Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 06–49

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

A. R. CONN, K. SCHEINBERG AND L. N. VICENTE

Abstract: In this paper we prove global convergence for first and second-order stationarity points of a class of derivative-free trust-region methods for unconstrained optimization. These methods are based on the sequential minimization of linear or quadratic 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.

Keywords: Trust-region methods, derivative-free optimization, nonlinear optimization, global convergence.
AMS Subject Classification (2000): 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 lead 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 problems 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]).

Date: October 26, 2006.
Support for this work was provided by Centro de Matemática da Universidade de Coimbra and by FCT under grant POCI/59442/MAT/2004.

One problem feature which frequently occurs 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 partial 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 partial derivatives are unavailable and hard to approximate directly, is a relatively recent topic [1, 3, 7, 10].

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 the specifics of model building, in fact is it not even required that these models are polynomial functions (as long as Cauchy and minimal eigenvalue decreases can be extracted from the trust-region subproblems), although they typically are. Instead, it is required that the derivative-free models have an 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 has been rigorously shown in [5, 4] how such *fully linear* and *fully quadratic* models can be constructed in the context of polynomial interpolation or regression.

In recent years there has 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 methods in [7, 10], 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 classical approach in derivative based methods and from the approach suggested in [2]. In derivative based methods the trust region becomes “irrelevant” when the iterates converge to a stationary point, 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 [10] suggests to use two different trust regions to separate these two roles. However, it 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, to ensure 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. On the other hand, when the measure of stationarity is large (the current iterate is far from a stationary point), then the trust region should be made comparably large to allow large steps. In particular, we observe 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 second-order global convergence. The trust region needs to be increased only when it is much smaller than the measure of stationarity.

Another new 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], classical 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 successful iterations (as long as it is not decreased). However, to derive the classical lim-type global convergence result [11] in the absence of sufficient decrease, some additional requirement must be imposed on the update of the trust-region radius at successful iterations, to avoid a cycling effect of the type described in [12]. The requirement that we use is that the trust-region radius is never increased except for the situations described in the previous paragraph (i.e., when the trust-region size is small compared to the measure of stationarity). We then show that even though the steps with simple decrease in the objective function are allowed, eventually they do not occur.

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 if not fully linear or fully quadratic (depending on the order of optimality desired) in a given iteration then the new iterate can be accepted only if it provides sufficient decrease in the objective function. This modification slightly complicates the convergence analysis, but it reflects the typical implementation of a trust-region derivative-free algorithm much better.

Finally, 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] that such analysis can be simply derived from the classical analysis for the derivative based case. However, as we mentioned 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.

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 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).

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 at least a reasonable approximation to a fully-linear model 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

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

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

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

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

$$\min_{s \in B(0;\Delta_k)} m_k(x_k + s), \tag{2}$$

where $m_k(x_k + s)$ is the model for the objective function given at (1) and $B(0; \Delta_k)$ is our trust region, centered at x_k with radius Δ_k, and now expressed in terms of $s = x - x_k$.

The Cauchy step. If we define

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

then the Cauchy step is a step given by

$$s_k^{\mathrm{C}} = -t_k^{\mathrm{C}} g_k. \tag{3}$$

A fundamental result that drives trust-region methods to first-order criticality is stated below (see [2]).

Theorem 2.1. *Consider the model (1) and the Cauchy step (3). Then,*

$$m_k(x_k) - m_k(x_k + s_k^C) \;\geq\; \tfrac{1}{2}\|g_k\| \min\left[\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right], \tag{4}$$

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,*

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \kappa_{fcd}\left[m_k(x_k) - m_k(x_k + s_k^C)\right], \tag{5}$$

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

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \frac{\kappa_{fcd}}{2}\|g_k\| \min\left[\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right]. \tag{6}$$

If $m_k(x_k + s)$ is not a linear or a quadratic function then Theorem 2.1 may no longer hold. Such models can be used in our framework if Assumption 2.1 is modified to directly impose (6) instead of (5), where now g_k and H_k are the gradient and Hessian of those models taken at $s = 0$.

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 τ_k be the most negative eigenvalue of H_k. In this case, we can determine a step of negative curvature s_k^{E}, such that

$$(s_k^{\mathrm{E}})^\top(g_k) \;\leq\; 0, \quad \|s_k^{\mathrm{E}}\| \;=\; \Delta_k, \quad \text{and} \quad (s_k^{\mathrm{E}})^\top H_k (s_k^{\mathrm{E}}) \;=\; \tau_k \Delta_k^2. \tag{7}$$

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 (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. The eigenstep induces the following decrease in the model (the proof is trivial and omitted).

Lemma 2.1. *Suppose that the model Hessian H_k has negative eigenvalues. Then we have that*

$$m_k(x_k) - m_k(x_k + s_k^E) \;\geq\; -\tfrac{1}{2}\tau_k \Delta_k^2. \tag{8}$$

The eigenstep plays a role similar to that of the Cauchy step, in that 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,*

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \kappa_{fod}\left[m_k(x_k) - \min\{m_k(x_k + s_k^C), m_k(x_k + s_k^E)\}\right], \tag{9}$$

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

A step satisfying this assumption is given by computing both the Cauchy step and the eigenstep and by choosing the one that provides the largest reduction in the model. By combining (4), (8), and (9), we obtain that

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \frac{\kappa_{fod}}{2} \max\left\{\|g_k\| \min\left[\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right], -\tau_k \Delta_k^2\right\}. \tag{10}$$

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)). Note that a fraction of optimal decrease condition is stronger than (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 2.1 may no longer hold. To be able to use such models in our framework, Assumption 2.2 needs to be modified to directly impose (10) instead of (9), where now g_k and H_k are the gradient and Hessian of those models taken at $s = 0$.

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 [5, 4]) depending upon the underlying basis and the number of function values available. What one requires in these cases 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 might be 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) constant Δ_{max}. Under this scenario, the region where f is sampled can be rigorously described as

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

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, that satisfies Assumption 3.1, be given. Let positive constants κ_{ef} and κ_{eg} be given and fixed. For any given $\Delta \in (0, \Delta_{max})$ and for any given $x \in L(x_0)$, consider a class of model functions $\mathcal{M} = \{m : \mathbb{R}^n \rightarrow \mathbb{R},\, m \in C^1\}$. The class $\mathcal{M}$ is called a fully linear class on $B(x; \Delta)$ if for any model function $m \in \mathcal{M}$*

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

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

and

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

$$|f(x+s) - m(x+s)| \;\leq\; \kappa_{ef}\,\Delta^2, \quad \forall s \in B(0;\Delta). \tag{12}$$

A model m that belongs to a fully-linear class $\mathcal{M}$ and, hence, satisfies (11) and (12) is called fully linear on $B(x;\Delta)$.

We next want to ensure that we can determine such a model.

Assumption 3.2. *For any given function f that satisfies Assumption 3.1, we assume that there exist suitable positive constants κ_{ef} and κ_{eg} such that, for any given $\Delta \in (0, \Delta_{max})$ and any given $x \in L(x_0)$, there exists a fully-linear class of models $\mathcal{M}$ on $B(x;\Delta)$, and that we can obtain a fully-linear model from this class in a finite, uniformly bounded for all x and Δ, number of improvement steps.*

Later in this section, we will indicate how this assumption can be satisfied in the particular context of polynomial interpolation and regression.

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

Assumption 3.3. *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 3.2. *Let a function f, that satisfies Assumption 3.3, be given. Let positive constants κ_{ef}, κ_{eg}, and κ_{eh} be given and fixed. For any given $\Delta \in (0, \Delta_{max})$ and for any given $x \in L(x_0)$, consider a class of model functions $\mathcal{M} = \{m : \mathbb{R}^n \to \mathbb{R},\, m \in C^2\}$. The class $\mathcal{M}$ is called a fully quadratic class on $B(x;\Delta)$ if for any model function $m \in \mathcal{M}$*

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

$$\|\nabla^2 f(x+s) - \nabla^2 m(x+s)\| \;\leq\; \kappa_{eh}\,\Delta, \quad \forall s \in B(0;\Delta), \tag{13}$$

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

$$\|\nabla f(x+s) - \nabla m(x+s)\| \;\le\; \kappa_{eg}\,\Delta^2, \quad \forall s \in B(0;\Delta), \tag{14}$$

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

$$|f(x+s) - m(x+s)| \;\le\; \kappa_{ef}\,\Delta^3, \quad \forall s \in B(0;\Delta). \tag{15}$$

Any model m that belongs to a fully-quadratic class $\mathcal{M}$ and, hence, satisfies (13)-(15) is called fully quadratic on $B(x;\Delta)$.

We again need to assume that such a model can be constructed.

Assumption 3.4. *For any given function f that satisfies Assumption 3.3, we assume that there exist suitable positive constants κ_{ef}, κ_{eg}, and κ_{eh} such that, for any given $\Delta \in (0, \Delta_{max})$ and any given $x \in L(x_0)$, there exists a fully-quadratic class of models $\mathcal{M}$ on $B(x;\Delta)$, and that, we can obtain a fully-quadratic model from this class in a finite, uniformly bounded for all x and Δ, number of improvement steps.*

Next we will indicate how this assumption can also be satisfied in the particular context of polynomial interpolation and regression.

Polynomial models. One way to ensure that a polynomial interpolation or regression model satisfies Taylor-like error bounds on the function value, on the gradient, and on the Hessian is to base this model on a Λ-poised sample set. Let us consider polynomial interpolation, discussed in [5]. The case of regression [4] is similar. We consider a Λ-poised set of interpolation points given by

$$Y \;=\; \left\{y^0, y^1, \ldots, y^p\right\},$$

where $p_1 = p + 1 = |Y|$ is a positive integer defining the number of points in the interpolation set. Let $m(y)$ denote an interpolating polynomial of degree d satisfying the interpolation conditions

$$m(y^i) \;=\; f(y^i),\; i = 0, \ldots, p.$$

Typically, $p_1 = p + 1$ is the dimension of the space of polynomials of given degree, so $p = n$, in the linear case, and $p = n + n(n+1)/2 = (n+1)(n+2)/2$, in the quadratic case.

We now show how these models fit into the framework described in the previous section. Assume that a constant $\Lambda \ge 1$ is given. Let $\mathcal{M}$ be the set of all quadratic interpolation polynomials each of which interpolates f on some set Y, Λ-poised on $B(x, \Delta)$. Using the error bounds in [5], we can conclude that $\mathcal{M}$ is a class of fully-quadratic models. From the results in [5], we know that we can select Λ in such a way that any interpolation set can be made Λ-poised in a finite number of improvement steps (in fact, in at most $(n+1)(n+1)/2 - 1$ steps). Hence, there exist suitable constants κ_{ef}, κ_{eg}, and κ_{eh} (dependent on Λ, but independent of x and Δ) such that a fully-quadratic model can be constructed for any given $\Delta \in (0, \Delta_{max})$ and any given $x \in L(x_0)$.

The case of fully-linear models and linear and quadratic regression fits into a similar framework. We conclude that polynomial interpolation and regression models can be chosen to satisfy Assumptions 3.2 and 3.4. But 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 3.2 and 3.4 (and Cauchy and minimal eigenvalue decreases can be extracted from the trust-region subproblems).

4. Derivative-free trust-region method (first order)

We will begin by formally stating the first-order version of the algorithm that we consider. The algorithm contemplates acceptance of new iterates based on simple decrease by selecting $\eta_0 = 0$. Note that accepting new iterates based on simple decrease is particularly appropriate in derivative-free optimization when function evaluations are expensive.

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

Step 0 (initialization): Choose an initial point x_0 and $\Delta_{max} > 0$. We assume that an initial model m_0 and a trust-region radius $\Delta_0 \in (0, \Delta_{max})$ are given.

The constants η_0, η_1, γ, γ_{inc} $\epsilon_c > 0$, β, and $\mu > 0$ are also given and satisfy the conditions $0 \leq \eta_0 \leq \eta_1 < 1$ (with $\eta_1 \neq 0$), $0 < \gamma < 1 < \gamma_{inc}$, $\epsilon_c > 0$, and $\mu > \beta > 0$. Set $k = 0$.

Step 1 (criticality step): If $\|g_k\| \leq \epsilon_c$, use Algorithm 4.2 (described below) to construct a model $\tilde{m}_k$, which is fully linear (for some constants κ_{ef} and κ_{eg}, which remain the same for all iterations of Algorithm 4.1) on the ball $B(x_k; \tilde{\Delta}_k)$ for some $\tilde{\Delta}_k \in (0, \mu\|\tilde{g}_k\|]$. Set $m_k = \tilde{m}_k$ and $\Delta_k = \min\{\tilde{\Delta}_k, \Delta_k\}$.

Step 2 (step calculation): Compute a step s_k that sufficiently reduces the model m_k (in the sense of (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 $\rho_k > \eta_1$ or if both $\rho_k > \eta_0$ and the model is fully linear (for the positive constants κ_{ef} and κ_{eg}) on $B(x_k; \Delta_k)$, then $x_{k+1} = x_k + s_k$ and the model is updated to take into consideration the new iterate, resulting in a new model m_{k+1}; otherwise the model and the iterate remain unchanged.

Step 4 (model improvement): If $\rho_k < \eta_1$ and if the model m_k is not fully linear on $B(x_k; \Delta_k)$, then make one or more suitable improvement steps. Define m_{k+1} to be the (possibly improved) model.

Step 5 (trust-region radius update): Set

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

Increment k by one and go to Step 1.

Note that in the algorithmic description above there could be two different models m_k, one given before and the other possibly after the criticality step. We will take this occurrence into account in our convergence analysis.

One possible procedure for 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 $\|g_k\| \leq \epsilon_c$ and one of the following holds: the model m_k is not fully linear or $\Delta_k > \mu\|g_k\|$. The constant $\alpha \in (0,1)$ is chosen at Step 0 of Algorithm 4.1.*

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

Repeat Increment i by one. Improve the previous model $m_k^{(i-1)}$ until it is fully linear on $B(x_k; \alpha^i \mu \|g_k^{(0)}\|)$ (notice that this can be done in a finite, uniformly bounded number of steps). Denote the new model by $m_k^{(i)}$. Set $\tilde{\Delta}_k = \alpha^i \mu \|g_k^{(0)}\|$ and $\tilde{m}_k = m_k^{(i)}$.

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

We will prove in the next section that Algorithm 4.2 terminates after a finite number of steps if $\|\nabla f(x_k)\| \neq 0$.

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 $\mathcal{S}$. Moreover, we will denote by $\mathcal{S}_+$ the subset of $\mathcal{S}$ for which $\Delta_k < \beta\|g_k\|$ and, hence, for which $\Delta_{k+1} = \gamma_{inc}\Delta_k$.
(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 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 (see Remark 4.1 below for further discussion). 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**.

Remark 4.1. *Notice that during a model-improvement step, Δ_k and x_k remain unchanged, and hence, in the absence of a successful iteration, there can be at most a finite, uniformly bounded, number of model-improvement steps before a fully-linear model is obtained. If we allow x_k to change during any model-improving iteration, then this property may no longer hold. However, if we apply any other mechanism which ensures that after a finite (uniformly bounded from above by some constant N) number of iterations either a successful iteration is encountered or a fully linear model is obtained, then it is possible to show that the convergence results which we prove below still hold. However, to simplify the proofs we will assume that x_k does not change during a model-improvement step.*

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] 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 satisfy the criticality test 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 either we do not have a fully-linear model m_k or that the radius Δ_k exceeds $\mu\|g_k\|$. We then define $m_k^{(0)} = m_k$ and the model is improved until it is fully linear on the ball $B(x_k; \alpha\mu\|g_k^{(0)}\|)$ (in a finite number of improvement steps). If the gradient $g_k^{(1)}$ of the resulting model $m_k^{(1)}$ satisfies $\|g_k^{(1)}\| \geq \alpha\|g_k^{(0)}\|$, the procedure stops with

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

Otherwise, that is if $\|g_k^{(1)}\| < \alpha\|g_k^{(0)}\|$, the model is improved until it is fully linear on the ball $B(x_k; \alpha^2\mu\|g_k^{(0)}\|)$. 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

$$\|g_k^{(i)}\| \;<\; \alpha^i\|g_k^{(0)}\| \;=\; \alpha^i\|g_k\|$$

for all $i \geq 0$, where $g_k^{(i)}$ is the gradient of the model $m_k^{(i)}$. This argument shows that $\lim_{i\to+\infty}\|g_k^{(i)}\| = 0$. Since each model $m_k^{(i)}$ was fully linear on $B(x_k; \alpha^i\mu\|g_k\|)$ then (11) with $s = 0$ and $x = x_k$ implies that

$$\|\nabla f(x_k) - g_k^{(i)}\| \;\leq\; \kappa_{eg}\alpha^i\mu\|g_k\|$$

for each $i \geq 0$. Thus, using the triangle inequality, it holds for all $i \geq 0$

$$\|\nabla f(x_k)\| \;\leq\; \|\nabla f(x_k) - g_k^{(i)}\| + \|g_k^{(i)}\| \;\leq\; (\kappa_{eg}\mu + 1)\,\alpha^i\|g_k\|.$$

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

The following simple lemma easily follows from the proof above.

Lemma 5.2. *Suppose that there exists a constant $\kappa_1 > 0$ such that $\|g_k\| \geq \kappa_1$ for all k. Then, the number of iterations of each execution of Algorithm 4.2 is uniformly bounded by $\lceil \log_\alpha(\frac{\kappa_1}{\epsilon_c}) \rceil$ for all k.*

Proof: If $\kappa_1 > \epsilon_c$ the criticality step is not entered and no such iterations are performed. So, let us focus on the case where $\kappa_1 \leq \epsilon_c$. It is trivial to observe that the number of iterations in each execution of Algorithm 4.2 is bounded by the number of times it requires to multiply ϵ_c by α to obtain a value not exceeding κ_1. In fact, suppose we entered the criticality step ($\|g_k\| \leq \epsilon_c$). After applying Algorithm 4.2 we have

$$\|\tilde{g}_k\| \;=\; \|g_k^{(i^*)}\| \;\geq\; \kappa_1 \;\geq\; \alpha^{i^*}\epsilon_c \;\geq\; \alpha^{i^*}\|g_k\|,$$

where $i^* = \lceil \log_\alpha(\frac{\kappa_1}{\epsilon_c}) \rceil$. ■

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, on the Lipschitz continuity of the gradient of f, and on the existence of fully linear models; i.e., Assumptions 3.1, 3.2, and 5.1. We also require the model Hessian $H_k = \nabla^2 m_k(x_k)$ to be uniformly bounded:

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 lemma.

Lemma 5.3. *If m_k is fully linear on $B(x_k; \Delta_k)$ and*

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

then the k-th iteration is successful.

Proof: Since

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

the fraction of Cauchy decrease condition (5)-(6) immediately gives that

$$m_k(x_k) - m_k(x_k+s_k) \;\geq\; \frac{\kappa_{fcd}}{2}\|g_k\| \min\left[\frac{\|g_k\|}{\kappa_{bhm}}, \Delta_k\right] \;=\; \frac{\kappa_{fcd}}{2}\|g_k\|\Delta_k. \quad (16)$$

On the other hand, since the current model is fully linear on $B(x_k;\Delta_k)$, then from the bound (12) on the error between the function and the model and from (16) 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 fact that $\Delta_k \leq \kappa_{fcd}\|g_k\|(1-\eta_1)/(4\kappa_{ef})$ to deduce the last inequality. Therefore, $\rho_k \geq \eta_1$, and iteration k is successful. ■

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

Lemma 5.4. *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*

$$\Delta_k \;>\; \kappa_2$$

for all k.

Proof: First, let us assume that $\epsilon_c > \|g_k\| \geq \kappa_1$. Then the criticality step is invoked. From Lemma 5.2 we know that the number of iterations in each execution of Algorithm 4.2 is bounded by $i^* = \lceil \log_\alpha(\frac{\kappa_1}{\epsilon_c})\rceil$. Hence, after each criticality step $\Delta_k = \tilde{\Delta}_k \geq \alpha^{i^*}\mu\|g_k\| \geq \alpha^{i^*}\mu\kappa_1$.

Now let us consider how Δ_k can change outside the criticality step (including the case when $\|g_k\| \geq \epsilon_c$ and the criticality step was not even entered at the beginning of the iteration). By Lemma 5.3 and by the assumption that $\|g_k\| \geq \kappa_1$ for all k, whenever Δ_k falls below a certain value

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

the k-th iteration has to be either successful or model improving (when it is not successful and m_k is not fully linear) and hence $\Delta_{k+1} \geq \Delta_k$. This, then, automatically implies that $\Delta_j \geq \Delta_k$ for all $j \geq k$ and $\Delta_j \geq \gamma\bar{\kappa}_2$ for all

$j \geq k$ due to the mechanism of Step 5. Combining the two bounds on Δ_k we conclude that $\Delta_k \geq \min\{\alpha^{i^*}\mu\kappa_1, \gamma\bar{\kappa}_2\} = \kappa_2$. ■

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

Lemma 5.5. *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. If an infinite loop occurs in Step 1, the result follows from Lemma 5.1.

Otherwise, we know that we can have only a finite (uniformly bounded, say by N) number of model-improvement 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. For any $k_0 > 0$ sufficiently large,

$$\sum_{k\geq k_0}^{+\infty} \Delta_k \;\leq\; N \sum_{i=1}^{+\infty} \gamma^i \Delta_{k_0} \;=\; \frac{N\gamma}{1-\gamma}\Delta_{k_0}.$$

Thus, Δ_k converges to zero.

Clearly, for any $i \geq k_0$ and $j \geq k_0$, we have

$$\|x_i - x_j\| \;\leq\; \sum_{k\geq k_0}^{+\infty} \Delta_k \;\leq\; \frac{N\gamma}{1-\gamma}\Delta_{k_0}.$$

And, as we let k_0 go to infinity, $\|x_i - x_j\| \to 0$ as i and j go to $+\infty$. 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.

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 (11) 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 5.3, since if it 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 will now show that the gradient of the model is not bounded away from zero even if the number of successful iterations is infinite.

Lemma 5.6. *If the number of successful iterations is infinite then*

$$\liminf_{k \to +\infty} \|g_k\| \;=\; 0. \tag{17}$$

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

$$\|g_k\| \;\geq\; \kappa_1 \tag{18}$$

for some $\kappa_1 > 0$. By Lemma 5.4 we have that $\Delta_k \geq \kappa_2$ for all k. Now consider a successful iteration of index k. The fact that $k \in \mathcal{S}$ implies that

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

By using the bound on the fraction of Cauchy decrease (5)-(6), we have that

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

Now, using the bounds $\|g_k\| \geq \kappa_1$, $\Delta_k \geq \kappa_2$, and $\|H_k\| \leq \kappa_{bhm}$, we have

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1 \kappa \frac{\kappa_{fcd}}{2} \min\left[\frac{\kappa_1}{\kappa_{bhm}}, \kappa_2\right] \;>\; 0.$$

Hence, at each successful iteration, the objective function f is decreased by a positive amount bounded away from zero. But since the number of successful iterations is infinite, that means that f is reduced by an infinite amount, which contradicts our assumption that f is bounded from below. The assumption (18) must therefore be false, which yields (17). ■

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*

$$\lim_{i \to +\infty} \|g_{k_i}\| \;=\; 0 \tag{19}$$

it also holds that

$$\lim_{i \to +\infty} \|\nabla f(x_{k_i})\| \;=\; 0. \tag{20}$$

Proof: By (19), $\|g_{k_i}\| \leq \epsilon_c$ for i sufficiently large, and by Lemma 5.1 the mechanism of Step 1 ensures that the model m_{k_i} is fully-linear on a ball $B(x_{k_i}; \Delta_{k_i})$ for some $\Delta_{k_i} \leq \mu \|g_{k_i}\|$ for all i sufficiently large if $\nabla f(x_{k_i}) \neq 0$. Then, using the bound (11) 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 (20). ■

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

Theorem 5.1. *Let Assumptions 3.1, 3.2, 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. We are now able to prove that all limit points of the sequence of iterates are first-order critical. This latter result needs an additional assumption on the algorithm.

Theorem 5.2. *Let Assumptions 3.1, 3.2, 5.1, and 5.2 hold. Assume in addition that the trust-region radius is never increased when $k \in \mathcal{S} \setminus \mathcal{S}_+$. Then,*

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

Proof: We have established by Lemma 5.5 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

$$\|\nabla f(x_{k_i})\| \;\geq\; \epsilon_0 \;>\; 0, \tag{21}$$

for some $\epsilon_0 > 0$ and for all i (we can ignore the other type 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

$$\epsilon \;\leq\; \min\left\{\frac{\epsilon_0}{2(2+\kappa_{eg}\mu)}, \epsilon_c\right\}. \tag{22}$$

Lemma 5.6 then ensures the existence, for each k_i in the subsequence, of a first iteration $\ell_i > k_i$ such that $\|g_{\ell_i}\| < \epsilon$. We thus obtain that there exists another subsequence indexed by $\{\ell_i\}$ such that

$$\|g_k\| \;\geq\; \epsilon \;\text{ for }\; k_i \leq k < \ell_i \;\text{ and }\; \|g_{\ell_i}\| \;<\; \epsilon, \tag{23}$$

for sufficiently large i.

We now restrict our attention to the set $\mathcal{K}$ which is the subsequence of successful or acceptable 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.

We now show that for any large enough $k \in \mathcal{K}$ the iteration k is successful. If $\eta_0 > 0$, for the purpose of this proof, we can assume that $\eta_0 = \eta_1$ and all iterations are successful.

Suppose now that $\eta_0 = 0$, recall that we then assume that Δ_k increases only when $k \in \mathcal{S}_+$, and suppose that there is an infinite number of acceptable iterations in $\mathcal{K}$. Without loss of generality we can then assume that there is an acceptable iteration somewhere between the iterations k_i and ℓ_i for all i. By Lemma 5.3 we know that once

$$\Delta_k \;\leq\; \bar{\kappa}_2 = \min\left[\frac{\epsilon}{\kappa_{bmh}}, \frac{\kappa_{fcd}\epsilon(1-\eta_1)}{4\kappa_{ef}}\right], \tag{24}$$

then we will either have a model-improving iteration (which we can ignore) or we a have successful iteration (instead of an acceptable one). Since we assumed that for all i there exists at least one acceptable iteration k such that $k_i \leq k < \ell_i$, then for this k we get $\Delta_k \geq \bar{\kappa}_2$. On the other hand, we know that during iterations whose indices k are not in $\mathcal{K}$ the criticality step is invoked and Δ_k is decreased to satisfy $\Delta_k \leq \mu\|g_k\|$. We also know that a subsequence of such $\|g_k\|$ converges to zero, hence for i large enough we know that on iterations between ℓ_i and k_{i+1} the trust-region radius has to be reduced at least once to a value below $\frac{\bar{\kappa}_2}{\gamma_{inc}}$. Since we assume that $\Delta_k \geq \bar{\kappa}_2$ for some $k_i \leq k < \ell_i$ for all i large enough, this means that for all i large enough Δ_k has to be increased at least once from the value above $\frac{\bar{\kappa}_2}{\gamma_{inc}}$ to the value above $\bar{\kappa}_2$. Recall that we assume that the only time when Δ_k can increase is when $k \in \mathcal{S}_+$. This means that for each i large enough there is a successful iteration $\ell_i \leq k < \ell_{i+1}$, $k \in \mathcal{S}_+$ with $\Delta_k \geq \frac{\bar{\kappa}_2}{\gamma_{inc}}$ and $\|g_k\| > \frac{\Delta_k}{\beta} \geq \frac{\bar{\kappa}_2}{\beta\gamma_{inc}}$. On each such iteration

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1[m_k(x_k) - m_k(x_k + s_k)] \;\geq\; \eta_1 \frac{\kappa_{fcd}}{2}\|g_k\| \min\left[\frac{\|g_k\|}{\kappa_{bhm}}, \Delta_k\right]. \tag{25}$$

This implies that there is an infinite number of successful iterations with both $\|g_k\|$ and Δ_k bounded away from zero from below, which, as we have observed before, contradicts the boundedness of $f(x)$ from below. Hence, we have a contradiction with the assumption that there is an infinite number of acceptable iterations in $\mathcal{K}$. After a certain large enough i all iterations in $\mathcal{K}$ are thus successful.

Moreover, it follows from the arguments above that

$$\lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \Delta_k \;=\; 0.$$

As a consequence we obtain $\Delta_k \leq \frac{\epsilon}{\kappa_{bhm}}$ for $k \in \mathcal{K}$ sufficiently large, and (25) and k being sufficiently large imply

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

We then deduce from this bound that, for i sufficiently large,

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

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 (23). For the second term we use the fact that, from (22), we know that the criticality step was invoked at iteration ℓ_i. Thus, the model m_{ℓ_i} is fully linear on $B(x_{\ell_i}; \mu\|g_{\ell_i}\|)$ and using (23), 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 (22) that

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

for i large enough, which contradicts (21). Hence our initial assumption must be false and the theorem follows. ∎

Remark 5.1. *Instead of assuming that Δ_k is not increased during iterations whose indices are in $\mathcal{S} \setminus \mathcal{S}_+$ we can assume that $\eta_0 \in (0, 1)$ and the above theorem will still hold. The proof is simpler than the proof above and is omitted.*

6. Derivative-free trust-region method (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 lines as 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. We state the version of the algorithm we wish to consider.

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

Step 0 (initialization): Choose an initial point x_0 and $\Delta_{max} > 0$. We assume that an initial model m_0 and a trust-region radius $\Delta_0 \in (0, \Delta_{max})$ are given.

The constants η_0, η_1, γ, γ_{inc} $\epsilon_c > 0$, β, and $\mu > 0$ are also given and satisfy the conditions $0 \leq \eta_0 \leq \eta_1 < 1$ (with $\eta_1 \neq 0$), $0 < \gamma < 1 < \gamma_{inc}$, $\epsilon_c > 0$, and $\mu > \beta > 0$. Set $k = 0$.

Step 1 (criticality test): If $\sigma_k^m \leq \epsilon_c$, use Algorithm 6.2 (described below) to construct a model $\tilde{m}_k$, which is fully quadratic (for some constants κ_{ef}, κ_{eg}, and κ_{eh}, which remain the same for all iterations of Algorithm 6.1) on the ball $B(x_k; \tilde{\Delta}_k)$ for some $\tilde{\Delta}_k \in (0, \mu\tilde{\sigma}_k^m]$. Set $m_k = \tilde{m}_k$ and $\Delta_k = \min\{\tilde{\Delta}_k, \Delta_k\}$.

Step 2 (step calculation): Compute a step s_k that sufficiently reduces the model m_k (in the sense of (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 $\rho_k > \eta_1$ or if both $\rho_k > \eta_0$ and the model is fully quadratic (for the positive constants κ_{ef}, κ_{eg}, and κ_{eh}) on $B(x_k; \Delta_k)$, then $x_{k+1} = x_k + s_k$ and the model is updated to take into consideration the new iterate resulting in a new model m_{k+1}; otherwise the model and the iterate remain unchanged.

Step 4 (model improvement): If $\rho_k < \eta_1$ and if the model m_k is not fully quadratic on $B(x_k; \Delta_k)$, then make it so by suitable improvement steps. Define m_{k+1} to be the (possibly improved) model.

Step 5 (trust-region radius update): Set

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

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 again 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 4.1 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 enforcing fully-quadratic models rather fully-linear ones.

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

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

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

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

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 3.3 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 (Assumption 3.4).

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 3.3 holds and m is a fully-quadratic model on $B(x;\Delta)$. Then, we have that*

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

Proof: The proof follows directly from the bound (13) on the error between the Hessians of m and f and the Wielandt-Hoffman Theorem (see, for example [6, Theorem 6.3.5]).

If we assume that the eigenvalues of $\nabla^2 f(x)$ are $\lambda_1 \leq \cdots \leq \lambda_n$ and of $\nabla^2 m(x)$ are $\mu_1 \leq \cdots \leq \mu_n$ then

$$\kappa_{eh}\,\Delta \geq \|\nabla^2 f(x) - \nabla^2 m(x)\| \;\geq\; \|\nabla^2 f(x) - \nabla^2 m(x)\|_F/\sqrt{n} \geq \left\{\sum_{i=1}^{n} |\lambda_i - \mu_i|^2/n\right\}^{\frac{1}{2}}$$
$$\geq \frac{|\lambda_1 - \mu_1|}{\sqrt{n}} = \frac{|\lambda_{min}(\nabla^2 f(x)) - \lambda_{min}(\nabla^2 m(x))|}{\sqrt{n}}$$

and the result follows. ■

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

Lemma 7.1. *Let Δ be bounded by Δ_{max}. Suppose that Assumption 3.3 holds and m is a fully-quadratic model on $B(x;\Delta)$. Then, we have that*

$$|\sigma(x) - \sigma^m(x)| \;\leq\; \kappa_\sigma \Delta. \tag{26}$$

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 (14) 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 $\sqrt{n}\kappa_{eh}\Delta$ because of Proposition 7.1. Finally we need only to write $\kappa_\sigma = \max\{\kappa_{eg}\Delta_{max}, \sqrt{n}\kappa_{eh}\}$ and the result follows. ■

The convergence theory will require the already mentioned Assumptions 3.3, 3.4, and 5.1, as well the uniform upper bound on the Hessians of the quadratic models (Assumption 5.2).

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

Lemma 7.2. *If $\sigma(x_k) \neq 0$, Step 1 of Algorithm 6.1 will satisfy the criticality test in a finite number of improvement steps (by applying Algorithm 6.2).*

Proof: The proof is practically identical to the proof of Lemma 5.1, with $\|g_k\|$ replaced by σ_k^m and $\nabla f(x_k)$ replaced by $\sigma(x_k)$. ■

We now state the second order analogue of Lemma 5.2.

Lemma 7.3. *Suppose that there exists a constant $\kappa_1 > 0$ such that $\sigma_k^m \geq \kappa_1$ for all k. Then, the number of iterations of each execution of Algorithm 6.2 is uniformly bounded by $\lceil \log_\alpha(\frac{\kappa_1}{\epsilon_c}) \rceil$ for all k.*

Proof: The proof is the exact repetition of the proof of Lemma 5.2 with $\|g_k\|$ replaced by σ_k^m. ■

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*

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

then the k-th iteration is successful.

Proof: The proof is similar to the proof of Lemma 5.3 for the first-order case, however now we need to take the second-order terms into account.

First we recall the fractions of Cauchy and eigenstep decreases (9)-(10), written after the application of Assumption 5.2,

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

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

$$m_k(x_k) - m_k(x_k+s_k) \;\geq\; \frac{\kappa_{fod}}{2}\|g_k\|\Delta_k \;=\; \frac{\kappa_{fod}}{2}\sigma_k^m\Delta_k. \tag{27}$$

On the other hand, since the current model is fully-quadratic on $B(x_k;\Delta_k)$, we may deduce from (27) and the bound (15) 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

$$m_k(x_k) - m_k(x_k+s_k) \;\geq\; -\tfrac{\kappa_{fod}}{2}\tau_k\Delta_k^2 \;=\; \tfrac{\kappa_{fod}}{2}\sigma_k^m\Delta_k^2. \tag{28}$$

But, since the current model is fully-quadratic on $B(x_k;\Delta_k)$, we deduce from (28) and the bound (15) 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 Lemmas 7.3 and 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*

$$\Delta_k \;>\; \kappa_2$$

for all k.

Proof: The proof is trivially derived by combining Lemmas 7.3 and 7.4 and the proof of Lemma 5.4. ■

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 5.5 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 Lemma 7.1. ■

In the case where the number of successful iterations is infinite we start by showing that the second-order stationarity of the model is not uniformly bounded away from zero.

Lemma 7.7. *If the number of successful iterations is infinite then*

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

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 $\Delta_k > \kappa_2$ for all k. For each successful iteration we have

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

Since $\sigma_k^m \geq \kappa_1$, then either $\|g_k\| \geq \kappa_1$ or $-\tau_k = -\lambda_{\min}(H_k) \geq \kappa_1$. That means that the right-hand side is bounded away from zero for all k, and, hence, so is $f(x_{k+1}) - f(x_k)$ for each successful iteration. That implies that the number of successful iterates cannot be infinite since f is bounded from below. We have arrived at a contradiction. ■

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

$$\lim_{i \to +\infty} \sigma_{k_i}^m \;=\; 0 \tag{29}$$

it also holds that

$$\lim_{i \to +\infty} \sigma(x_{k_i}) \;=\; 0. \tag{30}$$

Proof: By (29), $\sigma_{k_i}^m \leq \epsilon_c$ for i sufficiently large, and the mechanism of Step 1 ensures that the model m_{k_i} is fully quadratic on the ball $B(x_{k_i}; \Delta_{k_i})$ for some $\Delta_{k_i} \leq \mu \sigma_{k_i}^m$ (for all i sufficiently large). Now, using (26),

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

The limit (29) and this last bound then give (30). ■

Lemmas 7.7 and 7.8 immediately give the following global convergence result.

Theorem 7.1. *Let Assumptions 3.3, 3.4, 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 now able to prove that all limit points of the sequence of iterates are second-order critical. As in the first-order case, this latter result needs an additional assumption on the algorithm.

Theorem 7.2. *Let Assumptions 3.3, 3.4, 5.1, and 5.2 hold. Assume in addition that the trust-region radius is never increased when $k \in \mathcal{S} \setminus \mathcal{S}_+$. Then,*

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

Proof: We have established by Lemma 7.6 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

$$\sigma(x_{k_i}) \;\geq\; \epsilon_0 \;>\; 0, \tag{31}$$

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 7.8, 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

$$\epsilon \;\leq\; \min\left\{\frac{\epsilon_0}{2(2+\kappa_\sigma \mu)}, \epsilon_c\right\}. \tag{32}$$

Lemma 7.7 then ensures the existence, for each k_i in the subsequence, of a first successful or acceptable iteration $\ell_i > k_i$ such that $\sigma_{\ell_i}^m < \epsilon$. We thus obtain that there exists another subsequence indexed by $\{\ell_i\}$ such that

$$\sigma_k^m \;\geq\; \epsilon \;\text{ for }\; k_i \leq k < \ell_i \;\text{ and }\; \sigma_{\ell_i}^m \;<\; \epsilon, \tag{33}$$

for sufficiently large i.

We now restrict our attention to the set $\mathcal{K}$ which is defined as the subsequence of successful or acceptable 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.

We first show that for large enough $k \in \mathcal{K}$ the k-th iteration is successful, i.e., that there is only a finite number of acceptable iterations in $\mathcal{K}$. We omit the proof of this statement as it is a very close replica of the proof of a similar statement for the first-order case in the proof of Theorem 5.2.

We now observe that for large enough $k \in \mathcal{K}$ either $\|g_k\| > \epsilon$ in which case

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1 [m_k(x_k) - m_k(x_k + s_k)] \;\geq\; \eta_1 \frac{\kappa_{fod}}{2} \epsilon \min\left[\frac{\epsilon}{\kappa_{bhm}}, \Delta_k\right] \tag{34}$$

or $-\tau_k > \epsilon$ and

$$f(x_k) - f(x_{k+1}) \;\geq\; \eta_1 [m_k(x_k) - m_k(x_k + s_k)] \;\geq\; \eta_1 \frac{\kappa_{fod}}{2} \epsilon \Delta_k^2. \tag{35}$$

Since the sequence $\{f(x_k)\}$ is bounded from below (by Assumption 5.1) and monotonically decreasing (by construction), then it is convergent and the left-hand sides of both (34) and (35) must tend to zero when k tends to infinity. As a result, we get

$$\lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \Delta_k \;=\; 0.$$

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, for large enough $k \in \mathcal{K}$, k is also in $\mathcal{S}_+$.

From the scheme that updates Δ_j at successful iterations we can deduce that, for i large enough,

$$\|x_{k_i} - x_{\ell_i}\| \;\leq\; \sum_{j=k_i}^{\ell_i - 1} \Delta_j \;\leq\; \sum_{j=k_i}^{\ell_i - 1} \gamma_{inc}^{\ell_i - j} \Delta_j \leq \frac{\gamma_{inc}}{\gamma_{inc} - 1} \Delta_{\ell_i - 1}. \tag{36}$$

We conclude that $\|x_{k_i} - x_{\ell_i}\| \to 0$, from the fact that $\Delta_{\ell_i - 1} \to 0$. We therefore obtain that

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

Now,

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

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 bounded by ϵ by (33). For the second term we use the fact that from (32) we know that the criticality step was invoked at iteration ℓ_i. Thus, the model m_{ℓ_i} is fully quadratic on $B(x_{\ell_i}; \mu\sigma^m_{\ell_i})$ and using (33), 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 (32) that

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

for i large enough, which contradicts (31). Hence our initial assumption must be false and the theorem follows. ■

Remark 7.1. *Instead of assuming that Δ_k is not increased during iterations whose indices are in $\mathcal{S} \setminus \mathcal{S}_+$ we can assume that $\eta_0 \in (0, 1)$ and the above theorem will still hold. The proof is a relatively simple modification of the proof above.*

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 sample sets in derivative free optimization. Part II: polynomial regression and underdetermined interpolation*, Tech. Report 05-15, Departamento de Matemática, Universidade de Coimbra, Portugal, 2005.

[5] ——, *Geometry of interpolation sets in derivative free optimization*, Math. Program., (2006, to appear).

[6] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[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, 1999.

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

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

[12] 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.

A. R. Conn
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).

K. Scheinberg
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).

L. N. Vicente
Departamento de Matemática, Universidade de Coimbra, 3001-454 Coimbra, Portugal (lnv@mat.uc.pt).