Complexity and global rates of trust-region methods based on probabilistic models

S. Gratton* C. W. Royer† L. N. Vicente‡ Z. Zhang§

June 14, 2017

Abstract

Trust-region algorithms have been proved to globally converge with probability one when the accuracy of the trust-region models is imposed with a certain probability conditioning on the iteration history. In this paper, we study the complexity of such methods, providing global rates and worst case complexity bounds on the number of iterations (with overwhelmingly high probability), for both first and second order measures of optimality. Such results are essentially the same as the ones known for trust-region methods based on deterministic models. The derivation of the global rates and worst case complexity bounds follows closely from a study of direct-search methods based on the companion notion of probabilistic descent.

1 Introduction

Trust-region methods form a well established and understood class of methods for the minimization of a nonlinear (possibly nonsmooth) function subject or not to constraints on its variables (see the book [7] and the recent survey [28]). They have also been comprehensively studied in the context of derivative-free optimization (DFO), where the derivatives of the objective or constraint functions cannot be computed or approximated (see the book [12] and the recent survey [14]). In this paper we focus on the unconstrained minimization of a smooth objective function $f : \mathbb{R}^n \to \mathbb{R}$ without using its derivatives.

In the derivative-free setting, trust-region algorithms use sampled points to build a model of the objective function around the current iterate, typically by quadratic interpolation. The quality of these models is measured by the accuracy they provide relatively to a Taylor expansion in a ball $B(x, \delta)$ of center x and radius δ. Models that are as accurate as first-order Taylor expansions are called fully linear [9, 12].

*University of Toulouse, IRIT, 2 rue Charles Camichel, B.P. 7122 31071, Toulouse Cedex 7, France (serge.gratton@enseeiht.fr).

†Wisconsin Institute for Discovery, University of Wisconsin-Madison, 330 N Orchard Street, Madison WI 53715, USA (croyer2@wisc.edu). Support for this author was partially provided by Université Toulouse III Paul Sabatier under a doctoral grant.

‡CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra, Portugal (lnv@mat.uc.pt). Support for this author was partially provided by FCT under grants UID/MAT/00324/2013, SFRH/BSAB/114622/2016, and P2020 SAICTPAC/0011/2015.

§Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China (zaikun.zhang@polyu.edu.hk). Support for this author was provided by the Hong Kong Polytechnic University under the start-up grant 1-ZVHT.

Definition 1.1 *Given a function $f \in \mathcal{C}^1$ and constants $\kappa_{ef}, \kappa_{eg} > 0$, a $\mathcal{C}^1$ function $m : \mathbb{R}^n \to \mathbb{R}$ is called a* $(\kappa_{eg}, \kappa_{ef})$-fully linear model *of f on $B(x, \delta)$ if, for all $s \in B(0, \delta)$,*

$$\begin{array}{rcl} |m(s) - f(x+s)| & \leq & \kappa_{ef}\delta^2, \\ \|\nabla m(s) - \nabla f(x+s)\| & \leq & \kappa_{eg}\delta. \end{array}$$

Fully linear models are not necessarily linear or affine functions. Models that are as accurate as second-order Taylor expansions are called fully quadratic [9, 12]. Similarly, such models are not necessarily quadratic functions.

Definition 1.2 *Given a function $f \in \mathcal{C}^2$ and constants $\kappa_{ef}, \kappa_{eg}, \kappa_{eh} > 0$, a $\mathcal{C}^2$ function $m : \mathbb{R}^n \to \mathbb{R}$ is called a* $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic model *of f on $B(x, \delta)$ if, for all $s \in B(0, \delta)$,*

$$\begin{array}{rcl} |m(s) - f(x+s)| & \leq & \kappa_{ef}\delta^3, \\ \|\nabla m(s) - \nabla f(x+s)\| & \leq & \kappa_{eg}\delta^2, \\ \|\nabla^2 m(s) - \nabla^2 f(x+s)\| & \leq & \kappa_{eh}\delta. \end{array}$$

The construction of fully linear/quadratic models based on sampled sets raises a number of geometrical questions. Conn, Scheinberg, and Vicente [9, 10, 12] provided a systematic approach to the subject of deterministic sampling geometry in DFO, establishing error bounds for polynomial interpolation and regression models in terms of a constant measuring the quality or well poisedness of the corresponding sample set (ensuring then the fully linear/quadratic properties). They also showed how to deterministically build or update such sets to ensure that such a constant remains moderate in size. Some numerical studies pioneered by [15] have however shown that trust-region methods can tolerate the use of models updated without strict geometry requirements, although it is also known [25] that convergence cannot be ensured to first-order critical points without appealing to fully linear models when the size of the model gradient becomes small (a procedure known as the criticality step).

A DFO context of expensive function evaluations often makes it unaffordable to construct a deterministic model that is guaranteed to be fully quadratic, as such a process requires $(n+1)(n+2)/2$ function evaluations. Practical approaches rely on considerably fewer points (but at least $n+1$ to preserve fully linearity), and use the remaining degrees of freedom to minimize the norm of the model Hessian or its distance to the previous one. The most studied examples use minimum Frobenius type norms [8, 23], yet in [1] it was proposed to apply the theory of sparse ℓ_1-recovery to build quadratic models based on random sampling. Such models were proved to be fully quadratic with high probability even when considerably less than $(n+1)(n+2)/2$ points were used, depending on the sparsity of the Hessian of the objective.

Such findings have then called for a probabilistic analysis of derivative-free trust-region algorithms [2], where the accuracy of the models is only guaranteed with a certain probability. It was shown in [2] that the resulting trust-region methods converge with probability one to first and second order critical points. The main purpose of this paper is to establish (with overwhelmingly high probability) the rate under which these methods drive to zero the corresponding criticality measures: the norm of the gradient $\|\nabla f(x_k)\|$ (in the first-order case) and the maximum $\sigma(x_k) = \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\}$ between the gradient and the negative of the minimum Hessian eigenvalue (in the second-order case). The proofs rely heavily on the technique developed in [19] for establishing global rates and worst case complexity bounds for randomized algorithms in which the new iterate depends on some object (directions in [19], models here)

and the quality of the object is favorable with a certain probability. The technique is based on counting the number of iterations for which the quality is favorable and examining the probabilistic behavior of this number. Although the road map for our paper was described in [19, Section 6], its actual concretization poses a few delicate issues. In addition, this paper goes beyond [19, Section 6] by establishing results in expectation and addressing the second-order case.

In order to place our work in the existing literature, we briefly review what is known about the complexity of trust-region methods in the deterministic unconstrained case. Using exact first-order Taylor models, trust-region methods are known [20] to take at most $\mathcal{O}(\epsilon^{-2})$ iterations[1] to reduce the measure $\|\nabla f(\cdot)\|$ below $\epsilon \in (0,1)$. This first-order iteration complexity bound can be improved to $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(-\log(\epsilon))$, respectively under convexity and strong convexity assumptions [18]. A recent paper [13] also suggested a modification to the classical trust-region approach to improve the nonconvex case to $\mathcal{O}(\epsilon^{-1.5})$, a bound typical of cubic regularization methods. Using second-order Taylor models, it has been proved [4, 17] that at most $\mathcal{O}\left(\epsilon^{-3}\right)$ iterations are needed to reduce the second order criticality measure $\sigma(\cdot)$ below $\epsilon \in (0,1)$. For trust-region schemes exploiting derivatives, we point out that the above bounds are essentially valid for both iteration and function evaluation counts, as those methods typically compute one additional function value per iteration.

In the derivative-free case, we are particularly interested in counting function evaluations and in understanding the dependence of the complexity bounds in terms of the dimension n of the problem. Using fully linear models, the first-order iteration complexity bound is of the form $\mathcal{O}(\kappa_d^{-2}\epsilon^{-2})$ [16], where $\kappa_d = 1/\max\{\kappa_{ef}, \kappa_{eg}\}$ and κ_{ef}, κ_{eg} are the constants arising in Definition 1.1. Since interpolation techniques can ensure $\kappa_d^{-1} = \mathcal{O}(\sqrt{n})$ with at most $\mathcal{O}(n)$ evaluations per iteration, one recovers the bounds $\mathcal{O}(n\epsilon^{-2})$ for iterations and $\mathcal{O}(n^2\epsilon^{-2})$ for function evaluations, a result also derived for direct search [27]. Using fully quadratic models, the second-order iteration complexity bound is of the form $\mathcal{O}(\kappa_d^{-3}\epsilon^{-3})$ [16, 21], with $\kappa_d = 1/\max\{\kappa_{ef}, \kappa_{eg}, \kappa_{eh}\}$ and the $\kappa_{ef}, \kappa_{eg}, \kappa_{eh}$ constants defined as in Definition 1.2. Fully quadratic models constructed through interpolation satisfy $\kappa_d^{-1} = \mathcal{O}(n)$ and require $\mathcal{O}(n^2)$ evaluations per iteration, leading to a bound in $\mathcal{O}(n^3\epsilon^{-3})$ (resp. $\mathcal{O}(n^5\epsilon^{-3})$) for measuring the number of iterations (resp. function evaluations) needed to drive $\sigma(\cdot)$ below ϵ. We will see that the bounds developed in this paper are comparable to the above mentioned bounds known for deterministic derivative-free trust-region algorithms.

The remainder of the paper is organized as follows. Section 2 presents the results related to first-order stationarity, while Section 3 is concerned with the second-order counterpart. Finally, in Section 4, we comment on the extension of our work to other settings.

2 Complexity of first-order trust-region methods based on probabilistic models

We consider now the scenario analyzed in [2] where the models used in a trust-region method are randomly generated at each iteration. As a result, the iterates and trust-region radii produced by the algorithm will also be random. Upper case letters will be then used to designate random

[1] The notation $\mathcal{O}(A)$ stands for a scalar times A, with the scalar depending solely on the problem considered or constants from the algorithm (with no further dependence on A). The dependence on the problem dimension n will explicitly appear in A when considered appropriate.

variables and lower case their realizations. Hence, m_k, x_k, δ_k will denote respectively the realizations of the random model, iterate, and trust-region radius M_k, X_k, Δ_k at iteration k. The random models are then required to be fully linear with a certain favorable property regardless of the past iteration history. The following definition was proposed in [2] to analyze global convergence of the corresponding trust-region methods to first-order critical points.

Definition 2.1 *A sequence of random models $\{M_k\}$ is said to be (p)–probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear for a corresponding ball sequence $\{B(X_k, \Delta_k)\}$ if the events*

$$S_k \;=\; \{M_k \text{ is a } (\kappa_{eg}, \kappa_{ef})\text{-fully linear model of } f \text{ on } B(X_k, \Delta_k)\}$$

satisfy $\mathbb{P}(S_0) \geq p$ and, for each $k \geq 1$, the following submartingale-like condition

$$\mathbb{P}\left(S_k | M_0, \ldots, M_{k-1}\right) \;\geq\; p.$$

An example is given in [2] where, using random matrix theory, it was shown that linear interpolation based on Gaussian sample sets of cardinality $n+1$ (with a fixed point and the remaining n points being generated randomly from a standard Gaussian distribution) gives rise to fully linear models with a favorable probability (say $p > 1/2$). By using increasingly more sample points and building the models by linear regression it is possible to reach a probability as high as desired ([12, Chapter 4]; see also [22]).

2.1 Algorithm and assumptions

To simplify the presentation, we describe the trust-region methods under consideration (later given in Algorithm 2.1) for each realization of the model randomness. A few components of these methods are classical, with or without derivatives. At each iteration k, one minimizes a quadratic model

$$m_k(x_k + s) \;=\; f(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$$

in a trust region of the form $B(x_k, \delta_k)$. For global convergence to first-order criticality, the Hessian models are assumed to be uniformly bounded and the step s_k is asked to satisfy a fraction of the model decrease given by the negative model gradient within the trust region. These two assumptions are formalized below.

Assumption 2.1 *There exists a positive constant κ_{bhm} such that, for every k, the Hessians H_k of all realizations m_k of M_k satisfy*

$$\|H_k\| \;\leq\; \kappa_{bhm}.$$

Assumption 2.2 *For every k, and for all realizations m_k of M_k (and of X_k and Δ_k), we are able to compute a step s_k so that it satisfies a* fraction of Cauchy decrease, *i.e.,*

$$m(x_k) - m(x_k + s_k) \;\geq\; \frac{\kappa_{fcd}}{2} \|g_k\| \min\left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}, \tag{2.1}$$

for some constant $\kappa_{fcd} \in (0, 1]$, and with the convention that $\frac{\|g_k\|}{\|H_k\|} = \infty$ if $\|H_k\| = 0$.

Finally, the step acceptance and trust-region radius update are based on the ratio between the actual decrease in the objective function and the one predicted by the model, namely

$$\rho_k \;=\; \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

What is different now from the classical derivative-based case is that some form of criticality step has to be taken into account, where models are recomputed in regions small enough compared to the size of the model gradient. Following [2], the presentation and the analysis can be significantly simplified if this requirement is mitigated at each iteration. So, in Algorithm 2.1, the trust-region radius is reduced (at iterations where ρ_k is large enough and the step is taken) provided δ_k is too large compared to $\|g_k\|$.

Algorithm 2.1: A simple first-order derivative-free trust-region framework

Fix parameters $\eta_1, \eta_2, \delta_{\max} > 0$ and $0 < \gamma_1 < 1 < \gamma_2$. Select initial x_0 and $\delta_0 \in (0, \delta_{\max})$.
for $k = 0, 1, \ldots$ **do**
  Build a quadratic model $m_k(x_k + s)$ of f, and compute s_k by approximately minimizing m_k in $B(x_k, \delta_k)$ so that it satisfies (2.1). If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$ and

$$\delta_{k+1} \;=\; \begin{cases} \min\{\gamma_2 \delta_k, \delta_{\max}\} & \text{if } \|g_k\| \geq \eta_2 \delta_k, \\ \gamma_1 \delta_k & \text{otherwise.} \end{cases}$$

  Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \gamma_1 \delta_k$.
end

Note that we have slightly extended the framework in [2] by using two different parameters (namely γ_1 and γ_2) to update the trust-region radius, instead of using a single one and its inverse. As we will see, these parameters are intimately connected to the minimum probability with which the models are required to be fully linear. Also, the safeguard $\delta_{\max}$ is not used in the analysis of the first-order methods and is only there for coherence with the second-order case (where it appears in the analysis) as well as with [2].

The algorithm will be analyzed under the following two assumptions on f. As in [2] it would be enough to assume continuous differentiability in an enlarged initial level set, but we skip this detail for keeping the presentation simple.

Assumption 2.3 *The function f is continuously differentiable on $\mathbb{R}^n$, and its gradient is Lipschitz continuous.*

Assumption 2.4 *The objective function f is bounded from below on $\mathbb{R}^n$, and we denote by f_{low} a lower bound.*

At this point we can state a fundamental result for establishing the complexity of trust-region methods based on probabilistic models. As in the classical setting of trust-region methods, it ensures that the step is taken if the trust-region radius is small enough compared to the size of the true gradient. There are two differences between Lemma 2.1 and [2, Lemma 3.2]: first, the result is stated for the true gradient, in alignment to what is needed to establish complexity bounds; second, it is inferred additionally that the trust-region radius is increased under the same condition.

Lemma 2.1 *If m_k is $(\kappa_{eg}, \kappa_{ef})$-fully linear on $B(x_k, \delta_k)$ and*

$$\delta_k \;<\; \kappa \|\nabla f(x_k)\|, \tag{2.2}$$

where

$$\kappa \;=\; \left(\kappa_{eg} + \max\left\{\eta_2, \kappa_{bhm}, \frac{4\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\}\right)^{-1},$$

then at the k-th iteration the step is taken ($x_{k+1} = x_k + s_k$) and δ_k is increased.

Proof. From (2.2), one has

$$\kappa_{eg}\delta_k + \max\left\{\eta_2, \kappa_{bhm}, \frac{4\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\}\delta_k \;<\; \|\nabla f(x_k)\|,$$

and from this and Definition 1.1,

$$\max\left\{\eta_2, \kappa_{bhm}, \frac{4\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\}\delta_k \;<\; \|\nabla f(x_k)\| - \|g_k - \nabla f(x_k)\| \;\leq\; \|g_k\|.$$

Hence,

$$\delta_k \;<\; \min\left\{\frac{1}{\eta_2}, \frac{1}{\kappa_{bhm}}, \frac{\kappa_{fcd}(1-\eta_1)}{4\kappa_{ef}}\right\}\|g_k\|,$$

and from [12, Proof of Lemma 10.6] we obtain $\rho_k \geq \eta_1$. Since the first term in the minimum gives $\eta_2\delta_k < \|g_k\|$, the trust-region radius is increased. □

2.2 Behavior of the trust-region radius

We will now prove that the sequence of the trust-region radii is square summable and establish an explicit upper bound for the sum. Such a property is not standard in derivative-based trust-region algorithms as the trust-region radius is not expected to go to zero. Yet the summability holds for Algorithm 2.1 because the algorithm does not increase the trust region radius unless it is bounded by $\eta_2^{-1}\|g_k\|$ (and the reduction ratio ρ_k is at least η_1). The proof makes use of the set of indices corresponding to iterations where the trust-region radius is increased, that is

$$\mathcal{K} \;=\; \{k \in \mathbb{N} : \rho_k \geq \eta_1 \text{ and } \|g_k\| \geq \eta_2\delta_k\}. \tag{2.3}$$

Remark 2.1 *In the context of direct search based on probabilistic descent [19, Lemma 4.1], a similar result was proved for the sequence of the step size α_k. There, an iteration attains a decrease of the order of α_k^2 whenever the step is taken, in which case such direct-search algorithms always increase the step size. In the trust-region context, as detailed in the proof below, a decrease of the order of δ_k^2 corresponds only to the iterations where the trust-region radius is increased. There are iterations where the step is taken and the trust-region is decreased, in which case the decrease obtained is not necessarily of this order.*

Lemma 2.2 *For any realization of Algorithm 2.1,*

$$\sum_{k=0}^{\infty}\delta_k^2 \;\leq\; \beta \;:=\; \frac{\gamma_2^2}{1-\gamma_1^2}\left[\frac{\delta_0^2}{\gamma_2^2} + \frac{f_0 - f_{\text{low}}}{\eta}\right],$$

where $f_0 = f(x_0)$ and

$$\eta \;=\; \eta_1\eta_2\frac{\kappa_{fcd}}{2}\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}.$$

Proof. For any $k \in \mathcal{K}$, from (2.1),

$$\begin{aligned} f(x_k) - f(x_k + s_k) &\geq \eta_1 \left(m_k(x_k) - m_k(x_k + s_k)\right) \\ &\geq \eta_1 \frac{\kappa_{fcd}}{2} \eta_2 \min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\} \delta_k^2 \;=\; \eta \delta_k^2. \end{aligned}$$

Consequently, if we sum over all iteration indices in $\mathcal{K}$ less than a given integer k, we obtain

$$\eta \sum_{\substack{j \in \mathcal{K} \\ j \leq k}} \delta_j^2 \;\leq\; \sum_{\substack{j \in \mathcal{K} \\ j \leq k}} [f(x_j) - f(x_{j+1})] \;\leq\; \sum_{j \leq k} [f(x_j) - f(x_{j+1})] = f_0 - f(x_{k+1}) \;\leq\; f_0 - f_{\text{low}},$$

leading to

$$\sum_{k \in \mathcal{K}} \delta_k^2 \;\leq\; \frac{f_0 - f_{\text{low}}}{\eta}.$$

From now on the proof is as in the proof of [19, Lemma 4.1]. Suppose first that $\mathcal{K}$ has infinite cardinality, and let $\mathcal{K} = \{k_1, k_2, \ldots\}$ and, for auxiliary reasons, $k_0 = -1$ and $\delta_{-1} = \delta_0/\gamma_2$. The sum $\sum_{k=0}^{\infty} \delta_k^2$ can thus be rewritten as

$$\sum_{k=0}^{\infty} \delta_k^2 \;=\; \sum_{i=0}^{\infty} \sum_{k=k_i+1}^{k_{i+1}} \delta_k^2.$$

Besides, one has for each index i, $\delta_k \leq \gamma_2 (\gamma_1)^{k-k_i-1} \delta_{k_i}$ for $k = k_i + 1, \ldots, k_{i+1}$. Hence,

$$\sum_{k=k_i+1}^{k_{i+1}} \delta_k^2 \;\leq\; \frac{\gamma_2^2}{1-\gamma_1^2} \delta_{k_i}^2,$$

and we finally obtain the desired result:

$$\sum_{k=0}^{\infty} \delta_k^2 \;\leq\; \frac{\gamma_2^2}{1-\gamma_1^2} \sum_{i=0}^{\infty} \delta_{k_i}^2 = \frac{\gamma_2^2}{1-\gamma_1^2} \left[\delta_{-1}^2 + \sum_{k \in \mathcal{K}} \delta_k^2\right] \;\leq\; \frac{\gamma_2^2}{1-\gamma_1^2} \left[\frac{\delta_0^2}{\gamma_2^2} + \frac{f_0 - f_{\text{low}}}{\eta}\right].$$

In the case of a finite set $\mathcal{K} = \{k_1, k_2, \ldots, k_M\}$, defining $k_{M+1} = \infty$ allows to reproduce the same reasoning as above, leading to an identical bound. □

This is a stronger version of the result $\lim_{k \to \infty} \delta_k = 0$ proved in [2] for the purpose of global convergence.

2.3 Number of iterations with fully linear models

Recall the set of indices (2.3) corresponding to iterations where the step is taken and the trust-region radius is increased, and let y_k denote the corresponding indicator ($y_k = 1$ if $k \in \mathcal{K}$, $y_k = 0$ otherwise). Such choices will allow us to exploit the analysis of direct search based on probabilistic descent [19], with δ_k and $\mathcal{K}$ respectively playing the roles of the step size and the set of successful iterations in [19].

Let now $\nabla f(\tilde{x}_k)$, with $\tilde{x}_k \in \{x_0, \ldots, x_k\}$, represent a minimum norm gradient among $\nabla f(x_0), \ldots, \nabla f(x_k)$. In addition, we define

$$p_0 \;=\; \frac{\ln(\gamma_1)}{\ln(\gamma_1/\gamma_2)}. \tag{2.4}$$

The quantity $p_0 \in (0,1)$ will emerge naturally in our upcoming analysis, as we will need the probability p in Definition 2.1 to be higher than p_0 in order to render the complexity results. In practice, how to fulfill this requirement depends upon the model generation scenario. If the probability under which these models are sufficiently accurate can be adjusted in advance, then for any particular choices of γ_1 and γ_2 ($0 < \gamma_1 < 1 < \gamma_2$), we can adjust the probability so that it is higher than p_0. There may also exist scenarios where the model generation has a given non adjustable probability of being sufficiently accurate, in which case one can then set γ_1, γ_2 properly so that the quantity p_0 is smaller than this probability.

The next step in the analysis is to show, similarly to [19, Lemma 4.2], that the number of iterations that benefited from a fully linear model depends on the inverse of the square of $\|\nabla f(\tilde{x}_k)\|$ thus indicating that such a number is not necessarily low when the gradient is large[2]. The binary variable z_k below indicates whether m_k is $(\kappa_{eg}, \kappa_{ef})$-fully linear or not.

Lemma 2.3 *Given a realization of Algorithm 2.1 and a positive integer k,*

$$\sum_{l=0}^{k-1} z_l \;\leq\; \frac{\beta}{(\min\{\delta_0/\gamma_2, \kappa\|\nabla f(\tilde{x}_k)\|\})^2} + p_0 k.$$

Proof. For each $l \in \{0, 1, \ldots, k-1\}$, define

$$v_l \;=\; \begin{cases} 1 & \text{if } \delta_l \;<\; \min\{\gamma_2^{-1}\delta_0, \kappa\|\nabla f(\tilde{x}_k)\|\}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The proof relies then on $z_l \leq (1 - v_l) + v_l y_l$, which is true because, when $v_l = 1$, Lemma 2.1 implies that $y_l \geq z_l$ (since $\|\nabla f(\tilde{x}_k)\| \leq \|\nabla f(\tilde{x}_{k-1})\| \leq \|\nabla f(\tilde{x}_l)\|$). It suffices then to separately prove

$$\sum_{l=0}^{k-1} (1 - v_l) \;\leq\; \frac{\beta}{(\min\{\delta_0/\gamma_2, \kappa\|\nabla f(\tilde{x}_k)\|\})^2} \tag{2.6}$$

and

$$\sum_{l=0}^{k-1} v_l y_l \;\leq\; p_0 k. \tag{2.7}$$

Inequality (2.6) is justified by Lemma 2.2 and the fact that (2.5) implies

$$1 - v_l \;\leq\; \frac{\delta_l^2}{(\min\{\delta_0/\gamma_2, \kappa\|\nabla f(\tilde{x}_k)\|\})^2}.$$

The proof of inequality (2.7) is verbatim as in the proof of [19, Lemma 4.2], thanks to the definition of p_0 given by (2.4). □

[2] As opposed to Lemma 2.2 this result follows more directly from the theory in [19], given that the discrepancy mentioned in Remark 2.1 does not need special treatment.

2.4 Worst case complexity and global rate

We will now consider events belonging to the σ-algebra generated by the random model sequence $\{M_k\}$. The probabilities will then be expressed with respect to the corresponding probability space (see [24, Appendix A] for more details).

From Lemma 2.3, one then has the following inclusion of events

$$\left\{\|\nabla f(\tilde{X}_k)\| \geq \epsilon\right\} \subset \left\{\sum_{l=0}^{k-1} Z_l \leq \frac{\beta}{\kappa^2\epsilon^2} + p_0 k\right\}, \tag{2.8}$$

for any ϵ satisfying

$$0 < \epsilon < \frac{\delta_0}{\kappa\gamma_2}. \tag{2.9}$$

On the other hand the probabilistic behavior of the event on the right-hand side is known from the application of the Chernoff bound to the lower tail of $\sum_{l=0}^{k-1} Z_l$ (see, e.g., [19, Lemma 4.4]). Note that the conditioning on the past from Assumption 2.1 appears in the proof of the result below.

Lemma 2.4 *Suppose that $\{M_k\}$ is (p)–probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear and $\lambda \in (0, p)$. Then*

$$\pi_k(\lambda) := \mathbb{P}\left(\sum_{l=0}^{k-1} Z_l \leq \lambda k\right) \leq \exp\left[-\frac{(p-\lambda)^2}{2p}k\right].$$

Thus, when (2.9) holds and

$$k \geq \frac{2\beta}{(p-p_0)\kappa^2\epsilon^2}, \tag{2.10}$$

the inclusion (2.8) and the monotonicity of $\pi_k(\cdot)$ will give us (setting $\lambda = (p+p_0)/2$ in Lemma 2.4)

$$\begin{aligned}
\mathbb{P}\left(\|\nabla f(\tilde{X}_k)\| \leq \epsilon\right) \geq \mathbb{P}\left(\|\nabla f(\tilde{X}_k)\| < \epsilon\right) &\geq 1 - \mathbb{P}\left(\sum_{l=0}^{k-1} Z_l \leq \frac{\beta}{\kappa^2\epsilon^2} + p_0 k\right) \\
&\geq 1 - \pi_k\left(\frac{\beta}{k\kappa^2\epsilon^2} + p_0\right) \\
&\geq 1 - \pi_k\left(\frac{p-p_0}{2} + p_0\right) \\
&\geq 1 - \exp\left[-\frac{(p-p_0)^2}{8p}k\right],
\end{aligned} \tag{2.11}$$

leading to the following global rate result.

Theorem 2.1 *Suppose that $\{M_k\}$ is (p)–probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear with $p > p_0$ and*

$$k \geq \frac{2\beta\gamma_2^2}{(p-p_0)\delta_0^2}. \tag{2.12}$$

Then, the minimum gradient norm $\|\nabla f(\tilde{X}_k)\|$ satisfies

$$\mathbb{P}\left(\|\nabla f(\tilde{X}_k)\| \leq \frac{\sqrt{2}\beta^{\frac{1}{2}}}{\kappa(p-p_0)^{\frac{1}{2}}}\frac{1}{\sqrt{k}}\right) \geq 1 - \exp\left[-\frac{(p-p_0)^2}{8p}k\right].$$

Proof. Let

$$\epsilon \;=\; \frac{\sqrt{2}\beta^{\frac{1}{2}}}{\kappa(p-p_0)^{\frac{1}{2}}}\frac{1}{\sqrt{k}}.$$

Then (2.10) holds with equality, and $\epsilon < \delta_0/(\kappa\gamma_2)$ is guaranteed by (2.12). Hence (2.11) gives us the bound. □

We have thus proved a global rate of $1/\sqrt{k}$ for the norm of the gradient with overwhelmingly high probability.

Similarly, one can prove a worst-case bound of the order of $\mathcal{O}(\epsilon^{-2})$ for the first iteration index K_ϵ for which $\|\nabla f(\tilde{X}_{K_\epsilon})\| \leq \epsilon$, also with overwhelmingly high probability (and we note that K_ϵ is a random variable due to the randomness of the models). The proof relies on again applying (2.11), on the observation that $\mathbb{P}(K_\epsilon \leq k) = \mathbb{P}(\|\nabla f(\tilde{X}_k)\| \leq \epsilon)$, and on taking k as

$$k \;=\; \left\lceil \frac{2\beta}{(p-p_0)\kappa^2\epsilon^2} \right\rceil.$$

Theorem 2.2 *Suppose that $\{M_k\}$ is (p)–probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear with $p > p_0$ and that ϵ satisfies (2.9). Then, K_ϵ satisfies*

$$\mathbb{P}\left(K_\epsilon \;\leq\; \left\lceil \frac{2\beta}{(p-p_0)\kappa^2\epsilon^2} \right\rceil\right) \;\geq\; 1 - \exp\left[-\frac{\beta(p-p_0)}{4p\kappa^2\epsilon^2}\right].$$

Results in expectation are a natural byproduct of our analysis. It can be shown that $\mathbb{E}(\|\nabla f(\tilde{X}_k)\|)$ is bounded by a function of the order of $k^{-\frac{1}{2}} + \exp(-k)$ up to multiplicative constants (see [19, Proposition 5.2]). It is also possible to bound the expected value of K_ϵ [24].

Theorem 2.3 *Suppose that $\{M_k\}$ is (p)–probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear with $p > p_0$ and that ϵ satisfies (2.9). Then,*

$$\mathbb{E}\,(K_\epsilon) \;\leq\; c_1\epsilon^{-2} + \frac{1}{1-\exp(-c_2)},$$

where

$$c_1 \;=\; \frac{2\beta}{(p-p_0)\kappa^2}, \quad c_2 \;=\; \frac{(p-p_0)^2}{8p}.$$

Proof. Since K_ϵ only takes non-negative integer values, the expected value of K_ϵ satisfies:

$$\mathbb{E}\,(K_\epsilon) = \sum_{j=0}^{\infty} j\mathbb{P}\,(K_\epsilon = j) = \sum_{j\geq 0}\sum_{\substack{k\in\mathbb{N}\\ k<j}} \mathbb{P}\,(K_\epsilon = j) = \sum_{k\in\mathbb{N}}\sum_{\substack{j\geq 0\\ k<j}} \mathbb{P}\,(K_\epsilon = j) = \sum_{k\in\mathbb{N}} \mathbb{P}\,(K_\epsilon > k).$$

Hence,

$$\begin{aligned}
\mathbb{E}\,(K_\epsilon) \;&=\; \sum_{\substack{0\leq k<c_1\epsilon^{-2}\\ k\in\mathbb{N}}} \mathbb{P}\,(K_\epsilon > k) + \sum_{\substack{k\geq c_1\epsilon^{-2}\\ k\in\mathbb{N}}} \mathbb{P}\,(K_\epsilon > k)\\
&\leq\; c_1\epsilon^{-2} + 1 + \sum_{\substack{k\geq c_1\epsilon^{-2}\\ k\in\mathbb{N}}} \mathbb{P}\,(K_\epsilon > k)\\
&=\; c_1\epsilon^{-2} + 1 + \sum_{\substack{k\geq c_1\epsilon^{-2}\\ k\in\mathbb{N}}} \mathbb{P}\left(\|\nabla f(\tilde{X}_k)\| > \epsilon\right).
\end{aligned}$$

For any index $k \geq c_1\epsilon^{-2}$, similar to (2.11), we have

$$\mathbb{P}\left(\|\nabla f(\tilde{X}_k)\| > \epsilon\right) \;\leq\; \exp(-c_2 k).$$

As a result,

$$\mathbb{E}\left(K_\epsilon\right) \;\leq\; c_1\epsilon^{-2} + 1 + \sum_{\substack{k \geq c_1\epsilon^{-2} \\ k\in\mathbb{N}}} \exp(-c_2 k) \;\leq\; c_1\epsilon^{-2} + \sum_{k\in\mathbb{N}} \exp(-c_2 k) \;\leq\; c_1\epsilon^{-2} + \frac{1}{1-\exp(-c_2)},$$

which proves the desired result. □

The obtained bound is thus

$$\mathcal{O}\left(\frac{\kappa^{-2}\epsilon^{-2}}{p-p_0}\right) + \mathcal{O}(1),$$

which matches the results of [5] for line-search methods based on probabilistic models (where p_0 is taken as $1/2$).

More generally, one observes that for both Theorems 2.2 and 2.3 the dominant term in iteration effort is

$$\mathcal{O}\left((p-p_0)^{-1}\kappa^{-2}\epsilon^{-2}\right) \tag{2.13}$$

where $\kappa = \mathcal{O}\left(1/\max\{\kappa_{ef}, \kappa_{eg}\}\right)$ according to Lemma 2.1. As mentioned in the introductory section, one can ensure $\max\{\kappa_{ef}, \kappa_{eg}\} = \mathcal{O}(\sqrt{n})$ in a deterministic setting with $\mathcal{O}(n)$ points. Hence in a simple scenario where a random model is selected among a finite number of fixed fully linear models, the models are then almost always $(\kappa_{eg}, \kappa_{ef})$-fully linear, namely (p)-probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear with $p = 1$, and one recovers from (2.13) the same iteration and function evaluation complexity bounds mentioned in the introduction for the deterministic setting, up to a constant factor $1 - p_0$ that only involves algorithmic parameters. In a more general setting of probabilistic models, and when building linear models by randomly generating n sampling points, one can obtain (p)-probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear models with $\max\{\kappa_{ef}, \kappa_{eg}\} = \mathcal{O}(\sqrt{n})$ while $p > p_0$ independent of n (see [2, Section 6.2, "Analysis of poisedness of random sample sets"]). Hence from (2.13) one again recovers the same iteration and function evaluation complexity bounds known for the deterministic setting.

2.5 A note on global convergence

Our complexity theory implies (see [19, Proposition 5.1])

$$\mathbb{P}\left(\inf_{k\geq 0} \|\nabla f(X_k)\| = 0\right) \;=\; 1.$$

If we assume for all realizations of Algorithm 2.1 that the iterates never arrive at a stationary point in a finite number of iterations, then the events $\{\liminf_{k\to\infty} \|\nabla f(X_k)\| = 0\}$ and $\{\inf_{k\geq 0} \|\nabla f(X_k)\| = 0\}$ are identical and we recover the liminf result in probability one of [2, Theorem 4.2]. Note also that such a result could be derived even more directly by using the argument of [19, Lemma 3.2 and Theorem 3.1]. The paper [2] goes further into the analysis and also establishes a lim-type result.

3 Complexity of second-order trust-region methods based on probabilistic models

The proof technology of Section 2 enables us to derive a similar complexity study for the class of trust-region algorithms under consideration but now with the goal of approaching or converging to second-order critical points. To do so, additional assumptions need to be enforced regarding both the quality of the models and the properties of the step resulting from the approximate solution of the trust-region subproblem. Moreover, the objective function will be required to satisfy the following assumption.

Assumption 3.1 *The function f is twice continuously differentiable on $\mathbb{R}^n$, and its Hessian is Lipschitz continuous.*

The probabilistic counterpart to the concept of fully quadratic models is given by Definition 1.2.

Definition 3.1 *A sequence of random models $\{M_k\}$ is said to be (p)–probabilistically $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic for a corresponding ball sequence $\{B(X_k, \Delta_k)\}$ if the events*

$$S_k \;=\; \{M_k \text{ is a } (\kappa_{eh}, \kappa_{eg}, \kappa_{ef})\text{-fully quadratic model of } f \text{ on } B(X_k, \Delta_k)\}$$

satisfy $\mathbb{P}(S_0) \geq p$ and, for each $k \geq 1$, the following submartingale-like condition

$$\mathbb{P}\left(S_k | M_0, \ldots, M_{k-1}\right) \;\geq\; p.$$

The authors of [1] show how to build fully quadratic models with high probability from quadratic interpolation and uniformly generated sample sets. It was also proved there that such a procedure may recover such models with considerably less than $(n+1)(n+2)/2$ function evaluations when the Hessian of the function is sparse.

3.1 Algorithm and assumptions

As before, we consider quadratic models around the iterate x_k, with the same definitions for m_k, g_k, and H_k (and the imposition of Assumption 2.1). As curvature is now present in our analysis, we will make use of the notation $\tau_k = \lambda_{\min}(H_k)$. The solution of the trust-region subproblem has now to be second-order accurate, in the sense of Assumption 3.2.

Assumption 3.2 *For every k, and for all realizations m_k of M_k (and of X_k and Δ_k), we are able to compute a step s_k so that it satisfies both a* fraction of Cauchy decrease *and a* fraction of eigendecrease*, i.e.,*

$$m(x_k) - m(x_k + s_k) \;\geq\; \frac{\kappa_{fod}}{2} \max\left\{ \|g_k\| \min\left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}, -\tau_k \delta_k^2 \right\}. \tag{3.1}$$

for some constant $\kappa_{fod} \in (0, 1]$.

The first part of (3.1) can be satisfied by a Cauchy step, while considering a step of norm δ_k along an eigenvector of the model Hessian associated with the eigenvalue τ_k yields the second-order decrease in δ_k^2. A step satisfying (3.1) can be obtained by taking the one corresponding to the largest decrease caused in the model value.

Such considerations lead us from Algorithm 2.1 to Algorithm 3.1, preserving the overall structure of the method (and, in particular, the updating rules for the trust-region radius). As mentioned in the Introduction we make use of the second-order criticality measure

$$\sigma(x) \;=\; \max\left\{\|\nabla f(x)\|, -\lambda_{\min}(\nabla^2 f(x))\right\},$$

for which a natural estimator at x_k is

$$\sigma^m(x_k) \;:=\; \max\left\{\|g_k\|, -\tau_k\right\}.$$

In the case of a fully quadratic model, the two quantities are related as follows.

Lemma 3.1 *[12, Lemma 10.15] If m_k is $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic on $B(x_k, \delta_k)$, then*

$$|\sigma(x_k) - \sigma^m(x_k)| \;\leq\; \kappa_\sigma \delta_k, \tag{3.2}$$

where $\kappa_\sigma = \max\{\kappa_{eg}\delta_{\max}, \kappa_{eh}\}$.

Algorithm 3.1: A simple second-order derivative-free trust-region framework

---

Fix parameters $\eta_1, \eta_2, \delta_{\max} > 0$ and $0 < \gamma_1 < 1 < \gamma_2$. Select initial x_0 and $\delta_0 \in (0, \delta_{\max})$.
for $k = 0, 1, 2, \ldots$ **do**
  Build a quadratic model $m_k(x_k + s)$ of f, and compute s_k by approximately minimizing m_k in $B(x_k, \delta_k)$ so that it satisfies (3.1). If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$ and

$$\delta_{k+1} \;=\; \begin{cases} \min\{\gamma_2\delta_k, \delta_{\max}\} & \text{if } \sigma^m(x_k) \geq \eta_2\delta_k, \\ \gamma_1\delta_k & \text{otherwise.} \end{cases}$$

  Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \gamma_1\delta_k$.
end

Lemma 3.2 *If m_k is $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic on $B(x_k, \delta_k)$ and*

$$\delta_k \;<\; \kappa_{2nd}\sigma(x_k),$$

where

$$\kappa_{2nd} \;=\; \left(\kappa_\sigma + \max\left\{\eta_2, \kappa_{bhm}, \frac{4\kappa_{ef}\delta_{\max}}{\kappa_{fod}(1-\eta_1)}, \frac{4\kappa_{ef}}{\kappa_{fod}(1-\eta_1)}\right\}\right)^{-1},$$

then at the k-th iteration the step is taken ($x_{k+1} = x_k + s_k$) and δ_k is increased.

Proof. The proof is similar to the one of Lemma 2.1. Combining (3.2) with the error bound of Lemma 3.1 yields

$$\delta_k \;<\; \min\left\{\frac{1}{\eta_2}, \frac{1}{\kappa_{bhm}}, \frac{\kappa_{fod}(1-\eta_1)}{4\kappa_{ef}}, \frac{\kappa_{fod}(1-\eta_1)}{4\kappa_{ef}\delta_{\max}}\right\}\sigma^m(x_k).$$

This allows to directly conclude that $\rho_k \geq \eta_1$ (see [12, Lemma 10.17]). Besides, since the first term in the minimum gives $\eta_2\delta_k \leq \sigma^m(x_k)$, the trust-region radius is increased. □

3.2 Behavior of the trust-region radius

Given the decrease properties now enforced by the algorithm, similar to Lemma 2.2, we can prove that the sequence of the trust-region radii is cube summable. Let $\mathcal{K}_{2nd}$ be the set of indices corresponding to iterations where the step is taken and the trust-region radius is increased, i.e.,

$$\mathcal{K}_{2nd} \;=\; \{k \in \mathbb{N} : \rho_k \geq \eta_1 \text{ and } \sigma^m(x_k) \geq \eta_2 \delta_k\}\,.$$

Lemma 3.3 *For any realization of Algorithm 3.1,*

$$\sum_{k=0}^{\infty} \delta_k^3 \;\leq\; \beta_{2nd} \;:=\; \frac{\gamma_2^3}{1-\gamma_1^3}\left[\frac{\delta_0^3}{\gamma_2^3} + \frac{f_0 - f_{\text{low}}}{\eta_{2nd}}\right],$$

where

$$\eta_{2nd} \;=\; \eta_1\eta_2\frac{\kappa_{fod}}{2}\min\left\{\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\frac{1}{\delta_{\max}}, 1\right\}.$$

Proof. Similar to the proof of Lemma 2.2, we only need to consider the case where $|\mathcal{K}_{2nd}| = \infty$. For any $k \in \mathcal{K}_{2nd}$, if $\sigma^m(x_k) = \|g_k\|$, we have by Assumption 3.2 that

$$\begin{aligned}
f(x_k) - f(x_{k+1}) &\geq \eta_1\left[m_k(x_k) - m_k(x_k + s_k)\right] \\
&= \eta_1\frac{\kappa_{fod}}{2}\|g_k\|\min\left\{\frac{\|g_k\|}{\|H_k\|}, \delta_k\right\} \\
&\geq \eta_1\eta_2\frac{\kappa_{fod}}{2}\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\delta_k^2 \\
&\geq \eta_1\eta_2\frac{\kappa_{fod}}{2}\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\frac{1}{\delta_{\max}}\delta_k^3.
\end{aligned}$$

Meanwhile, if $\sigma^m(x_k) = -\tau_k$, a similar reasoning leads to:

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1\eta_2\frac{\kappa_{fod}}{2}\delta_k^3.$$

As a result, for any index $k \in \mathcal{K}_{2nd}$, one has

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1\eta_2\frac{\kappa_{fod}}{2}\min\left\{\min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}\frac{1}{\delta_{\max}}, 1\right\}\delta_k^3 \;=\; \eta_{2nd}\delta_k^3.$$

The rest of the proof follows the lines of Lemma 2.2 replacing the squares of the trust-region radii by cubes. □

3.3 Number of iterations with fully quadratic models

An upper bound on the number of iterations where the models are fully quadratic is derived similarly as in the first-order case. The main changes consist in defining $\tilde{x}_k \in \{x_0, \ldots, x_k\}$ such that $\sigma(\tilde{x}_k)$ is a minimum value among $\{\sigma(x_0), \ldots, \sigma(x_k)\}$ and z_k as the binary variable indicating if m_k is $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic or not.

Lemma 3.4 *Given a realization of Algorithm 3.1 and a positive integer k,*

$$\sum_{l=0}^{k-1} z_l \;\leq\; \frac{\beta_{2nd}}{(\min\{\delta_0/\gamma_2, \kappa_{2nd}\sigma(\tilde{x}_k)\})^3} + p_0 k.$$

3.4 Worst case complexity and global rate

The derivation of the second-order complexity theory is based on observing that Lemma 3.4 now implies

$$\left\{\|\sigma(\tilde{X}_k)\| \geq \epsilon\right\} \subset \left\{\sum_{l=0}^{k-1} Z_l \leq \frac{\beta_{2nd}}{\kappa_{2nd}^3 \epsilon^3} + p_0 k\right\},$$

for any ϵ satisfying

$$0 < \epsilon < \frac{\delta_0}{\kappa_{2nd}\gamma_2}. \tag{3.3}$$

Then a result identical to Lemma 2.4 can be ensured considering the definition of Z_l based on fully quadratic models and replacing the probabilistic fully linear assumption by the probabilistic fully quadratic one. The rest of the analysis proceeds similarly with minor changes in constants. The global rate is now $1/\sqrt[3]{k}$, as shown below.

Theorem 3.1 *Suppose that $\{M_k\}$ is (p)–probabilistically $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic with $p > p_0$ and*

$$k \geq \frac{2\beta_{2nd}\gamma_2^3}{(p-p_0)\delta_0^3}.$$

Then, the minimum second-order criticality measure $\sigma(\tilde{X}_k)$ satisfies

$$\mathbb{P}\left(\sigma(\tilde{X}_k) \leq \frac{\sqrt[3]{2}\beta_{2nd}^{\frac{1}{3}}}{\kappa_{2nd}(p-p_0)^{\frac{1}{3}}} \frac{1}{\sqrt[3]{k}}\right) \geq 1 - \exp\left[-\frac{(p-p_0)^2}{8p}k\right].$$

Similar conclusions as those of Subsection 2.5 can be drawn here regarding global convergence results of the liminf type that can be deduced from the above theorem, and regarding their interplay with the second-order convergence theory of [2]. However, a significant difference with the first-order case is the absence of a rigorous proof for deriving a lim-type result [2].

As in the first-order case, we can establish also a worst-case complexity bound with overwhelming high probability, but now of the order of ϵ^{-3}.

Theorem 3.2 *Suppose that $\{M_k\}$ is (p)–probabilistically $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic with $p > p_0$ and that ϵ satisfies (3.3). Let K_ϵ be the first iteration index for which $\sigma(\tilde{X}_{K_\epsilon}) \leq \epsilon$. Then, K_ϵ satisfies*

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{2\beta_{2nd}}{(p-p_0)\kappa_{2nd}^3\epsilon^3} \right\rceil\right) \geq 1 - \exp\left[-\frac{\beta_{2nd}(p-p_0)}{4p\kappa_{2nd}^3\epsilon^3}\right].$$

As in Theorem 2.3, we can also bound the expected number of iterations needed to reach the desired accuracy.

Theorem 3.3 *Suppose that $\{M_k\}$ is (p)–probabilistically $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic with $p > p_0$ and that ϵ satisfies (3.3). Then,*

$$\mathbb{E}(K_\epsilon) \leq c_3\epsilon^{-3} + \frac{1}{1-\exp(-c_2)},$$

where

$$c_3 = \frac{2\beta_{2nd}}{(p-p_0)\kappa_{2nd}^3},$$

and c_2 is defined as in Theorem 2.3.

The iteration complexity bound is thus essentially $\mathcal{O}\left((p-p_0)^{-1}\kappa_{2nd}^{-3}\epsilon^{-3}\right)$, where we observe from Lemma 3.2 that $\kappa_{2nd} = \mathcal{O}\left(1/\max\{\kappa_{ef}, \kappa_{eg}, \kappa_{eh}\}\right)$. Similarly to the first-order case, we can recover the deterministic bounds for iteration and function evaluation complexity mentioned in the introduction when the models are almost always $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic. Such models are (p)-probabilistically $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic with $p = 1$ and $\max\{\kappa_{ef}, \kappa_{eh}, \kappa_{eg}\} = \mathcal{O}(n)$, and can be built selecting among a finite number of fixed fully quadratic models using $\mathcal{O}(n^2)$ points.

The implication of Theorems 3.2 and 3.3 on the function evaluation complexity becomes more interesting in a broader probabilistic setting. It was proved in [1] that $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic models with $\max\{\kappa_{eh}, \kappa_{eg}, \kappa_{ef}\} = \mathcal{O}(n)$ can be obtained with high probability (yet less than 1) using fewer than $\mathcal{O}(n^2)$ sample points if the objective function has a sparse Hessian. For instance, when the Hessian is tridiagonal, only $\mathcal{O}(n(\log^4 n))$ points are needed, and p can be chosen so that $p - p_0$ is positive and independent of n. Hence Theorems 3.2 and 3.3 lead to a function evaluation complexity bound of

$$\mathcal{O}\big(n(\log^4 n)\max\{\kappa_{ef}^3, \kappa_{eg}^3, \kappa_{eh}^3\}\epsilon^{-3}\big) \;=\; \mathcal{O}\big((n\log n)^4\epsilon^{-3}\big)$$

which compares favorably in terms of n with the bound mentioned in the introduction for the deterministic case.

4 Remarks on open questions

Recently a number of papers have appeared proposing and analyzing derivative-free trust-region methods for the unconstrained optimization of a stochastic function. In this setting, what is observable is $\tilde{f}(x, \varepsilon(\omega))$, where ε is a random variable. The objective function $f(x)$ may be given by $\mathbb{E}(\tilde{f}(x, \varepsilon))$. One approach [26] extended the framework [11] using Sample-Average Approximation (SAA). The number of observations in each Monte Carlo oracle may be up to $\mathcal{O}(\delta_k^{-4})$. First-order global convergence was proved with probability one but for algorithmic parameters that depend on unknown problem constants. Another approach [6] extended trust-region methods based on probabilistic models [2] to cover also probabilistic estimates of the objective function. In the non-biased case with $f(x) = \mathbb{E}(\tilde{f}(x, \varepsilon))$, the probabilistic assumptions can be ensured by SAA within $\mathcal{O}(\delta_k^{-4})$ observations. This approach can also handle biased cases like failures in function evaluations or even processor failures (thus accommodating gradient failures when using finite differences). First-order global convergence was also proved with probability one but again for algorithmic parameters that depend on unknown problem constants. A similar approach [22] led to first-order global convergence in probability (weaker than with probability one), but under more practical assumptions. Very recently, a paper [3] developed a complexity analysis for the approach in [6] showing a complexity bound of $\mathcal{O}(\epsilon^{-2})$ on the expected number of iterations needed to satisfy $\|\nabla f(X_k)\| \leq \epsilon$ (again under unverifiable assumptions on algorithmic parameters).

It is an open question whether our proof technology can improve upon these stochastic optimization approaches in the sense of establishing global convergence with probability one and global rates and complexity bounds with overwhelmingly high probability without unverifiable assumptions on algorithmic parameters. Another challenging prospect for future work is to develop better rates and bounds for the convex and strongly convex cases for either deterministic or stochastic functions.

References

[1] A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization. *Math. Program.*, 134:223–257, 2012.

[2] 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.

[3] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method for nonconvex optimization. arXiv:1609.07428v1, 2016.

[4] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *J. Complexity*, 28:93–108, 2012.

[5] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Math. Program.*, 2017 (to appear).

[6] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Math. Program.*, 2017 (to appear).

[7] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2000.

[8] A. R. Conn, K. Scheinberg, and Ph. L. Toint. A derivative free optimization algorithm in practice. In *Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, St Louis, Missouri, USA, September 1998. Multidisciplinary Analysis Optimization Conferences.

[9] A. R. Conn, K. Scheinberg, and L. N. Vicente. Geometry of interpolation sets in derivative-free optimization. *Math. Program.*, 111:141–172, 2008.

[10] A. R. Conn, K. Scheinberg, and L. N. Vicente. Geometry of sample sets in derivative-free optimization: Polynomial regression and underdetermined interpolation. *IMA J. Numer. Anal.*, 28:721–748, 2008.

[11] A. R. Conn, K. Scheinberg, and L. N. Vicente. Global convergence of general derivative-free trust-region algorithms to first and second order critical points. *SIAM J. Optim.*, 20:387–415, 2009.

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

[13] F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $O\left(\epsilon^{-3/2}\right)$ for nonconvex optimization. *Math. Program.*, 162:1–32, 2017.

[14] A. L. Custódio, K. Scheinberg, and L. N. Vicente. Methodologies and software for derivative-free optimization. In *Chapter 37 of Advances and Trends in Optimization with Engineering Applications, MOS-SIAM Book Series on Optimization*. SIAM, Philadelphia, 2017.

[15] G. Fasano, J. L. Morales, and J. Nocedal. On the geometry phase in model-based algorithms for derivative-free optimization. *Optim. Methods Softw.*, 24:145–154, 2009.

[16] R. Garmanjani, D. Júdice, and L. N. Vicente. Trust-region methods without using derivatives: Worst case complexity and the non-smooth case. *SIAM J. Optim.*, 26:1987–2011, 2016.

[17] G. N. Grapiglia, J. Yuan, and Y. Yuan. Nonlinear stepsize control algorithms: Complexity bounds for first- and second-order optimality. *J. Optim. Theory Appl.*, 171:980–997, 2016.

[18] G. N. Grapiglia, J. Yuan, and Y. Yuan. On the worst-case complexity of nonlinear stepsize control algorithms for convex unconstrained optimization. *Optim. Methods Softw.*, 31:591–604, 2016.

[19] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Direct search based on probabilistic descent. *SIAM J. Optim.*, 25:1515–1541, 2015.

[20] S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM J. Optim.*, 19:414–444, 2008.

[21] D. Júdice. *Trust-Region Methods without using Derivatives: Worst Case Complexity and the Non-smooth Case*. PhD thesis, Dept. Mathematics, Univ. Coimbra, 2015.

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

[23] M. J. D. Powell. Least Frobenius norm updating of quadratic models that satisfy interpolation conditions. *Math. Program.*, 100:183–215, 2004.

[24] C. W. Royer. *Derivative-free Optimization Methods based on Probabilistic and Deterministic Properties: Complexity Analysis and Numerical Relevance*. PhD thesis, Université de Toulouse, November 2016.

[25] K. Scheinberg and Ph. L. Toint. Self-correcting geometry in model-based algorithms for derivative-free unconstrained optimization. *SIAM J. Optim.*, 20:3512–3532, 2010.

[26] S. Shashaani, F. Hashemi, and R. Pasupathy. ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. arXiv:1610.06506v1, 2016.

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

[28] Y. Yuan. Recent avances in trust region algorithms. *Math. Program.*, 151:249–281, 2015.