Derivative-free separable quadratic modeling and cubic regularization for unconstrained optimization

A. L. Custódio* R. Garmanjani † M. Raydan ‡

July 16, 2021

Abstract

We present a derivative-free separable quadratic modeling and cubic regularization technique for solving smooth unconstrained minimization problems. The derivative-free approach is mainly concerned with building a quadratic model that could be generated by numerical interpolation or using a minimum Frobenious norm approach, when the number of points available does not allow to build a complete quadratic model. This model plays a key role to generate an approximated gradient vector and Hessian matrix of the objective function at every iteration. We add a specialized cubic regularization strategy to minimize the quadratic model at each iteration, that makes use of separability. We discuss convergence results, including worst case complexity, of the proposed schemes to first-order stationary points. Some preliminary numerical results are presented to illustrate the robustness of the specialized separable cubic algorithm.

AMS Subject Classification: 90C30, 65K05, 90C56, 65D05.

Keywords: Derivative-free optimization, fully-linear models, fully-quadratic models, cubic regularization, worst-case complexity.

1 Introduction

We consider unconstrained minimization problems of the form

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

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable in $\mathbb{R}^n$. However, we assume that the derivatives of f are not available and that cannot be easily approximated by

*Department of Mathematics, FCT-UNL-CMA, Campus de Caparica, 2829-516 Caparica, Portugal (alcustodio@fct.unl.pt). Support for this author was provided by national funds through FCT – Fundação para a Ciência e a Tecnologia I. P., under the scope of projects PTDC/MAT-APL/28400/2017 and UIDB/MAT/00297/2020.

†Centro de Matemática e Aplicações (CMA), FCT, UNL, 2829-516 Caparica, Portugal (r.garmanjani@fct.unl.pt). Support for this author was provided by national funds through FCT – Fundação para a Ciência e a Tecnologia I. P., under the scope of projects PTDC/MAT-APL/28400/2017 and UIDB/MAT/00297/2020.

‡Centro de Matemática e Aplicações (CMA), FCT, UNL, 2829-516 Caparica, Portugal (m.raydan@fct.unl.pt). This author was financially supported by Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the projects UIDB/MAT/00297/2020 (Centro de Matemática e Aplicações) and CEECIND/02211/2017.

finite difference methods. This situation frequently arises when f must be evaluated through black-box simulation packages, and each function evaluation may be costly and/or contaminated with noise [12].

Recently [4, 22, 23], in a derivative-based context, several separable models combined with either a variable-norm trust-region strategy or with a cubic regularization scheme were proposed for solving (1), and their standard asymptotic convergence results were established. The main idea of these separable model approaches is to minimize a quadratic (or a cubic) model at each iteration, in which the quadratic part is the second-order Taylor approximation of the objective function. With a suitable change of variables, based on the Schur factorization, the solution of these subproblems is trivialized and an adequate choice of the norm at each iteration permits the employment of a trust-region reduction procedure that ensures the fulfillment of global convergence to second-order stationary points [4, 22]. In that case, the separable model method with a trust-region strategy has the same asymptotic convergence properties as the trust-region Newton method. Later in [23], starting with the same modeling introduced in [22], the trust-region scheme was replaced with a separable cubic regularization strategy. Adding convenient regularization terms, the standard asymptotic convergence results were retained, and moreover the complexity of the cubic strategy for finding approximate first-order stationary points became $O(\varepsilon^{-3/2})$. For the separable cubic regularization approach used in [23], complexity results with respect to second-order stationarity were also established. We note that regularization procedures serve to the same purpose and are strongly related to trust-region schemes, with the advantage of possessing improved worst-case complexity (WCC) bounds; see, e.g., [1, 5, 6, 8, 18, 20, 21, 25].

However, as previously mentioned, the separable cubic approaches developed in [4, 22, 23] are based on the availability of the exact gradient vector and the exact Hessian matrix at every iteration. When exact derivatives are not available, quadratic models which are based only on the objective function values, computed at sample points, can be obtained retaining good quality of approximation of the gradient and the Hessian of the objective function. These derivative-free models can be constructed by means of polynomial interpolation or regression or by any other approximation technique. These models are called, depending on their accuracy, fully-linear or fully-quadratic; see [9, 10, 12] for details.

Fully-linear and fully-quadratic models are the basis for derivative-free optimization trust-region methods [11, 12, 27] and have also been successfully used in the definition of a search step for unconstrained directional direct search algorithms [13]. In the latter, minimum Frobenious norm approaches are adopted, when the number of points available does not allow the computation of a determined interpolation model.

This state of affairs motivated us to develop a derivative-free separable version of the regularized method introduced in [23]. This means that we will start with a derivative-free quadratic model, which can be obtained by different schemes, to obtain an approximated gradient vector and Hessian matrix per iteration, and then we will add the separable regularization cubic terms associated with an adaptive regularization parameter to guarantee convergence to stationary points.

The paper is organized as follows. In Section 2 we present the main ideas behind the derivative-based separable modeling approaches. Section 3 revises several derivative-free schemes for building quadratic models. In Section 4 we describe our proposed derivative-free separable cubic regularization strategy, and discuss the associated convergence properties. Section 5 reports numerical results to give further insight into the proposed approach. Finally, in Section 6

we present some concluding remarks.

2 Separable cubic modeling

Let us start by recalling that in the standard derivative-based quadratic modeling approach, for solving (1), a quadratic model of $f(x)$ around x_k is constructed by defining the model of the objective function as

$$M_k(s) \;=\; f_k + g_k^\top s + \frac{1}{2} s^\top H_k s, \tag{2}$$

where $g_k = \nabla f(x_k)$ is the gradient vector at x_k, and H_k is either the Hessian of f at x_k or a symmetric approximation to the Hessian $\nabla^2 f(x_k)$. The step s_k is the minimizer of $M_k(s)$.

In [22], instead of using the standard quadratic model associated with Newton's method, the equivalent separable quadratic model

$$MS_k(y) \;=\; f_k + (Q_k^T g_k)^\top y + \frac{1}{2} y^\top D_k y \tag{3}$$

was considered to approximate the objective function f around the iterate x_k. In (3), the change of variables $y = Q_k^T s$ is used, where the spectral (or Schur) factorization of H_k:

$$H_k \;=\; Q_k D_k Q_k^T, \tag{4}$$

is computed at every iteration. In (4), Q_k is an orthogonal $n \times n$ matrix whose columns are the eigenvectors of H_k, and D_k is a real diagonal $n \times n$ matrix whose diagonal entries are the eigenvalues of H_k. Let us note that since H_k is symmetric then (4) is well-defined for all k. We also note that (3) may be non-convex, i.e., some of the diagonal entries of D_k could be negative.

For the separable regularization counterpart in [23], the separable model (3) is kept and a cubic regularization term is added:

$$MSreg_k(y) \;=\; f_k + (Q_k^T g_k)^\top y + \frac{1}{2} y^\top D_k y + \sigma_k \frac{1}{6} \sum_{i=1}^{n} |y_i|^3, \tag{5}$$

where $\sigma_k \geq 0$ is dynamically obtained. Notice that a $1/6$ factor is included in the last term of (5) to simplify derivative expressions.

As a consequence, at every iteration k the subproblem

$$\min_{y \in \mathbb{R}^n} MSreg_k(y) \;\text{ subject to }\; \|y\|_\infty \;\leq\; \Delta_k, \tag{6}$$

is solved to compute the vector y_k, and then the step will be recovered as

$$s_k \;=\; Q_k y_k.$$

The gradient of the model $MSreg_k(y)$, given by (5), can be written as follows:

$$\nabla MSreg_k(y) \;=\; Q_k^T g_k + D_k y + \frac{\sigma_k}{2} \widehat{u}_k,$$

where the i-th entry of the n-dimensional vector $\widehat{u}_k$ is equal to $|y_i| y_i$. Similarly, the Hessian of (5) is given by

$$\nabla^2 MSreg_k(y) \;=\; D_k + \mathrm{diag}(|y_i|).$$

Notice that, since D_k is diagonal, the model (5) is separable. Hence, to solve $\nabla MSreg_k(y) = 0$, and find the critical points, we only need to independently minimize n one-dimensional special functions in the closed interval $[-\Delta, \Delta]$. These special one-variable functions are of the following form

$$h(z) \;=\; c_0 + c_1 z + c_2 z^2 + c_3 |z|^3.$$

The details on how to find the global minimizer of $h(z)$ on the closed and bounded interval $[-\Delta, \Delta]$, for $\Delta > 0$, are fully described in [23, Sec. 3].

In the next section, we will describe several derivative-free alternatives to compute a model of type (2), to be incorporated in the separable model (5).

3 Fully-linear and fully-quadratic derivative-free models

Interpolation or regression based models are commonly used in derivative-free optimization as surrogates of the objective function. The simplex gradient, which is the gradient of a linear interpolation model, is used in the implicit filtering method to compute a search direction [3, 19] and quadratic interpolation models are used as replacement of Taylor models in derivative-free trust-region approaches [11, 27].

The terminology fully-linear and fully-quadratic, to describe a derivative-free model that retains Taylor-like bounds, was first proposed in [12]. Definitions 3.1 and 3.2 provide a slightly modified version of it, suited for the present work.

Assumption 3.1 *Let f be a continuously differentiable function with Lipschitz continuous gradient (with constant L_g).*

Definition 3.1 *[12, Definition 6.1] Let a function $f : \mathbb{R}^n \to \mathbb{R}$, that satisfies Assumption 3.1, be given. A set of model functions $M = \{m : \mathbb{R}^n \to \mathbb{R},\, m \in C^1\}$ is called a fully-linear class of models if:*

1. *There exist positive constants κ_{ef} and κ_{eg} such that for any $x \in \mathbb{R}^n$ and $\Delta \in (0, \Delta_{max}]$ there exists a model function $m(x+s)$ in M, with Lipschitz continuous gradient, and such that*

   - *the error between the gradient of the model and the gradient of the function satisfies*

   $$\|\nabla f(x+s) - \nabla m(x+s)\| \;\leq\; \kappa_{eg}\, \Delta, \quad \forall s \in B(0;\Delta), \tag{7}$$

   and

   - *the error between the model and the function satisfies*

   $$|f(x+s) - m(x+s)| \;\leq\; \kappa_{ef}\, \Delta^2, \quad \forall s \in B(0;\Delta). \tag{8}$$

   Such a model m is called fully-linear on $B(x;\Delta)$.

2. *For this class M there exists an algorithm, which we will call a 'model-improvement' algorithm, that in a finite, uniformly bounded (with respect to x and Δ) number of steps can*

- *either establish that a given model $m \in M$ is fully-linear on $B(x;\Delta)$ (we will say that a certificate has been provided),*
- *or find a model $m \in M$ that is fully-linear on $B(x;\Delta)$.*

For fully-quadratic models, stronger assumptions on the smoothness of the objective function are required.

Assumption 3.2 *Let f be a twice continuously differentiable function with Lipschitz continuous Hessian (with constant L_H).*

Definition 3.2 *[12, Definition 6.2] Let a function $f : \mathbb{R}^n \to \mathbb{R}$, that satisfies Assumption 3.2, be given. A set of model functions $M = \{m : \mathbb{R}^n \to \mathbb{R},\ m \in C^2\}$ is called a fully-quadratic class of models if:*

1. *There exist positive constants κ_{ef}, κ_{eg}, and κ_{eh} such that for any $x \in \mathbb{R}^n$ and $\Delta \in (0, \Delta_{max}]$ there exists a model function $m(x+s)$ in M, with Lipschitz continuous Hessian, and such that*

   - *the error between the Hessian of the model and the Hessian of the function satisfies*

   $$\|\nabla^2 f(x+s) - \nabla^2 m(x+s)\| \ \leq\ \kappa_{eh}\,\Delta, \quad \forall s \in B(0;\Delta), \tag{9}$$

   - *the error between the gradient of the model and the gradient of the function satisfies*

   $$\|\nabla f(x+s) - \nabla m(x+s)\| \ \leq\ \kappa_{eg}\,\Delta^2, \quad \forall s \in B(0;\Delta), \tag{10}$$

   and

   - *the error between the model and the function satisfies*

   $$|f(x+s) - m(x+s)| \ \leq\ \kappa_{ef}\,\Delta^3, \quad \forall s \in B(0;\Delta). \tag{11}$$

   Such a model m is called fully-quadratic on $B(x;\Delta)$.

2. *For this class M there exists an algorithm, which we will call a 'model-improvement' algorithm, that in a finite, uniformly bounded (with respect to x and Δ) number of steps can*

   - *either establish that a given model $m \in M$ is fully-quadratic on $B(x;\Delta)$ (we will say that a certificate has been provided),*
   - *or find a model $m \in M$ that is fully-quadratic on $B(x;\Delta)$.*

Algorithms for model certification or for improving the quality of a given model can be found in [12]. This quality is directly related to the geometry of the sample set used in its computation [9, 10]. However, some practical approaches have reported good numerical results related to implementations that do not consider a strict geometry control [2, 14].

4 Derivative-free separable cubic regularization approach

In a derivative-free optimization setting, instead of (2), we will consider the following quadratic model

$$\tilde{M}_k(s) \;=\; f_k + \tilde{g}_k^\top s + \frac{1}{2} s^\top \tilde{H}_k s, \tag{12}$$

where $\tilde{g}_k$ and $\tilde{H}_k$ are good quality approximations of g_k and H_k, respectively, built using interpolation or a minimum Frobenius norm approach (see Chapters 3 and 5 in [12]). Hence, analogous to the discussion in Section 2, by using the change of variables $y = \tilde{Q}_k^T s$, where $\tilde{H}_k = \tilde{Q}_k \tilde{D}_k \tilde{Q}_k^T$, with $\tilde{Q}_k$ an orthogonal $n \times n$ matrix whose columns are the eigenvectors of $\tilde{H}_k$, and $\tilde{D}_k$ is a real diagonal $n \times n$ matrix whose diagonal entries are the eigenvalues of $\tilde{H}_k$, the equivalent separable quadratic model

$$\tilde{MS}_k(y) \;=\; f_k + (\tilde{Q}_k^T \tilde{g}_k)^\top y + \frac{1}{2} y^\top \tilde{D}_k y \tag{13}$$

is used for the approximation of the objective function f around the iterate x_k. We then regularize (13) by adding a cubic or a quadratic term, depending on having been able to compute a fully-quadratic or a fully-linear model, respectively:

$$\tilde{MSreg}_k(y) \;=\; f_k + (\tilde{Q}_k^T \tilde{g}_k)^\top y + \frac{1}{2} y^\top \tilde{D}_k y + \sigma_k \frac{1}{p!} \sum_{i=1}^{n} |y_i|^p, \tag{14}$$

where $p \in \{2, 3\}$ and $\sigma_k \geq 0$ is dynamically obtained.

As a consequence, at every iteration k the subproblem

$$\min_{y \in \mathbb{R}^n} \tilde{MSreg}_k(y) \;\text{ subject to }\; \frac{\xi}{\sigma_k} \;\leq\; \|y\|_\infty \;\leq\; \Delta_k, \tag{15}$$

is solved to compute the vector y_k, and then the step will be recovered as

$$s_k \;=\; \tilde{Q}_k y_k.$$

Again, the constraint $\|y\|_\infty \;\leq\; \Delta_k$ will ensure the existence of solution for problem (15). The additional constraint $\|y\|_\infty \;\geq\; \frac{\xi}{\sigma_k}$ relates the stepsize with the regularization parameter and is required to establish the convergence results. A similar strategy has been used in [8] when building models using a probabilistic approach.

In this case, by solving n one-dimensional independent minimization problems in the closed intervals $[-\Delta, -\xi/\sigma]$ and $[\xi/\sigma, \Delta]$, we are being more demanding than the original constraint. These one-variable functions are of the form

$$h(z) \;=\; c_0 + c_1 z + c_2 z^2 + c_3 |z|^3.$$

The details on how to find the global minimizer of $h(z)$ on the closed and bounded intervals $[-\Delta, -\xi/\sigma]$ and $[\xi/\sigma, \Delta]$, for $\Delta > 0$, are similar to the ones described in [23, Sec. 3]. A practical approach for the resolution of (15) will be suggested and tested in Section 5.

The following algorithm is an adaptation of Algorithm 2.1 in [23], for the derivative-free case.

Algorithm 1

Let $\alpha > 0$, $\sigma_{small} > 0$, $\eta > 1$, and $\xi > 0$ be algorithmic parameters. Assume that $x_0 \in \mathbb{R}^n$ is a given initial approximation to the solution of problem (1). Initialize $k \leftarrow 0$.

Step 1: Choose $\sigma > 0$ and $\Delta > \frac{\xi}{\sigma}$.

Step 2: Build a quadratic polynomial model $\tilde{M}_k(s) = f_k + \tilde{g}_k^\top s + \frac{1}{2} s^\top \tilde{H}_k s$, by selecting points in $B(x_k, \frac{\xi}{\sigma})$ (fully-linear, minimum Frobenious norm models or fully-quadratic polynomial models can be considered, depending on the number of points available for reuse or on the effort allowed in terms of number of function evaluations). Set $p = 2$ (respectively $p = 3$) if the computed model is fully-linear (respectively fully-quadratic).

Step 3: Compute an approximate solution s_{trial} of

$$\text{Minimize } \tilde{g}_k^\top s + \frac{1}{2} s^\top \tilde{H}_k s + \frac{\sigma}{p!} \sum_{i=1}^{n} |[\tilde{Q}_k^\top s]_i|^p \text{ subject to } \frac{\xi}{\sigma} \ \leq \ \|\tilde{Q}_k^\top s\|_\infty \ \leq \ \Delta, \tag{16}$$

where $\tilde{H}_k = \tilde{Q}_k^\top \tilde{D}_k \tilde{Q}_k$ is a Schur factorization of $\tilde{H}_k$.

Step 4: Test the sufficient decrease condition

$$f(x_k + s_{trial}) \leq f(x_k) - \alpha \sum_{i=1}^{n} |[\tilde{Q}_k^\top s_{trial}]_i|^p. \tag{17}$$

If (17) is fulfilled, define $s_k = s_{trial}$, $x_{k+1} = x_k + s_k$, update $k \leftarrow k + 1$ and go to Step 1. Otherwise define $\sigma_{new} \in [\eta\sigma, 2\eta\sigma]$, update $\sigma \leftarrow \max\{\sigma_{small}, \sigma_{new}\}$, and go to Step 2.

In the following subsections, the convergence and worst-case behavior of Algorithm 1 will be analyzed independently for the fully-linear and fully-quadratic cases.

4.1 Fully-linear approach

This subsection will be devoted to the analysis of the WCC of Algorithm 1 when fully-linear models are used. For that, we need the following technical lemma.

Lemma 4.1 *[26, Lemma 1.2.3] Let Assumption 3.1 hold. Then, we have*

$$\left| f(x+s) - f(x) - \nabla f(x)^\top s \right| \ \leq \ \frac{L_g}{2} \|s\|^2. \tag{18}$$

As it is common in nonlinear optimization, we assume that the norm of the Hessian of each model is bounded.

Assumption 4.1 *Assume that the norm of the Hessian of the model is bounded, i.e.,*

$$\|\tilde{H}_k\| \ \leq \ \kappa_{\tilde{H}}, \quad \forall k \geq 0 \tag{19}$$

for some $\kappa_{\tilde{H}} > 0$.

We also assume that the trial point provides decrease to the current model.

Assumption 4.2 *Assume that*

$$\tilde{g}_k^\top s_{trial} + \frac{1}{2} s_{trial}^\top \tilde{H}_k s_{trial} + \frac{\sigma}{2} \sum_{i=1}^{n} [\tilde{Q}_k^\top s_{trial}]_i^2 \ \leq \ 0. \tag{20}$$

In the following lemma, we will derive an upper bound on the number of function evaluations required to satisfy the sufficient decrease condition (17), which in turn guarantees that every iteration of Algorithm 1 is well-defined. Moreover, we also obtain an upper bound for the regularization parameter.

Lemma 4.2 *Let Assumptions 3.1, 4.1, and 4.2 hold and assume that at Step 2 of Algorithm 1 a fully-linear model is always used. In order to satisfy condition* (17)*, with $p = 2$, Algorithm 1 needs at most*

$$\left\lceil \frac{\log\left(\left[2\left(\alpha + \frac{L_g}{2} + \kappa_{eg} + \frac{\kappa_{\tilde{H}}}{2}\right)\right]/\sigma_{small}\right)}{\log \eta} \right\rceil + 1 \tag{21}$$

function evaluations, not accounting for the ones required for model computation. In addition, the maximum value of σ for which (17) *is satisfied, is given by*

$$\sigma_{max} \ = \ \max\left\{\sigma_{small}, 4\eta\left(\alpha + \frac{L_g}{2} + \kappa_{eg} + \frac{\kappa_{\tilde{H}}}{2}\right)\right\}. \tag{22}$$

Proof. First, we will show that if

$$\sigma \ \geq \ 2\left(\alpha + \frac{L_g}{2} + \kappa_{eg} + \frac{\kappa_{\tilde{H}}}{2}\right) \tag{23}$$

then the sufficient decrease condition (17) of Algorithm 1 is satisfied for $p = 2$.

In view of (20), we have

$$\begin{aligned} f(x_k + s_{trial}) - f(x_k) \ &\leq \ f(x_k + s_{trial}) - f(x_k) - \tilde{g}_k^\top s_{trial} - \frac{1}{2} s_{trial}^\top \tilde{H}_k s_{trial} - \frac{\sigma}{2} \sum_{i=1}^{n} [\tilde{Q}_k^\top s_{trial}]_i^2 \\ &\leq \ |f(x_k + s_{trial}) - f(x_k) - \nabla f(x_k)^\top s_{trial}| + |(\nabla f(x_k) - \tilde{g}_k)^\top s_{trial}| \\ &\quad + \left|\frac{1}{2} s_{trial}^\top \tilde{H}_k s_{trial}\right| - \frac{\sigma}{2} \sum_{i=1}^{n} [\tilde{Q}_k^\top s_{trial}]_i^2. \end{aligned}$$

Thus, since the models are fully-linear, by using (18), (19), and $\|s_{trial}\| \geq \frac{\xi}{\sigma}$, we obtain

$$\begin{aligned} f(x_k + s_{trial}) - f(x_k) \ &\leq \ \left(\frac{L_g}{2} + \kappa_{eg} + \frac{\kappa_{\tilde{H}}}{2}\right) \|s_{trial}\|^2 - \frac{\sigma}{2} \sum_{i=1}^{n} [\tilde{Q}_k^\top s_{trial}]_i^2 \\ &= \ \left(\frac{L_g}{2} + \kappa_{eg} + \frac{\kappa_{\tilde{H}}}{2} - \frac{\sigma}{2}\right) \sum_{i=1}^{n} [\tilde{Q}_k^\top s_{trial}]_i^2 \\ &\leq \ -\alpha \sum_{i=1}^{n} [\tilde{Q}_k^\top s_{trial}]_i^2, \end{aligned}$$

where the last inequality holds due to (23).

Now, from the way as σ is updated at Step 4 of Algorithm 1, it can be easily seen that for the fulfillment of (17) with $p = 2$ we need

$$\left\lceil \frac{\log\left(\left[2\left(\alpha + \frac{L_g}{2} + \kappa_{eg} + \frac{\kappa_{\tilde{H}}}{2}\right)\right]/\sigma_{small}\right)}{\log \eta} \right\rceil + 1$$

function evaluations, and, additionally, the upper bound on σ at (22) is derived from (23). ■

The following assumption, which holds for global solutions of subproblem (16) (with $p = 2$), is central in establishing our WCC results.

Assumption 4.3 *Assume that, for all $k \geq 0$,*

$$\begin{aligned} \|\tilde{Q}_k^\top s_{trial}\|_\infty \;&=\; \frac{\xi}{\sigma_k}, \;\; or \;\|\tilde{Q}_k^\top s_{trial}\|_\infty \;=\; \Delta, \\ or \;\left\| \nabla_s \left[\tilde{g}_k^\top s + \frac{1}{2} s^\top \tilde{H}_k s + \frac{\sigma}{2} \sum_{i=1}^{n} [\tilde{Q}_k^\top s]_i^2 \right]_{s=s_{trial}} \right\| \;&\leq\; \beta_1 \|s_{trial}\|, \end{aligned} \tag{24}$$

for some $\beta_1 > 0$.

Under this assumption, we are able to prove that, when the trial point is not on the boundary of the feasible region of (16) (with $p = 2$), then the norm of the gradient of the objective function at the new point is of the same order as the norm of the trial point.

Lemma 4.3 *Let Assumptions 3.1, 4.1, 4.2, and 4.3 hold. Then, we have*

$$\|\tilde{Q}_k^\top s_{trial}\|_\infty \;=\; \frac{\xi}{\sigma_k} \;\; or \;\|\tilde{Q}_k^\top s_{trial}\|_\infty \;=\; \Delta, \tag{25}$$

or

$$\|\nabla f(x_k + s_{trial})\| \;\leq\; \kappa_1 \|s_{trial}\|, \tag{26}$$

where $\kappa_1 = L_g + \kappa_{eg} + \kappa_{\tilde{H}} + \sigma_{max} + \beta_1$, and σ_{max} was defined in Lemma 4.2.

Proof. Assume that none of the equalities at (25) hold. We have $\nabla_s \tilde{MSreg}_k(s_{trial}) = \tilde{g}_k + \tilde{H}_k s_{trial} + r(s_{trial})$, where

$$r(s_{trial}) \;=\; \sigma \tilde{Q}_k \left([\tilde{Q}_k^\top s_{trial}]_1, \ldots, [\tilde{Q}_k^\top s_{trial}]_n \right)^\top.$$

Now, by using (18), (19), and (7), we have

$$\begin{aligned} \left\| \nabla f(x_k + s_{trial}) - \nabla_s \tilde{MSreg}_k(s_{trial}) \right\| \;&= \left\| \nabla f(x_k + s_{trial}) - \left(\tilde{g}_k + \tilde{H}_k s_{trial} + r(s_{trial}) \right) \right\| \\ &\leq\; \|\nabla f(x_k + s_{trial}) - \nabla f(x_k)\| + \|\nabla f(x_k) - \tilde{g}_k\| \\ &\quad + \left\| \tilde{H}_k s_{trial} \right\| + \|r(s_{trial})\| \\ &\leq\; \left(L_g + \kappa_{eg} + \kappa_{\tilde{H}} + \sigma_{max} \right) \|s_{trial}\|. \end{aligned}$$

Therefore, in view of (24), we have

$$\begin{aligned} \|\nabla f(x_k + s_{trial})\| \;&\leq\; \|\nabla f(x_k + s_{trial}) - \nabla_s \tilde{MSreg}_k(s_{trial})\| + \|\nabla_s \tilde{MSreg}_k(s_{trial})\| \\ &\leq\; \left(L_g + \kappa_{eg} + \kappa_{\tilde{H}} + \sigma_{max} + \beta_1 \right) \|s_{trial}\|, \end{aligned}$$

which completes the proof. ■

Now, we have all the ingredients to derive an upper bound on the number of iterations required by Algorithm 1 to find a point at which the norm of the gradient is below some given positive threshold.

Theorem 4.1 *Given $\epsilon > 0$, let Assumptions 3.1, 4.1, 4.2, and 4.3 hold. Let $\{x_k\}$ be the sequence of iterates generated by Algorithm 1 and assume that $\|\nabla f(x_{k+1})\| > \epsilon$ and $f(x_{k+1}) > f_{min}$. Then, we have*

$$k+1 \;\leq\; \frac{f(x_0) - f_{min}}{\alpha \min\left\{\left(\frac{\xi}{\sigma_{max}}\right)^2, (\frac{\epsilon}{\kappa_1})^2\right\}}, \tag{27}$$

where σ_{max} and κ_1 were defined in Lemmas 4.2 and 4.3, respectively.

Proof. In view of Lemma 4.3, we have

$$\|s_k\| \;\geq\; \min\left\{\frac{\xi}{\sigma_k}, \Delta, \frac{\|\nabla f(x_{k+1})\|}{\kappa_1}\right\}. \tag{28}$$

Hence, since $\|\nabla f(x_{k+1})\| > \epsilon$ and $\Delta > \frac{\xi}{\sigma_k}$, we obtain

$$\|s_k\| \;\geq\; \min\left\{\frac{\xi}{\sigma_k}, \frac{\epsilon}{\kappa_1}\right\} \;\geq\; \min\left\{\frac{\xi}{\sigma_{max}}, \frac{\epsilon}{\kappa_1}\right\}. \tag{29}$$

On the other hand, due to the sufficient decrease condition (17), we obtain

$$\begin{aligned} f(x_{k+1}) \;&\leq\; f(x_k) - \alpha \sum_{i=1}^{n} [\tilde{Q}_k^\top s_k]_i^2 \\ &\leq\; f(x_k) - \alpha \|\tilde{Q}_k^\top s_k\|_2^2 \\ &\leq\; f(x_k) - \alpha \min\left\{\left(\frac{\xi}{\sigma_{max}}\right)^2, \left(\frac{\epsilon}{\kappa_1}\right)^2\right\}. \end{aligned}$$

By summing up these inequalities, for $0, 1, \ldots, k$, we obtain

$$k+1 \;\leq\; \frac{f(x_0) - f_{min}}{\alpha \min\left\{\left(\frac{\xi}{\sigma_{max}}\right)^2, \left(\frac{\epsilon}{\kappa_1}\right)^2\right\}},$$

which concludes the proof. ■

In view of Theorem 4.1, it is readily resulted that there is a subsequence of the points generated by Algorithm 1 for which the norm of the gradient goes to zero.

Corollary 4.1 *Let all the assumptions of Theorem 4.1 hold, and $f(x_k) > f_{min}$ for all $k \in \mathbb{N}$. Then,*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| \;=\; 0.$$

Since $\kappa_{eg} = \mathcal{O}(\sqrt{n})$ (see Chapter 2 in [12]), we have $\kappa_1 = \mathcal{O}(\sqrt{n})$. Now, if ξ is chosen such that $\frac{\xi}{\sigma_{max}} \geq \frac{\epsilon}{\kappa_1}$, then the dependency of the upper bound given at (27) on n is $\mathcal{O}(n)$. Furthermore, for building a fully-linear model we need $\mathcal{O}(n)$ function evaluations. Combining these facts with Theorem 4.1, we can derive an upper bound on the number of function evaluations that Algorithm 1 needs for driving the first-order stationarity measure below some given positive threshold.

Corollary 4.2 *Given $\epsilon > 0$, let Assumptions 3.1, 4.2, and 4.3 hold. Let $\{x_k\}$ be the sequence of iterates generated by Algorithm 1 and assume that $\|\nabla f(x_{k+1})\| > \epsilon$ and $f(x_{k+1}) > f_{min}$. Then, Algorithm 1 needs at most $\mathcal{O}\left(n^2\epsilon^{-2}\right)$ function evaluations for driving the norm of the gradient below ϵ.*

The complexity bound derived here matches the one derived in [15] for derivative-free trust-region methods.

4.2 Fully-quadratic approach

In this subsection, we will analyze the WCC of Algorithm 1 when we build fully-quadratic models. The following lemma is essential for establishing such bounds.

Lemma 4.4 *[26, Lemma 1.2.4] Let Assumption 3.2 hold. Then, we have*

$$\left| f(x+s) - f(x) - \nabla f(x)^\top s - \frac{1}{2} s^\top \nabla^2 f(x) s \right| \ \leq \ \frac{L_H}{6} \|s\|^3, \tag{30}$$

and

$$\left\| \nabla f(x+s) - \nabla f(x) - \nabla^2 f(x) s \right\| \ \leq \ \frac{L_H}{2} \|s\|^2. \tag{31}$$

Similarly to the fully-linear case, we assume that the value of the model at the trial point is less than or equal to the current function value.

Assumption 4.4 *Assume that*

$$\tilde{g}_k^\top s_{trial} + \frac{1}{2} s_{trial}^\top \tilde{H}_k s_{trial} + \frac{\sigma}{6} \sum_{i=1}^{n} |[\tilde{Q}_k^\top s_{trial}]_i|^3 \ \leq \ 0. \tag{32}$$

With this assumption, we are able to obtain upper bounds on the number of function evaluations required for satisfying the sufficient decrease condition (17) and on the regularization parameter.

Lemma 4.5 *Let Assumptions 3.2 and 4.4 hold and assume that at Step 2 of Algorithm 1 a fully-quadratic model is always used. In order to satisfy condition* (17)*, with $p = 3$, Algorithm 1 needs at most*

$$\left\lceil \frac{\log\left(\left[6\left(\alpha + \sqrt{n}\left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right)\right)\right] / \sigma_{small}\right)}{\log \eta} \right\rceil + 1 \tag{33}$$

function evaluations, not considering the ones required for model computation. In addition, the maximum value of σ for which (17) *is satisfied, is given by*

$$\sigma_{max} \ = \ \max\left\{\sigma_{small}, 12\eta\left[\alpha + \sqrt{n}\left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right)\right]\right\}. \tag{34}$$

Proof. First, we will show that if

$$\sigma \ \geq \ 6\left(\alpha + \sqrt{n}\left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right)\right) \tag{35}$$

then the sufficient decrease condition (17) of Algorithm 1 is satisfied, with $p = 3$.

In view of (30), we have

$$\begin{aligned} f(x_k + s_{trial}) - f(x_k) \;&\leq\; \nabla f(x_k)^\top s_{trial} + \frac{1}{2} s_{trial}^\top \nabla^2 f(x_k) s_{trial} + \frac{L_H}{6}\|s_{trial}\|^3 \\ &\leq\; \tilde{g}_k^\top s_{trial} + \frac{1}{2} s_{trial}^\top \tilde{H}_k s_{trial} + \frac{L_H}{6}\|s_{trial}\|^3 + |(\nabla f(x_k) - \tilde{g}_k)^\top s_{trial}| \\ &\quad + \frac{1}{2}|s_{trial}^\top (\nabla^2 f(x_k) - \tilde{H}_k) s_{trial}|. \end{aligned}$$

Thus, by using (9), (10), as the models are fully-quadratic and since $\|s_{trial}\| \geq \frac{\epsilon}{\sigma}$,we obtain

$$f(x_k + s_{trial}) - f(x_k) \;\leq\; \tilde{g}_k^\top s_{trial} + \frac{1}{2} s_{trial}^\top \tilde{H}_k s_{trial} + \left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right)\|s_{trial}\|^3.$$

Now, by applying (32), we have

$$f(x_k + s_{trial}) - f(x_k) \;\leq\; -\frac{\sigma}{6}\sum_{i=1}^{n} |[\tilde{Q}_k^\top s_{trial}]_i|^3 + \left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right)\|s_{trial}\|^3,$$

which, in view of the inequality $\|\cdot\|_3 \geq n^{-1/6}\|\cdot\|_2$ (see Theorem 16 on page 26 in [17]), leads to

$$\begin{aligned} f(x_k + s_{trial}) - f(x_k) \;&\leq\; -\frac{\sigma}{6}\sum_{i=1}^{n} |[\tilde{Q}_k^\top s_{trial}]_i|^3 + \sqrt{n}\left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right)\|\tilde{Q}_k^\top s_{trial}\|_3^3 \\ &=\; \left(\sqrt{n}\left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right) - \frac{\sigma}{6}\right)\sum_{i=1}^{n} |[\tilde{Q}_k^\top s_{trial}]_i|^3 \\ &\leq\; -\alpha\sum_{i=1}^{n} |[\tilde{Q}_k^\top s_{trial}]_i|^3, \end{aligned}$$

where the last inequality holds due to (35).

Now, from the way σ is updated at Step 4 of Algorithm 1, it can easily be seen that for the fulfillment of (35) we need

$$\left\lceil \frac{\log\left(\left[6\left(\alpha + \sqrt{n}\left(\frac{L_H}{6} + \kappa_{eg} + \frac{\kappa_{eh}}{2}\right)\right)\right]/\sigma_{small}\right)}{\log \eta} \right\rceil + 1$$

function evaluations, and, additionally, the upper bound on σ at (34) is derived from (35). ■

The following assumption is quite similar to condition (14) given in [23], and it holds for a global solution of subproblem (16) (with $p = 3$).

Assumption 4.5 *Assume that, for all $k \geq 0$,*

$$\begin{aligned} &\|\tilde{Q}_k^\top s_{trial}\|_\infty \;=\; \frac{\xi}{\sigma_k},\ \textit{or}\ \|\tilde{Q}_k^\top s_{trial}\|_\infty \;=\; \Delta, \\ \textit{or}\ &\left\|\nabla_s\left[\tilde{g}_k^\top s + \frac{1}{2} s^\top \tilde{H}_k s + \sum_{i=1}^{n}\frac{\sigma}{6}|[\tilde{Q}_k^\top s]_i|^3\right]_{s=s_{trial}}\right\| \;\leq\; \beta_2\|s_{trial}\|^2, \end{aligned} \tag{36}$$

for some $\beta_2 > 0$.

Again, we are able to prove that, when the trial point is not on the boundary of the feasible region of (16) (with $p = 3$), then the norm of the gradient of the function computed at the new point is of the order of the squared norm of the trial point.

Lemma 4.6 *Let Assumptions 3.2, 4.4, and 4.5 hold. Then, we have*

$$\|\tilde{Q}_k^\top s_{trial}\|_\infty \;=\; \frac{\xi}{\sigma_k} \;or\; \|\tilde{Q}_k^\top s_{trial}\|_\infty \;=\; \Delta, \tag{37}$$

or

$$\|\nabla f(x_k + s_{trial})\| \;\leq\; \kappa_2 \|s_{trial}\|^2, \tag{38}$$

where $\kappa_2 = \frac{L_H}{2} + \kappa_{eg} + \kappa_{eh} + \frac{\sigma_{max}}{2} + \beta_2$, *and* σ_{max} *was defined in Lemma 4.5.*

Proof. Assume that none of the equalities at (37) hold. We have $\nabla_s M\tilde{Sreg}_k(s_{trial}) = \tilde{g}_k + \tilde{H}_k s + r(s_{trial})$, where

$$r(s_{trial}) \;=\; \frac{\sigma}{2}\tilde{Q}_k \left(\text{sign}\left([\tilde{Q}_k^\top s_{trial}]_1\right)[\tilde{Q}_k^\top s_{trial}]_1^2, \ldots, \text{sign}\left([\tilde{Q}_k^\top s_{trial}]_n\right)[\tilde{Q}_k^\top s_{trial}]_n^2\right)^\top.$$

Now, by using (31), (9), and (10), we have

$$\begin{aligned}
\left\|\nabla f(x_k + s_{trial}) - \nabla_s M\tilde{Sreg}_k(s_{trial})\right\| \;&= \left\|\nabla f(x_k + s_{trial}) - \left(\tilde{g}_k + \tilde{H}_k s_{trial} + r(s_{trial})\right)\right\| \\
&\leq \left\|\nabla f(x_k + s_{trial}) - \nabla f(x_k) - \nabla^2 f(x_k) s_{trial}\right\| \\
&\quad + \|\nabla f(x_k) - \tilde{g}_k\| + \left\|\left(\nabla^2 f(x_k) - \tilde{H}_k\right) s_{trial}\right\| + \|r(s_{trial})\| \\
&\leq \left(\frac{L_H}{2} + \kappa_{eg} + \kappa_{eh} + \frac{\sigma_{max}}{2}\right)\|s_{trial}\|^2.
\end{aligned}$$

Therefore, in view of (36), we have

$$\begin{aligned}
\|\nabla f(x_k + s_{trial})\| \;&\leq\; \|\nabla f(x_k + s_{trial}) - \nabla_s M\tilde{Sreg}_k(s_{trial})\| + \|\nabla_s M\tilde{Sreg}_k(s_{trial})\| \\
&\leq\; \left(\frac{L_H}{2} + \kappa_{eg} + \kappa_{eh} + \frac{\sigma_{max}}{2} + \beta_2\right)\|s_{trial}\|^2,
\end{aligned}$$

which completes the proof. ■

Now, we have all the supporting results to establish the WCC bound of Algorithm 1 for the fully-quadratic case.

Theorem 4.2 *Given* $\epsilon > 0$, *let Assumptions 3.2, 4.4, and 4.5 hold. Let* $\{x_k\}$ *be the sequence of iterates generated by Algorithm 1 and assume that* $\|\nabla f(x_{k+1})\| > \epsilon$ *and* $f(x_{k+1}) > f_{min}$. *Then, we have*

$$k + 1 \;\leq\; \frac{\sqrt{n}(f(x_0) - f_{min})}{\alpha \min\left\{\left(\frac{\xi}{\sigma_{max}}\right)^3, (\frac{\epsilon}{\kappa_2})^{3/2}\right\}}, \tag{39}$$

where σ_{max} *and* κ_2 *were defined in Lemmas 4.5 and 4.6, respectively.*

Proof. In view of Lemma 4.6, we have

$$\|s_k\| \;\geq\; \min\left\{\frac{\xi}{\sigma_k}, \Delta, \sqrt{\frac{\|\nabla f(x_{k+1})\|}{\kappa_2}}\right\}. \tag{40}$$

Hence, since $\|\nabla f(x_{k+1})\| > \epsilon$ and $\Delta > \frac{\xi}{\sigma_k}$, we obtain

$$\|s_k\| \;\geq\; \min\left\{\frac{\xi}{\sigma_k}, \sqrt{\frac{\epsilon}{\kappa_2}}\right\} \;\geq\; \min\left\{\frac{\xi}{\sigma_{max}}, \sqrt{\frac{\epsilon}{\kappa_2}}\right\}. \tag{41}$$

On the other hand, due to the sufficient decrease condition (17) and the inequality $\|\cdot\|_3 \geq n^{-1/6}\|\cdot\|_2$, we obtain

$$\begin{aligned} f(x_{k+1}) \;&\leq\; f(x_k) - \alpha \sum_{i=1}^{n} |[\tilde{Q}_k^\top s_k]_i|^3 \\ &\leq\; f(x_k) - \frac{\alpha \|\tilde{Q}_k^\top s_k\|_2^3}{\sqrt{n}} \\ &\leq\; f(x_k) - \frac{\alpha \min\left\{\left(\frac{\xi}{\sigma_{max}}\right)^3, (\frac{\epsilon}{\kappa_2})^{3/2}\right\}}{\sqrt{n}}. \end{aligned}$$

By summing up these inequalities, for $0, 1, \ldots, k$, we obtain

$$k+1 \;\leq\; \frac{\sqrt{n}(f(x_0) - f_{min})}{\alpha \min\left\{\left(\frac{\xi}{\sigma_{max}}\right)^3, (\frac{\epsilon}{\kappa_2})^{3/2}\right\}},$$

which concludes the proof. ■

The following corollary is an immediate result of Theorem 4.2.

Corollary 4.3 *Let all the assumptions of Theorem 4.2 hold, and $f(x_k) > f_{min}$ for all $k \in \mathbb{N}$. Then,*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| \;=\; 0.$$

Similarly to what we saw before for the fully-linear case, since $\kappa_{eg} = \mathcal{O}(n)$ and $\kappa_{eh} = \mathcal{O}(n)$ (see Chapter 3 in [12]), we have $\kappa_2 = \mathcal{O}(n)$. By choosing ξ in a way such that $\frac{\xi}{\sigma_{max}} \geq \sqrt{\frac{\epsilon}{\kappa_2}}$, the dependency of the upper bound given at (39) on n becomes of the order $\mathcal{O}(n^2)$. Additionally, for building a fully-quadratic model we need $\mathcal{O}(n^2)$ function evaluations. Combining these facts with Theorem 4.2, we can establish a WCC bound for finding a stationary point as follows.

Corollary 4.4 *Given $\epsilon > 0$, let Assumptions 3.2, 4.4, and 4.5 hold. Let $\{x_k\}$ be the sequence of iterates generated by Algorithm 1 and assume that $\|\nabla f(x_{k+1})\| > \epsilon$ and $f(x_{k+1}) > f_{min}$. Then, Algorithm 1 needs at most $\mathcal{O}\left(n^4\epsilon^{-3/2}\right)$ function evaluations for driving the norm of the gradient below ϵ.*

In terms of ϵ, the derived complexity bound matches the one established in [7] for a derivative-free adaptive cubic with regularization method. The dependency of the bound derived here on n is worse than the one derived in [7]. However, we have explicitly taken into account the dependency of the constants κ_{eg} and κ_{eh} on n.

5 Numerical results

In this section we start by providing a practical implementation of Algorithm 1, which will be used to illustrate the different options to build the quadratic models at Step 2 of Algorithm 1 and two different strategies to address the subproblems (15).

Algorithm 2 (Practical implementation of Algorithm 1)

Given $x_0 \in \mathbb{R}^n$, select values for the constants $\epsilon > 0$, $\delta_{ini} \geq \xi > 0$, $\sigma_{small} > 0$, $\eta > 1$, $\alpha > 0$, and $\Delta > 0$. Define $n+1 < p_{min} \leq p_{max} \leq (n+1)(n+2)/2$ as the minimum and maximum number of points allowed in model computation. Set $k = 0$ and $stop = 0$.

while $stop = 0$, **do**

Step 1: **set** $\frac{\xi}{\sigma} = 0$, $\delta = \delta_{ini}$, and $\sigma = 0$.

Step 2: **solve** subproblem (15) **for** y_k:

Compute a quadratic model $\tilde{M}_k$, by **selecting** $p_{min} \leq p_k \leq p_{max}$ **points in** $B(x_k; \delta)$.
if $\|\tilde{g}_k\| < \epsilon$, **set** $stop = 1$ and go to **Step 5.**

end if

Compute $\tilde{H}_k = \tilde{Q}_k \tilde{D}_k \tilde{Q}_k^T$, with $\tilde{Q}_k$ an orthonormal matrix,

and **set** $b_k = \tilde{Q}_k^T \tilde{g}_k$.

for $i = 1, 2, \ldots, n$

set $(y_k)_i = z_*$ where z_* is the bounded global minimizer of the related ith one-variable function.

end for

Step 3: **set** $s_{trial} = \tilde{Q}_k y_k$, and **compute**: $\vartheta = \alpha \sum_{i=1}^{n} |(y_k)_i|^p$

if $f(x_k + s_{trial}) > f(x_k) - \vartheta$, **set** $\sigma = \max\{\sigma_{small}, \eta\sigma\}$ and $\delta = \frac{\delta_{ini}}{\sigma}$.

Go to Step 2.

end if

Step 4: **set** $s_k = s_{trial}$, $x_{k+1} = x_k + s_k$ and $k = k + 1$.

end while

Step 5: **set** $x_* = x_k$.

Model computation is a key issue for the success of the algorithm. However, in Derivative-free Optimization, saving in function evaluations by reusing previously evaluated points is a main concern. At each evaluation of a new point, the corresponding function value is stored in a list, of maximum size equal to $(n+1)(n+2)$, for possible future use in model computation. If new points need to be generated with the solely purpose of model computation, the center, 'extreme' points and 'mid-points' of the set defined by $x_k + \delta[I \ \ -I]$ are considered. Inspired by the works

of [2, 14], no explicit control of geometry is kept (in fact, we also tried the approach suggested by [27], but the results did not improve). If a new point is evaluated and the maximum number of points allowed in the list has been reached, then the point farther away from the current iterate will be replaced by the new one. Points are always selected in $B(x_k;\delta)$ for model computation. The option for δ larger than $\frac{\xi}{\sigma}$ allows a better reuse of the function values previously computed, avoiding an excessive number of function evaluations just for model computation. Additionally, the update of δ ensures that if the regularization parameter increases, the size of the neighborhood in which the points are selected decreases, a mechanism that resembles the behavior of trust-region radius in derivative-based optimization.

Fully-linear and fully-quadratic models can be considered at all iterations, as well as hybrid versions, where depending on the number of points available for reuse inside $B(x_k;\delta)$ the option for a fully-linear or a fully-quadratic model is taken (thus, some iterations will use a fully-linear model and others a fully-quadratic model). Fully-quadratic models always require $(n+1)(n+2)/2$ points for computation. Fully-linear models are built using all the points available in $B(x_k;\delta)$, once that this number is at least $n+2$ and does not exceed $(n+1)(n+2)/2-1$. In this case, a minimum Frobenius norm approach in taken to solve the linear system that provides the model coefficients (see [12, Section 5.3]). At each iteration, the initial choice $\sigma = 0$ allows to take advantage of the local properties of the "pure" quadratic models. In this case, no lower bound is considered when solving subproblem (15).

In fact, regarding the solution of this subproblem, the imposed lower bound causes additional difficulties to the separability approach. Two strategies were considered to address it. In the first, every one-dimensional problem considers the corresponding lower and upper bounds. This approach is not equivalent to the original formulation. It imposes a stronger condition since any vector y computed with this approach will satisfy $\|y\|_\infty \geq \frac{\xi}{\sigma}$, but there could be a vector y satisfying $\|y\|_\infty \geq \frac{\xi}{\sigma}$, which does not satisfy $|y_i| \geq \frac{\xi}{\sigma}, \forall i \in \{1,\ldots,n\}$. The second approach adopted disregards the lower bound condition, only considering $\|y\|_\infty \leq \Delta$ when solving subproblem (15). After computing y, the lower bound condition is tested and, if not satisfied, $\max_{i=1,\ldots,n} |y_i|$ is set equal to $\frac{\xi}{\sigma}$ to force the obtained vector y to also satisfy the lower bound constraint at (15).

Algorithm 2 was implemented in Matlab 2021a. The experiments were executed in a laptop computer with CPU Intel core i7 1.99 GHz, RAM memory of 16 GB, running Windows 10 64-bits. As test sets, we considered the smooth and nonsmooth collections proposed in [24]. Each test set has 53 problems, with a number of variables comprised between 2 and 12. Computational code for the problems and initial points can be found at `https://www.mcs.anl.gov/ more/dfo`.

Parameters in Algorithm 2 were set to the following values: $\delta_{ini} = 1$, $\xi = 10^{-5}$, $\sigma_{small} = 0.1$, $\eta = 8$, $\alpha = 10^{-4}$, and $\Delta = 10$. As stoping criteria, we consider $\epsilon = 10^{-5}$ or a maximum of 1500 function evaluations. Regarding model computation, four variants were tested, depending on using fully-linear or fully-quadratic models and also on the value of p in the sufficient decrease condition used to accept new points at Step 3 of Algorithm 2. `Fully-quadratic` variant always computes a fully-quadratic model, built using $(n+1)(n+2)/2$ points, with a cubic sufficient decrease condition ($p = 3$). `Fully-linear` always computes a quadratic model, using $n+2$ points, under a minimum Frobenious norm approach. In this case, the sufficient decrease condition considers $p = 2$. Hybrid versions compute fully-quadratic models, using $(n+1)(n+2)/2$ points or fully-linear minimum Frobenious norm models, with at least $n+2$ points and a maximum of $(n+1)(n+2)/2-1$ points (depending on the number of points available in $B(x_k;\delta)$). In this case, variant `Hybrid_p3` always uses a cubic sufficient decrease condition to

accept new points, whereas variant `Hybrid_p23` selects a quadratic or cubic sufficient decrease condition, depending on the type of model that could be computed at the current iteration ($p = 2$ for fully-linear and $p = 3$ for fully-quadratic).

Results are reported using data profiles [24]. In a simplified way, a data profile provides the percentage of problems solved by a given algorithmic variant inside a given computational budget (expressed in sets of $n_p + 1$ function evaluations). Let $\mathcal{S}$ and $\mathcal{P}$ represent the set of solvers, associated to the different algorithmic variants considered, and the set of problems to be tested, respectively. If $h_{p,s}$ represents the number of function evaluations required by algorithm $s \in \mathcal{S}$ to solve problem $p \in \mathcal{P}$ (up to a certain accuracy), the data profile cumulative function is given by

$$d_s(\zeta) \;=\; \frac{1}{|\mathcal{P}|}\left|\left\{p \in \mathcal{P} : \frac{h_{p,s}}{n_p + 1} \leq \zeta\right\}\right|. \tag{42}$$

With this purpose, a problem is considered to be solved to an accuracy level τ if the decrease obtained from the initial objective function value ($f(x_0) - f(x)$) is at least $1 - \tau$ of the best decrease obtained for all the solvers considered ($f(x_0) - f_L$), meaning:

$$f(x_0) - f(x) \;\geq\; (1 - \tau)[f(x_0) - f_L].$$

In the numerical experiments reported, the accuracy level was set equal to 10^{-5}.

Figure 1 reports the results obtained when considering different strategies for building the quadratic models. In this case, the stricter approach is used for solving subproblem (15), always imposing the lower bound for each entry of the vector y at each one-dimensional minimization.

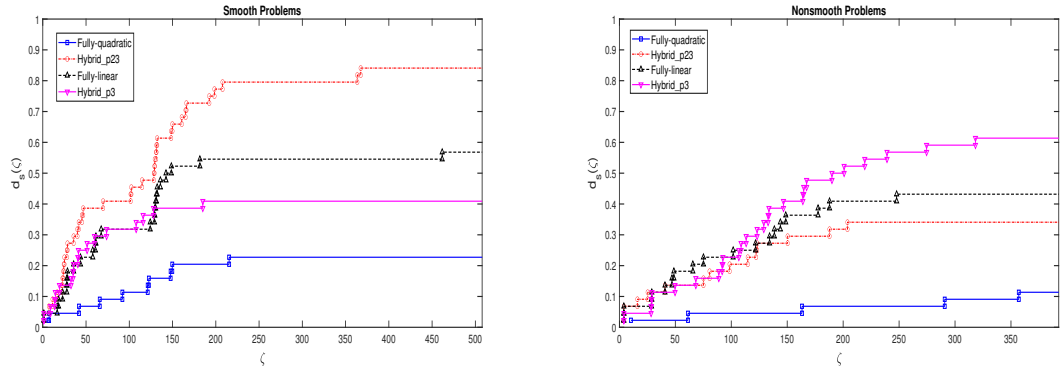

Figure 1: Data profiles comparing the use of different strategies for the computation of the quadratic models. Results on the smooth test set are on the left and on the nonsmooth test set are on the right.

It is clear that hybrid versions present the best performance and always using a fully-quadratic model shows the worst behavior. However, depending on the level of smoothness of the objective functions, the best hybrid version differs. Using a cubic sufficient decrease condition to accept new points corresponds to the variant with the best performance for the nonsmooth test set, whereas the version that adapts the type of sufficient decrease required, depending on the type of model built seems to be more adequate for smooth problems.

For the best variant in each of the two test sets, we considered the second approach to the solution of problem (15), where the lower bound constraint is initially ignored, being the

computed solution y modified *a posteriori*, if it would not satisfy the lower bound constraint. We denote these variants by adding the word `projection`. Results can be found in Figure 2.

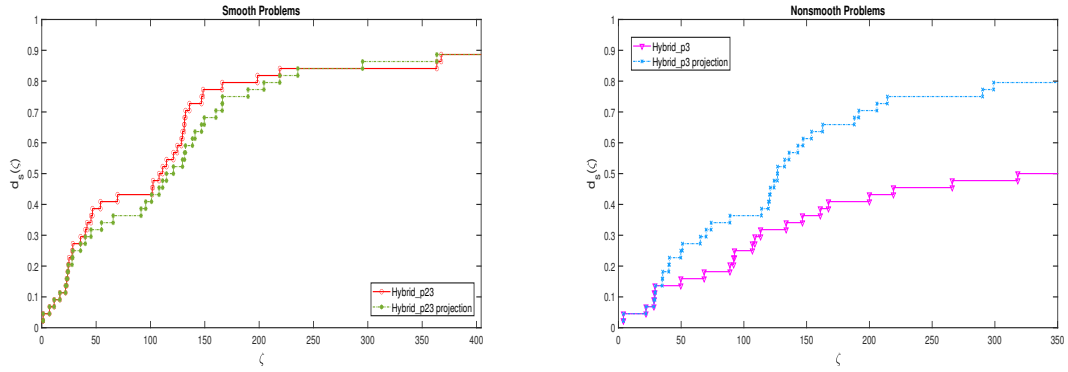

Figure 2: Data profiles comparing two different strategies to address the solution of subproblem (15). Results on the smooth test set are on the left and on the nonsmooth test set, on the right.

It is worthwhile to mention that, for both considered test sets, it was never required to modify the final solution computed by ignoring the lower bound constraint, which means that, in all our experiments by only taking into account the upper bound constraint, subproblems (15) were in fact fully solved. Thus, in the nonsmooth test set, differences between the two curves result from the stricter approach, which forces every component of the solution to satisfy the lower bound, being too conservative.

6 Conclusions and final remarks

We present and analyze a derivative-free separable regularization approach for solving smooth unconstrained minimization problems. At each iteration we build a quadratic model of the objective function using only function evaluations. Several variants have been considered for this task, from a less expensive minimum Frobenius norm approach, to a more expensive fully-quadratic model, or a hybrid version that combines the previous approaches depending on the number of available useful points from previous iterations.

For each one of the variants, we add to the model either a separable quadratic or a separable cubic regularization term to guarantee convergence to stationary points. Moreover, for each option we present a WCC analysis and we establish that, for driving the norm of the gradient below $\epsilon > 0$, the fully-quadratic and the Frobenius norm regularized approaches need at most $\mathcal{O}\left(n^4\epsilon^{-3/2}\right)$ or $\mathcal{O}\left(n^2\epsilon^{-2}\right)$ function evaluations, respectively.

The application of a convenient change of variables, based on the Schur factorization of the approximate Hessian matrices, trivializes the computation of the minimizer of the regularized models required at each iteration. In fact, the solution of the subproblem required at each iteration is reduced to the minimization of n independent one-dimensional simple functions (a polynomial of degree 2 plus a term of the form $|z|^3$) on a closed and bounded set on the real line. It is worth noticing that, for the typical low-dimensions used in DFO, the computational cost of one Schur factorization per iteration is insignificant, as compared to the cost associated with the function evaluations required to build the quadratic model.

We also present a variety of numerical experiments to add understanding and illustrate the behavior of all the different options considered for model computation. In general, we noticed that all the options show a robust performance. However, the worst behavior, concerning the required number of function evaluations, is consistently observed when using the fully-quadratic approach, and the best performance is observed when using the hybrid versions combined with a separable cubic regularization term and with the less restrictive `projection` scheme for solving the subproblems.

Concerning the worst case complexity (WCC) results obtained for the considered approaches, a few comments are in order. Even though these results are of a theoretical nature and in general pessimistic in relation to the practical behavior of the methods, it is interesting to analyze which of the two considered approaches produces a better WCC result. For that, it is convenient to use their leading terms, i.e., $n^4\epsilon^{-3/2}$ for the one using the fully-quadratic model and $n^2\epsilon^{-2}$ for the one using the Frobenius norm model. After some simple algebraic manipulations, we obtain that for the fully-quadratic approach to be better (that is, to require fewer function evaluations), it must hold that $n < \epsilon^{-1/4}$ or equivalently that $\epsilon < 1/n^4$. Therefore, if n is relatively small and ϵ is not very large (for example $n \leq 12$ and $\epsilon \leq 10^{-5}$) then the combined scheme that is based on the fully-quadratic model has a better WCC result than the scheme based on the minimum Frobenius norm approach. Notice that, indeed, in our numerical experiments $n \leq 12$ and for our stopping criterion we fix $\epsilon = 10^{-5}$, and hence from the theoretical WCC point of view, the best option is the one based on the fully-quadratic model. However, in our computational experiments the worst practical performance is clearly associated with the combination that uses the fully-quadratic model. We also note that if we choose a more tolerant stopping criterion (say $\epsilon = 10^{-2}$), then for most of the same considered small dimensional problems we have that $\epsilon > 1/n^4$, and so the scheme that uses the minimum Frobenius norm model exhibits simultaneously the best theoretical WCC result as well as the best practical performance.

References

[1] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models, *Math. Program.*, 163, pp. 359–368, 2017.

[2] 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, pp. 223–257, 2012.

[3] D. M. Bortz and C. T. Kelley, The simplex gradient and noisy optimization problems, *Computational Methods in Optimal Design and Control, Progress in Systems and Control Theory*, J. T. Borggaard, J. Burns, E. Cliff, and S. Schreck eds., Birkhäuser, Boston, 24, pp. 77–90, 1998.

[4] C. P. Brás, J.M. Martínez, and M. Raydan, Large-scale unconstrained optimization using separable cubic modeling and matrix-free subspace minimization, *Comput. Optim. Appl.*, 75, pp. 169–205, 2020.

[5] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results, *Math. Program.*, 127, pp. 245–295, 2011.

[6] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity, *Math. Program.*, 130, pp. 295–319, 2011.

[7] C. Cartis, N. I. M. Gould, and Ph. L. Toint, On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization, *SIAM J. Optim.*, 22, pp. 66–86, 2012.

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

[9] A. R. Conn, K. Scheinberg, and L. N. Vicente, Geometry of interpolation sets in derivative free optimization, *Math. Program.*, 111, pp. 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, pp. 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, pp. 387–415, 2009.

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

[13] A. L. Custódio, H. Rocha, and L. N. Vicente, Incorporating minimum Frobenius norm models in direct search, *Comput. Optim. Appl.*, 46, pp. 265–278, 2010.

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

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

[16] G. N. Grapiglia, J. Yuan, and Y.-X. Yuan, On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization, *Math. Program.*, 152, pp. 491–520, 2015.

[17] G.H Hardy, J.E Littlewood, G Pólya. *Inequalities*. Cambridge Univ. Press, New York, 1934.

[18] E. W. Karas, S. A. Santos, and B. F. Svaiter, Algebraic rules for quadratic regularization of Newton's method, *Comput. Optim. Appl.*, 60, pp. 343–376, 2015.

[19] C. T. Kelley, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999.

[20] S. Lu, Z. Wei and L. Li, A trust region algorithm with adaptive cubic regularization methods for nonsmooth convex minimization, *Comput. Optim. Appl.*, 51, pp. 551–573, 2012.

[21] J.M. Martínez, On high-order model regularization for constrained optimization, *SIAM J. Optim.*, 27, pp. 2447–2458, 2017.

[22] J.M. Martínez and M. Raydan, Separable cubic modeling and a trust-region strategy for unconstrained minimization with impact in global optimization, *J. Global Optim.* 63, pp. 319–342, 2015.

[23] J.M. Martínez and M. Raydan, Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization, *J. Global Optim.*, 68, pp. 367–385, 2017.

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

[25] Y. Nesterov and B. T. Polyak, Cubic regularization of Newton method and its global performance, *Math. Program.*, 108, pp. 177–205, 2006.

[26] Y. Nesterov, Introductory Lectures on Convex Optimization, Kluwer Academic Publishers, Dordrecht, 2004.

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