GLOBAL CONVERGENCE OF GENERAL DERIVATIVE-FREE TRUST-REGION ALGORITHMS TO FIRST AND SECOND ORDER CRITICAL POINTS

ANDREW R. CONN*, KATYA SCHEINBERG†, AND LUíS N. VICENTE‡

Abstract. In this paper we prove global convergence for first and second-order stationary points of a class of derivative-free trust-region methods for unconstrained optimization. These methods are based on the sequential minimization of quadratic (or linear) models built from evaluating the objective function at sample sets. The derivative-free models are required to satisfy Taylor-type bounds but, apart from that, the analysis is independent of the sampling techniques.

A number of new issues are addressed, including global convergence when acceptance of iterates is based on simple decrease of the objective function, trust-region radius maintenance at the criticality step, and global convergence for second-order critical points.

Key words. Trust-Region Methods, Derivative-Free Optimization, Nonlinear Optimization, Global Convergence.

AMS subject classifications. 65D05, 90C30, 90C56

1. Introduction. Trust-region methods are a well studied class of algorithms for the solution of nonlinear programming problems [2, 8]. These methods have a number of attractive features. The fact that they are intrinsically based on quadratic models makes them particularly attractive to deal with curvature information. Their robustness is partially associated with the regularization effect of minimizing quadratic models over regions of predetermined size. Extensive research on solving trust-region subproblems and related numerical issues has led to efficient implementations and commercial codes. On the other hand, the convergence theory of trust-region methods is both comprehensive and elegant in the sense that it covers many problem classes and particularizes from one problem class to a subclass in a natural way. Many extensions have been developed and analyzed to deal with different algorithmic adaptations or problem features (see [2]).

One problem feature which frequently appears in computational science and engineering is the unavailability of derivative information, which can occur in several forms and degrees. Trust-region methods have been designed since the beginning of their development to deal with the absence of second-order derivatives and to incorporate quasi-Newton techniques. However, the design and analysis of rigorous trust-region methods for derivative-free optimization, when both first and second-order derivatives are unavailable and hard to approximate directly, is a relatively recent topic [1, 3, 7, 12].

In this paper we address trust-region methods for unconstrained derivative-free optimization. These methods maintain linear or quadratic models which are based only on the objective function values computed at sample points. The corresponding models can be constructed by means of polynomial interpolation or regression or by any other approximation technique. The approach taken in this paper abstracts from

*Department of Mathematical Sciences, IBM T.J. Watson Research Center, Route 134, P.O. Box 218, Yorktown Heights, New York 10598, USA (arconn@us.ibm.com).

†Department of Mathematical Sciences, IBM T.J. Watson Research Center, Route 134, P.O. Box 218, Yorktown Heights, New York 10598, USA (katya@us.ibm.com).

‡CMUC, Department of Mathematics, University of Coimbra, 3001-454 Coimbra, Portugal (lnv@mat.uc.pt). Support for this author was provided by FCT under grant POCI/59442/MAT/2004 and PTDC/MAT/64838/2006.

the specifics of model building. In fact, it is not even required that these models are polynomial functions as long as Cauchy and eigenvalue decreases can be extracted from the trust-region subproblems. Instead, it is required that the derivative-free models have a uniform local behavior (possibly after a finite number of modifications of the sample set) similar to what is observed by Taylor models in the presence of derivatives. We call such models, depending on their accuracy, *fully linear* and *fully quadratic*. It is shown in [4, 5] how such *fully-linear* and *fully-quadratic* models can be constructed in the context of polynomial interpolation or regression.

In recent years there have been a number of trust-region based methods for derivative-free optimization. These methods can be classified into two categories: the methods which target good practical performance, such as the methods in [7, 12], and which, up to now, had no supporting convergence theory; and the methods for which global convergence was shown, but at the expense of practicality, such as described in [2, 3]. In this paper we are trying to bridge the gap by describing an algorithmic framework in the spirit of the first category of methods, while retaining all the same global convergence properties of the second category. We list next the features that make our algorithm closer to a practical one when compared to the methods in [2, 3].

The trust-region maintenance in this paper is different from the approaches in derivative-based methods [2]. In derivative-based methods, under appropriate conditions, the trust-region radius becomes bounded away from zero when the iterates converge to a local minimizer [2, Theorem 6.5.5], hence, its radius can remain unchanged or increase near optimality. This is not the case in trust-region derivative-free methods. The trust region for these methods serves two purposes: it restricts the step size to the neighborhood where the model is assumed to be good, and it also defines the neighborhood in which the points are sampled for the construction of the model. Powell in [12] suggests to use two different trust regions, which makes the method and its implementation more complicated. We choose to maintain only one trust region. However, it is important to keep the radius of the trust region comparable to some measure of stationarity so that when the measure of stationarity is close to zero (that is the current iterate may be close to a stationary point) the models become more accurate, a procedure that is accomplished by the so-called criticality step [3]. The update of the trust-region radius at the criticality step forces it to converge to zero, hence defining a natural stopping criterion for this class of methods.

Another feature of our algorithm is the acceptance of new iterates that provide simple decrease in the objective function, rather than a sufficient decrease. This feature is of particular relevance in the derivative-free context, especially when function evaluations are expensive. As in the derivative case [9], the standard liminf-type results are obtained for general trust-region radius updating schemes. In particular, it is possible to update the trust-region radius freely at the end of successful iterations (as long as it is not decreased). However, to derive the classical lim-type global convergence result [13] (see also [2, Theorem 6.4.6]) an additional requirement is imposed on the update of the trust-region radius at successful iterations, to avoid a cycling effect of the type described in [14]. But, because of the update of the trust-region radius at the criticality step mentioned in the previous paragraph, such provisions are not needed to achieve lim-type global convergence to first-order critical points even when iterates are accepted based on simple decrease. (We point out that a modification to derivative-based trust-region algorithms based on a criticality step would produce a similar lim-type result. However, forcing the trust-region radius to converge to zero may jeopardize the fast rates of local convergence in the presence of derivatives.)

In our framework it is possible to make steps, and for the algorithm to progress, without insisting that the model is made fully linear or fully quadratic on *every* iteration. In contrast with [2] and [3], we only require (i) that the models can be made fully linear or fully quadratic during a finite, uniformly bounded, number of iterations and (ii) that if a model is not fully linear or fully quadratic (depending on the order of optimality desired) in a given iteration then the new iterate can be accepted as long as it provides decrease in the objective function (sufficient decrease for the lim-result). This modification slightly complicates the convergence analysis, but it reflects much better the typical implementation of a trust-region derivative-free algorithm.

As far as we are aware, we provide the first comprehensive analysis of global convergence of trust-region derivative-free methods to second-order stationary points. It is mentioned in [2, Pages 321–322] that such analysis can be simply derived from the classical analysis for the derivative-based case. However, as we remarked above, the algorithms in [2, 3] are not as close to a practical one as the one suggested here and, moreover, the details of adjusting a 'classical' derivative-based convergence analysis to the derivative-free case are not as trivial as one might expect, even without the additional 'practical' changes to the algorithm. We observe, for instance, that it is not necessary to increase the trust-region radius on every successful iteration, as it is done in classical derivative-based methods to ensure lim-type global convergence to second-order critical points (even when iterates are accepted based on simple decrease of the objective function). In fact, in the case of the second-order analysis, the trust region needs to be increased only when it is much smaller than the measure of stationarity, to allow large steps when the current iterate is far from a stationary point and the trust-region radius is too small.

The trust-region framework we propose and analyze is sufficiently general to cover a wide class of derivative-free methods. The focus of the paper, however, is on global convergence (convergence to some form of stationarity from arbitrary starting points). We provide no analysis for local rates of convergence. As a result, the fully-linear models constructed with $n+1$ points, $2n+1$ points or $(n+1)(n+2)/2-1$ points, for instance, are treated exactly the same by our theory, while it is clear that the corresponding local convergence rates might differ significantly. As mentioned earlier, the theory supporting global convergence to first-order stationary points presented in this paper only requires that fully-linear models are constructed in a finite and uniformly bounded number of iterations. While fully-quadratic models are required for global convergence to second-order stationary points they may require an excessive number of sample points (of the order of n^2). Our framework does not enforce fully-quadratic models on every iteration, but does not eliminate the necessity of these models to achieve second-order global convergence. In cases when n is too large to allow for the use of fully-quadratic models, underdetermined quadratic models can be successfully used and the first-order global convergence theory applies.

The paper is organized as follows. In Section 2 we review the basic concepts of trust-region methods needed in this paper. The properties of fully-linear and fully-quadratic models are discussed in Section 3. Then, in Section 4 we introduce a general derivative-free trust-region method. The corresponding analysis of global convergence for first-order stationary points is given in Section 5. The second-order case is covered in Section 6 (algorithm description) and in Section 7 (analysis of global convergence to second-order stationary points).

Notation. There are several constants used in this paper which are denoted by κ with acronyms for the subscripts that are meant to be helpful. We collected their definition in this subsection, for convenience. The actual meaning of the constants will become clear when each of them is introduced in the paper.

κ_{fcd}	'fraction of Cauchy decrease'
κ_{fed}	'fraction of eigenstep decrease'
κ_{fod}	'fraction of optimal decrease'
κ_{blg}	'bound on the Lipschitz constant of the gradient of the models'
κ_{blh}	'bound on the Lipschitz constant of the Hessian of the models'
κ_{ef}	'error in the function value'
κ_{eg}	'error in the gradient'
κ_{eh}	'error in the Hessian'
κ_{bhm}	'bound on the Hessian of the models'

2. The trust-region framework basics. The problem we are considering is

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

where f is a real-valued function, assumed once (or twice) continuously differentiable and bounded from below.

As in traditional derivative-based trust-region methods, the main idea is to use a model for the objective function which one, hopefully, is able to trust in a neighborhood of the current point. The model has to be fully linear in order to ensure global convergence to a first-order critical point. One would also like to have something approaching a fully-quadratic model, to allow global convergence to a second-order critical point (and to speed up local convergence). Typically, the model is a quadratic, written in the form

$$m_k(x_k + s) \;=\; m_k(x_k) + s^\top g_k + \tfrac{1}{2} s^\top H_k s, \tag{2.1}$$

where x_k is the current iterate, $g_k \in \mathbb{R}^n$, and H_k is a symmetric matrix in $\mathbb{R}^{n\times n}$. The derivatives of this quadratic model with respect to the s variables are given by $\nabla m_k(x_k + s) = H_k s + g_k$, $\nabla m_k(x_k) = g_k$, and $\nabla^2 m_k(x_k) = H_k$.

At each iterate k, we consider the model $m_k(x_k + s)$ that is intended to approximate the true objective f within a suitable neighborhood of x_k — the trust region. This region is taken for simplicity as the set of all points

$$B(x_k; \Delta_k) \;=\; \{x \in \mathbb{R}^n :\; \|x - x_k\| \le \Delta_k\},$$

where Δ_k is called the trust-region radius, and where $\|\cdot\|$ could be an iteration dependent norm, but usually is fixed and in our case will be taken as the standard Euclidean norm.

Thus, in the unconstrained case, the local model problem we are considering is stated as

$$\min_{s\in B(0;\Delta_k)} m_k(x_k + s), \tag{2.2}$$

where $m_k(x_k + s)$ is the model for the objective function given at (2.1) and $B(0; \Delta_k)$ is our trust region, now centered at 0 and expressed in terms of $s = x - x_k$.

The Cauchy step. If we define

$$t_k^{\mathrm{C}} \;=\; \mathrm{argmin}_{t\geq 0: x_k - tg_k \in B(x_k;\Delta_k)} m_k(x_k - tg_k),$$

then the Cauchy step is a step given by

$$s_k^{\mathrm{C}} = -t_k^{\mathrm{C}} g_k. \tag{2.3}$$

A fundamental result that drives trust-region methods to first-order criticality is stated below (see [2, Theorem 6.3.3] for a proof).

THEOREM 2.1. *Consider the model (2.1) and the Cauchy step (2.3). Then,*

$$m_k(x_k) - m_k(x_k + s_k^C) \;\geq\; \tfrac{1}{2}\|g_k\| \min\left[\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right], \tag{2.4}$$

where we assume that $\|g_k\|/\|H_k\| = +\infty$ *when* $H_k = 0$.

In fact, it is not necessary to actually find the Cauchy step to achieve global convergence to first-order stationarity. It is sufficient to relate the step computed to the Cauchy step and thus what is required is the following assumption.

ASSUMPTION 2.1. *For all iterations* k,

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \kappa_{fcd}\left[m_k(x_k) - m_k(x_k + s_k^C)\right], \tag{2.5}$$

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

The steps computed under Assumption 2.1 will therefore provide a fraction of Cauchy decrease, which from Theorem 2.1 can be bounded below as

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \frac{\kappa_{fcd}}{2}\|g_k\| \min\left[\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right]. \tag{2.6}$$

If $m_k(x_k + s)$ is not a linear or a quadratic function then Theorem 2.1 is not directly applicable. In this case one could, for instance, define a Cauchy step by applying a line search at $s = 0$ along $-g_k$ to the model $m_k(x_k+s)$, stopping when some type of sufficient decrease condition is satisfied (see [2, Section 6.3.3]). Calculating a step yielding a decrease better than the Cauchy decrease could be achieved whenever possible by approximately solving the trust-region subproblem, which involves now the minimization of a nonlinear function within a trust region.

The eigenstep. When considering a quadratic model and global convergence to second-order critical points, the model reduction that is required can be achieved along a direction related to the greatest negative curvature. Let us assume that H_k has at least one negative eigenvalue and let $\tau_k < 0$ be the most negative eigenvalue of H_k. In this case, we can determine a step of negative curvature s_k^{E}, such that

$$(s_k^{\mathrm{E}})^\top (g_k) \;\leq\; 0, \quad \|s_k^{\mathrm{E}}\| \;=\; \Delta_k, \quad \text{and} \quad (s_k^{\mathrm{E}})^\top H_k (s_k^{\mathrm{E}}) \;=\; \tau_k \Delta_k^2. \tag{2.7}$$

We refer to s_k^{E} as the eigenstep.

The eigenstep s_k^{E} is the eigenvector of H_k corresponding to the most negative eigenvalue τ_k, whose sign and scale are chosen to ensure that the first two parts of (2.7) are satisfied. Note that due to the presence of negative curvature, s_k^{E} is the minimizer of the quadratic function along that direction inside the trust region. Also we do not have to insist that we use the eigenvector corresponding to the most negative eigenvalue, any direction with sufficient negative curvature would be suitable,

whereupon the lemma that follows would provide a fraction of the same decrease. The eigenstep s_k^{E} induces the following decrease in the model (the proof is trivial and omitted).

LEMMA 2.2. *Suppose that the model Hessian H_k has negative eigenvalues. Then we have that*

$$m_k(x_k) - m_k(x_k + s_k^E) \;\geq\; -\tfrac{1}{2}\tau_k \Delta_k^2. \tag{2.8}$$

The eigenstep plays a role similar to that of the Cauchy step, in that, provided negative curvature is present in the model, we now require the model decrease at $x_k + s_k$ to satisfy

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \kappa_{fed}[m_k(x_k) - m_k(x_k + s_k^{\mathrm{E}})],$$

for some constant $\kappa_{fed} \in (0, 1]$. Since we also want the step to yield a fraction of Cauchy decrease, we will consider the following assumption.

ASSUMPTION 2.2. *For all iterations k,*

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \kappa_{fod}\left[m_k(x_k) - \min\{m_k(x_k + s_k^C), m_k(x_k + s_k^E)\}\right], \tag{2.9}$$

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

A step satisfying this assumption is given, for instance, by computing both the Cauchy step and, in the presence of negative curvature in the model, the eigenstep, and by choosing the one that provides the larger reduction in the model. By combining (2.4), (2.8), and (2.9), we obtain that

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \frac{\kappa_{fod}}{2} \max\left\{ \|g_k\| \min\left[\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right], -\tau_k \Delta_k^2 \right\}. \tag{2.10}$$

In some trust-region literature what is required for global convergence to second-order critical points is a fraction of the decrease obtained by the optimal trust-region step (i.e, an optimal solution of (2.2)). Note that a fraction of optimal decrease condition is stronger than (2.10) for the same value of κ_{fod}.

If $m_k(x_k+s)$ is not a quadratic function then Theorem 2.1 and Lemma 2.2 are not directly applicable. Similarly to the Cauchy step case, one could here define an eigenstep by applying a line search to the model $m_k(x_k+s)$, at $s = 0$ and along a direction of negative (or most negative) curvature of H_k, stopping when some type of sufficient decrease condition is satisfied (see [2, Section 6.6.2]). Calculating a step yielding a decrease better than the Cauchy and eigen decreases could be achieved whenever possible by approximately solving the trust-region subproblem, which, again, involves now the minimization of a nonlinear function within a trust region.

3. Conditions on derivative-free models. Since we cannot use Taylor models, the most obvious replacement is a polynomial interpolation model. In fact, in what follows we may use polynomial interpolation or regression models (see [4, 5]) depending upon the underlying basis and the number of function values available. What one requires in these cases for the theory to hold is Taylor-like error bounds with a uniformly bounded constant that characterizes the geometry of the sample sets.

In this paper we will abstract from the specifics of the models that we use. We will only impose those requirements on the models that are essential for the convergence

theory. We will then indicate that polynomial interpolation and regression models, in particular, satisfy our requirements.

We will now discuss the assumptions on the models which we use to prove the convergence of our derivative-free trust-region framework.

Fully-linear models. For the purposes of convergence to first-order critical points, we assume that the function f and its gradient are Lipschitz continuous in regions considered by a potential algorithm. To better define this region, we suppose that x_0 (the initial iterate) is given and that new iterates correspond to reductions in the value of the objective function. Thus, the iterates must necessarily belong to the level set

$$L(x_0) \;=\; \{x \in \mathbb{R}^n :\; f(x) \leq f(x_0)\}\,.$$

However, when considering models based on sampling it is possible (especially at the early iterations) that the function f is evaluated outside $L(x_0)$. Let us assume that sampling is restricted to regions of the form $B(x_k; \Delta_k)$ and that Δ_k never exceeds a given (possibly large) positive constant Δ_{max}. Under this scenario, the region where f is sampled is within the set

$$L_{enl}(x_0) \;=\; L(x_0) \cup \bigcup_{x \in L(x_0)} B(x; \Delta_{max}) \;=\; \bigcup_{x \in L(x_0)} B(x; \Delta_{max}).$$

For fully-linear models and global convergence to first-order critical points we require the existence of the first-order derivatives and their Lipschitz continuity.

ASSUMPTION 3.1. *Suppose x_0 and Δ_{max} are given. Assume that f is continuously differentiable in an open domain containing the set $L_{enl}(x_0)$ and that ∇f is Lipschitz continuous on $L_{enl}(x_0)$.*

Now we discuss the corresponding assumptions on the models, by introducing the abstract concept of a fully-linear model.

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

1. *There exist positive constants κ_{ef}, κ_{eg}, and κ_{blg} such that for any $x \in L(x_0)$ and $\Delta \in (0, \Delta_{max}]$ there exists a model function $m(x+s)$ in $\mathcal{M}$, with Lipschitz continuous gradient and corresponding Lipschitz constant bounded by κ_{blg}, and such that*
   - *the error between the gradient of the model and the gradient of the function satisfies*

$$\|\nabla f(x+s) - \nabla m(x+s)\| \;\leq\; \kappa_{eg}\,\Delta, \quad \forall s \in B(0;\Delta), \tag{3.1}$$

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

$$|f(x+s) - m(x+s)| \;\leq\; \kappa_{ef}\,\Delta^2, \quad \forall s \in B(0;\Delta). \tag{3.2}$$

   Such a model m is called fully linear on $B(x;\Delta)$.
2. *For this class $\mathcal{M}$ 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 $m \in \mathcal{M}$ is fully linear on $B(x;\Delta)$ (we will say that a certificate has been provided and the model is certifiably fully linear),*
- *or find a model $\tilde{m} \in \mathcal{M}$ that is fully linear on $B(x;\Delta)$.*

If a model is fully linear on $B(x;\bar{\Delta})$ with respect to some (large enough) constants κ_{ef}, κ_{eg}, and κ_{blg} and for some $\bar{\Delta} \in (0, \Delta_{max}]$, then it is also fully linear on $B(x;\Delta)$ for any $\Delta \in [\bar{\Delta}, \Delta_{max}]$, with the same constants. This result is stated next. The proof is omitted since it can be derived easily from the proof of the fully-quadratic case (see Lemma 3.4).

LEMMA 3.2. *Consider a function f satisfying Assumption 3.1 and a model m fully linear, with respect to constants κ_{ef}, κ_{eg}, and κ_{blg} on $B(x;\bar{\Delta})$, with $x \in L(x_0)$ and $\bar{\Delta} \leq \Delta_{max}$.*

Assume also, without loss of generality, that κ_{eg} is no less than the sum of κ_{blg} and the Lipschitz constant of the gradient of f, and that $\kappa_{ef} > (1/2)\kappa_{eg}$.

Then m is fully linear on $B(x;\Delta)$, for any $\Delta \in [\bar{\Delta}, \Delta_{max}]$, with respect to the same constants κ_{ef}, κ_{eg}, and κ_{blg}.

For the remainder of the paper we assume, without loss of generality, that the constants κ_{ef}, κ_{eg}, and κ_{blg} of any fully-linear class $\mathcal{M}$ which we use in our algorithm are such that Lemma 3.2 holds.

The algorithmic framework which we describe and analyze in Sections 4 and 5 relies on a fully-linear class $\mathcal{M}$. To prove global convergence all that is needed is that the models used in the algorithm belong to such a class and that Assumption 2.1 is satisfied. We allow as much flexibility for the choice of models as we can, while retaining the convergence properties.

As a consequence of this flexibility some of the model classes that fit in the framework are usually of no interest for a practical algorithm. For instance, consider $\mathcal{M} = \{f\}$ — a class consisting of the function f itself. Clearly, by Definition 3.1 such an $\mathcal{M}$ is a fully-linear class of models, since f is a fully-linear model of itself for any x and Δ and since the algorithm for verifying that f is fully-linear is trivial. However, in derivative-free optimization, $m = f$ is not expected to be a quadratic function. We already discussed in Section 2 how to compute Cauchy steps and eigensteps for non-quadratic models based on existing model gradients (which in this case would amount to gradients of the function f itself). And, even if some model gradient is available to extract, for instance, some form of fraction of Cauchy decrease by line search, improving this decrease by approximately solving the trust-region subproblem, $\min f(x_k + s)$ s.t. $s \in B(0;\Delta_k)$, seems a problem nearly as complicated as the original one.

Another source of impractical fully-linear classes is the flexibility in the choice of a model-improvement algorithm. The definition requires the existence of a finite procedure which either certifies that a model is fully linear or produces such a model. For example, Taylor models based on suitably chosen finite-differences gradient evaluations are a fully-linear class of models, but a model-improvement algorithm needs to build such models 'from scratch' for each new x and Δ. In a derivative-free algorithm with expensive (and often noisy) function evaluations this approach is typically impractical. However, our framework still supports such an approach and guarantees its convergence, provided that all necessary assumptions are satisfied.

To justify the usefulness of our framework we will show at the end of this section that, reasonable, practical fully-linear model classes exist, i.e., such classes for which the fraction of Cauchy decrease is easy to obtain and improve by approximately solving

the trust-region subproblem, and for which there exists a practical model-improvement algorithm.

First, we extend Definition 3.1 to fully-quadratic classes of models.

Fully-quadratic models. For global convergence to second-order critical points, we will need an assumption on the Hessian of f.

ASSUMPTION 3.2. *Suppose x_0 and Δ_{max} are given. Assume that f is twice continuously differentiable in an open domain containing the set $L_{enl}(x_0)$ and that $\nabla^2 f$ is Lipschitz continuous on $L_{enl}(x_0)$.*

We will now introduce formally the concept of fully-quadratic classes and models.

DEFINITION 3.3. *Let a function f, that satisfies Assumption 3.2, be given. A set of model functions $\mathcal{M} = \{m : \mathbb{R}^n \to \mathbb{R},\, m \in C^2\}$ is called a fully-quadratic class of models if*

1. *There exist positive constants κ_{ef}, κ_{eg}, κ_{eh}, and κ_{blh}, such that for any $x \in L(x_0)$ and $\Delta \in (0, \Delta_{max}]$ there exists a model function $m(x+s)$ in $\mathcal{M}$, with Lipschitz continuous Hessian and corresponding Lipschitz constant bounded by κ_{blh}, and such that*
   - *the error between the Hessian of the model and the Hessian of the function satisfies*

$$\|\nabla^2 f(x+s) - \nabla^2 m(x+s)\| \;\leq\; \kappa_{eh}\,\Delta, \quad \forall s \in B(0;\Delta), \tag{3.3}$$

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

$$\|\nabla f(x+s) - \nabla m(x+s)\| \;\leq\; \kappa_{eg}\,\Delta^2, \quad \forall s \in B(0;\Delta), \tag{3.4}$$

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

$$|f(x+s) - m(x+s)| \;\leq\; \kappa_{ef}\,\Delta^3, \quad \forall s \in B(0;\Delta). \tag{3.5}$$

   Such a model m is called fully quadratic on $B(x;\Delta)$.
2. *For this class $\mathcal{M}$ 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 $m \in \mathcal{M}$ is fully quadratic on $B(x;\Delta)$ (we will say that a certificate has been provided and the model is certifiably fully quadratic),*
   - *or find a model $\tilde{m} \in \mathcal{M}$ that is fully quadratic on $B(x;\Delta)$.*

We will now show that if a model is fully quadratic on $B(x;\bar{\Delta})$ with respect to some (large enough) constants κ_{ef}, κ_{eg}, κ_{eh}, and κ_{blh} and for some $\bar{\Delta} \in (0, \Delta_{max}]$, then it is also fully quadratic on $B(x;\Delta)$ for any $\Delta \in [\bar{\Delta}, \Delta_{max}]$, with the same constants.

LEMMA 3.4. *Consider a function f satisfying Assumption 3.2 and a model m fully quadratic, with respect to constants κ_{ef}, κ_{eg}, κ_{eh}, and κ_{blh} on $B(x;\bar{\Delta})$, with $x \in L(x_0)$ and $\bar{\Delta} \leq \Delta_{max}$.*

Assume also, without loss of generality, that κ_{eh} is no less than the sum of κ_{blh} and the Lipschitz constant of the Hessian of f, and that $\kappa_{eg} \geq (1/2)\kappa_{eh}$ and $\kappa_{ef} \geq (1/3)\kappa_{eg}$.

Then m is fully quadratic on $B(x;\Delta)$, for any $\Delta \in [\bar{\Delta}, \Delta_{max}]$, with respect to the same constants κ_{ef}, κ_{eg}, κ_{eh}, and κ_{blh}.

Proof. Let us consider any $\Delta \in [\bar{\Delta}, \Delta_{max}]$. Consider, also, an s such that $\bar{\Delta} \leq \|s\| \leq \Delta$, and let $\theta = \bar{\Delta}/\|s\|$. Since $x + \theta s \in B(x; \bar{\Delta})$ then, due to the model being fully quadratic on $B(x; \bar{\Delta})$, we know that

$$\|\nabla^2 f(x + \theta s) - \nabla^2 m(x + \theta s)\| \ \leq \ \kappa_{eh} \bar{\Delta}.$$

Since $\nabla^2 f$ and $\nabla^2 m$ are Lipschitz continuous and since κ_{eh} is no less than the sum of the corresponding Lipschitz constants, we have

$$\|\nabla^2 f(x + s) - \nabla^2 f(x + \theta s) - \nabla^2 m(x + \theta s) + \nabla^2 m(x + s)\| \ \leq \ \kappa_{eh}(\|s\| - \bar{\Delta}).$$

Thus, by combining the above expressions we obtain

$$\|\nabla^2 f(x + s) - \nabla^2 m(x + s)\| \ \leq \ \kappa_{eh}\|s\| \ \leq \ \kappa_{eh}\Delta. \tag{3.6}$$

Now let us consider the vector function $g(\alpha) = \nabla f(x+\alpha s) - \nabla m(x+\alpha s)$, $\alpha \in [0, 1]$. From the fact that m is a fully-quadratic model on $B(x; \bar{\Delta})$ we have $\|g(\theta)\| \leq \kappa_{eg}\bar{\Delta}^2$. We are interested in bounding $\|g(1)\|$, which can be achieved by bounding $\|g(1) - g(\theta)\|$ first. By applying the integral mean value theorem componentwise, we obtain

$$\|g(1) - g(\theta)\| \ = \ \left\| \int_\theta^1 g'(\alpha) d\alpha \right\| \ \leq \ \int_\theta^1 \|g'(\alpha)\| d\alpha.$$

Now, using (3.6) we have

$$\begin{aligned} \int_\theta^1 \|g'(\alpha)\| d\alpha &\leq \int_\theta^1 \|s\| \|\nabla^2 f(x + \alpha s) - \nabla^2 m(x + \alpha s)\| \ d\alpha \\ &\leq \int_\theta^1 \alpha \kappa_{eh} \|s\|^2 d\alpha \ = \ (1/2)\kappa_{eh}(\|s\|^2 - \bar{\Delta}^2). \end{aligned}$$

Hence from $\kappa_{eg} \geq 1/2\kappa_{eh}$ we obtain

$$\|\nabla f(x + s) - \nabla m(x + s)\| \ \leq \ \|g(1) - g(\theta)\| + \|g(\theta)\| \ \leq \ \kappa_{eg}\|s\|^2 \ \leq \ \kappa_{eg}\Delta^2. \tag{3.7}$$

Finally, we consider the function $\phi(\alpha) = f(x + \alpha s) - m(x + \alpha s)$, $\alpha \in [0, 1]$. From the fact that m is a fully-quadratic model on $B(x; \bar{\Delta})$, we have $|\phi(\theta)| \leq \kappa_{ef}\bar{\Delta}^3$. We are interested in bounding $|\phi(1)|$, which can be achieved by bounding $|\phi(1) - \phi(\theta)|$ first by using (3.7):

$$\begin{aligned} \left| \int_\theta^1 \phi'(\alpha) d\alpha \right| &\leq \int_\theta^1 \|s\| \|\nabla f(x + \alpha s) - \nabla m(x + \alpha s)\| \ d\alpha \\ &\leq \int_\theta^1 \alpha^2 \kappa_{eg} \|s\|^3 d\alpha \ = \ (1/3)\kappa_{eg}(\|s\|^3 - \bar{\Delta}^3). \end{aligned}$$

Hence, from $\kappa_{ef} \geq (1/3)\kappa_{eg}$ we obtain

$$|f(x + s) - m(x + s)| \ \leq \ |\phi(1) - \phi(\theta)| + |\phi(\theta)| \ \leq \ \kappa_{ef}\|s\|^3 \ \leq \ \kappa_{ef}\Delta^3.$$

The proof is complete. □

For the remainder of the paper we assume, without loss of generality, that the constants κ_{ef}, κ_{eg}, κ_{eh}, κ_{blh} of any fully-quadratic class $\mathcal{M}$ which we use in our algorithm are such that Lemma 3.4 holds.

The discussion after the definition of fully-linear class of models applies to the fully-quadratic case almost word for word. In particular, this means that there is much flexibility in the definition of the fully-quadratic class of models which allows for both practical and generally impractical choices. We justify our definition by showing that the classical choice of derivative-free models — quadratic interpolation polynomials — form a practical fully-quadratic class (and, hence, a practical fully-linear class as well).

Polynomial models. Given a function f that satisfies Assumption 3.2 let us consider the set of all quadratic functions that interpolate f at exactly $(n+1)(n+2)/2$ distinct points. Given x and Δ, let $Y \in B(x;\Delta)$ be a set of interpolation points.

DEFINITION 3.5. *Given a set of interpolation points $Y = \{y^0, y^1, \ldots, y^p\}$, with $p = (n+1)(n+2)/2 - 1$, a basis of $p+1$ polynomials $\ell_j(x)$, $j = 0, \ldots, p$, of degree ≤ 2, is called a basis of Lagrange polynomials if*

$$\ell_j(y^i) \;=\; \delta_{ij} \;=\; \begin{cases} 1 & \text{if } \; i = j, \\ 0 & \text{if } \; i \neq j. \end{cases}$$

A set Y is called poised if and only if the basis of Lagrange polynomials exists and is unique (see [6, Chapter 3]). Given $\Lambda > 0$, we say that a poised set Y is Λ–poised in $B(x;\Delta)$ if $Y \subset B(x;\Delta)$ and

$$\Lambda \;\geq\; \max_{0 \leq i \leq p} \max_{\|s\| \leq \Delta} |\ell_i(x+s)|.$$

It is known (for example, see [4]) that if Y is Λ–poised in $B(x;\Delta)$ then the corresponding interpolating polynomial $m(x+s)$ exists, is unique, and satisfies (3.3)–(3.5) (for this given x and $\Delta \leq \Delta_{max}$ and for some constants κ_{ef}, κ_{eg}, and κ_{eh} which depend only on Λ, n, and the Lipschitz constant of $\nabla^2 f$ in the Assumption 3.2). Hence, we conclude that any quadratic polynomial which interpolates f on any Λ-poised interpolation set Y is an element of the same fully-quadratic class.

We now discuss possible model-improvement algorithms to construct Λ–poised interpolation sets with uniformly bounded Λ. There are two main requirements for a set Y to be Λ–poised in $B(x;\Delta)$:

1. $Y \subset B(x;\Delta)$.
2. $\max_{0 \leq i \leq p} \max_{\|s\| \leq \Delta} |\ell_i(x+s)| \leq \Lambda$.

The first condition is easy to check and to enforce, at least in theory, by replacing at most p points (it is usually assumed that $x \in Y$, hence at least one interpolation point is always in $B(x;\Delta)$).

Ensuring the bound on the Lagrange polynomials, on the other hand, requires significant effort. In [4] two algorithms are proposed based on QR or LU factorizations of a multivariate version of a Vandermonde matrix to determine whether a given set Y is Λ–poised. It is shown that if the pivot values encountered during such a factorization remain (in absolute value) above a certain fixed positive threshold, then the set is Λ–poised for some large enough Λ, whose value depends on the pivot threshold. Each pivot corresponds to an interpolation point, in fact it is a value of a certain polynomial, let us call it a *pivot polynomial*, at this interpolation point. At each step of the factorization algorithm such a pivot polynomial is generated, and is evaluated at all remaining interpolation points. The point which gives the largest pivot value is selected and if the pivot value is above the threshold, then the point it accepted and the next factorization step begins. If the pivot value is too small, then

a new point is generated. It is shown in [4] that if the threshold is reasonably small (smaller than 1/4 in the quadratic case), then it is always possible to find a point in $B(x;\Delta)$ for which the absolute value of this pivot polynomial is above the threshold. Moreover, such a point can be obtained by a simple enumerating scheme. Hence, if small pivots are encountered during the factorization, then the 'unacceptable' points are replaced by 'acceptable' ones and after the factorization is completed the resulting set Y is Λ–poised, with Λ independent of x and Δ.

Let us discuss whether such procedure is 'practical'. Derivative-free optimization problems typically address functions whose evaluation is expensive, hence a practical approach should attempt to economize on function evaluations. The first question about the pivoting algorithm is whether too many interpolation points need to be replaced at each iteration.

If an interpolation point is outside $B(x;\Delta)$, then it has to be replaced. Our algorithmic framework allows replacing only one point per iteration, hence allowing for the possibility of further progress even before a fully-quadratic model is constructed. Another situation when an interpolation point needs to be replaced is when a new iterate is found and needs to be included in the interpolation set. In this case the factorization algorithm will simply start by choosing the new iterate to generate the first pivot and then proceed by choosing points which produce the best pivot value until the factorization is complete. The remaining unused point will be the one which is replaced.

If at a given step of the factorization algorithm one cannot find an interpolation point which gives the pivot value above the threshold then a new interpolation point needs to be generated. Our framework again allows generating only one such new point per iteration. In practice, it turns out that if care is taken when replacing 'far away' points, it is rarely necessary to replace points because of bad pivot values. It is often beneficial to replace points anyway to improve overall poisedness, but this can be done in an economical manner, in the sense that at most one point per iteration gets replaced. Hence we claim that this procedure is reasonably efficient in practice in terms of the number of function evaluations.

In terms of the linear algebra cost involved in completing the factorization procedure, this cost can be as high as $\mathcal{O}(n^6)$ per iteration to recompute all $(n+1)(n+2)/2-1$ pivot values. This cost is acceptable for many derivative-free applications, where the cost of function evaluations is dominant and the dimension n is not large. However, there are some cases when n is of the order of 100 and the cost of a function evaluation is not as high as the cost of linear algebra per iteration.

An alternative method of maintaining interpolation models was suggested by Powell in [10], [11], and in [12]. His method is based on considering the absolute value of the Lagrange polynomials as the criterion for the acceptance of new interpolation points. There are two possible situations when interpolation points are replaced.

1. A new interpolation point has to be included in the interpolation set (because it is the new iterate). It replaces an interpolation point whose corresponding Lagrange polynomial has a large absolute value at the new point.
2. A model is suspected of being inaccurate. Then a point furthest from the current iterate is replaced by a point within $B(x;\Delta)$, which maximizes, possibly approximately, the absolute value of the corresponding Lagrange polynomial.

Both of these actions are aimed at keeping points within $B(x;\Delta)$, and at reducing the maximum value Λ of the Lagrange polynomials. This approach is efficient in that it only replaces one or two points per iteration and the update of all Lagrange

polynomial coefficients requires at most $\mathcal{O}(n^4)$ per iteration. However, in addition we need to globally optimize an absolute value of a Lagrange polynomial. See [6, Chapter 6] for more details.

In [12] Powell suggests using $2n+1$ points to construct quadratic models based on the minimization of the Frobenius norm of the change of the model Hessian. This ensures the reduction of the linear algebra per-iteration cost, while still providing adequate quadratic models. Similar techniques can be used in conjunction with the algorithm in [4] to reduce the cost of the linear algebra. If appropriate care is taken, the models based on $2n+1$ points can be guaranteed to be fully linear.

We conclude this section by noting that the case of fully-linear models fits into a similar framework. In fact, linear interpolation models can be chosen to satisfy the requirements of Definition 3.1 (see [4]). In addition, linear and quadratic regression polynomial models can also be chosen to satisfy the requirements of Definitions 3.1 and 3.3, respectively (see [5]). We have therefore shown the existence of several classes of models which fit into our algorithmic framework.

The purpose of our abstraction of fully-linear and fully-quadratic models is to allow for the use of models different from polynomial interpolation and regression, as long as these models satisfy Assumptions 2.1 and 2.2 and fit the Definitions 3.1 and 3.3. The abstraction highlights, in our opinion, the fundamental requirements for obtaining the appropriate convergence results.

4. Derivative-free trust-region methods (first order). We now formally state the first-order version of the algorithm that we consider. We point out that the model m_k and the trust-region radius Δ_k are only set at the end of the criticality step (Step 1). The iteration ends by defining an incumbent model m_{k+1}^{icb} and an incumbent trust-region radius Δ_{k+1}^{icb} for the next iteration, which might then be changed or not by the criticality step.

ALGORITHM 4.1 (**Derivative-free trust-region method (1st order)**).

Step 0 (initialization): Choose a fully-linear class of models $\mathcal{M}$ and a corresponding model-improvement algorithm (see, e.g., [4]). Choose an initial point x_0 and $\Delta_{max} > 0$. We assume that an initial model m_0^{icb} (with gradient and possibly the Hessian at $s = 0$ given by g_0^{icb} and H_0^{icb}, respectively) and a trust-region radius $\Delta_0^{icb} \in (0, \Delta_{max}]$ are given.
The constants η_0, η_1, γ, γ_{inc} ϵ_c, β, μ, and α are also given and satisfy the conditions $0 \leq \eta_0 \leq \eta_1 < 1$ (with $\eta_1 \neq 0$), $0 < \gamma < 1 < \gamma_{inc}$, $\epsilon_c > 0$, $\mu > \beta > 0$, and $\alpha \in (0,1)$. Set $k = 0$.

Step 1 (criticality step): If $\|g_k^{icb}\| > \epsilon_c$ then $m_k = m_k^{icb}$ and $\Delta_k = \Delta_k^{icb}$.
If $\|g_k^{icb}\| \leq \epsilon_c$ then proceed as follows. Call the model-improvement algorithm to attempt to certify if the model m_k^{icb} is fully linear on $B(x_k; \Delta_k^{icb})$. If at least one of the following conditions holds,
- the model m_k^{icb} is not certifiably fully linear on $B(x_k; \Delta_k^{icb})$,
- $\Delta_k^{icb} > \mu \|g_k^{icb}\|$,

then apply Algorithm 4.2 (described below) to construct a model $\tilde{m}_k(x_k + s)$ (with gradient and possibly the Hessian at $s = 0$ given by $\tilde{g}_k$ and $\tilde{H}_k$, respectively), which is fully linear (for some constants κ_{ef}, κ_{eg}, and κ_{blg}, which remain the same for all iterations of Algorithm 4.1) on the ball $B(x_k; \tilde{\Delta}_k)$,

for some $\tilde{\Delta}_k \in (0, \mu\|\tilde{g}_k\|]$ given by Algorithm 4.2. In such a case set[1]

$$m_k \;=\; \tilde{m}_k \;\text{ and }\; \Delta_k \;=\; \min\{\max\{\tilde{\Delta}_k, \beta\|\tilde{g}_k\|\}, \Delta_k^{icb}\}.$$

Otherwise set $m_k = m_k^{icb}$ and $\Delta_k = \Delta_k^{icb}$.

Step 2 (step calculation): Compute a step s_k that sufficiently reduces the model m_k (in the sense of (2.5)) and such that $x_k + s_k \in B(x_k; \Delta_k)$.

Step 3 (acceptance of the trial point): Compute $f(x_k + s_k)$ and define

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

If $\rho_k \geq \eta_1$ or if both $\rho_k \geq \eta_0$ and the model is fully linear (for the positive constants κ_{ef}, κ_{eg}, and κ_{blg}) on $B(x_k; \Delta_k)$, then $x_{k+1} = x_k + s_k$ and the model is updated to include the new iterate into the sample set, resulting in a new model m_{k+1}^{icb} (with gradient and possibly the Hessian at $s = 0$ given by g_{k+1}^{icb} and H_{k+1}^{icb}, respectively); otherwise the model and the iterate remain unchanged ($m_{k+1}^{icb} = m_k$ and $x_{k+1} = x_k$).

Step 4 (model improvement): If $\rho_k < \eta_1$ use the model-improvement algorithm to

- attempt to certify that m_k is fully linear on $B(x_k; \Delta_k)$,
- if such a certificate is not obtained, we say that m_k is not certifiably fully linear and make one or more suitable improvement steps.

Define m_{k+1}^{icb} to be the (possibly improved) model.

Step 5 (trust-region radius update): Set

$$\Delta_{k+1}^{icb} \in \begin{cases} [\Delta_k, \min\{\gamma_{inc}\Delta_k, \Delta_{max}\}] & \text{if } \rho_k \geq \eta_1, \\ \{\gamma\Delta_k\} & \text{if } \rho_k < \eta_1 \text{ and } m_k \text{ is fully linear}, \\ \{\Delta_k\} & \text{if } \rho_k < \eta_1 \text{ and } m_k \\ & \text{is not certifiably fully linear.} \end{cases}$$

Increment k by one and go to Step 1.

The procedure invoked in the criticality step (Step 1 of Algorithm 4.1) is described in the following algorithm.

ALGORITHM 4.2 (**Criticality step: 1st order**). *This algorithm is only applied if $\|g_k^{icb}\| \leq \epsilon_c$ and at least one of the following holds: the model m_k^{icb} is not certifiably fully linear on $B(x_k; \Delta_k^{icb})$ or $\Delta_k^{icb} > \mu\|g_k^{icb}\|$. The constant $\alpha \in (0, 1)$ is chosen at Step 0 of Algorithm 4.1.*

Initialization: Set $i = 0$. Set $m_k^{(0)} = m_k^{icb}$.

Repeat Increment i by one. Use the model-improvement algorithm to improve the previous model $m_k^{(i-1)}$ until it is fully linear on $B(x_k; \alpha^{i-1}\Delta_k^{icb})$ (notice that this can be done in a finite, uniformly bounded number of steps given the choice of the model-improvement algorithm in Step 0 of Algorithm 4.1). Denote the new model by $m_k^{(i)}$. Set $\tilde{\Delta}_k = \alpha^{i-1}\Delta_k^{icb}$ and $\tilde{m}_k = m_k^{(i)}$.

Until $\tilde{\Delta}_k \leq \mu\|g_k^{(i)}\|$.

Note that if $\|g_k^{icb}\| \leq \epsilon_c$ in the criticality step of Algorithm 4.1 and Algorithm 4.2 is invoked, the model m_k is fully linear on $B(x_k; \tilde{\Delta}_k)$ with $\tilde{\Delta}_k \leq \Delta_k$. Then, by Lemma 3.2, m_k is also fully linear on $B(x_k; \Delta_k)$ (as well as on $B(x_k; \mu\|g_k\|)$).

[1] Note that Δ_k is selected to be the number in $[\tilde{\Delta}_k, \Delta_k^{icb}]$ closest to $\beta\|\tilde{g}_k\|$.

We will prove in the next section that Algorithm 4.2 terminates after a finite number of steps if $\|\nabla f(x_k)\| \neq 0$. If $\|\nabla f(x_k)\| = 0$, then we will cycle in the criticality step until some stopping criterion is met.

An analogue of this step can be found in Powell's work (e.g., [12]), and is related to improving geometry when the step s_k is much smaller than Δ_k, which occurs when the gradient of the model is small relative to the Hessian. Here we use the size of the gradient as the criticality test. Scaling with respect to the size of the Hessian is also possible, as long as arbitrarily small or large scaling factors are not allowed.

After Step 3 of Algorithm 4.1, we may have the following possible situations at each iteration:

1. $\rho_k \geq \eta_1$, hence, the new iterate is accepted and the trust-region radius is retained or increased. We will call such iterations **successful**. We will denote the set of indices of all successful iterations by $\mathcal{S}$.
2. $\eta_1 > \rho_k \geq \eta_0$ and m_k is fully linear. Hence, the new iterate is accepted and the trust-region radius is decreased. We will call such iterations **acceptable**. (There are no acceptable iterations when $\eta_0 = \eta_1 \in (0, 1)$.)
3. $\eta_1 > \rho_k$ and m_k is not certifiably fully linear. Hence, the model is improved. The new point might be included in the sample set but is not accepted as a new iterate. We will call such iterations **model-improving**.
4. $\rho_k < \eta_0$ and m_k is fully linear. This is the case when no (acceptable) decrease was obtained and there is no need to improve the model. The trust-region radius is reduced and nothing else changes. We will call such iterations **unsuccessful**.

5. Global convergence for first-order critical points. We will first show that unless the current iterate is a first-order stationary point then the algorithm will not loop infinitely in the criticality step of Algorithm 4.1 (Algorithm 4.2). The proof is very similar to the one in [3, Lemma 5.iii] but we repeat the details here for completeness.

LEMMA 5.1. *If $\nabla f(x_k) \neq 0$, Step 1 of Algorithm 4.1 will terminate in a finite number of improvement steps (by applying Algorithm 4.2).*

Proof. Assume that the loop in Algorithm 4.2 is infinite. We will show that $\nabla f(x_k)$ has to be zero in this case. At the start, we know that we do not have a certifiably fully-linear model m_k^{icb} or that the radius Δ_k^{icb} exceeds $\mu\|g_k^{icb}\|$. We then define $m_k^{(0)} = m_k^{icb}$ and the model is improved until it is fully linear on the ball $B(x_k; \alpha^0\Delta_k^{icb})$ (in a finite number of improvement steps). If the gradient $g_k^{(1)}$ of the resulting model $m_k^{(1)}$ satisfies $\mu\|g_k^{(1)}\| \geq \alpha^0\Delta_k^{icb}$, the procedure stops with

$$\tilde{\Delta}_k^{icb} \;=\; \alpha^0\Delta_k^{icb} \;\leq\; \mu\|g_k^{(1)}\|.$$

Otherwise, that is if $\mu\|g_k^{(1)}\| < \alpha^0\Delta_k^{icb}$, the model is improved until it is fully linear on the ball $B(x_k; \alpha\Delta_k^{icb})$. Then, again, either the procedure stops or the radius is again multiplied by α, and so on.

The only way for this procedure to be infinite (and to require an infinite number of improvement steps) is if

$$\mu\|g_k^{(i)}\| \;<\; \alpha^{i-1}\Delta_k^{icb},$$

for all $i \geq 1$, where $g_k^{(i)}$ is the gradient of the model $m_k^{(i)}$. This construction implies that $\lim_{i\to+\infty} \|g_k^{(i)}\| = 0$. Since each model $m_k^{(i)}$ was fully linear on $B(x_k; \alpha^{i-1}\Delta_k^{icb})$

then (3.1) with $s = 0$ and $x = x_k$ provide

$$\|\nabla f(x_k) - g_k^{(i)}\| \ \leq \ \kappa_{eg} \alpha^{i-1} \Delta_k^{icb}$$

for each $i \geq 1$. Thus, using the triangle inequality, it holds for all $i \geq 1$

$$\|\nabla f(x_k)\| \ \leq \ \|\nabla f(x_k) - g_k^{(i)}\| + \|g_k^{(i)}\| \ \leq \ \left(\kappa_{eg} + \frac{1}{\mu}\right) \alpha^{i-1} \Delta_k^{icb}.$$

Since $\alpha \in (0, 1)$, this implies that $\nabla f(x_k) = 0$. □

We will prove now the results related to global convergence to first-order critical points. For minimization we need to assume that f is bounded from below.

ASSUMPTION 5.1. *Assume f is bounded below on $L(x_0)$, that is there exists a constant κ_* such that, for all $x \in L(x_0)$, $f(x) \geq \kappa_*$.*

We will make use of the assumptions on the boundedness of f from below and on the Lipschitz continuity of the gradient of f (i.e., Assumptions 3.1 and 5.1), and of the existence of fully-linear models (Definition 3.1). For simplicity of the presentation, we also require the model Hessian $H_k = \nabla^2 m_k(x_k)$ to be uniformly bounded. In general, fully-linear models are only required to have continuous first-order derivatives (κ_{bhm} below can then be regarded as a bound on the Lipschitz constant of the gradient of these models).

ASSUMPTION 5.2. *There exists a constant $\kappa_{bhm} > 0$ such that, for all x_k generated by the algorithm,*

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

We start the main part of the analysis with the following key lemma.

LEMMA 5.2. *If m_k is fully linear on $B(x_k; \Delta_k)$ and*

$$\Delta_k \ \leq \ \min\left[\frac{1}{\kappa_{bhm}}, \frac{\kappa_{fcd}(1-\eta_1)}{4\kappa_{ef}}\right] \|g_k\|,$$

then the k-th iteration is successful.

Proof. Since

$$\Delta_k \ \leq \ \frac{\|g_k\|}{\kappa_{bhm}},$$

the fraction of Cauchy decrease condition (2.5)–(2.6) immediately gives that

$$m_k(x_k) - m_k(x_k + s_k) \ \geq \ \frac{\kappa_{fcd}}{2} \|g_k\| \min\left[\frac{\|g_k\|}{\kappa_{bhm}}, \Delta_k\right] \ = \ \frac{\kappa_{fcd}}{2} \|g_k\| \Delta_k. \tag{5.1}$$

On the other hand, since the current model is fully linear on $B(x_k; \Delta_k)$, then from the bound (3.2) on the error between the function and the model and from (5.1) we have

$$\begin{aligned} |\rho_k - 1| \ &\leq \ \left|\frac{f(x_k + s_k) - m_k(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}\right| + \left|\frac{f(x_k) - m_k(x_k)}{m_k(x_k) - m_k(x_k + s_k)}\right| \\ &\leq \ \frac{4\kappa_{ef}\Delta_k^2}{\kappa_{fcd}\|g_k\|\Delta_k} \\ &\leq \ 1 - \eta_1, \end{aligned}$$

where we have used the assumption $\Delta_k \leq \kappa_{fcd}\|g_k\|(1-\eta_1)/(4\kappa_{ef})$ to deduce the last inequality. Therefore, $\rho_k \geq \eta_1$, and iteration k is successful. □

It now follows that if the gradient of the model is bounded away from zero then so is the trust-region radius.

LEMMA 5.3. *Suppose that there exists a constant $\kappa_1 > 0$ such that $\|g_k\| \geq \kappa_1$ for all k. Then, there exists a constant $\kappa_2 > 0$ such that*

$$\Delta_k \ \geq \ \kappa_2$$

for all k.

Proof. We know from Step 1 of Algorithm 4.1 (independently of whether Algorithm 4.2 has been invoked) that

$$\Delta_k \ \geq \ \min\{\beta\|g_k\|, \Delta_k^{icb}\}.$$

Thus,

$$\Delta_k \ \geq \ \min\{\beta\kappa_1, \Delta_k^{icb}\}. \tag{5.2}$$

By Lemma 5.2 and by the assumption that $\|g_k\| \geq \kappa_1$ for all k, whenever Δ_k falls below a certain value given by

$$\bar{\kappa}_2 \ = \ \min\left[\frac{\kappa_1}{\kappa_{bhm}}, \frac{\kappa_{fcd}\kappa_1(1-\eta_1)}{4\kappa_{ef}}\right],$$

the k-th iteration has to be either successful or model improving (when it is not successful and m_k is not certifiably fully linear) and hence, from Step 5, $\Delta_{k+1}^{icb} \geq \Delta_k$. We conclude from this, (5.2), and the rules of Step 5 that $\Delta_k \geq \min\{\Delta_0^{icb}, \beta\kappa_1, \gamma\bar{\kappa}_2\} = \kappa_2$. □

We will now consider what happens when the number of successful iterations is finite.

LEMMA 5.4. *If the number of successful iterations is finite then*

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

Proof. Let us consider iterations that come after the last successful iteration. We know that we can have only a finite (uniformly bounded, say by N) number of model-improving iterations before the model becomes fully linear and, hence, there is an infinite number of iterations that are either acceptable or unsuccessful and in either case the trust region is reduced. Since there are no more successful iterations, then Δ_k is never increased for sufficiently large k. Moreover, Δ_k is decreased at least once every N iterations by a factor of γ. Thus, Δ_k converges to zero.

Now, for each j, let i_j be the index of the first iteration after the j-th iteration for which the model m_j is fully linear. Then

$$\|x_j - x_{i_j}\| \ \leq \ N\Delta_j \ \to 0$$

as j goes to $+\infty$.

Let us now observe that

$$\|\nabla f(x_j)\| \ \leq \ \|\nabla f(x_j) - \nabla f(x_{i_j})\| + \|\nabla f(x_{i_j}) - g_{i_j}\| + \|g_{i_j}\|.$$

What remains to show is that all three terms on the right-hand side are converging to zero. The first term converges to zero because of the Lipschitz continuity of ∇f and the fact that $\|x_{i_j} - x_j\| \to 0$. The second term is converging to zero because of the bound (3.1) on the error between the gradients of a fully-linear model and the function f and the fact that m_{i_j} is fully linear. Finally, the third term can be shown to converge to zero by Lemma 5.2, since if $\|g_{i_j}\|$ was bounded away from zero for a subsequence, then for small enough Δ_{i_j} (recall that $\Delta_{i_j} \to 0$), i_j would be a successful iteration, which would then yield a contradiction. □

We now prove that the trust-region radius converges to zero, which is particularly relevant in the derivative-free context.

LEMMA 5.5.

$$\lim_{k \to +\infty} \Delta_k \;=\; 0. \tag{5.3}$$

Proof. When $\mathcal{S}$ is finite the result is shown in the proof of Lemma 5.4. Let us consider the case when $\mathcal{S}$ is infinite. For any $k \in \mathcal{S}$ we have

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1[m_k(x_k) - m_k(x_k + s_k)].$$

By using the bound on the fraction of Cauchy decrease (2.6), we have

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1 \frac{\kappa_{fcd}}{2} \|g_k\| \min\left[\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right].$$

Due to Step 1 of Algorithm 4.1 we have $\|g_k\| \geq \min\{\epsilon_c, \mu^{-1}\Delta_k\}$, hence

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1 \frac{\kappa_{fcd}}{2} \min\{\epsilon_c, \mu^{-1}\Delta_k\} \min\left[\frac{\min\{\epsilon_c, \mu^{-1}\Delta_k\}}{\|H_k\|}, \Delta_k\right].$$

Since $\mathcal{S}$ is infinite and f is bounded from below, the right-hand side of the above expression has to converge to zero. Hence $\lim_{k \in \mathcal{S}} \Delta_k \;=\; 0$, and the proof is complete if all iterations are successful. Now recall that the trust-region radius can only be increased during a successful iteration, and it can only be increased by a ratio of at most γ_{inc}. Let $k \notin \mathcal{S}$ be the index of an iteration (after the first successful one). Then $\Delta_k \leq \gamma_{inc}\Delta_{s_k}$, where s_k is the index of the last successful iteration before k. Since $\Delta_{s_k} \to 0$, then $\Delta_k \to 0$, for $k \notin \mathcal{S}$. □

The following lemma now follows.

LEMMA 5.6.

$$\liminf_{k \to +\infty} \|g_k\| \;=\; 0. \tag{5.4}$$

Proof. Assume, for the purpose of deriving a contradiction, that, for all k,

$$\|g_k\| \;\geq\; \kappa_1 \tag{5.5}$$

for some $\kappa_1 > 0$. By Lemma 5.3 we have that $\Delta_k \geq \kappa_2$ for all k. We obtain a contradiction with Lemma 5.5. □

We now show that if the model gradient $\|g_k\|$ converges to zero on a subsequence then so does the true gradient $\|\nabla f(x_k)\|$.

LEMMA 5.7. *For any subsequence $\{k_i\}$ such that*

$$\lim_{i \to +\infty} \|g_{k_i}\| \;=\; 0 \tag{5.6}$$

it also holds that

$$\lim_{i \to +\infty} \|\nabla f(x_{k_i})\| \;=\; 0. \tag{5.7}$$

Proof. First we note that, by (5.6), $\|g_{k_i}\| \leq \epsilon_c$ for i sufficiently large. Thus, the mechanism of the criticality step (Step 1) ensures that the model m_{k_i} is fully linear on a ball $B(x_{k_i}; \Delta_{k_i})$ with $\Delta_{k_i} \leq \mu \|g_{k_i}\|$ for all i sufficiently large (if $\nabla f(x_{k_i}) \neq 0$). Then, using the bound (3.1) on the error between the gradients of the function and the model, we have

$$\|\nabla f(x_{k_i}) - g_{k_i}\| \;\leq\; \kappa_{eg} \Delta_{k_i} \;\leq\; \kappa_{eg} \mu \|g_{k_i}\|.$$

As a consequence, we have

$$\|\nabla f(x_{k_i})\| \;\leq\; \|\nabla f(x_{k_i}) - g_{k_i}\| + \|g_{k_i}\| \;\leq\; (\kappa_{eg}\mu + 1)\|g_{k_i}\|,$$

for all i sufficiently large. But since $\|g_{k_i}\| \to 0$ then this implies (5.7). ☐

Lemmas 5.6 and 5.7 immediately give the following global convergence result.

THEOREM 5.8. *Let Assumptions 3.1, 5.1, and 5.2 hold. Then,*

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

If the sequence of iterates is bounded then this result implies the existence of one limit point that is first-order critical. In fact we are able to prove that all limit points of the sequence of iterates are first-order critical.

THEOREM 5.9. *Let Assumptions 3.1, 5.1, and 5.2 hold. Then,*

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

Proof. Lemma 5.4 establishes that in the case when $\mathcal{S}$ is finite the theorem holds. Hence, we will assume that $\mathcal{S}$ is infinite. Suppose, for the purpose of establishing a contradiction, that there exists a subsequence $\{k_i\}$ of successful or acceptable iterations such that

$$\|\nabla f(x_{k_i})\| \;\geq\; \epsilon_0 \;>\; 0, \tag{5.8}$$

for some $\epsilon_0 > 0$ and for all i (we can ignore the other types of iterations, since x_k does not change during such iterations). Then, because of Lemma 5.7, we obtain that

$$\|g_{k_i}\| \;\geq\; \epsilon \;>\; 0,$$

for some $\epsilon > 0$ and for all i sufficiently large. Without loss of generality, we pick ϵ such that

$$\epsilon \;\leq\; \min\left\{ \frac{\epsilon_0}{2(2 + \kappa_{eg}\mu)}, \epsilon_c \right\}. \tag{5.9}$$

Lemma 5.6 then ensures the existence, for each k_i in the subsequence, of a first iteration $\ell_i > k_i$ such that $\|g_{\ell_i}\| < \epsilon$. By removing elements from $\{k_i\}$, without loss of

generality and without a change of notation, we thus obtain that there exists another subsequence indexed by $\{\ell_i\}$ such that

$$\|g_k\| \geq \epsilon \text{ for } k_i \leq k < \ell_i \text{ and } \|g_{\ell_i}\| < \epsilon, \tag{5.10}$$

for sufficiently large i, with inequality (5.8) being retained.

We now restrict our attention to the set $\mathcal{K}$ corresponding to the subsequence of iterations whose indices are in the set

$$\cup_{i\in\mathbb{N}_0}\{k \in \mathbb{N}_0 : k_i \leq k < \ell_i\},$$

where k_i and ℓ_i belong to the two subsequences given above in (5.10).

We know that $\|g_k\| \geq \epsilon$ for $k \in \mathcal{K}$. From Lemma 5.5 $\lim_{k\to+\infty} \Delta_k = 0$ and by Lemma 5.2 we conclude that for any large enough $k \in \mathcal{K}$ the iteration k is either successful, if the model is fully linear, or model improving, otherwise.

Moreover, for each $k \in \mathcal{K} \cap \mathcal{S}$ we have

$$f(x_k) - f(x_{k+1}) \geq \eta_1[m_k(x_k) - m_k(x_k + s_k)] \geq \eta_1 \frac{\kappa_{fcd}}{2}\|g_k\| \min\left[\frac{\|g_k\|}{\kappa_{bhm}}, \Delta_k\right]. \tag{5.11}$$

and for any such k large enough, $\Delta_k \leq \frac{\epsilon}{\kappa_{bhm}}$. Hence, we have for $k \in \mathcal{K}\cap\mathcal{S}$ sufficiently large,

$$\Delta_k \leq \frac{2}{\eta_1\kappa_{fcd}\epsilon}[f(x_k) - f(x_{k+1})].$$

Since for any $k \in \mathcal{K}$ large enough the iteration is either successful or model improving and since for a model improving iteration $x_k = x_{k+1}$ we have, for all i sufficiently large,

$$\|x_{k_i} - x_{\ell_i}\| \leq \sum_{\substack{j=k_i \\ j\in\mathcal{K}\cap\mathcal{S}}}^{\ell_i-1} \|x_j - x_{j+1}\| \leq \sum_{\substack{j=k_i \\ j\in\mathcal{K}\cap\mathcal{S}}}^{\ell_i-1} \Delta_j \leq \frac{2}{\eta_1\kappa_{fcd}\epsilon}[f(x_{k_i}) - f(x_{\ell_i})].$$

Since the sequence $\{f(x_k)\}$ is bounded below (Assumption 5.1) and monotonic decreasing, we see that the right-hand side of this inequality must converge to zero, and we therefore obtain that

$$\lim_{i\to+\infty} \|x_{k_i} - x_{\ell_i}\| = 0.$$

Now,

$$\|\nabla f(x_{k_i})\| \leq \|\nabla f(x_{k_i}) - \nabla f(x_{\ell_i})\| + \|\nabla f(x_{\ell_i}) - g_{\ell_i}\| + \|g_{\ell_i}\|.$$

The first term of the right-hand side tends to zero because of the Lipschitz continuity of the gradient of f (Assumption 3.1), and is thus bounded by ϵ for i sufficiently large. The third term is bounded by ϵ by (5.10). For the second term we use the fact that from (5.9) and the mechanism of the criticality step (Step 1) at iteration ℓ_i, the model m_{ℓ_i} is fully linear on $B(x_{\ell_i}; \mu\|g_{\ell_i}\|)$. Thus, using using (3.1) and (5.10), we also deduce that the second term is bounded by $\kappa_{eg}\mu\epsilon$ (for i sufficiently large). As a consequence, we obtain from these bounds and (5.9) that

$$\|\nabla f(x_{k_i})\| \leq (2 + \kappa_{eg}\mu)\epsilon \leq \tfrac{1}{2}\epsilon_0$$

for i large enough, which contradicts (5.8). Hence our initial assumption must be false and the theorem follows. □

REMARK 5.1. *This last theorem is the only result for which we need to use the fact that $x_k = x_{k+1}$ at the model-improving iterations. So, this requirement could be lifted from the algorithm if only a liminf-type result is desired. The advantage of this is that it becomes possible to accept simple decrease in the function value even when the model is not fully linear. The disadvantage, aside from the weaker convergence result, is in the inherent difficulty of producing fully-linear models after at most N consecutive model-improvement steps when the region where each such model has to be fully linear can change at each iteration.*

6. Derivative-free trust-region methods (second order). In order to achieve global convergence to second-order critical points, the algorithm must attempt to drive to zero a quantity that expresses second-order stationarity. Following [2, Section 9.3], one possibility is to work with

$$\sigma_k^m \;=\; \max\left\{\|g_k\|, -\lambda_{\min}(H_k)\right\},$$

which measures the second-order stationarity of the model.

The algorithm follows mostly the same arguments as those of Algorithm 4.1. One fundamental difference is that σ_k^m now plays the role of $\|g_k\|$. Another is the need to work with fully-quadratic models. A third main modification is the need to be able to solve the trust-region subproblem better, so that the step yields both a fraction of Cauchy decrease and a fraction of the eigenstep decrease when negative curvature is present. Finally, to prove the lim-type convergence result in the second-order case, we also need to increase the trust-region radius on some of the successful iterations, whereas in the first-order case that was optional. Unlike the case of traditional trust-region methods [2, Page 158] that seek second-order convergence results we do not increase the trust-region radius on *every* successful iteration. We only insist on such an increase when the size of the trust-region radius is small when compared to the measure of stationarity.

We state the version of the algorithm that we consider.

ALGORITHM 6.1 (**Derivative-free trust-region method (2nd order)**).

Step 0 (initialization): Choose a fully-quadratic class of models $\mathcal{M}$ and a corresponding model-improvement algorithm (see, e.g., [4]). Choose an initial point x_0 and $\Delta_{max} > 0$. We assume that an initial model m_0^{icb} $m_0^{icb}(x_0 + s)$ (with gradient and Hessian at $s = 0$ given by g_0^{icb} and H_0^{icb}, respectively), with $\sigma_0^{m,icb} = \max\{\|g_0^{icb}\|, -\lambda_{\min}(H_0^{icb})\}$, and a trust-region radius $\Delta_0^{icb} \in (0, \Delta_{max}]$ are given.
The constants η_0, η_1, γ, γ_{inc} ϵ_c, β, μ, and α are also given and satisfy the conditions $0 \leq \eta_0 \leq \eta_1 < 1$ (with $\eta_1 \neq 0$), $0 < \gamma < 1 < \gamma_{inc}$, $\epsilon_c > 0$, $\mu > \beta > 0$, and $\alpha \in (0, 1)$. Set $k = 0$.

Step 1 (criticality step): If $\sigma_k^{m,icb} > \epsilon_c$ then $m_k = m_k^{icb}$ and $\Delta_k = \Delta_k^{icb}$.
If $\sigma_k^{m,icb} \leq \epsilon_c$ then proceed as follows. Call the model-improvement algorithm to attempt to certify if the model m_k^{icb} is fully quadratic on $B(x_k; \Delta_k^{icb})$. If at least one of the following conditions holds,
- the model m_k^{icb} is not certifiably fully quadratic on $B(x_k; \Delta_k^{icb})$,
- $\Delta_k^{icb} > \mu\sigma_k^{m,icb}$,

then apply Algorithm 6.2 (described below) to construct a model $\tilde{m}_k(x_k + s)$ (with gradient and Hessian at $s = 0$ given by $\tilde{g}_k$ and $\tilde{H}_k$, respectively),

with $\tilde{\sigma}_k^m = \max\{\|\tilde{g}_k\|, -\lambda_{\min}(\tilde{H}_k)\}$, which is fully quadratic (for some constants κ_{ef}, κ_{eg}, κ_{eh}, and κ_{blh}, which remain the same for all iterations of Algorithm 6.1) on the ball $B(x_k; \tilde{\Delta}_k)$ for some $\tilde{\Delta}_k \in (0, \mu\tilde{\sigma}_k^m]$ given by Algorithm 6.2. In such a case set[2]

$$m_k = \tilde{m}_k \text{ and } \Delta_k = \min\{\max\{\tilde{\Delta}_k, \beta\tilde{\sigma}_k^m\}, \Delta_k^{icb}\}.$$

Otherwise set $m_k = m_k^{icb}$ and $\Delta_k = \Delta_k^{icb}$.

Step 2 (step calculation): Compute a step s_k that sufficiently reduces the model m_k (in the sense of (2.9)) and such that $x_k + s_k \in B(x_k; \Delta_k)$.

Step 3 (acceptance of the trial point): Compute $f(x_k + s_k)$ and define

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

If $\rho_k \geq \eta_1$ or if both $\rho_k \geq \eta_0$ and the model is fully quadratic (for the positive constants κ_{ef}, κ_{eg}, κ_{eh}, and κ_{blh}) on $B(x_k; \Delta_k)$, then $x_{k+1} = x_k + s_k$ and the model is updated to include the new iterate into the sample set resulting in a new model m_{k+1}^{icb} (with gradient and Hessian at $s = 0$ given by g_{k+1}^{icb} and H_{k+1}^{icb}, respectively), with $\sigma_{k+1}^{m,icb} = \max\{\|g_{k+1}^{icb}\|, -\lambda_{\min}(H_{k+1}^{icb})\}$; otherwise the model and the iterate remain unchanged ($m_{k+1}^{icb} = m_k$ and $x_{k+1} = x_k$).

Step 4 (model improvement): If $\rho_k < \eta_1$ use the model-improvement algorithm to

- attempt to certify that m_k is fully quadratic on $B(x_k; \Delta_k)$,
- if such a certificate is not obtained, we say that m_k is not certifiably fully quadratic and make one or more suitable improvement steps.

Define m_{k+1}^{icb} to be the (possibly improved) model.

Step 5 (trust-region radius update): Set

$$\Delta_{k+1}^{icb} \in \begin{cases} \{\min\{\gamma_{inc}\Delta_k, \Delta_{max}\}\} & \text{if } \rho_k \geq \eta_1 \text{ and } \Delta_k < \beta\sigma_k^m, \\ [\Delta_k, \min\{\gamma_{inc}\Delta_k, \Delta_{max}\}] & \text{if } \rho_k \geq \eta_1 \text{ and } \Delta_k \geq \beta\sigma_k^m, \\ \{\gamma\Delta_k\} & \text{if } \rho_k < \eta_1 \text{ and } m_k \\ & \text{is fully quadratic,} \\ \{\Delta_k\} & \text{if } \rho_k < \eta_1 \text{ and } m_k \\ & \text{is not certifiably fully quadratic.} \end{cases}$$

Increment k by one and go to Step 1.

We need to recall for Algorithm 6.1 the definitions of **successful**, **acceptable**, **model-improving**, and **unsuccessful** iterations which we stated for the sequence of iterations generated by Algorithm 4.1. We will use the same definitions here, adapted to the quadratic models. We denote the set of all successful iterations by $\mathcal{S}$ and the set of all such iterations when $\Delta_k < \beta\sigma_k^m$ by $\mathcal{S}_+$.

As in the first-order case, during a model-improvement step, Δ_k and x_k remain unchanged, hence there can only be a finite number of model-improvement steps before a fully-quadratic model is obtained. The comments outlined in Remark 5.1 about possibly changing x_k at any model-improving iteration, suitably modified, apply in the fully-quadratic case as well.

The criticality step can be implemented following a procedure similar to the one described in Algorithm 4.2, essentially by replacing $\|g_k\|$ by σ_k^m and by using fully-quadratic models rather than fully-linear ones.

[2]Note that Δ_k is selected to be the number in $[\tilde{\Delta}_k, \Delta_k^{icb}]$ closest to $\beta\|\tilde{\sigma}_k^m\|$.

ALGORITHM 6.2 (**Criticality step: 2nd order**). *This algorithm is only applied if* $\sigma_k^{m,icb} \leq \epsilon_c$ *and at least one the following holds: the model* m_k^{icb} *is not certifiably fully quadratic on* $B(x_k; \Delta_k^{icb})$ *or* $\Delta_k^{icb} > \mu\sigma_k^{m,icb}$. *The constant* $\alpha \in (0,1)$ *is chosen at Step 0 of Algorithm 6.1.*

Initialization: Set $i = 0$. Set $m_k^{(0)} = m_k^{icb}$.

Repeat Increment i by one. Improve the previous model $m_k^{(i-1)}$ until it is fully quadratic on $B(x_k; \alpha^{i-1}\Delta_k^{icb})$ (notice that this can be done in a finite, uniformly bounded number of steps, given the choice of the model-improvement algorithm in Step 0 of Algorithm 6.1). Denote the new model by $m_k^{(i)}$. Set $\tilde{\Delta}_k = \alpha^{i-1}\Delta_k^{icb}$ and $\tilde{m}_k = m_k^{(i)}$.

Until $\tilde{\Delta}_k \leq \mu(\sigma_k^m)^{(i)}$.

Note that if $\sigma_k^{m,icb} \leq \epsilon_c$ in the criticality step of Algorithm 6.1 and Algorithm 6.2 is invoked, the new model m_k is fully quadratic on $B(x_k; \tilde{\Delta}_k)$ with $\tilde{\Delta}_k \leq \Delta_k$. Then, by Lemma 3.4, m_k is also fully quadratic on $B(x_k; \Delta_k)$ (as well as on $B(x_k; \mu\sigma_k^m)$).

7. Global convergence for second-order critical points. For global convergence to second-order critical points, we will need one more order of smoothness, namely Assumption 3.2 on the Lipschitz continuity of the Hessian of f. It will be also necessary to assume that the function f is bounded from below (Assumption 5.1). Naturally, we will also assume the existence of fully-quadratic models.

We start by introducing the notation

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

and

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

It will be important to bound the difference between the true $\sigma(x)$ and the model $\sigma^m(x)$. For that purpose, we first derive a bound on the difference between the smallest eigenvalues of a function and of a corresponding fully-quadratic model.

PROPOSITION 7.1. *Suppose that Assumption 3.2 holds and* m *is a fully-quadratic model on* $B(x; \Delta)$. *Then,*

$$|\lambda_{\min}(\nabla^2 f(x)) - \lambda_{\min}(\nabla^2 m(x))| \;\leq\; \kappa_{eh}\Delta.$$

Proof. The proof follows directly from the bound (3.3) on the error between the Hessians of m and f and the simple observation that if v is a normalized eigenvector corresponding to the smallest eigenvalue of $\nabla^2 m(x)$ then

$$\begin{aligned}\lambda_{\min}(\nabla^2 f(x)) - \lambda_{\min}(\nabla^2 m(x)) &\leq v^\top[\nabla^2 f(x) - \nabla^2 m(x)]v \\ &\leq \|\nabla^2 f(x) - \nabla^2 m(x)\| \\ &\leq \kappa_{eh}\,\Delta.\end{aligned}$$

Analogously, letting v be a normalized eigenvector corresponding to the smallest eigenvalue of $\nabla^2 f(x)$, we obtain

$$\lambda_{\min}(\nabla^2 m(x)) - \lambda_{\min}(\nabla^2 f(x)) \;\leq\; \kappa_{eh}\,\Delta$$

and the result follows. ☐

The following lemma shows that the difference between the true $\sigma(x)$ and the model $\sigma^m(x)$ is of the order of Δ.

LEMMA 7.2. *Let Δ be bounded by Δ_{max}. Suppose that Assumption 3.2 holds and m is a fully-quadratic model on $B(x;\Delta)$. Then, we have that*

$$|\sigma(x) - \sigma^m(x)| \;\leq\; \kappa_\sigma \Delta, \tag{7.1}$$

for some $\kappa_\sigma > 0$.

Proof. It follows that

$$\begin{aligned} |\sigma(x) - \sigma^m(x)| &= \big|\max\left\{\|\nabla f(x)\|, \max\{-\lambda_{\min}(\nabla^2 f(x)), 0\}\right\} \\ &\quad - \max\left\{\|\nabla m(x)\|, \max\{-\lambda_{\min}(\nabla^2 m(x)), 0\}\right\}\big| \\ &\leq \max\left\{\big|\, \|\nabla f(x)\| - \|\nabla m(x)\| \,\big|, \right. \\ &\quad \left. \big|\max\{-\lambda_{\min}(\nabla^2 f(x)), 0\} - \max\{-\lambda_{\min}(\nabla^2 m(x)), 0\}\big|\right\}. \end{aligned}$$

The first argument $|\, \|\nabla f(x)\| - \|\nabla m(x)\| \,|$ is bounded above by $\kappa_{eg}\Delta_{max}\Delta$, because of the error bound (3.4) between the gradients of f and m, and from the bound $\Delta \leq \Delta_{max}$. The second argument is clearly dominated by $|\lambda_{\min}(\nabla^2 f(x)) - \lambda_{\min}(\nabla^2 m(x))|$, which is bounded above by $\kappa_{eh}\Delta$ because of Proposition 7.1. Finally we need only to write $\kappa_\sigma = \max\{\kappa_{eg}\Delta_{max}, \kappa_{eh}\}$ and the result follows. ☐

The convergence theory will require the already mentioned Assumptions 3.2, and 5.1, as well as the uniform upper bound on the Hessians of the quadratic models (Assumption 5.2).

As for the first-order case, we begin by noting that the criticality step can be successfully executed in a finite number of improvement steps.

LEMMA 7.3. *If $\sigma(x_k) \neq 0$, Step 1 of Algorithm 6.1 will terminate in a finite number of improvement steps (by applying Algorithm 6.2).*

Proof. The proof is analogous to the proof of Lemma 5.1, with $\|g_k^{(i)}\|$ replaced by $(\sigma_k^m)^{(i)}$ and $\nabla f(x_k)$ replaced by $\sigma(x_k)$. ☐

We now show that an iteration must be successful if the current model is fully quadratic and the trust-region radius is small enough with respect to σ_k^m.

LEMMA 7.4. *If m_k is fully quadratic on $B(x_k;\Delta_k)$ and*

$$\Delta_k \;\leq\; \min\left[\frac{1}{\kappa_{bhm}}, \frac{\kappa_{fod}(1-\eta_1)}{4\kappa_{ef}\Delta_{max}}, \frac{\kappa_{fod}(1-\eta_1)}{4\kappa_{ef}}\right]\sigma_k^m,$$

then the k-th iteration is successful.

Proof. The proof is similar to the proof of Lemma 5.2 for the first-order case, except that now we need to take the second-order terms into account.

First we recall the fractions of Cauchy and eigenstep decreases (2.10), which provide

$$m_k(x_k) - m_k(x_k + s_k) \geq \tfrac{\kappa_{fod}}{2} \max\left\{\|g_k\| \min\left[\tfrac{\|g_k\|}{\kappa_{bhm}}, \Delta_k\right], -\tau_k \Delta_k^2\right\}.$$

From the expression for σ_k^m, one of the two cases has to hold: either $\|g_k\| = \sigma_k^m$ or $-\tau_k = -\lambda_{\min}(H_k) = \sigma_k^m$.

In the first case, using the fact that $\Delta_k \leq \sigma_k^m/\kappa_{bhm}$, we conclude that

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \frac{\kappa_{fod}}{2}\|g_k\|\Delta_k \;=\; \frac{\kappa_{fod}}{2}\sigma_k^m \Delta_k. \tag{7.2}$$

On the other hand, since the current model is fully quadratic on $B(x_k;\Delta_k)$, we may deduce from (7.2) and the bound (3.5) on the error between the model m_k and f that

$$\begin{aligned} |\rho_k - 1| &\leq \left|\frac{f(x_k+s_k)-m_k(x_k+s_k)}{m_k(x_k)-m_k(x_k+s_k)}\right| + \left|\frac{f(x_k)-m_k(x_k)}{m_k(x_k)-m_k(x_k+s_k)}\right| \\ &\leq \frac{4\kappa_{ef}\Delta_k^3}{(\kappa_{fod}\sigma_k^m)\Delta_k} \\ &\leq \frac{4\kappa_{ef}\Delta_{max}}{\kappa_{fod}\sigma_k^m}\Delta_k \\ &\leq 1-\eta_1. \end{aligned}$$

In the case when $-\tau_k = \sigma_k^m$, we first write

$$m_k(x_k) - m_k(x_k+s_k) \;\geq\; -\tfrac{\kappa_{fod}}{2}\tau_k\Delta_k^2 \;=\; \tfrac{\kappa_{fod}}{2}\sigma_k^m\Delta_k^2. \tag{7.3}$$

But, since the current model is fully quadratic on $B(x_k;\Delta_k)$, we deduce from (7.3) and the bound (3.5) on the error between m_k and f that

$$\begin{aligned} |\rho_k - 1| &\leq \left|\frac{f(x_k+s_k)-m_k(x_k+s_k)}{m_k(x_k)-m_k(x_k+s_k)}\right| + \left|\frac{f(x_k)-m_k(x_k)}{m_k(x_k)-m_k(x_k+s_k)}\right| \\ &\leq \frac{4\kappa_{ef}\Delta_k^3}{(\kappa_{fod}\sigma_k^m)\Delta_k^2} \\ &\leq 1-\eta_1. \end{aligned}$$

In either case $\rho_k \geq \eta_1$ and iteration k is, thus, successful. □

As in the first-order case, the following result follows readily from Lemma 7.4.

LEMMA 7.5. *Suppose that there exists a constant $\kappa_1 > 0$ such that $\sigma_k^m \geq \kappa_1$ for all k. Then, there exists a constant $\kappa_2 > 0$ such that*

$$\Delta_k \;\geq\; \kappa_2$$

for all k.

Proof. The proof is trivially derived by combining Lemma 7.4 and the proof of Lemma 5.3. □

We are now able to show that if there are only finitely many successful iterations then we approach a second-order stationary point.

LEMMA 7.6. *If the number of successful iterations is finite then*

$$\lim_{k\to+\infty} \sigma(x_k) \;=\; 0.$$

Proof. The proof of this lemma is virtually identical to that of Lemma 5.4 for the first-order case, with $\|g_k\|$ being substituted by σ_k^m and $\|\nabla f(x_k)\|$ being substituted by $\sigma(x_k)$ and by using Lemmas 7.2 and 7.4. □

We now prove that the whole sequence of trust-region radii converges to zero.

LEMMA 7.7.

$$\lim_{k\to+\infty} \Delta_k \;=\; 0. \tag{7.4}$$

Proof. When $\mathcal{S}$ is finite the proof is as in the proof of Lemma 5.4 (the argument is exactly the same). Let us consider the case when $\mathcal{S}$ is infinite. For any $k \in \mathcal{S}$ we have

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \eta_1 [m(x_k) - m(x_k + s_k)] \\ &\geq \eta_1 \tfrac{\kappa_{fod}}{2} \max\left\{ \|g_k\| \min\left[\tfrac{\|g_k\|}{\kappa_{bhm}}, \Delta_k \right], -\tau_k \Delta_k^2 \right\}. \end{aligned}$$

Due to Step 1 of Algorithm 6.1 we have that $\sigma_k^m \geq \min\{\epsilon_c, \mu^{-1}\Delta_k\}$. If on iteration k $\|g_k\| \geq \max\{-\tau_k, 0\} = \{-\lambda_{\min}(H_k), 0\}$, then $\sigma_k^m = \|g_k\|$ and

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \frac{\kappa_{fod}}{2} \min\{\epsilon_c, \mu^{-1}\Delta_k\} \min\left[\frac{\min\{\epsilon_c, \mu^{-1}\Delta_k\}}{\kappa_{bhm}}, \Delta_k \right]. \tag{7.5}$$

If, on the other hand, $\|g_k\| < -\tau_k$, then $\sigma_k^m = -\tau_k$ and

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \frac{\kappa_{fod}}{2} \min\{\epsilon_c, \mu^{-1}\Delta_k\} \Delta_k^2. \tag{7.6}$$

There are two subsequences of successful iterations, possibly overlapping, $\{k_i^1\}$, for which (7.5) holds, and $\{k_i^2\}$, for which (7.6) holds. The union of these subsequences contains all successful iterations. Since $\mathcal{S}$ is infinite and f is bounded from below, then either the corresponding subsequence $\{k_i^1\}$ (resp. $\{k_i^2\}$) is finite or the right-hand side of (7.5) (resp. (7.6)) has to converge to zero. Hence $\lim_{k \in \mathcal{S}} \Delta_k = 0$, and the proof is complete if all iterations are successful. Now recall that the trust-region radius can only be increased during a successful iteration, and it can only be increased by a ratio of at most γ_{inc}. Let $k \notin \mathcal{S}$ be the index of an iteration (after the first successful one). Then $\Delta_k \leq \gamma_{inc}\Delta_{s_k}$, where s_k is the index of the last successful iteration before k. Since $\Delta_{s_k} \to 0$, then $\Delta_k \to 0$, for $k \notin \mathcal{S}$. □

We obtain the following lemma as a simple corollary.

LEMMA 7.8.

$$\liminf_{k \to +\infty} \sigma_k^m = 0.$$

Proof. Assume, for the purpose of deriving a contradiction, that, for all k,

$$\sigma_k^m \geq \kappa_1$$

for some $\kappa_1 > 0$. Then by Lemma 7.5 there exists a constant κ_2 such that $\Delta_k \geq \kappa_2$ for all k. We obtain contradiction with Lemma 7.7. □

We now verify that the criticality step (Step 1 of Algorithm 6.1) ensures that a subsequence of the iterates approach second-order stationarity, by means of the following auxiliary result.

LEMMA 7.9. *For any subsequence $\{k_i\}$ such that*

$$\lim_{i \to +\infty} \sigma_{k_i}^m = 0 \tag{7.7}$$

it also holds that

$$\lim_{i \to +\infty} \sigma(x_{k_i}) = 0. \tag{7.8}$$

Proof. From (7.7), $\sigma_{k_i}^m \leq \epsilon_c$ for i sufficiently large. The mechanism of the criticality step (Step 1) ensures then that the model m_{k_i} is fully quadratic on the ball $B(x_{k_i}; \Delta_{k_i})$ with $\Delta_{k_i} \leq \mu\sigma_{k_i}^m$ for all i sufficiently large (if $\sigma_{k_i}^m \neq 0$). Now, using (7.1),

$$\sigma(x_{k_i}) \;=\; \big(\sigma(x_{k_i}) - \sigma_{k_i}^m\big) + \sigma_{k_i}^m \;\leq\; (\kappa_\sigma \mu + 1)\sigma_{k_i}^m.$$

The limit (7.7) and this last bound then give (7.8). □

Lemmas 7.8 and 7.9 immediately give the following global convergence result.

THEOREM 7.10. *Let Assumptions 3.2, 5.1, and 5.2 hold. Then,*

$$\liminf_{k \to +\infty} \sigma(x_k) \;=\; 0.$$

If the sequence of iterates is bounded this result implies the existence of at least one limit point that is second-order critical. We are, in fact, able to prove that all limit points of the sequence of iterates are second-order critical. In this proof we make use of the additional requirement on Step 5, which imposes in successful iterations an increase on the trust-region radius Δ_k if it is too small compared to σ_k^m.

THEOREM 7.11. *Let Assumptions 3.2, 5.1, and 5.2 hold. Then,*

$$\lim_{k \to +\infty} \sigma(x_k) \;=\; 0.$$

Proof. Lemma 7.6 establishes that in the case when $\mathcal{S}$ is finite the theorem holds. Hence, we will assume that $\mathcal{S}$ is infinite. Suppose, for the purpose of establishing a contradiction, that there exists a subsequence $\{k_i\}$ of successful or acceptable iterations such that

$$\sigma(x_{k_i}) \;\geq\; \epsilon_0 \;>\; 0, \tag{7.9}$$

for some $\epsilon_0 > 0$ and for all i (as in the first-order case, we can ignore the other iterations, since x_k does not change during such iterations). Then, because of Lemma 7.9, we obtain that

$$\sigma_{k_i}^m \;\geq\; \epsilon \;>\; 0,$$

for some $\epsilon > 0$ and for all i sufficiently large. Without loss of generality, we pick ϵ such that

$$\epsilon \;\leq\; \min\left\{ \frac{\epsilon_0}{2(2 + \kappa_\sigma \mu)}, \epsilon_c \right\}. \tag{7.10}$$

Lemma 7.8 then ensures the existence, for each k_i in the subsequence, of a first successful or acceptable iteration $\ell_i > k_i$ such that $\sigma_{\ell_i}^m < \epsilon$. By removing elements from $\{k_i\}$, without loss of generality and without a change of notation, we thus obtain that there exists another subsequence indexed by $\{\ell_i\}$ such that

$$\sigma_k^m \;\geq\; \epsilon \;\text{ for }\; k_i \leq k < \ell_i \;\text{ and }\; \sigma_{\ell_i}^m \;<\; \epsilon, \tag{7.11}$$

for sufficiently large i, with inequality (7.9) being retained.

We now restrict our attention to the set $\mathcal{K}$ which is defined as the subsequence of iterations whose indices are in the set

$$\cup_{i \in \mathbb{N}_0} \{k \in \mathbb{N}_0 : \; k_i \leq k < \ell_i\},$$

where k_i and ℓ_i belong to the two subsequences defined above in (7.11).

From Lemmas 7.4 and 7.7, just as in the proof of Theorem 5.9, it follows that for large enough $k \in \mathcal{K}$ the k-th iteration is either successful, if the model is fully linear, or model improving, otherwise, i.e., that there is only a finite number of acceptable iterations in $\mathcal{K}$.

Let us now consider the situation where an index k is in $\mathcal{K} \cap \mathcal{S} \setminus \mathcal{S}_+$. In this case, $\Delta_k \geq \beta\sigma_k^m \geq \beta\epsilon$. It immediately follows from $\Delta_k \to 0$ for $k \in \mathcal{K}$ that $\mathcal{K} \cap \mathcal{S} \setminus \mathcal{S}_+$ contains only a finite number of iterations. Hence, $k \in \mathcal{K} \cap \mathcal{S}$ is also in $\mathcal{S}_+$ when k is sufficiently large.

Let us now show that for $k \in \mathcal{K} \cap \mathcal{S}_+$ sufficiently large it holds that $\Delta_{k+1} = \gamma_{inc}\Delta_k$ (when the last successful iteration in $[k_i, \ell_i - 1]$ occurs before $\ell_i - 1$). We know that since $k \in \mathcal{S}_+$, then $\Delta_{k+1}^{icb} = \gamma_{inc}\Delta_k$ after execution of Step 5. However, Δ_{k+1}^{icb} may be reduced during Step 1 of the $k+1$-st iteration (or any subsequent iteration). By examining the assignments at the end of Step 1, we see that on any iteration $k+1 \in \mathcal{K}$, the radius Δ_{k+1}^{icb} is only reduced when $\Delta_{k+1} \geq \beta\tilde{\sigma}_{k+1}^m = \beta\sigma_{k+1}^m \geq \beta\epsilon$, but this can only happen a finite number of times, due to the fact that $\Delta_k \to 0$. Hence for large enough $k \in \mathcal{K} \cap \mathcal{S}_+$, we obtain $\Delta_{k+1} = \gamma_{inc}\Delta_k$.

Let $\mathcal{S}_+^i = [k_i, \ell_i - 1] \cap \mathcal{S}_+ = \{j_i^1, j_i^2, \ldots, j_i^*\}$ be the set of all indices of the successful iterations that fall in the interval $[k_i, \ell_i - 1]$. From the scheme that updates Δ_k at successful iterations, and from the fact that $x_k = x_{k+1}$ and $\Delta_{k+1} = \Delta_k$ for model improving steps, we can deduce that, for i large enough,

$$\|x_{k_i} - x_{\ell_i}\| \;\leq\; \sum_{j \in \mathcal{S}_+^i} \Delta_j \;\leq\; \sum_{j \in \mathcal{S}_+^i} (1/\gamma_{inc})^{j_i^* - j} \Delta_{j_i^*} \leq \frac{\gamma_{inc}}{\gamma_{inc} - 1} \Delta_{j_i^*}.$$

Thus, from the fact that $\Delta_{j_i^*} \to 0$, we conclude that $\|x_{k_i} - x_{\ell_i}\| \to 0$. We therefore obtain that

$$\lim_{i \to +\infty} \|x_{k_i} - x_{\ell_i}\| = 0.$$

Now,

$$\sigma(x_{k_i}) \;=\; (\sigma(x_{k_i}) - \sigma(x_{\ell_i})) + \left(\sigma(x_{\ell_i}) - \sigma_{\ell_i}^m\right) + \sigma_{\ell_i}^m.$$

The first term of the right-hand side tends to zero because of the Lipschitz continuity of $\sigma(x)$, and is thus bounded by ϵ for i sufficiently large. The third term is at most ϵ by (7.11). For the second term we use the fact that from (7.10) and the mechanism of the criticality step (Step 1) at iteration ℓ_i, the model m_{ℓ_i} is fully quadratic on $B(x_{\ell_i}; \mu\sigma_{\ell_i}^m)$. Using (7.1) and (7.11), we also deduce that the second term is bounded by $\kappa_\sigma\mu\epsilon$ (for i sufficiently large). As a consequence, we obtain from these bounds and (7.10) that

$$\sigma(x_{k_i}) \;\leq\; (2 + \kappa_\sigma\mu)\epsilon \;\leq\; \tfrac{1}{2}\epsilon_0$$

for i large enough, which contradicts (7.9). Hence our initial assumption must be false and the theorem follows. □

8. Acknowledgement. We are grateful to Michael Powell for many helpful comments on the earlier version of the manuscript. In particular his observations led to an improved version of Step 1, to Lemma 3.2, and to the insertion of Lemma 5.5 which helped to shorten and strengthen the proof of the lim-result. We are also grateful to the other anonymous referee for numerous helpful suggestions.

REFERENCES

[1] B. Colson and Ph. L. Toint, *Optimizing partially separable functions without derivatives*, Optim. Methods Softw., 20 (2005), pp. 493–508.

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

[3] A. R. Conn, K. Scheinberg, and Ph. L. Toint, *On the convergence of derivative-free methods for unconstrained optimization*, in Approximation Theory and Optimization, Tributes to M. J. D. Powell, edited by M. D. Buhmann and A. Iserles, Cambridge University Press, Cambridge, 1997, pp. 83–108.

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

[5] ———, *Geometry of sample sets in derivative-free optimization: Polynomial regression and underdetermined interpolation*, IMA J. Numer. Anal., 28 (2008), pp. 721–748.

[6] ———, *Introduction to Derivative-Free Optimization*, MPS-SIAM Series on Optimization, Philadelphia, 2009.

[7] M. Marazzi and J. Nocedal, *Wedge trust region methods for derivative free optimization*, Math. Program., 91 (2002), pp. 289–300.

[8] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, Berlin, second ed., 2006.

[9] M. J. D. Powell, *A new algorithm for unconstrained optimization*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, New York, 1970.

[10] ———, *Direct search algorithms for optimization calculations*, Acta Numerica, 7 (1998), pp. 287–336.

[11] ———, *On the Lagrange functions of quadratic models that are defined by interpolation*, Optim. Methods Softw., 16 (2001), pp. 289–309.

[12] ———, *On trust region methods for unconstrained minimization without derivatives*, Math. Program., 97 (2003), pp. 605–623.

[13] S. W. Thomas, *Sequential Estimation Techniques for Quasi-Newton Algorithms*, PhD thesis, Cornell University, Ithaca, New York, 1975.

[14] Y.-X. Yuan, *An example of non-convergence of trust region algorithms*, in Advances in Nonlinear Programming, Y.-X. Yuan, ed., Kluwer Academic, Dordercht, 1998, pp. 205–215.