Chapter 10

Trust-region methods based on derivative-free models

Trust-region methods are a well-studied class of algorithms for the solution of nonlinear programming problems [57, 178]. 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 [57]).

In this chapter we address trust-region methods for unconstrained derivative-free optimization. These methods maintain quadratic (or linear) 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 chapter abstracts from the specifics of model building. In fact, it is not even required that these models be polynomial functions as long as appropriate decreases (such as Cauchy and eigenstep 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. In Chapter 6, we called such models, depending on their accuracy, *fully linear* and *fully quadratic*. It has been rigorously shown in Chapters 3–6 how such *fully linear* and *fully quadratic* models can be constructed in the context of polynomial interpolation or regression.

Again, the problem we are considering is (1.1), where f is a real-valued function, assumed to be once (or twice) continuously differentiable and bounded from below.

10.1 The trust-region framework basics

The fundamentals of trust-region methods are rather simple. 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. To be useful it must be significantly easier to optimize the model within the neighborhood than solving the original problem. The neighborhood considered is called the trust region. 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 quadratic, written in the form

$$m_k(x_k + s) \;=\; m_k(x_k) + s^\top g_k + \frac{1}{2} s^\top H_k s. \tag{10.1}$$

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 $\nabla^2 m_k(x_k) = H_k$. Clearly, g_k is the gradient of the model at $s = 0$.

If m_k is a first-order Taylor model, then $m_k(x_k) = f(x_k)$ and $g_k = \nabla f(x_k)$, and if it is a second-order Taylor model, one has, in addition, $H_k = \nabla^2 f(x_k)$. In general, even in the derivative case, H_k is a symmetric approximation to $\nabla^2 f(x_k)$. In the derivative-free case, we use models where $H_k \neq \nabla^2 f(x_k)$, $g_k \neq \nabla f(x_k)$, and, in the absence of interpolation, $m_k(x_k) \neq f(x_k)$.

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. Figure 10.1 illustrates a linear model and a quadratic model of a nonlinear function in a trust region (a simple ball), both built by interpolation. As expected, the quadratic model captures the curvature of the function.

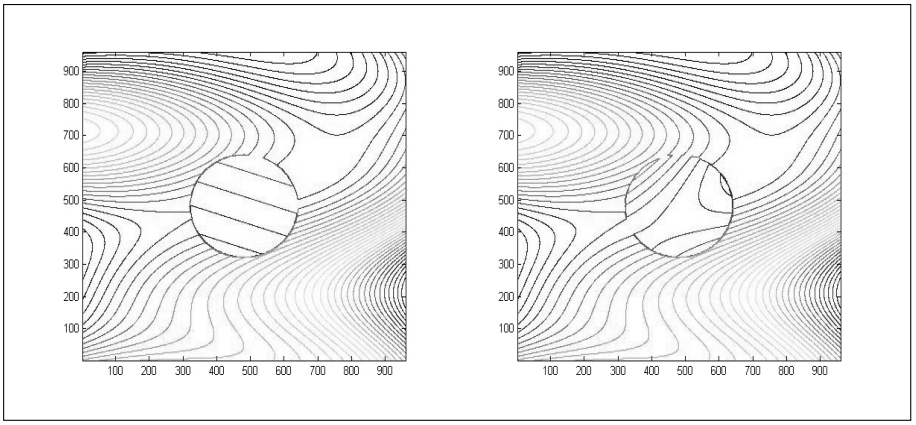

Figure 10.1. *Contours of a linear model (left) and a quadratic model (right) of a nonlinear function in a trust region.*

Thus, in the unconstrained case, the local model problem (called trust-region subproblem) we are considering is stated as

$$\min_{s\in B(0;\Delta_k)} m_k(x_k+s), \tag{10.2}$$

where $m_k(x_k+s)$ is the model for the objective function given at (10.1) and $B(0;\Delta_k)$ is our trust region of radius Δ_k, now centered at 0 and expressed in terms of $s = x - x_k$.

The Cauchy step

In some sense, the driving force for all optimization techniques is steepest descent since it defines the locally best direction of descent. It turns out that it is crucial, from the point of view of global convergence, that one minimizes the model at least as well as something related to steepest descent. On this basis, one defines something called the Cauchy step s_k^{C}, which is actually the step to the minimum of the model along the steepest descent direction within the trust region. Thus, if we define

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

then the Cauchy step is a step given by

$$s_k^{\mathrm{C}} = -t_k^{\mathrm{C}} g_k. \tag{10.3}$$

A fundamental result that drives trust-region methods to first-order criticality is stated and proved below.

Theorem 10.1. *Consider the model* (10.1) *and the Cauchy step* (10.3)*. Then*

$$m_k(x_k) - m_k(x_k + s_k^C) \;\geq\; \frac{1}{2}\|g_k\| \min\left\{\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right\}, \tag{10.4}$$

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

Proof. We first note that

$$m_k(x_k - \alpha g_k) \;=\; m_k(x_k) - \alpha\|g_k\|^2 + \frac{1}{2}\alpha^2 g_k^\top H_k g_k. \tag{10.5}$$

In the case where the curvature of the model along the steepest descent direction $-g_k$ is positive, that is, when $g_k^\top H_k g_k > 0$, we know that the model is convex along that direction, and so a stationary point will necessarily be the global minimizer in that direction. Denoting the optimal parameter by α_k^* we have that $-\|g_k\|^2 + \alpha_k^* g_k^\top H_k g_k = 0$ and

$$\alpha_k^* = \frac{\|g_k\|^2}{g_k^\top H_k g_k}.$$

Thus, if $\| - \alpha_k^* g_k\| = \|g_k\|^3 / g_k^\top H_k g_k \leq \Delta_k$, then the unique minimizer lies in the trust region and we can conclude that

$$m_k(x_k + s_k^{\mathrm{C}}) - m_k(x_k) \;=\; -\frac{1}{2}\frac{\|g_k\|^4}{g_k^\top H_k g_k},$$

and, consequently,

$$m_k(x_k) - m_k(x_k + s_k^C) \geq \frac{1}{2}\|g_k\|^2/\|H_k\|. \tag{10.6}$$

If $\|-\alpha_k^* g_k\| > \Delta_k$, then we take $\|-t_k^C g_k\| = \Delta_k$; i.e., we go to the boundary of the trust region. In this case

$$m_k(x_k + s_k^C) - m_k(x_k) = -\frac{\Delta_k\|g_k\|^2}{\|g_k\|} + \frac{1}{2}\frac{\Delta_k^2\|g_k\|}{\|g_k\|^3} g_k^\top H_k g_k.$$

But $\|g_k\|^3/g_k^\top H_k g_k > \Delta_k$ (since the optimal step was outside the trust region) then gives

$$m_k(x_k) - m_k(x_k + s_k^C) \geq \frac{1}{2}\Delta_k\|g_k\|. \tag{10.7}$$

It remains to consider the case when the one-dimensional problem in α is not convex. In this case, again we know that we will terminate at the boundary of the trust region (because the second and third terms on the right-hand side of (10.5) are both negative for all positive α). Furthermore, $g_k^\top H_k g_k \leq 0$ implies that

$$m_k(x_k + s_k^C) - m_k(x_k) \leq -t_k^C\|g_k\|^2 = -\Delta_k\|g_k\| < -\frac{1}{2}\Delta_k\|g_k\|. \tag{10.8}$$

The derived bounds (10.6), (10.7), and (10.8) on the change in the model imply that (10.4) holds. □

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 10.1. *For all iterations k,*

$$m_k(x_k) - m_k(x_k + s_k) \geq \kappa_{fcd}\left[m_k(x_k) - m_k(x_k + s_k^C)\right] \tag{10.9}$$

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

The steps computed under Assumption 10.1 will therefore provide a fraction of Cauchy decrease, which from Theorem 10.1 can be bounded from below as

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{fcd}}{2}\|g_k\| \min\left\{\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right\}. \tag{10.10}$$

If $m_k(x_k + s)$ is not a linear or a quadratic function, then Theorem 10.1 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 [57, Section 6.3.3] or Section 12.2). Calculating a step yielding a decrease better than the Cauchy decrease could be achieved by approximately solving the trust-region subproblem, which now involves the minimization of a nonlinear function within a trust region.

Assumption 10.1 is the minimum requirement for how well one has to do at solving (10.2) to achieve global convergence to first-order critical points. If we would like to guarantee more, then we must drive the algorithm with more than just the steepest descent direction. We will consider this case next.

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 $\tau_k = \lambda_{min}(H_k)$ be the most negative eigenvalue of H_k. In this case, we can determine a step of negative curvature s_k^{E}, such that

$$(s_k^{\mathrm{E}})^\top (g_k) \leq 0, \quad \|s_k^{\mathrm{E}}\| = \Delta_k, \quad \text{and} \quad (s_k^{\mathrm{E}})^\top H_k (s_k^{\mathrm{E}}) = \tau_k \Delta_k^2. \tag{10.11}$$

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 (10.11) 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.

Lemma 10.2. *Suppose that the model Hessian H_k has negative eigenvalues. Then we have that*

$$m_k(x_k) - m_k(x_k + s_k^E) \geq -\frac{1}{2}\tau_k \Delta_k^2. \tag{10.12}$$

Proof. It suffices to point out that

$$\begin{aligned} m_k(x_k) - m_k(x_k + s_k^{\mathrm{E}}) &= -(s_k^{\mathrm{E}})^\top (g_k) - \tfrac{1}{2}(s_k^{\mathrm{E}})^\top H_k (s_k^{\mathrm{E}}) \\ &\geq -\tfrac{1}{2}(s_k^{\mathrm{E}})^\top H_k (s_k^{\mathrm{E}}) \\ &= -\tfrac{1}{2}\tau_k \Delta_k^2. \quad \square \end{aligned}$$

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 10.2. *For all iterations k,*

$$m_k(x_k) - m_k(x_k + s_k) \geq \kappa_{fod}\left[m_k(x_k) - \min\{m_k(x_k + s_k^C), m_k(x_k + s_k^E)\}\right] \tag{10.13}$$

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

A step satisfying this assumption is given 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 (10.4), (10.12), and (10.13), we obtain that

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{fod}}{2} \max\left\{\|g_k\| \min\left\{\frac{\|g_k\|}{\|H_k\|}, \Delta_k\right\}, -\tau_k \Delta_k^2\right\}. \tag{10.14}$$

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 (10.2)). Note that a fraction of optimal decrease condition is stronger than (10.14) for the same value of κ_{fod}.

If $m_k(x_k+s)$ is not a quadratic function, then Theorem 10.1 and Lemma 10.2 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 [57, Section 6.6.2] or Section 12.2). Calculating a step yielding a decrease better than the Cauchy and eigenstep decreases could be achieved by approximately solving the trust-region subproblem, which, again, now involves the minimization of a nonlinear function within a trust region.

The update of the trust-region radius

The other essential ingredient of a trust-region method is the so-called trust-region management. The basic idea is to compare the truth, that is, the actual reduction in the objective function, to the predicted reduction in the model. If the comparison is good, we take the new step and (possibly) increase the trust-region radius. If the comparison is bad, we reject the new step and decrease the trust-region radius. Formally, this procedure can be stated as follows. We introduce a distinction between simple and sufficient decreases in the objective function. In the former case, when $\eta_0 = 0$, the step is accepted as long as it provides a simple decrease in the objective function, which might be a natural thing to do in derivative-free optimization when functions are expensive to evaluate.

Suppose that the current iterate is x_k and that the candidate for the next iterate is x_k+s_k. Assume, also, that one is given constants η_0, η_1, and γ satisfying $\gamma \in (0,1)$ and $0 \leq \eta_0 \leq \eta_1 < 1$ (with $\eta_1 \neq 0$).

Truth versus prediction: Define

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

Step acceptance (sufficient decrease): If $\rho_k \geq \eta_1$, then accept the new point x_k+s_k; otherwise, the new candidate point is rejected (and the trust-region radius is reduced, as below).

Step acceptance (possibly based on simple decrease): If $\rho_k \geq \eta_0$, then accept the new point x_k+s_k (but reduce the trust-region radius if $\rho_k < \eta_1$, as below); otherwise, the new candidate point is rejected (and the trust-region radius is reduced, as below).

Trust-region management: Set

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, +\infty) & \text{if } \rho_k \geq \eta_1, \\ \{\gamma \Delta_k\} & \text{if } \rho_k < \eta_1. \end{cases}$$

An important property which ensures convergence is the following: if the model is based on, for example, (some reasonable approximation to) a truncated Taylor series expansion, then we know that as the trust-region radius becomes smaller the model necessarily

becomes better. This guarantees that the trust-region radius is bounded away from zero, away from stationarity. In what follows we are not using Taylor models because we are interested in the case where although derivatives may exist they are not available.

10.2 Conditions on the trust-region models

In the derivative-free case, away from stationary points, it is necessary to ensure that the trust-region radius remains bounded away from zero. In the case of models based on a Taylor series approximation, this is ensured by the fact that the model becomes progressively better as the neighborhood (trust region) becomes smaller. The management of the trust-region radius guarantees that it stays bounded away from zero as long as the current iterate is not a stationary point. Models like polynomial interpolation or regression models do not necessarily become better when the radius of the trust region is reduced. Hence, we have to ensure that we reduce only the trust-region radius when we are certain that the failure of a current step is due to the size of the trust region and not to the poor quality of the model itself. With this safeguard, we can prove, once again, that, as the neighborhood becomes small, the prediction becomes good and thus the trust-region radius remains bounded away from zero and one can obtain, via Assumptions 10.1 and 10.2, similar results to those one is able to obtain in the case with derivatives. As we know from Chapters 3–6, 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 the remainder of this section, we will describe the assumptions on the function and on the models which we use, in this chapter, to prove the global convergence of the derivative-free trust-region algorithms. We will impose only those requirements on the models that are essential for the convergence theory. The models might not necessarily be quadratic functions as mentioned in Section 10.1. (We will cover the use of nonlinear models in trust-region methods in Section 12.2.)

For the purposes of convergence to first-order critical points, we assume that the function f and its gradient are Lipschitz continuous in the domain considered by the algorithms. 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

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

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

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

Thus, what we need are the requirements already stated in Assumption 6.1, making sure, however, that the open domain mentioned there contains the larger set $L_{enl}(x_0)$.

Assumption 10.3. *Suppose x_0 and Δ_{max} are given. Assume that f is continuously differentiable with Lipschitz continuous gradient in an open domain containing the set $L_{enl}(x_0)$.*

The algorithmic framework which will be described in Section 10.3 for interpolation-based trust-region methods also requires the selection of a fully linear class $\mathcal{M}$—see Definition 6.1. We reproduce this definition below in the notation of this chapter (where y is given by $x+s$ and the set S is $L(x_0)$).

Definition 10.3. *Let a function $f : \mathbb{R}^n \to \mathbb{R}$, that satisfies Assumption* 10.3*, 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 the following hold:*

1. *There exist positive constants κ_{ef}, κ_{eg}, and ν_1^m 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 ν_1^m, and such that*

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

   $$\|\nabla f(x+s) - \nabla m(x+s)\| \leq \kappa_{eg}\, \Delta \quad \forall s \in B(0;\Delta), \tag{10.15}$$

   and

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

   $$|f(x+s) - m(x+s)| \leq \kappa_{ef}\, \Delta^2 \quad \forall s \in B(0;\Delta). \tag{10.16}$$

   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 $B(x;\Delta)$.*

For the remainder of this chapter we will assume, without loss of generality, that the constants κ_{ef}, κ_{eg}, and ν_1^m of any fully linear class $\mathcal{M}$ which we use in our algorithmic framework are such that Lemma 10.25 below holds. In this way, we make sure that if a model is fully linear in a ball, it will be so in any larger concentric one, as happens with Taylor models defined by first-order derivatives.

To analyze the convergence to second-order critical points, we require, in addition, the Lipschitz continuity of the Hessian of f. The overall smoothness requirement has been stated in Assumption 6.2, but we need to consider here, however, that the open domain mentioned there now contains the larger set $L_{enl}(x_0)$.

Assumption 10.4. *Suppose x_0 and Δ_{max} are given. Assume that f is twice continuously differentiable with Lipschitz continuous Hessian in an open domain containing the set $L_{enl}(x_0)$.*

The algorithmic framework which will be described in Section 10.5 for interpolation-based trust-region methods also requires the selection of a fully quadratic class $\mathcal{M}$—see Definition 6.2. We repeat this definition below using the notation of this chapter (where y is given by $x+s$ and the set S is $L(x_0)$).

Definition 10.4. *Let a function f, that satisfies Assumption* 10.4*, 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 the following hold:*

1. *There exist positive constants κ_{ef}, κ_{eg}, κ_{eh}, and ν_2^m 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 ν_2^m, and such that*

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

   $$\|\nabla^2 f(x+s) - \nabla^2 m(x+s)\| \leq \kappa_{eh}\,\Delta \quad \forall s \in B(0;\Delta), \tag{10.17}$$

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

   $$\|\nabla f(x+s) - \nabla m(x+s)\| \leq \kappa_{eg}\,\Delta^2 \quad \forall s \in B(0;\Delta), \tag{10.18}$$

   and

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

   $$|f(x+s) - m(x+s)| \leq \kappa_{ef}\,\Delta^3 \quad \forall s \in B(0;\Delta). \tag{10.19}$$

   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 $B(x;\Delta)$.*

For the remainder of this chapter we will assume, without loss of generality, that the constants κ_{ef}, κ_{eg}, κ_{eh}, and ν_2^m of any fully quadratic class $\mathcal{M}$ which we use in our algorithmic framework are such that Lemma 10.26 below holds. By proceeding in this way, we guarantee that if a model is fully quadratic in a ball, it remains so in any larger concentric ball (as in Taylor models defined by first- and second-order derivatives).

10.3 Derivative-free trust-region methods (first order)

Derivative-free trust-region methods can be roughly classified into two categories: the methods which target good practical performance, such as the methods in [163, 190] (see

Chapter 11); and the methods for which global convergence was shown, but at the expense of practicality, such as described in [57, 59]. In this book we try 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 this algorithmic framework closer to a practical one when compared to the methods in [57, 59].

The trust-region maintenance that we will use is different from the approaches in derivative-based methods [57]. In derivative-based methods, under appropriate conditions, the trust-region radius becomes bounded away from zero when the iterates converge to a local minimizer [57]; 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 [190] suggests using 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*. 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 this algorithmic framework is the acceptance of new iterates that provide a 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 [184], the standard liminf-type results are obtained for general trust-region radius updating schemes (such as the simple one described in Section 10.1). 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 [214] in the derivative case, 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 [236]. But, as we will see, because of the update of the trust-region radius at the criticality step mentioned in the previous paragraph, such further provisions are not needed to achieve lim-type global convergence to first-order critical points even when iterates are accepted based on simple decrease.[15]

In our framework it is possible to take steps, and for the algorithm to progress, without insisting that the model be made fully linear or fully quadratic on *every* iteration. In contrast with [57, 59], we require only (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 a 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.

[15] 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 under the presence of derivatives.

We now formally state the first-order version of the algorithm that we consider. The algorithm contemplates acceptance of new iterates based on simple decrease by selecting $\eta_0 = 0$. We have already noted that accepting new iterates when function evaluations are expensive based on simple decrease is particularly appropriate in derivative-free optimization. We also point out that the model m_k and the trust-region radius Δ_k are set only 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 then might be changed or might not by the criticality step.

Algorithm 10.1 (Derivative-free trust-region method (first order)).

Step 0 (initialization): Choose a fully linear class of models $\mathcal{M}$ and a corresponding model-improvement algorithm (see, e.g., Chapter 6). Choose an initial point x_0 and $\Delta_{max} > 0$. We assume that an initial model $m_0^{icb}(x_0 + s)$ (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 ω are also given and satisfy the conditions $0 \le \eta_0 \le \eta_1 < 1$ (with $\eta_1 \ne 0$), $0 < \gamma < 1 < \gamma_{inc}$, $\epsilon_c > 0$, $\mu > \beta > 0$, and $\omega \in (0,1)$. 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 $\|g_k^{icb}\| \le \epsilon_c$, 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 10.2 (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 ν_1^m, which remain the same for all iterations of Algorithm 10.1) on the ball $B(x_k; \tilde{\Delta}_k)$, for some $\tilde{\Delta}_k \in (0, \mu\|\tilde{g}_k\|]$ given by Algorithm 10.2. In such a case set[16]

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

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 (10.9)) 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

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

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

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 ν_1^m) 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 $m_{k+1}^{icb}(x_{k+1}+s)$ (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 10.1) is described in the following algorithm.

Algorithm 10.2 (Criticality step: first order). *This algorithm is applied only 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* $\omega \in (0,1)$ *is chosen at Step* 0 *of Algorithm* 10.1.

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 $B(x_k;\omega^{i-1}\Delta_k^{icb})$ (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 10.1). Denote the new model by $m_k^{(i)}$. Set $\tilde{\Delta}_k = \omega^{i-1}\Delta_k^{icb}$ 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 10.2 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.

Note that if $\|g_k^{icb}\| \leq \epsilon_c$ in the criticality step of Algorithm 10.1 and Algorithm 10.2 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 10.25, m_k is also fully linear on $B(x_k;\Delta_k)$ (as well as on $B(x_k;\mu\|g_k\|)$).

After Step 3 of Algorithm 10.1, 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**.

10.4 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 10.1 (Algorithm 10.2). The proof is very similar to the one in [59], but we repeat the details here for completeness.

Lemma 10.5. *If $\nabla f(x_k) \neq 0$, Step 1 of Algorithm 10.1 will terminate in a finite number of improvement steps (by applying Algorithm 10.2).*

Proof. Assume that the loop in Algorithm 10.2 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 $B(x_k; \omega^0 \Delta_k^{icb})$ (in a finite number of improvement steps). If the gradient $g_k^{(1)}$ of the resulting model $m_k^{(1)}$ satisfies $\mu \|g_k^{(1)}\| \geq \omega^0 \Delta_k^{icb}$, the procedure stops with

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

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

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

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

for all $i \geq 1$, 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 $B(x_k; \omega^{i-1} \Delta_k^{icb})$, (10.15) with $s = 0$ and $x = x_k$ implies that

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

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

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

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

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

Assumption 10.5. *Assume that f is bounded from 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 10.5 and 10.3) and of the existence of fully linear models (Definition 10.3). 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 required only 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 10.6. *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 10.6. *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 kth iteration is successful.

Proof. Since

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

the fraction of Cauchy decrease condition (10.9)–(10.10) immediately gives that

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \frac{\kappa_{fcd}}{2}\|g_k\| \min\left\{\frac{\|g_k\|}{\kappa_{bhm}}, \Delta_k\right\} \;=\; \frac{\kappa_{fcd}}{2}\|g_k\|\Delta_k. \tag{10.20}$$

On the other hand, since the current model is fully linear on $B(x_k;\Delta_k)$, then from the bound (10.16) on the error between the function and the model and from (10.20) 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 10.7. *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 10.1 (independently of whether Algorithm 10.2 has been invoked) that

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

Thus,

$$\Delta_k \geq \min\{\beta\kappa_1, \Delta_k^{icb}\}. \tag{10.21}$$

By Lemma 10.6 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 kth 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, (10.21), 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 10.8. *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, 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 jth 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 be shown 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 (10.15) on the error between the gradients of a fully linear model and the function f and because of the fact that m_{i_j} is fully linear. Finally, the third term can be shown to converge to zero by Lemma 10.6, 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 now prove another useful lemma, namely, that the trust-region radius converges to zero, which is particularly relevant in the derivative-free context.

Lemma 10.9.

$$\lim_{k \to +\infty} \Delta_k = 0. \tag{10.22}$$

Proof. When $\mathcal{S}$ is finite the result is shown in the proof of Lemma 10.8. 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 (10.10), we have that

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

Due to Step 1 of Algorithm 10.1 we have that $\|g_k\| \geq \min\{\epsilon_c, \mu^{-1}\Delta_k\}$; 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, and by using Assumption 10.6, 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 completed if all iterations are successful. Now recall that the trust-region radius can be increased only during a successful iteration, and it can be increased only 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 easily.

Lemma 10.10.

$$\liminf_{k \to +\infty} \|g_k\| = 0. \tag{10.23}$$

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

$$\|g_k\| \geq \kappa_1 \tag{10.24}$$

for some $\kappa_1 > 0$. By Lemma 10.7 we have that $\Delta_k \geq \kappa_2$ for all k. We obtain a contradiction with Lemma 10.9. ∎

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

$$\lim_{i \to +\infty} \|g_{k_i}\| = 0 \tag{10.25}$$

it also holds that

$$\lim_{i \to +\infty} \|\nabla f(x_{k_i})\| = 0. \tag{10.26}$$

Proof. First, we note that, by (10.25), $\|g_{k_i}\| \leq \epsilon_c$ 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 (10.15) 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 (10.26). ∎

Lemmas 10.10 and 10.11 immediately give the following global convergence result.

Theorem 10.12. *Let Assumptions* 10.3*,* 10.5*, and* 10.6 *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 10.13. *Let Assumptions* 10.3*,* 10.5*, and* 10.6 *hold. Then*

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

Proof. We have established by Lemma 10.8 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

$$\|\nabla f(x_{k_i})\| \geq \epsilon_0 > 0 \tag{10.27}$$

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 10.11, 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

$$\epsilon \leq \min\left\{\frac{\epsilon_0}{2(2+\kappa_{eg}\mu)}, \epsilon_c\right\}. \tag{10.28}$$

Lemma 10.10 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

$$\|g_k\| \geq \epsilon \text{ for } k_i \leq k < \ell_i \text{ and } \|g_{\ell_i}\| < \epsilon \tag{10.29}$$

for sufficiently large i.

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

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

where k_i and ℓ_i belong to the two subsequences given above in (10.29).

We know that $\|g_k\| \geq \epsilon$ for $k \in \mathcal{K}$. From Lemma 10.9, $\lim_{k\to+\infty} \Delta_k = 0$, and by Lemma 10.6 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

$$f(x_k) - f(x_{k+1}) \geq \eta_1[m_k(x_k) - m_k(x_k+s_k)] \geq \eta_1 \frac{\kappa_{fcd}}{2}\|g_k\| \min\left\{\frac{\|g_k\|}{\kappa_{bhm}}, \Delta_k\right\}, \tag{10.30}$$

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 from below (Assumption 10.5) 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 10.3), and it is thus bounded by ϵ for i sufficiently large. The third term is bounded by ϵ by (10.29). For the second term we use the fact that from (10.28) 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 (10.15) and (10.29), 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 (10.28) that

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

for i large enough, which contradicts (10.27). Hence, our initial assumption must be false, and the theorem follows. □

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.

10.5 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 [57, Section 9.3], one possibility is to work with

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

which measures the second-order stationarity of the model.

The algorithm follows mostly the same arguments as those of Algorithm 10.1. 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 that seek second-order convergence results [57], we do not increase the trust-region radius on *every* successful

iteration. We insist on such an increase only when the size of the trust-region radius is small when compared to the measure of stationarity.

We state the version of the algorithm we wish to consider.

Algorithm 10.3 (Derivative-free trust-region method (second order)).

Step 0 (initialization): Choose a fully quadratic class of models $\mathcal{M}$ and a corresponding model-improvement algorithm (see, e.g., Chapter 6). Choose an initial point x_0 and $\Delta_{max} > 0$. We assume that an initial model $m_0^{icb}(x_0+s)$ (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 ω are also given and satisfy the conditions $0 \le \eta_0 \le \eta_1 < 1$ (with $\eta_1 \neq 0$), $0 < \gamma < 1 < \gamma_{inc}$, $\epsilon_c > 0$, $\mu > \beta > 0$, and $\omega \in (0,1)$. 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 $\sigma_k^{m,icb} \le \epsilon_c$, 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 10.4 (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 ν_2^m, which remain the same for all iterations of Algorithm 10.3) on the ball $B(x_k; \tilde{\Delta}_k)$ for some $\tilde{\Delta}_k \in (0, \mu\tilde{\sigma}_k^m]$ given by Algorithm 10.4. In such a case set[17]

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

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 (10.13)) 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 $\rho_k \ge \eta_1$ or if both $\rho_k \ge \eta_0$ and the model is fully quadratic (for the positive constants κ_{ef}, κ_{eg}, κ_{eh}, and ν_2^m) 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 $m_{k+1}^{icb}(x_{k+1}+s)$ (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 $x_{k+1} = x_k$).

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

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 10.3 the definitions of **successful**, **acceptable**, **model improving**, and **unsuccessful** iterations which we stated for the sequence of iterations generated by Algorithm 10.1. We will use the same definitions here, adapted to the fully 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 be only a finite number of model-improvement steps before a fully quadratic model is obtained. The comments outlined after Theorem 10.13 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 10.2, essentially by replacing $\|g_k\|$ by σ_k^m and by using fully quadratic models rather than fully linear ones.

Algorithm 10.4 (Criticality step: second order). *This algorithm is applied only 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 $\omega \in (0,1)$ is chosen at Step 0 of Algorithm 10.3.*

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 $B(x_k; \omega^{i-1}\Delta_k^{icb})$ (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 10.3). Denote the new model by $m_k^{(i)}$. Set $\tilde{\Delta}_k = \omega^{i-1}\Delta_k^{icb}$ 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 10.3 and Algorithm 10.4 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 10.26, m_k is also fully quadratic on $B(x_k; \Delta_k)$ (as well as on $B(x_k; \mu\sigma_k^m)$).

10.6 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 10.4 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 10.5). 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 10.14. *Suppose that Assumption* 10.4 *holds and m is a fully quadratic model on $B(x;\Delta)$. Then we have that*

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

Proof. The proof follows directly from the bound (10.17) 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 would 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 10.15. *Let Δ be bounded by Δ_{max}. Suppose that Assumption* 10.4 *holds and m is a fully quadratic model on $B(x;\Delta)$. Then we have, for some $\kappa_\sigma > 0$, that*

$$|\sigma(x) - \sigma^m(x)| \;\leq\; \kappa_\sigma\,\Delta. \tag{10.31}$$

Proof. It follows that

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

The first argument $|\,\|\nabla f(x)\| - \|\nabla m(x)\|\,|$ is bounded from above by $\kappa_{eg}\Delta_{max}\Delta$, because of the error bound (10.18) 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 from above by $\kappa_{eh}\Delta$ because of Proposition 10.14. Finally, we need only write $\kappa_\sigma = \max\{\kappa_{eg}\Delta_{max}, \kappa_{eh}\}$, and the result follows. ☐

The convergence theory will require the already mentioned assumptions (Assumptions 10.4 and 10.5), as well as the uniform upper bound on the Hessians of the quadratic models (Assumption 10.6).

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 10.16. *If $\sigma(x_k) \neq 0$, Step 1 of Algorithm 10.3 will terminate in a finite number of improvement steps (by applying Algorithm 10.4).*

Proof. The proof is practically identical to the proof of Lemma 10.5, 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 10.17. *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 kth iteration is successful.

Proof. The proof is similar to the proof of Lemma 10.6 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 (10.14),

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

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

$$m_k(x_k) - m_k(x_k+s_k) \geq \frac{\kappa_{fod}}{2}\|g_k\|\Delta_k = \frac{\kappa_{fod}}{2}\sigma_k^m\Delta_k. \tag{10.32}$$

On the other hand, since the current model is fully quadratic on $B(x_k;\Delta_k)$, we may deduce from (10.32) and the bound (10.19) 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

$$m_k(x_k) - m_k(x_k+s_k) \geq -\tfrac{\kappa_{fod}}{2}\tau_k\Delta_k^2 = \tfrac{\kappa_{fod}}{2}\sigma_k^m\Delta_k^2. \tag{10.33}$$

But, since the current model is fully quadratic on $B(x_k;\Delta_k)$, we deduce from (10.33) and the bound (10.19) 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 10.17.

Lemma 10.18. *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 a combination of Lemma 10.17 and the proof of Lemma 10.7. □

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

Lemma 10.19. *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 10.8 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 10.15 and 10.17. □

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

Lemma 10.20.

$$\lim_{k \to +\infty} \Delta_k = 0. \tag{10.34}$$

Proof. When $\mathcal{S}$ is finite the proof is as in the proof of Lemma 10.8 (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 10.3 we have that $\sigma_k^m \geq \min\{\epsilon_c, \mu^{-1}\Delta_k\}$. If on iteration k we have $\|g_k\| \geq \max\{-\tau_k, 0\} = \{-\lambda_{min}(H_k), 0\}$, then $\sigma_k^m = \|g_k\|$ and

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \frac{\kappa_{fod}}{2} \min\{\epsilon_c, \mu^{-1}\Delta_k\} \min \left\{ \frac{\min\{\epsilon_c, \mu^{-1}\Delta_k\}}{\kappa_{bhm}}, \Delta_k \right\}. \tag{10.35}$$

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

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \frac{\kappa_{fod}}{2} \min\{\epsilon_c, \mu^{-1}\Delta_k\} \Delta_k^2. \tag{10.36}$$

There are two subsequences of successful iterations, possibly overlapping, $\{k_i^1\}$, for which (10.35) holds, and $\{k_i^2\}$, for which (10.36) 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 (10.35) (resp., (10.36)) has to converge to zero. Hence, $\lim_{k \in \mathcal{S}} \Delta_k = 0$, and the proof is completed if all iterations are successful. Now recall that the trust-region radius can be increased only during a successful iteration, and it can be increased only 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 10.21.

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

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 10.18 there exists a constant κ_2 such that $\Delta_k \geq \kappa_2$ for all k. We obtain contradiction with Lemma 10.20. ∎

We now verify that the criticality step (Step 1 of Algorithm 10.3) ensures that a subsequence of the iterates approaches second-order stationarity, by means of the following auxiliary result.

Lemma 10.22. *For any subsequence* $\{k_i\}$ *such that*

$$\lim_{i \to +\infty} \sigma_{k_i}^m = 0 \tag{10.37}$$

it also holds that

$$\lim_{i \to +\infty} \sigma(x_{k_i}) = 0. \tag{10.38}$$

Proof. From (10.37), $\sigma_{k_i}^m \leq \epsilon_c$ for i sufficiently large. The mechanism of the criticality step (Step 1) then ensures 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 (10.31),

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

The limit (10.37) and this last bound then give (10.38). ∎

Lemmas 10.21 and 10.22 immediately give the following global convergence result.

Theorem 10.23. *Let Assumptions* 10.4*,* 10.5*, and* 10.6 *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 of Algorithm 10.3, 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 10.24. *Let Assumptions* 10.4*,* 10.5*, and* 10.6 *hold. Then*

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

Proof. We have established by Lemma 10.19 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

$$\sigma(x_{k_i}) \geq \epsilon_0 > 0 \tag{10.39}$$

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 10.22, 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

$$\epsilon \leq \min\left\{\frac{\epsilon_0}{2(2+\kappa_\sigma\mu)}, \epsilon_c\right\}. \tag{10.40}$$

Lemma 10.21 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

$$\sigma_k^m \geq \epsilon \text{ for } k_i \leq k < \ell_i \text{ and } \sigma_{\ell_i}^m < \epsilon \tag{10.41}$$

for sufficiently large i.

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

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

where k_i and ℓ_i belong to the two subsequences defined above in (10.41).

From Lemmas 10.17 and 10.20, just as in the proof of Theorem 10.13, it follows that for large enough $k \in \mathcal{K}$ the kth 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 reduced only when $\Delta_{k+1} \geq \beta\tilde{\sigma}_{k+1}^m = \beta\sigma_{k+1}^m \geq \beta\epsilon$, but this can happen only 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,

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

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

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

The first term of the right-hand side tends to zero because of the Lipschitz continuity of $\sigma(x)$, and it is thus bounded by ϵ for i sufficiently large. The third term is bounded by ϵ by (10.41). For the second term we use the fact that, from (10.40) and the mechanism of the criticality step (Step 1) at iteration ℓ_i, the model m_{ℓ_i} is fully quadratic on $B(x_{\ell_i}; \mu\sigma^m_{\ell_i})$. Using (10.31) and (10.41), 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 (10.40) that

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

for i large enough, which contradicts (10.39). Hence, our initial assumption must be false, and the theorem follows. □

10.7 Model accuracy in larger concentric balls

We will show next that if a model is fully linear on $B(x;\bar{\Delta})$ with respect to some (large enough) constants κ_{ef}, κ_{eg}, and v_1^m 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 was needed in this chapter for the analysis of the global convergence properties of the trust-region methods. Such a property reproduces what is known for Taylor models and is a clear indication of the appropriateness of the definition of fully linear models.

Lemma 10.25. *Consider a function f satisfying Assumption* 6.1 *and a model m fully linear, with respect to constants κ_{ef}, κ_{eg}, and v_1^m 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 v_1^m and the Lipschitz constant of the gradient of f, and that $\kappa_{ef} \geq (1/2)\kappa_{eg}$.

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 v_1^m.

Proof. We start by considering any $\Delta \in [\bar{\Delta}, \Delta_{max}]$. Then we consider an s such that $\bar{\Delta} \leq \|s\| \leq \Delta$ and let $\theta = \bar{\Delta}/\|s\|$. Since $x+\theta s \in B(x;\bar{\Delta})$ and the model is fully linear on $B(x;\bar{\Delta})$, we obtain

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

By using the Lipschitz continuity of ∇f and ∇m and the assumption that κ_{eg} is no less than the sum of the corresponding Lipschitz constants, we derive

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

Thus, by combining the above expressions we obtain

$$\|\nabla f(x+s) - \nabla m(x+s)\| \leq \kappa_{eg}\|s\| \leq \kappa_{eg}\Delta. \tag{10.42}$$

In the second part of the proof, we consider the function $\phi(\alpha) = f(x+\alpha s) - m(x+\alpha s)$, $\alpha \in [0,1]$. We want to bound $|\phi(1)|$. From the fact that m is a fully linear model on $B(x;\bar{\Delta})$, we have $|\phi(\theta)| \leq \kappa_{ef}\bar{\Delta}^2$. To bound $|\phi(1)|$, we bound $|\phi(1) - \phi(\theta)|$ first by

using (10.42):

$$\left|\int_\theta^1 \phi'(\alpha)d\alpha\right| \le \int_\theta^1 \|s\|\|\nabla f(x+\alpha s)-\nabla m(x+\alpha s)\|\, d\alpha$$
$$\le \int_\theta^1 \alpha\kappa_{eg}\|s\|^2 d\alpha \;=\; (1/2)\kappa_{eg}(\|s\|^2-\bar\Delta^2).$$

Using the assumption $\kappa_{ef} \ge (1/2)\kappa_{eg}$, we finally get

$$|f(x+s)-m(x+s)| \;\le\; |\phi(1)-\phi(\theta)|+|\phi(\theta)| \;\le\; \kappa_{ef}\|s\|^2 \;\le\; \kappa_{ef}\Delta^2. \quad \square$$

Similar to the linear case, 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 v_2^m 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 10.26. *Consider a function f satisfying Assumption* 6.2 *and a model m fully quadratic, with respect to constants κ_{ef}, κ_{eg}, κ_{eh}, and v_2^m on $B(x;\bar\Delta)$, with $x \in L(x_0)$ and $\bar\Delta \le \Delta_{max}$.*

Assume also, without loss of generality, that κ_{eh} is no less than the sum of v_2^m and the Lipschitz constant of the Hessian of f, and that $\kappa_{eg} \ge (1/2)\kappa_{eh}$ and $\kappa_{ef} \ge (1/3)\kappa_{eg}$.

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 v_2^m.

Proof. Let us consider any $\Delta \in [\bar\Delta,\Delta_{max}]$. Consider, also, an s such that $\bar\Delta \le \|s\| \le \Delta$, 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

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

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

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

Thus, by combining the above expressions we obtain

$$\|\nabla^2 f(x+s)-\nabla^2 m(x+s)\| \;\le\; \kappa_{eh}\|s\| \;\le\; \kappa_{eh}\Delta. \tag{10.43}$$

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 $\|g(\theta)\| \le \kappa_{eg}\bar\Delta^2$. 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

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

Now using (10.43) we have

$$\int_\theta^1 \|g'(\alpha)\| d\alpha \le \int_\theta^1 \|s\| \|\nabla^2 f(x+\alpha s) - \nabla^2 m(x+\alpha s)\| \, d\alpha$$
$$\le \int_\theta^1 \alpha \kappa_{eh} \|s\|^2 d\alpha \;=\; (1/2)\kappa_{eh}(\|s\|^2 - \bar{\Delta}^2).$$

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

$$\|\nabla f(x+s) - \nabla m(x+s)\| \;\le\; \|g(1) - g(\theta)\| + \|g(\theta)\| \;\le\; \kappa_{eg}\|s\|^2 \;\le\; \kappa_{eg}\Delta^2. \quad (10.44)$$

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 $|\phi(\theta)| \le \kappa_{ef}\bar{\Delta}^3$. We are interested in bounding $|\phi(1)|$, which can be achieved by bounding $|\phi(1) - \phi(\theta)|$ first by using (10.44):

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

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

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

The proof is complete. □

10.8 Trust-region subproblem

An extensive and detailed analysis as to how the trust-region subproblem (10.2) can be solved more or less exactly when the model function is quadratic is given in [57, Chapter 7] for the ℓ_∞ and ℓ_2 trust-region norms. Of course, this is at some computational cost over the approximate solutions (satisfying, for instance, Assumptions 10.1 and 10.2), but particularly in the context of modestly dimensioned domains, such as one might expect in derivative-free optimization, the additional work may well be desirable at times because of the expected faster convergence rate. Although we will not go into the same level of detail, it does seem appropriate to at least indicate how one can solve the more popular ℓ_2-norm trust-region subproblem.

The basic driver, as one might expect for such a relatively simple problem, is the optimality conditions. However, for this specially structured problem we have much more than just the first-order necessary conditions in that we are able to characterize the global solution(s) of a (possibly) nonconvex problem.

Theorem 10.27. *Any global minimizer s_* of $m(x+s) = m(x) + s^\top g + \frac{1}{2}s^\top H s$, subject to $\|s\| \le \Delta$, satisfies the equation*

$$[H + \lambda_* I]s_* \;=\; -g, \quad (10.45)$$

where $H + \lambda_* I$ is positive semidefinite, $\lambda_* \geq 0$, and

$$\lambda_*(\|s_*\| - \Delta) = 0. \tag{10.46}$$

If $H + \lambda_* I$ is positive definite, then s_* is unique.

If Δ is large enough and H is positive definite, the complementarity conditions (10.46) are satisfied with $\lambda_* = 0$ and the unconstrained minimum lies within the trust region. In all other circumstances, a solution lies on the boundary of the trust region and $\|s_*\| = \Delta$. Suppose that H has an eigendecomposition

$$H = QEQ^\top,$$

where E is a diagonal matrix of eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, and Q is an orthogonal matrix of associated eigenvectors. Then

$$H + \lambda I = Q(E + \lambda I)Q^\top.$$

Theorem 10.27 indicates that the value of λ we seek must satisfy $\lambda \geq -\lambda_1$ (as only then is $H + \lambda I$ positive semidefinite), and, if $\lambda > -\lambda_1$, the model minimizer is unique (as this ensures that $H + \lambda I$ is positive definite).

Suppose that $\lambda > -\lambda_1$. Then $H + \lambda I$ is positive definite, and thus (10.45) has a unique solution,

$$s(\lambda) = -[H + \lambda I]^{-1} g = -Q(E + \lambda I)^{-1} Q^\top g.$$

However, the solution we are looking for depends upon the nonlinear inequality

$$\|s(\lambda)\| \leq \Delta.$$

Now

$$\|s(\lambda)\|^2 = \|Q(E + \lambda I)^{-1} Q^\top g\|^2 = \|(E + \lambda I)^{-1} Q^\top g\|^2 = \sum_{i=1}^{n} \frac{\gamma_i^2}{(\lambda_i + \lambda)^2},$$

where γ_i is $(Q^\top g)_i$, the ith component of $Q^\top g$. It is now apparent that if $\lambda > -\lambda_1$, then $\|s(\lambda)\|$ is a continuous, nonincreasing function of λ on $(-\lambda_1, +\infty)$ that tends to zero as λ tends to $+\infty$. Moreover, provided $\gamma_j \neq 0$, then $\lim_{\lambda \to -\lambda_j} \|s(\lambda)\| = +\infty$. Thus, provided $\gamma_1 \neq 0$, $\|s(\lambda)\| = \Delta$ for a unique value of $\lambda \in (-\lambda_1, +\infty)$.

When H is positive definite and $\|H^{-1} g\| \leq \Delta$ the solution corresponds to $\lambda = 0$, as we have already mentioned. Otherwise, when H is positive definite and $\|H^{-1} g\| > \Delta$ there is a unique solution to (10.45) in $(0, +\infty)$.

When H is not positive definite and $\gamma_1 \neq 0$ we need to find a solution to (10.45) with $\lambda > -\lambda_1$. Because of the high nonlinearities in the neighborhood of $-\lambda_1$ it turns out to be preferable to solve the so-called secular equation

$$\frac{1}{\|s(\lambda)\|} = \frac{1}{\Delta},$$

which is close to linear in the neighborhood of the optimal λ. Because of the near linearity in the region of interest, it is reasonable to expect fast convergence of Newton's method.

But, as is well known, an unsafeguarded Newton method may fail to converge, so care and ingenuity must be taken to safeguard the method.

When H is not positive definite and $\gamma_1 = 0$ we have the so-called hard case, since there is no solution to (10.45) in $(-\lambda_1, +\infty)$ when $\Delta > \|s(-\lambda_1)\|$. However, there is a solution at $\lambda = -\lambda_1$, but it includes an eigenvector of H corresponding to the eigenvalue λ_1, which thus has to be estimated.

The reader is referred to [57] for all the (considerable) details. A more accessible but less complete reference is [178, Chapter 4].

10.9 Other notes and references

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. The idea of minimizing a quadratic interpolation model within a *region of validity* goes back to Winfield [227, 228] in 1969. Glad and Goldstein [106] have also suggested minimizing regression quadratic models obtained by sampling over sets defined by positive integer combinations of $D_{\oplus}$, as in directional direct-search methods. However, the design and analysis of 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. The first attempts in these directions have been presented by Powell in the 5th Stockholm Optimization Days in 1994 and in the 5th SIAM Conference on Optimization in 1996, using quadratic interpolation. Conn and Toint [58] in 1996 have reported encouraging numerical results for quadratic interpolation models. Around the same time, Elster and Neumaier [88] developed and analyzed an algorithm based on the minimization of quadratic regression models within trust regions built by sampling over box-type grids, also reporting good numerical results.

Conn, Scheinberg, and Toint [59] (see also [57]) introduced the criticality step and designed and analyzed the first interpolation-based derivative-free trust-region method globally convergent to first-order critical points. Most of the other issues addressed in this chapter, including the appropriate incorporation of fully linear and fully quadratic models, global convergence when acceptance of iterates is based on simple decrease of the objective function, and global convergence for second-order critical points, were addressed by Conn, Scheinberg, and Vicente [62]. This paper provided the first comprehensive analysis of global convergence of trust-region derivative-free methods to second-order stationary points. It was mentioned in [57] that such analysis could be simply derived from the classical analysis for the derivative-based case. However, as we remarked during this chapter, the algorithms in [57, 59] are not as close to a practical one as the one described in this chapter, 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. As we have seen in Sections 10.5 and 10.6, it is not necessary to increase the trust-region radius on every successful iteration, as 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, as described in these sections, 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 small.

Other authors have addressed related issues. Colson [55] and Colson and Toint [56] showed how to take advantage of partial separability of functions in the development and implementation of interpolation-based trust-region methods. The wedge algorithm of Marazzi and Nocedal [163] and the least Frobenius norm updating algorithm of Powell [191] will be covered in detail in Chapter 11.

Finally, we would like to point out that derivative-based trust-region methods have been analyzed under the influence of inexactness of gradient values [51, 57] and inexactness of function values [57]. The influence of inexact function values, in particular, is relevant also in the derivative-free case since the objective function can be subject to noise or inaccuracy.

10.10 Exercises

1. Prove that when H_k has a negative eigenvalue, s_k^{E} is the minimizer of the quadratic model $m_k(x_k + s)$ along that direction and inside the trust region $B(0; \Delta_k)$.

2. Prove that if s_k satisfies a fraction of optimal decrease condition,

$$m_k(x_k) - m_k(x_k + s_k) \;\geq\; \kappa_{fod}[m_k(x_k) - m_k(x_k + s_k^*)],$$

   where $\kappa_{fod} \in (0, 1]$ and s_k^* is an optimal solution of (10.2), then it also satisfies (10.14).

3. Show that when the sequence of iterates is bounded, Theorem 10.12 (resp., Theorem 10.23) implies the existence of one limit point of the sequence of iterates $\{x_k\}$ that is a first-order (resp., second-order) stationary point.

4. Prove Lemma 10.16.