bound $\mathcal{O}(\varepsilon^{1-n}\max\{r^{-p}, \varepsilon^{-p/(p-1)}\})$, where tested-point complexity counts the number of points at which stochastic function estimates are queried. The analysis relies exclusively on nonsmooth tools, that is Clarke and Goldstein subdifferentials and the Lebourg mean value theorem (see, e.g., [5,11, 14] for further details on these theoretical tools). To replace gradient-based descent estimates in the complexity argument, we test multiple *dense random directions* per iteration (see, e.g., [21]) and establish a geometric success guarantee: with probability strictly larger than $1/2$, at least one sampled direction falls in a spherical cap aligned with a Goldstein-descent direction. This spherical-cap mechanism is fundamentally different from the positive-spanning arguments used in smooth direct-search analyses (see, e.g., [7]). To the best of our knowledge, this is the first convergence and expected-complexity analysis for an *extrapolation-based direct-search* method in a *nonsmooth* setting. A recent work by De Santis, Liuzzi and Lucidi [24] studies a stochastic line-search scheme in a smooth setting, assuming continuously differentiable objectives with L-Lipschitz gradients and fixed-confidence accuracy for the function estimates; in that framework, the sampling strategy enforces a variance-stepsize coupling (of order Δ_k^4) and the resulting complexity guarantees are stated in terms of the expected gradient norm (e.g., $\mathcal{O}(\varepsilon^{-2})$). The present work addresses a complementary regime, focusing on merely locally Lipschitz, fully nonsmooth objectives and purely zeroth-order stochastic information. In contrast to analyses that assume a variance-stepsize coupling directly on the stochastic estimates, we impose conditional polynomial-tail accuracy conditions. We combine a line search with dense random directions and obtain convergence and expected-complexity results in terms of Clarke/Goldstein stationarity.

1.3 Paper Structure

The paper proceeds as follows. Section 2 presents the algorithm, the stochastic line search, and notation. Section 3 states the standing assumptions (compact-

ness, conditional independence, and p-tail). Section 4 proves a uniformly positive conditional expected merit-function decrease proportional to Δ_k^p, shows $\sum_k \Delta_k^p < \infty$ for $p > 1$ almost surely and hence $\Delta_k \to 0$ almost surely. This establishes convergence to Clarke stationary point almost surely. Section 5 proves that the algorithm reaches an (r, ε)-Goldstein stationary point in expected $\mathcal{O}\left(\max\left\{r^{-p}, \varepsilon^{-\frac{p}{p-1}}\right\}\right)$ iterations using a supermartingale stopping-time analysis. Furthermore, it derives the corresponding expected tested-point complexity. Section 6 presents numerical experiments showing that the proposed method is competitive with state-of-the-art stochastic derivative-free algorithms on a standard nonsmooth benchmark. Section 7 finally summarizes the theoretical results and discusses possible future extensions of the method. The appendix gathers auxiliary proofs.

2 Algorithm: Direct Search with Extrapolation for Nonsmooth Objectives

We consider a direct-search method with extrapolation for possibly nonsmooth stochastic zeroth-order optimization. The outer scheme is given in Algorithm 1, and the stochastic line-search subroutine is given in Algorithm 2. We fix $p \in (1, 2]$, $\theta > 0$, $\gamma \in (0, 1)$, $\bar{m} \in \mathbb{N}$, and $\bar{\imath} \in \mathbb{N}$. Random quantities are denoted by uppercase letters and their realizations by lowercase letters. At the beginning of iteration k, the current iterate and stepsize are $X_k \in \mathbb{R}^n$ and $\Delta_k > 0$. The algorithm then samples $\bar{m}$ candidate directions $D_{k,1}, \ldots, D_{k,\bar{m}} \in \mathbb{S}^{n-1}$, where we denote by $\mathbb{S}^{n-1} := \{d \in \mathbb{R}^n : \|d\| = 1\}$ the unit sphere in $\mathbb{R}^n$. We use two sigma-algebras. Let $\mathcal{G}_{k-1}$ denote the information available before sampling the directions at iteration k, so that X_k and Δ_k are $\mathcal{G}_{k-1}$-measurable. Conditionally on $\mathcal{G}_{k-1}$, the directions $D_{k,1}, \ldots, D_{k,\bar{m}}$ are sampled independently and uniformly on $\mathbb{S}^{n-1}$. After sampling them, but before drawing any stochastic function estimates at iteration k, we set

$$\mathcal{F}_{k-1} := \mathcal{G}_{k-1} \vee \sigma(D_{k,1}, \ldots, D_{k,\bar{m}}).$$

Thus X_k, Δ_k, and all directions $D_{k,m}$ are $\mathcal{F}_{k-1}$-measurable, while the estimates generated during iteration k are not. Oracle assumptions are stated conditionally on $\mathcal{F}_{k-1}$. For $m = 1, \ldots, \bar{m}$ and $i \geq 0$, define

$$\Delta_{k,i} := \gamma^{-i}\Delta_k, \qquad X_{k,m}^{(i)} := X_k + \Delta_{k,i} D_{k,m}.$$

We also set $X_k^{(-1)} := X_k$ and $\Delta_{k,-1} := 0$. The corresponding true and estimated values are

$$f_{k,m}^{(i)} := f(X_{k,m}^{(i)}), \qquad \tilde{f}_{k,m}^{(i)} := \tilde{f}(X_{k,m}^{(i)}), \qquad i \geq 0,$$

and

$$f_k^{(-1)} := f(X_k), \qquad \tilde{f}_k^{(-1)} := \tilde{f}(X_k).$$

At iteration k, the algorithm first samples $\bar{m}$ search directions independently, then it observes $\tilde{f}_k^{(-1)}$ and finally tests the directions $D_{k,1}, \dots, D_{k,\bar{m}}$ sequentially. Direction m is successful at depth $i \geq 0$ if

$$\tilde{f}_k^{(-1)} - \tilde{f}_{k,m}^{(i)} \geq \theta \Delta_{k,i}^p.$$

The algorithm accepts the first direction for which the line search succeeds. We denote its index by m_k^* and the last successful extrapolation depth by $H_k \geq 0$. If no direction succeeds, we set $H_k = -1$. Thus, on a successful iteration,

$$D_k := D_{k,m_k^*}, \qquad X_k^{(i)} := X_{k,m_k^*}^{(i)}, \qquad f_k^{(i)} := f_{k,m_k^*}^{(i)}, \qquad \tilde{f}_k^{(i)} := \tilde{f}_{k,m_k^*}^{(i)}.$$

The accepted step length is

$$\Delta_k^{\mathrm{acc}} := \begin{cases} 0, & H_k = -1, \\ \gamma^{-H_k} \Delta_k, & H_k \geq 0. \end{cases}$$

Hence $X_{k+1} = X_k + \Delta_k^{\mathrm{acc}} D_k$, with the convention that $X_{k+1} = X_k$ when $H_k = -1$. The stepsize update is

$$\Delta_{k+1} = \begin{cases} \gamma \Delta_k, & H_k = -1, \\ \gamma^{-1} \Delta_k, & H_k = 0, \\ \gamma^{-H_k} \Delta_k, & H_k \geq 1. \end{cases}$$

Equivalently, on successful iterations, $\Delta_{k+1} = \max\{\Delta_k^{\mathrm{acc}}, \gamma^{-1}\Delta_k\}$. Thus every unsuccessful iteration contracts the radius by γ, whereas every successful iteration expands the next radius by at least γ^{-1}.

Algorithm 1 DSE

Inputs: $x_0 \in \mathbb{R}^n$, $\delta_0 > 0$, $\theta > 0$, $\gamma \in (0,1)$, $\bar{m} \in \mathbb{N}$, $\bar{\imath} \in \mathbb{N}$
for $k = 0, 1, 2, \dots$ **do**
  Sample $d_{k,1}, \dots, d_{k,\bar{m}} \in \mathbb{S}^{n-1}$ independently
  Observe the baseline estimate $\tilde{f}_k^{(-1)} \leftarrow \tilde{f}(x_k)$
  for $m = 1$ **to** $\bar{m}$ **do**
    $(\delta^{\mathrm{acc}}, \delta^{\mathrm{next}}) \leftarrow$ **StochasticLinesearch** $\left(x_k, \delta_k, d_{k,m}, \theta, \gamma, \tilde{f}_k^{(-1)}, \bar{\imath}\right)$
    if $\delta^{\mathrm{acc}} > 0$ **then**
      break
    end if
  end for
  $d_k = d_{k,m}$
  $\delta_{k+1} \leftarrow \delta^{\mathrm{next}}$
  $x_{k+1} \leftarrow x_k + \delta^{\mathrm{acc}}\, d_k$
end for

Algorithm 2 StochasticLinesearch

1: **Input:** $x \in \mathbb{R}^n$, $\delta > 0$, $d \in \mathbb{R}^n$, $\theta > 0$, $\gamma \in (0,1)$, $\tilde{f}_k^{(-1)}, \bar{\imath}$
2: Set $\delta_0 = \delta$, $\delta_i = \gamma^{-i}\delta, \quad i = 0, \ldots, \bar{\imath}$
3: **if** $\tilde{f}(x + \delta_0 d) > \tilde{f}_k^{(-1)} - \theta\delta_0^p$ **then**
4: **return** $(0, \gamma\delta_0)$
5: **end if**
6: $h = \max\left\{j \in [0:\bar{\imath}] : \tilde{f}(x + \delta_i d) \leq \tilde{f}_k^{(-1)} - \theta\delta_i^p, \forall\, i \in [0:j]\right\}$
7: **if** $h = 0$ **then**
8: **return** $(\delta_0, \delta_0/\gamma)$
9: **else**
10: **return** (δ_h, δ_h)
11: **end if**

We would like to highlight that the maximization problem defined in Step 6 of the stochastic linesearch procedure is only a compact way of describing a sequential run-until-failure procedure. In practice, after the level $i = 0$ test succeeds, the line search evaluates the levels $i = 1, 2, \ldots, \bar{\imath}$ one at a time and stops as soon as the sufficient-decrease test fails. The returned value of h is the last consecutive level for which all tests from 0 up to h have succeeded. Thus, the algorithm does not evaluate extrapolation levels beyond the first failed test. The following remark finally clarifies the role of the maximum extrapolation depth $\bar{\imath}$ in Algorithm 2.

Remark 2.1 (Role of the extrapolation cap) The extrapolation cap $\bar{\imath}$ is mainly an analytical device and is not restrictive in practice. Indeed, any implementation is run under a finite oracle-call budget, and $\bar{\imath}$ can be chosen as large as this budget allows. Its role in the analysis is to keep the extrapolation search tree uniformly finite, which yields a simple maximal-error bound and a direct tested-point complexity estimate (See Sections 3-5).

3 Assumptions and Preliminaries

We now state the main assumptions on the objective and the stochastic oracle, deriving the key error bounds needed for the analysis. Throughout, we condition on $\mathcal{F}_{k-1}$, which contains X_k, Δ_k, and the sampled directions $D_{k,1}, \ldots, D_{k,\bar{m}}$, but not the stochastic estimates generated at iteration k. We assume noisy function estimates with the following basic structure.

Assumption 3.1 (Compactness of the iterates) *There exists a nonempty compact set $\mathcal{C} \subset \mathbb{R}^n$ such that $X_k \in \mathcal{C}$ a.s. for all $k \geq 0$. Moreover, the initial stepsize $\Delta_0 > 0$ is deterministic and finite.*

Remark 3.1 (Compactness of the tested points) Under the fixed maximum extrapolation depth $\bar{\imath} \in \mathbb{N}_0$, and Assumption 3.1, all tested points lie in a compact enlargement of $\mathcal{C}$. Indeed, let

$$\Delta_{\mathcal{C}} := \operatorname{diam}(\mathcal{C}), \qquad \bar{\Delta} := \max\{\Delta_0, \gamma^{-1}\Delta_{\mathcal{C}}\}.$$

We first prove that $\Delta_k \le \bar{\Delta},\ \forall k \ge 0$. The claim is true for $k = 0$ by the definition of $\bar{\Delta}$. Suppose that $\Delta_k \le \bar{\Delta}$. If iteration k is unsuccessful, then

$$\Delta_{k+1} = \gamma \Delta_k \le \Delta_k \le \bar{\Delta}.$$

If iteration k is successful with $H_k = 0$, then $X_{k+1} = X_k + \Delta_k D_k$ is the accepted point. Since $X_k, X_{k+1} \in \mathcal{C}$, we have

$$\Delta_k = \|X_{k+1} - X_k\| \le \Delta_{\mathcal{C}}.$$

Therefore,

$$\Delta_{k+1} = \gamma^{-1}\Delta_k \le \gamma^{-1}\Delta_{\mathcal{C}} \le \bar{\Delta}.$$

Finally, if iteration k is successful with $H_k \ge 1$, then $X_{k+1} = X_k + \gamma^{-H_k}\Delta_k D_k$, and hence

$$\Delta_{k+1} = \gamma^{-H_k}\Delta_k = \|X_{k+1} - X_k\| \le \Delta_{\mathcal{C}} \le \bar{\Delta}.$$

Thus, by induction, $\Delta_k \le \bar{\Delta},\ \forall k \ge 0$. Consequently, for every tested point,

$$\|X_{k,m}^{(i)} - X_k\| = \gamma^{-i}\Delta_k \le \gamma^{-\bar{\imath}}\bar{\Delta}.$$

Therefore

$$X_{k,m}^{(i)} \in \mathcal{C}_{enl} := \mathcal{C} + \gamma^{-\bar{\imath}}\bar{\Delta}\,\mathbb{B}, \qquad 1 \le m \le \bar{m}, \quad 0 \le i \le \bar{\imath}.$$

Here, $\mathbb{B} := \{z \in \mathbb{R}^n : \|z\| \le 1\}$ denotes the closed unit ball in $\mathbb{R}^n$, and the symbol $+$ denotes the Minkowski sum of sets, namely $A + B := \{a + b : a \in A,\ b \in B\}$. Since $\mathcal{C}$ is compact and $\gamma^{-\bar{\imath}}\bar{\Delta}\,\mathbb{B}$ is compact, the set $\mathcal{C}_{enl}$ is compact. Hence, all iterates and all tested trial points generated by the algorithm lie in the compact set $\mathcal{C}_{enl}$.

Assumption 3.2 (Conditional independence) *At iteration k, conditionally on $\mathcal{F}_{k-1}$, the stochastic function estimates associated with the base point and with the search tree,*

$$\tilde{f}_k^{(-1)} \quad \textit{and} \quad \left\{\tilde{f}_{k,m}^{(i)} : 1 \le m \le \bar{m},\ 0 \le i \le \bar{\imath}\right\},$$

are mutually independent.

Assumption 3.3 (p-tail for single-point errors) *There exists $\varepsilon_h > 0$ such that, for every $\alpha \ge \varepsilon_h$:*

$$\begin{aligned}
&\mathbb{P}\Big(|\tilde{f}_k^{(-1)} - f_k^{(-1)}| \ge \alpha\,\Delta_k^p \;\Big|\; \mathcal{F}_{k-1}\Big) \le \frac{\varepsilon_h}{\alpha^{p/(p-1)}},\\
&\mathbb{P}\Big(|\tilde{f}_{k,m}^{(i)} - f_{k,m}^{(i)}| \ge \alpha\,\Delta_k^p \;\Big|\; \mathcal{F}_{k-1}\Big) \le \frac{\varepsilon_h}{\alpha^{p/(p-1)}}, \quad i \in [0:\bar{\imath}],\ m \in [1:\bar{m}].
\end{aligned} \tag{2}$$

To simplify the notation, from now on we set $\bar{E}^{(i)}_{k,m} := \tilde{f}^{(i)}_{k,m} - f^{(i)}_{k,m}$, with $\bar{E}^{(-1)}_{k} := \tilde{f}^{(-1)}_{k} - f^{(-1)}_{k}$, and define the error difference estimator

$$E^{(i)}_{k,m} := (\tilde{f}^{(-1)}_{k} - \tilde{f}^{(i)}_{k}) - (f^{(-1)}_{k} - f^{(i)}_{k}) = \bar{E}^{(-1)}_{k} - \bar{E}^{(i)}_{k,m},$$

Assumptions 3.1 and 3.2 are standard in stochastic derivative-free optimization (see, e.g., [1,23]). In particular, compactness of the iterate set allows us to invoke tools from nonsmooth analysis (such as the Clarke subdifferential) and ensures, under mild regularity (e.g., continuity on $\mathcal{C}_{enl}$), that the objective f attains its minimum. For notational simplicity, since our analysis will be restricted to a compact set, we will often write f is L-Lipschitz. Assumptions 3.3, instead, is a technical condition that provides quantitative control of the stochastic estimation error $\tilde{f}$ and is needed to establish the merit-function drift bounds used throughout the analysis. From this basic set of hypotheses we derive several consequences that will be used in the convergence and complexity proofs. For readability, we list the statements here and defer the proofs to the Appendix A. Similarly, from Assumption 3.3 we can get several other useful results by simply integrating the tails.

Lemma 3.2 (Conditional Mean Bound) *Let Assumption 3.3 hold. For each k, $i \in [-1 : \bar{\imath}]$ and $m \in [1 : \bar{m}]$ recall the estimation error $\bar{E}^{(i)}_{k,m} := \tilde{f}^{(i)}_{k,m} - f^{(i)}_{k,m}$. Then, there exists a constant $\mu_h > 0$ such that the following bound on the conditional mean holds, uniformly in k and i*

$$\mathbb{E}\Big[\, |\bar{E}^{(i)}_{k,m}| \;\Big|\; \mathcal{F}_{k-1} \Big] \;\le\; \mu_h\, \Delta_k^p,$$

with μ_h a suitably chosen positive parameter.

From the previous lemma we can also bound the conditional expectation of the difference estimator $E^{(i)}_{k,m}$.

Lemma 3.3 (Conditional L^1 bound for the error difference estimator) *Under Assumption 3.3, for every outer iteration k and every inner index $i \in \{-1, 0, 1, \dots, \bar{\imath}\}$,*

$$\mathbb{E}\Big[\, |E^{(i)}_{k,m}| \,\Big|\, \mathcal{F}_{k-1} \Big] \;\le\; 2\,\mu_h\, \Delta_k^p \qquad a.s., \tag{3}$$

with μ_h as in Lemma 3.2. Consequently, the signed conditional mean is also controlled:

$$\Big|\mathbb{E}\Big[\, E^{(i)}_{k,m} \,\Big|\, \mathcal{F}_{k-1} \Big]\,\Big| \;\le\; 2\,\mu_h\, \Delta_k^p \qquad a.s.. \tag{4}$$

Lemma 3.4 (Bounded maximal iteration error) *At iteration k, after sampling the trial directions $D_{k,1}, \dots, D_{k,\bar{m}}$, define the maximal difference-estimation error over the potential search tree by*

$$E_k^{\max} := \max_{1 \le m \le \bar{m}} \max_{0 \le i \le \bar{\imath}} \Big| E^{(i)}_{k,m} \Big|.$$

If Assumption 3.3 holds, then via Lemma 3.3, one has

$$\mathbb{E}\left[E_k^{\max} \mid \mathcal{F}_{k-1}\right] \leq c_{\max}\Delta_k^p \qquad a.s.,$$

where $c_{\max} = 2\mu_h \bar{m}(\bar{\imath}+1)$.

Finally, we note that the conditional p-tail bound in Assumption 3.3 directly implies the commonly used notion of β-probabilistically ε_f-accurate function estimates, as made precise in the following remark.

Remark 3.5 (From p-tails to β-probabilistic ε_f-accuracy) Assumption 3.3 states that, for some $\varepsilon_h > 0$ and all $i \in \{-1, 0, 1, \ldots, \bar{\imath}\}$,

$$\mathbb{P}\Big(|\tilde{f}_k^{(i)} - f_k^{(i)}| \;\geq\; \alpha\,\Delta_k^p \;\Big|\; \mathcal{F}_{k-1}\Big) \;\leq\; \frac{\varepsilon_h}{\alpha^{p/(p-1)}} \qquad \text{for all } \alpha \geq \varepsilon_h.$$

Fix a target confidence $\beta \in (0,1)$. Choose

$$\varepsilon_f \;\geq\; \max\left\{\varepsilon_h,\; \left(\frac{\varepsilon_h}{1-\beta}\right)^{\frac{p-1}{p}}\right\}.$$

Then, setting $\alpha = \varepsilon_f$ in the tail bound gives

$$\mathbb{P}\Big(|\tilde{f}_k^{(i)} - f_k^{(i)}| \;\leq\; \varepsilon_f\,\Delta_k^p \;\Big|\; \mathcal{F}_{k-1}\Big) \;\geq\; 1 - \frac{\varepsilon_h}{\varepsilon_f^{p/(p-1)}} \;\geq\; \beta,$$

i.e., the usual β-probabilistically ε_f-accurate condition holds uniformly in i. If the tail in Assumption 3.3 holds for all $\alpha > 0$, the simpler choice $\varepsilon_f = (\varepsilon_h/(1-\beta))^{\frac{p-1}{p}}$ suffices.

We conclude by observing that Assumption 3.3 is not restrictive in practice. Indeed, standard sampling strategies in stochastic derivative-free optimization often rely on averaging multiple independent realizations of the stochastic oracle; see, for example, [1,4,25]. Under the classical assumptions commonly imposed on such sample-average estimators, one obtains the stronger conditions typically adopted in the stochastic derivative-free optimization literature, which in turn imply Assumption 3.3. For completeness, Appendix B establishes this connection for a standard sample-average estimator and derives the corresponding batch-size requirements.

4 Main Convergence Result

Following the trust-region/direct-search literature (see,e.g., [16,23]), we analyze progress through the *merit function*

$$\Phi_k \;=\; f(X_k) - f_{min} \;+\; \eta\,\Delta_k^p, \tag{5}$$

where $f_{\min}$ is a lower bound for f on the compact region containing the iterates, and η is a positive parameter. The term $f(X_k) - f^\star$ measures objective decrease, while $\eta\,\Delta_k^p$ penalizes overly aggressive step-size growth during

successful extrapolations. This choice ensures that the *merit-function drift* $\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}]$ trades off true decrease against step-size changes in a way that yields a uniform lower bound proportional to Δ_k^p. At iteration k, define $H_k \in \{-1, 0, \dots, \bar{\imath}\}$, where $H_k = -1$ denotes failure of the whole iteration, while $H_k = i \geq 0$ denotes success at extrapolation depth i. For $i = 0, \dots, \bar{\imath}$, set

$$\Delta_{k,i} := \gamma^{-i}\Delta_k, \qquad \pi_i := \mathbb{P}(H_k = i \mid \mathcal{F}_{k-1}),$$

and also

$$\pi_{-1} := \mathbb{P}(H_k = -1 \mid \mathcal{F}_{k-1}).$$

Thus

$$\pi_{-1} + \sum_{i=0}^{\bar{\imath}} \pi_i = 1.$$

On a successful iteration, let m_k^* be the accepted direction index, so that

$$X_{k+1} = X_{k,m_k^*}^{(H_k)} = X_k + \Delta_{k,H_k} D_{k,m_k^*}.$$

On a failed iteration, $X_{k+1} = X_k$. The stepsize update is

$$\Delta_{k+1} = \begin{cases} \gamma\Delta_k, & H_k = -1, \\ \gamma^{-1}\Delta_k, & H_k = 0, \\ \gamma^{-i}\Delta_k, & H_k = i \geq 1. \end{cases}$$

Recall the maximal difference-estimation error

$$E_k^{\max} := \max_{1 \leq m \leq \bar{m}} \max_{0 \leq i \leq \bar{\imath}} \left| \left(\tilde{f}_k^{(-1)} - \tilde{f}_{k,m}^{(i)} \right) - \left(f_k^{(-1)} - f_{k,m}^{(i)} \right) \right|.$$

Lemma 4.1 (Merit-function drift) *Let Assumption 3.3 hold. Then, by Lemma 3.4, there exists $c_{\max} > 0$ such that*

$$\mathbb{E}[E_k^{\max} \mid \mathcal{F}_{k-1}] \leq c_{\max}\Delta_k^p \qquad \textit{a.s.}$$

Assume that

$$\eta > \frac{c_{\max}}{1 - \gamma^p}, \tag{6}$$

and

$$\theta \geq \eta \max\{1, \gamma^{-p} - \gamma^p\}. \tag{7}$$

Then

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}] \geq \left(\eta(1 - \gamma^p) - c_{\max}\right) \Delta_k^p > 0 \qquad \text{a.s.} \tag{8}$$

Proof All expectations are taken conditionally on $\mathcal{F}_{k-1}$. We decompose

$$\Phi_k - \Phi_{k+1} = f(X_k) - f(X_{k+1}) + \eta(\Delta_k^p - \Delta_{k+1}^p).$$

Step 1: True-reduction term. Fix $i \in [0 : \bar{\imath}]$. Define

$$E_{k,m_k^*}^{(i)} := \big(\tilde{f}_k - \tilde{f}_{k,m_k^*}^{(i)}\big) - \big(f_k - f_{k,m_k^*}^{(i)}\big).$$

Then

$$f_k - f_{k,m_k^*}^{(i)} = \big(\tilde{f}_k - \tilde{f}_{k,m_k^*}^{(i)}\big) - E_{k,m_k^*}^{(i)}.$$

On $\{H_k = i\}$, the sufficient decrease holds, so

$$\tilde{f}_k - \tilde{f}_{k,m_k^*}^{(i)} \geq \theta \Delta_{k,i}^p.$$

Therefore

$$\mathbf{1}_{\{H_k=i\}}\big(f_k - f_{k,m_k^*}^{(i)}\big) \geq \mathbf{1}_{\{H_k=i\}}\Big(\theta\Delta_{k,i}^p - E_{k,m_k^*}^{(i)}\Big).$$

Taking conditional expectations and using

$$\Delta_{k,i}^p = \gamma^{-ip}\Delta_k^p, \qquad \mathbb{E}[\mathbf{1}_{\{H_k=i\}} \mid \mathcal{F}_{k-1}] = \pi_i,$$

we obtain

$$\mathbb{E}\Big[\mathbf{1}_{\{H_k=i\}}\big(f_k - f_{k,m_k^*}^{(i)}\big) \,\Big|\, \mathcal{F}_{k-1}\Big] \geq \pi_i\theta\Delta_{k,i}^p - \mathbb{E}\Big[\mathbf{1}_{\{H_k=i\}}E_{k,m_k^*}^{(i)} \,\Big|\, \mathcal{F}_{k-1}\Big].$$

Since $|E_{k,m_k^*}^{(i)}| \leq E_k^{\max}$, it follows that

$$\mathbb{E}\Big[\mathbf{1}_{\{H_k=i\}}\big(f_k - f_{k,m_k^*}^{(i)}\big) \,\Big|\, \mathcal{F}_{k-1}\Big] \geq \pi_i\theta\Delta_{k,i}^p - \mathbb{E}\big[\mathbf{1}_{\{H_k=i\}}E_k^{\max} \,\big|\, \mathcal{F}_{k-1}\big].$$

Summing over $i \in [0 : \bar{\imath}]$ gives

$$\sum_{i=0}^{\bar{\imath}} \mathbb{E}\Big[\mathbf{1}_{\{H_k=i\}}\big(f_k - f_{k,m_k^*}^{(i)}\big) \,\Big|\, \mathcal{F}_{k-1}\Big] \geq \sum_{i=0}^{\bar{\imath}} \pi_i\theta\Delta_{k,i}^p - \mathbb{E}\big[\mathbf{1}_{\{H_k\geq 0\}}E_k^{\max} \,\big|\, \mathcal{F}_{k-1}\big].$$

By Lemma 3.4,

$$\mathbb{E}\big[\mathbf{1}_{\{H_k\geq 0\}}E_k^{\max} \,\big|\, \mathcal{F}_{k-1}\big] \leq \mathbb{E}[E_k^{\max} \mid \mathcal{F}_{k-1}] \leq c_{\max}\Delta_k^p.$$

Hence

$$\sum_{i=0}^{\bar{\imath}} \mathbb{E}\Big[\mathbf{1}_{\{H_k=i\}}\big(f_k - f_{k,m_k^*}^{(i)}\big) \,\Big|\, \mathcal{F}_{k-1}\Big] \geq \left(\sum_{i=0}^{\bar{\imath}} \pi_i\theta\gamma^{-pi} - c_{\max}\right)\Delta_k^p. \tag{9}$$

Step 2: Stepsize term. By the modified stepsize update,

$$\Delta_{k+1} = \mathbf{1}_{\{H_k=-1\}}(\gamma\Delta_k) + \mathbf{1}_{\{H_k=0\}}(\gamma^{-1}\Delta_k) + \sum_{i=1}^{\bar{\imath}} \mathbf{1}_{\{H_k=i\}}(\gamma^{-i}\Delta_k).$$

Therefore

$$\eta(\Delta_k^p - \Delta_{k+1}^p) = \eta\Delta_k^p \Bigg[\mathbf{1}_{\{H_k=-1\}}(1-\gamma^p) + \mathbf{1}_{\{H_k=0\}}(1-\gamma^{-p}) + \\ + \sum_{i=1}^{\bar{\imath}} \mathbf{1}_{\{H_k=i\}}\big(1-\gamma^{-pi}\big) \Bigg].$$

Taking conditional expectations yields

$$\mathbb{E}\big[\eta(\Delta_k^p - \Delta_{k+1}^p) \,\big|\, \mathcal{F}_{k-1}\big] = \eta\Delta_k^p \Bigg[\pi_{-1}(1-\gamma^p) + \pi_0(1-\gamma^{-p}) + \\ + \sum_{i=1}^{\bar{\imath}} \pi_i(1-\gamma^{-pi}) \Bigg]. \tag{10}$$

Step 3: Combining the two estimates. Combining (9) and (10), and recalling that the true-reduction term is zero on $\{H_k = -1\}$, we obtain

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \,|\, \mathcal{F}_{k-1}] \geq \Bigg[\pi_{-1}\eta(1-\gamma^p) + \pi_0\Big(\theta + \eta(1-\gamma^{-p})\Big) + \\ + \sum_{i=1}^{\bar{\imath}} \pi_i\Big(\theta\gamma^{-pi} + \eta(1-\gamma^{-pi})\Big) - c_{\max} \Bigg] \Delta_k^p. \tag{11}$$

Define

$$A := \eta(1-\gamma^p), \quad B := \theta + \eta(1-\gamma^{-p}) = \theta - \eta(\gamma^{-p} - 1),$$

and, for $i \geq 1$,

$$C_i := \theta\gamma^{-pi} + \eta(1-\gamma^{-pi}) = \eta + (\theta-\eta)\gamma^{-pi}.$$

Then, under (7),

$$B \geq A, \qquad C_i \geq A \quad \forall i \geq 1.$$

Indeed, $B \geq A$ is equivalent to

$$\theta - \eta(\gamma^{-p} - 1) \geq \eta(1-\gamma^p),$$

that is,

$$\theta \geq \eta(\gamma^{-p} - \gamma^p),$$

which follows from (7). Moreover, (7) also implies $\theta \geq \eta$. Hence, for $i \geq 1$,

$$C_i = \eta + (\theta-\eta)\gamma^{-pi} \geq \eta \geq \eta(1-\gamma^p) = A.$$

Therefore

$$\pi_{-1}A + \pi_0 B + \sum_{i=1}^{\bar{\imath}} \pi_i C_i \geq \left(\pi_{-1} + \pi_0 + \sum_{i=1}^{\bar{\imath}} \pi_i\right) A = A.$$

Using this in (11), we obtain

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \,|\, \mathcal{F}_{k-1}] \geq (A - c_{\max})\, \Delta_k^p.$$

Substituting $A = \eta(1-\gamma^p)$ gives

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \,|\, \mathcal{F}_{k-1}] \geq (\eta(1-\gamma^p) - c_{\max})\, \Delta_k^p.$$

By (6), the coefficient is strictly positive. This concludes the proof.

Lemma 4.2 *Under the hypotheses of Lemma 4.1, it holds*

$$\sum_{k=0}^{\infty} \Delta_k^p < \infty \qquad a.s.$$

In particular,

$$\Delta_k \to 0 \qquad a.s.\ as\ k \to \infty.$$

Proof Define $\Theta := \eta(1-\gamma^p) - c_{\max},\ \Theta > 0$. Taking total expectations in the drift inequality gives

$$\mathbb{E}[\Phi_k - \Phi_{k+1}] \geq \Theta\, \mathbb{E}[\Delta_k^p] \qquad \forall k \in \mathbb{N}_0.$$

Summing from $k = 0$ to N yields

$$\Theta \sum_{k=0}^{N} \mathbb{E}[\Delta_k^p] \leq \sum_{k=0}^{N} \mathbb{E}[\Phi_k - \Phi_{k+1}] = \mathbb{E}[\Phi_0 - \Phi_{N+1}].$$

Since

$$\Phi_{N+1} = f_{N+1} - f_{\min} + \eta \Delta_{N+1}^p \geq 0,$$

we obtain

$$\Theta \sum_{k=0}^{N} \mathbb{E}[\Delta_k^p] \leq \mathbb{E}[\Phi_0].$$

Therefore,

$$\sum_{k=0}^{N} \mathbb{E}[\Delta_k^p] \leq \frac{\mathbb{E}[\Phi_0]}{\Theta} \qquad \forall N \in \mathbb{N}_0.$$

Letting $N \to \infty$ and using monotone convergence,

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \Delta_k^p\right] = \sum_{k=0}^{\infty} \mathbb{E}[\Delta_k^p] \leq \frac{\mathbb{E}[\Phi_0]}{\Theta} < \infty.$$

Since the random variable $\sum_{k=0}^{\infty} \Delta_k^p$ is nonnegative and has finite expectation, it is finite almost surely. Hence $\sum_{k=0}^{\infty} \Delta_k^p < \infty$ almost surely. Finally, since $\Delta_k^p \geq 0$ and $p > 1$, summability implies $\Delta_k^p \to 0$ a.s. and therefore $\Delta_k \to 0$ a.s. because $p > 0$.

Thus, from the merit-function drift we have $\sum_{k=0}^{\infty} \Delta_k^p < \infty$ a.s., hence $\Delta_k \to 0$ along the sequence almost surely. This in turn guarantees the existence of a subsequence of unsuccessful iterations.

Lemma 4.3 (Existence of infinitely many unsuccessful iterations) *Let*

$$\mathcal{U} := \{k \in \mathbb{N}_0 : \textit{iteration } k \textit{ is unsuccessful}\}$$

be the random set of indices of unsuccessful iterations. Assume that

$$\sum_{k=0}^{\infty} \Delta_k^p < \infty \qquad \textit{a.s.}$$

Then, with probability one, the set $\mathcal{U}$ *is infinite.*

Proof Argue by contradiction. Suppose that, with strictly positive probability, there are only finitely many unsuccessful iterations. On this event, there exists a finite random index k^u such that every iteration $k \geq k^u$ is successful. By the stepsize update rule, on a successful iteration we have

$$\Delta_{k+1} = \begin{cases} \gamma^{-1}\Delta_k, & H_k = 0, \\ \gamma^{-H_k}\Delta_k, & H_k \geq 1. \end{cases}$$

Since $0 < \gamma < 1$, it follows in both cases that

$$\Delta_{k+1} \geq \gamma^{-1}\Delta_k > \Delta_k.$$

In particular, on the event under consideration, the tail sequence $\{\Delta_k\}_{k \geq k^u}$ is strictly increasing. Since $\Delta_{k^u} > 0$, we get

$$\Delta_k \geq \Delta_{k^u} > 0 \qquad \forall k \geq k^u.$$

Therefore

$$\sum_{k=k^u}^{\infty} \Delta_k^p \geq \sum_{k=k^u}^{\infty} \Delta_{k^u}^p = \infty,$$

which contradicts the assumption

$$\sum_{k=0}^{\infty} \Delta_k^p < \infty \qquad \text{a.s.}$$

Hence the probability that $\mathcal{U}$ is finite must be zero. Therefore $\mathcal{U}$ is infinite almost surely.

It is on these unsuccessful iterations that our convergence analysis will concentrate.

We say that a subsequence $\{x_k\}_{k \in L}$ is refining if it converges to x^* and if $\{d_k\}_{k \in L}$ is dense in the unit sphere. We will show that under (6), (7), if Assumptions 3.1, 3.2, 3.3 hold, then with probability 1 all refining subsequences converge to a Clarke stationary point. We start by recalling the definition of Clarke generalized directional derivative and Clarke subdifferential.

Definition 4.4 (Clarke generalized directional derivative) Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near x. Its *Clarke generalized directional derivative* at x in the direction h is

$$f^\circ(x; h) := \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + t\,h) - f(y)}{t}.$$

Definition 4.5 (Clarke subdifferential) Under the same hypotheses, the *Clarke subdifferential* of f at x is

$$\partial_C f(x) := \big\{ v \in \mathbb{R}^n : f^\circ(x; h) \;\geq\; \langle v,\, h \rangle \quad \forall\, h \in \mathbb{R}^n \big\}.$$

Equivalently, one shows

$$\partial_C f(x) = \operatorname{conv}\Big\{ \lim_{i \to \infty} \nabla f(x_i) : x_i \to x,\ f \text{ is differentiable at } x_i \Big\}.$$

Definition 4.6 (Clarke-stationary point) Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and let $\partial_C f(x)$ denote its Clarke subdifferential. A point $x^\star \in \mathbb{R}^n$ is *Clarke-stationary* if and only if $0 \in \partial_C f(x^\star)$. Equivalently, the Clarke directional derivative satisfies

$$f^\circ(x^\star; d) \;\geq\; 0 \qquad \text{for all } d \in \mathbb{R}^n.$$

Now some preliminary results. The next lemma extends Assumption 3.3 to the case where α is a random variable.

Lemma 4.7 *Let A be an $\mathcal{F}_{k-1}$-measurable random variable such that $A \geq \varepsilon_h$ almost surely. If Assumption 3.3 holds, then, for every $i \in [0 : \bar{\imath}]$ and $m \in [1 : \bar{m}]$, almost surely*

$$\mathbb{P}\left(|\tilde{f}_k^{(-1)} - f_k^{(-1)}| \geq A \Delta_k^p \,\middle|\, \mathcal{F}_{k-1} \right) \leq \frac{\varepsilon_h}{A^{p/(p-1)}},$$
$$\mathbb{P}\Big(|\tilde{f}_{k,m}^{(i)} - f_{k,m}^{(i)}| \geq A\, \Delta_k^p \;\Big|\; \mathcal{F}_{k-1} \Big) \leq \frac{\varepsilon_h}{A^{p/(p-1)}}.$$

Proof Similar to [23, Lemma 2.3]. It is enough to test the desired conditional inequality against arbitrary events $G \in \mathcal{F}_{k-1}$. For simple thresholds $A = \sum_j a_j \mathbf{1}_{G_j}$, with $G_j \in \mathcal{F}_{k-1}$ and $a_j \geq \varepsilon_h$, the claim follows by applying Assumption 3.3 on each $G \cap G_j$. The general case $A \geq \varepsilon_h$ follows by taking simple $\mathcal{F}_{k-1}$-measurable thresholds A_n such that $\varepsilon_h \leq A_n \leq A$ and $A_n \uparrow A$, and then passing to the limit by dominated convergence.

To prove $0 \in \partial_C f(\bar{x})$ at a limit point $\bar{x}$ of $\{X_k\}$, we proceed in two steps. First, we show that along the (infinite) set $\mathcal{U}$ of *unsuccessful* iterations the forward difference in the sampled direction is asymptotically nonnegative,

$$\liminf_{k \in \mathcal{U},\ k \to \infty} \frac{f(X_k + \Delta_k D_k) - f(X_k)}{\Delta_k} \;\geq\; 0 \quad \text{a.s.}$$

which is Lemma 4.8 below. Second, we exploit that the distribution of D_k has *dense support* on the unit sphere: for any fixed unit vector u, there exist indices

$k_j \in \mathcal{U}$ with $D_{k_j} \to u$ and $\Delta_{k_j} \to 0$. Passing to the limit and using the outer semicontinuity of the Clarke directional derivative then yields $f^\circ(\bar{x}; u) \geq 0$ for every u, which is equivalent to $0 \in \partial_C f(\bar{x})$. The next lemma establishes the first step.

Lemma 4.8 *Let $\mathcal{U}$ be the (random) set of indices of unsuccessful iterations. Under Assumptions 3.1, 3.2, 3.3, and the stepsize/merit conditions ensuring $\sum_k \mathbb{E}[\Delta_k^p] < \infty$ (hence $\Delta_k \to 0$ a.s.), we have a.s.*

$$\liminf_{k \in \mathcal{U},\, k \to \infty} \frac{f(X_k + \Delta_k D_k) - f(X_k)}{\Delta_k} \;\geq\; 0.$$

Proof Fix $l \in \mathbb{N}$. On an unsuccessful outer iteration $k \in \mathcal{U}$, the inner extrapolation does not start and for every tested direction $D_{k,m}$, with $m \in [1 : \bar{m}]$, the level $i = 0$ sufficient-decrease test fails. Since, on unsuccessful iterations, no direction is accepted, we define the direction used in the refining subsequence as the last tested direction, namely $D_k := D_{k,\bar{m}}$. Hence

$$\tilde{f}_k^{(-1)} = \tilde{f}(X_k), \qquad \tilde{f}_k^{(0)} = \tilde{f}(X_k + \Delta_k D_k).$$

Because the iteration is unsuccessful, the sufficient-decrease test fails at level $i = 0$ for this direction. Therefore,

$$\tilde{f}_k^{(0)} - \tilde{f}_k^{(-1)} > -\theta \Delta_k^p. \tag{12}$$

Apply Assumption 3.3 with the $\mathcal{F}_{k-1}$-measurable threshold $\Upsilon_k := \frac{1}{2l}\Delta_k^{1-p}$ (note $\Upsilon_k \geq \varepsilon_h$ for all k large enough since $\Delta_k \to 0$ a.s.). Conditionally on $\mathcal{F}_{k-1}$,

$$\mathbb{P}\left(|\tilde{f}_k^{(i)} - f_k^{(i)}| \;\geq\; \frac{\Delta_k}{2l} \;\middle|\; \mathcal{F}_{k-1} \right) \;\leq\; \frac{\varepsilon_h}{\Upsilon_k^{p/(p-1)}} \;=\; \varepsilon_h \, (2l)^{p/(p-1)} \, \Delta_k^p, \; i \in \{-1, 0\}.$$

Taking expectations and summing (Tonelli) over k and $i \in \{-1, 0\}$ gives

$$\sum_{k \geq 0} \sum_{i \in \{-1,0\}} \mathbb{P}\left(|\tilde{f}_k^{(i)} - f_k^{(i)}| \;\geq\; \frac{\Delta_k}{2l} \right) \;\leq\; \varepsilon_h (2l)^{p/(p-1)} \, \mathbb{E}\left[\sum_{k \geq 0} \Delta_k^p \right] \;<\; \infty.$$

By Borel–Cantelli (first lemma), there is an a.s.-finite random index $\bar{k}$ such that for all $k \geq \bar{k}$ (in particular, for all large $k \in \mathcal{U}$),

$$|\tilde{f}_k^{(i)} - f_k^{(i)}| \;\leq\; \frac{\Delta_k}{2l}, \qquad i \in \{-1, 0\}. \tag{13}$$

For $k \in \mathcal{U}$ and $k \geq \bar{k}$, decompose

$$\begin{aligned} f(X_k + \Delta_k D_k) - f(X_k) &= \big(f(X_k + \Delta_k D_k) - \tilde{f}_k^{(0)}\big) + \big(\tilde{f}_k^{(0)} - \tilde{f}_k^{(-1)}\big) + \big(\tilde{f}_k^{(-1)} - f(X_k)\big) \\ &\geq -\frac{\Delta_k}{2l} \;+\; \big(\tilde{f}_k^{(0)} - \tilde{f}_k^{(-1)}\big) \;-\; \frac{\Delta_k}{2l} \qquad \text{by (13)} \\ &\geq -\,\theta\, \Delta_k^p \;-\; \frac{\Delta_k}{l} \qquad \text{by (12)}. \end{aligned}$$

Dividing by $\Delta_k > 0$ yields, for all large $k \in \mathcal{U}$,

$$\frac{f(X_k + \Delta_k D_k) - f(X_k)}{\Delta_k} \;\geq\; -\theta\, \Delta_k^{p-1} \;-\; \frac{1}{l}.$$

Since $\Delta_k \to 0$ a.s., we obtain a.s.

$$\liminf_{k \in \mathcal{U},\, k \to \infty} \frac{f(X_k + \Delta_k D_k) - f(X_k)}{\Delta_k} \;\geq\; -\frac{1}{l}.$$

As $l \in \mathbb{N}$ was arbitrary, letting $l \to \infty$ gives the claim.

We now report the main convergence result for our stochastic direct-search scheme. The result requires the existence of accumulation points for the sequence $\{x_k\}$, which can be obtained assuming that the iterates generated by the algorithm lie in a compact set (see Assumption 3.1). We omit the proof, as it follows verbatim from [23, Theorem 3.3].

Theorem 4.9 *Assume that f is Lipschitz continuous in a neighborhood of every limit point of the sequence of iterates $\{X_k\}$. Let $\mathcal{U}$ denote the random set of indices of unsuccessful iterations. Suppose that Assumptions 3.1, 3.2, 3.3, and conditions* (6) *and* (7) *hold. Then, with probability one, the following implication holds: if $V \subseteq \mathcal{U}$ is a random index set such that $\{D_k\}_{k \in V}$ is dense in $\mathbb{S}^{n-1}$ and $\lim_{\substack{k \to \infty \\ k \in V}} X_k = X^*$, then X^* is Clarke stationary, namely, $f^\circ(X^*; d) \geq 0$ for every $d \in \mathbb{R}^n$.*

5 Expected Complexity with Goldstein-stationarity stopping

To establish the expected iteration complexity of the proposed algorithm, we adapt the supermartingale stopping-time framework introduced in [3]. This framework allows us to bound the expected number of iterations required to reach an approximate stationary condition. In contrast to the almost-sure convergence analysis of the previous section, we now characterize stationarity via the Goldstein r-subdifferential, $\partial_r f$, rather than the Clarke subdifferential, $\partial_C f$. The reason for this shift is detailed in Remark 5.7. We start by recalling the definition of stopping times.

Definition 5.1 (Stopping time) Let $(\mathcal{A}_k)_{k \geq 0}$ be a filtration on $(\Omega, \mathcal{A}, \mathbb{P})$. A random variable $T : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ is called a stopping time with respect to $(\mathcal{A}_k)_{k \geq 0}$ if

$$\{T \leq k\} \in \mathcal{A}_k \qquad \forall k \in \mathbb{N}_0.$$

Equivalently, T is a stopping time if

$$\{T = k\} \in \mathcal{A}_k \qquad \forall k \in \mathbb{N}_0.$$

For the complexity analysis, we use the pre-direction algorithmic-history filtration

$$\mathcal{A}_k := \mathcal{G}_{k-1},$$

where $\mathcal{G}_{k-1}$ is the information available at the beginning of iteration k, before sampling the directions and before drawing the oracle estimates of iteration k. Thus X_k, Δ_k, and Φ_k are $\mathcal{A}_k$-measurable. Moreover, with the notation of Section 2,

$$\mathcal{A}_k = \mathcal{G}_{k-1} \subseteq \mathcal{F}_{k-1} = \mathcal{G}_{k-1} \vee \sigma(D_{k,1}, \dots, D_{k,\bar{m}}) \subseteq \mathcal{A}_{k+1} = \mathcal{G}_k.$$

The filtration $(\mathcal{F}_{k-1})$ is used for oracle-conditioning arguments after the directions have been sampled, whereas $(\mathcal{A}_k)$ is used for the stopping-time and radius-dynamics arguments. Since f is possibly nonsmooth, the stopping time is defined in terms of the r-Goldstein subdifferential.

Definition 5.2 (Goldstein subdifferential) For $r > 0$, let $B_r(x) := \{y \in \mathbb{R}^n : \|y - x\| \le r\}$ denote the closed ball of radius r centered at x. The *r-Goldstein subdifferential* of f at x is defined as the convex hull of the Clarke subdifferentials at all points $y \in B_r(x)$, namely

$$\partial_r f(x) := \operatorname{conv}\left(\bigcup_{y \in B_r(x)} \partial_C f(y) \right).$$

For fixed tolerances $r > 0$ and $\varepsilon > 0$, we define the stopping time as

$$T_{r,\varepsilon} = \inf\left\{ k \in \mathbb{N}_0 : \min_{g \in \partial_r f(X_k)} \|g\| \le \varepsilon \right\}. \tag{14}$$

Equivalently, $T_{r,\varepsilon}$ is the first iteration at which X_k is a (r,ε)-Goldstein stationary point.
Since the definition only depends on (X_k) which are $\mathcal{A}_k$ measurable, it follows that $T_{r,\varepsilon}$ is indeed a $(\mathcal{A}_k)$-stopping time. Adapting the framework of [3], we can provide a bound on $\mathbb{E}[T_{r,\varepsilon}]$. At a given iteration k, define the base-estimate event

$$\mathcal{E}_{k,0} := \left\{ |\tilde{f}(X_k) - f(X_k)| \le \varepsilon_f \Delta_k^p \right\},$$

and, for each $m = 1, \dots, \bar{m}$, the trial-estimate event

$$\mathcal{E}_{k,m} := \left\{ |\tilde{f}(X_k + \Delta_k D_{k,m}) - f(X_k + \Delta_k D_{k,m})| \le \varepsilon_f \Delta_k^p \right\}.$$

Then, fix a stationarity tolerance $\varepsilon > 0$ and a Goldstein neighborhood radius $r > 0$. Let

$$G_{r,k}^* \in \operatorname*{argmin}_{g \in \partial_r f(X_k)} \|g\|,$$

and, on the event $\{T_{r,\varepsilon} > k\}$, define

$$D_k^* := -\frac{G_{r,k}^*}{\|G_{r,k}^*\|}.$$

Indeed, on this event we have $\|G^*_{r,k}\| > \varepsilon$, so D^*_k is well-defined. The direction D^*_k is the Goldstein descent direction at iteration k. Since the algorithm samples random directions, we only need one sampled direction to be sufficiently aligned with D^*_k. To quantify the probability that a randomly sampled direction is sufficiently aligned with a target descent direction, we use the standard formula for the surface measure of a spherical cap.

Definition 5.3 (Spherical cap) Fix $0 < \rho < 1$ and a unit vector $u \in \mathbb{S}^{n-1}$. The *spherical cap with center u and alignment parameter ρ* is

$$\mathcal{C}_\rho(u) := \{d \in \mathbb{S}^{n-1} : \langle u, d\rangle \geq \rho\}.$$

Equivalently, $\mathcal{C}_\rho(u)$ is the cap with half-angle $\arccos(\rho)$.

Thus, $D_{k,m} \in \mathcal{C}_\rho(D^*_k)$ means that the sampled direction $D_{k,m}$ is sufficiently aligned with the Goldstein descent direction at iteration k. Now, define the good event that both estimates are accurate and $\exists\ m \in \{1, \ldots, \bar{m}\} : D_{k,m} \in \mathcal{C}_\rho(D^*_k)$, with D^*_k being the Goldstein descent direction at iteration k, and $C_\rho(D^*_k)$ being the spherical cap therein centered. That is, define

$$\mathcal{I}_k := \mathcal{E}_{k,0} \cap \left[\bigcup_{m=1}^{\bar{m}} \left(\{D_{k,m} \in \mathcal{C}_\rho(D^*_k)\} \cap \mathcal{E}_{k,m}\right)\right],$$

and let $J_k := \mathbf{1}_{\mathcal{I}_k}$. We then set

$$W_{k+1} := 2J_k - 1 \in \{-1, +1\}. \tag{15}$$

Hence $W_{k+1} = +1$ if the good event $\mathcal{I}_k$ occurs, and $W_{k+1} = -1$ otherwise. In particular, we will show that, whenever $T_{r,\varepsilon} > k$ and $\Delta_k \leq \bar{\Delta}_{r,\varepsilon}$, the occurrence of $\mathcal{I}_k$ implies that iteration k is successful. We set $W_0 = 1$ only for initialization. Notice that W_k is $\mathcal{A}_k$-measurable, whereas W_{k+1} is not $\mathcal{A}_k$-measurable. We now state the abstract conditions needed to apply the stopping-time framework (see also [3, Assumption 1]).

Assumption 5.1 *Let $T_{r,\varepsilon}$ be the stopping time defined in* (14)*. Let the stochastic process $\{(\Phi_k, \Delta_k, W_k)\}_{k\geq 0}$ be adapted to $(\mathcal{A}_k)_{k\geq 0}$. The following conditions hold.*

(i) There exist constants $\lambda > 0$ and $\Delta_{\max} = \Delta_0 e^{\lambda j_{max}} > 0$, $j_{max} \in \mathbb{Z}$ such that

$$\lambda = -\log\gamma \qquad \text{and} \qquad \Delta_k \leq \Delta_{\max} \quad \text{for all } k \geq 0.$$

(ii) There exist a threshold

$$\bar{\Delta}_{r,\varepsilon} = \Delta_0 e^{\lambda j_{r,\varepsilon}}, \qquad j_{r,\varepsilon} \in \mathbb{Z}, \quad j_{r,\varepsilon} \leq 0,$$

and a constant $q > 1/2$ such that

$$\mathbb{P}(W_{k+1} = +1 \mid \mathcal{A}_k) = q, \qquad \mathbb{P}(W_{k+1} = -1 \mid \mathcal{A}_k) = 1 - q,$$

and, for every $k \geq 0$,

$$\mathbf{1}_{\{T_{r,\varepsilon} > k\}} \Delta_{k+1} \geq \mathbf{1}_{\{T_{r,\varepsilon} > k\}} \min\left\{\Delta_k e^{\lambda W_{k+1}}, \bar{\Delta}_{r,\varepsilon}\right\}. \tag{16}$$

(iii) There exist a nondecreasing function $h : [0,\infty) \to [0,\infty)$ *and a constant* $\Theta > 0$ *such that, for every* $k \geq 0$,

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{A}_k]\mathbf{1}_{\{T_{r,\varepsilon}>k\}} \geq \Theta h(\Delta_k)\mathbf{1}_{\{T_{r,\varepsilon}>k\}} \qquad a.s. \tag{17}$$

In particular, we take $h(\Delta) = \Delta^p$.

Assumption 5.1 states that, before the stopping time, the nonnegative stochastic process Φ_k has an expected decrease of at least $\Theta h(\Delta_k)$ at each iteration. Moreover, before the stopping time and whenever $\Delta_k \leq \bar{\Delta}_{r,\varepsilon}$, the radius has an upward tendency: the good event $\mathcal{I}_k$, which implies $W_{k+1} = +1$, occurs with conditional probability larger that $1/2$, namely

$$\mathbb{P}(W_{k+1} = +1 \mid \mathcal{A}_k) = q > 1/2.$$

Our goal is to bound $\mathbb{E}[T_{r,\varepsilon}]$ in terms of $h(\bar{\Delta}_{r,\varepsilon})$. What happens is that, on average, $\Delta_k \geq \bar{\Delta}_{r,\varepsilon}$ frequently, and hence, $\mathbb{E}[\Phi_{k+1} - \Phi_k]$ can be bounded by a negative fixed value (dependent on r, ε), sufficiently frequently, which will allow us to apply Wald's identity and derive the bound on $\mathbb{E}[T_{r,\varepsilon}]$ (for more details refer to [3, Theorem 2]). The next result is a direct specialization of the supermartingale stopping-time bound of Blanchet, Cartis, Menickelly and Scheinberg [3, Theorem 2] to our framework.

Theorem 5.4 *Let Assumption 5.1 hold. Then*

$$\mathbb{E}[T_{r,\varepsilon}] \leq \frac{q}{2q-1} \cdot \frac{\Phi_0}{\Theta h(\bar{\Delta}_{r,\varepsilon})} + 1.$$

Assumption 5.1(i) follows from the compactness assumption (see Remark 3.1). For Assumption 5.1(iii), the merit-function drift established in Lemma 4.1 gives

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}] \geq \Theta \Delta_k^p,$$

where $\Theta := \eta(1-\gamma^p) - c_{\max} > 0$. Since $\mathcal{A}_k \subseteq \mathcal{F}_{k-1}$, the tower property gives

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{A}_k] = \mathbb{E}[\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}] \,|\, \mathcal{A}_k] \geq \Theta \Delta_k^p.$$

Multiplying by the $\mathcal{A}_k$-measurable indicator $\mathbf{1}_{\{T_{r,\varepsilon}>k\}}$, we obtain

$$\mathbb{E}[\Phi_k - \Phi_{k+1} \mid \mathcal{A}_k]\mathbf{1}_{\{T_{r,\varepsilon}>k\}} \geq \Theta \Delta_k^p \mathbf{1}_{\{T_{r,\varepsilon}>k\}},$$

which is Assumption 5.1(iii).
It remains to verify Assumption 5.1(ii), namely the radius dynamics and the lower bound

$$\mathbb{P}(W_{k+1} = +1 \mid \mathcal{A}_k) \geq q > 1/2$$

before the stopping time and for sufficiently small Δ_k. The next two lemmas provide exactly this. The first lemma proves a uniform lower bound on the probability that at least one of the $\bar{m}$ sampled directions yields a sufficient-decrease step. The second lemma shows that, with the given stepsize update rule, the radius process satisfies the required comparison inequality. Before stating the lemmas, we report the following definition that collects the notation for the regularized incomplete Beta function and the cap probability under the uniform distribution of directions on the sphere.

Definition 5.5 (Regularized incomplete Beta function and Spherical-cap probability) For $a, b > 0$ and $z \in [0, 1]$, the *regularized incomplete Beta function* is

$$I_z(a, b) := \frac{B(z; a, b)}{B(a, b)}, \qquad B(z; a, b) := \int_0^z t^{a-1}(1-t)^{b-1}\, dt,$$

where

$$B(a, b) := \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Equivalently, if $U \sim \mathrm{Beta}(a, b)$, then

$$I_z(a, b) = \mathbb{P}(U \le z).$$

Now, assume $n \ge 2$. By standard properties of a uniformly distributed point on the sphere (see, e.g., [19, Chapter 9]), the surface measure of the spherical cap $\mathcal{C}_\rho(u)$ is independent of its center u. More precisely, if $D \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, then

$$\mathbb{P}(D \in \mathcal{C}_\rho(u)) = p_\rho := \frac{1}{2} I_{1-\rho^2}\left(\frac{n-1}{2}, \frac{1}{2}\right). \tag{18}$$

Moreover, $p_\rho > 0$ for every fixed $\rho \in (0, 1)$, and $p_\rho \downarrow 0$ as $\rho \uparrow 1$.

If $D_{k,m}$ is uniformly distributed on $\mathbb{S}^{n-1}$, then $\mathbb{P}(D_{k,m} \in \mathcal{C}_\rho(D_k^*) \mid \mathcal{A}_k) = p_\rho$. We first show that, before the stopping time and for sufficiently small stepsizes, the occurrence of the good event $\mathcal{I}_k$ forces the algorithmic sufficient-decrease test to succeed. This is the analogue, in the present Goldstein-stationarity setting, of the key implication used in Dzahini's analysis [7]: if the stationarity measure is still larger than the prescribed tolerance, the stepsize is small enough, and the estimates are accurate, then the iteration must be successful. Here, the role of the descent direction is played by any sampled direction lying in the spherical cap $\mathcal{C}_\rho(D_k^*)$.

Lemma 5.6 (Sufficient decrease on the good event) *Assume*

$$c_1 := \rho\varepsilon - \sqrt{1-\rho^2}\, L > 0.$$

Let $\theta > 0$ be the sufficient-decrease parameter used by the algorithm, and choose

$$\Delta_{r,\varepsilon} < \min\left\{ r, \left(\frac{c_1}{\theta + 2\varepsilon_f}\right)^{1/(p-1)} \right\}.$$

Let

$$B_k := \{T_{r,\varepsilon} > k\} \cap \{\Delta_k \le \Delta_{r,\varepsilon}\}.$$

Then, on the event $B_k \cap \mathcal{I}_k$, there exists $m \in \{1, \dots, \bar{m}\}$ such that

$$D_{k,m} \in \mathcal{C}_\rho(D_k^*)$$

and both the base estimate and the corresponding trial estimate are ε_f-accurate. For such an index m,

$$\tilde{f}(X_k + \Delta_k D_{k,m}) - \tilde{f}(X_k) \le -\theta\Delta_k^p.$$

Consequently, iteration k is successful.

Proof We argue on the event B_k. By definition of $T_{r,\varepsilon}$,

$$\min_{g\in\partial_r f(X_k)} \|g\| > \varepsilon.$$

Choose

$$G^*_{r,k} \in \operatorname*{argmin}_{g\in\partial_r f(X_k)} \|g\|, \qquad D^*_k := -\frac{G^*_{r,k}}{\|G^*_{r,k}\|}.$$

Then $\|G^*_{r,k}\| > \varepsilon$.

Fix $m \in \{1,\dots,\bar{m}\}$ such that $D_{k,m} \in \mathcal{C}_\rho(D^*_k)$, and set $S_{k,m} := \Delta_k D_{k,m}$. By Lebourg's mean-value theorem, there exist

$$\bar{X}_{k,m} = X_k + \Xi_{k,m} S_{k,m}, \qquad \Xi_{k,m} \in (0,1),$$

and

$$\bar{G}_{k,m} \in \partial_C f(\bar{X}_{k,m})$$

such that

$$f(X_k + S_{k,m}) - f(X_k) = \langle \bar{G}_{k,m}, S_{k,m}\rangle = \Delta_k \langle \bar{G}_{k,m}, D_{k,m}\rangle.$$

Since $\Delta_k \le \Delta_{r,\varepsilon} \le r$, we have

$$\|\bar{X}_{k,m} - X_k\| \le \Delta_k \le r.$$

Therefore $\bar{X}_{k,m} \in B_r(X_k)$, and hence

$$\bar{G}_{k,m} \in \partial_r f(X_k).$$

By the projection theorem applied to the closed convex set $\partial_r f(X_k)$,

$$\langle \bar{G}_{k,m}, G^*_{r,k}\rangle \ge \|G^*_{r,k}\|^2.$$

Consequently,

$$\langle \bar{G}_{k,m}, D^*_k\rangle = -\frac{\langle \bar{G}_{k,m}, G^*_{r,k}\rangle}{\|G^*_{r,k}\|} \le -\|G^*_{r,k}\| < -\varepsilon.$$

Write

$$D_{k,m} = R_{k,m} D^*_k + \sqrt{1 - R^2_{k,m}}\, V_{k,m},$$

where

$$R_{k,m} := \langle D^*_k, D_{k,m}\rangle, \qquad V_{k,m} \perp D^*_k, \qquad \|V_{k,m}\| = 1.$$

Since $D_{k,m} \in \mathcal{C}_\rho(D^*_k)$, we have $R_{k,m} \ge \rho$. Moreover, since f is L-Lipschitz on the relevant neighborhood, $\|\bar{G}_{k,m}\| \le L$. Thus

$$\begin{aligned}
\langle \bar{G}_{k,m}, D_{k,m}\rangle &= R_{k,m}\langle \bar{G}_{k,m}, D^*_k\rangle + \sqrt{1 - R^2_{k,m}}\langle \bar{G}_{k,m}, V_{k,m}\rangle \\
&\le -R_{k,m}\varepsilon + \sqrt{1 - R^2_{k,m}}\, L \\
&\le -\rho\varepsilon + \sqrt{1-\rho^2}\, L \\
&= -c_1.
\end{aligned}$$

Hence

$$f(X_k + \Delta_k D_{k,m}) - f(X_k) \leq -c_1 \Delta_k.$$

If the two estimates are ε_f-accurate, then

$$\begin{aligned}\tilde{f}(X_k + \Delta_k D_{k,m}) - \tilde{f}(X_k) &\leq f(X_k + \Delta_k D_{k,m}) - f(X_k) + 2\varepsilon_f \Delta_k^p \\ &\leq -c_1 \Delta_k + 2\varepsilon_f \Delta_k^p.\end{aligned}$$

By the definition of $\Delta_{r,\varepsilon}$,

$$\Delta_k^{p-1} \leq \Delta_{r,\varepsilon}^{p-1} < \frac{c_1}{\theta + 2\varepsilon_f}.$$

Therefore

$$c_1 \Delta_k > (\theta + 2\varepsilon_f)\Delta_k^p.$$

It follows that

$$\tilde{f}(X_k + \Delta_k D_{k,m}) - \tilde{f}(X_k) \leq -\theta \Delta_k^p.$$

Thus the sufficient-decrease test succeeds for direction m, and iteration k is successful.

Remark 5.7 (Overcoming the coarse Clarke bound via Goldstein subdifferentials) In a pure Clarke framework, bounding the Lebourg subgradient relies on the coarse Lipschitz bound $\|\bar{G}_{k,m} - G_k\| \leq 2L$. This requires a descent margin of $\rho\varepsilon - 2L > 0$, forcing $\rho > 2L/\varepsilon$. As $\varepsilon \to 0$, this constraint inevitably demands $\rho \geq 1$. The spherical cap becomes empty, collapsing the success probability p_ρ to zero, and restricting the analysis to low-accuracy targets where $\varepsilon > 2L$. Shifting to the Goldstein r-subdifferential resolves this topological barrier. By targeting an (ε, r)-Goldstein stationary point and restricting $\Delta_k \leq r$, the intermediate Lebourg point is trapped in $B_r(X_k)$, ensuring $\bar{G}_{k,m} \in \partial_r f(X_k)$. The projection theorem then eliminates the $2L$ residual, refining the descent condition to $c_1 := \rho\varepsilon - \sqrt{1-\rho^2}L > 0$. This yields $\rho > \frac{L}{\sqrt{\varepsilon^2 + L^2}}$, which is strictly less than 1 for all $\varepsilon > 0$, guaranteeing a positive success probability as $\varepsilon \to 0$. In the smoother regime $f \in C^{1,1}$, define $G_k := \nabla f(X_k)$, $\bar{G}_{k,m} := \nabla f(\bar{X}_{k,m})$, where $\bar{X}_{k,m}$ is the intermediate point on the segment $[X_k, X_k + \Delta_k D_{k,m}]$. If ∇f is $\bar{L}$-Lipschitz, then

$$\left|\langle \bar{G}_{k,m} - G_k, D_{k,m} \rangle\right| \leq \bar{L}\Delta_k.$$

The $2L$ penalty is replaced by $\mathcal{O}(\Delta_k)$, yielding $c_{1,k} = \rho\varepsilon - \mathcal{O}(\Delta_k)$. Since $\Delta_k \to 0$ almost surely, $c_{1,k} > 0$ eventually holds regardless of ε, as established in [7].

We next quantify the probability of the good event introduced above. More specifically, the following lemma shows that, under conditional independence and accuracy assumptions, this event occurs with probability bounded from below uniformly in k.

Lemma 5.8 (Probability of the good event) *Assume that, conditionally on $\mathcal{A}_k$, the directions $D_{k,1}, \dots, D_{k,\bar{m}}$ are independent and uniformly distributed on $\mathbb{S}^{n-1}$. Assume also that the estimates are β-probabilistically ε_f-accurate, namely*

$$\mathbb{P}(\mathcal{E}_{k,0} \mid \mathcal{A}_k) \geq \beta, \qquad \mathbb{P}(\mathcal{E}_{k,m} \mid \mathcal{A}_k, D_{k,m}) \geq \beta, \quad m = 1, \dots, \bar{m}.$$

Moreover, assume that, conditionally on $\mathcal{A}_k$, the base-estimate event $\mathcal{E}_{k,0}$ is independent of the trial events, and that the pairs

$$\{(D_{k,m}, \mathcal{E}_{k,m})\}_{m=1}^{\bar{m}}$$

are conditionally independent.
Then, we have

$$\mathbb{P}(\mathcal{I}_k \mid \mathcal{A}_k) = \mathbb{P}(W_{k+1} = +1 \mid \mathcal{A}_k) \geq q_{\bar{m}}$$

where

$$q_{\bar{m}} := \beta \left[1 - (1 - p_\rho \beta)^{\bar{m}}\right].$$

Proof Since $D_{k,m}$ is uniformly distributed on $\mathbb{S}^{n-1}$ conditionally on $\mathcal{A}_k$, and since the spherical cap has probability p_ρ, we have

$$\mathbb{P}(D_{k,m} \in \mathcal{C}_\rho(D_k^*) \mid \mathcal{A}_k) = p_\rho.$$

Define $\mathcal{S}_{k,m} := \{D_{k,m} \in \mathcal{C}_\rho(D_k^*)\} \cap \mathcal{E}_{k,m}$. Using the conditional uniformity of $D_{k,m}$ and the conditional β-accuracy of the trial estimate,

$$\begin{aligned}
\mathbb{P}(\mathcal{S}_{k,m} \mid \mathcal{A}_k) &= \mathbb{E}\Big[\mathbf{1}_{\{D_{k,m} \in \mathcal{C}_\rho(D_k^*)\}} \mathbb{P}(\mathcal{E}_{k,m} \mid \mathcal{A}_k, D_{k,m}) \,\Big|\, \mathcal{A}_k\Big] \\
&\geq \beta\, \mathbb{P}(D_{k,m} \in \mathcal{C}_\rho(D_k^*) \mid \mathcal{A}_k) \\
&= p_\rho \beta.
\end{aligned}$$

By conditional independence, the events $\mathcal{S}_{k,1}, \dots, \mathcal{S}_{k,\bar{m}}$ are conditionally independent given $\mathcal{A}_k$. Hence

$$\mathbb{P}\left(\bigcup_{m=1}^{\bar{m}} \mathcal{S}_{k,m} \,\middle|\, \mathcal{A}_k\right) \geq 1 - (1 - p_\rho \beta)^{\bar{m}}.$$

Furthermore, $\mathcal{E}_{k,0}$ is conditionally independent of these trial events and satisfies

$$\mathbb{P}(\mathcal{E}_{k,0} \mid \mathcal{A}_k) \geq \beta.$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\mathcal{I}_k \mid \mathcal{A}_k) &= \mathbb{P}\left(\mathcal{E}_{k,0} \cap \bigcup_{m=1}^{\bar{m}} \mathcal{S}_{k,m} \,\middle|\, \mathcal{A}_k\right) \\
&\geq \beta \left[1 - (1 - p_\rho \beta)^{\bar{m}}\right] \\
&= q_{\bar{m}}.
\end{aligned}$$

Since $W_{k+1} = +1$ if and only if $\mathcal{I}_k$ occurs, we obtain

$$\mathbb{P}(W_{k+1} = +1 \mid \mathcal{A}_k) = \mathbb{P}(\mathcal{I}_k \mid \mathcal{A}_k) \geq q_{\bar{m}},$$

which gives the claim.

Remark 5.9 (Polling capacity $\bar{m}$ and dimensional complexity) To apply the expected complexity bound of Theorem 5.4 with constants bounded uniformly, it is convenient to fix a number $q_0 \in (1/2, \beta)$ and choose $\bar{m}$ so that

$$q_{\bar{m}} := \beta\,[\,1 - (1 - p_\rho\beta)^{\bar{m}}\,] \geq q_0.$$

This is possible only if $\beta > 1/2$. Solving the inequality $q_{\bar{m}} \geq q_0$ gives

$$(1 - p_\rho\beta)^{\bar{m}} \leq 1 - \frac{q_0}{\beta},$$

and hence

$$\bar{m} \geq \frac{\log(1 - q_0/\beta)}{\log(1 - p_\rho\beta)}.$$

Thus, for example, it is sufficient to choose

$$\bar{m} > \bar{m}_{\min}(q_0) := \left\lceil \frac{\log(1 - q_0/\beta)}{\log(1 - p_\rho\beta)} \right\rceil.$$

With this choice, $q_{\bar{m}} \geq q_0 > 1/2$, satisfying the supermartingale condition. However, targeting the Goldstein r-subdifferential requires the descent margin $c_1 > 0$, forcing $\rho > L/\sqrt{\varepsilon^2 + L^2}$. As the stationarity tolerance $\varepsilon \downarrow 0$, the required cap alignment satisfies $\rho \uparrow 1$, and the cap probability p_ρ goes to zero. To quantify the resulting dependence of $\bar{m}$ on ε, choose ρ so that

$$1 - \rho^2 \asymp \left(\frac{\varepsilon}{L}\right)^2.$$

Equivalently, the cap remains nonempty but becomes narrower as $\varepsilon \downarrow 0$. Let $z := 1 - \rho^2$. Using the small-z asymptotics of the regularized incomplete Beta function from Definition 5.5, we have

$$p_\rho = \frac{1}{2} I_z\left(\frac{n-1}{2}, \frac{1}{2}\right) \asymp z^{(n-1)/2}.$$

Therefore

$$p_\rho \asymp \left(\frac{\varepsilon}{L}\right)^{n-1}.$$

Since $\log(1 - p_\rho\beta) \asymp -p_\rho\beta$ as $p_\rho \downarrow 0$, the sufficient polling budget satisfies

$$\bar{m}_{\min}(q_0) = \mathcal{O}\left(\frac{1}{p_\rho}\right) = \mathcal{O}\left(\left(\frac{L}{\varepsilon}\right)^{n-1}\right).$$

Thus, for fixed L, choosing

$$\bar{m} = \mathcal{O}(\varepsilon^{1-n})$$

is sufficient to keep the success probability uniformly bounded away from $1/2$. This dimensional dependence reflects the geometric difficulty of finding a descent direction in a purely nonsmooth landscape via a dense set of random directions. The above argument, together with the use of the spherical cap, is somehow related to the probabilistic-descent framework of Gratton et al. [12,

Appendix B]. There, for smooth objectives, random polling directions are used to obtain, with high probability, a direction lying in a spherical cap around the negative gradient; and a fixed positive cosine threshold (for instance of order $1/\sqrt{n}$) is sufficient to guarantee descent. In our nonsmooth Goldstein setting, the spherical-cap idea is used, but nonsmoothness lets the cap shrink with the stationarity tolerance, leading to the polling budget $\bar{m} = O(\varepsilon^{1-n})$. We finally note that the constant $c_{\max} = 2\mu_h \bar{m}(\bar{\imath}+1)$ depends on $\bar{m}$. We can however guarantee that $c_{\max}$ remains uniformly bounded by tightening of the sample-average accuracy, that is by increasing the batch size in the stochastic estimates construction, if needed.

Recall that an iteration is called *successful* if at least one of the $\bar{m}$ tested directions satisfies the sufficient-decrease test, and *unsuccessful* otherwise. Equivalently, success corresponds to $H_k \geq 0$, whereas failure corresponds to $H_k = -1$. By the stepsize update rule, an unsuccessful iteration contracts the polling radius by γ, while a successful iteration expands the next polling radius by at least γ^{-1}. The good-event indicator used in the stopping-time argument is instead $J_k = \mathbf{1}_{\mathcal{I}_k}$. Hence $J_k = 1$ does not define success in general; rather, on the event $\{T_{r,\varepsilon} > k\} \cap \{\Delta_k \leq \Delta_{r,\varepsilon}\}$, Lemma 5.6 shows that $\mathcal{I}_k$ implies $H_k \geq 0$. The following lemma verifies the radius-dynamics condition in Assumption 5.1.

Lemma 5.10 (Verification of Assumption 5.1(ii)) *Let $\gamma \in (0,1)$ and set $\lambda := -\log\gamma > 0$. Let*

$$\bar{\Delta}_{r,\varepsilon} = \Delta_0 \gamma^{j_{r,\varepsilon}}, \qquad j_{r,\varepsilon} \in \mathbb{Z}, \qquad j_{r,\varepsilon} \leq 0,$$

be such that

$$\bar{\Delta}_{r,\varepsilon} \leq \Delta_{r,\varepsilon},$$

where $\Delta_{r,\varepsilon}$ is defined in Lemma 5.6. Then Assumption 5.1(ii) holds with

$$q = q_{\bar{m}} := \beta\left[1 - (1 - p_\rho \beta)^{\bar{m}}\right].$$

Proof The inequality in Assumption 5.1(ii) is trivial when $\mathbf{1}_{\{T_{r,\varepsilon} > k\}} = 0$. Hence, assume that $T_{r,\varepsilon} > k$. Since the stepsizes lie on the geometric mesh, we can write

$$\Delta_k = \Delta_0 \gamma^{i_k} \qquad \text{for some } i_k \in \mathbb{Z}.$$

If $\Delta_k > \bar{\Delta}_{r,\varepsilon}$, then

$$\Delta_k \geq \gamma^{-1} \bar{\Delta}_{r,\varepsilon}.$$

Moreover, from the stepsize update rule, we always have

$$\Delta_{k+1} \geq \gamma \Delta_k.$$

Thus

$$\Delta_{k+1} \geq \gamma \Delta_k \geq \bar{\Delta}_{r,\varepsilon},$$

and therefore

$$\Delta_{k+1} \geq \min\left\{\Delta_k e^{\lambda W_{k+1}}, \bar{\Delta}_{r,\varepsilon}\right\}.$$

Now assume that

$$\Delta_k \le \bar{\Delta}_{r,\varepsilon} \le \Delta_{r,\varepsilon}.$$

If $W_{k+1} = +1$, then $\mathcal{I}_k$ occurs. By Lemma 5.6, iteration k is successful. Hence

$$\Delta_{k+1} \ge \gamma^{-1}\Delta_k = \Delta_k e^{\lambda} = \Delta_k e^{\lambda W_{k+1}}.$$

If $W_{k+1} = -1$, then, independently of whether the iteration is successful or unsuccessful, the update rule gives

$$\Delta_{k+1} \ge \gamma\Delta_k = \Delta_k e^{-\lambda} = \Delta_k e^{\lambda W_{k+1}}.$$

Therefore, in both cases,

$$\Delta_{k+1} \ge \min\left\{\Delta_k e^{\lambda W_{k+1}}, \bar{\Delta}_{r,\varepsilon}\right\}.$$

Finally, by Lemma 5.8,

$$\mathbb{P}(W_{k+1} = +1 \mid \mathcal{A}_k) = \mathbb{P}(\mathcal{I}_k \mid \mathcal{A}_k) \ge \beta\left[1 - (1 - p_\rho\beta)^{\bar{m}}\right] = q_{\bar{m}}.$$

If $\bar{m}$ is chosen so that $q_{\bar{m}} > 1/2$, then Assumption 5.1(ii) holds with $q = q_{\bar{m}}$.

Remark 5.11 (Choice of $\bar{\Delta}_{r,\varepsilon}$ and mesh index $j_{r,\varepsilon}$) Lemma 5.6 and Lemma 5.8 provide an analytic threshold $\Delta_{r,\varepsilon}$ such that, for $k < T_{r,\varepsilon}$, the spherical-cap argument yields a uniformly positive probability of success whenever $\Delta_k \le \Delta_{r,\varepsilon}$. For the stopping-time theorem, we need a threshold belonging to the geometric mesh generated by the radius updates. Since the updates multiply the radius by powers of γ, the radii belong to the mesh

$$\Delta_0\gamma^j, \qquad j \in \mathbb{Z}.$$

We therefore define

$$j_{r,\varepsilon} := \min\left\{j \in \mathbb{Z} : \Delta_0\gamma^j \le \Delta_{r,\varepsilon}\right\},$$

and set

$$\bar{\Delta}_{r,\varepsilon} := \Delta_0\gamma^{j_{r,\varepsilon}}.$$

Then

$$\bar{\Delta}_{r,\varepsilon} \le \Delta_{r,\varepsilon},$$

so Lemma 5.6 and Lemma 5.8 apply whenever $\Delta_k \le \bar{\Delta}_{r,\varepsilon}$.

We are now ready to combine the merit-function drift, the radius recursion, and the lower bound on the good-event probability within the stopping-time framework. This gives an expected bound on the number of outer iterations needed to reach (r, ε)-Goldstein stationarity. It is important to highlight that assuming c_{max} is uniformly bounded (see end of Remark 5.9) keeps the constants in the outer-iteration bound of Corollary 5.12 uniform in ε.

Corollary 5.12 (Expected stopping time) *Assume the merit-function drift condition given in Lemma 4.1 holds with* $\Theta := \eta(1-\gamma^p) - c_{\max} > 0$, *and let the radius recursion and success-probability bound be as in Lemma 5.6 and Lemma 5.8. Suppose that* $\bar{m}$ *is chosen so that*

$$q_{\bar{m}} := \beta\left[1-(1-p_\rho\beta)^{\bar{m}}\right] > q_0 > \frac{1}{2}.$$

Then we obtain the expected iteration complexity bound

$$\mathbb{E}[T_{r,\varepsilon}] = \mathcal{O}\left(\max\left\{r^{-p}, \varepsilon^{-p/(p-1)}\right\}\right).$$

Proof By Assumption (i) (bounded step sizes) and Lemma 4.1 (merit-function drift with $h(\Delta) = \Delta^p$), items (i) and (iii) of Assumptions 5.1 hold. Lemma 5.10 verifies item (ii) with the Bernoulli driver $W_{k+1} \in \{-1, +1\}$, parameter $q = q_{\bar{m}} > \frac{1}{2}$, and $\lambda = \log(\gamma^{-1})$, together with the cap at $\bar{\Delta}_{r,\varepsilon}$. Substituting these constants into the renewal-reward bound (Theorem 5.4) gives the stated inequality. More specifically, Theorem 5.4 applies with

$$h(\Delta) = \Delta^p, \qquad q = q_{\bar{m}}, \qquad \Delta_{r,\varepsilon} = \bar{\Delta}_{r,\varepsilon},$$

and gives

$$\mathbb{E}[T_{r,\varepsilon}] \le \frac{q_{\bar{m}}}{2q_{\bar{m}}-1}\frac{\Phi_0}{\Theta\bar{\Delta}^p_{r,\varepsilon}} + 1. \tag{19}$$

By construction of the analytic threshold in Lemma 5.6, we have that $\Delta_{r,\varepsilon} \asymp \min\left\{r, \varepsilon^{1/(p-1)}\right\}$. Therefore, it is easy to see that also $\bar{\Delta}_{r,\varepsilon} \asymp \min\left\{r, \varepsilon^{1/(p-1)}\right\}$. It thus follows that

$$\frac{1}{\bar{\Delta}^p_{r,\varepsilon}} = \mathcal{O}\left(\max\left\{r^{-p}, \varepsilon^{-p/(p-1)}\right\}\right).$$

Substituting this estimate into (19) gives

$$\mathbb{E}[T_{r,\varepsilon}] = \mathcal{O}\left(\max\left\{r^{-p}, \varepsilon^{-p/(p-1)}\right\}\right),$$

which proves the claim.

The previous corollary counts outer iterations. We now translate this bound into a tested-point complexity estimate by using the fact that, at each iteration, the algorithm evaluates at most $\bar{m}$ directions and at most $\bar{\imath}+1$ extrapolation levels per direction. Also in this case, we note that tightening the sample-average accuracy to make c_{max} uniformly bounded may increase the raw stochastic sample complexity, but not the tested-point complexity considered in Corollary 5.13.

Corollary 5.13 *[Expected tested-point complexity to Goldstein stationarity] Let $T_{r,\varepsilon}$ be the stopping time defined in* (14)*. Assume the hypotheses of Corollary 5.12. Fix $q_0 \in (1/2, \beta)$, and choose $\bar{m}$ large enough so that $q_{\bar{m}} \geq q_0$. Suppose moreover that $\bar{\imath}$ is fixed and that the line search tests at most $\bar{m}$ directions per outer iteration and at most $\bar{\imath}+1$ extrapolation levels per direction. Assume $\bar{m}$ is chosen according to Remark 5.9. Then*

$$\mathbb{E}\left[\sum_{k=0}^{T_{r,\varepsilon}-1}(1+P_k)\right] = \mathcal{O}\left(\varepsilon^{1-n}\max\left\{r^{-p}, \varepsilon^{-p/(p-1)}\right\}\right),$$

where P_k denotes the number of trial points evaluated during iteration k, excluding the baseline point X_k.

Proof At every outer iteration, the algorithm tests at most $\bar{m}$ directions. For each direction, the line search evaluates at most the levels $i = 0, 1, \ldots, \bar{\imath}$. Therefore the number of trial points evaluated during iteration k satisfies

$$P_k \leq \bar{m}(\bar{\imath}+1) \qquad \text{a.s.}$$

Summing up to the stopping time gives

$$\sum_{k=0}^{T_{r,\varepsilon}-1} P_k \leq \bar{m}(\bar{\imath}+1)T_{r,\varepsilon}.$$

Taking expectations yields

$$\mathbb{E}\left[\sum_{k=0}^{T_{r,\varepsilon}-1} P_k\right] \leq \bar{m}(\bar{\imath}+1)\mathbb{E}[T_{r,\varepsilon}].$$

If the baseline estimate is counted once per outer iteration, the same argument gives

$$\mathbb{E}\left[\sum_{k=0}^{T_{r,\varepsilon}-1}(1+P_k)\right] \leq \big(1+\bar{m}(\bar{\imath}+1)\big)\mathbb{E}[T_{r,\varepsilon}].$$

The rest follows by substituting the bound from Corollary 5.12. Finally, if $\bar{m}$ is chosen minimally to ensure $q_{\bar{m}} > 1/2$, the scaling $\bar{m} = \mathcal{O}(\varepsilon^{1-n})$ follows from Remark 5.9. Thus the result is proved.

Remark 5.14 (Tested points versus raw stochastic oracle samples) Corollary 5.13 counts calls to the estimator $\tilde{f}$, or equivalently the number of tested points. It does not count the raw stochastic oracle samples used to construct each estimate. If $\tilde{f}$ is computed by leveraging a batch size m_k of function evaluations at iteration k, then the raw sample cost is proportional to

$$\sum_{k<T_{r,\varepsilon}} m_k P_k,$$

where P_k is the number of estimator calls at iteration k. Thus Corollary 5.13 should be interpreted as a tested-point complexity bound, not as a raw stochastic-oracle sample complexity bound.

Remark 5.15 (Comparison with smooth stochastic direct search) To put the expected total tested-point complexity of Corollary 5.13 into perspective, consider the symmetric target regime where we seek an $(\varepsilon, \varepsilon)$-Goldstein stationary point (i.e., setting $r = \varepsilon$). The complexity bound algebraically factors as:

$$\mathcal{O}\Big(\varepsilon^{1-n} \max\big\{\varepsilon^{-p},\ \varepsilon^{-\frac{p}{p-1}}\big\}\Big) = \mathcal{O}\Big(\varepsilon^{1-n} \cdot \varepsilon^{-\frac{p}{\min(p-1,1)}}\Big).$$

This factorization isolates the theoretical cost of our direct nonsmooth approach. The second term, $\mathcal{O}(\varepsilon^{-p/\min(p-1,1)})$, is exactly the expected iteration-complexity bound established by [7] for stochastic directional direct search on smooth ($C^{1,1}$) objective functions. In that setting, passing from iterations to tested points only multiplies the bound by the cardinality of the positive spanning set, which is uniformly bounded (for instance, of order n for standard positive bases). By contrast, in our nonsmooth random-polling analysis, such a polling-set factor is replaced by $\mathcal{O}(\varepsilon^{1-n})$. This term represents the geometric cost of sampling enough uniform random directions to hit a shrinking spherical cap aligned with a Goldstein descent direction, without the deterministic alignment guarantee provided by the cosine measure of positive spanning sets in smooth settings.

It is also useful to make a comparison with randomized-smoothing zeroth-order methods, such as those of Nesterov and Spokoiny [21] and Kornowski and Shamir [15]. Those methods replace the original objective f by a smoothed surrogate f_μ and use random finite-difference or directional-derivative estimators to approximate gradients of the smoothed function. This approach can lead to sharper worst-case bounds in regimes where the smoothed problem is an appropriate proxy for the original one. However, the resulting performance depends on the choice of the smoothing radius μ and on the stochastic oracle structure. A larger value of μ gives a smoother and more stable surrogate but may introduce a larger bias between f_μ and f; a smaller value of μ reduces this bias but can make gradient estimators more sensitive to stochastic noise. Thus, as μ is reduced to improve the approximation of f by f_μ, one may need additional sampling or variance-control mechanisms to stabilize the estimate. By contrast, the DSE method analyzed here does not optimize a smoothed surrogate and does not attempt to reconstruct a gradient. It evaluates candidate points and accepts steps through a sufficient-decrease test on stochastic function estimates directly. The price paid for this direct nonsmooth treatment is the worse cap-probability factor in the expected complexity bound. This factor should however be understood as a worst-case geometric penalty. Moreover, for locally Lipschitz functions, points of nondifferentiability are negligible by Rademacher's theorem (see, e.g., [5, Section 2.5]). Hence random trial points typically lie in smooth pieces of the objective, where the set of successful directions may be larger than the single certified cap used in the proof. Thus the true chance of success can be substantially larger than what the conservative lower bound used in the analysis says. The advantage in our framework is that no smoothing radius has to be tuned. This helps explain why, despite

the more conservative worst-case complexity, the method is competitive with random-gradient smoothing techniques in our empirical benchmarks.

6 Numerical Results

We compare our direct search with extrapolation (DSE) against four baselines: the stochastic direct search (SDS) of [23], StoMADS [1], the zeroth-order gradient-smoothing method (GS) of Nesterov and Spokoiny [21], and the Optimal Stochastic Nonsmooth Nonconvex Optimization Algorithm (OSNNOA) of Kornowski and Shamir [15]. Unless specified otherwise, we choose the sufficient-decrease exponent as $p = 2$ for all methods, and we use the same budget and profiling protocol as in [23]. The implementation of the proposed method and the code used to reproduce the numerical experiments are available on Github: `https://github.com/APalmier99/DSE_StDFOpt.git`.

6.1 DSE Implementation and Solvers Compared

The DSE implementation used in the experiments is a practical variant of the theoretical scheme analyzed in previous sections. As in the SDS implementation from [23, Section 5], we adopt a multi-phase strategy that improves robustness and speed in practice combining line searches along coordinates with line searches along dense directions. SDS follows the implementation choices reported in [23, Section 5]. For StoMADS we use the authors' reference settings. For GS we set parameters as discussed in [21]. For OSNNOA, we implemented the randomized-smoothing scheme described in [15], using the proposed stochastic gradient estimator and parameter choices as guidance. Amount of noise, oracle budget, starting point, initial stepsizes are fixed across all algorithms for fairness following again the guidelines in [23, Section 5].

6.2 Data and Performance Profiles

Following [20], a run on problem p is declared successful at iteration k if

$$f(x_k) \;\le\; f_L \;+\; \tau\big(f(x_0) - f_L\big), \tag{20}$$

where f_L is the best objective value achieved by any solver on p. We use the standard *performance* and *data* profiles, with time counters $t_{p,s}$ defined as the number of function evaluations used by solver s to satisfy (20). The performance profile of s is

$$\rho_s(\alpha) \;=\; \frac{1}{|\mathcal{P}|}\,\Big|\,\big\{\, p \in \mathcal{P} :\; t_{p,s} \;\le\; \alpha \cdot \min_{s'} t_{p,s'} \;\big\}\Big|,$$

and the data profile is

$$d_s(\kappa) \;=\; \frac{1}{|\mathcal{P}|}\,\Big|\,\{\, p \in \mathcal{P} :\; t_{p,s} \;\le\; \kappa\,(n_p+1)\;\}\Big|,$$

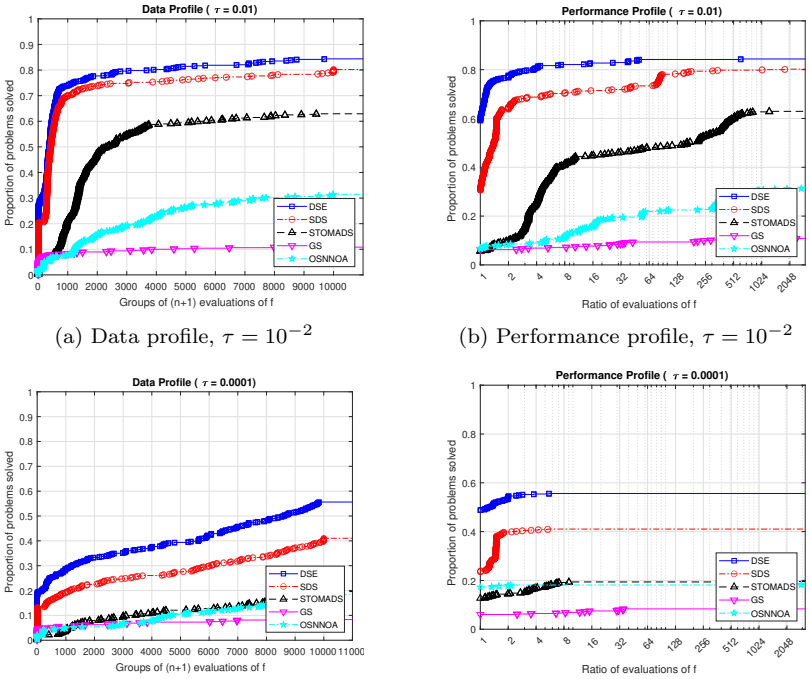
where n_p is the dimension of problem p [23, Section 5]. All profiles are computed with the *true* objective values while the solvers operate on their (possibly stochastic) estimates, and the evaluation budget is fixed as in [23, Section 5]. We enforce a budget of $10^4(n_p+1)$ function evaluations per run. We report results for two tolerances $\tau \in \{10^{-2}, 10^{-4}\}$.
To evaluate performances we use a standard nonsmooth DFO test set (unconstrained), aggregating all runs across problems and random seeds when building the profiles, as in [23, Section 5]. The benchmark set comprises 96 instances of nonsmooth DFO problems (see [23, Table 1]). To stabilize comparisons, we solve each instance *five* times per solver with different seeds, yielding $|P| = 96 \times 5 = 480$ runs per solver. We simulated noise after averaging p_k independent samples by adding to the objective $N(0, 1/p_k)$ distributed random variables. The remaining parameters were tuned with a basic grid search.
In Figure 1, we report data and performance profiles. It is easy to see that, for both tolerances, DSE outperforms the competitors for the whole range of budgets/ratios. More specifically,

- **Data profiles.** The DSE curve (blue) lies strictly above those of SDS (red), StoMADS (black), OSNNOA (china blue) and GS (purple), for essentially all budgets. At $\tau = 10^{-2}$ the gap is pronounced, with DSE solving the largest fraction of problems at small budgets; Even at $\tau = 10^{-4}$, DSE continues to have a distinct advantage.
- **Performance profiles.** DSE attains the highest proportion of "best-within-a-factor" wins across ratios α. The separation from SDS is visible already near $\alpha \approx 1$, and remains throughout the range; For the majority of problems, StoMADS, GS, and OSNNOA perform worse.

Overall, the profiles indicate that DSE is highly effective in practice and the multi-phase strategy helps to improve performances: the initial coordinate searches efficiently capture simple descent directions, while the dense directions give strong thorough exploration once stepsizes are small.

7 Conclusions and Future Directions

We proposed and analyzed DSE, a stochastic direct-search method with extrapolation for unconstrained nonsmooth zeroth-order optimization. The method combines random polling directions with a stochastic sufficient-decrease line search and uses only noisy function-value estimates. Under conditional tail and independence assumptions on the oracle errors, we proved almost-sure convergence of refining subsequences to Clarke stationary points.
We further established expected complexity bounds for reaching (r, ε)-Goldstein



(a) Data profile, $\tau = 10^{-2}$

(b) Performance profile, $\tau = 10^{-2}$

(c) Data profile, $\tau = 10^{-4}$

(d) Performance profile, $\tau = 10^{-4}$

Fig. 1: Data (left) and performance (right) profiles. Top row: $\tau = 10^{-2}$. Bottom row: $\tau = 10^{-4}$.

stationarity. The proof combines a merit-function drift argument with a supermartingale stopping-time framework and a spherical-cap probability estimate for the random polling directions. At the outer-iteration level, we obtain an expected iteration complexity $\mathcal{O}\left(\max\left\{r^{-p}, \varepsilon^{-p/(p-1)}\right\}\right)$. Moreover, after choosing $\bar{m}$, the maximum number of polling directions to be used at each iteration, large enough to keep the success probability uniformly bounded away from $1/2$, the expected tested-point complexity satisfies $\mathcal{O}\left(\varepsilon^{1-n}\max\left\{r^{-p}, \varepsilon^{-p/(p-1)}\right\}\right)$. The numerical results show that extrapolation can substantially improve the practical performance of stochastic direct-search schemes on nonsmooth benchmark problems. In particular, the proposed implementation compares favorably with SDS, StoMADS, and randomized-smoothing baselines on the tested instances.

Several directions remain open. First, it would be interesting to sharpen the dimension dependence in the nonsmooth complexity bound or to identify structural conditions under which the conservative spherical-cap factor can be improved. Second, extending the analysis to variance-reduction mechanisms, common-random-number estimators, and constrained nonsmooth stochastic problems would further broaden the applicability of the approach.

A Proofs of the Theoretical Results

Proof (Lemma 3.2) Work conditionally on $\mathcal{F}_{k-1}$ throughout. Introduce the normalized error

$$Z_{k,i} \;:=\; \frac{|\bar{E}_{k,i}|}{\Delta_k^p} \;\geq 0.$$

Assumption 3.3 is exactly the conditional tail bound

$$\mathbb{P}\big(Z_{k,i} \;\geq\; \alpha \;\big|\; \mathcal{F}_{k-1}\big) \;\leq\; \frac{\varepsilon_h}{\alpha^{p/(p-1)}} \qquad \forall\, \alpha \geq \varepsilon_h.$$

By the layer-cake representation,

$$\mathbb{E}\big[Z_{k,i} \;\big|\; \mathcal{F}_{k-1}\big] = \int_0^\infty \mathbb{P}\big(Z_{k,i} \geq t \;\big|\; \mathcal{F}_{k-1}\big)\; dt.$$

Fix any cutoff $c \geq \varepsilon_h$ and split the integral:

$$\mathbb{E}[Z_{k,i} \mid \mathcal{F}_{k-1}] \;\leq\; \int_0^c 1\, dt \;+\; \int_c^\infty \frac{\varepsilon_h}{t^{p/(p-1)}}\, dt \;=\; c + \varepsilon_h (p-1) c^{-1/(p-1)}.$$

The right-hand side is minimized (over $c \geq \varepsilon_h$) at $c^\star = \max\{\varepsilon_h^{(p-1)/p},\ \varepsilon_h\}$ so

$$\min_{c \geq \varepsilon_h} \Big(c + \varepsilon_h (p-1) c^{-1/(p-1)}\Big) = \begin{cases} p\varepsilon_h^{(p-1)/p}, & \varepsilon_h \leq 1, \quad \textit{case A} \\ \varepsilon_h \left[1 + (p-1)\varepsilon_h^{-1/(p-1)}\right], & \varepsilon_h > 1 \quad \textit{case B.} \end{cases}$$

Therefore $\mathbb{E}[Z_{k,i} \mid \mathcal{F}_{k-1}] \leq \mu_h$ with μ_h defined as above. Multiplying back by Δ_k^p yields $\mathbb{E}[|\bar{E}_{k,i}| \mid \mathcal{F}_{k-1}] \leq \mu_h\, \Delta_k^p$, uniformly in k and i, and the result is proved.

Proof (Lemma 3.3) Work conditionally on $\mathcal{F}_{k-1}$ throughout. By the previous lemma, we have

$$\mathbb{E}\big[\,|\bar{E}_k^{(-1)}|\,\big|\,\mathcal{F}_{k-1}\big] \;\leq\; \mu_h\, \Delta_k^p, \qquad \mathbb{E}\big[\,|\bar{E}_{k,m}^{(i)}|\,\big|\,\mathcal{F}_{k-1}\big] \;\leq\; \mu_h\, \Delta_k^p \quad \forall i,\ \forall m.$$

Using $E_{k,m}^{(i)} = \bar{E}_k^{(-1)} - \bar{E}_{k,m}^{(i)}$ and the triangle inequality,

$$\mathbb{E}\big[\,|E_{k,m}^{(i)}|\,\big|\,\mathcal{F}_{k-1}\big] \;\leq\; \mathbb{E}\big[\,|\bar{E}_k^{(-1)}|\,\big|\,\mathcal{F}_{k-1}\big] + \mathbb{E}\big[\,|\bar{E}_{k,m}^{(i)}|\,\big|\,\mathcal{F}_{k-1}\big] \;\leq\; 2\,\mu_h\, \Delta_k^p,$$

which is (3). The signed bound (4) follows since $|\mathbb{E}[E_{k,m}^{(i)} \mid \mathcal{F}_{k-1}]| \leq \mathbb{E}[|E_{k,m}^{(i)}| \mid \mathcal{F}_{k-1}]$ by Jensen inequality.

Finally, bounding $E_k^{\max} \leq \sum_m \sum_i |E_{k,m}^{(i)}|$ and applying the previous result, immediately yields Lemma 3.4 as well.

B Random Estimates Construction

Standard sampling strategies in stochastic derivative-free optimization often rely on averaging multiple independent realizations of the stochastic oracle. Under the classical assumptions commonly imposed on such sample-average estimators, one can establish the stronger oracle conditions typically assumed in the stochastic derivative-free optimization literature, which in turn imply Assumption 3.3. The purpose of this appendix is to make this connection explicit and derive the corresponding batch-size requirements.

Proposition B.1 (Rosenthal batch size for Assumption 3.3) *Let $p \in (1, 2]$ and set*

$$r := \frac{p}{p-1} \geq 2.$$

Let $\mathcal{H}$ be a sigma-algebra, let X be an $\mathcal{H}$-measurable random point, and let $\Delta > 0$ be an $\mathcal{H}$-measurable random variable. Suppose that, conditionally on $\mathcal{H}$, the oracle samples

$$F(X, \zeta_1), \dots, F(X, \zeta_W)$$

are independent and satisfy, for some $\sigma > 0$,

$$\mathbb{E}[F(X, \zeta_j) - f(X) \,|\, \mathcal{H}] = 0, \qquad \mathbb{E}[|F(X, \zeta_j) - f(X)|^r \,|\, \mathcal{H}] \leq \sigma^r,$$

for every $j = 1, \dots, W$, with W a positive, integer-valued, $\mathcal{H}$-measurable random variable. Define the sample-average estimator

$$\tilde{f}(X) := \frac{1}{W} \sum_{j=1}^{W} F(X, \zeta_j).$$

Then there exists a constant $C_r > 0$, depending only on r, such that, if

$$W \geq \left(\frac{C_r \sigma^r}{\varepsilon_h} \right)^{2/r} \Delta^{-2p},$$

then

$$\mathbb{P}\Big(|\tilde{f}(X) - f(X)| \geq \alpha \Delta^p \,\Big|\, \mathcal{H} \Big) \leq \frac{\varepsilon_h}{\alpha^{p/(p-1)}} \qquad \text{for every } \alpha > 0.$$

Consequently, a batch size $W = \mathcal{O}(\Delta^{-2p})$ is sufficient to satisfy Assumption 3.3.

Proof Define

$$\xi_j := F(X, \zeta_j) - f(X), \qquad j = 1, \dots, W.$$

Then

$$\tilde{f}(X) - f(X) = \frac{1}{W} \sum_{j=1}^{W} \xi_j.$$

Conditionally on $\mathcal{H}$, the random variables $\xi_1, \dots, \xi_W$ are independent and centered, and satisfy

$$\mathbb{E}[|\xi_j|^r \,|\, \mathcal{H}] \leq \sigma^r.$$

Since $r \geq 2$, the conditional version of Rosenthal's inequality gives

$$\mathbb{E}\left[\left| \sum_{j=1}^{W} \xi_j \right|^r \,\middle|\, \mathcal{H} \right] \leq C_r \sigma^r W^{r/2},$$

where $C_r > 0$ depends only on r. Therefore,

$$\mathbb{E}\Big[|\tilde{f}(X) - f(X)|^r \,\Big|\, \mathcal{H} \Big] \leq C_r \sigma^r W^{-r/2}.$$

By conditional Markov's inequality,

$$\mathbb{P}\Big(|\tilde{f}(X) - f(X)| \geq \alpha \Delta^p \,\Big|\, \mathcal{H} \Big) \leq \frac{C_r \sigma^r W^{-r/2}}{\alpha^r \Delta^{pr}}.$$

If

$$W \geq \left(\frac{C_r \sigma^r}{\varepsilon_h} \right)^{2/r} \Delta^{-2p},$$

then

$$C_r \sigma^r W^{-r/2} \Delta^{-pr} \leq \varepsilon_h.$$

Hence

$$\mathbb{P}\Big(|\tilde{f}(X) - f(X)| \geq \alpha \Delta^p \,\Big|\, \mathcal{H} \Big) \leq \frac{\varepsilon_h}{\alpha^r}.$$

Since $r = p/(p-1)$, the claim follows.

Remark B.2 (Application to the algorithm) At iteration k, Proposition B.1 is applied with

$$\mathcal{H} = \mathcal{F}_{k-1}, \qquad \Delta = \Delta_k.$$

For the base-point estimate, take $X = X_k$, while for the trial-point estimate associated with direction m and extrapolation level i, take $X = X_{k,m}^{(i)}$. Thus, define

$$\tilde{f}_k^{(-1)} := \frac{1}{W_k^{(-1)}} \sum_{j=1}^{W_k^{(-1)}} F\big(X_k, \zeta_{k,j}^{(-1)}\big)$$

and

$$\tilde{f}_{k,m}^{(i)} := \frac{1}{W_{k,m}^{(i)}} \sum_{j=1}^{W_{k,m}^{(i)}} F\big(X_{k,m}^{(i)}, \zeta_{k,m,j}^{(i)}\big).$$

Here,

$$\Big\{\zeta_{k,j}^{(-1)}\Big\}_j \quad \text{and} \quad \Big\{\zeta_{k,m,j}^{(i)}\Big\}_{m,i,j}$$

denote conditionally independent copies of the generic oracle random variable ζ, given $\mathcal{F}_{k-1}$. These estimates satisfy Assumption 3.3 whenever

$$W_k^{(-1)} = \mathcal{O}(\Delta_k^{-2p}), \qquad W_{k,m}^{(i)} = \mathcal{O}(\Delta_k^{-2p}),$$

uniformly over $m \in [1 : \bar{m}]$ and $i \in [0 : \bar{\imath}]$.
Moreover, using fresh conditionally independent batches for the base point and for all potential trial points ensures that Assumption 3.2 is also satisfied.

Remark B.3 For $p = 2$, one has $r = 2$, and the result reduces to the usual finite-variance case with $W_k = \mathcal{O}(\Delta_k^{-4})$. For $p \in (1, 2)$, one has $r > 2$, so a moment assumption stronger than finite variance is needed in order to obtain the tail exponent required by Assumption 3.3.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Audet, C., Dzahini, K.J., Kokkolaras, M., Le Digabel, S.: Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. Comput. Optim. Appl. **79**(1), 1–34 (2021). DOI 10.1007/s10589-020-00249-0. URL `https://doi.org/10.1007/s10589-020-00249-0`
2. Bandeira, A.S., Scheinberg, K., Vicente, L.N.: Convergence of trust-region methods based on probabilistic models. SIAM J. Optim. **24**(3), 1238–1264 (2014). DOI 10.1137/130915984. URL `https://doi.org/10.1137/130915984`
3. Blanchet, J., Cartis, C., Menickelly, M., Scheinberg, K.: Convergence rate analysis of a stochastic trust-region method via supermartingales. INFORMS J. Optim. **1**(2), 92–119 (2019). DOI 10.1287/ijoo.2019.0016. URL `https://doi.org/10.1287/ijoo.2019.0016`
4. Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models. Math. Program. **169**(2), 447–487 (2018). DOI 10.1007/s10107-017-1141-8. URL `https://doi.org/10.1007/s10107-017-1141-8`
5. Clarke, F.H.: Optimization and Nonsmooth Analysis. Society for Industrial and Applied Mathematics (1990)

6. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to derivative-free optimization, *MPS/SIAM Series on Optimization*, vol. 8. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA (2009). DOI 10.1137/1.9780898718768. URL https://doi.org/10.1137/1.9780898718768
7. Dzahini, K.J.: Expected complexity analysis of stochastic direct-search. Comput. Optim. Appl. **81**(1), 179–200 (2022). DOI 10.1007/s10589-021-00329-9. URL https://doi.org/10.1007/s10589-021-00329-9
8. Dzahini, K.J., Kokkolaras, M., Le Digabel, S.: Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. Math. Program. **198**(1), 675–732 (2023). DOI 10.1007/s10107-022-01787-7. URL https://doi.org/10.1007/s10107-022-01787-7
9. Dzahini, K.J., Rinaldi, F., Royer, C.W., Zeffiro, D.: Direct-search methods in the year 2025: theoretical guarantees and algorithmic paradigms. EURO J. Comput. Optim. **13**, Paper No. 100110, 24 (2025). DOI 10.1016/j.ejco.2025.100110. URL https://doi.org/10.1016/j.ejco.2025.100110
10. Dzahini, K.J., Wild, S.M.: Stochastic trust-region algorithm in random subspaces with convergence and expected complexity analyses. SIAM J. Optim. **34**(3), 2671–2699 (2024). DOI 10.1137/22M1524072. URL https://doi.org/10.1137/22M1524072
11. Goldstein, A.A.: Optimization of lipschitz continuous functions. Mathematical Programming **13**(1), 14–22 (1977)
12. Gratton, S., Royer, C.W., Vicente, L.N., Zhang, Z.: Direct search based on probabilistic descent. SIAM Journal on Optimization **25**(3), 1515–1541 (2015). DOI 10.1137/140961602
13. Ha, Y., Shashaani, S., Pasupathy, R.: Complexity of zeroth- and first-order stochastic trust-region algorithms. SIAM J. Optim. **35**(3), 2098–2127 (2025). DOI 10.1137/24M1664484. URL https://doi.org/10.1137/24M1664484
14. Jahn, J.: Introduction to the Theory of Nonlinear Optimization. Springer Berlin Heidelberg (1996)
15. Kornowski, G., Shamir, O.: An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. Journal of Machine Learning Research **25**(122), 1–14 (2024). URL http://jmlr.org/papers/v25/23-1159.html
16. Larson, J., Billups, S.C.: Stochastic derivative-free optimization using a trust region framework. Comput. Optim. Appl. **64**(3), 619–645 (2016). DOI 10.1007/s10589-016-9827-z. URL https://doi.org/10.1007/s10589-016-9827-z
17. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numer. **28**, 287–404 (2019). DOI 10.1017/s0962492919000060. URL https://doi.org/10.1017/s0962492919000060
18. Lin, T., Zheng, Z., Jordan, M.I.: Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In: Advances in Neural Information Processing Systems (2022)
19. Mardia, K.V., Jupp, P.E.: Directional Statistics. John Wiley & Sons, Chichester (2000)
20. Moré, J.J., Wild, S.M.: Benchmarking derivative-free optimization algorithms. SIAM Journal on Optimization **20**(1), 172–191 (2009)
21. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017). DOI 10.1007/s10208-015-9296-2. URL https://doi.org/10.1007/s10208-015-9296-2
22. Paquette, C., Scheinberg, K.: A stochastic line search method with expected complexity analysis. SIAM J. Optim. **30**(1), 349–376 (2020). DOI 10.1137/18M1216250. URL https://doi.org/10.1137/18M1216250
23. Rinaldi, F., Vicente, L.N., Zeffiro, D.: Stochastic trust-region and direct-search methods: a weak tail bound condition and reduced sample sizing. SIAM J. Optim. **34**(2), 2067–2092 (2024). DOI 10.1137/22M1543446. URL https://doi.org/10.1137/22M1543446
24. Santis, A.D., Liuzzi, G., Lucidi, S.: A linesearch-based derivative-free method for noisy black-box problems (2025). URL https://arxiv.org/abs/2508.00495
25. Shashaani, S., Hashemi, F.S., Pasupathy, R.: ASTRO-DF: a class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. SIAM J. Optim. **28**(4), 3145–3176 (2018). DOI 10.1137/15M1042425. URL https://doi.org/10.1137/15M1042425