INDUSTRIAL AND SYSTEMS ENGINEERING

ISE

Stochastic trust-region and direct-search methods: A weak tail bound condition and reduced sample sizing

F. Rinaldi[1], L. N. Vicente[2], and D. Zeffiro[1]

[1]Università di Padova

[2]Lehigh University

ISE Technical Report 22T-002

LEHIGH UNIVERSITY

Stochastic trust-region and direct-search methods: A weak tail bound condition and reduced sample sizing

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

June 15, 2023

Abstract

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

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

1 Introduction

We consider the following unconstrained optimization problem

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

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

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

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

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

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

with the black-box oracle given by sampling on the ξ space. When dealing with statistical learning problems [21], the function $F(x, \xi)$ evaluates the loss of the decision rule parametrized by x on a data point ξ. In simulation-based engineering applications [1], the function $F(x, \xi)$ is simply related to some noisy computable version of the original function. In this case, ξ represents the random variable that induces the noise, with a classic example given by Monte Carlo simulations. When this random variable is exact in expected value, problem (1.1) turns out to be the expected loss formulation

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

a case addressed in recent literature, see, e.g., [22, 36] for further details.

1.1 A short review of stochastic derivative-free optimization

Although the role of derivative-free optimization is particularly important when the black box representing the function is somehow noisy or, in general, of a stochastic type, traditional DFO methods have been developed primarily for deterministic functions, and only recently adapted to deal with stochastic observations (see, e.g., [9] for a detailed discussion on this matter). We give here a brief overview of the main results available in the literature by first focusing on *model-based* strategies and then moving to *direct-search* approaches. Further details on these two classes of methods can be found, e.g., in [3, 11].

In [22], the authors describe a trust-region algorithm to handle noisy objectives and prove convergence when f is sufficiently smooth (i.e., with Lipschitz continuous gradient) and the noise is drawn independently from a distribution with zero mean and finite variance, that is they aim at solving a smooth version of problem (1.2), when ξ is additive noise. In the same line of research, the authors in [36] developed a class of derivative-free trust-region algorithms, called ASTRO-DF, for unconstrained optimization problems whose objective function has Lipschitz continuous gradient and can only be implicitly expressed via a Monte Carlo oracle. The authors consider again an objective with noise drawn independently from a distribution with zero mean, finite variance and a bound on the $4v$-th moment (with $v \geq 2$), and prove the almost sure convergence of their method when using stochastic polynomial interpolation models. Another relevant reference in this context is given by [9], where the authors analyze a trust-region model-based algorithm for solving unconstrained stochastic optimization problems. They consider random models of a smooth objective function, obtained from stochastic observations of the function or its gradient. Convergence rates for this class of methods are reported in [7]. The frameworks analyzed in [7, 8, 9] extend the trust-region DFO method based on probabilistic models described in [5]. It is important to notice that the randomness

in the models described in [5] comes from the way sample points are chosen, rather than from noise in the function evaluations. All the above-mentioned model-based approaches consider functions with a certain degree of smoothness (e.g., with Lipschitz continuous gradient) and assume that a probabilistically accurate gradient estimate (e.g., some kind of probabilistically fully-linear model) can be generated, while of course such an estimate is not available when dealing with non-smooth functions.

A detailed convergence rate analysis of stochastic direct-search variants is reported in [13] for the smooth case, i.e., for an objective function with Lipschitz continuous gradient. The main theoretical results are obtained by suitably adapting the supermartingale-based framework proposed in [7]. A stochastic mesh adaptive direct search for black-box nonsmooth optimization is proposed in [2]. The authors prove convergence with probability one to a Clarke stationary point [10] of the objective function by assuming that stochastic observations are sufficiently accurate and satisfy a variance condition. The considered analysis adapts to the direct-search gradient-free framework the theoretical analysis given in [32] for a class of stochastic gradient-based methods. It was extended in [14] to the constrained case.

In a different line of work, zeroth-order methods, first analyzed in [31] for stochastic objectives, make use of two point estimates to approximate the gradient of a smoothed version of the objective. In [16] and [31], complexity bounds are given in the stochastic smooth non-convex setting and the stochastic convex non-smooth setting respectively. In [24], such bounds are extended to the stochastic non-smooth non-convex setting, measuring convergence with the (δ, ε)-Goldstein subdifferential. For a survey of zeroth order methods with applications to machine learning problems we refer the reader to [23]. Other approaches recently adapted from the deterministic setting to stochastic derivative-free/zeroth-order optimization include quasi-Newton methods [29], the stochastic cubic regularized Newton [33], and adaptive regularization methods with cubics [34], requiring stochastic estimates of both the objective gradient and also of the objective Hessian in the latter two cases.

1.2 The contributions of this manuscript

A main goal of this manuscript is to introduce a tail bound probabilistic condition leading to a reduced number of samples per iteration when dealing with a stochastic black-box function in general direct-search and trust-region schemes. This probabilistic condition focuses on the *reduction estimate*, that is the estimate of the difference between the function at the current iterate and at a potential next iterate, used in the acceptance test of those derivative-free algorithms. It expresses a bound on the probability that the reduction estimate error is greater than a fraction of a stepsize power characterizing the sufficient decrease needed for trial-point acceptance, and can therefore be easily adapted to different choices of the power defining such a sufficient decrease.

Our condition enables us to define a trade-off between noise, algorithm parameters, and number of samples per iteration needed to achieve global convergence, which in this context should be intended as convergence to stationary points regardless of the starting point chosen [27]. One of our results is that if all the noise moments are finite, like in the case of Gaussian noise, we only need $O(\Delta_k^{-2-\varepsilon})$ samples, where Δ_k is the stepsize at iteration k, as

described in Corollary 2.5. Here, $\varepsilon > 0$ can be made arbitrarily small by selecting the sufficient decrease power arbitrarily close to 1. This result compares to the $O(\Delta_k^{-4})$ number of samples required in previous works on stochastic trust-region methods [7, 9, 36] and stochastic direct-search methods [2, 13, 14], under a finite variance assumption for the noise. In those works, the sufficient decrease power is taken equal to 2, with the exception of [13] where the power is considered greater than 1. This article also shows that the number of samples needed can be lowered to $O(\Delta_k^{-\varepsilon})$ when the sampling errors are suitably correlated and the random number generator is known, and in particular under a Lipschitz continuity assumption used in the analysis of zeroth-order methods, as it is proved in Corollary 2.9.

We introduce two different algorithmic schemes, namely a simple stochastic direct-search strategy and a stochastic version of the basic deterministic trust-region scheme reported in [25]. Both schemes work as follows: they randomly generate a direction (direct search) or a linear term (trust region); then generate the new iterate by either moving along the direction (direct search) or by solving a trust-region subproblem (trust region); finally they use a sufficient decrease acceptance test to decide if the new point can be accepted (successful iteration) or not. In this work, we use stochastic function estimates in the acceptance tests rather than exact values. Our tail bound condition applies to the function reduction estimates of both schemes, and it allows us to deduce global convergence and to take advantage of the improvement in the number of samples per iteration. We point out that this is the first time global convergence is proved for a stochastic derivative-free trust-region algorithm for non-smooth unconstrained optimization problems. We also remark that the convergence analysis of our trust-region scheme is developed under a new bound on the Hessian of the quadratic model which allows us to generate non-unit linear terms, and thus generalizing the deterministic version given in [25].

Lastly, we show that, for suitable choices of the algorithmic parameters, our tail bound condition is implied by the variance conditions considered in [2] and by the probabilistically accurate function estimate assumption used in [2, 9, 32]. It is also interesting to notice that the finite variance oracle usually considered in the literature (see, e.g., [22, 36]) can be replaced by a more general finite moment oracle (see Subsection 2.2 for further details) when constructing estimates satisfying our conditions.

1.3 Outline of the manuscript

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

and additional numerical results to an appendix.

2 A weak tail bound probabilistic condition for function estimation

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

Our algorithms take a step along a certain direction, which can be a direct-search direction or a trust-region step, and in both cases there is a suitable stepsize quantifying the displacement. The algorithms generate a random process, as described in detail for analogous methods, e.g., in [2, Section 2.2] and [9, Section 3]. The random quantity realizations of the process are indicated as follows. The random direction, the stepsize, and the current point are denoted by G_k, Δ_k, and X_k, with realizations g_k, δ_k, and x_k respectively. The random estimates of $f(X_k)$ and $f(X_k + \Delta_k G_k)$ are denoted by F_k and F_k^g, with realizations f_k and f_k^g respectively. In the direct-search case, the acceptance criterion will be defined as

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

for some $\theta > 0$ and $q > 1$, with δ_k replaced by the norm of the step $\|s_k\|$ in the trust-region case. $\mathcal{F}_{k-1}$ is defined as the σ-algebra of events up to the choice of G_k, so that in particular, G_k is always measurable with respect to $\mathcal{F}_{k-1}$, which will be considered in the proof of Theorem 3.1. This σ-algebra will be used to formalize conditioning on the "past history" of the algorithm up to the choice of G_k. More explicitly, $\mathcal{F}_{k-1}$ is defined as the σ-algebra generated by $(F_j, F_j^g)_{j=0}^{k-1}$ and $(G_j)_{j=0}^{k}$. $\mathbb{E}$ is used to denote expectation and conditional expectation, $\hat{v}$ as a shorthand for $v/\|v\|$, with $\hat{v} = 0$ for $v = 0$, a.s. as a shorthand for *almost surely*, and $[1 : p]$ to denote the integers in the interval $[1, p]$. The starting stepsize Δ_0 is assumed to be deterministic, so that in particular $\mathbb{E}[\Delta_0] < +\infty$, implying that the conditional expectations appearing in the rest of the article are well defined.

2.1 The weak tail bound probabilistic condition

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

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

$$\mathbb{P}\left(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \alpha \Delta_k^q \mid \mathcal{F}_{k-1}\right) \leq \frac{\varepsilon_q}{\alpha^{q/(q-1)}} \tag{2.2}$$

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

The above assumption is in particular a power law [38] tail bound with exponent $q/(q-1)+1$. Notice that an error bound is only assumed for the estimate of the difference $f(X_k)-f(X_k+\Delta_k G_k)$ and not for the estimates of $f(X_k)$ and $f(X_k+\Delta_k G_k)$ taken individually; basically, this bounds the probability that the error in that estimate is large, as such an estimation plays a crucial role in the acceptance tests of the algorithms of this work. It will be clear from Sections 3.2 and 4.2 that the knowledge of an upper bound on ε_q is needed in order to ensure convergence in the proposed algorithms.

Remark 2.2. As described in Section 2.2, Assumption 2.1 can be made for any q, if the r-th moment of the evaluation noise is finite, for $r = q/(q-1)$. Furthermore, for $q \in (1, 2]$, the number of samples needed to satisfy Assumption 2.1 is just $O(\Delta_k^{-2q})$ rather than the standard $O(\Delta_k^{-4})$ required under finite variance assumptions [2] with exponent 2 in the sufficient decrease condition (2.1). This improvement is possible thanks to the relation between the tail bound (2.2) and the acceptance criterion (2.1), together with classic results from probability theory on the convergence rate for the law of large numbers. More precisely, this property will be used: for an average A of m i.i.d. samples with finite r-th finite moment, there is a tail bound of the form $\mathbb{P}(A \geq \alpha) \leq K_{m,r}/\alpha^r$ with $K_{m,r} \propto m^{-\frac{r}{2}}$, as a consequence of Rosenthal's inequality [18] and where $\propto$ stands for "proportional to". Details about these inequalities will be discussed in the appendix.

For convergence purposes, a variant of Assumption 2.1 where the real number α is replaced with a $\mathcal{F}_{k-1}$-measurable random variable A will be needed. This is justified by the following lemma.

Lemma 2.3. *Let* A *be a nonnegative* $\mathcal{F}_{k-1}$ *measurable random variable. If* (2.2) *holds, then*

$$\mathbb{P}\left(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \mathrm{A}\Delta_k^q \ |\mathcal{F}_{k-1}\right) \leq \varepsilon(\mathrm{A}) := +\infty \mathbb{1}_{\{0\}} + \frac{\varepsilon_q}{\mathrm{A}^{q/(q-1)}}\mathbb{1}_{(0,+\infty)} \tag{2.3}$$

Proof. Let $Y = |F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))|/\Delta_k^q$, and $r = \frac{q}{q-1}$. We prove that in this case that for every $F \in \mathcal{F}_{k-1}$:

$$\mathbb{E}\left[\mathbb{1}_F \mathbb{1}_{\{Y \geq \mathrm{A}\}}\right] \leq \mathbb{E}\left[\mathbb{1}_F \varepsilon(\mathrm{A})\right] . \tag{2.4}$$

We prove this intermediate result in the case where A is a discrete random variable with a countable set of possible realizations $\{a_i\}_{i\in\mathbb{N}}$, and then extend the result to the general case by approximation. Indeed we have

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}_F \mathbb{1}_{\{Y \geq \mathrm{A}\}}\right] &= \sum_{i\in\mathbb{N}} \mathbb{E}\left[\mathbb{1}_F \mathbb{1}_{\{Y \geq \mathrm{A}\}} \mathbb{1}_{\{\mathrm{A}=a_i\}}\right] = \sum_{i\in\mathbb{N}} \mathbb{E}\left[\mathbb{1}_{F\cap\{\mathrm{A}=a_i\}} \mathbb{1}_{\{Y \geq a_i\}}\right] \\ &\leq \sum_{i\in\mathbb{N}} \mathbb{E}\left[\mathbb{1}_{F\cap\{\mathrm{A}=a_i\}} \varepsilon(a_i)\right] = \sum_{i\in\mathbb{N}} \mathbb{E}\left[\mathbb{1}_F \mathbb{1}_{\{\mathrm{A}=a_i\}} \varepsilon(\mathrm{A})\right] = \mathbb{E}\left[\mathbb{1}_F \varepsilon(\mathrm{A})\right] \end{aligned} \tag{2.5}$$

as desired, where we used that $F \cap \{\mathrm{A} = a_i\}$ is measurable w.r.t. $\mathcal{F}_{k-1}$ together with (2.2) for $\alpha = a_i$ in the inequality. Notice that if $a_i = 0$ then by assumption $\varepsilon(a_i) = +\infty$ so the inequality is trivial.

Let now A be a general positive random variable, and $\{A_i\}_{i\in\mathbb{N}}$ be a decreasing sequence of discrete random variables converging to A (e.g., $A_i = \sum_{j=0}^{+\infty} \mathbb{1}_{A\in[j/2^i,(j+1)/2^i)} \frac{j+1}{2^i}$). Then $\{\frac{\varepsilon_q}{A_i^r}\}$ is non decreasing and converges a.s. to $\varepsilon(\mathrm{A})$, so we have all the assumptions needed to apply Beppo Levi's Lemma and get

$$\lim_{i\to\infty} \mathbb{E}\left[\mathbb{1}_F \frac{\varepsilon_q}{\mathrm{A}_i^r}\right] = \mathbb{E}\left[\mathbb{1}_F \varepsilon(\mathrm{A})\right] . \tag{2.6}$$

Therefore

$$\mathbb{E}\left[\mathbb{1}_F \mathbb{1}_{\{Y\geq \mathrm{A}\}}\right] = \lim_{i\to\infty} \mathbb{E}\left[\mathbb{1}_F \mathbb{1}_{\{Y\geq \mathrm{A}_i\}}\right] \leq \lim_{i\to\infty} \mathbb{E}[\mathbb{1}_F \varepsilon(\mathrm{A}_i)] = \mathbb{E}\left[\mathbb{1}_F \varepsilon(\mathrm{A})\right] , \tag{2.7}$$

where we used the dominated convergence theorem in the first equality, (2.5) in the first inequality, and (2.6) in the second equality. We have thus proved (2.4) in the general case. Now, let $Z = \mathbb{P}(\mathbb{1}_{Y\geq A} \mid \mathcal{F}_{k-1}) = \mathbb{E}\left[\mathbb{1}_{Y\geq A} \mid \mathcal{F}_{k-1}\right]$. We have, for every $F \in \mathcal{F}_{k-1}$,

$$\mathbb{E}\left[Z\mathbb{1}_F\right] = \mathbb{E}\left[\mathbb{1}_{Y\geq A}\mathbb{1}_F\right] \leq \mathbb{E}\left[\varepsilon(\mathrm{A})\mathbb{1}_F\right] , \tag{2.8}$$

where the first equality follows by definition of conditional expectation and the inequality follows by (2.4). Since both Z and $\frac{\varepsilon_q}{\mathrm{A}^r}$ are $\mathcal{F}_{k-1}$ measurable, from (2.8) we get $Z \leq \frac{\varepsilon_q}{\mathrm{A}^r}$ a.s. as desired. □

The proof is technical and can be found in the Appendix.

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

2.2 Sampling improvement under the new condition

We will show that our tail bound condition can be satisfied under a reduced number of function samples.

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

$$f(x) = \mathbb{E}_\xi[F(x,\xi)] , \qquad \mathrm{E}_\xi\left[|F(x,\xi) - f(x)|^r\right] \leq M_r < +\infty . \tag{2.9}$$

Recall that finite r-th moment implies finite r'-th moment for any $r' \in (1, r]$. Thus for $r < 2$ assumption (2.9) is weaker than assuming finite variance, while for $r > 2$ (2.9) is stronger than assuming finite variance. The next result describes the number of samples needed asymptotically to satisfy the tail bound conditions as a function of r.

Theorem 2.4. *Assume that (2.9) holds with $r = \frac{q}{q-1}$. If $q > 2$, then Assumption 2.1 can be satisfied using $O(\Delta_k^{-q^2})$ i.i.d. samples, while if $q \in (1, 2]$, it can be satisfied using $O(\Delta_k^{-2q})$ i.i.d. samples.*

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

Corollary 2.5. *Let $\varepsilon \in (0,2]$. Then, for $q = 1 + \varepsilon/2$, $O(\Delta_k^{-2-\varepsilon})$ samples are sufficient to satisfy Assumption 2.1, under the finite moment assumption (2.9) for $r = \frac{q}{q-1}$.*

In the rest of this section we assume that the objective is given in the form (1.2), and that the CRN (common number generator) framework can be applied, that is different x can be sampled with fixed ξ. Let now $\bar{F}(x,\xi) = F(x,\xi) - f(x)$ be the sampling error. The sampling errors of close points are assumed to be correlated in the following way:

$$\mathbb{E}_\xi \left[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^r \right] \leq D_r \|x-y\|^r \tag{2.10}$$

for some $D_r > 0$. First, we prove that (2.10) is satisfied if $F(\cdot,\xi)$ is Lipschitz continuous, uniformly in ξ. We remark that uniform Lipschitz continuity assumptions analogous to the one made here are standard in the analysis of zeroth-order methods [24, 31]. In the finite sum setting, this assumption is equivalent to the Lipschitz continuity of every summand.

Proposition 2.6. *Assume that $|F(x,\xi) - F(y,\xi)| \leq L_f \|x-y\|$ for every ξ, and for every $x, y \in \mathbb{R}^n$. Then (2.10) holds for every r, with $D_r = 2^r L_f^r$.*

Proof. Notice that from (1.2) and the uniform L_f Lipschitz continuity assumption it follows that f is L_f Lipschitz continuous as well. Hence, we can write

$$\begin{aligned} |\bar{F}(x,\xi) - \bar{F}(y,\xi)| &= |F(x,\xi) - F(y,\xi) + (f(y) - f(x))| \\ &\leq |f(x) - f(y)| + |F(x,\xi) - F(y,\xi)| \leq 2L_f \|x-y\|\,, \end{aligned}$$

and conclude

$$\mathbb{E}_\xi \left[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^r \right] \leq \mathbb{E}_\xi \left[2^r L_f^r \|x-y\|^r \right] = 2^r L_f^r \|x-y\|^r\,,$$

as desired. □

We now present another example where (2.10) is satisfied, with the noise modelled as a Gaussian process, as is common practice in Bayesian optimization (see, e.g., [35]).

Proposition 2.7. *Assume that $\{F(x,\xi)\}$ is a Gaussian process with expectation $f(x)$, exponentiated kernel with amplitude $\sigma > 0$ and lengthscale $l > 0$, so that in particular*

$$\mathrm{Cov}_\xi(F(x,\xi), F(y,\xi)) = \sigma^2 \exp\left(-\frac{\|x-y\|^2}{2l^2} \right) \tag{2.11}$$

for every $x, y \in \mathbb{R}^n$. Then assumption (2.10) is satisfied for every $r \geq 2$ (with D_r depending on r).

We now show how the bound given in Theorem 2.4 improves under (2.10), for $r \geq 2$.

Theorem 2.8. *If the random number generator is known and (2.10) holds with $r = \frac{q}{q-1}$, then Assumption 2.1 can be satisfied for $q \in (1,2]$ using $O(\Delta_k^{2-2q})$ i.i.d. samples.*

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

Corollary 2.9. *If $q = 1 + \frac{\varepsilon}{2}$ with $\varepsilon \in (0,2]$ then $O(\Delta_k^{-\varepsilon})$ samples are sufficient to satisfy Assumption 2.1 under (2.10) for $r = \frac{q}{q-1}$.*

2.3 Comparison with existing conditions

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

$$\begin{aligned} \mathbb{E}\left[|F_k^g - f(X_k + \Delta_k G_k)|^2 \mid \mathcal{F}_{k-1}\right] &\leq k_f^2 \Delta_k^4 \\ \mathbb{E}\left[|F_k - f(X_k)|^2 \mid \mathcal{F}_{k-1}\right] &\leq k_f^2 \Delta_k^4 \,, \end{aligned} \tag{2.12}$$

for a constant $k_f > 0$. The k_f-variance condition in (2.12) is a gradient-free version of [32, Assumption 2.4, (iii)], and more precisely can be obtained from the latter by removing the gradient related terms in the right-hand side. It is important to note here that in [32] as well as in other works on smooth stochastic derivative free optimization (see, e.g., [9, 22, 36] and references therein), a probabilistically accurate gradient estimate is also used, while of course such an estimate is not available in a possibly non-smooth setting.

Proposition 2.10. *Condition (2.12) implies Assumption 2.1 for $\varepsilon_q = 4k_f^2$ and $q = 2$.*

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

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

We now describe the relation between our assumption and the β-probabilistic accuracy assumption

$$\mathbb{P}\left(\{|F_k - f(X_k)| \leq \tau_f \Delta_k^2\} \cap \{|F_k^g - f(X_k + \Delta_k G_k)| \leq \tau_f \Delta_k^2|\} \mid \mathcal{F}_{k-1}\right) \geq \beta \,, \tag{2.13}$$

used in [2, 9, 32] in combination with other assumptions. In particular, conditions (2.12) are used in [2] and [32] (as discussed above), and a probabilistic assumption on the accuracy of random models for the objective is considered in [9].

We show that if (2.13) is satisfied for every β in a certain interval, with τ_f depending on β and an accuracy parameter ε, then also our assumption is satisfied with ε_q dependent on ε.

Proposition 2.12. *Let $\varepsilon > 0$ and $\bar{p} \in (0, 1)$. Assume that (2.13) holds for every $\beta \in [1-\bar{p}, 1)$, with $\tau_f = \tau_f(\beta) < \frac{1}{2}\sqrt{\frac{\varepsilon}{1-\beta}}$. Then Assumption 2.1 holds with $\varepsilon_q = \frac{\varepsilon}{\bar{p}}$ and $q = 2$.*

The proposition above follows from the inclusion

$$\begin{aligned} &\{|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| < \alpha \Delta_k^2\} \\ &\supset \{|F_k - f(X_k)| \leq \tau_f(\beta) \Delta_k^2\} \cap \{|F_k^g - f(X_k + \Delta_k G_k)| \leq \tau_f(\beta) \Delta_k^2\} \,, \end{aligned} \tag{2.14}$$

whenever $\tau_f(\beta) < \frac{\alpha}{2}$. A detailed proof is presented in the appendix.

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

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

3.1 The stochastic direct-search scheme

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

Algorithm 1 `Stochastic direct search`

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

In order for the method to convergence to Clarke stationary points, certain subsequences of $\{g_k\}$ must be dense in the unit sphere as described in Theorem 3.3. As a remark, a dense sequence in the unit sphere can be generated using a suitable quasirandom sequence [17, 25].

3.2 Convergence analysis under the tail bound probabilistic condition

The following theorem, which implies that the stepsize sequence $\{\Delta_k\}$ converges to zero almost surely, is a key result in the convergence analysis. In the proof, Assumption 2.1 makes it

possible to unify the argument for unsuccessful and successful steps.

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

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

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

Theorem 3.1. *Under Assumption 2.1, if Inequality (3.1) holds then* $\sum_{k\in\mathbb{N}_0} \mathbb{E}\left[\Delta_k^q\right] < \infty$.

Proof. Let $\varepsilon_f = \sqrt[r(q)]{\varepsilon_q}$, $\Phi_k = f(X_k) - f^* + \eta\Delta_k^q$, with $\eta = \frac{\theta}{\bar{\tau}_q}$, and $\varepsilon = -\varepsilon_f + \tau_q^-\theta/\bar{\tau}_q > 0$ where the inequality follows by (3.1).

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

$$\mathbb{E}\left[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}\right] \geq \varepsilon\Delta_k^q\,. \tag{3.2}$$

The thesis then follows as in [13, Theorem 3].

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

$$\begin{aligned}
&\mathbb{E}\left[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}\right] = \mathbb{E}\left[(\Phi_k - \Phi_{k+1})(\mathbb{1}_{J_k} \quad + (1-\mathbb{1}_{J_k}))|\mathcal{F}_{k-1}\right] \\
&= (f(X_k) - f(X_{k+1}) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&\quad + (f(X_k) - f(X_{k+1}) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}\left[\ 1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&= (f(X_k) - f(X_k + \Delta_k G_k) + \eta(\Delta_k^q - \Delta_{k+1}^q))\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&\quad + \eta(\Delta_k^q - \Delta_{k+1}^q)\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&\geq (((\theta - Z_k) - \eta\tau_q^+)\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right])\Delta_k^q,
\end{aligned} \tag{3.3}$$

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

$$\begin{aligned}
&(((\theta - Z_k) - \eta\tau_q^+)\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right])\Delta_k^q \\
&= ((\theta - Z_k - \eta\bar{\tau}_q)\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-)\Delta_k^q \\
&= -Z_k\Delta_k^q\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\Delta_k^q\,,
\end{aligned} \tag{3.4}$$

where we used $\mathbb{E}[1-\mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = 1-\mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}]$ in the first equality, and $\theta = \eta\bar{\tau}_q$ in the second one. By combining (3.3) and (3.4) we can therefore conclude

$$\mathbb{E}\left[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}\right] \geq -Z_k\Delta_k^q\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\Delta_k^q\,. \tag{3.5}$$

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

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

In particular $J_k \subset \{|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq Z_k \Delta_k^q\}$ and we can write, for $Z_k^+ = Z_k \mathbb{1}_{Z_k > 0}$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] = \mathbb{E}\left[\mathbb{1}_{J_k}\mathbb{1}_{\{Z_k>0\}} + \mathbb{1}_{J_k}\mathbb{1}_{\{Z_k\leq 0\}}|\mathcal{F}_{k-1}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{J_k\cap\{Z_k>0\}}|\mathcal{F}_{k-1}\right] + \mathbb{1}_{\{Z_k\leq 0\}}\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \qquad (3.6)\\
&\leq \mathbb{P}\left(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq Z_k^+ \Delta_k^q |\mathcal{F}_{k-1}\right) + \mathbb{1}_{\{Z_k\leq 0\}}\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right],
\end{aligned}$$

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

$$\begin{aligned}
&-\rho_k \mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \geq -\rho_k^+ \mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&\geq -\rho_k^+ (\mathbb{P}\left(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \rho_k^+ \Delta_k^q |\mathcal{F}_{k-1}\right) + \mathbb{1}_{\{\rho_k\leq 0\}}\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right]) \\
&= -\rho_k^+ \mathbb{P}\left(|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \geq \rho_k^+ \Delta_k^q |\mathcal{F}_{k-1}\right) \\
&\geq -\rho_k^+ \min\left(1, \varepsilon(\rho_k^+)\right) = -\rho_k^+ \min\left(1, \varepsilon(\rho_k^+)\right) \geq -\rho_k^+ \min\left(1, \varepsilon_1(\rho_k^+)\right) \geq -\varepsilon_f\,, \qquad (3.7)
\end{aligned}$$

where we applied (3.6) in the first inequality, the second inequality is a direct consequence of Lemma 2.3 for $\mathrm{A} = Z_k^+$, and $\varepsilon_1(t) = +\infty \mathbb{1}_{\{0\}} + \varepsilon_q / t \cdot \mathbb{1}_{(0,+\infty)}$. Hence,

$$-Z_k \Delta_k^q \mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta \tau_q^- \Delta_k^q \geq (-\varepsilon_f + \eta\tau_q^-)\Delta_k^q = \varepsilon \Delta_k^q\,, \qquad (3.8)$$

where we used (3.7) in the inequality.

Claim (3.2) can finally be obtained by concatenating (3.5) and (3.8). □

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

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

$$\liminf_{k\in K,\, k\to\infty} \frac{f(X_k + \Delta_k G_k) - f(X_k)}{\Delta_k} \geq 0\,. \qquad (3.9)$$

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

$$\liminf_{k\in K,\, k\to\infty} \frac{f(X_k + \Delta_k G_k) - f(X_k)}{\Delta_k} \geq -\frac{1}{m}\,. \qquad (3.10)$$

To start with, by applying Lemma 2.3 with $\mathrm{A} = \frac{\Delta_k^{1-q}}{m}$ we have

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

and therefore taking expectations on both sides

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

We can now deduce

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

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

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

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

$$|F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))| \leq \frac{\Delta_k}{m} \quad \text{for } k \text{ large enough.} \tag{3.11}$$

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

$$\begin{aligned}\frac{f(X_k + \Delta_k G_k) - f(X_k)}{\Delta_k} &\geq \frac{F_k^g - F_k - |F_k - F_k^g - (f(X_k) - f(X_k + \Delta_k G_k))|}{\Delta_k} \\ &\geq -\theta\Delta_k - \frac{1}{m}\,,\end{aligned} \tag{3.12}$$

where we used (3.11) combined with the unsuccessful step condition of Algorithm 1 in the second inequality. Finally, (3.10) follows passing to the liminf for $k \to \infty$ in (3.12). □

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 as in [2, Assumption 1].

Theorem 3.3. *Assume that f is Lipschitz continuous with constant L_f^* around any limit point of the sequence of iterates $\{X_k\}$. Let K be the random set of indices of unsuccessful iterations. Let Assumptions 2.1 and (3.1) hold. Then, the following property holds a.s. in Ω: if $L \subset K$ is a random set such that the sequence $\{G_k\}_{k\in L}$ is dense in the unit sphere and $\lim_{k\in L,\, k\to\infty} X_k = X^*$, then the point X^* is Clarke stationary, i.e., $f^\circ(X^*, d) \geq 0$ for every $d \in \mathbb{R}^n$.*

Proof. We refer to $\mathcal{V}$ as the event with probability one that (3.9) holds, and assume $\omega \in \mathcal{V}$ in the rest of the proof, with $L(\omega), K(\omega)$ satisfying the assumption described in the statement. Let d be a direction in the unit sphere, and let $S(\omega) \subset L(\omega)$ be such that $\lim_{k\in S(\omega),\, k\to\infty} G_k(\omega) = d$. By definition of Clarke stationarity, since

$$f^\circ(X^*(\omega), d) \geq \limsup_{k\in S(\omega),\, k\to\infty} \frac{f(X_k(\omega) + \Delta_k(\omega)d) - f(X_k(\omega))}{\Delta_k(\omega)}\,,$$

we just need to prove that on $\mathcal{V}$, and therefore a.s.,

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

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

$$\begin{aligned} &\limsup_{k \in S(\omega),\, k \to \infty} \frac{f(X_k(\omega) + \Delta_k(\omega) G_k(\omega)) - f(X_k(\omega))}{\Delta_k(\omega)} \\ &\geq \liminf_{k \in K(\omega),\, k \to \infty} \frac{f(X_k(\omega) + \Delta_k(\omega) G_k(\omega)) - f(X_k(\omega))}{\Delta_k(\omega)} \geq 0 \,, \end{aligned} \tag{3.13}$$

where the last inequality follows by (3.9).

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

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

Passing to the limsup for $k \in S(\omega)$ we get

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

for every $\omega \in \mathcal{V}$, where we used $\|G_k(\omega) - d\| \to 0$ by construction in the first inequality and (3.13) in the second. □

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

After having analyzed a simple stochastic direct-search method, we focus on a stochastic version of the Basic DFO-TRNS presented in [25], and analyze its convergence properties under tail bound probabilistic conditions like the ones used in Section 3. Some minor changes in notation are convenient and will be introduced with a clear reference to the corresponding elements of Algorithm 1.

4.1 The stochastic trust-region scheme

As already mentioned, the simple trust-region algorithm that we reported here is a minor modification of the Basic DFO-TRNS algorithm proposed in [25]. Indeed, there are two differences between the Basic DFO-TRNS algorithm and its stochastic counterpart.

The first difference is in the updating rule related to the trust-region radius. In the modification presented in this work, $\tau \in (0, 1)$ is chosen, with $1 - \tau$ corresponding contraction factor and $\bar{\tau} \in [1, 1 + \tau]$ as expansion factor.

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

The detailed scheme is reported in Algorithm 2. At every iteration k, a symmetric matrix B_k is built from interpolation or regression on a sample set of points. The linear term g_k needs to randomly cover the unit sphere when normalized. By using these quantities, a quadratic model of the objective function around x_k is built. The step s_k is obtained by solving the trust-region subproblem, i.e., by minimizing the quadratic model within the spherical trust-region constraint. Once the current step has been computed, the algorithm generates an estimate of the true objective function f at the trial point $x_k + s_k$ and recomputes a new estimate at x_k, after which the acceptance ratio $\bar{\rho}_k$ is computed. Note that, as in [25], the non-standard acceptance ratio is motivated by convergence requirements. In this scheme, realizations related to the estimates of the function values $f(x_k)$ at the current iterate and $f(x_k + s_k)$ at the potential next iterate are indicated with f_k and f_k^s, thus replacing f_k^g used in the direct-search scheme, as a shorthand for $F_k(\omega)$ and $F_k^s(\omega)$, respectively.

Algorithm 2 `Stochastic DFO Trust-Region Algorithm`

1 **Initialization.** Select $x_0 \in \mathbb{R}^n$, $\theta > 0$, $\tau \in (0, 1)$, $\bar{\tau} \in [1, 1 + \tau]$, $\delta_0 > 0$, $q > 1$.

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

3 Select a direction $g_k \neq 0$ and build a symmetric matrix B_k.

4 Compute
$$s_k \in \arg\min_{\|s\| \leq \delta_k} g_k^\top s + \frac{1}{2} s^\top B_k s.$$

5 Compute estimates f_k, f_k^s for f at x_k, $x_k + s_k$, respectively, and let
$$\bar{\rho}_k = \frac{f_k - f_k^s}{\theta \|s_k\|^q}.$$

6 **If** $\bar{\rho}_k \geq 1$ **Then** set `SUCCESS` = `true`, $x_{k+1} = x_k + s_k$, $\delta_{k+1} = \bar{\tau}\delta_k$.

7 **Else** set `SUCCESS` = `false`, $x_{k+1} = x_k$, $\delta_{k+1} = (1 - \tau)\delta_k$.

8 **End If**

9 **End For**

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

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

When $\|G_k\| = 1$, the above assumption is essentially saying that B_k can be unbounded as long as it does not go to infinity faster than $1/\Delta_k$, that is a weaker version of [25, Assumption 2.1].

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

Lemma 4.2. *Under Assumption 4.1,* $\|S_k\| \geq \rho\Delta_k$.

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

4.2 Convergence analysis under the tail bound probabilistic condition

In order to analyze the method introduced above, we adapt Assumption 2.1, replacing G_k with $\hat{S}_k$, using the $\hat{\cdot}$ notation introduced at the beginning of Section 2, and Δ_k with $\|S_k\|$. Now Δ_k stands for the trust-region radius. Hence, we obtain the following tail bound condition.

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

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

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

Importantly, if B_k is random the definition of $\mathcal{F}_{k-1}$ must be modified as the σ-algebra of events up to the generation of B_k and g_k.
The next theorem implies convergence of the series of trust-region radii elevated to the q almost surely. This obviously implies that the trust-region radius converges to zero almost surely.

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

$$\theta > \frac{\left(\rho^q\tau_q^- + \tau_q^+\right) \sqrt[r(q)]{\varepsilon_q}}{\rho^q\tau_q^-}, \tag{4.1}$$

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

Proof. Let $\varepsilon_f = \sqrt[r(q)]{\varepsilon_q}$, $\Phi_k = f(X_k) - f^* + \eta\|S_k\|^q$, with $\eta = \frac{\theta\rho^q}{\tau_q^+ + \rho^q\tau_q^-}$. Let also $\varepsilon = -\varepsilon_f + \frac{\rho^q\theta}{\tau_q^+ + \rho^q\tau_q^-} > 0$, where the inequality follows by (4.1). We will prove, for every $k \geq 0$, that

$$\mathbb{E}\left[\Phi_k - \Phi_{k+1} \mid \mathcal{F}_{k-1}\right] \geq \varepsilon\|S_k\|^q. \tag{4.2}$$

Then for the same reasons stated in the proof of Theorem 3.1, with $\|S_k\|$ instead of Δ_k, we get

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

and therefore

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

where we used Assumption 4.1 in the inequality. Let Z_k be the random variable such that $f(X_k) - f(X_k + S_k) = (\theta - Z_k)\|S_k\|^q$, and let J_k be the event that the step k is successful. We have

$$\begin{aligned}
&\mathbb{E}\left[(\Phi_k - \Phi_{k+1})|\mathcal{F}_{k-1}\right] \\
&\geq (f(X_k) - f(X_k + S_k) - \eta\tau_q^+\Delta_k^q)\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&\quad + \eta(\Delta_k^q - \Delta_{k+1}^q)\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&\geq (\theta - Z_k)\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right]\|S_k\|^q - \eta\tau_q^+\Delta_k^q\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\Delta_k^q\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&\geq (\theta - Z_k)\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right]\|S_k\|^q - \eta\frac{\tau_q^+}{\rho^q}\|S_k\|^q\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\|S_k\|^q\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] \\
&= \left((\theta - Z_k) - \eta\frac{\tau_q^+}{\rho^q}\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right]\right)\|S_k\|^q
\end{aligned} \tag{4.3}$$

where the first inequality follows as in (3.3), the second inequality by definition of Z_k, and the third inequality we use (4.1) on the second summand and $\|S_k\| \leq \Delta_k$ in the third summand. In turn,

$$\begin{aligned}
&\left((\theta - Z_k) - \eta\frac{\tau_q^+}{\rho^q}\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\mathbb{E}\left[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right]\right)\|S_k\|^q \\
&= \left(\theta - Z_k - \eta\left(\frac{\tau_q^+}{\rho^q} + \tau_q^-\right)\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\right)\|S_k\|^q \\
&= -Z_k\|S_k\|^q\mathbb{E}\left[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}\right] + \eta\tau_q^-\|S_k\|^q\,,
\end{aligned} \tag{4.4}$$

where we used $\mathbb{E}[1 - \mathbb{1}_{J_k}|\mathcal{F}_{k-1}] = 1 - \mathbb{E}[\mathbb{1}_{J_k}|\mathcal{F}_{k-1}]$ in the first equality, and $\eta = \frac{\theta\rho^q}{\tau_q^+ + \rho^q\tau_q^-}$ in the second one. By combining (4.3) and (4.4) we therefore get

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

The conclusion now follows as in the proof of Theorem 3.1, replacing Δ_k with $\|S_k\|$. □

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

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

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

Proof. Follows analogously to Lemma 3.2. □

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

Theorem 4.6. *Assume that f is Lipschitz continuous with constant L_f^* around any accumulation point of the sequence of iterates $\{X_k\}$. Let K be the random set of indices of unsuccessful iterations. Let Assumptions 4.1, 4.3, and (4.1) hold. Then, the following property holds a.s. in Ω: if $L \subset K$ is a random set such that $\{\hat{S}_k\}_{k\in L}$ is dense in the unit sphere and $\lim_{k\in L,\, k\to\infty} X_k = X^*$, then the point X^* is Clarke stationary, i.e., $f^\circ(X^*, d) \geq 0$ for every $d \in \mathbb{R}^n$.*

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

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

Assumption 4.7. *For some positive sequence of uniformly bounded random variables $\{A_k\}$ such that $A_k \to 0$ a.s., it holds a.s. $\|B_k\| \leq A_k\|G_k\|/\Delta_k$.*

Trivially, Assumption 4.7 implies Assumption 4.1, with $\rho = \frac{1}{\max(\{\|A_k\|_\infty\})}$.

Proposition 4.8. *Let Assumptions 4.3, 4.7, and (4.1) hold. Then a.s. $\lim_{k\to\infty} \hat{G}_k + \hat{S}_k = 0$.*

Proof. First, notice that $\|\hat{G}_k\| = 1$, as well as $\|\hat{S}_k\| = 1$ since G_k must be always different from 0 and therefore S_k as well. Now define F_k^m as the local model $F_k^m(s) = G_k^\top s + \frac{1}{2}s^\top B_k s$, and let $\Gamma_k = \hat{G}_k^\top \hat{S}_k$ be the cosine of the angle between $\hat{G}_k$ and $\hat{S}_k$. We need to prove $\Gamma_k \to -1$ almost surely.

We have on the one hand

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

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

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

Putting (4.5) and (4.6) together we obtain

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

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

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

Since the right-hand side of (4.8) converges to 0 a.s. thanks to Assumption 4.7, we get $1 + \min(0, \gamma_k) \to 0$ a.s., and we can conclude $\Gamma_k \to -1$ a.s. as desired. □

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

Corollary 4.9. *Assume that f is Lipschitz continuous with constant L_f^* around any accumulation point of the sequence of iterates $\{X_k\}$. Let K be the random set of indices of unsuccessful iterations. Let Assumptions 4.3, 4.7 and (4.1) hold. Then, the following property holds a.s. in Ω: if $L \subset K$ is a random set such that the sequence $\{\hat{G}_k\}_{k\in L}$ is dense in the unit sphere and $\lim_{k\in L, k\to\infty} X_k = X^*$ then the point X^* is Clarke stationary, i.e., $f^\circ(X^*, d) \geq 0$ for all $d \in \mathbb{R}^n$.*

Proof. Thanks to Proposition 4.8, for almost every ω in Ω, if the sequence $\{\hat{G}_k(\omega)\}_{k\in L}$ is dense in the unit sphere $\{\hat{S}_k(\omega)\}_{k\in L}$ also is, and we can therefore apply Theorem 4.6. □

5 Numerical results

We report here some numerical results, first comparing the performance of Algorithm 1 for different choices of q, and then comparing Algorithms 1 and 2 to StoMADS from [2].
To compare the performance of the algorithms, we use data and performance profiles as defined in [30]. Their definitions are briefly recalled here. Given a set S of algorithms and a set P of problems, for $s \in S$ and $p \in P$, let $t_{p,s}$ be the number of function evaluations required by algorithm s on problem p to satisfy the condition

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

where $\gamma_p > 0$ and f_L is the best objective function value achieved by any solver on problem p. Then, the performance and data profiles of solver s are generated using

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

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

5.1 Algorithm 1 for different choices of q

In this section, we compare two basic instances of Algorithm 1, obtained choosing uniformly at random the search direction in the unit sphere, for different choices of the sufficient decrease parameter and sampling strategies, corresponding to different values of q and r in the

algorithmic scheme and in the assumptions. The main goal is to provide further evidence that choosing q smaller than 2 and using fewer samples per iteration as suggested by the theory can improve numerical performance. In particular, we will show that the claim remains true also in the case of correlated errors discussed in Section 2.2.

Remark 5.1. It is of course not always the case that an improvement in number of samples per iteration leads to an improvement in the solution found with a fixed budget of samples, since using lower values of q might increase the iteration complexity. For instance, for smooth objectives with deterministic oracles, a complexity of $O(\epsilon^{-\frac{q}{q-1}})$ was proved in [37] for a scheme analogous to Algorithm 1, with $q \in (1, 2]$. Then in this case, the lower number of samples per iteration for q approaching 1 comes at the price of a potentially much higher iteration complexity. However, it is important to note that the complexity bounds from [37] heavily rely on the Lipschitz continuity of the gradient, so that this trade-off does not necessarily generalize to potentially non-smooth objectives.

The basic version of Algorithm 1 used here is referred to as SDSq for $q \in \{2, 1.5\}$. We are therefore comparing a standard choice [2, 13, 14] to one allowing the use of a lower number of samples per iteration as proved in Theorems 2.4 and 2.8. The noise on the objective was assumed to be 0 in expectation and normally distributed with standard deviation 0.1. By Theorem 2.4, $O(\Delta_k^{-2q})$ samples are needed to satisfy the weak tail bound assumptions. Given that the Gaussian noise has finite r-th moment for every r, Theorem 2.4 can be applied with $r = q/(q-1)$. The number of samples needed per iteration is then $O(\Delta_k^{-4})$ and $O(\Delta_k^{-3})$ respectively for $q = 2$ and $q = 1.5$. Thus, we simulated the resulting noise after averaging p_k independent samples by adding to the objective $N(0, 1/\sqrt{p_k})$ distributed random variables. The remaining parameters were tuned with a basic grid search to obtain good performances for both instances of Algorithm 1 to $\tau = 0.001$, $\bar{\tau} = 1.001$, $\theta = 0.5$ and $\delta_0 = 2$.

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

Data and performance profile in the general case of finite r-th moment and correlated errors are reported in Figures 1 and 2 respectively. In the case of finite r-th moment, the number of samples was set equal to $p_k = \lceil 0.01\delta_k^{-4} \rceil$ and $p_k = \lceil 0.01\delta_k^{-3} \rceil$ for $q = 2$ and $q = 1.5$ respectively, consistently with the bound proved in Theorem 2.4. In the correlated error case, at the iteration k we add noise with standard deviation $0.1\delta_k$ in the estimate of the difference and set the number of samples to $\lceil 0.01\delta_k^{-2} \rceil$ and $\lceil 0.01\delta_k^{-1} \rceil$ for $q = 2$ and $q = 1.5$ respectively, consistently with the bound proved in Theorem 2.8. For both cases it can clearly be seen how choosing $q = 1.5$ rather than $q = 2$ leads to a better performance.

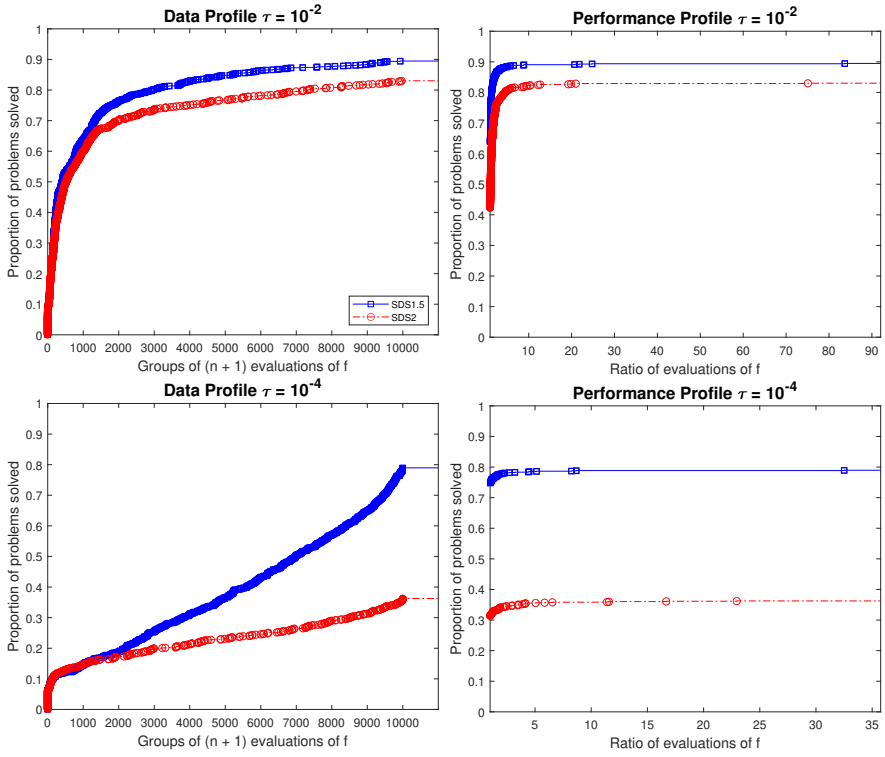

Figure 1: Data and performance profiles for Algorithm 1 with $q \in \{2, 1.5\}$ and in the finite $r-$th moment setting, corresponding to SDS2 and SDS1.5, on the set of problems reported in Table 1.

5.2 Comparison of Algorithms 1 and 2 with STOMADS

We describe in this section numerical results comparing a modified version of Algorithm 1, Algorithm 2, and the StoMADS algorithm from [2]. The version of Algorithm 1 considered here is obtained alternating coordinate search directions with random directions after the stepsize falls below a certain threshold $\bar{\Delta}$, like it was done in the deterministic case, e.g., in [15, 20]. The convergence result of Theorem 3.3 extends to this variant in a straightforward way. In the tests, we used the threshold $\bar{\Delta} = 0.5$. As for Algorithm 2, we adopt at all iterations the approach described in [6, Section 5] to build the model at iteration 0, i.e., we build a minimum Frobenius norm model using the sample set $\{x_k\} \cup \{x_k \pm \delta_k e_i\}_{i\in[1:n]}$ at all iterations k. Unlike in [6, Algorithm 5.1] we do not add or subtract any point to the sample set, and rebuild from scratch the model at every iteration.

For both Algorithm 1 and Algorithm 2, two choices of the parameter q were tested, that is $q = 2$ and $q = 1.5$. The two instances of the modified version of Algorithm 1 are referred to as SDS+q for $q \in \{2, 1.5\}$, and analogously the two instances of Algorithm 2 are referred to as STRq for $q \in \{2, 1.5\}$. In accordance with this bound and the one in our Theorem 2.4, the number of samples was set equal to $p_k = \lceil 0.01\delta_k^{-4} \rceil$ and $p_k = \lceil 0.01\delta_k^{-3} \rceil$ for $q = 2$ and $q = 1.5$ respectively, like in the previous section. In the case of StoMADS, when using the default choice $q = 2$ for the frame size exponent in the acceptance criterion, the theory in [2] suggests

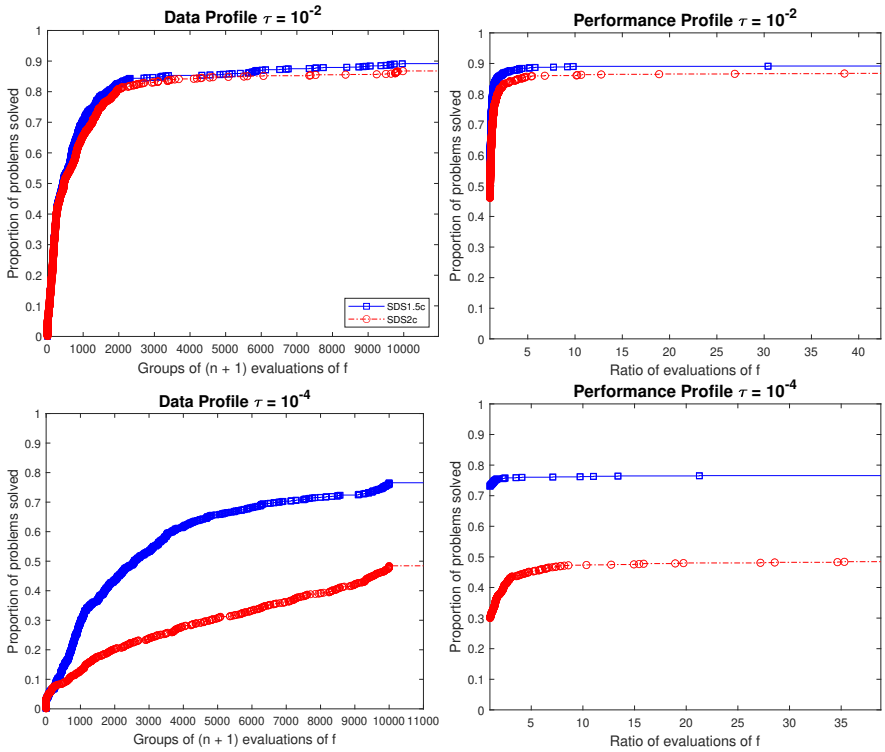

Figure 2: Data and performance profiles for Algorithm 1 with $q \in \{2, 1.5\}$ and in the correlated error setting, corresponding to SDS2c and SDS1.5c, on the set of problems reported in Table 1.

the use of $O(\Delta_k^{-4})$ samples per iteration.

By taking a look at the profiles, it can be easily seen that SDS+1.5 and STR1.5 outperform the other methods for $\gamma_p = 10^{-2}$ and $\gamma_p = 10^{-4}$ respectively. This suggests that the algorithms analyzed in this work can outperform StoMADS, and that using $q = 1.5$ with fewer samples per iteration can give better performances both for Algorithm 1 and 2. The trust-region method also appear to show better performance when considering lower accuracy parameters. The trus-region approach seems to work better than the direct-search one, which is not surprising even considering fixed geometry model building.

6 Concluding remarks and future work

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

Our tail bound assumption allowed us to obtain convergence of both a direct-search and a trust-region method using a reduced number of samples per iteration. Surprisingly, unlike

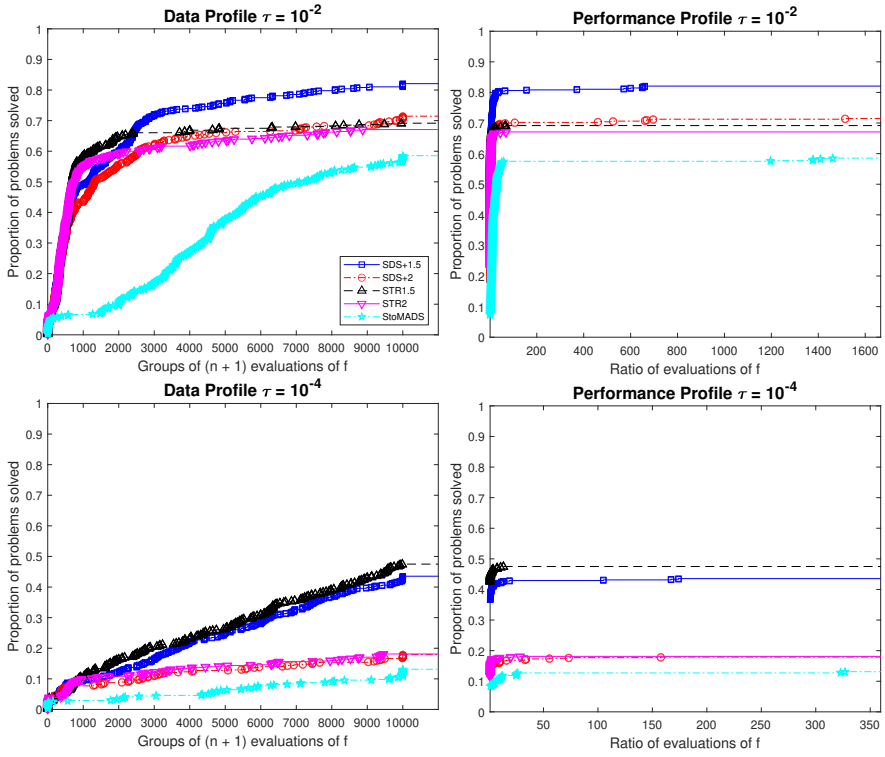

Figure 3: Data and performance profiles for Algorithm 1 with $q \in \{2, 1.5\}$, Algorithm 2 with $q \in \{2, 1.5\}$ and StoMADS.

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

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

References

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

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

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

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

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

[6] Afonso S Bandeira, Katya Scheinberg, and Luís Nunes Vicente. Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization. *Math. Program.*, 134:223–257, 2012.

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

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

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

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

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

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

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

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

[15] G. Fasano, G. Liuzzi, S. Lucidi, and F. Rinaldi. A linesearch-based derivative-free approach for nonsmooth constrained optimization. *SIAM J. Optim.*, 24:959–992, 2014.

[16] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23:2341–2368, 2013.

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

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

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

[20] Vyacheslav Kungurtsev, Francesco Rinaldi, and Damiano Zeffiro. Retraction based direct search methods for derivative free riemannian optimization. *arXiv preprint arXiv:2202.11052*, 2022.

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

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

[23] Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Stochastic zeroth-order riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 2022, to appear.

[24] Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35:26160–26175, 2022.

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

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

[27] S. Lucidi and M. Sciandrone. On the global convergence of derivative-free methods for unconstrained optimization. *SIAM J. Optim.*, 13:97–116, 2002.

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

[29] Matt Menickelly, Stefan M Wild, and Miaolan Xie. A stochastic quasi-Newton method in the absence of common random numbers. *arXiv preprint arXiv:2302.09128*, 2023.

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

[31] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17:527–566, 2017.

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

[33] Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. Stochastic zeroth-order optimization under nonstationarity and nonconvexity. *Journal of Machine Learning Research*, 23:1–47, 2022.

[34] Katya Scheinberg and Miaolan Xie. Stochastic adaptive regularization method with cubics: A high probability complexity bound. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

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

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

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

[38] Y. S. Virkar. *Power-law distributions and binned empirical data*. PhD thesis, University of Colorado at Boulder, 2012.

A Appendix

In this appendix, we report the missing proofs and some additional numerical results.

A.1 Proofs

We now recall the Rosenthal inequality [18, Equation (1)], together with a corollary useful for several results of Section 2.2. This inequality states that, if $\{Z_i\}_{i\in[1:p]}$ is a sequence of independent random variables with 0 mean and finite r-th moment, $r > 2$, and $S = \frac{1}{p}\sum_{i=1}^{p} Z_i$, one has

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

for some constant $C_r > 0$ depending from r.
We report here the corollary, which concerns the special case of i.i.d. samples.

Proposition A.1. *If $\{Z_i\}$ is a sequence of independent copies of a random variable Z, $r \geq 2$ and S are defined as above, then*

$$\mathbb{E}\left[|S|^r\right] \leq C_r p^{-\frac{r}{2}} \mathbb{E}\left[|Z|^r\right] . \tag{A.2}$$

Proof. For $r = 2$, the result trivially holds with $C_r = 1$, since

$$\mathbb{E}\left[|S|^2\right] = p^{-2} \sum_{i=1}^{p} \mathbb{E}\left[Z_i^2\right] = p^{-1} \mathbb{E}\left[Z^2\right] .$$

Under the assumptions of this proposition (A.1) reduces to

$$\mathbb{E}\left[|S|^r\right] \leq p^{-r} C_r \max\left(p\mathbb{E}\left[|Z|^r\right], p^{\frac{r}{2}} \mathbb{E}\left[|Z|^2\right]^{\frac{r}{2}}\right) . \tag{A.3}$$

Now,

$$\begin{aligned} &p^{-r} C_r \max\left(p\mathbb{E}\left[|Z|^r\right], p^{\frac{r}{2}} \mathbb{E}\left[|Z|^2\right]^{\frac{r}{2}}\right) \\ &\leq C_r p^{-r} \max\left(p\mathbb{E}\left[|Z|^r\right], p^{\frac{r}{2}} \mathbb{E}\left[|Z|^r\right]\right) \leq C_r p^{-\frac{r}{2}} \mathbb{E}\left[|Z|^r\right] , \end{aligned} \tag{A.4}$$

where Jensen's inequality is used on the second argument of the max operator in the first inequality and $r \geq 2$ and $p \geq 1$ in the second inequality. By concatenating (A.3) and (A.4), (A.2) is proved. □

Proof of Theorem 2.4. Let $\bar{F}_k = F_k - f(X_k)$ and $\bar{F}_k^g = F_k^g - f(X_k + \Delta_k G_k)$, for F_k and F_k^g average of p_k samples, with $\xi_{k,i}$ and $\xi_{k,i}^g$ independent samples for $i \in [1 : p]$:

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

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

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

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

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

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

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

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

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

$$\mathbb{E}\left[|\bar{F}_k|^r \mid \mathcal{F}_{k-1}\right] \leq C_r p_k^{-\frac{r}{2}} \mathbb{E}\left[|Z|^r \mid \mathcal{F}_{k-1}\right] \leq C_r M_r p_k^{-\frac{r}{2}}, \tag{A.8}$$

where we used (2.9) in the second inequality. Of course (A.8) holds with $\bar{F}_k^g$ instead of $\bar{F}_k$ as well. Then, reasoning as in (A.5), we get

$$\mathbb{E}\left[|A_k|^r \mid \mathcal{F}_{k-1}\right] \leq 2^r C_r M_r p_k^{-\frac{r}{2}},$$

and analogously to (A.7):

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

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

Proof of Proposition 2.7. By setting $x = y$ in (2.11) we get

$$\mathrm{Var}_\xi[F(x,\xi)] = \sigma^2. \tag{A.9}$$

Moreover, we have

$$\mathrm{Cov}_\xi(F(x,\xi),F(y,\xi)) = \sigma^2 \exp\left(-\frac{\|x-y\|^2}{2l^2}\right) \geq \sigma^2\left(1 - \frac{\|x-y\|^2}{2l^2}\right), \tag{A.10}$$

where we used (2.11) in the equality and $e^x \geq 1+x$ in the inequality. We therefore have

$$\begin{aligned}
&\mathbb{E}_\xi\left[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^2\right] = \mathbb{E}_\xi\left[\bar{F}(x,\xi)^2\right] + \mathbb{E}_\xi\left[\bar{F}(y,\xi)^2\right] - 2\mathbb{E}_\xi\left[\bar{F}(x,\xi)\bar{F}(y,\xi)\right] \\
&= \mathrm{Var}_\xi(F(x,\xi)) + \mathrm{Var}_\xi(F(y,\xi)) - 2\mathrm{Cov}_\xi(F(x,\xi),F(y,\xi)) \\
&\leq 2\sigma^2 - 2\sigma^2\left(1 - \frac{\|x-y\|^2}{2l^2}\right) = \frac{\sigma^2}{l^2}\|x-y\|^2.
\end{aligned} \tag{A.11}$$

where we applied (A.9) and (A.10) in the last inequality. Let now $V_{x,y} = \mathbb{E}_\xi\left[|\bar{F}(x,\xi) - \bar{F}(y,\xi)|^2\right] = \mathrm{Var}_\xi\left[\bar{F}(x,\xi) - \bar{F}(y,\xi)\right]$. $(F(x,\xi),F(y,\xi))$ is a bivariate Gaussian vector since $\{(F(x,\xi)\}$ is a Gaussian process. Thus, the linear combination $F(x,\xi) - F(y,\xi)$ is still Gaussian, whence $\bar{F}(x,\xi) - \bar{F}(y,\xi)$ is Gaussian with mean 0. In particular, we can write $\bar{F}(x,\xi) - \bar{F}(y,\xi) = \sqrt{V_{x,y}}N$, with N having standard normal distribution. We conclude by noticing, for $r \geq 2$,

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

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

Proof of Theorem 2.8. Let A_k be an estimate of the difference between the errors at the current and the tentative points obtained with p_k samples:

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

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

$$\mathbb{E}\left[|A_k|^r \mid \mathcal{F}_{k-1}\right] \le C_r p_k^{-\frac{r}{2}}\mathbb{E}\left[|Z|^r \mid \mathcal{F}_{k-1}\right] \le D_r C_r p_k^{-\frac{r}{2}}\|\Delta_k G_k\|^r = D_r C_r p_k^{-\frac{r}{2}}\Delta_k^r\,, \tag{A.12}$$

where we used the conditional version of (A.2) in the first inequality, (2.10) in the second inequality, and $\|G_k\| = 1$ in the equality.

We thus have

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

where we used the conditional Chebyshev's inequality in the first inequality, and (A.12) in the last inequality. Hence we obtain Assumption 2.1 for $p_k = O(\Delta_k^{-\frac{2}{r-1}}) = O(\Delta_k^{2-2q})$ as desired. □

Proof of Proposition 2.10. First, notice that

$$\begin{aligned}&\mathbb{E}\left[|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))|^2 \mid \mathcal{F}_{k-1}\right]\\ &\le 2(\mathbb{E}\left[|F_k^g - f(X_k+\Delta_k G_k)|^2 \mid \mathcal{F}_{k-1}\right] + \mathbb{E}\left[|F_k - f(X_k)|^2 \mid \mathcal{F}_{k-1}\right])\\ &\le 4k_f^2\Delta_k^4\,,\end{aligned} \tag{A.13}$$

where we used $(a+b)^2 \le 2(a^2+b^2)$ for $a, b \in \mathbb{R}$ in the first inequality, and (2.12) in the second. We now have

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

where we used the conditional Chebyshev's inequality in the first inequality, and (A.13) in the second inequality. By setting $\varepsilon_q = 4k_f^2$ in the above equation we obtain

$$\mathbb{P}[|F_k - F_k^g - (f(X_k) - f(X_k+\Delta_k G_k))| \ge \alpha\Delta_k^2 \mid \mathcal{F}_{k-1}] \le \frac{\varepsilon_q}{\alpha^2}$$

as desired. □

Proof of Proposition 2.12. Notice that (2.2) is trivially satisfied for $\alpha < \sqrt{\varepsilon_q}$. We then just need to deal with the case $\alpha \geq \sqrt{\varepsilon_q}$. First observe that by the triangular inequality

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

which proves in particular (2.14). Let $\alpha \geq \sqrt{\varepsilon_q}$ be arbitrary. For $\beta = 1 - \frac{\varepsilon_q}{\alpha^2}\bar{p} \in [1-\bar{p}, 1)$,

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

where the second inequality follows from (2.13), and we were able to apply (2.14) in the first inequality since by assumption

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

using $\varepsilon_q = \frac{\varepsilon}{\bar{p}}$ in the last equality. Given that $\alpha \geq \sqrt{\varepsilon_q}$ is arbitrary, this concludes the proof. □

A.2 Benchmark problems

Table 1: Problems used in numerical experiments.

name	dimension	reference
crescent	2	[30]
cb2	2	[28]
charconn1	2	[26]
charconn2	2	[26]
demyanov-malozemov	2	[26]
dennis-woods	2	[12]
wong1	7	[28]
wong2	10	[28]
wong3	20	[28]
elattar	6	[28]
goffin	50	[28]
hald-madsen 1	2	[26]
lq	2	[28]
ql	2	[28]
maxl	20	[28]
maxq	20	[30]
mifflin 1	2	[19]
mifflin 2	2	[19]
rosen-suzuki	4	[28]
wf	2	[28]
spiral	2	[28]
evd 52	3	[28]
kowalik-osborne	4	[28]
oet 5	4	[28]
oet 6	4	[28]
gamma	4	[28]
exp	5	[28]
pbc1	5	[28]
evd61	6	[28]
filter	9	[28]
polak 2	10	[28]
polak 3	11	[28]
polak 6	4	[28]

name	dimension	reference
watson	20	[28]
osborne 2	11	[28]
shor	5	[28]
colville 1	5	[28]
hs 78	5	[28]
maxquad	10	[28]
gill	10	[28]
mxhilb	50	[19]
l1hilb	50	[28]
davidon 2	4	[28]
shelldual	15	[28]
steiner 2	12	[28]
transformer	6	[28]
polak 6.10	1	[28]
wild1	20	[19]
wild2	20	[19]
wild3	20	[19]
wild19	20	[19]
wild11	20	[19]
wild16	20	[19]
wild20	20	[19]
wild15	20	[19]
wild21	20	[19]
maxq	{10, 20, 30, 40}	[19]
l1hilb	{10, 20, 30, 40}	[30]
lq	{10, 20, 30, 40}	[30]
cb3	{10, 20, 30, 40}	[30]
cb32	{10, 20, 30, 40}	[30]
af	{10, 20, 30, 40}	[30]
brown	{10, 20, 30, 40}	[30]
mifflin2	{10, 20, 30, 40}	[30]
crescent	{10, 20, 30, 40}	[30]
crescent2	{10, 20, 30, 40}	[30]