

RECURSIVE TRUST-REGION METHODS FOR
MULTILEVEL NONLINEAR OPTIMIZATION (PART I):
GLOBAL CONVERGENCE AND COMPLEXITY

by S. Gratton¹, A. Sartenaer² and Ph. L. Toint²

Report 04/06

August 11, 2004

¹ CERFACS,
av. G. Coriolis, Toulouse, France,
Email: serge.gratton@cerfacs.fr

² Department of Mathematics,
University of Namur,
61, rue de Bruxelles, B-5000 Namur, Belgium,
Email: annick.sartenaer@fundp.ac.be, philippe.toint@fundp.ac.be

Recursive Trust-Region Methods for Multilevel Nonlinear Optimization (Part I): Global Convergence and Complexity

Serge Gratton, Annick Sartenaer and
Philippe L. Toint

10 August 2004

Abstract

A class of trust-region methods is presented for solving unconstrained nonlinear and possibly nonconvex discretized optimization problems, like those arising in systems governed by partial differential equations. The algorithms in this class make use of the discretization level as a mean of speeding up the computation of the step. This use is recursive, leading to true multilevel optimization methods reminiscent of multigrid methods in linear algebra and the solution of partial-differential equations. Global convergence of the recursive algorithm is proved to first-order stationary points on the fine grid. A new theoretical complexity result is also proved for single- as well as multilevel trust-region algorithms, that gives a bound on the number of iterations that are necessary to reduce the norm of the gradient below a given threshold.

Keywords: nonlinear optimization, multilevel problems, simplified models, recursive algorithms, convergence theory.

1 Introduction

Many large-scale finite-dimensional optimization problems arise from the discretization of infinite-dimensional problems, a primary example being optimal-control problems defined in terms of either ordinary or partial differential equations. While the direct solution of such problems for a discretization level yielding the desired accuracy is often possible using existing packages for large-scale numerical optimization, this technique typically does make very little use of the fact that there is an underlying infinite-dimensional problem for which several discretization levels are possible, and the approach thus rapidly becomes cumbersome. This observation motivates the developments presented here, where we explore the theoretical properties of a class of algorithms which, at variance with the technique just described, makes explicit use of this fact in the hope to allow better efficiency and, possibly, enhance reliability.

Using the different possible levels⁽¹⁾ of discretization for an infinite-dimensional problem is not a new idea. A simple first approach is to use coarser grids in order to compute approximate solutions which can then be used as starting points for the optimization problem on a finer grid (see Griewank and Toint, 1982, Banks, Gill and Marcia, 2003, Betts and Erb, 2003 or Benson, McInnes, Moré and Sarich, 2004, for instance). However, potentially more efficient techniques are inspired from the multigrid paradigm in the solution of partial differential equations and associated systems of linear algebraic equations (see, for example, Brandt, 1977, Bramble, 1993, Hackbusch, 1994, Hackbusch, 1995 or Briggs, Henson and McCormick, 2000, for descriptions and references for this much studied topic), and have only been discussed relatively recently in the optimization community. The work presented in this note was in particular motivated by the “generalized truncated Newton algorithm” presented in Fisher (1998), a talk by Moré (2003) and the contributions by Nash (1999) and Lewis and Nash (2003). These latter two papers present the description of MG/OPT, a linesearch-based recursive algorithm, an outline of its convergence properties and impressive numerical results. The generalized truncated Newton algorithm and MG/OPT are very similar and, like many linesearch methods, naturally suited to convex problems. Their extension to nonconvex ones is possible but requires an explicit convexification of the subproblems, a technique which is always delicate as one has to guess a numerically suitable convexification level.

The class of algorithms discussed in this note can be viewed as an alternative where one uses the trust-region technology whose efficiency and reliability in the solution of nonconvex problems is well-known (we refer the reader to Conn, Gould and Toint, 2000 for a more complete coverage of this subject). Our developments are organized as follows. We first describe our class of multilevel trust-region algorithms in Section 2, and show in Section 3 that it is well defined and globally convergent to first-order critical points. A new complexity result for local minimization by trust-region methods is also derived. Some conclusions are drawn and perspectives presented in Section 4.

2 Recursive multilevel trust-region algorithms

We start by considering the solution of the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2.1}$$

where f is a twice-continuously differentiable objective function which maps \mathbb{R}^n into \mathbb{R} and is bounded below. The trust-region methods which we investigate are iterative and, given an initial point x_0 , produce a sequence $\{x_k\}$ of iterates (hopefully) converging to

⁽¹⁾We stress that the multilevel optimization problems explored in this paper are different in spirit from bilevel (and hence, possibly, multilevel) programs, when these are understood as optimization problems where “the constraint domain (...) is implicitly determined by a series of optimization problems which must be solved in a predetermined sequence” (see Migdalas, Pardalos and Värbrand, 1998). The interested reader is referred to the volume just cited for further developments and references.

a local stationary point for the problem, i.e., to a point where $g(x) \stackrel{\text{def}}{=} \nabla_x f(x) = 0$. At each iterate x_k , classical trust-region methods build a model $m_k(x_k + s)$ of $f(x_k + s)$. This model is then assumed to be adequate in a “trust region”, defined as a sphere of radius $\Delta_k > 0$ centered at x_k , and a step s_k is then computed that sufficiently reduces this model in the region. The objective function is computed at the trial point $x_k + s_k$ and this trial point is accepted as the next iterate if and only if the ratio

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

is larger than a small positive constant η_1 . The value of the radius is finally updated to ensure that it is decreased when the trial point cannot be accepted as the next iterate, and is increased or unchanged if ρ_k is sufficiently large. In many practical trust-region algorithms, the model $m_k(x_k + s)$ is quadratic and takes the form

$$m_k(x_k + s) = f(x_k) + \langle g_k, s \rangle + \frac{1}{2} \langle s, H_k s \rangle, \quad (2.2)$$

where $g_k \stackrel{\text{def}}{=} \nabla_x f(x_k)$, H_k is a symmetric $n \times n$ approximation of $\nabla_{xx} f(x_k)$ and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. Obtaining sufficient decrease on this model then amounts to (approximately) solving

$$\min_{\|s\| \leq \Delta_k} m_k(x_k + s) = \min_{\|s\| \leq \Delta_k} f(x_k) + \langle g_k, s \rangle + \frac{1}{2} \langle s, H_k s \rangle, \quad (2.3)$$

where $\|\cdot\|$ is the Euclidean norm.

Such methods are known to be efficient and reliable. They provably converge to first-order critical points whenever the sequence $\{\|H_k\|\}$ is uniformly bounded above, i.e., when there is a constant $\kappa_H \geq 1$ such that $1 + \|H_k\| \leq \kappa_H$ for all k . Besides computing the value $f(x_k + s_k)$, the cost per iteration is dominated by the numerical solution of the subproblem (2.3), which involves either a few symmetric matrix factorizations or the application of a (preconditioned) conjugate gradient procedure. In both cases, the dimension n of the problem is crucial. If we now assume that (2.1) results from the discretization of some infinite-dimensional problem on a relatively fine grid, it is clear that n is typically large, since it usually grows as some power of the inverse of the mesh-size. The cost of solving (2.1) by the approach described above is therefore often significant.

In what follows, we investigate what can be done to reduce this cost if one attempts to exploit the knowledge of alternative simplified expressions of the objective function, when available. More specifically, we assume that we know a collection of functions $\{f_i\}_{i=0}^r$ such that each f_i is a twice continuously differentiable function from \mathbb{R}^{n_i} to \mathbb{R} (with $n_i \leq n_{i-1}$), the connection with our original problem being that $n_r = n$ and $f_r(x) = f(x)$ for all $x \in \mathbb{R}^n$. We will also assume that, for each $i = 1, \dots, r$, f_i is “more costly” to minimize than f_{i-1} . This may be because f_i has more variables than f_{i-1} (as would typically be the case if the f_i represent increasingly finer discretizations of the same infinite-dimensional objective), or because the structure (in terms of partial separability, sparsity or eigenstructure) of f_i is more complex than that of f_{i-1} , or for

any other reason. To fix terminology, we will refer to a particular i as a *level*. We use the first subscript i in all subsequent subscripted symbols to denote a quantity corresponding to the i -th level (meaning in particular, if applied to a vector, that this vector belongs to \mathbb{R}^{n_i}).

Of course, for f_{i-1} to be useful at all in minimizing f_i , there should be some relation between the variables of these two functions. We henceforth assume that, for each $i = 1, \dots, r$, there exists a full-rank linear operator R_i from \mathbb{R}^{n_i} into $\mathbb{R}^{n_{i-1}}$ (the restriction) and another full-rank operator P_i from $\mathbb{R}^{n_{i-1}}$ into \mathbb{R}^{n_i} (the prolongation) such that

$$P_i = R_i^T. \quad (2.4)$$

This last assumption is realistic and is often expressed, in the context of multigrid algorithms, in the form $\sigma_i P_i = R_i^T$ for some known constants $\sigma_i > 0$, where P_i and R_i are interpreted as restriction and prolongation between a fine and a coarse grid (see, for instance, see Briggs et al., 2000). This alternative form was also used in Nash (1999). In order to keep our notation simple, we nevertheless prefer to use condition (2.4), which can be directly obtained from the more usual form by scaling P_i and/or R_i .

The idea is then to use f_{r-1} to construct an alternative model h_{r-1} for $f_r = f$ in the neighbourhood of the current iterate, that is cheaper than (2.2), and to use this alternative model to define the step in the trust-region algorithm whenever possible. If more than two levels are available ($r > 1$), this can be done recursively, the approximation process stopping at level 0, where (2.2) is always used. The notation necessary to fully describe this procedure is unfortunately very cumbersome, and we have decided to use a simple technique where quantities of interest have a double subscript i, k . The first, i , is the level index ($0 \leq i \leq r$) and the second, k , the index of the current iteration *within level i* , and is *reset to 0 each time level i is entered*⁽²⁾.

Consider now some iteration k at level i (with current iterate $x_{i,k}$) and suppose that one decides to use the lower level model h_{i-1} based on f_{i-1} to compute a step. The first task is to restrict $x_{i,k}$ to create the starting iterate $x_{i-1,0}$ at level $i-1$, that is

$$x_{i-1,0} = R_i x_{i,k}. \quad (2.5)$$

We then define the lower level model as the function

$$h_{i-1}(x_{i-1,0} + s_{i-1}) \stackrel{\text{def}}{=} f_{i-1}(x_{i-1,0} + s_{i-1}) + \langle v_{i-1}, s_{i-1} \rangle \quad (2.6)$$

where

$$v_{i-1} = R_i g_{i,k} - \nabla_{x_{i-1}} f_{i-1}(x_{i-1,0}) \quad (2.7)$$

with $g_{i,k} \stackrel{\text{def}}{=} \nabla_{x_i} h_i(x_{i,k})$. By convention, we set $v_r = 0$, such that

$$h_r(x_{r,0} + s_r) = f_r(x_{r,0} + s_r) = f(x_0 + s) \quad \text{and} \quad g_{r,k} = \nabla_{x_r} h_r(x_{r,k}) = \nabla_x f(x_k) = g_k.$$

⁽²⁾We are well aware that this creates some ambiguities, since a sequence of indices i, k can occur more than once if level i ($i < r$) is used more than once, implying the existence of more than one starting iterate $x_{i,0}$ at this level.

The function h_i therefore corresponds to a modification of f_i by a linear term that enforces the relation

$$g_{i-1,0} = \nabla_{x_{i-1}} h_{i-1}(x_{i-1,0}) = R_i g_{i,k}. \quad (2.8)$$

Because our technique involves minimizing h_i at level i , we need this function to be bounded below. We therefore make this assumption for every $i = 0, \dots, r$. The first-order modification (2.6) is not unusual in multigrid applications in the context of the “full approximation scheme” (see, for instance, Chapter 3 of Briggs et al., 2000 or Hemker and Johnson, 1987) and was also used in Fisher (1998) and Nash (1999). It will play an crucial role in our development, since it ensures that the first-order behaviours of h_i and h_{i-1} are coherent in a neighbourhood of $x_{i,k}$ and $x_{i-1,0}$, respectively: indeed, one verifies that, if s_i and s_{i-1} satisfy

$$s_i = P_i s_{i-1}, \quad (2.9)$$

then

$$\langle g_{i,k}, s_i \rangle = \langle g_{i,k}, P_i s_{i-1} \rangle = \langle R_i g_{i,k}, s_{i-1} \rangle = \langle g_{i-1,0}, s_{i-1} \rangle \quad (2.10)$$

where we have used (2.9), (2.4) and (2.8) successively.

Our task, when entering level $i = 0, \dots, r$, is then to (locally) minimize h_i starting from $x_{i,0}$. At iteration k of this minimization, we first choose, at iterate $x_{i,k}$, a model between $h_{i-1}(x_{i-1,0} + s_{i-1})$ (given by (2.6)) and

$$m_{i,k}(x_{i,k} + s_i) = h_i(x_{i,k}) + \langle g_{i,k}, s_i \rangle + \frac{1}{2} \langle s_i, H_{i,k} s_i \rangle \quad (2.11)$$

where the latter is the usual truncated Taylor series in which $H_{i,k}$ is a symmetric $n_i \times n_i$ approximation to the second derivatives of h_i (which is also the second derivative of f_i) at $x_{i,k}$ such that, for some $\kappa_H \geq 1$,

$$1 + \|H_{i,k}\| \leq \kappa_H \quad (2.12)$$

for all k and all $i = 0, \dots, r$. Once the model is chosen (we will return to the conditions of this choice below), we then compute a step $s_{i,k}$ that generates a “sufficient decrease” on this model within a trust region defined by

$$\mathcal{B}_{i,k} \stackrel{\text{def}}{=} \{s_i \mid \|s_i\|_i \leq \Delta_{i,k}\}, \quad (2.13)$$

for some trust-region radius $\Delta_{i,k} > 0$. The norm $\|\cdot\|_i$ in this last expression is level-dependent and defined, for some symmetric positive definite matrix M_i , as

$$\|s_i\|_i \stackrel{\text{def}}{=} \sqrt{\langle s_i, M_i s_i \rangle} \stackrel{\text{def}}{=} \|s_i\|_{M_i}. \quad (2.14)$$

The “sufficient decrease” of the model $m_{i,k}$ is understood here in its usual meaning for trust-region methods, which is to say that $s_{i,k}$ is such that

$$m_{i,k}(x_{i,k}) - m_{i,k}(x_{i,k} + s_{i,k}) \geq \kappa_{\text{red}} \|g_{i,k}\| \min \left[\frac{\|g_{i,k}\|}{1 + \|H_{i,k}\|}, \Delta_{i,k} \right] \quad (2.15)$$

for some constant $\kappa_{\text{red}} \in (0, 1)$. This condition is known as the ‘‘Cauchy point’’ condition. Chapter 7 of Conn et al. (2000) reviews several techniques that enforce it, including the exact minimization of $m_{i,k}$ within the trust region or an approximate minimization using (possibly preconditioned) Krylov space methods.

If the model (2.11) is chosen⁽³⁾, this is nothing but a usual ellipsoidal trust-region subproblem solution yielding a step $s_{i,k}$. On the other hand, if h_{i-1} is chosen, the minimization of this latter model (hopefully) produces a new point $x_{i-1,*}$ such that $h_{i-1}(x_{i-1,*}) < h_{i-1}(x_{i-1,0})$ and a corresponding step $x_{i-1,*} - x_{i-1,0}$ which must then be brought back to level i by the transformation (2.9). Since

$$\|s_i\|_i = \|s_i\|_{M_i} = \|P_i s_{i-1}\|_{M_i} = \|s_{i-1}\|_{P_i^T M_i P_i} \stackrel{\text{def}}{=} \|s_{i-1}\|_{M_{i-1}} = \|s_{i-1}\|_{i-1} \quad (2.16)$$

(which is well-defined since P_i is full-rank), the trust-region constraint (2.13) at level $i - 1$ then becomes

$$\|x_{i-1,*} - x_{i-1,0}\|_{i-1} \leq \Delta_{i,k}. \quad (2.17)$$

This in turn implies that, for $i = 0 \dots, r - 1$,

$$M_i = Q_i^T Q_i \text{ where } Q_i = P_r \dots P_{i+2} P_{i+1}, \quad (2.18)$$

while we define $M_r = I$ for consistency.

Is the cheaper model h_{i-1} always useful? The answer to this question is obviously negative, as it may happen for instance that $g_{i,k}$ lies in the nullspace of R_i (assuming for this example that $n_{i-1} < n_i$) and thus that $R_i g_{i,k}$ is zero while $g_{i,k}$ is not. In this case, the current iterate appears to be first-order critical for h_{i-1} in $\mathbb{R}^{n_{i-1}}$ while it is not for h_i in \mathbb{R}^{n_i} . Using the model h_{i-1} is hence potentially useful only if $\|g_{i-1,0}\| = \|R_i g_{i,k}\|$ is large enough compared to $\|g_{i,k}\|$. The developments below show that the appropriate technical condition turns out to be that

$$\|R_i g_{i,k}\| \geq \kappa_g \|g_{i,k}\| \text{ and } \|R_i g_{i,k}\| > \epsilon_{i-1}^g \quad (2.19)$$

for some constant $\kappa_g \in (0, 1)$ and where $\epsilon_{i-1}^g \in (0, 1)$ is a measure of the first-order criticality that is judged sufficient at level $i - 1$. Note that, given $g_{i,k}$, P_i and R_i , this condition is easy to check before even attempting to compute a step at level $i - 1$.

We are now in position to describe our recursive multilevel trust-region (RMTR) algorithm more formally (see page 7). In this description, we use the constants η_1 , η_2 , γ_1 and γ_2 satisfying the conditions

$$0 < \eta_1 \leq \eta_2 < 1, \text{ and } 0 < \gamma_1 \leq \gamma_2 < 1.$$

It is assumed that the prolongations/restrictions P_i and R_i are known, as the description of the levels $i = 0, \dots, r$. An initial trust-region radius for each level $\Delta_i^g > 0$ is also defined, as well as level-dependent gradient norm tolerances $\epsilon_i^g \in (0, 1)$ and trust-region tolerances $\epsilon_i^\Delta \in (0, 1)$ for $i = 0, \dots, r$. The algorithm’s initial data consists of the level index i ($0 \leq i \leq r$), a starting point $x_{i,0}$, the gradient $g_{i,0}$ at this point, the radius Δ_{i+1} of the level $i + 1$ trust region (by convention, we set $\Delta_{r+1} = \infty$) and the tolerances ϵ_i^g and ϵ_i^Δ .

Algorithm 2.1: RMTR($i, x_{i,0}, g_{i,0}, \Delta_{i+1}, \epsilon_i^g, \epsilon_i^\Delta$)

Step 0: Initialization.

Compute $v_i = g_{i,0} - \nabla_{x_i} f_i(x_{i,0})$ and $h_i(x_{i,0})$. Set $\Delta_{i,0} = \min[\Delta_i^s, \Delta_{i+1}]$ and $k = 0$.

Step 1: Model choice.

If $i = 0$ or if (2.19) fails, go to Step 3. Otherwise, choose to go to Step 2 (recursive step) or to Step 3 (Taylor step).

Step 2: Recursive step computation.

Call Algorithm RMTR($i - 1, R_i x_{i,k}, R_i g_{i,k}, \Delta_{i,k}, \epsilon_{i-1}^g, \epsilon_{i-1}^\Delta$), yielding an approximate solution $x_{i-1,*}$. Then define $s_{i,k} = P_i(x_{i-1,*} - R_i x_{i,k})$, set $\delta_{i,k} = h_{i-1}(R_i x_{i,k}) - h_{i-1}(x_{i-1,*})$ and go to Step 4.

Step 3: Taylor step computation.

Choose $H_{i,k}$ in view of (2.12) and compute a step $s_{i,k} \in \mathbb{R}^{n_i}$ that sufficiently reduces the model $m_{i,k}$ (given by (2.11)) in the sense of (2.15) and such that $\|s_{i,k}\|_i \leq \Delta_{i,k}$. Set $\delta_{i,k} = m_{i,k}(x_{i,k}) - m_{i,k}(x_{i,k} + s_{i,k})$.

Step 4: Acceptance of the trial point

Compute $h_i(x_{i,k} + s_{i,k})$ and define

$$\rho_{i,k} = \frac{h_i(x_{i,k}) - h_i(x_{i,k} + s_{i,k})}{\delta_{i,k}}. \quad (2.20)$$

If $\rho_{i,k} \geq \eta_1$, then define $x_{i,k+1} = x_{i,k} + s_{i,k}$; otherwise define $x_{i,k+1} = x_{i,k}$.

Step 5: Termination.

Compute $g_{i,k+1}$. If $\|g_{i,k+1}\| \leq \epsilon_i^g$ or $\|x_{i,k+1} - x_{i,0}\|_i > (1 - \epsilon_i^\Delta)\Delta_{i+1}$, then return with the approximate solution $x_{i,*} = x_{i,k+1}$.

Step 6: Trust-region radius update.

Set

$$\Delta_{i,k}^+ \in \begin{cases} [\Delta_{i,k}, +\infty) & \text{if } \rho_{i,k} \geq \eta_2, \\ [\gamma_2 \Delta_{i,k}, \Delta_{i,k}] & \text{if } \rho_{i,k} \in [\eta_1, \eta_2), \\ [\gamma_1 \Delta_{i,k}, \gamma_2 \Delta_{i,k}] & \text{if } \rho_{i,k} < \eta_1. \end{cases} \quad (2.21)$$

and

$$\Delta_{i,k+1} = \min \left[\Delta_{i,k}^+, \Delta_{i+1} - \|x_{i,k+1} - x_{i,0}\|_i \right]. \quad (2.22)$$

Increment k by one and go to Step 1.

Some comments on this algorithm are now necessary.

1. The recursive nature of this Algorithm RMTR is clear from the fact that it calls

⁽³⁾Observe that this is the only possible choice for $i = 0$.

itself in Step 2. It is, in that sense, reminiscent of multigrid methods for the solution of linear systems (see Hackbusch, 1994) and is close in spirit to the MG/OPT method by Nash (1999). However, this latter method differs from ours in two main respects: Algorithm RMTR does not make any convexity assumption, nor does it need to modify second derivatives approximations if they turn out not to be positive definite, and the new algorithm does not require the solution of the subproblems to be “sufficiently exact”.

2. The original task of minimizing $f(x) = f_r(x_r) = h_r(x_r)$ (up to the gradient norm tolerance $\epsilon_r^g < \|g_0\|$) is achieved by calling $\text{RMTR}(r, x_0, g_0, \infty, \epsilon_r^g, \epsilon_r^\Delta)$. For coherence of notations, we view this call as being made from some (virtual) iteration 0 at level $r + 1$.
3. In the case where $r = 0$, that is if there is only one level in the problem, the algorithm reduces to the well-known usual trust-region method (see p. 116 of Conn et al., 2000) and enjoys all the desirable properties of this method.
4. The choice

$$\eta_1 = 0.01, \quad \eta_2 = 0.95, \quad \gamma_1 = 0.05 \quad \text{and} \quad \gamma_2 = 0.25 \quad (2.23)$$

is most often appropriate. The choice of Δ_r^s , the initial trust-region radius at level r , is slightly more problem dependent. Although the value 1 often gives reasonable performance, more elaborate strategies exist for choosing it in the context of single level algorithms (see, for instance, Powell, 1970, Sartenaer, 1997, or Sections 10.5.1 and 17.2 in Conn et al., 2000).

5. As with usual trust-region methods, we call iteration k at level i *successful* if $\rho_{i,k} \geq \eta_1$, that is if the trial point $x_{i,k} + s_{i,k}$ is accepted as the next iterate $x_{i,k+1}$. The iteration is said to be *very successful* if $\rho_{i,k} \geq \eta_2$, implying that $\Delta_{i,k}^+ \geq \Delta_{i,k}$.
6. The motivation for (2.22) and the termination test $\|x_{i,k+1} - x_{i,0}\|_i > (1 - \epsilon_i^\Delta)\Delta_{i+1}$ in Step 5 is to guarantee that iterates at a lower level in a recursion remain in the trust region defined at the calling level. We verify this property in Lemma 2.1.
7. When $i > 0$ and (2.19) holds, Step 1 offers the choice between the expensive and cheap step computations. Instead of systematically choosing the cheap one, we leave this option open to keep our class of algorithms as flexible as possible.

Before we go any further in the analysis of Algorithm RMTR, we need to introduce some additional concepts and notation.

1. Iteration k at level i , associated with the computation of the step $s_{i,k}$, will be referred to as iteration (i, k) . It will be called a *Taylor iterations* if Step 3 is used (that is if Taylor’s model $m_{j,\ell}(x_{j,\ell} + s_j)$ is chosen at Step 1). If Step 2 is used instead, iteration (i, k) will then be called a *recursive iteration*.

2. When Step 2 is used at iteration (i, k) , we say that it initiates a *minimization sequence* at level $i - 1$, which consists of all successive iterations *at this level* (starting from the point $x_{i-1,0} = R_i x_{i,k}$) until a return is made to level i within iteration (i, k) . In this case, we also say that iteration (i, k) is the *predecessor* of the minimizing sequence at level $i - 1$. If $(i - 1, \ell)$ belongs to this minimizing sequence, we use the notation

$$(i, k) = \pi(i - 1, \ell)$$

to denote this relation. Observe that ℓ is arbitrary in the minimization sequence.

3. At a given iteration (i, k) , we associate the set

$$\mathcal{R}(i, k) \stackrel{\text{def}}{=} \{(j, \ell) \mid \text{iteration } (j, \ell) \text{ occurs within iteration } (i, k)\}. \quad (2.24)$$

The set $\mathcal{R}(i, k)$ always contains the pair (i, k) and only contains that pair if Step 3 is used at iteration (i, k) . If Step 2 is used instead of Step 3, then it additionally contains the pairs of level and iteration numbers of all iterations that occur in the potential recursion started in Step 2 and terminating on return within iteration (i, k) . Because $\mathcal{R}(i, k)$ is defined in terms of iterations, it does *not* contain the pairs of indices corresponding to the terminating iterates $(j, *)$ of its (internal) minimization sequences. One easily verifies that $j \leq i$ for every j such that $(j, \ell) \in \mathcal{R}(i, k)$ for some non-negative k and ℓ . The mechanism of the algorithm also ensures that

$$\Delta_{j,\ell} \leq \Delta_{i,k} \text{ whenever } (j, \ell) \in \mathcal{R}(i, k), \quad (2.25)$$

because of the choice of $\Delta_{j,0}$ in Step 0 and (2.22).

Note that $\mathcal{R}(i, k)$ always contains a single minimization sequence at level $i - 1$, but may contain more than one such sequence at level $i - 2$, since each iteration at level $i - 1$ may generate its own.

4. For any iteration $(j, \ell) \in \mathcal{R}(i, k)$, there exists a unique *path* from (j, ℓ) to (i, k) defined by taking the predecessor of iteration (j, ℓ) , say $(j + 1, q) = \pi(j, \ell)$, and then the predecessor of $(j + 1, q)$ and so on until iteration (i, k) .

We also define

$$d(i, k) = \min_{(j,\ell) \in \mathcal{R}(i,k)} j, \quad (2.26)$$

which is the index of the deepest level reached by the potential recursion of iteration (i, k) . The path from $(d(i, k), \ell)$ to (i, k) is the longest in all paths defined in $\mathcal{R}(i, k)$.

5. We use the notation

$$\mathcal{T}(i, k) \stackrel{\text{def}}{=} \{(j, \ell) \in \mathcal{R}(i, k) \mid \text{iteration } (j, \ell) \text{ uses Step 3}\},$$

to denote the subset of *Taylor's iterations* in $\mathcal{R}(i, k)$, that is iterations at which Taylor's model $m_{j,\ell}(x_{j,\ell} + s_j)$ is chosen.

We conclude this presentation of Algorithm RMTR by proving that it has a central property of trust-region methods, namely that the steps remain in the trust region.

Lemma 2.1 *The mechanism of Algorithm RMTR guarantees that, for each iteration (i, k) ,*

$$\|s_{i,k}\|_i \leq \Delta_{i,k}. \quad (2.27)$$

Moreover, if $\Delta_{j+1,q}$ is the trust-region radius of iteration $(j+1, q) = \pi(j, \ell)$, we have that, for each $(j, \ell) \in \mathcal{R}(i, k)$,

$$\|x_{j,\ell} - x_{j,0}\|_j \leq \Delta_{j+1,q} \quad \text{and} \quad \|x_{j,*} - x_{j,0}\|_j \leq \Delta_{j+1,q}. \quad (2.28)$$

Proof. The constraint (2.27) is explicit for Taylor iterations. We therefore only have to verify that it holds if Step 2 is chosen at iteration (i, k) . If this is the case, consider $j = d(i, k)$, and consider the first time it occurs in $\mathcal{R}(i, k)$. Assume furthermore that $x_{j,*} = x_{j,p}$. Because no recursion occurs to a level lower than j , one must have (from Step 3) that

$$\|s_{j,\ell}\|_j \leq \Delta_{j,\ell} \quad (\ell = 0, \dots, p-1). \quad (2.29)$$

Then we obtain, for $\ell = 1, \dots, p$, that, if iteration $(j, \ell-1)$ is successful,

$$\|x_{j,\ell} - x_{j,0}\|_j = \|x_{j,\ell-1} - x_{j,0} + s_{j,\ell-1}\|_j \leq \|x_{j,\ell-1} - x_{j,0}\|_j + \|s_{j,\ell-1}\|_j,$$

because of the triangular inequality, while

$$\|x_{j,\ell} - x_{j,0}\|_j = \|x_{j,\ell-1} - x_{j,0}\|_j \leq \|x_{j,\ell-1} - x_{j,0}\|_j + \|s_{j,\ell-1}\|_j,$$

if it is unsuccessful. Combining these two bounds and (2.29), we have that

$$\begin{aligned} \|x_{j,\ell} - x_{j,0}\|_j &\leq \|x_{j,\ell-1} - x_{j,0}\|_j + \Delta_{j,\ell-1} \\ &\leq \|x_{j,\ell-1} - x_{j,0}\|_j + \Delta_{j+1,q} - \|x_{j,\ell-1} - x_{j,0}\|_j \\ &= \Delta_{j+1,q} \end{aligned} \quad (2.30)$$

for $\ell = 1, \dots, p$, where the last inequality results from (2.22). We then verify, using (2.16), that

$$\|s_{j+1,q}\|_{j+1} = \|P_{j+1}(x_{j,*} - x_{j,0})\|_{j+1} = \|x_{j,p} - x_{j,0}\|_j = \|x_{j,*} - x_{j,0}\|_j \leq \Delta_{j+1,q},$$

which is nothing but the inequality of (2.29) at iteration $(j+1, q)$. The same reasoning may then be applied to each iteration at level $j+1$ that uses Step 2. Since the inequality in (2.29) is guaranteed for all other iterations of that level by Step 3, we obtain that (2.29) also holds with j replaced by $j+1$. The same must therefore be true for (2.30). The induction can then be continued up to level i , yielding both (2.27) and (2.28) (for which the case $\ell = 0$ is obvious). \square

In the same vein, the algorithm also ensures the following two properties.

Lemma 2.2 *The mechanism of Algorithm RMTR guarantees that, for each iterate of index (j, ℓ) such that $(j, \ell) \neq (j, *)$ (i.e., for all iterates at level j but the last one),*

$$\|g_{j,\ell}\| > \epsilon_j^g \quad (2.31)$$

and

$$\|x_{j,\ell} - x_{j,0}\|_j \leq (1 - \epsilon_j^\Delta) \Delta_{j+1,q}, \quad (2.32)$$

where $\Delta_{j+1,q}$ is the trust-region radius of iteration $(j+1, q) = \pi(j, \ell)$.

Proof. These bounds directly follow from the stopping criteria for minimization at level j , in Step 5 of the algorithm. \square

3 Global convergence

We now investigate the global convergence properties of our recursive multilevel algorithm. Our exposition starts with the analysis of properties that are specific to Algorithm RMTR, and subsequently revisits the main concepts and developments of Section 6.4 in Conn et al. (2000) to conclude in the specific case of the multilevel algorithm. Interestingly, the techniques of proof are different and lead to a new complexity result that is also valid in the classical single-level case.

We complete our assumption by supposing that the Hessian of each h_i is bounded above by the constant κ_H (possibly increasing this constant), i.e., more formally, that

$$1 + \|\nabla_{x_i} h_i(x_i)\| \leq \kappa_H \quad (3.1)$$

for all $x_i \in \mathbb{R}^{n_i}$ and all i . We also define the constants

$$\kappa_{\text{PR}} \stackrel{\text{def}}{=} \max \left[1, \max_{i=1,\dots,r} \|P_i\| \right] = \max \left[1, \max_{i=1,\dots,r} \|R_i\| \right] \quad (3.2)$$

(where we used (2.4) to deduce the second equality), and

$$\kappa_\sigma \stackrel{\text{def}}{=} \min \left[1, \min_{i=0,\dots,r} \sigma_{\min}(M_i) \right] > 0, \quad (3.3)$$

where $\sigma_{\min}(A)$ denotes the smallest singular value of the matrix A . We finally define the constants

$$\Delta_{\min}^s = \min_{i=0,\dots,r} \Delta_i^s, \quad \epsilon_{\min}^g = \min_{i=0,\dots,r} \epsilon_i^g \quad \text{and} \quad \epsilon_{\min}^\Delta = \min_{i=0,\dots,r} \epsilon_i^\Delta. \quad (3.4)$$

We start by proving some useful bounds on the gradient norms for all iterates that belong to a recursion process initiated within a sufficiently small trust region.

Lemma 3.1 *Assume that $r > 0$, and that, for some iteration (i, k) ,*

$$\Delta_{i,k} \leq \frac{\sqrt{\kappa_\sigma} \kappa_g^r}{2r \kappa_H} \|g_{i,k}\| \stackrel{\text{def}}{=} \kappa_1 \|g_{i,k}\|, \quad (3.5)$$

where $\kappa_1 \in (0, 1)$. Then one has that

$$\frac{1}{2}\kappa_g^r \|g_{i,k}\| \leq \|g_{j,\ell}\| \leq \kappa_{PR}^r (1 + \frac{1}{2}\kappa_g^r) \|g_{i,k}\| \quad (3.6)$$

for all $(j, \ell) \in \mathcal{R}(i, k)$.

Proof. The result is obvious for $(j, \ell) = (i, k)$ since, by definition, $\kappa_g < 1$ and $\kappa_{PR} \geq 1$. Let us now consider some iteration $(j, \ell) \in \mathcal{R}(i, k)$ with $j < i$. From the mean-value theorem, we know that, for any iteration (j, ℓ) ,

$$g_{j,\ell} = g_{j,0} + G_{j,\ell}(x_{j,\ell} - x_{j,0}) \quad (3.7)$$

where

$$G_{j,\ell} = \int_0^1 \nabla_{x_j x_j} h_j(x_{j,0} + t(x_{j,\ell} - x_{j,0})) dt. \quad (3.8)$$

But

$$\|G_{j,\ell}\| \leq \max_{t \in [0,1]} \|\nabla_{x_j x_j} h_j(x_{j,0} + t(x_{j,\ell} - x_{j,0}))\| \leq \kappa_H, \quad (3.9)$$

and hence, by definition of the norms and (3.3),

$$\|g_{j,\ell}\| \geq \|g_{j,0}\| - \kappa_H \|x_{j,\ell} - x_{j,0}\| \geq \|g_{j,0}\| - \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \|x_{j,\ell} - x_{j,0}\|_j \quad (3.10)$$

for all (j, ℓ) . On the other hand, if $(j+1, q) = \pi(j, \ell)$, we have also that, for all $(j, \ell) \in \mathcal{R}(i, k)$,

$$\|x_{j,\ell} - x_{j,0}\|_j \leq \Delta_{j+1,q} \leq \Delta_{i,k} \quad (3.11)$$

because of (2.28) and (2.25) (as $(j+1, q) \in \mathcal{R}(i, k)$). Combining (3.10) and (3.11), we obtain that, for all $(j, \ell) \in \mathcal{R}(i, k)$,

$$\|g_{j,\ell}\| \geq \|g_{j,0}\| - \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k}. \quad (3.12)$$

Consider now the path from (j, ℓ) to (i, k) in $\mathcal{R}(i, k)$. Let this path consists of the iterations $(j, \ell), (j+u, t_{j+u})$ for $u = 1, \dots, i-j-1$ and (i, k) . We then have that

$$\begin{aligned} \|g_{j,\ell}\| &\geq \|g_{j,0}\| - \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\geq \kappa_g \|g_{j+1,t_{j+1}}\| - \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\geq \kappa_g \|g_{j+1,0}\| - 2 \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\geq \kappa_g^2 \|g_{j+2,t_{j+2}}\| - 2 \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\geq \kappa_g^r \|g_{i,k}\| - r \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k}, \end{aligned}$$

where we successively used (3.12), the first part of (2.19) and the inequality $\kappa_g < 1$. We then deduce the first inequality of (3.6) from (3.5).

To prove the second, we re-use (3.7)–(3.9) to obtain that

$$\|g_{j,\ell}\| \leq \|g_{j,0}\| + \kappa_H \|x_{j,\ell} - x_{j,0}\| \leq \|g_{j,0}\| + \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \|x_{j,\ell} - x_{j,0}\|_j. \quad (3.13)$$

Combining this with (3.11), we conclude that

$$\|g_{j,\ell}\| \leq \|g_{j,0}\| + \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k}. \quad (3.14)$$

We now retrack the iteration path from (j, ℓ) back to (i, k) as above, and successively deduce from (3.14), (2.8) and (3.2) that

$$\begin{aligned} \|g_{j,\ell}\| &\leq \|g_{j,0}\| + \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\leq \kappa_{\text{PR}} \|g_{j+1,\ell_{j+1}}\| + \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\leq \kappa_{\text{PR}} \|g_{j+1,0}\| + (\kappa_{\text{PR}} + 1) \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\leq \kappa_{\text{PR}}^2 \|g_{j+2,\ell_{j+2}}\| + 2 \frac{\kappa_{\text{PR}} \kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\leq \kappa_{\text{PR}}^r \|g_{i,k}\| + r \frac{\kappa_{\text{PR}}^{r-1} \kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \\ &\leq \kappa_{\text{PR}}^r \left[\|g_{i,k}\| + r \frac{\kappa_H}{\sqrt{\kappa_\sigma}} \Delta_{i,k} \right], \end{aligned}$$

using $\kappa_{\text{PR}} \geq 1$. We may now use the bound (3.5) to conclude that the second inequality of (3.6) must hold. \square

We now investigate what happens at non-critical points if the trust-region radius $\Delta_{i,k}$ is small enough. This investigation is conducted by considering the subset $\mathcal{V}(i, k)$ of $\mathcal{R}(i, k)$ defined by

$$\mathcal{V}(i, k) = \left\{ (j, \ell) \in \mathcal{R}(i, k) \mid \delta_{j,\ell} \geq \frac{1}{2} \kappa_{\text{red}} \kappa_g^r \kappa_\epsilon^{j-d(i,k)} \|g_{i,k}\| \Delta_{j,\ell} \right\}, \quad (3.15)$$

where

$$\kappa_\epsilon \stackrel{\text{def}}{=} \eta_2 \epsilon_{\min}^\Delta < 1. \quad (3.16)$$

$\mathcal{V}(i, k)$ is the subset of iterations within the recursion at iteration (i, k) for which the model decrease is bounded below by a (level-dependent) factor times the product of the gradient norm $\|g_{i,k}\|$ and the trust-region radius $\Delta_{j,\ell}$. Note that, if iteration (j, ℓ) belongs to $\mathcal{V}(i, k)$, this implies that $\delta_{j,\ell}$ can be computed in a finite number of iterations, and thus that $\mathcal{R}(j, \ell)$ is finite. The idea of the next two results is to show that $\mathcal{V}(i, k)$ and $\mathcal{R}(i, k)$ coincide for a sufficiently small radius $\Delta_{i,k}$.

Theorem 3.2 *Consider an iteration (i, k) for which $\|g_{i,k}\| > 0$ and*

$$\begin{aligned} \Delta_{i,k} &\leq \min \left[\Delta_{\min}^s, \min \left(\kappa_1, \frac{\kappa_{\text{red}} \kappa_\sigma \kappa_g^r \kappa_\epsilon^r (1 - \eta_2)}{2 \kappa_H} \right) \|g_{i,k}\| \right] \\ &\stackrel{\text{def}}{=} \min[\Delta_{\min}^s, \kappa_2 \|g_{i,k}\|], \end{aligned} \quad (3.17)$$

where $\kappa_2 \in (0, 1)$. Then the following conclusions hold:

1. every iteration using Taylor's model belongs to (3.15), that is

$$\mathcal{T}(i, k) \subseteq \mathcal{V}(i, k), \quad (3.18)$$

2. iteration (j, ℓ) is very successful for every $(j, \ell) \in \mathcal{V}(i, k)$.

Moreover, if all iterations (j, ℓ) of a minimization sequence at level $j < i$ belong to $\mathcal{V}(i, k)$ and if $\pi(j, \ell) = (j + 1, q)$, then

3. the decrease in the objective function at level j satisfies

$$h_j(x_{j,0}) - h_j(x_{j,\ell}) \geq \frac{1}{2} \kappa_{\text{red}} \kappa_g^r \kappa_\epsilon^{j-d(i,k)+1} \ell \|g_{i,k}\| \Delta_{j+1,q} \quad (3.19)$$

for each $\ell > 0$,

4. there are at most

$$p_* \stackrel{\text{def}}{=} \left\lceil \frac{\kappa_{PR}^r \sqrt{\kappa_\sigma} (2 + \kappa_g^r) + \kappa_2 \kappa_H}{\kappa_{\text{red}} \kappa_\sigma \kappa_g^r \kappa_\epsilon^r} \right\rceil \quad (3.20)$$

iterations in the minimization sequence at level j ,

5. we have that

$$(j + 1, q) \in \mathcal{V}(i, k). \quad (3.21)$$

Proof. [1.] We start by proving (3.18), and note that, for $(j, \ell) \in \mathcal{R}(i, k)$, (2.25), the bound $\kappa_\sigma \leq 1$, (3.17), the left inequality in (3.6) and (2.12) allow us to conclude that

$$\Delta_{j,\ell} \leq \Delta_{i,k} \leq \frac{\kappa_g^r}{2\kappa_H} \|g_{i,k}\| \leq \frac{\|g_{j,\ell}\|}{1 + \|H_{j,\ell}\|}. \quad (3.22)$$

If we now assume that $(j, \ell) \in \mathcal{T}(i, k)$, the sufficient decrease condition (2.15) must hold at this iteration, which, together with the left part of (3.6) and (3.22), gives that

$$\delta_{j,\ell} = m_{j,\ell}(x_{j,\ell}) - m_{j,\ell}(x_{j,\ell} + s_{j,\ell}) \geq \kappa_{\text{red}} \|g_{j,\ell}\| \Delta_{j,\ell} \geq \frac{1}{2} \kappa_{\text{red}} \kappa_g^r \|g_{i,k}\| \Delta_{j,\ell}, \quad (3.23)$$

which then implies (3.18) since $\kappa_\epsilon < 1$.

[2.] We prove item 2 separately for $(j, \ell) \in \mathcal{T}(i, k)$ and for $(j, \ell) \in \mathcal{V}(i, k) \setminus \mathcal{T}(i, k)$. Consider the case where $(j, \ell) \in \mathcal{T}(i, k)$ first. We deduce from Taylor's theorem that, for $(j, \ell) \in \mathcal{T}(i, k)$,

$$|h_j(x_{j,\ell} + s_{j,\ell}) - m_{j,\ell}(x_{j,\ell} + s_{j,\ell})| \leq \kappa_H \left(\frac{\|s_{j,\ell}\|}{\|s_{j,\ell}\|_j} \right)^2 \Delta_{j,\ell}^2, \quad (3.24)$$

(see, for instance, Theorem 6.4.1 on p. 133 of Conn et al., 2000). But, by definition of the norms and (3.3), we know that $\|s_{j,\ell}\|_j \geq \sqrt{\kappa_\sigma} \|s_{j,\ell}\|$. Hence, (3.24) becomes

$$|h_j(x_{j,\ell} + s_{j,\ell}) - m_{j,\ell}(x_{j,\ell} + s_{j,\ell})| \leq \frac{\kappa_H}{\kappa_\sigma} \Delta_{j,\ell}^2.$$

Combining this last bound with (3.23), we obtain from (2.20) that

$$|\rho_{j,\ell} - 1| \leq \left| \frac{h_j(x_{j,\ell} + s_{j,\ell}) - m_{j,\ell}(x_{j,\ell} + s_{j,\ell})}{m_{j,\ell}(x_{j,\ell}) - m_{j,\ell}(x_{j,\ell} + s_{j,\ell})} \right| \leq \frac{2\kappa_H}{\kappa_{\text{red}} \kappa_\sigma \kappa_g^r \|g_{i,k}\|} \Delta_{j,\ell} \leq 1 - \eta_2,$$

where the last inequality is deduced from (2.25) and the fact that (3.17) implies the bound

$$\Delta_{i,k} \leq \frac{\kappa_{\text{red}} \kappa_\sigma \kappa_g^r \|g_{i,k}\| (1 - \eta_2)}{2\kappa_H}$$

since $\kappa_\epsilon < 1$. Hence

$$\rho_{j,\ell} \geq \eta_2 \quad (3.25)$$

and iteration $(j, \ell) \in \mathcal{T}(i, k)$ is thus very successful, as requested in item 2 of the theorem's statement.

We next prove item 2 for $(j, \ell) \in \mathcal{V}(i, k) \setminus \mathcal{T}(i, k)$, which implies, in particular, that $\mathcal{R}(j, \ell)$ is finite and $x_{j-1,*}$ well-defined. If we consider iteration (j, ℓ) , we may still deduce from the mean-value theorem that

$$h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell}) = -\langle g_{j,\ell}, s_{j,\ell} \rangle - \frac{1}{2} \langle s_{j,\ell}, \nabla_{x_j x_j} h_j(\xi_j) s_{j,\ell} \rangle$$

for some $\xi_j \in [x_{j,\ell}, x_{j,\ell} + s_{j,\ell}]$, and also that

$$h_{j-1}(x_{j-1,0}) - h_{j-1}(x_{j-1,*}) = -\langle g_{j-1,0}, z_{j-1} \rangle - \frac{1}{2} \langle z_{j-1}, \nabla_{x_{j-1} x_{j-1}} h_{j-1}(\xi_{j-1}) z_{j-1} \rangle$$

for some $\xi_{j-1} \in [x_{j-1,0}, x_{j-1,0} + z_{j-1}]$, where

$$z_{j-1} = x_{j-1,*} - x_{j-1,0} = x_{j-1,*} - R_j x_{j,\ell}.$$

Now, because $s_{j,\ell} = P_j z_{j-1}$, we deduce from (2.10) that

$$\langle g_{j,\ell}, s_{j,\ell} \rangle = \langle g_{j-1,0}, z_{j-1} \rangle,$$

and therefore that

$$\begin{aligned} h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell}) &= h_{j-1}(x_{j-1,0}) - h_{j-1}(x_{j-1,*}) \\ &\quad - \frac{1}{2} \langle s_{j,\ell}, \nabla_{x_j x_j} h_j(\xi_j) s_{j,\ell} \rangle \\ &\quad + \frac{1}{2} \langle z_{j-1}, \nabla_{x_{j-1} x_{j-1}} h_{j-1}(\xi_{j-1}) z_{j-1} \rangle. \end{aligned} \quad (3.26)$$

But Lemma 2.1 implies that

$$\|s_{j,\ell}\|_j \leq \Delta_{j,\ell} \quad \text{and} \quad \|z_{j-1}\|_{j-1} \leq \Delta_{j,\ell},$$

which in turn, with the Cauchy-Schwarz inequality, gives that

$$|\langle s_{j,\ell}, \nabla_{x_j x_j} h_j(\xi_j) s_{j,\ell} \rangle| \leq \kappa_{\mathbf{H}} \left(\frac{\|s_{j,\ell}\|}{\|s_{j,\ell}\|_j} \right)^2 \Delta_{j,\ell}^2 \leq \frac{\kappa_{\mathbf{H}}}{\kappa_\sigma} \Delta_{j,\ell}^2. \quad (3.27)$$

Similarly,

$$|\langle z_{j-1}, \nabla_{x_{j-1} x_{j-1}} h_{j-1}(\xi_{j-1}) z_{j-1} \rangle| \leq \frac{\kappa_{\mathbf{H}}}{\kappa_\sigma} \Delta_{j,\ell}^2. \quad (3.28)$$

Combining (3.26), (3.27), (3.28) and the definition of $\delta_{j,\ell}$, we obtain that

$$h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell}) \geq \delta_{j,\ell} - \frac{\kappa_{\mathbf{H}}}{\kappa_\sigma} \Delta_{j,\ell}^2. \quad (3.29)$$

But since $(j, \ell) \in \mathcal{V}(i, k)$ and $\kappa_\epsilon < 1$, we have that

$$\delta_{j,\ell} \geq \kappa_{\text{red}} \kappa_{\mathbf{g}}^r \kappa_\epsilon^{j-d(i,k)} \|g_{i,k}\| \Delta_{j,\ell} \geq \kappa_{\text{red}} \kappa_{\mathbf{g}}^r \kappa_\epsilon^r \|g_{i,k}\| \Delta_{j,\ell} > 0$$

and we conclude from (3.29), the definition of $\rho_{j,\ell}$ and this last bound that

$$\rho_{j,\ell} = \frac{h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell})}{\delta_{j,\ell}} \geq 1 - \frac{\kappa_{\text{H}} \Delta_{j,\ell}^2}{\kappa_{\sigma} \delta_{j,\ell}} \geq 1 - \frac{\kappa_{\text{H}} \Delta_{j,\ell}}{\kappa_{\text{red}} \kappa_{\sigma} \kappa_{\text{g}}^r \kappa_{\epsilon}^r \|g_{i,k}\|}.$$

Noting now that (3.17) implies the inequality

$$\Delta_{i,k} \leq \frac{\kappa_{\text{red}} \kappa_{\sigma} \kappa_{\text{g}}^r \kappa_{\epsilon}^r \|g_{i,k}\| (1 - \eta_2)}{\kappa_{\text{H}}}$$

and using the bound (2.25), we obtain that $\rho_{j,\ell} \geq \eta_2$. Iteration (j, ℓ) is thus very successful, which completes the proof of item 2.

[3.] We now assume that all iterations (j, ℓ) of a minimization sequence at level $j < i$ belong to $\mathcal{V}(i, k)$ with $(j + 1, q) = \pi(j, \ell)$. We first notice that $(j + 1, q) \in \mathcal{R}(i, k)$, (2.25), (3.17) and (3.4) imply that

$$\Delta_{j+1,q} \leq \Delta_{i,k} \leq \Delta_{\min}^s \leq \Delta_j^s.$$

Hence Step 0 gives that $\Delta_{j,0} = \Delta_{j+1,q}$ and since all iterations at level j are very successful because of item 2, we have from Step 6 that, for all (j, ℓ) with $\ell > 0$,

$$\begin{aligned} \Delta_{j,\ell} &= \min \left[\Delta_{j,\ell-1}^+, \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j \right] \\ &\geq \min \left[\Delta_{j,\ell-1}, \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j \right] \\ &= \min \left[\min[\Delta_{j,\ell-2}^+, \Delta_{j+1,q} - \|x_{j,\ell-1} - x_{j,0}\|_j], \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j \right] \\ &\geq \min \left[\Delta_{j,\ell-2}, \Delta_{j+1,q} - \max_{p=\ell-1,\ell} \|x_{j,p} - x_{j,0}\|_j \right] \\ &\geq \min \left[\Delta_{j,0}, \Delta_{j+1,q} - \max_{p=1,\dots,\ell} \|x_{j,p} - x_{j,0}\|_j \right] \\ &= \Delta_{j+1,q} - \max_{p=1,\dots,\ell} \|x_{j,p} - x_{j,0}\|_j \\ &\geq \epsilon_j^{\Delta} \Delta_{j+1,q}, \end{aligned}$$

where we used (2.32) to deduce the last inequality. Note that $\Delta_{j,0} = \Delta_{j+1,q} > \epsilon_j^{\Delta} \Delta_{j+1,q}$, covering the case where $\ell = 0$. Combining these bounds with the very successful nature of each iteration at level j , we obtain that, for each (j, p) with $p = 0, \dots, \ell - 1$,

$$\begin{aligned} h_j(x_{j,p}) - h_j(x_{j,p} + s_{j,p}) &\geq \eta_2 \delta_{j,p} \\ &\geq \frac{1}{2} \eta_2 \kappa_{\text{red}} \kappa_{\text{g}}^r \kappa_{\epsilon}^{j-d(i,k)} \|g_{i,k}\| \Delta_{j,p} \\ &\geq \frac{1}{2} \kappa_{\text{red}} \kappa_{\text{g}}^r \kappa_{\epsilon}^{j-d(i,k)} \eta_2 \epsilon_j^{\Delta} \|g_{i,k}\| \Delta_{j+1,q} \\ &\geq \frac{1}{2} \kappa_{\text{red}} \kappa_{\text{g}}^r \kappa_{\epsilon}^{j-d(i,k)+1} \|g_{i,k}\| \Delta_{j+1,q}. \end{aligned}$$

Summing now over iterations $p = 0, \dots, \ell - 1$ at level j , we obtain that

$$\begin{aligned} h_j(x_{j,0}) - h_j(x_{j,\ell}) &= \sum_{p=0}^{\ell-1} [h_j(x_{j,p}) - h_j(x_{j,p} + s_{j,p})] \\ &\geq \frac{1}{2} \kappa_{\text{red}} \kappa_{\text{g}}^r \kappa_{\epsilon}^{j-d(i,k)+1} \ell \|g_{i,k}\| \Delta_{j+1,q}, \end{aligned}$$

yielding (3.19).

[4.] In order to prove item 4, we start by proving that the total decrease in h_j (the objective function for the considered minimization sequence at the j -th level) is bounded above by some multiple of $\|g_{i,k}\|$ and $\Delta_{j+1,q}$. We first note that the mean-value theorem gives that

$$h_j(x_{j,0} + s_{j,\min}) = h_j(x_{j,0}) + \langle g_{j,0}, s_{j,\min} \rangle + \frac{1}{2} \langle s_{j,\min}, \nabla_{x_j x_j} h_j(\xi_j) s_{j,\min} \rangle$$

for some $\xi_j \in [x_{j,0}, x_{j,0} + s_{j,\min}]$, where we have defined

$$s_{j,\min} = \arg \min_{\|s_j\|_j \leq \Delta_{j+1,q}} h_j(x_{j,0} + s_j).$$

Hence, we obtain that, for all s_j such that $\|s_j\|_j \leq \Delta_{j+1,q}$,

$$h_j(x_{j,0}) - h_j(x_{j,0} + s_j) \leq h_j(x_{j,0}) - h_j(x_{j,0} + s_{j,\min}) \leq \frac{\|g_{j,0}\|}{\sqrt{\kappa_\sigma}} \Delta_{j+1,q} + \frac{\kappa_H}{2\kappa_\sigma} \Delta_{j+1,q}^2.$$

But we have that $\|x_{j,\ell} - x_{j,0}\|_j \leq \Delta_{j+1,q}$ because of (2.28) and therefore the right inequality of (3.6), (2.25) and (3.17) now give that

$$h_j(x_{j,0}) - h_j(x_{j,\ell}) \leq \left[\frac{\kappa_{\text{PR}}^r + \frac{1}{2} \kappa_{\text{PR}}^r \kappa_g^r}{\sqrt{\kappa_\sigma}} + \frac{\kappa_2 \kappa_H}{2\kappa_\sigma} \right] \|g_{i,k}\| \Delta_{j+1,q} \quad (3.30)$$

for all (j, ℓ) with $\ell \geq 0$. Combining now this bound with (3.19) and remembering that $\kappa_\epsilon < 1$, we deduce that item 4 must hold with (3.20).

[5.] Finally, since the minimization sequence at level j is guaranteed to terminate after a finite number of iterations $1 \leq \ell \leq p_*$, we deduce from (3.19) and the definition of $\delta_{j+1,q}$ that

$$\delta_{j+1,q} \geq \frac{1}{2} \kappa_{\text{red}} \kappa_g^r \kappa_\epsilon^{j+1-d(i,k)} \|g_{i,k}\| \Delta_{j+1,q},$$

and (3.21) then immediately follows. \square

We may deduce the following important corollary from this theorem.

Corollary 3.3 *Assume (3.17) holds for some iteration (i, k) for which $\|g_{i,k}\| > 0$. Then all iterations $(j, \ell) \in \mathcal{R}(i, k)$ are very successful. Moreover, the total number of iterations in $\mathcal{R}(i, k)$ is finite and $\Delta_{i,k}^+ \geq \Delta_{i,k}$.*

Proof. As suggested above, we proceed by showing that $\mathcal{V}(i, k) = \mathcal{R}(i, k)$, working from the deepest recursion level upwards. Thus consider level $j = d(i, k)$ first. At this level, all iterations (j, ℓ) belong to $\mathcal{T}(i, k)$ and thus, by (3.18), to $\mathcal{V}(i, k)$. If $j = i$, we have achieved our objective. Assume therefore that $j < i$ and consider level $j + 1$. Using (3.21), we see that all iterations involving a recursion to level j must belong to $\mathcal{V}(i, k)$, and are thus very successful because of item 2 of Theorem 3.2. Again, if $j + 1 = i$, we have proved that $\mathcal{V}(i, k) = \mathcal{R}(i, k)$ and all iterations in this set are very successful. If $j + 1 < i$, we may then apply the

same reasoning to level $j + 2$, and so on until level i is reached. We may thus conclude that $\mathcal{V}(i, k)$ and $\mathcal{R}(i, k)$ always coincide and only contain very successful iterations. Furthermore, using item 4 of Theorem 3.2, we see that the total number of iterations in $\mathcal{R}(i, k)$ is bounded above by

$$\sum_{l=0}^r p_*^l \leq r p_*^r + 1.$$

Finally, the fact that $\Delta_{i,k}^+ \geq \Delta_{i,k}$ then results from the mechanism of Step 6 of the algorithm and the very successful nature of iteration $(i, k) \in \mathcal{R}(i, k)$. \square

This last result guarantees the finiteness of the recursion at iteration (i, k) (and thus finiteness of the computation of $s_{i,k}$) if $\Delta_{i,k}$ is small enough. It also ensures the following useful consequence.

Lemma 3.4 *Each minimization sequence contains at least one successful iteration.*

Proof. This follows from the fact that unsuccessful iterations cause the trust-region radius to decrease, until (3.17) is eventually satisfied and a (very) successful iteration occurs. \square

We now investigate the consequence of the above results on the trust-region radius at each minimization level.

Lemma 3.5 *For every iteration (j, ℓ) , with $j < r$, we have that*

$$\Delta_{j,\ell} \geq \gamma_1 \min [\Delta_{\min}^s, \kappa_2 \epsilon_j^g, \epsilon_j^\Delta \Delta_{j+1,q}], \quad (3.31)$$

where $(j + 1, q) = \pi(j, \ell)$.

Proof. Consider the minimization sequence at level $j < r$ initiated from iteration $(j + 1, q)$, and assume, for the purpose of obtaining a contradiction, that iteration (j, ℓ) is the first such that

$$\Delta_{j,\ell} < \gamma_1 \min [\Delta_{\min}^s, \kappa_2 \epsilon_j^g, \epsilon_j^\Delta \Delta_{j+1,q}]. \quad (3.32)$$

Note that, because $\epsilon_j^\Delta < 1$ and $\gamma_1 < 1$,

$$\Delta_{j,0} = \min[\Delta_j^s, \Delta_{j+1,q}] \geq \min[\Delta_{\min}^s, \epsilon_j^\Delta \Delta_{j+1,q}] > \gamma_1 \min [\Delta_{\min}^s, \kappa_2 \epsilon_j^g, \epsilon_j^\Delta \Delta_{j+1,q}],$$

which ensures that $\ell > 0$ and hence that $\Delta_{j,\ell}$ is computed by applying Step 6 of the algorithm at iteration $(j, \ell - 1)$. Suppose now that

$$\Delta_{j,\ell} = \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j, \quad (3.33)$$

i.e., the second term is active in (2.22). Then, using (2.32), the definition of γ_1 and (3.32), we deduce that

$$\Delta_{j,\ell} \geq \Delta_{j+1,q} - (1 - \epsilon_j^\Delta) \Delta_{j+1,q} = \epsilon_j^\Delta \Delta_{j+1,q} > \gamma_1 \epsilon_j^\Delta \Delta_{j+1,q} > \Delta_{j,\ell},$$

which is impossible. Hence (3.33) cannot hold and we obtain that

$$\Delta_{j,\ell} = \Delta_{j,\ell-1}^+ \geq \gamma_1 \Delta_{j,\ell-1},$$

where the last inequality results from (2.21) and (2.22). Combining this bound with (3.32) and (2.31), we deduce that

$$\Delta_{j,\ell-1} \leq \min [\Delta_{\min}^s, \kappa_2 \epsilon_j^g, \epsilon_j^\Delta \Delta_{j+1,q}] \leq \min [\Delta_{\min}^s, \kappa_2 \|g_{j,\ell-1}\|].$$

Hence we may apply Corollary 3.3 and conclude that iteration $(j, \ell - 1)$ is very successful and that

$$\Delta_{j,\ell-1} \leq \Delta_{j,\ell-1}^+ = \Delta_{j,\ell}.$$

As a consequence, iteration (j, ℓ) cannot be the first such that (3.32) holds. This contradiction now implies that (3.32) is impossible, which completes the proof. \square

Thus trust-region radii are bounded away from zero by a level-dependent factor. We now verify that this factor may be made independent of the level.

Theorem 3.6 *There exists a constant $\Delta_{\min} \in (0, \min[\Delta_{\min}^s, 1])$ such that*

$$\Delta_{j,\ell} \geq \Delta_{\min} \tag{3.34}$$

for every iteration (j, ℓ) .

Proof. Observe first that Lemma 3.5 ensures the bound

$$\Delta_{r,k} \geq \gamma_1 \min[\Delta_{\min}^s, \kappa_2 \epsilon_r^g] \stackrel{\text{def}}{=} \gamma_1 \mu \tag{3.35}$$

for all $k \geq 0$, because we have assumed that the call to the uppermost level is made with an infinite trust-region radius. Note that

$$\mu \in (0, 1) \tag{3.36}$$

because of (3.4) and (3.17). Suppose now that, for some iteration (j, ℓ) ,

$$\Delta_{j,\ell} < \gamma_1^{r+2} (\epsilon_{\min}^\Delta)^r \mu. \tag{3.37}$$

If $j = r$, this contradicts (3.35). Hence $0 \leq j < r$. Lemma 3.5 and the definition of μ in (3.35) then imply that

$$\min[\mu, \epsilon_j^\Delta \Delta_{j+1,q}] < \gamma_1^{r+1} (\epsilon_{\min}^\Delta)^r \mu, \tag{3.38}$$

where, as above, iteration $(j + 1, q) = \pi(j, \ell)$. If $\min[\mu, \epsilon_j^\Delta \Delta_{j+1,q}] = \mu$, then $\mu < \gamma_1^{r+1} (\epsilon_{\min}^\Delta)^r \mu$, which is impossible because $\gamma_1^{r+1} (\epsilon_{\min}^\Delta)^r < 1$. As a consequence,

$$\epsilon_j^\Delta \Delta_{j+1,q} = \min[\mu, \epsilon_j^\Delta \Delta_{j+1,q}] < \gamma_1^{r+1} (\epsilon_{\min}^\Delta)^r \mu \leq \gamma_1^{r+1} (\epsilon_{\min}^\Delta)^{r-1} \epsilon_j^\Delta \mu,$$

because of (3.4), and hence

$$\Delta_{j+1,q} < \gamma_1^{r+1} (\epsilon_{\min}^\Delta)^{r-1} \mu.$$

This condition is entirely similar to (3.37), but one level higher. We may therefore repeat the reasoning at levels $j + 1, \dots, r - 1$, yielding the bound

$$\Delta_{r,k} < \gamma_1^{r+2-(r-j)} (\epsilon_{\min}^\Delta)^{r-(r-j)} \mu = \gamma_1^{j+2} (\epsilon_{\min}^\Delta)^j \mu < \gamma_1 \mu.$$

But this last inequality contradicts (3.35), and we therefore deduce that (3.37) never holds. This proves (3.34) with

$$\Delta_{\min} \stackrel{\text{def}}{=} \gamma_1^{r+2} (\epsilon_{\min}^\Delta)^r \min[\Delta_{\min}^s, \kappa_2 \epsilon_r^g] \quad (3.39)$$

and the bounds $\gamma_1 \in (0, 1)$, $\epsilon_{\min}^\Delta \in (0, 1)$, $\kappa_2 \in (0, 1)$ and $\epsilon_r^g \in (0, 1)$ together imply that $\Delta_{\min} \in (0, \min[\Delta_{\min}^s, 1])$, as requested. \square

This result must be compared to Theorem 6.4.3 on p. 135 of Conn et al. (2000), keeping (2.31) in mind with the fact that we have called the uppermost minimization level with some nonzero tolerance ϵ_r^g . Also note in (3.39) that Δ_{\min} is linearly proportional to ϵ_r^g for small enough values of this threshold.

The next crucial step of our analysis is to show that the algorithm is well-defined in that all the recursions are finite.

Theorem 3.7 *The number of iterations at each level is finite. Moreover, there exists $\kappa_h \in (0, 1)$ such that, for every minimization sequence at level $i = 0, \dots, r$,*

$$h_i(x_{i,0}) - h_i(x_{i,p+1}) \geq \tau_{i,p} \eta_1^{i+1} \kappa_h,$$

where $\tau_{i,p}$ is the total number of successful iterations in $\bigcup_{\ell=0}^p \mathcal{T}(i, \ell)$.

Proof. We prove the desired result by induction on higher and higher levels from 0 to r . We start by defining $\omega_{i,\ell}$ to be the number of successful iterations in $\mathcal{T}(i, \ell)$ as well as the number of successful iterations in the set $\bigcup_{\ell=0}^p \mathcal{T}(i, \ell)$:

$$\tau_{i,p} = \sum_{\ell=0}^p \omega_{i,\ell}. \quad (3.40)$$

Note that $\omega_{i,\ell} \geq 1$ if iteration (i, ℓ) is successful.

Consider first an arbitrary minimization sequence at level 0 (if any). Every iteration in this sequence must be a Taylor iteration, which means that every successful iteration in the sequence satisfies

$$\begin{aligned} h_0(x_{0,\ell}) - h_0(x_{0,\ell+1}) &\geq \eta_1 \kappa_{\text{red}} \epsilon_0^g \min \left[\frac{\epsilon_0^g}{\kappa_H}, \Delta_{\min} \right] \\ &\geq \omega_{0,\ell} \eta_1 \kappa_{\text{red}} \epsilon_{\min}^g \min \left[\frac{\epsilon_{\min}^g}{\kappa_H}, \Delta_{\min} \right], \end{aligned} \quad (3.41)$$

where we have used (2.15), (2.31), (2.12), Theorem 3.6, (3.4) and the fact that $\omega_{0,\ell} = 1$ for every successful iteration $(0, \ell)$ because $\mathcal{T}(0, \ell) = \{(0, \ell)\}$. Since we know from Lemma 3.4 that there is at least one such iteration for every minimization

sequence, we may now sum the objective decreases at level 0 and obtain from (3.41) that

$$h_0(x_{0,0}) - h_0(x_{0,p+1}) = \sum_{\ell=0}^p \text{^{(S)}} [h_0(x_{0,\ell}) - h_0(x_{0,\ell+1})] \geq \tau_{0,p} \eta_1 \kappa_h, \quad (3.42)$$

where the sum with superscript (S) is restricted to successful iterations and where

$$\kappa_h \stackrel{\text{def}}{=} \kappa_{\text{red}} \epsilon_{\text{min}}^g \min \left[\frac{\epsilon_{\text{min}}^g}{\kappa_H}, \Delta_{\text{min}} \right] \in (0, 1). \quad (3.43)$$

But we have assumed that h_0 is bounded below, and therefore (3.42) imposes that $\tau_{0,p}$ must be finite. Since $\tau_{0,p}$ accounts for all successful iterations in the sequence, we obtain that there must be a last finite successful iteration $(0, \ell_0)$. If the sequence were nevertheless infinite, this would mean that every iteration $(0, \ell)$ is unsuccessful for all $\ell > \ell_0$, causing $\Delta_{j,\ell}$ to converge to zero, which is impossible in view of Theorem 3.6. Hence the minimization sequence is finite. The same reasoning may be applied to every such sequence at level 0.

Now consider an arbitrary minimization sequence at level i and assume that each minimization sequence at level $i - 1$ is finite and also that each successful iteration $(i - 1, u)$ in every minimization sequence at this lower level satisfies

$$h_{i-1}(x_{i-1,u}) - h_{i-1}(x_{i-1,u+1}) \geq \omega_{i-1,u} \eta_1^i \kappa_h, \quad (3.44)$$

which is the direct generalization of (3.41) at level $i - 1$. Consider a successful iteration (i, ℓ) , whose existence is ensured by Lemma 3.4. If it is a Taylor iteration (i.e., if $(i, \ell) \in \mathcal{T}(i, \ell)$), we obtain as above that

$$h_i(x_{i,\ell}) - h_i(x_{i,\ell+1}) \geq \eta_1 \kappa_h \geq \eta_1^{i+1} \kappa_h = \omega_{i,\ell} \eta_1^{i+1} \kappa_h \quad (3.45)$$

since $\eta_1 \in (0, 1)$ and $\omega_{i,\ell} = 1$ for every successful Taylor iteration. If, on the other hand, iteration (i, ℓ) uses Step 2, then, assuming $x_{i-1,*} = x_{i-1,t+1}$, we obtain that

$$\begin{aligned} h_i(x_{i,\ell}) - h_i(x_{i,\ell+1}) &\geq \eta_1 [h_{i-1}(x_{i-1,0}) - h_{i-1}(x_{i-1,*})] \\ &= \eta_1 \sum_{u=0}^t \text{^{(S)}} [h_{i-1}(x_{i-1,u}) - h_{i-1}(x_{i-1,u+1})]. \end{aligned}$$

Observing that $\omega_{i,\ell} = \tau_{i-1,t}$, our induction assumption (3.44) and (3.40) then give that

$$h_i(x_{i,\ell}) - h_i(x_{i,\ell+1}) \geq \eta_1^{i+1} \kappa_h \sum_{u=0}^t \omega_{i-1,u} = \tau_{i-1,t} \eta_1^{i+1} \kappa_h = \omega_{i,\ell} \eta_1^{i+1} \kappa_h. \quad (3.46)$$

Combining (3.45) and (3.46), we see that (3.44) again holds at level i instead of $i - 1$. Moreover, as above,

$$h_i(x_{i,0}) - h_i(x_{i,p+1}) = \sum_{\ell=0}^p \text{^{(S)}} [h_i(x_{i,\ell}) - h_i(x_{i,\ell+1})] \geq \tau_{i,p} \eta_1^{i+1} \kappa_h,$$

for the minimization sequence including iteration (i, ℓ) , and the number of successful iterations in this sequence must again be finite since h_i is bounded below. As for level 0, we may then conclude that the number of iterations (both successful and unsuccessful) in the sequence is finite. Moreover, the same reasoning holds for every minimization sequence at level i , and the induction is complete. \square

A first remarkable consequence of this theorem is an upper bound on the number of iterations needed by the trust-region algorithm to reduce the gradient norm at level r below a given threshold value.

Corollary 3.8 *Assume that one knows a constant f_{low} such that $h_r(x_r) = f(x) \geq f_{\text{low}}$ for every $x \in \mathbb{R}^n$ and that $\epsilon_r^g = \epsilon_{\min}^g$. Then Algorithm RMTR needs at most*

$$\left\lceil \frac{f(x_0) - f_{\text{low}}}{\theta(\epsilon_r^g)} \right\rceil$$

successful Taylor iterations at any level to obtain an iterate $x_{r,k}$ such that $\|g_{r,k}\| \leq \epsilon_r^g$, where

$$\theta(\epsilon) = \eta_1^{r+1} \kappa_{red} \epsilon \min \left[\frac{\epsilon}{\kappa_H}, \gamma_1^{r+2} (\epsilon_{\min}^{\Delta})^r \min[\Delta_{\min}^s, \kappa_2 \epsilon] \right].$$

Proof. The desired bound directly follows from Theorem 3.7, (3.43) and (3.39). (To keep the expression manageable, we have refrained from substituting the value of κ_2 from (3.17) and, in this value, that of κ_1 from (3.5), all this values being independent of ϵ .) \square

Of course, the bound provided by this corollary may be very pessimistic and not all the constants in the definition of $\theta(\epsilon)$ may be known in practice, but this loose complexity result is nevertheless theoretically interesting as it applies to rather general nonconvex problems. The condition that $\epsilon_r^g = \epsilon_{\min}^g$ is not restrictive as it does not make much practical sense to require an accuracy at lower levels that is higher than at upper ones. One should also note that the bound is in terms of iteration numbers, and only implicitly accounts for the cost of computing a Taylor step satisfying (2.15).

Corollary 3.8 suggests several comments.

1. The bound involves the number of successful Taylor iterations, that is successful iterations where the trial step is computed without resorting to further recursion. This provides an adequate measure of the linear algebra effort for all successful iterations, since successful iterations using the recursion of Step 2 cost very little beyond the evaluation of the level-dependent objective function and its gradient. Moreover, the number of such iterations is, by construction, at most equal to r times that of Taylor iterations (in the worst case where each iteration at level r includes a full recursion to level 0 with a single successful iteration at each level $j > 0$).

Hence the result shows that the number of necessary successful iterations, all levels included, is of order $1/\epsilon^2$ for small values of ϵ . This order is not qualitatively

altered by the inclusion of unsuccessful iterations either, provided we replace the very successful trust-region radius update (top case in (2.21)) by

$$\Delta_{i,k}^+ \in [\Delta_{i,k}, \gamma_3 \Delta_{i,k}] \quad \text{if } \rho_{i,k} \geq \eta_2,$$

for some $\gamma_3 > 1$. Indeed, Theorem 3.6 imposes that the decrease in radius caused by unsuccessful iterations must asymptotically be compensated by an increase at successful ones, irrespective of the fact that Δ_{\min} depends on ϵ by (3.39). This is to say that, if α is the average number of unsuccessful iterations per successful one at any level, then one must have that $\gamma_3 \gamma_2^\alpha \geq 1$, and therefore that $\alpha \leq -\log(\gamma_3)/\log(\gamma_2)$. Thus the complexity bound in $1/\epsilon^2$ for small ϵ is only modified by a constant factor if all iterations (successful and unsuccessful) are considered.

2. It is also very interesting to note that the bound involves the number of successful Taylor iterations *summed up on all levels* (as a result of Theorem 3.7). Thus successful such iterations at cheap low levels decrease the number of necessary expensive ones at higher levels, and the multilevel algorithm requires (at least in the theoretical worst case) less Taylor iterations at the upper level than the single-level variant. This provides some theoretical backing to the practical experience that the structure of multilevel unconstrained optimization problems can be used to advantage.
3. Finally, we note that little effort has been made in the above proofs to keep the value $\theta(\epsilon)$ as large as possible (and hence the upper bound on the number of iterations as small as possible) for a given ϵ . However, we observe with interest that the constants involved in this definition do not depend on the problem dimension, but rather on the properties of the problem ($r, \kappa_H, \kappa_\sigma$) or of the algorithm itself ($\kappa_{\text{red}}, \kappa_g, \gamma_1, \eta_1, \eta_2, \epsilon_{\min}^\Delta, \Delta_{\min}^s$).

A second important consequence of Theorem 3.7 is that the algorithm is globally convergent, in the sense that it generates a subsequence of iterates whose gradients converge to zero if run with $\epsilon_r^g = 0$.

Corollary 3.9 *Assume that Algorithm RMTR is called at the uppermost level with $\epsilon_r^g = 0$. Then*

$$\liminf_{k \rightarrow \infty} \|g_{r,k}\| = 0. \tag{3.47}$$

Proof. We first observe that the sequence of iterates $\{x_{r,k}\}$ generated by the algorithm called with $\epsilon_r^g = 0$ is identical to that generated as follows. We consider, at level r , a sequence of gradient tolerances $\{\epsilon_{r,j}^g\} \in (0, 1)$ monotonically converging to zero, start the algorithm with $\epsilon_r^g = \epsilon_{r,0}^g$ and alter slightly the mechanism of Step 5 (at level r only) to reduce ϵ_r^g from $\epsilon_{r,j}^g$ to $\epsilon_{r,j+1}^g$ as soon as $\|g_{r,k+1}\| \leq \epsilon_{r,j}^g$. The calculation is then continued with this more stringent threshold until it is also

attained, ϵ_r^g is then again reduced and so on. Since $\Delta_{r+1} = \infty$, each successive minimization at level r can only stop at iteration k if

$$\|g_{r,k+1}\| \leq \epsilon_{r,j}^g. \quad (3.48)$$

Theorem 3.7 then ensures that there are only finitely many successful iterations between two reductions of ϵ_r^g . We therefore obtain that for each $\epsilon_{r,j}^g$ there is an arbitrarily large k such that (3.48) holds. The desired result then follows immediately from our assumption that $\{\epsilon_{r,j}^g\}$ converges to zero. \square

The interest of this result is mostly theoretical, since most practical applications of Algorithm RMTR consider a nonzero gradient tolerance ϵ_r^g .

The attentive reader may have noticed that our theory still applies when we modify the technique described at the start of Corollary 3.9 by allowing a reduction of all the ϵ_i^g to zero at the same time⁽⁴⁾, instead of merely reducing the uppermost one. If this modified technique is used, each minimization sequence in the algorithm becomes infinite (as if it were initiated with a zero gradient threshold) and recursion to lower levels remain possible for arbitrarily small gradients, and may therefore occur arbitrarily far in the sequence of iterates. Moreover, we may still apply Corollary 3.9 at each level and deduce that

$$\liminf_{k \rightarrow \infty} \|g_{i,k}\| = 0 \quad (3.49)$$

for all $i = 0, \dots, r$.

As is the case for single-level trust-region algorithms, we now would like to prove that the limit inferior in (3.47) (and possibly (3.49)) can be replaced by a true limit, while still allowing recursion for very small gradients. We start by deriving a variant of Theorem 3.7 that does not assume that *all* gradient norms remain above some threshold to obtain a measure of the predicted decrease at some iteration (i, k) .

Lemma 3.10 *There exists a constant $\kappa_3 \in (0, 1)$ such that, for all (i, k) ,*

$$\delta_{i,k} \geq \kappa_{red} \eta_1^r \gamma_1^r \kappa_g^r \|g_{i,k}\| \min[\Delta_{\min}^s, \kappa_3 \|g_{i,k}\|, \Delta_{i,k}]. \quad (3.50)$$

Proof. Consider iteration (i, k) . If it is a Taylor iteration, then, if we set

$$\kappa_3 = \min \left[\frac{\kappa_g^r}{\kappa_H}, \kappa_2 \kappa_g^r \right] = \kappa_2 \kappa_g^r \in (0, 1), \quad (3.51)$$

(3.50) immediately follows from (2.15), (2.12) and the bounds $\kappa_g \in (0, 1)$, $\eta_1 \in (0, 1)$ and $\gamma_1 \in (0, 1)$. Otherwise define the iteration (j, ℓ) (with $j < i$) to be the deepest successful iteration in $\mathcal{R}(i, k)$ such that

$$g_{j,0} = g_{j,1} = \dots = g_{j,\ell} = R_{j+1} \dots R_i g_{i,k}$$

and such that all iterations $(j+1, t_{j+1}), (j+2, t_{j+2}), \dots$, up to $(i-1, t_{i-1})$ of the path from (j, ℓ) to (i, k) are successful (meaning that iterations (j, u) are unsuccessful

⁽⁴⁾The ratios $\epsilon_i^g / \epsilon_r^g$ could for instance be fixed or kept within prescribed bounds.

for $u = 0, \dots, \ell - 1$, if any, and that iterations (p, u) are also unsuccessful for $p = j + 1, \dots, i - 1$ and $u = 0, \dots, t_p - 1$, if any). Note that such a path is guaranteed to exist because of Lemma 3.4. Using the first part of (2.19), we then obtain that

$$\|g_{j,0}\| = \|g_{j,1}\| = \dots = \|g_{j,\ell}\| = \|R_{j+1} \dots R_i g_{i,k}\| \geq \kappa_g^r \|g_{i,k}\|. \quad (3.52)$$

If $\ell = 0$, then

$$\Delta_{j,\ell} = \min[\Delta_j^s, \Delta_{j+1,t_{j+1}}] \geq \min[\Delta_{\min}^s, \Delta_{j+1,t_{j+1}}]. \quad (3.53)$$

If, on the other hand, $\ell > 0$, we know that iterations $(j, 0)$ to $(j, \ell - 1)$ are unsuccessful. Corollary 3.3 then implies for iteration $(j, \ell - 1)$ that

$$\Delta_{j,\ell-1} > \min[\Delta_{\min}^s, \kappa_2 \|g_{j,\ell-1}\|] = \min[\Delta_{\min}^s, \kappa_2 \|g_{j,0}\|].$$

But this inequality, (2.21), (2.22), the unsuccessful nature of the first ℓ iterations at level j , (3.52) and the bound $\gamma_1 < 1$ then yield that

$$\begin{aligned} \Delta_{j,\ell} &\geq \min[\gamma_1 \Delta_{j,\ell-1}, \Delta_{j+1,t_{j+1}} - \|x_{j,0} - x_{j,\ell}\|_j] \\ &= \min[\gamma_1 \Delta_{j,\ell-1}, \Delta_{j+1,t_{j+1}}] \\ &\geq \min[\gamma_1 \min(\Delta_{\min}^s, \kappa_2 \|g_{j,0}\|), \Delta_{j+1,t_{j+1}}] \\ &\geq \min[\gamma_1 \min(\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|), \Delta_{j+1,t_{j+1}}] \\ &\geq \gamma_1 \min[\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|, \Delta_{j+1,t_{j+1}}]. \end{aligned}$$

Combining this last inequality with (3.53), we conclude that

$$\Delta_{j,\ell} \geq \gamma_1 \min[\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|, \Delta_{j+1,t_{j+1}}]$$

for $\ell \geq 0$. Our choice of iteration (j, ℓ) also ensures that the same reasoning can now be applied not only to iteration (j, ℓ) , but also to every iteration in the path $(j + 1, t_{j+1}), \dots, (i - 1, t_{i-1})$, because the first part of (2.19) implies that

$$\|g_{p,0}\| = \|R_{p+1} \dots R_i g_{i,k}\| \geq \kappa_g^r \|g_{i,k}\|,$$

for all $j \leq p < i$. Thus we obtain that

$$\Delta_{j+u,t_{j+u}} \geq \gamma_1 \min[\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|, \Delta_{j+u+1,t_{j+u+1}}]$$

for $u = 1, \dots, i - j - 1$ (where we identify $t_i = k$ for $u = i - j - 1$). We may then use these bounds recursively level by level and deduce that

$$\begin{aligned} \Delta_{j,\ell} &\geq \gamma_1 \min[\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{j,k}\|, \Delta_{j+1,t_{j+1}}] \\ &\geq \gamma_1 \min[\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|, \gamma_1 \min(\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|, \Delta_{j+2,t_{j+2}})] \\ &\geq \gamma_1^2 \min[\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|, \Delta_{j+2,t_{j+2}}] \\ &\geq \gamma_1^r \min[\Delta_{\min}^s, \kappa_2 \kappa_g^r \|g_{i,k}\|, \Delta_{i,k}] \end{aligned} \quad (3.54)$$

because $\gamma_1 < 1$. On the other hand, $(j, \ell) \in \mathcal{T}(i, k)$ by construction, and we therefore obtain from (2.15) and (2.12) that

$$\delta_{j,\ell} \geq \kappa_{\text{red}} \|g_{j,\ell}\| \min \left[\frac{\|g_{j,\ell}\|}{\kappa_{\text{H}}}, \Delta_{j,\ell} \right]. \quad (3.55)$$

Gathering now (3.52), (3.54) and (3.55), we obtain that

$$\delta_{j,\ell} \geq \kappa_{\text{red}} \kappa_{\text{g}}^r \|g_{i,k}\| \min \left[\frac{\kappa_{\text{g}}^r \|g_{i,k}\|}{\kappa_{\text{H}}}, \gamma_1^r \min[\Delta_{\text{min}}^{\text{s}}, \kappa_2 \kappa_{\text{g}}^r \|g_{i,k}\|, \Delta_{i,k}] \right],$$

and thus, using (3.51), that

$$\delta_{j,\ell} \geq \kappa_{\text{red}} \kappa_{\text{g}}^r \gamma_1^r \|g_{i,k}\| \min[\Delta_{\text{min}}^{\text{s}}, \kappa_3 \|g_{i,k}\|, \Delta_{i,k}]. \quad (3.56)$$

But the fact that all iterations on the path from (j, ℓ) to (i, k) are successful also implies that

$$\begin{aligned} \delta_{i,k} &= h_{i-1}(x_{i-1,0}) - h_{i-1}(x_{i-1,*}) \\ &\geq h_{i-1}(x_{i-1,t_{i-1}}) - h_{i-1}(x_{i-1,t_{i-1}+1}) \\ &\geq \eta_1 \delta_{i-1,t_{i-1}} \\ &= \eta_1 [h_{i-2}(x_{i-2,0}) - h_{i-2}(x_{i-2,*})] \\ &\geq \eta_1 [h_{i-2}(x_{i-2,t_{i-2}}) - h_{i-2}(x_{i-2,t_{i-2}+1})] \\ &\geq \eta_1^2 \delta_{i-2,t_{i-2}} \\ &\geq \eta_1^r \delta_{j,\ell}. \end{aligned}$$

The bound (3.50) then follows from this last inequality and (3.56). \square

All the elements are now in place to show that, if the algorithm is run with $\epsilon_r^{\text{g}} = 0$, then gradients at level r converge to zero.

Theorem 3.11 *Assume that Algorithm RMTR is called at the uppermost level with $\epsilon_r^{\text{g}} = 0$. Then*

$$\lim_{k \rightarrow \infty} \|g_{r,k}\| = 0. \quad (3.57)$$

Proof. The proof is identical to that of Theorem 6.4.6 on p. 137 of Conn et al. (2000), with (3.50) (with $i = r$) now playing the role of the sufficient model reduction condition AA.1 at level r . \square

This last result implies, in particular, that any limit point of the infinite sequence $\{x_{r,k}\}$ is first-order critical for problem (2.1). But we may draw stronger conclusions. If we assume again that all ϵ_i^{g} ($i = 0, \dots, r-1$) are driven down to zero together with ϵ_r^{g} (thus allowing recursion even for very small gradients), then, as explained above, each minimization sequence in the algorithm becomes infinite, and we may apply Theorem 3.11 to each of them, concluding that

$$\lim_{k \rightarrow \infty} \|g_{i,k}\| = 0$$

for every level $i = 0, \dots, r$. The behaviour of Algorithm RMTR is therefore truly coherent with its multilevel formulation, since the same convergence results hold for each level.

While this property is encouraging from the theoretical point of view, allowing longer and longer minimization sequences at lower levels may not be the most efficient strategy in practice. Instead, one would rather look for ways to prematurely terminate such sequences if significant progress at the higher level is not likely. There are of course many possible ways to achieve this goal, but they all rely on truncating the lower level sequence possibly before the termination tests of Step 5 are satisfied. For instance, one might think to stop a minimization sequence after a preset number of successful iterations: in combination with the freedom left at Step 1 to choose the model whenever (2.19) holds, this strategy allows a straightforward implementation of fixed lower-iterations patterns, like the V or W cycles in multigrid methods (see Briggs et al., 2000). Alternatively, one might replace the termination test on the absolute accuracy on the gradient norm $\|g_{i,k+1}\| \leq \epsilon_i^g$ in Step 5 by the more elaborate test

$$\|g_{i,k+1}\| \leq \min[\epsilon_i^{\text{ag}}, \epsilon_i^{\text{rg}}\|g_{i,0}\|],$$

where ϵ_i^{ag} and ϵ_i^{rg} now define the absolute and relative gradient accuracy that is required for terminating the minimization sequence at level i . Fortunately, “premature” termination does not affect the convergence results at the upper level, provided each minimization sequence contains at least one successful iteration. The key to this observation is that Lemmas 2.1 and 2.2 do not depend on the actual stopping criterion used, and that all subsequent proofs do not depend on it either. Hence we see that Algorithm RMTR covers a wide class of possible implementations. As the initial point is left at the user choice, our algorithm also allows for simpler mesh refinement schemes for computing good initial estimates.

Finally, we note that the conditions that we have used to describe the relations between the different levels are not the most general possible. For instance, our theory remains essentially unchanged if we merely insist on first-order coherence (i.e., conditions (2.8) and (2.10)) to hold only for small enough trust-region radii $\Delta_{i,k}$, or only up to a perturbation of the order of $\Delta_{i,k}$ or $\|g_{i,k}\|\Delta_{i,k}$. Other generalizations may be possible along those lines, in particular for simplified models of other types than arising from a coarser discretization. Similarly, although we have assumed for motivation purposes that each f_i is “more costly” to minimize than f_{i-1} , we have not used this feature in the theory presented above.

4 Comments and perspectives

We have defined a recursive trust-region algorithm that is able to exploit cheap lower levels models in a multilevel optimization problem. This algorithm has been proved to be well-defined and globally convergent to first-order. We have also presented a theoretical complexity result giving a bound on the number of iterations that are required

by the algorithm to find an approximate critical point of the objective function within prescribed accuracy. This last result also shows that the total complexity of solving an unconstrained multilevel problem can be shared amongst the levels, exploiting the structure to advantage.

The theory developed here provides a set of very weak conditions linking neighbouring unconstrained optimization problems, ensuring that they can be used in the solution of one of them. In particular, these conditions only focus on the coherence of first-order information, which is then enough to derive first-order convergence results. This weak relationship between the considered problems is very pessimistic in practice, since one might typically expect problems arising from different discretization of the same underlying infinite-dimensional problem to share more than local first-order behaviour. In particular, curvature information might be preserved across discretizations, which gives hopes that a recursive second-order convergence theory might also be developed.

Although the example of discretized problems has been used as a major motivation for our work, this is not the only case where our theory can be applied. We think in particular of cases where different models of the true objective function might live in the same space, but involve different levels of complexity and/or cost. This is for instance of interest in a number of problems arising from physics, like data assimilation in weather forecasting (see Fisher, 1998), where different models may involve different levels of sophistication in the physical modelling itself. More generally, the algorithm and theory presented here is relevant in most areas where simplified models are considered, such as multidisciplinary optimization (see Alexandrov, Dennis, Lewis and Torczon, 1998) or PDE-constrained problems (see Arian, Fahl and Sachs, 2000, Fahl and Sachs, 2003).

Following some recent research, we may also think to search for even more efficient algorithms by combining the trust-region framework developed here with other globalization techniques, like linesearches (see Toint, 1983, Nocedal and Yuan, 1998, Gertz, 1999), non-monotone techniques (see Xiao and Zhou, 1992, Xiao and Chu, 1995, Toint, 1997, Ulbrich, 1999) or filter methods (see Gould, Sainvitu and Toint, 2004). While this might add yet another level of technicity of the convergence proofs, we expect such extensions to be possible and the resulting algorithms to be of practical interest.

Applying recursive trust-region methods of the type discussed here to constrained problems is another potentially useful development. Although we anticipate the associated convergence theory to be again more technically difficult, intuition suggests that the power of such methods should also be exploitable in this case.

A number of practical issues related to Algorithm RMTR have not been discussed although they may be crucial in practice. A first such issue is trust-region scaling/preconditioning, which we think can be achieved by using a matrix M_r that is different from the identity. A second practical issue is the determination of the various thresholds ϵ_i^g and ϵ_i^A . There is in fact no need to make them depend only on the level, as above, but our developments may be trivially adjusted if one allows these thresholds to depend also on the iteration, provided they remain bounded below by some level-

dependent values. The same comment holds for the default radii $\Delta_i^{\$}$ at the start of each minimization sequence. Our intention is to discuss these points along with other practical considerations in a forthcoming paper (Gratton, Sartenaer and Toint, 2004) describing preliminary numerical experience with Algorithm RMTR.

Acknowledgements

The third author would like to thank Arnaud Fréville and the Nord-Pas-de-Calais region for scriptural support.

References

- N. M. Alexandrov, J. E. Dennis, R. M. Lewis, and V. Torczon. A trust region framework for managing the use of approximation models. *Structural Optimization*, **15**(1), 16–23, 1998.
- E. Arian, M. Fahl, and E. W. Sachs. Trust-region proper orthogonal decomposition for flow control. Technical Report 2000-25, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center Hampton, Virginia, USA, 2000.
- R. E. Banks, P. E. Gill, and R. F. Marcia. Interior point methods for a class of elliptic variational inequalities. *in* Biegler, Ghattas, Heinkenschloss and Van Bloemen Waanders (2003), pp. 218–235.
- S. J. Benson, L. C. McInnes, J. Moré, and J. Sarich. Scalable algorithms in optimization: Computational experiments. Preprint ANL/MCS-P1175-0604, Mathematics and Computer Science, Argonne National Laboratory, Argonne, Illinois, USA, 2004. To appear in the Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization (MA&O) Conference, August 30 - September 1, 2004.
- J. T. Betts and S. O. Erb. Optimal low thrust trajectory to the moon. *SIAM Journal on Applied Dynamical Systems*, **2**(2), 144–170, 2003.
- T. Biegler, O. Ghattas, M. Heinkenschloss, and B. Van Bloemen Waanders, editors. *High performance algorithms and software for nonlinear optimization*, Springer Verlag, Heidelberg, Berlin, New York, 2003.
- J. H. Bramble. *Multigrid Methods*. Longman Scientific and Technical, New York, 1993.
- A. Brandt. Multi-level adaptive solutions to boundary value problems. *Mathematics of Computation*, **31**(138), 333–390, 1977.
- W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial*. SIAM, Philadelphia, USA, 2nd edn, 2000.

- A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 01 in ‘MPS-SIAM Series on Optimization’. SIAM, Philadelphia, USA, 2000.
- M. Fahl and E. Sachs. Reduced order modelling approaches to PDE-constrained optimization based on proper orthogonal decomposition. *in* Biegler et al. (2003), pp. 268–281.
- M. Fisher. Minimization algorithms for variational data assimilation. *In* ‘Recent Developments in Numerical Methods for Atmospheric Modelling’, pp. 364–385. ECMWF, 1998.
- E. M. Gertz. *Combination Trust-Region Line-Search Methods for Unconstrained Optimization*. PhD thesis, Department of Mathematics, University of California, San Diego, California, USA, 1999.
- N. I. M. Gould, C. Sainvitu, and Ph. L. Toint. A filter-trust-region method for unconstrained optimization. Technical Report 04/03, Department of Mathematics, University of Namur, Namur, Belgium, 2004.
- S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multi-level nonlinear optimization (Part II): Numerical experience. Technical Report (in preparation), Department of Mathematics, University of Namur, Namur, Belgium, 2004.
- A. Griewank and Ph. L. Toint. Local convergence analysis for partitioned quasi-Newton updates. *Numerische Mathematik*, **39**, 429–448, 1982.
- W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*. Springer Series in Applied Mathematical Sciences. Springer Verlag, Heidelberg, Berlin, New York, 1994.
- W. Hackbusch. *Multi-grid Methods and Applications*. Number 4 in ‘Springer Series in Computational Mathematics’. Springer Verlag, Heidelberg, Berlin, New York, 1995.
- P. W. Hemker and G. M. Johnson. Multigrid approach to Euler equations. *In* S. F. McCormick, ed., ‘Multigrid methods’, Vol. 3 of *Frontiers in Applied Mathematics*, pp. 57–72, SIAM, Philadelphia, USA, 1987.
- M. Lewis and S. G. Nash. Model problems for the multigrid optimization of systems governed by differential equations. Technical report, Systems Engineering and Operations Research Department, George Mason University, Fairfax, Virginia, USA, December 2003.
- A. Migdalas, P. Pardalos, and P. Värbrand, editors. *Multilevel Optimization: Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

- J. J. Moré. Terascale optimal PDE solvers. Talk at the ICIAM 2003 Conference in Sydney, 2003.
- S. G. Nash. A multigrid approach to discretized optimization problems. Technical report, Systems Engineering and Operations Research Department, George Mason University, Fairfax, Virginia, USA, February 1999.
- J. Nocedal and Y. Yuan. Combining trust region and line search techniques. *In* Y. Yuan, ed., ‘Advances in Nonlinear Programming’, pp. 153–176, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
- M. J. D. Powell. A Fortran subroutine for unconstrained minimization requiring first derivatives of the objective function. Technical Report R-6469, AERE Harwell Laboratory, Harwell, Oxfordshire, England, 1970.
- A. Sartenaer. Automatic determination of an initial trust region in nonlinear programming. *SIAM Journal on Scientific Computing*, **18**(6), 1788–1803, 1997.
- Ph. L. Toint. VE08AD, a routine for partially separable optimization with bounded variables. *Harwell Subroutine Library*, **2**, 1983.
- Ph. L. Toint. A non-monotone trust-region algorithm for nonlinear optimization subject to convex constraints. *Mathematical Programming*, **77**(1), 69–94, 1997.
- M. Ulbrich. Non-monotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems. Technical Report TUM-M9906, Faculty of Mathematics, Technische Universität München, 1999.
- Y. Xiao and E. K. W. Chu. Nonmonotone trust region methods. Technical Report 95/17, Monash University, Clayton, Australia, 1995.
- Y. Xiao and F. Zhou. Nonmonotone trust region methods with curvilinear path in unconstrained optimization. *Computing*, **48**(3–4), 303–317, 1992.