Convergence of a regularized Euclidean residual algorithm for nonlinear least-squares

Bellavia, S.; Morini, B.; Cartis, C.; Gould, N.I.M.; Toint, Ph.L.

Published in:
SIAM Journal on Numerical Analysis

DOI:
10.1137/080732432

Publication date:
2010

Document Version
Early version, also known as pre-print

Convergence of a Regularized Euclidean Residual Algorithm for Nonlinear Least-Squares

by S. Bellavia[1], C. Cartis[2], N. I. M. Gould[3], B. Morini[1] and Ph. L. Toint[4]

8 October 2009

[1] Dipartimento di Energetica "S. Stecco",
Università di Firenze,
via C. Lombroso 6/17, 50134 Firenze, Italia,
Email: stefania.bellavia@unifi.it, benedetta.morini@unifi.it.

[2] School of Mathematics,
University of Edinburgh,
Edinburgh, EH9 3JZ, Scotland, United Kingdom.
Email: coralia.cartis@ed.ac.uk.

[3] Computational Science and Engineering Department,
Rutherford Appleton Laboratory,
Chilton, Oxfordshire, OX11 0QX, England.
Email: nick.gould@stfc.ac.uk.

[4] Department of Mathematics,
FUNDP-University of Namur,
61, rue de Bruxelles, B-5000 Namur, Belgium.
Email: philippe.toint@fundp.ac.be

Convergence of a Regularized Euclidean Residual Algorithm for Nonlinear Least-Squares

S. Bellavia, C. Cartis, N. I. M. Gould, B. Morini and Ph. L. Toint

8 October 2009

Abstract

The convergence properties of the new Regularized Euclidean Residual method for solving general nonlinear least-squares and nonlinear equations problems are investigated. This method, derived from a proposal by Nesterov (2007), uses a model of the objective function consisting of the unsquared Euclidean linearized residual regularized by a quadratic term. At variance with previous analysis, its convergence properties are here considered without assuming uniformly nonsingular globally Lipschitz continuous Jacobians, nor exact subproblem solution. It is proved that the method is globally convergent to first-order critical points, and, under stronger assumptions, to roots of the underlying system of nonlinear equations. The rate of convergence is also shown to be quadratic under stronger assumptions.

Keywords: Nonlinear least-squares, systems of nonlinear equations, numerical algorithms, global convergence.

1 Introduction

Finding values that minimize a specified norm $\|F(x)\|$ of a given vector-valued continuously-differentiable function $F \in \mathbb{R}^m$ of several variables $x \in \mathbb{R}^n$ is one of the corner-stones of computational mathematics. Although other norms are of interest, we shall concentrate here on the Euclidean-norm case, for this then leads to the equivalent nonlinear least-squares problem

$$\min_{x \in \mathbb{R}^n} f(x) \stackrel{\rm def}{=} \tfrac{1}{2}\|F(x)\|^2 \tag{1.1}$$

involving the continuously differentiable $f(x)$. This problem is not only of practical importance for its own sake in applications such as parameter identification, image registration and data assimilation, but its solution also forms the basis of many methods for solving optimization problems involving constraints. It is, for example, crucial when reducing constraint violation in several sequential quadratic programming (SQP) techniques (see Celis, 1985, Byrd, Schnabel and Shultz, 1987, Omojokun, 1989, Vardi, 1985, Powell and Yuan, 1990, etc.). The central problem of solving systems of nonlinear equations

$$F(x) = 0 \tag{1.2}$$

for $n = m$ is also covered by the formulation (1.1) in the sense that one then wishes to reduce the objective function $f(x)$ to zero. More generally, first-order optimality conditions for (1.1) require that

$$g(x) \equiv J^T(x)F(x) = 0$$

involving the Jacobian $J(x)$ of $F(x)$, and it is these conditions that we seek to satisfy.

Nearly all efficient methods for the solution of this problem are variants of Newton's method, in which (structured) quadratic approximations to $f(x)$ are minimized. However, such methods are well-known not to be globally convergent, and must then be modified to include safeguards to guarantee convergence from arbitrary starting points. Linesearch and trust regions (in which the quadratic model is minimized in a restricted neighbourhood of the current iterate) offer two standard safeguards (see Nocedal and Wright, 1999, or Conn, Gould and Toint, 2000, for more detail).

Interestingly, other techniques are possible, and methods based on adaptive *regularisation* have recently created some interest (see Griewank, 1981, Nesterov and Polyak, 2006, Weiser, Deuflhard and Erdmann, 2007, or Cartis, Gould and Toint, 2009*a*). In such methods, the smooth objective function's model is minimized in a neighbourhood implicitly defined by a regularisation term which penalizes the third power of the step length. In this paper, we consider another new technique proposed by Nesterov (2007) for the special case of nonlinear systems (1.2). This technique is different from previous approaches in that it uses a non-smooth model of $\|F(x)\|$, based on a linearization of F, rather than the smooth $f(x)$ which is then regularized by a quadratic term. However, as is the case for the cubic regularisation, this model can be consistently interpreted as an overestimation of the function $\|F(x)\|$ when the Jacobian matrix F is Lipschitz continuous, an intuitively appealing property. Another interesting[1] feature of this technique is that it only involves the Jacobian of F and thus that its expected performance depends on the condition number of that matrix (rather than on its square) for zero-residual problems. In his paper, Nesterov proves interesting global complexity results and fast local rate of convergence, under rather restrictive assumptions requiring that $m \leq n$, that $J(x)$ is uniformly full-rank and globally Lipschitz continuous, and that the model is globally minimized exactly at every iteration. His global convergence analysis to first-order critical points is more general, but still requires the latter and global Lipschitz continuity of the Jacobian. Note that this is a simplified version of the generalised class of proximal-point methods (e.g., Rockafellar, 1976) applied to the model rather than actual objective $\|F(x)\|$.

As discussed at the end of our paper (see Section 5), this new class of methods appear to compare favourably with a modern trust-region code in some nontrivial examples. It is thus of interest to investigate its convergence properties, especially in a weaker setting than that considered by Nesterov, allowing now for general (possibly over- or under-determined) nonlinear least-squares for both the global and local analyses, and without requiring the exact solution of subproblems. Furthermore, global convergence is proved without concerns for the global or local Lipschitz continuity or full-rank property of the Jacobian, while the fast local rate analysis requires only local full-rank property, and minimal Lipschitz continuity, of the Jacobian. Section 2 first describes the method in more detail. The global convergence analysis to first-order critical points is then carried out in Section 3 under very weak assumptions on the step calculation. Section 4 then investigates how a more accurate step can be computed and the implication of this improvement on the local convergence properties. Preliminary numerical experience is presented in Section 5. Finally, some conclusions and perspectives are discussed in Section 6.

Throughout the paper, a subscript will denote an iteration counter, and for a particular iterate x_k and relevant function $h(x)$, h_k will be shorthand for $h(x_k)$. The (appropriately-dimensioned) identity matrix will be denoted by I.

2 The method

We start by introducing the "modified Gauss-Newton method" proposed by Nesterov (2007) and its extension to the general nonlinear least-squares case, which uses the same motivation. If we assume that $J(x)$ is globally Lipschitz continuous (with constant $2L$) and since Taylor's theorem gives that, for some iterate x_k,

$$F(x_k + p) = F(x_k) + J(x_k)p + \int_0^1 (J(x_k + tp) - J(x_k))p\, dt,$$

we deduce from the triangle inequality that

$$\begin{aligned} \|F(x_k + p)\| &\leq \|F(x_k) + J_k p\| + \|p\| \int_0^1 \|J(x_k + tp) - J_k\|\, dt \qquad (2.1) \\ &\leq \|F(x_k) + J_k p\| + L\|p\|^2 \stackrel{\text{def}}{=} m_k^{\text{N}}(p). \qquad (2.2) \end{aligned}$$

Therefore, if we knew the constant L and if we were able to compute a step p_k minimizing the model $m_k^{\text{N}}(p)$, then the point $x_{k+1} = x_k + p_k$ must improve $\|F(x)\|$ and hence the objective function $f(x)$ of (1.1). Here we follow a more indirect approach suggested by Griewank (1981), Cartis et al. (2009*a*)

[1] Although not unique: several variants of the trust-region Gauss-Newton method share this property. See Cartis, Gould and Toint (2009*b*), for instance.

and (in a simpler form) by Nesterov and Polyak (2006) and Nesterov (2007), and introduce a dynamic positive parameter σ_k and the non-smooth model

$$m_k^0(p) \stackrel{\rm def}{=} \|F(x_k) + J_k p\| + \sigma_k \|p\|^2 \tag{2.3}$$

of $\|F(x)\|$ around x_k. Cartis et al. (2009*a*) provide rules for adapting the parameter σ_k in a numerically efficient manner. In this regard, it is important to note that the model (2.3) is an exact penalty function for the problem

$$\min_{p\in\mathbb{R}^n} \|p\|^2 \quad \text{subject to} \quad J_k p = -F(x_k),$$

and for all σ_k sufficiently small its minimizer solves $F(x_k) + J_k p = 0$, if such system is compatible (see Nocedal and Wright, 1999, §15.3). We would thus expect the Newton step (satisfying $J_k p = -F(x_k)$,) to be taken asymptotically for small enough σ_k.

In Nesterov (2007), the solution of the subproblem

$$\min_{x\in\mathbb{R}^n} m_k^{\rm N}(p) \tag{2.4}$$

is expressed in terms of the solution of a one-dimensional optimization problem with a non-negative simple bound while Cartis et al. (2009*b*) rely, for the minimization of m_k^0 in (2.3), on the equivalent differentiable constrained optimization problem

$$\min_{x\in\mathbb{R}^n,\nu\in\mathbb{R}} \quad \nu + \sigma_k\|p\|^2, \quad \text{subject to} \quad \|F(x_k) + J_k p\|^2 = \nu^2 \tag{2.5}$$

for some $\nu \geq 0$. The first-order optimality conditions for (2.5) take the form

$$\begin{pmatrix} \sigma_k p \\ 1 \end{pmatrix} = \xi \begin{pmatrix} {J_k}^T(F(x_k) + J_k p) \\ -2\nu \end{pmatrix}, \tag{2.6}$$

for any p such that the residual $\nu = \|F_k + J_k p\|$ is nonzero[2] and for some multiplier ξ. Letting

$$B_k \stackrel{\rm def}{=} {J_k}^T J_k, \quad F_k \stackrel{\rm def}{=} F(x_k) \quad \text{and} \quad g_k \stackrel{\rm def}{=} J_k^T F_k, \tag{2.7}$$

the vector p solves (2.6) if $p = p(\lambda)$ where $\lambda > 0$ and

$$(B_k + \lambda I)p = -g_k, \quad \lambda = 2\sigma_k \|F_k + J_k p\|. \tag{2.8}$$

Note that if there is a p for which $F_k + J_k p = 0$, then this p satisfies (2.8) along with $\lambda = 0$ and this case must be checked for before attempting to find another vector $\tilde{p}$ and a scalar $\tilde{\lambda} > 0$ which solve (2.8).

In this paper, we consider a slightly more general model of the form

$$m_k(p) \stackrel{\rm def}{=} \sqrt{\|F_k + J_k p\|^2 + \mu_k\|p\|^2} + \sigma_k\|p\|^2 \tag{2.9}$$

for some scalar $\mu_k \geq 0$ and attempt to find a step p by (possibly approximately) solving

$$\min_{p\in\mathbb{R}^n} m_k(p). \tag{2.10}$$

If $\mu_k > 0$ and $F_k \neq 0$, the model $m_k(p)$ is continuously differentiable, but this is not the case if $\mu_k = 0$ since its first and second derivatives are both undefined when $F_k + J_k p = 0$. However, it always enjoys the following desirable property.

Lemma 2.1 *Suppose that $\sigma_k > 0$. Then the model $m_k(p)$ is strictly convex for all $\mu_k \geq 0$.*

Proof. Indeed, since $m_k(p) = \phi(p) + \sigma_k\|p\|^2$, where

$$\phi(p) \stackrel{\rm def}{=} \sqrt{\|F_k + J_k p\|^2 + \mu_k\|p\|^2}, \tag{2.11}$$

and the function $\sigma_k\|p\|^2$ is strictly convex, $m_k(p)$ is strictly convex if $\phi(p)$ is convex. But the functions $g_1(p) = \|F_k + J_k p\|$, $g_2(p) = \sqrt{\mu_k}\|p\|$ are convex and nonnegative for all $p \in \mathbb{R}^n$. It then follows that $\phi(p)$ is convex. □

[2] If $F_k + J_k p = 0$, the constraint qualification (LICQ) fails for (2.5) and the first-order conditions (2.6) do not apply.

If $J(x)$ is globally Lipschitz continuous, then (2.1) and the inequality $m_k^0(p) \leq m_k(p)$ again ensure that $m_k(p)$ consistently overestimates $\|F(x+p)\|$ if $\sigma_k \geq L$.

The algorithm adopted to solve (1.1) then uses the model m_k along the lines of the adaptive cubic overestimation method proposed by Cartis et al. (2009*a*). As in this method, an approximate solution of (2.10) is allowed, in the sense that one accepts any step p such that the model (2.9) at p produces a value of the model smaller than that achieved by the *Cauchy point* given by

$$p_k^c = -\alpha_k g_k, \qquad \alpha_k = \operatorname*{argmin}_{\alpha \geq 0} m_k(-\alpha g_k), \tag{2.12}$$

with g_k being given by (2.7). Observe that α_k is uniquely defined in this last expression since m_k is strictly convex.

We may now state our algorithm more formally as Algorithm RER on this page. As is usual in trust-region methods, the iteration k will be called *successful* if $\rho_k \geq \eta_1$ and *unsuccessful* otherwise.

Algorithm 2.1: Regularized Euclidean Residual (RER) Algorithm

An initial point x_0 and the constants $\mu_0 \geq 0$, $\sigma_0 > 0$, $1 > \eta_2 > \eta_1 > 0$, $\gamma_2 \geq \gamma_1 > 1$, $\gamma_3 > 0$, $\epsilon_g > 0$ and $\epsilon_F > 0$ are given.

For $k = 0, 1, \ldots$ until $\|g_k\| < \epsilon_g$ or $\|F(x_k)\| < \epsilon_F$

Step 1: Compute an approximate minimizer p_k of $m_k(p)$ such that

$$m_k(p_k) \leq m_k(p_k^c), \tag{2.13}$$

where p_k^c is given in (2.12).

Step 2: Compute

$$\rho_k = \frac{\|F(x_k)\| - \|F(x_k + p_k)\|}{\|F(x_k)\| - m_k(p_k)}, \tag{2.14}$$

Step 3: Set

$$x_{k+1} = \begin{cases} x_k + p_k & \text{if } \rho_k \geq \eta_1, \\ x_k & \text{otherwise.} \end{cases}$$

Step 4: Set

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k] & \text{if } \rho_k \geq \eta_2 \quad \text{(very successful)}, \\ [\sigma_k, \gamma_1 \sigma_k) & \text{if } \eta_1 \leq \rho_k < \eta_2 \quad \text{(successful)}, \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k) & \text{otherwise} \quad \text{(unsuccessful)}. \end{cases} \tag{2.15}$$

Step 5: Set

$$\mu_{k+1} = \begin{cases} \min(\mu_k, \gamma_3 \|F_{k+1}\|) & \text{if } \rho_k \geq \eta_1, \\ \mu_k & \text{otherwise.} \end{cases} \tag{2.16}$$

It is important to note at this point that the denominator in (2.14) is always strictly positive whenever the current iterate is not a first-order critical point (this is proved in Lemma 3.2 below), and hence that the algorithm is well defined. Moreover, this also ensures that the sequence of successive $\|F(x_k)\|$ is non-increasing. For future reference, we also state the following simple properties of Algorithm 2.1.

Lemma 2.2

i) The sequence $\{\mu_k\}$ is non-negative, monotonically non-increasing and such that $\mu_k \leq \min[\mu_0, \gamma_3 \|F_k\|]$. As a consequence, the initial choice $\mu_0 = 0$ implies that $\mu_k = 0$ for all k, in which case $m_k(p) = m_k^0(p)$ at every iteration.

ii) If there exists a limit point x^* of the sequence $\{x_k\}$ of iterates generated by Algorithm RER such that $F(x^*) = 0$, then all limit points of $\{x_k\}$ are roots of $F(x) = 0$.

Proof. *ii)* Since the sequence $\{\|F_k\|\}$ is non-increasing and bounded below, it is convergent. Then, the existence of a limit point x^* such that $F(x^*) = 0$ implies that all limit points share this property, and thus every limit point of $\{x_k\}$ is a zero of F. □

Nesterov (2007) also proposes a simpler dynamic update of the regularisation parameter, which, in our notation, amounts to choosing $\eta_1 = \eta_2 = 1$, $\gamma_1 = \gamma_2 = 2$ and $\sigma_{k+1} = \frac{1}{2}\sigma_k$ whenever $\rho_k \geq 1$. When the Jacobian is Lipschitz continuous, this strategy may be proved to be efficient in the sense that it needs on average one increase in σ_k for one decrease. A similar property holds for our present choices (see Theorem 6.1 in Cartis et al., 2009*a*), and we believe that our greater flexibility (and, in particular, the possibility to choose η_1 to be a small positive constant) is important for the practical efficiency of the algorithm.

3 Global Convergence Analysis

We first make note of a simple bounding result, whose proof follows by inspection.

Lemma 3.1 *For all $\alpha \in [0, 1]$, we have that $\frac{1}{2}\alpha \leq 1 - \sqrt{1-\alpha} \leq \alpha$.*

In order to prove global convergence to first-order critical points, we first derive an easy consequence of the fact that an iterate is not first-order critical.

Lemma 3.2 *Assume that $g_k \neq 0$. Then, for $\mu_k \geq 0$,*

$$F_k \neq 0, \quad \langle g_k, (B_k + \mu_k I)g_k\rangle > 0, \qquad \text{and} \quad \langle p_k, (B_k + \mu_k I)p_k\rangle > 0 \tag{3.1}$$

and also that

$$m_k(p_k^C) < \|F_k\|. \tag{3.2}$$

Proof. The first statement in (3.1) immediately results from our assumption that $g_k = J_k^T F_k \neq 0$, from which we also deduce that $\|B_k\| = \|J_k^T J_k\| > 0$. Moreover, $J_k^T F_k \in \text{range}(J_k^T) = \text{null}(J_k)^\perp$ and thus $J_k g_k = J_k J_k^T F_k$ is nonzero. The first inequality in (3.1) then results from the identity $\langle g_k, (B_k + \mu_k I)g_k\rangle = \|J_k J_k^T F_k\|^2 + \mu_k \|g_k\|^2$. We also observe that the inequality $\langle \nabla_x m_k(0), g_k\rangle = \|g_k\|^2/\|F_k\| > 0$ ensures that $-g_k$ is a descent direction for m_k at 0, and thus that (3.2) follows from (2.12). Finally, $m_k(p_k) \leq m_k(p_k^C) < \|F_k\|$ because of (3.2) and (2.13). Thus $J_k p_k$ is nonzero and the last inequality of (3.1) follows from $\langle p_k, (B_k + \mu_k I)p_k\rangle = \|J_k p_k\|^2 + \mu_k \|p_k\|^2$. □

We now provide a lower bound on the decrease attained at the Cauchy step.

Lemma 3.3 *Assume that $g_k \neq 0$. Then we have that*

$$\|F_k\| - m_k(p_k) \geq \|F_k\| - m_k(p_k^c) \geq \frac{\|g_k\|^2}{4\|F_k\|} \min\left[\frac{1}{2\sigma_k\|F_k\|}, \frac{1}{\|B_k + \mu_k I\|}\right], \tag{3.3}$$

where we consider the Euclidean matrix norm.

Proof. For any $\alpha \geq 0$, we deduce from (2.9) and (2.11) that

$$m_k(-\alpha g_k) \quad = \quad \phi(-\alpha g_k) + \sigma_k \alpha^2 \|g_k\|^2 = \|F_k\|\sqrt{1 - \pi(\alpha)} + \sigma_k \alpha^2 \|g_k\|^2, \tag{3.4}$$

where

$$\pi(\alpha) = \frac{2\alpha\|g_k\|^2 - \alpha^2 g_k^T (B_k + \mu_k I)g_k}{\|F_k\|^2},$$

and the denominator of this last expression is nonzero because of Lemma 3.2. Trivially, we have that $1 - \pi(\alpha) \geq 0$; moreover $\pi(\alpha) > 0$ for any $\alpha \in (0, \bar{\alpha})$ where

$$\bar{\alpha} = \frac{2\|g_k\|^2}{g_k^T (B_k + \mu_k I)g_k},$$

which is also well-defined for $\mu_k \geq 0$ because of Lemma 3.2. Choosing $\alpha \in (0, \bar{\alpha})$, it follows that $0 < \pi(\alpha) \leq 1$. By Lemma 3.1, this implies that $\sqrt{1-\pi(\alpha)} \leq 1 - \pi(\alpha)/2$, and (3.4) then yields that

$$\begin{aligned} m_k(-\alpha g_k) - \|F_k\| &\leq \|F_k\| \left(1 - \frac{2\alpha\|g_k\|^2 - \alpha^2 g_k^T(B_k+\mu_k I)g_k}{2\|F_k\|^2}\right) + \sigma_k\alpha^2\|g_k\|^2 - \|F_k\| \\ &= -\frac{\alpha}{2\|F_k\|}\left[2\|g_k\|^2 - \alpha g_k^T(B_k+\mu_k I)g_k\right] + \sigma_k\alpha^2\|g_k\|^2 \\ &\leq \frac{\alpha\|g_k\|^2}{\|F_k\|}\big(-1 + \frac{\alpha}{2}\|B_k+\mu_k I\| + \sigma_k\alpha\|F_k\|\big). \end{aligned} \tag{3.5}$$

The right hand side of the last inequality is negative for any $\alpha \in (0, \hat{\alpha})$ with

$$\hat{\alpha} = \frac{2}{\|B_k+\mu_k I\| + 2\sigma_k\|F_k\|}.$$

Note that $\bar{\alpha} > \hat{\alpha}$ as

$$\bar{\alpha} \geq \frac{2\|g_k\|^2}{\|B_k+\mu_k I\|\|g_k\|^2} > \frac{2}{\|B_k+\mu_k I\| + 2\sigma_k\|F_k\|}.$$

Now, introduce

$$\alpha^* = \frac{1}{2\max(2\sigma_k\|F_k\|, \|B_k+\mu_k I\|)}. \tag{3.6}$$

Clearly, $\alpha^* < \hat{\alpha}$. Then, from (3.5) we obtain

$$\begin{aligned} m_k(-\alpha^* g_k) - \|F_k\| &\leq \frac{\alpha^*\|g_k\|^2}{\|F_k\|}\left(-1 + \frac{\alpha^*}{2}\|B_k+\mu_k I\| + \sigma_k\alpha^*\|F_k\|\right) \\ &\leq \frac{\alpha^*\|g_k\|^2}{\|F_k\|}\left(-1 + \frac{1}{2}\right) \\ &= -\frac{\|g_k\|^2}{4\|F_k\|}\frac{1}{\max(2\sigma_k\|F_k\|, \|B_k+\mu_k I\|)}, \end{aligned}$$

which completes the proof since $m_k(p_k) \leq m_k(p_k^c) \leq m_k(-\alpha^* g_k)$ because of (2.12) and (2.13). □

Using a similar methodology, we now derive a bound on the step.

Lemma 3.4 *Assume that $g_k \neq 0$. Then we have that*

$$\|p_k\| \leq \frac{2\|g_k\|}{\sigma_k\|F_k\|}. \tag{3.7}$$

Proof. The fact that $m_k(p_k) < \|F_k\|$ gives that $\|p_k\| > 0$, $\langle g_k, p_k\rangle \leq 0$ and

$$\|F_k\|\sqrt{1-\tau_k(p_k)} + \sigma_k\|p_k\|^2 < \|F_k\|, \tag{3.8}$$

where

$$\tau_k(p_k) \stackrel{\text{def}}{=} -\frac{2\langle g_k, p_k\rangle + \langle p_k, (B_k+\mu_k I)p_k\rangle}{\|F_k\|^2} = \frac{2|\langle g_k, p_k\rangle| - \langle p_k, (B_k+\mu_k I)p_k\rangle}{\|F_k\|^2}, \tag{3.9}$$

and also that $0 < \tau_k(p_k) \leq 1$. Note that $\tau_k(p_k)$ is well-defined because of Lemma 3.2. Hence, we have that

$$\sigma_k\|p_k\|^2 < \|F_k\|\left[1 - \sqrt{1-\tau_k(p_k)}\right] \leq \|F_k\|\tau_k(p_k) = \frac{2|\langle g_k, p_k\rangle| - \langle p_k, (B_k+\mu_k I)p_k\rangle}{\|F_k\|},$$

where we have used Lemma 3.1. This yields, using (3.1) and the Cauchy-Schwarz inequality, that

$$\sigma_k\|F_k\|\,\|p_k\|^2 < 2|\langle g_k, p_k\rangle| - \langle p_k, (B_k+\mu_k I)p_k\rangle \leq 2|\langle g_k, p_k\rangle| \leq 2\|g_k\|\,\|p_k\|. \tag{3.10}$$

Dividing both sides by $\|p_k\|$ then gives (3.7). □

To proceed in our analysis we make a further assumption on the Jacobian of $F(x)$.

Assumption 3.1 *Let $\{x_k\}$ be the sequence generated by the RER Algorithm. Then, there exists a positive constant $\kappa_{\rm J}$ such that, for all $k \geq 0$ and all $x \in [x_k, x_k + p_k]$,*

$$\|J(x)\| \leq \kappa_{\rm J}. \tag{3.11}$$

Note that the monotonic nature of the sequence $\{\|F_k\|\}$ implies that, for all k,

$$\|F_k\| \leq \|F_0\|. \tag{3.12}$$

We then immediately deduce from (3.11) and Lemma 2.2 that, for $k \geq 0$,

$$\|B_k + \mu_k I\| \leq \kappa_{\rm J}^2 + \gamma_3 \|F_0\| \stackrel{\rm def}{=} \kappa_{\rm D}. \tag{3.13}$$

Another consequence of (3.11) is that in the conditions of Lemma 3.4,(3.7) implies

$$\|p_k\| \leq \frac{2\kappa_{\rm J}}{\sigma_k}, \quad k \geq 0. \tag{3.14}$$

Next we give a bound on the error between the objective function and the model at the new candidate iterate.

Lemma 3.5 *Suppose that $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ is continuously differentiable. Then*

$$\|F(x_k + p_k)\| - m_k(p_k) \leq \|p_k\| \int_0^1 \|J(x_k + tp_k) - J_k\| \, dt - \sigma_k \|p_k\|^2, \quad k \geq 0. \tag{3.15}$$

Furthermore, if Assumption 3.1 holds, then

$$\|F(x_k + p_k)\| - m_k(p_k) \ \leq \frac{2\kappa_{\rm J}}{\sigma_k} \int_0^1 \|J(x_k + tp_k) - J_k\| \, dt, \quad k \geq 0. \tag{3.16}$$

Proof. The mean-value theorem implies that

$$F(x_k + p_k) = F(x_k) + J_k p_k + \int_0^1 \big[J(x_k + tp_k) - J_k\big] p_k \, dt,$$

which further gives that

$$\|F(x_k + p_k)\| \leq \|F_k + J_k p_k\| + \|p_k\| \int_0^1 \|J(x_k + tp_k) - J_k\| \, dt.$$

Therefore, using (2.11) and the inequality $\sqrt{a^2 + b^2} \geq a$ for all $a, b \geq 0$, we obtain that

$$\begin{aligned} \|F(x_k + p_k)\| - m_k(p_k) &= \|F(x_k + p_k)\| - \phi(p_k) - \sigma_k \|p_k\|^2 \\ &\leq \|F_k + J_k p_k\| - \phi(p_k) + \|p_k\| \int_0^1 \|J(x_k + tp_k) - J_k\| \, dt \\ &\quad - \sigma_k \|p_k\|^2 \\ &\leq \|p_k\| \int_0^1 \|J(x_k + tp_k) - J_k\| \, dt - \sigma_k \|p_k\|^2, \end{aligned}$$

as desired. The bound (3.16) now follows from (3.14) and (3.15). □

Next we show that provided there are only finitely many successful iterations, all later iterates are first-order critical.

Theorem 3.6 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable. Suppose that Assumption 3.1 holds and that there are only finitely many successful or very successful iterations. Then $x_k = x^*$ for all sufficiently large k and $g(x^*) = 0$.*

Proof. After the last successful iterate is computed, indexed by say k_0, the construction of the algorithm implies that $x_{k_0+1} = x_{k_0+i} \stackrel{\rm def}{=} x^*$, for all $i \geq 1$. Since all iterations $k \geq k_0 + 1$ are unsuccessful, the updating rule (2.15) implies that $\sigma_{k+1} \geq \gamma_1 \sigma_k$, with $\gamma_1 > 1$, for all $k \geq k_0 + 1$, and so

$$\sigma_k \to +\infty, \ \text{ as } \ k \to \infty. \tag{3.17}$$

If $\|g_{k_0+1}\| > 0$, then $\|g_k\| = \|g_{k_0+1}\| \stackrel{\rm def}{=} \epsilon > 0$, for all $k \geq k_0 + 1$. It now follows from (3.3), (3.12) and (3.13) that

$$\|F_k\| - m_k(p_k) \geq \frac{\epsilon^2}{8\|F_0\|^2} \min\left\{ \frac{1}{\sigma_k}, \frac{2\|F_0\|}{\kappa_{\rm D}} \right\}, \quad k \geq k_0 + 1,$$

and so, as (3.17) implies $1/\sigma_k \to 0$, we have

$$\|F_k\| - m_k(p_k) \geq \frac{\epsilon^2}{8\|F_0\|^2 \sigma_k}, \ \text{ for all } \ k \geq k_0 + 1 \ \text{ sufficiently large.} \tag{3.18}$$

This, (2.14) and (3.16) imply

$$0 \leq 1 - \rho_k = \frac{\|F(x_k + p_k)\| - m_k(p_k)}{\|F_k\| - m_k(p_k)} \leq 16\kappa_{\rm J}\|F_0\|^2 \epsilon^{-2} \int_0^1 \|J(x_k + tp_k) - J_k\| \, dt,$$

for all $k \geq k_0 + 1$ sufficiently large; the first inequality above holds, since $\rho_k \geq 1$ implies that k is very successful, which contradicts $k \geq k_0 + 1$ unsuccessful. Note that $x_k + tp_k = x^* + tp_k$ for all $k \geq k_0 + 1$, and that due to (3.14), (3.17) and $t \in [0, 1]$, we have $x^* + tp_k \to x^*$ as $k \to \infty$. Since $J_k = J_*$, $k \geq k_0 + 1$, and J continuous, we now conclude that

$$\|J(x_k + tp_k) - J_k\| \to 0, \quad k \to \infty, \quad t \in [0, 1],$$

and so $\rho_k \to 1$ as $k \to \infty$. This implies that for all k sufficiently large, $\rho_k \geq \eta_2$ and thus k is very successful. This contradicts $k \geq k_0 + 1$ unsuccessful. Thus $g_{k_0+1} = g_* = 0$. □

The following theorem states that at least one limit point of the sequence $\{x_k\}$ is a stationary point of problem (1.1).

Theorem 3.7 *Assume $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ is continuously differentiable and that Assumption 3.1 holds. Then*

$$\liminf_{k \to \infty} \|g_k\| = 0. \tag{3.19}$$

Proof. Note that if $g_k = 0$ for some k, then the RER Algorithm terminates and (3.19) holds (finitely). Also, if there are finitely many successful iterations, Theorem 3.6 implies the above. Thus without loss of generality, we may assume that $g_k \neq 0$ for all k and that there are infinitely many successful iterations, and let

$$\mathcal{S} = \{k \geq 0 \mid \text{ iteration } k \text{ is successful or very successful}\}.$$

To show that $\{\|g_k\|\}$ is not bounded away from zero, let us assume the contrary, namely, that there exists $\epsilon > 0$ such that

$$\|g_k\| \geq \epsilon, \ \text{ for all } \ k \geq 0. \tag{3.20}$$

Let us first prove that (3.20) implies that

$$\sum_{k \in \mathcal{S}} \frac{1}{\sigma_k} < +\infty. \tag{3.21}$$

It follows from (2.14), (2.15), (3.3) and (3.20) that

$$\|F_k\| - \|F_{k+1}\| \geq \eta_1 \frac{\epsilon^2}{4\|F_k\|} \min\left\{ \frac{1}{2\sigma_k \|F_k\|}, \frac{1}{\|B_k + \mu_k I\|} \right\}, \quad k \in \mathcal{S},$$

and furthermore, from (3.12) and (3.13),

$$\|F_k\| - \|F_{k+1}\| \geq \frac{\eta_1 \epsilon^2}{8\|F_0\|^2} \min\left\{\frac{1}{\sigma_k}, \frac{2\|F_0\|}{\kappa_{\rm D}}\right\}, \quad k \in \mathcal{S}. \tag{3.22}$$

Since $\{\|F_k\|\}$ is bounded below and monotonically non-increasing, it is convergent and hence the minimum in the right-hand side of (3.22) will be attained at $1/\sigma_k$ as the left-hand side of (3.22) converges to zero. Thus we have

$$\|F_k\| - \|F_{k+1}\| \geq \frac{c_0}{\sigma_k}, \quad k \in \mathcal{S} \text{ sufficiently large},$$

where $c_0 \stackrel{\rm def}{=} \eta_1 \epsilon^2/(8\|F_0\|^2)$, which summed up over all $k \geq 0$ sufficiently large, larger than some k_0, gives

$$\|F_{k_0}\| - \lim_{k\to\infty} \|F_k\| \geq c_0 \sum_{k\in\mathcal{S}, k=k_0}^{\infty} \frac{1}{\sigma_k},$$

and so, since $\{\|F_k\|\}$ is convergent, (3.21) holds.

Next we estimate the ratio ρ_k in (2.14). For its denominator, note that (3.21) implies

$$1/\sigma_k \to 0, \quad k \in \mathcal{S}, \quad k \to \infty. \tag{3.23}$$

Thus (3.3), (3.12), (3.13) and (3.20) imply, similarly to (3.18), that

$$\|F_k\| - m_k(p_k) \geq \frac{\epsilon^2}{8\|F_0\|^2\sigma_k}, \text{ for all } k \in \mathcal{S} \text{ sufficiently large.} \tag{3.24}$$

It follows from (2.14), (3.16) and (3.24) that

$$1 - \rho_k = \frac{\|F(x_k + p_k)\| - m_k(p_k)}{\|F_k\| - m_k(p_k)} \leq 16\kappa_{\rm J}\|F_0\|^2\epsilon^{-2} \int_0^1 \|J(x_k + tp_k) - J_k\|\, dt, \tag{3.25}$$

for all $k \in \mathcal{S}$ sufficiently large. Now let us argue that the sequence of iterates $\{x_k\}$, $k \geq 0$, is a Cauchy sequence, and hence convergent. The construction of the algorithm, (3.14) and (3.21) imply

$$\|x_{k+l} - x_k\| \leq \sum_{i=k}^{k+l-1} \|x_{i+1} - x_i\| = \sum_{i=k, i\in\mathcal{S}}^{k+l-1} \|p_i\| \leq 2\kappa_{\rm J} \sum_{i=k, i\in\mathcal{S}}^{k+l-1} \frac{1}{\sigma_i} \to 0, \text{ as } k \to \infty,$$

and hence $\{x_k\}$ converges to some $\tilde{x}$. Furthermore, $\|x_k + tp_k - \tilde{x}\| \leq \|x_k - \tilde{x}\| + \|p_k\|$, for all $t \in [0, 1]$. Also, (3.14) and (3.23) imply that $\|p_k\| \to 0$, $k \in \mathcal{S}$, $k \to \infty$. Thus

$$x_k + tp_k \to \tilde{x}, \quad k \in \mathcal{S}, \quad k \to \infty, \text{ for all } t \in [0, 1],$$

and we conclude

$$\|J(x_k + tp_k) - J_k\| \leq \|J(x_k + tp_k) - J(\tilde{x})\| + \|J_k - J(\tilde{x})\| \to 0, k \in \mathcal{S}, k \to \infty, \forall t \in [0, 1],$$

which implies, together with (3.25), that either $\rho_k \geq 1$ or $\rho_k \to 1$, $k \in \mathcal{S}$, $k \to \infty$. Both these conditions imply that k is a very successful iteration for $k \in \mathcal{S}$ sufficiently large, which together with (2.15), gives that $\sigma_{k+1} \leq \sigma_k$, $k \in \mathcal{S}$ sufficiently large. Now, if all k belong to $\mathcal{S}$ for k sufficiently large (i. e., there are no unsuccessful iterations for k sufficiently large), then the latter inequality contradicts (3.23), and so (3.20) cannot hold. Otherwise, recalling that we assumed $\mathcal{S}$ to be infinite (which implies not all iterations can be consecutively unsuccessful for all k sufficiently large), let $\{k_i\}$ denote an (infinite) subsequence of very successful iterations such that $\{k_i - 1\}$ is unsuccessful for all i (since all $k \in \mathcal{S}$ are very successful for all k sufficiently large, without loss of generality, we can ignore successful iterates; also, if such a subsequence $\{k_i\}$ does not exist, then we are in the

previous case of all iterates being very successful for all k sufficiently large). Then, from (2.15), we have $\sigma_{k_i} \leq \gamma_2 \sigma_{k_i-1}$, for all i, which together with (3.23), implies that

$$1/\sigma_{k_i-1} \to 0, \quad i \to \infty. \tag{3.26}$$

It follows that the inequality in (3.24) holds for k replaced by k_i-1, for all i sufficiently large. Hence, (3.25) holds for k_i-1, for all i sufficiently large. Further, (3.14) and (3.26) imply $\|p_{k_i-1}\| \to 0$, $i \to \infty$, and thus, since $x_k \to \tilde{x}$, $k \to \infty$, we have $x_{k_i-1}+tp_{k_i-1} \to \tilde{x}$, $i \to \infty$. As above, we can now conclude that either $\rho_{k_i-1} \geq 1$ or $\rho_{k_i-1} \to 1$, $i \to \infty$. But this implies that $k_i - 1$ is a very successful iteration for all i sufficiently large. This contradicts our assumption that $k_i - 1$ is an unsuccessful iteration for all i. Thus all iterations are very successful for sufficiently large k, a case which we have already addressed. □

Note that Theorems 3.6 and 3.7 only required J_k to be bounded above; the bound (3.11), however, will be needed next.

To be able to show that the whole sequence $\{g_k\}$ converges to zero, we employ the additional assumption below.

Assumption 3.2 *The Jacobian J is uniformly continuous on the sequence of iterates $\{x_k\}$, i. e.,*

$$\|J(x_{t_i}) - J(x_{l_i})\| \to 0, \quad \textit{whenever } \|x_{t_i} - x_{l_i}\| \to 0, \ i \to \infty, \tag{3.27}$$

where $\{x_{t_i}\}$ and $\{x_{l_i}\}$ are subsequences of $\{x_k\}$.

Clearly, Assumption 3.2 is satisfied if J is uniformly continuous on $\mathbb{R}^n$; it is also satisfied in J is Lipschitz continuous on $\mathbb{R}^n$.

The next theorem states that all limit points of the sequence $\{x_k\}$ are stationary points of problem (1.1). It also indicates a case where such limit points solve the problem of finding a root of $F(x) = 0$.

Theorem 3.8 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable and suppose that Assumptions 3.1 and 3.2 hold. Then,*

$$\lim_{k\to\infty} \|g_k\| = 0. \tag{3.28}$$

Furthermore, if $m \leq n$ and there exists a limit point x^* of the sequence $\{x_k\}$ of iterates generated by Algorithm RER such that $F(x^*) = 0$ and $J(x^*)$ is full-rank, then all limit points of $\{x_k\}$ are roots of $F(x) = 0$.

Proof. To prove (3.28), assume that there exists an infinite subsequence $\{t_i\} \subset \mathcal{S}$ such that

$$\|g_{t_i}\| \geq 2\epsilon, \ \text{ for all } \ i, \tag{3.29}$$

for some $\epsilon > 0$. By (3.19), for each t_i there is a first successful iteration $l_i > t_i$ such that $\|g_{l_i}\| < \epsilon$. Thus $\{l_i\} \subseteq \mathcal{S}$ and

$$\|g_k\| \geq \epsilon, \quad t_i \leq k < l_i \quad \text{and} \quad \|g_{l_i}\| < \epsilon. \tag{3.30}$$

Letting $\mathcal{K} = \{k \in \mathcal{S} \mid t_i \leq k < l_i\}$, we observe that this index subset is also infinite. Moreover, (2.14), (3.3), (3.12), (3.13) and (3.7) imply that

$$\|F_k\| - \|F_{k+1}\| \geq \frac{\eta_1\epsilon}{16\|F_0\|} \min\left[\frac{2\|g_k\|}{\sigma_k\|F_k\|}, \frac{4\epsilon}{\kappa_{\text{D}}}\right] \geq \frac{\eta_1\epsilon}{16\|F_0\|} \min\left[\|p_k\|, \frac{4\epsilon}{\kappa_{\text{D}}}\right], \ k \in \mathcal{K}. \tag{3.31}$$

The sequence $\{\|F_k\|\}$ is monotonically non-increasing and bounded below, hence it converges, and so the left-hand side of (3.31) converges to zero, implying that

$$\|p_k\| \to 0, \quad k \in \mathcal{K}, \quad k \to \infty,$$

on the right-hand side of (3.31). Thus (3.31) becomes

$$\|F_k\| - \|F_{k+1}\| \geq \kappa_{\text{g}}\|p_k\|, \ \text{ for all } \ t_i \leq k < l_i, \ k \in \mathcal{S}, \ i \ \text{ sufficiently large}, \tag{3.32}$$

where $\kappa_{\rm g} \stackrel{\rm def}{=} \eta_1\epsilon/(16\|F_0\|)$. Summing up (3.32) and using $x_{k+1} = x_k + p_k$ gives

$$\|F_{t_i}\| - \|F_{l_i}\| \geq \kappa_{\rm g} \sum_{k=t_i, k\in\mathcal{S}}^{l_i-1} \|p_k\| = \kappa_{\rm g} \sum_{k=t_i}^{l_i-1} \|x_{k+1} - x_k\| \geq \|x_{t_i} - x_{l_i}\|, \tag{3.33}$$

for all i sufficiently large. Again using that $\{\|F_k\|\}$ is convergent, the left-hand side of (3.33) converges to zero, and thus

$$\sum_{k=t_i, k\in\mathcal{S}}^{l_i-1} \|p_k\| \to 0 \ \text{ and } \ \|x_{t_i} - x_{l_i}\| \to 0, \ \text{ as } \ i \to \infty. \tag{3.34}$$

We now show that the second limit in (3.34) implies that

$$\|g_{t_i} - g_{l_i}\| \to 0, \ \text{ as } \ i \to \infty. \tag{3.35}$$

We have

$$\|g_{t_i} - g_{l_i}\| \leq \|J_{t_i}^T\| \cdot \|F_{t_i} - F_{l_i}\| + \|F_{l_i}\| \cdot \|J_{t_i} - J_{l_i}\|, \ \text{ for all } \ i.$$

Recalling (3.11), (3.12) and (3.27), (3.35) holds provided $\|F_{t_i} - F_{l_i}\| \to 0$. To see the latter, employ Taylor's theorem and (3.11) to get

$$\|F_{t_i} - F_{l_i}\| \leq \sum_{k=t_i, k\in\mathcal{S}}^{l_i-1} \|F_k - F_{k+1}\| \leq \kappa_{\rm J} \sum_{k=t_i, k\in\mathcal{S}}^{l_i-1} \|p_k\|,$$

whose right-hand side tends to zero due to (3.34). This proves (3.35). We have now reached a contradiction since (3.29) and (3.30) imply $\|g_{t_i} - g_{l_i}\| \geq \|g_{t_i}\| - \|g_{l_i}\| \geq \epsilon$. Hence (3.29) cannot hold and we conclude that (3.28) must hold.

Finally, assume that $\{x_{k_j}\}$ converges to x^* with $J(x^*)$ being full-rank and $m \leq n$. Then (3.28) ensures that $\|F_{k_j}\|$ converges to zero because the singular values of J_{k_j} must remain uniformly bounded away from zero by continuity (for j large enough). We may now conclude our proof by using Lemma 2.2 *ii)*. □

Note that roots of $F(x) = 0$ must be second-order critical points of problem (1.1), and our last theorem may then be interpreted as guaranteeing convergence to such points if the Jacobian remains uniformly full-rank over the iterates. Of course, a guarantee of convergence to second-order points that are not roots of F cannot be given in the framework of the present first-order Gauss-Newton-like method, where the model ignores all second-derivative terms $\nabla_{xx}F_i(x)$.

Note that Theorem 3.8 still holds if we require only J_k to be bounded above and J to be uniformly continuous also on the line segments in between successful iterates.

Theorems 3.6, 3.7 and 3.8 extend results concerning the convergence of trust-region methods given in Moré (1983); see also Thomas (1975).

4 Beyond the Cauchy Point

In practice, more model reduction is sought than that achieved at the Cauchy step, with the objective of improving the speed of convergence. We thus need to investigate the properties of the model further and to describe how a better step can be computed before stating improved convergence results.

4.1 The model and its minimizer

In this section we characterize the minimizer of the model $m_k(p)$.

Lemma 4.1 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable and assume $\|g_k\| \neq 0$.*

i) If the vector p_k^* is the solution of (2.10), then there is a nonnegative λ_k^* such that (p_k^*, λ_k^*) solves

$$(B_k + \lambda I)p = -g_k, \tag{4.1}$$
$$\lambda = \mu_k + 2\sigma_k\phi(p), \tag{4.2}$$

where $\phi(p)$ is given in (2.11).

ii) If $\mu_k > 0$, then there exists a unique solution (p_k^*, λ_k^*) of (4.1) and (4.2), and p_k^* solves (2.10).

iii) If $\mu_k = 0$ and there exists a solution (p_k^*, λ_k^*) of (4.1) and (4.2) with $\lambda_k^* > 0$, then p_k^* solves (2.10). Otherwise, the solution of (2.10) is given by the minimum norm solution of the linear system $B_k p = -g_k$.

Proof. i) If p_k^* solves (2.10) and $\mu_k = 0$ then (4.1) and (4.2) follow from (2.8). On the other hand, if $\mu_k > 0$, then $\phi(p)$ is positive, $m_k(p)$ is differentiable for any p and the gradient $\nabla m_k(p)$ has the form

$$\nabla m_k(p) = \frac{g_k + B_k p + \mu_k p}{\phi(p)} + 2\sigma_k p. \tag{4.3}$$

Thus, $\nabla m_k(p)$ vanishes when (4.1) and (4.2) are satisfied.

As $\nabla m_k(p_k^*) = 0$, it follows that p_k^* solves (4.1) and (4.2) along with

$$\lambda_k^* = \mu_k + 2\sigma_k\phi(p_k^*). \tag{4.4}$$

ii) If $\mu_k > 0$, as $m_k(p)$ is differentiable for any p, it follows that a solution (p_k^*, λ_k^*) of (4.1)-(4.2) satisfies

$$\nabla m_k(p_k^*) = 0.$$

Then, the strict convexity of $m_k(p)$ implies that p_k^* solves (2.10) and (p_k^*, λ_k^*) is the unique solution of (4.1) and (4.2).

iii) Let $\mu_k = 0$. We recall that the first-order conditions (2.6) holds for any p such that $\nu = \|J_k p + F_k\| \neq 0$. Then, if there exist p_k^* and $\lambda_k^* > 0$ satisfying (4.1)-(4.2), p_k^* solves (2.6) with $\nu \neq 0$ and this implies that p_k^* solves (2.10). On the other hand, if all the solutions (p_k^*, λ_k^*) of (4.1)-(4.2) are such that $\lambda_k^* = 0$, then p_k^* satisfies $J_k p_k^* = -F_k$. This implies that there is no solution of the first-order conditions and constraint qualification (LICQ) must fail. Thus $\nu = 0$ in (2.5) and the minimizer p_k^* satisfies $J_k p_k^* = -F_k$. Since $m_k(p) = \sigma_k\|p_k\|^2$ for all p such that $J_k p = -F_k$, we can conclude that the solution p_k^* of (2.10) is the minimum norm solution to $B_k p = -g_k$. □

Next, we let $p(\lambda)$ be the minimum-norm solution of (4.1) for a given $\lambda \geq 0$ and $p_k^* = p(\lambda_k^*)$, the minimum of $m_k(p)$. The following lemma is an intermediate result towards proving an upper bound on the scalar λ_k^*.

Lemma 4.2 *Assume $\|g_k\| \neq 0$ and let $p(\lambda)$ be the minimum norm solution of (4.1) with $\lambda \geq 0$. Assume furthermore that J_k is of rank ℓ and its singular-value decomposition is given by $U_k\Sigma_k V_k^T$ where $\Sigma_k = \mathrm{diag}(\varsigma_1, \ldots, \varsigma_\nu)$, with $\nu = \min(m, n)$. Then, denoting $r = U_k^T F_k$, we have that*

$$\|p(\lambda)\|^2 = \sum_{i=1}^{\ell} \frac{\varsigma_i^2 r_i^2}{(\varsigma_i^2 + \lambda)^2} \quad \text{and} \quad \|F_k + J_k p(\lambda)\|^2 = \sum_{i=1}^{\ell} \frac{\lambda^2 r_i^2}{(\varsigma_i^2 + \lambda)^2} + \sum_{i=\ell+1}^{m} r_i^2. \tag{4.5}$$

Proof. (See also Lemmas 2.2 and 4.1 in Cartis et al., 2009*b*). The defining equation (4.1) and the singular-value decomposition of J_k give that

$$p(\lambda) = -V_k(\Sigma_k^T\Sigma_k + \lambda I)^+\Sigma_k^T r$$

where the superscript $^+$ denotes the Moore-Penrose generalized inverse. Taking the square norm of this expression then yields the first part of (4.5). We also deduce that

$$F_k + J_k p(\lambda) = U_k\big(r - \Sigma_k(\Sigma_k^T\Sigma_k + \lambda I)^+\Sigma_k^T r\big),$$

whose squared norm then gives the second part of (4.5). □

Lemma 4.3 *Assume $\|g_k\| \neq 0$ and let $p(\lambda)$ be the minimum norm solution of (4.1) with $\lambda \geq 0$. Then, the function $\phi(p(\lambda))$ is monotonically increasing in $(\mu_k, +\infty)$ and*

$$\phi(p(\lambda)) \leq \|F_k\|. \tag{4.6}$$

Moreover, if $p_k^* = p(\lambda_k^*)$ is the minimizer of $m_k(p)$, then

$$\lambda_k^* \in [\mu_k, \mu_k + 2\sigma_k \|F_k\|\,]. \tag{4.7}$$

Proof. Using (2.11) and (4.5), we deduce that

$$\phi(p(\lambda)) = \sqrt{\sum_{i=1}^{\ell} \frac{(\lambda^2 + \mu_k \varsigma_i^2)}{(\lambda + \varsigma_i^2)^2}\, r_i^2 + \sum_{i=\ell+1}^{m} r_i^2}. \tag{4.8}$$

and

$$\phi'(p(\lambda)) = \frac{1}{\phi(p(\lambda))} \sum_{i=1}^{\ell} \frac{(\lambda - \mu_k)\varsigma_i^2}{(\lambda + \varsigma_i^2)^3}\, r_i^2.$$

Thus $\phi(p(\lambda))$ is monotonically increasing in $(\mu_k, +\infty)$. Moreover, we deduce from (4.8) that

$$\lim_{\lambda \to \infty} \phi(p(\lambda)) = \sqrt{\sum_{i=1}^{m} r_i^2} = \|F_k\|, \tag{4.9}$$

and we conclude that (4.6) holds. Finally, (4.7) trivially follows from (4.2) and (4.6). □

Note that if $\phi(p(\lambda_k^*)) > 0$ then it follows from (4.2) that $\lambda_k^* > 0$; this is the case whenever $\mu_k > 0$.

4.2 Computing the Trial Step Using Factorizations

We now consider tools for computing an approximate minimizer p_k of the model m_k (see Step 1 of Algorithm RER). In practice, we look for a step p_k satisfying the sufficient decrease condition (2.13) and such that

$$p_k = p(\lambda_k), \qquad (B_k + \lambda_k I)p_k = -g_k, \tag{4.10}$$

where λ_k is an approximation to λ_k^* in (4.4). Our procedure is based on the observation that the optimal scalar λ_k^* solves the so-called secular equation given in (4.2), i.e.,

$$\rho(\lambda) = \lambda - \mu_k - 2\sigma_k \phi(p(\lambda)) = 0. \tag{4.11}$$

In what follows, we suppose that $\rho(\lambda)$ admits a positive root and we explore ways to solve (4.11) by root-finding methods and propose alternative one-dimension nonlinear equations in the variable λ. It is easy to see that $\rho'(\lambda)$ may change sign in $(\mu_k, +\infty)$, while $\zeta(\lambda)$ in the equation

$$\zeta(\lambda) \stackrel{\text{def}}{=} (\lambda - \mu_k)^2 - 4\sigma_k^2 (\phi(p(\lambda)))^2 = 0,$$

is increasing for $\lambda \in [\lambda_k^*, +\infty)$ but is not guaranteed to be convex. Therefore, applying Newton's method to these nonlinear equations safely needs an accurate initial guess. As an alternative to the secular equation (4.11), we consider the problem of finding the positive root of the function $-\rho(\lambda)/\lambda$, i.e.,

$$\psi(\lambda) = 2\sigma_k \frac{\phi(p(\lambda))}{\lambda} + \frac{\mu_k}{\lambda} - 1 = 0. \tag{4.12}$$

The following result establishes desirable properties of this formulation.

Lemma 4.4 *The function $\psi(\lambda)$ is convex and strictly decreasing in $(\mu_k, +\infty)$ and Newton method applied to (4.12) will converge globally and monotonically to the positive root λ_k^* of (4.2) for any initial guess $\lambda^{(0)} \in (\mu_k, \lambda_k^*)$. The secant method has the same properties for any initial guesses $\lambda^{(0)}, \lambda^{(1)}$ such that $\mu_k < \lambda^{(0)} < \lambda^{(1)} \leq \lambda_k^*$.*

Proof. By (4.5) and a result by Boyd and Vandenberghe (2004) [p. 87], we verify that the functions of λ given by

$$\frac{\|F_k + J_k p(\lambda)\|}{\lambda} = \sqrt{\sum_{i=1}^{\ell} \left(\frac{r_i}{\varsigma_i^2 + \lambda}\right)^2 + \sum_{i=\ell+1}^{m} \left(\frac{r_i}{\lambda}\right)^2}$$

and

$$\|p(\lambda)\| = \sqrt{\sum_{i=1}^{\ell} \left(\frac{\varsigma_i r_i}{\varsigma_i^2 + \lambda}\right)^2}$$

are convex and nonnegative on $(\mu_k, +\infty)$. (See also Lemma 4.1 in Cartis et al., 2009*b* for the case where $\mu_k = 0$). Moreover,

$$\left(\|p(\lambda)\|\right)' = -\frac{1}{\|p(\lambda)\|} \sum_{i=1}^{\ell} \frac{\varsigma_i^2 r_i^2}{(\varsigma_i^2 + \lambda)^3} < 0,$$

and hence $\|p(\lambda)\|$ is decreasing. As a consequence, $\sqrt{\mu_k}\|p(\lambda)\|/\lambda$ is also convex and nonnegative. Applying again the cited result by Boyd and Vandenberghe (2004) [p. 87], we deduce that

$$\frac{\phi(p(\lambda))}{\lambda} = \sqrt{\left(\frac{\|F_k + J_k p(\lambda)\|}{\lambda}\right)^2 + \left(\frac{\sqrt{\mu_k}\|p(\lambda)\|}{\lambda}\right)^2}.$$

is convex and the convexity of μ_k/λ finally ensures that of $\psi(\lambda)$.

Now, since $\psi(\lambda) > -1$ for all $\lambda \in (\mu_k, \infty)$ and has a horizontal asymptote at -1 for $\lambda \to \infty$, we deduce that $\psi(\lambda)$ must be strictly decreasing in (μ_k, ∞). Thus λ_k^* (whose existence is assumed) is the unique positive root of (4.12) and the convergence properties of both the Newton method and the secant method applied to (4.12) follow from Lemma A.1 in Cartis et al. (2009*b*). □

In order to apply the Newton method to (4.12), we need

$$\psi'(\lambda) = -\frac{2\sigma_k}{\lambda^2}\phi(p(\lambda)) + \frac{2\sigma_k}{\lambda}\phi'(p(\lambda)) - \frac{\mu_k}{\lambda^2}. \tag{4.13}$$

Differentiating (4.1) with respect to λ we get

$$(B_k + \lambda I)\nabla_\lambda p(\lambda) + p(\lambda) = 0, \tag{4.14}$$

where $\nabla_\lambda p(\lambda)$ is the gradient of $p(\lambda)$. Furthermore, by using (4.1), we obtain that

$$\begin{aligned}
\phi'(p(\lambda)) &= \frac{2(B_k p(\lambda) + g_k)^T \nabla_\lambda p(\lambda) + 2\mu_k p(\lambda)^T \nabla_\lambda p(\lambda)}{2\phi(p(\lambda))} \\
&= \frac{(\mu_k - \lambda) p(\lambda)^T \nabla_\lambda p(\lambda)}{\phi(p(\lambda))} \\
&= \frac{(\lambda - \mu_k) p(\lambda)^T (B_k + \lambda I)^{-1} p(\lambda)}{\phi(p(\lambda))}.
\end{aligned}$$

If the Cholesky factorization $B_k + \lambda I = R^T R$ is available, then $\psi'(\lambda)$ takes the form

$$\psi'(\lambda) = -\frac{2\sigma_k}{\lambda^2}\,\phi(p(\lambda)) + \frac{2\sigma_k(\lambda - \mu_k)\|R^{-T} p(\lambda)\|^2}{\lambda\phi(p(\lambda))} - \frac{\mu_k}{\lambda^2}, \tag{4.15}$$

and we have all the necessary ingredients for computing the Newton method for (4.12). We also observe that, since $\psi(\lambda)$ is convex in $(\mu_k, +\infty)$, a Newton step from an initial $\lambda^{(0)}$ with $\psi(\lambda^{(0)}) < 0$ will underestimate the root λ_k^* and a suitable value between μ_k and λ_k^* can therefore always be found, possibly by bisection. The complete strategy then gives Algorithm 4.1 on the next page.

Algorithm 4.1: Newton method for (4.12) using Cholesky factorization.
An initial $\lambda^{(0)} > \mu_k$ is given. For $\ell = 0, 1, \ldots$ until convergence

1. Compute $B_k + \lambda^{(\ell)} I = R^T R$.

2. Solve $R^T R\, p(\lambda^{(\ell)}) = -g_k$.

3. Solve $R^T\, z(\lambda^{(\ell)}) = p(\lambda^{(\ell)})$.

4. Compute $\psi(\lambda^{(\ell)})$ and $\psi'(\lambda^{(\ell)})$ given in (4.12) and (4.15).

5. Set $\bar{\lambda}^{(\ell)} = \lambda^{(\ell)} - \dfrac{\psi(\lambda^{(\ell)})}{\psi'(\lambda^{(\ell)})}$.

6. If $\bar{\lambda}^{(\ell)} > \mu_k$, set $\lambda^{(\ell+1)} = \bar{\lambda}^{(\ell)}$. Otherwise, set $\lambda^{(\ell+1)} = \frac{1}{2}(\mu_k + \lambda^{(\ell)})$.

Practical versions of the above algorithms should not iterate until convergence to λ_k^* is obtained with high accuracy but return an approximate solution to λ_k^*, producing an approximate step p_k. In practice, Algorithm 4.1 must iterate until

$$\lambda_k \in (\mu_k, \lambda_k^*] \tag{4.16}$$

and condition (2.13) is met. Unfortunately, this requires the computation of p_k^c. Note that, because

$$\begin{aligned} m_k(-\alpha g_k) &= \phi(-\alpha g_k) + \sigma_k \alpha^2 \|g_k\|^2 \\ &= \sqrt{\|F_k\|^2 - 2\alpha\|g_k\|^2 + \alpha^2 g_k^T (B_k + \mu_k I) g_k} + \sigma_k \alpha^2 \|g_k\|^2, \end{aligned}$$

it follows that $p_k^{\mathrm{C}} = -\alpha_k^{\mathrm{C}} g_k$ where $\alpha_k^{\mathrm{C}} \in (0, \|g_k\|^2 / g_k^T (B_k + \mu_k I) g_k)$ is the unique solution to the scalar nonlinear equation

$$2\sigma_k \|g_k\|^2\, \alpha_k \phi(-\alpha_k g_k) = \|g_k\|^2 - \alpha_k g_k^T (B_k + \mu_k I) g_k. \tag{4.17}$$

In practice, α_k^{C} can be computed solving this equation by a root-finding method, at the cost of computing the Hessian-vector product in the last term of (4.17). Note also that (4.16) holds as soon as bisection stops in Algorithm 4.1, and this may be encouraged by choosing $\lambda^{(0)}$ very close to μ_k.

4.2.1 Local Convergence Analysis

We may now complete our convergence results under the condition that an approximate model minimizer is computed.

Assumption 4.1 *The step p_k is computed to satisfy (2.13), (4.10) and (4.16).*

More specifically, we are able to prove that, when $\{x_k\}$ admit a limit point x^* such that $F(x^*) = 0$ and $J(x^*)$ is of full rank, then the iterations must be very successful for k sufficiently large, irrespective of the relative values of m and n. The following assumption is needed for the latter to hold.

Assumption 4.2 *Let $\{x_k\}$ be the sequence generated by the RER Algorithm. Then there exists a constant $\kappa_{\mathrm{S}} > 0$ such that, if $\|x - x_k\| \leq \kappa_{\mathrm{S}}$ and $x \in [x_k, x_k + p_k]$, then*

$$\|J(x) - J(x_k)\| \leq 2\kappa_{\mathrm{L}} \|x - x_k\|, \quad \textit{for all } k. \tag{4.18}$$

Clearly, (4.18) is automatically satisfied when $J(x)$ is globally Lipschitz-continuous over $\mathbb{R}^n$. Note that if Assumption 4.2 replaces Assumption 3.2 in the conditions of Theorem 3.8, the latter still holds. To see this, note that the first limit in (3.27) for the subsequences of interest in the proof of Theorem 3.8 is implied by (4.18) and the first limit in (3.34).

We first prove that the error between the objective function and the model decreases quickly enough with the steplength.

Lemma 4.5 *Assume that $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ is continuously differentiable and that Assumption 4.2 holds. If $\|p_k\| \leq \kappa_{\rm S}$, then*

$$\|F(x_k + p_k)\| - m_k(p_k) \leq (\kappa_{\rm L} - \sigma_k)\|p_k\|^2. \tag{4.19}$$

Proof. The bound (4.19) follows from (3.15) since (4.18) applies for $x = x_k + tp_k$ due to $\|p_k\| \leq \kappa_{\rm S}$. □

We now prove that the iteration must be very successful when σ_k is sufficiently large.

Lemma 4.6 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable and suppose that Assumptions 3.1 and 4.2 hold. Assume that $g_k \neq 0$ and that*

$$\sigma_k \geq \max\left[\kappa_{\rm L}, \frac{2\kappa_{\rm J}}{\kappa_{\rm S}}\right]. \tag{4.20}$$

Then $\rho_k \geq 1$, iteration k is very successful and $\sigma_{k+1} \leq \sigma_k$.

Proof. Note that (3.14) and the second term in the maximum in (4.20) imply that

$$\|p_k\| \leq \frac{2\kappa_{\rm J}}{\sigma_k} \leq \kappa_{\rm S}.$$

We can now apply Lemma 4.5 and deduce that (4.19) holds. But the right-hand side of this inequality is non-positive because of (4.20), and hence, since

$$1 - \rho_k = \frac{\|F(x_k + p_k)\| - m_k(p_k)}{\|F_k\| - m_k(p_k)},$$

and since $\|F_k\| - m_k(p_k) > 0$ by construction (also see (3.3)), we deduce that $\rho_k \geq 1 > \eta_2$. The conclusion then follows from the mechanism of the algorithm. □

The following result then shows that the sequence of parameters $\{\sigma_k\}$ is bounded above.

Lemma 4.7 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable. Suppose that Assumptions 3.1 and 4.2 hold and that $g_k \neq 0$ for all k. Then there exists a constant $\sigma_{\max} > 0$ such that, for all $k \geq 0$,*

$$\sigma_k \leq \sigma_{\max}. \tag{4.21}$$

Proof. Note that for any $k \geq 0$, we know from Lemma 4.6 that (4.20) implies that $\sigma_{k+1} \leq \sigma_k$. Hence, applying the updating rule (2.15), the parameter σ_k cannot be larger than γ_2 times the right-hand side of (4.20). Since the initial value σ_0 may exceed this value, the bound on σ_k takes the form

$$\sigma_k \leq \max\left[\gamma_2\kappa_{\rm L}, \frac{2\gamma_2\kappa_{\rm J}}{\kappa_{\rm S}}, \sigma_0\right] \stackrel{\rm def}{=} \sigma_{\max}.$$

□

The next lemma gives useful asymptotic bounds on quantities of interest.

Lemma 4.8 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable. If x^* is a limit point of the sequence $\{x_k\}$ such that $F(x^*) = 0$ and $J(x^*)$ is of full rank, then, for x_k sufficiently close to x^*,*

$$\|F_k\| \leq \theta\|x_k - x^*\| \tag{4.22}$$

$$\|g_k\| \leq \|J_k\|\,\|F_k\| \leq \theta\|F_k\| \tag{4.23}$$

where $\theta \stackrel{\rm def}{=} 2\max[\|J(x^*)\|, \|J(x^*)^+\|]$. If Assumption 4.1 holds in addition, then

$$\|p_k\| \leq \theta^2\|g_k\| \leq \theta^3\|F_k\| \leq \theta^4\|x_k - x^*\|. \tag{4.24}$$

Moreover, if Assumptions 3.1 and 4.2 hold, then

$$\lambda_k \leq \chi\|F_k\|, \tag{4.25}$$

with $\chi \stackrel{\rm def}{=} \gamma_3 + 2\sigma_{max}$, and iteration k is very successful.

Proof. Since $J(x^*)$ is of full rank, we may choose ϵ to be a positive scalar such that, for any $x_k \in S(x^*, \epsilon)$, J_k is full rank, $\|J_k\| \leq \theta$ and $\|{J_k}^+\| \leq \theta$. Consequently, $\|B_k^+\| \leq \theta^2$, for $x_k \in S(x^*, \epsilon)$. For such an x_k, we see that

$$\|F_k\| \leq \|F(x^*) + \int_0^1 \big[J(x^* + t(x_k - x^*))\big](x_k - x^*)\, dt\| \leq \theta \|x_k - x^*\|.$$

which is (4.22). Using the definition of θ, we then have that

$$\|g_k\| \leq \|J_k\|\, \|F_k\| \leq \theta \|F_k\|,$$

which is (4.23), and, by (4.10) (as implied by Assumption 4.1), that

$$\|p_k\| \leq \|(B_k + \lambda_k I)^+\|\, \|g_k\| \leq \theta^2 \|g_k\| \leq \theta^3 \|F_k\| \leq \theta^4 \|x_k - x^*\|,$$

proving (4.24). Suppose now that Assumptions 3.1 and 4.2 hold, and reduce ϵ if necessary to ensure that

$$\|F_k\| \leq \min\left[\frac{1}{2\sigma_{\max}\theta^2}, \frac{\kappa_{\rm S}}{\theta^3}, \frac{1-\eta_2}{4\kappa_{\rm D}\kappa_{\rm L}\theta^4}\right] \tag{4.26}$$

for all $x_k \in S(x^*, \epsilon)$, where $\kappa_{\rm S}$ is given by Assumption 4.2. Then (4.16) (also implied by Assumption 4.1), (4.7), Lemma 4.7 and Lemma 2.2 give that

$$\lambda_k \leq \lambda_k^* \leq \mu_k + 2\sigma_k \|F_k\| \leq \chi \|F_k\|,$$

which is (4.25). Observing that (4.24) and (4.26) imply that $\|p_k\| \leq \kappa_{\rm S}$, we may also verify that

$$\rho_k = 1 - \frac{\|F(x_k + p_k)\| - m_k(p_k)}{\|F_k\| - m_k(p_k)} \geq 1 - \frac{(\kappa_{\rm L} - \sigma_k)\|p_k\|^2}{\|F_k\| - m_k(p_k)} \geq 1 - \frac{\kappa_{\rm L}\|p_k\|^2}{\|F_k\| - m_k(p_k)},$$

where we used Lemma 4.5 to derive the first inequality. But the bound $\|B_k^+\| \leq \theta^2$ ensures that the minimum singular value of B_k is larger or equal to $1/\theta^2$, and therefore, because of (4.26), that

$$\|B_k + \mu_k I\| \geq \|B_k\| \geq \frac{1}{\theta^2} \geq 2\sigma_{\max}\|F_k\| \geq 2\sigma_k \|F_k\|.$$

As a consequence, the first term in the minimum of (3.3) is the largest and we deduce, using (3.13), that

$$\|F_k\| - m_k(p_k^c) \geq \frac{\|g_k\|^2}{4\|F_k\|\|B_k + \mu_k I\|} \geq \frac{\|g_k\|^2}{4\kappa_{\rm D}\|F_k\|}.$$

Using this inequality, (2.13) and (4.24), we then obtain that

$$\rho_k \geq 1 - \frac{4\kappa_{\rm D}\kappa_{\rm L}\|p_k\|^2}{\|g_k\|^2}\|F_k\| \geq 1 - 4\kappa_{\rm D}\kappa_{\rm L}\theta^4\|F_k\|,$$

i. e., $\rho_k \geq \eta_2$, because of (4.26). □

We now prove that, if $m \geq n$ and there exists a limit point x^* such that $F(x^*) = 0$ and $J(x^*)$ is of full rank, then x^* is an isolated solution of $F(x) = 0$ and the complete sequence $\{x_k\}$ converges to x^*.

Theorem 4.9 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable and suppose that Assumption 4.1 holds and that $m \geq n$. If x^* is a limit point of the sequence $\{x_k\}$ such that $F(x^*) = 0$ and $J(x^*)$ is of full rank, then $\{x_k\}$ converges to x^*.*

Proof. Since $J(x^*)$ is of full rank n, $J(x^*)^+ J(x^*) = I_n$. Thus, by continuity $\|I_n - J(x^*)^+ J(x^* + t(x - x^*))\|$ becomes arbitrarily small in a suitable neighbourhood of x^*. For any x sufficiently close to x^* to ensure that $\|I_n - J(x^*)^+ J(x^* + t(x - x^*))\| \leq 1/2$, the mean value theorem then yields that

$$\begin{aligned} \|J(x^*)^+ F(x)\| &= \|(x - x^*) - \int_0^1 \left(I_n - J(x^*)^+ J(x^* + t(x - x^*))\right)(x - x^*)\, dt\|, \\ &\geq \left(1 - \int_0^1 \frac{1}{2}\, dt\right)\|x - x^*\|. \end{aligned}$$

Using this inequality, we then obtain that, for any such x,

$$\|F(x)\| \geq \frac{\|J(x^*)^+F(x)\|}{\|J(x^*)^+\|} \geq \frac{1}{2\|J(x^*)^+\|}\|x-x^*\|, \tag{4.27}$$

and we conclude that x^* is an isolated limit point of the sequence $\{x_k\}$. Consider now a subsequence $\{x_{k_j}\}$ converging to x^* We may then apply (4.24) for j sufficiently large and deduce that $\|p_{k_j}\|$ converges to zero. Using Lemma 4.10 in Moré and Sorensen (1983), we finally conclude that $\{x_k\}$ converges to x^*. □

In the following result we consider the case where x^* is an isolated solution of the overdetermined ($m \geq n$) system $F(x) = 0$, and show that convergence is fast in this case if one is ready to strengthen somewhat the assumptions on the Jacobian.

Theorem 4.10 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable. Suppose that $m \geq n$ and that Assumptions 3.1, 4.1 and 4.2 hold. Assume that x^* is a limit point of the sequence $\{x_k\}$ such that $F(x^*) = 0$ and $J(x^*)$ is of full rank. Suppose moreover that $J(x)$ is Lipschitz continuous (with constant κ_*) in a neighbourhood of x^* if $m > n$. Then $\{x_k\}$ converges to x^* Q-quadratically.*

Proof. From Theorem 4.9 we know that $\{x_k\}$ converges to x^*. Let ϵ, θ and χ be chosen as in Lemma 4.8 to ensure that (4.18), (4.22)-(4.25) and (4.26) hold, which ensure that iteration k is successful and that $\|p_k\| \leq \kappa_{\rm S}$. By (4.27), we obtain, for $x_k \in S(x^*, \epsilon)$, that

$$\begin{aligned} \|x_k + p_k - x^*\| &\leq 2\theta\|F(x_k+p_k)\| \\ &\leq 2\theta(\|F(x_k+p_k) - F_k - J_kp_k\| + \|F_k + J_kp_k\|) \\ &\leq 2\theta(\kappa_{\rm L}\|p_k\|^2 + \|F_k + J_kp_k\|). \end{aligned} \tag{4.28}$$

Because (4.24) gives that $\|p_k\| \leq \theta^4\|x_k - x^*\|$, we only need to bound $\|F_k + J_kp_k\|$ to prove Q-quadratic convergence.

Let $J_k = U_k\Sigma_kV_k^T = (U_{k,1}, U_{k,2})\Sigma_kV_k^T$ where $U_{k,1} \in \mathbb{R}^{m\times n}$, $U_{k,2} \in \mathbb{R}^{m\times(m-n)}$ and $\Sigma_k = \mathrm{diag}(\varsigma_1, \ldots, \varsigma_n)$. Then we have that

$$U_{k,1}^T = U_{k,1}^T(J_k^T)^+J_k^T$$

because $(J_k^T)^+J_k^T$ is the orthogonal projection onto the range of J_k. As a consequence, we may write that

$$\|U_{k,1}^T(F_k + J_kp_k)\| = \|U_{k,1}^T\big[(J_k^T)^+(B_kp_k + g_k)\big]\|.$$

If we substitute (4.10) in the right-hand side and use (4.22), (4.24) and (4.25), we obtain that

$$\|U_{k,1}^T(F_k + J_kp_k)\| \leq \chi\theta\|F_k\|\,\|p_k\| \leq \chi\theta^6\|x_k - x^*\|^2. \tag{4.29}$$

Moreover, if $m > n$, we verify easily that

$$\|U_{k,2}^T(F_k + J_kp_k)\| = \|U_{k,2}^TF_k\|. \tag{4.30}$$

We now bound this last quantity as in Fan and Yuan (2005). Specifically, let $q_k = -J_k^+F_k$, in which case $J_kq_k = -J_kJ_k^+F_k = -U_{k,1}U_{k,1}^TF_k$. Since q_k minimizes $\|F_k + J_kp\|$, we obtain that

$$\|U_{k,2}^TF_k\| = \|U_{k,2}U_{k,2}^TF_k\| = \|F_k + J_kq_k\| \leq \|F_k + J_k(x_k - x^*)\| \leq \kappa_*\|x_k - x^*\|^2. \tag{4.31}$$

Combining together the triangle inequality and (4.29)-(4.31), we find that

$$\|F_k + J_kp_k\| \leq \|U_{k,1}^T(F_k + J_kp_k)\| + \|U_{k,2}^T(F_k + J_kp_k)\| \leq (\chi\theta^6 + \kappa_*)\|x_k - x^*\|^2,$$

which concludes the proof in view of (4.24) and (4.28). □

The final theorem in this section studies the local convergence for underdetermined systems, that is when $m \leq n$. In this case, if x^* is a limit point of the sequence $\{x_k\}$ and $J(x^*)$ is of full rank, then $F(x^*) = 0$, but in general x^* is not an isolated solution of $F(x) = 0$.

Theorem 4.11 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable. Suppose that $m \leq n$ and that Assumptions 3.1, 4.1 and 4.2 hold. If x^* is a limit point of the sequence $\{x_k\}$ and $J(x^*)$ is of full rank (and thus $F(x^*) = 0$), then $\{x_k\}$ converges to x^* Q-quadratically.*

Proof. Again let ϵ and θ and χ be chosen as in Lemma 4.8 to ensure that (4.18), (4.22)-(4.25) and (4.26) hold, which ensure that iteration k is successful and that $\|p_k\| \leq \kappa_{\rm S}$. If necessary, reduce ϵ further to ensure that

$$\theta^3\epsilon(\chi\theta^2 + \kappa_{\rm L}\theta^4) \leq \frac{1}{2}. \tag{4.32}$$

Let ψ be a positive scalar such that

$$\psi \leq \frac{\epsilon}{1 + 2\theta^4} \tag{4.33}$$

and assume $x_k \in S(x^*, \psi)$ for some $k \geq k_0$, in which case (4.24) immediately gives that

$$\|p_k\| \leq \theta^4\psi. \tag{4.34}$$

To ensure that the sequence $\{x_k\}$ is convergent, we need to show that it is a Cauchy sequence. We achieve this objective by proving, by recurrence, that, if $x_k \in S(x^*, \psi)$, then

$$x_{k+\ell} \in S(x^*, \epsilon) \quad \text{and} \quad \|p_{k+\ell+1}\| \leq \frac{1}{2}\|p_{k+\ell}\| \tag{4.35}$$

for all $\ell \geq 0$. Consider the case $\ell = 0$ first. Since $(J_k^T)^+ J_k^T = I_m$, we deduce from (4.10) and (4.25) that

$$\|F_k + J_k p_k\| \leq \|(J_k^T)^+\| \, \|B_k p_k + g_k\| \leq \theta \, \|B_k p_k + g_k\| \leq \chi\theta \, \|F_k\| \, \|p_k\|. \tag{4.36}$$

Thus using successively the triangle inequality, (4.34) and (4.33), we verify that

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\| + \|p_k\| \leq (\psi + \theta^4\psi) \leq \epsilon \tag{4.37}$$

i.e. $x_{k+1} \in S(x^*, \epsilon)$. Then, (4.24) yields that, for any such iterate,

$$\|p_{k+1}\| \leq \theta^3\|F_{k+1}\| = \theta^3\|F(x_k + p_k)\|, \tag{4.38}$$

since iteration k is successful. As a consequence, we see that

$$\begin{aligned} \|p_{k+1}\| &\leq \theta^3\|F_k + J_k p_k + (F(x_k + p_k) - F_k - J_k p_k)\| \\ &\leq \theta^3(\chi\theta\|F_k\| + \kappa_{\rm L}\|p_k\|) \, \|p_k\|, \end{aligned} \tag{4.39}$$

where we used (4.18) and (4.36). Using now (4.22), (4.24), (4.34) and the bound $\psi \leq \epsilon$ implied by (4.33), we have that, whenever $x_{k+1} \in S(x^*, \epsilon)$,

$$\|p_{k+1}\| \leq \theta^3\epsilon(\chi\theta^2 + \kappa_{\rm L}\theta^4) \, \|p_k\| \leq \frac{1}{2}\|p_k\|, \tag{4.40}$$

where the last inequality results from (4.32). Hence (4.35) holds for $\ell = 0$. Assume now that (4.35) holds for iterations $k + j$, $j = 0, \ldots, \ell - 1$. Using this assumption, the convergence of the geometric progression of factor $\frac{1}{2}$ and (4.34), we obtain that

$$\|x_{k+\ell} - x^*\| \leq \psi + \sum_{j=0}^{\ell-1} \|p_{k+j}\| \leq \psi + \sum_{j=0}^{\ell-1} \left(\frac{1}{2}\right)^j \|p_k\| \leq \psi + 2\,\|p_k\| \leq \psi + 2\theta^4\psi,$$

and hence $x_{k+\ell} \in S(x^*, \epsilon)$ because of (4.33). As for $\ell = 0$, we then use (4.34) and the successful nature of iteration $k + \ell$ (itself implied by the inclusion $x_{k+\ell} \in S(x^*, \epsilon)$) to deduce that

$$\|p_{k+\ell+1}\| \leq \theta^3\|F(x_{k+\ell} + p_{k+\ell})\| \leq \theta^3(\chi\theta\|F_{k+\ell}\| + \kappa_{\rm L}\|p_{k+\ell}\|) \, \|p_{k+\ell}\|.$$

But, by (4.24) and our recurrence assumption,

$$\|p_{k+\ell}\| \leq \theta^4\|x_{k+\ell} - x^*\| \leq \theta^4\epsilon$$

and thus, using (4.22), we deduce that

$$\|p_{k+\ell+1}\| \leq \theta^3 \epsilon(\chi\theta^2 + \kappa_{\rm L}\theta^4)\,\|p_{k+\ell}\| \leq \frac{1}{2}\,\|p_{k+\ell}\|,$$

which concludes our proof of (4.35). We may thus conclude from (4.34) and (4.35) that, if $x_k \in S(x^*, \psi)$, the successive steps after k satisfy the inequalities

$$\|p_k\| \leq \theta^4\psi \quad \text{and} \quad \|p_{k+\ell+1}\| \leq \frac{1}{2}\,\|p_{k+\ell}\|, \;\; \ell = 0, 1, \ldots. \tag{4.41}$$

This in turn implies that $\{x_k\}$ is a Cauchy sequence and, as a consequence, that $\{x_k\}$ converges. Since x^* is a limit point of the sequence, we deduce that $\lim_{k\to\infty} x_k = x^*$.

We finally show the Q-quadratic convergence rate by noting that, because of (4.41),

$$\|x_{k+1} - x^*\| \leq \sum_{j=k+1}^{\infty} \|p_j\| \leq \sum_{j=0}^{\infty} \left(\frac{1}{2}\right)^j \|p_{k+1}\| = 2\,\|p_{k+1}\|.$$

But (4.39), (4.22) and (4.24) together imply that

$$\|p_{k+1}\| \leq \theta^3(\chi\theta^2\|x_k - x^*\| + \kappa_{\rm L}\theta^4\|x_k - x^*\|)\,\theta^4\|x_k - x^*\| = \theta^9(\chi + \kappa_{\rm L}\theta^2)\|x_k - x^*\|^2.$$

Combining these last two inequalities then completes the proof. □

4.3 Computing the Trial Step in a Subspace

If the factorization of B_k is unavailable because of cost or memory limitations, an alternative approach to compute a trial step consists in minimizing $m_k(p)$ over a sequence of nested Krylov subspaces (see Cartis et al., 2009*a*, 2009*b*). In each subspace a secular equation is solved and the dimension of the subspace is progressively increased until the gradient of the model is sufficiently small. Suitable strategies are then adopted to recover an approximate solution at a low computational cost. The requirement to satisfy the Cauchy condition (2.13) is then automatically fulfilled by including g_k in each subspace, which is obtained by initializing the Krylov sequence with that vector. Note however that (4.10) no longer holds in this framework, making the analysis of Section 4.2.1 inapplicable.

Our development of this approach parallels that of Cartis et al. (2009*b*), but is briefly restated here because it now includes the case where $\mu_k > 0$ which was not considered in this reference. Applying Golub-Kahan bi-diagonalization algorithm at iteration k, we get matrices $W_j \in \mathbb{R}^{m\times j}$, $Q_j \in \mathbb{R}^{n\times j}$ and $C_j \in \mathbb{R}^{(j+1)\times j}$ such that

$$J_k Q_j = W_{j+1} C_j, \tag{4.42}$$

where $Q_j^T Q_j = I$, $W_j^T W_j = I$ and C_j is bidiagonal (note that this technique uses J_k only, not $J_k^T J_k$). Then a sequence of minimizers of $m_k(Q_j\, y)$ in the expanding subspaces $p = Q_j\, y$, $j = 1, 2, \ldots$, are sought. In fact, the solution to (2.10) reduces to

$$\min_{y\in\mathbb{R}^j} m_k(Q_j\, y) = \sqrt{\|C_j y - \beta_1 e_1\|^2 + \mu_k\|y\|^2} + \sigma_k\|y\|^2, \tag{4.43}$$

with $\beta_1 = \|F_k\|$. The minimizer y_j to (4.43) is the vector $y_j = y_j(\lambda_j)$ satisfying

$$(C_j^T C_j + \lambda I)y = \beta_1 C_j^T e_1, \tag{4.44}$$

$$\lambda = \mu_k + 2\sigma_k\sqrt{\|C_j y - \beta_1 e_1\|^2 + \mu_k\|y\|^2}. \tag{4.45}$$

Algorithm 4.1 may be used to solve (4.45) accurately. Nested subspaces are constructed for increasing j until $p_j = Q_j\, y_j$ satisfies

$$\|\nabla m_k(p(\lambda_j))\| \leq \omega_k \tag{4.46}$$

for an iteration-dependent tolerance $\omega_k > 0$, at which point the step p_k is then taken as the last computed p_j. We now study properties of the sequence $\{x_k\}$ generated using this approach.

Lemma 4.12 *Let x^* be such that $F(x^*) = 0$ and $J(x^*)$ is of full rank. Suppose moreover that $J(x)$ is Lipschitz continuous (with constant κ_*) in a neighbourhood of x^* if $m > n$. Then there exist constants χ, ϵ and θ such that, if $x_k \in S(x^*, \epsilon)$ and $\omega_k \leq 1/(2\theta)$, we have that*

$$\|F_k + J_k p_k\| \leq \frac{\theta}{1-\theta\omega_k}(\omega_k\sqrt{\mu_k} + \lambda_k)\|p_k\|, \quad \textit{if} \quad m \leq n; \tag{4.47}$$

$$\|F_k + J_k p_k\| \leq \frac{\theta}{1-\theta\omega_k}\left[(\omega_k\sqrt{\mu_k} + \lambda_k)\|p_k\| + \kappa_*\|x_k - x^*\|^2\right], \ \textit{if} \quad m > n; \tag{4.48}$$

$$\|p_k\| \leq \|(B_k + \lambda_k I)^+\|\,\|g_k\| \leq \theta^2\,\|g_k\| \leq \theta^4\|x_k - x^*\|. \tag{4.49}$$

Moreover, if Assumptions 3.1 and 4.2 hold, then

$$\lambda_k \leq \chi\|F_k\|, \tag{4.50}$$

with $\chi \stackrel{\rm def}{=} \gamma_3 + 2\sigma_{\max}$, and iteration k is very successful.

Proof. As above, let θ and ϵ be positive scalars such that for any $x_k \in S(x^*, \epsilon)$, J_k is of full rank, $\|J_k\| \leq \theta$, $\|{J_k}^+\| \leq \theta$. By (4.3) and (4.46) we have

$$\|(B_k + \lambda_k I)p_k + g_k\| \leq \omega_k\phi(p_k)$$

whenever $\|F_k + J_k p_k\| > 0$, since this last inequality implies that $\phi(p_k) > 0$. Let $J_k = U_k\Sigma_k V_k^T = (U_{k,1}, U_{k,2})\Sigma_k V_k^T$ where $U_{k,1} \in \mathbb{R}^{m\times\nu}$, $U_{k,2} \in \mathbb{R}^{m\times(m-\nu)}$, $\Sigma_k = \mathrm{diag}(\varsigma_1, \ldots, \varsigma_\nu)$ and $\nu = \min(m, n)$. We may then derive that

$$\begin{aligned}\|U_{k,1}^T(F_k + J_k p_k)\| &\leq \|(J_k^T)^+(B_k p_k + g_k)\| \\ &\leq \theta\big(\|(B_k + \lambda_k I)p_k + g_k\| + \lambda_k\|p_k\|\big) \\ &\leq \theta\big(\omega_k\phi(p_k) + \lambda_k\|p_k\|\big),\end{aligned}$$

and this inequality also obviously holds if $\|F_k + J_k p_k\| = 0$. Then, using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we deduce that

$$\|U_{k,1}^T(F_k + J_k p_k)\| \leq \theta\big(\omega_k\|F_k + J_k p_k\| + \omega_k\sqrt{\mu_k}\|p_k\| + \lambda_k\|p_k\|\big).$$

If $m \leq n$, since $U_{k,1}$ belongs to $\mathbb{R}^{m\times m}$, $\|U_{k,1}^T(F_k + J_k p_k)\| = \|F_k + J_k p_k\|$. Moreover, we note that (4.31) remains valid whenever $m > n$. Thus, if $\omega_k < 1/(2\theta)$, we then obtain (4.47) and (4.48). Regarding (4.49), note that, by (4.44),

$$\|p_k\| = \|p_j\| = \|y_j\| \leq \|(C_j^T C_j + \lambda I)^{-1}\|\,\|\beta_1 C_j^T e_1\| \tag{4.51}$$

and also that (4.42) gives the relations

$$C_j^T C_j + \lambda I = Q_j^T({J_k}^T J_k + \lambda I)Q_j \quad \text{and} \quad \beta_1 C_j^T e_1 = -Q_j^T g_k. \tag{4.52}$$

Now observe that the columns of Q_j are by construction orthogonal to the nullspace of J_k and hence the eigenvalues of $Q_j^T({J_k}^T J_k + \lambda I)Q_j$ are interlaced between the nonzero eigenvalues of $({J_k}^T J_k + \lambda I)$; they are therefore bounded above and below by the largest and smallest nonzero eigenvalues of this last matrix. Using (4.51) with (4.52), the definition of $B_k = J_k^T J_k$ and this last observation, we then deduce that (4.49) holds. Finally, suppose that Assumptions 3.1, 4.1 and 4.2 hold and consider (4.50). By (4.45) and (4.52) we have that

$$\begin{aligned}\lambda_k &= \mu_k + 2\sigma_k\sqrt{\|C_j y_j - \beta_1 e_1\|^2 + \mu_k\|y_j\|^2} \\ &= \mu_k + 2\sigma_k\sqrt{y_j^T C_j^T C_j y_j + 2y_j^T Q_j^T g_k + \|F_k\|^2 + \mu_k\|y_j\|^2}\end{aligned}$$

where y_j satisfies (4.44). Using the singular value decomposition of C_j, we deduce, as in Lemma 4.3, that

$$\sqrt{\|C_j y_j(\lambda) - \beta_1 e_1\|^2 + \mu_k\|y_j(\lambda)\|^2}$$

is monotonically increasing as a function of λ and converges to $\|F_k\|$ for λ going to infinity, which then, together with the upper bound (4.21) on σ_k, yields (4.50). The proof of the very successful nature of iteration k is identical to that given in Lemma 4.8. □

Following the lines of Theorem 4.10, we may now obtain the local convergence results corresponding to Theorems 4.10 and 4.11 for the case where the step is computed in a subspace.

Theorem 4.13 *Assume that $m \geq n$ and that x^* is a limit point of the sequence $\{x_k\}$ such that $F(x^*) = 0$ and $J(x^*)$ is nonsingular. Assume also that Assumptions 3.1 and 4.2 hold. Suppose moreover that $J(x)$ is Lipschitz continuous (with constant κ_*) in a neighbourhood of x^* if $m > n$. Then, if the scalar ω_k in (4.46) is such that $\omega_k \leq \kappa_\omega \sqrt{\|F_k\|}$ for some $\kappa_\omega > 0$, the sequence $\{x_k\}$ converges to x^* Q-quadratically.*

Proof. The proof follows the same steps as those of Theorem 4.10, taking into account that our assumptions on μ_k and ω_k, the convergence of $\|F_k\|$ to zero, (4.47) and (4.48) together yield that, for k large enough,

$$\begin{aligned} \|F_k + J_k p_k\| &\leq 2\theta\big[\,(\kappa_\omega\sqrt{\gamma_3}\,\|F_k\| + \lambda_k)\|p_k\| + \kappa_*\|x_k - x^*\|^2\big] \\ &\leq 2\theta\big[\,(\kappa_\omega\sqrt{\gamma_3}\,\theta\|x_k - x^*\| + \chi\theta\|x_k - x^*\|)\,\theta^2\|g_k\| + \kappa_*\|x_k - x^*\|^2\big] \\ &\leq 2\theta\big[\theta^5(\kappa_\omega\sqrt{\gamma_3} + \chi) + \kappa_*\big]\|x_k - x^*\|^2, \end{aligned}$$

where we have used (4.22), (4.23), (4.49) and (4.50). Inserting this bound in (4.28) then ensures the desired rate of convergence. □

Theorem 4.14 *Let $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ be continuously differentiable. Suppose that $m \leq n$ and that Assumptions 3.1 and 4.2 hold. If x^* is a limit point of the sequence $\{x_k\}$ and $J(x^*)$ is of full rank (and thus $F(x^*) = 0$), then, if the scalar ω_k in (4.46) is such that $\omega_k \leq \kappa_\omega \sqrt{\|F_k\|}$ for some $\kappa_\omega > 0$, the sequence $\{x_k\}$ converges to x^* Q-quadratically.*

Proof. The proof parallels that of Theorem 4.11, where we first replace (4.36) by the inequality

$$\|F_k + J_k p_k\| \leq 2\theta\,(\kappa_\omega\sqrt{\gamma_3} + \chi)\,\|F_k\|\,\|p_k\|,$$

which follows, for k sufficiently large, from (4.47), our assumptions on μ_k and ω_k, the convergence of $\|F_k\|$ to zero and (4.50). After deriving (4.34) and (4.37), (4.38) now results from (4.49) and the successful nature of iteration k. The rest of the proof then follows that of Theorem 4.11 step by step, with (4.49) replacing (4.24) and (4.50) replacing (4.25). □

5 Numerical results

In this section we present some numerical results obtained when solving nonlinear least-squares problems from the CUTEr collection with Algorithm RER. All runs were performed using a Fortran 95 code on a Intel Xeon (TM) 3.4 Ghz, 1GB RAM. A key role in the performance of the Algorithm RER is played by the regularisation parameter σ_k in Step 4. Here, $\sigma_0 = 1$ and on very successful iterations we set $\sigma_{k+1} = \max(\min(\sigma_k, \|g_k\|), \epsilon_M)$, where $\epsilon_M \simeq 10^{-16}$ is the relative machine precision. For other successful iterations σ_k is left unchanged, while in case of unsuccessful iterations σ_k is doubled.

The approximate minimizer p_k in Step 1 of the RER algorithm was computed minimizing $m_k(p)$ over a sequence of nested Krylov subspaces. This computation is carried out using the module L2RT (Cartis et al., 2009*b*) from the GALAHAD library (see Gould, Orban and Toint, 2003). The approximate minimizer p_k satisfies the accuracy requirement (4.46) with

$$\omega_k = \min(0.1, \|\nabla m_k(0)\|^{1/2})\|\nabla m_k(0)\|. \tag{5.1}$$

First we run the RER algorithm with $\mu_0 = 0$. Then, the tests have been repeated with $\mu_0 = 10^{-4}$ and

$$\mu_{k+1} = \begin{cases} \max[\min(\mu_k, 10^{-3}\|F_{k+1}\|), \epsilon_M] & \text{if } \rho_k \geq \eta_1, \\ \mu_k & \text{otherwise,} \end{cases}$$

Test problem	n	m	NNLSTR with ST point		NNLSTR beyond ST point		RER with $\mu_0 = 0$		RER with $\mu_0 = 10^{-4}$	
			Oiter	Iiter	Oiter	Iiter	Oiter	Iiter	Oiter	Iiter
ARGTRIG	200	200	9	931	9	931	9	875	9	866
ARWHDNE	500	998	321	322	232	318	230	368	197	293
BROYDNBD	1000	1000	18	76	18	80	13	91	13	91
INTEGREQ	102	100	4	8	4	8	4	7	4	7
YATP1SQ	2600	2600	40	46	28	40	20	30	21	32

Table 1: The columns contain the name of the problem, its dimensions, the number of outer (Oiter) and inner (Iiter) iterations performed.

which corresponds to the choice $\gamma_3 = 10^{-3}$ in (2.16).

The RER method is compared with NNLSTR, a trust-region method which has been implemented following the standard scheme (see Conn et al., 2000, Alg. 6.1.1). In NNLSTR, the approximate solution p_k of the trust-region problem is computed using the module LSTR (Cartis et al., 2009*b*) from the GALAHAD library with the stopping criterion (4.46) and the tolerance ω_k defined by (5.1); note that the LSTR technique is a first-order approach. When the solution of the trust-region subproblem lies on the trust-region boundary, the Steihaug-Toint point is computed (Steihaug, 1983, Toint, 1981). We also assessed whether there is any gain in iterating beyond the Steihaug-Toint point in the solution of the trust-region subproblem. On very successful iterations, the trust-region radius Δ_{k+1} is set to $\max\{\Delta_k, 2\|p_k\|\}$, while it is left unchanged on successful iterations, and it is halved otherwise. The initial trust-region radius is set to 1. The two variants of the NNLSTR code (with and without exploration beyond the Steihaug-Toint point) can therefore be considered as a modern trust-region codes for unconstrained optimization.

The RER and trust-region algorithms are stopped whenever the criterion

$$\|F_k\| \leq \max(10^{-6}, 10^{-12}\|F_0\|) \quad \text{or} \quad \|g_k\| \leq \max(10^{-6}, 10^{-12}\|g_0\|)$$

is met.

In Table 1 we give the results obtained on the following five CUTEr test examples: the three square nonlinear systems ARGTRIG, BROYDNBD, YATP1SQ, the underdetermined problem INTEGREQ and the overdetermined test ARWHDNE. The number of outer iterations performed by RER with positive μ_0 is the same as in the case $\mu_0 = 0$, except for problems YATP1SQ and ARWHDNE. These exceptions point out the advantage that can be gained sometimes by employing a positive regularisation μ_0.

The convergence history plot for problem YATP1SQ in Figure 1 illustrates the fast asymptotic rate predicted by our theoretical results.

The numerical results we obtained are encouraging, as the RER Algorithm requires a low number of outer iterations except for problem ARWHDNE. This is important in practice because each outer iteration involves one evaluation of $F(x)$ (and possibly of its Jacobian), the cost of which often dominates the whole solution process. Reducing the number of outer iterations thus often results in significant computational savings. The slow convergence on ARWHDNE may be ascribed to the fact that the methods converge to a nonzero residual solution with a final value of $\|F\| \simeq 0.12 \times 10^2$. This illustrates that the first-order Gauss-Newton-like model employed by the algorithms discussed here may be not appropriate to handle this situation. Furthermore, our implementations of Newton-like cubic overestimation or trust-region schemes on this problem terminate in 6–7 outer iterations, implying that significant gain can be made from using second-order information when solving such nonzero-residual problems.

6 Conclusions and perspectives

We have described a variant of the Gauss-Newton algorithm for nonlinear least-squares, inspired by ideas of Nesterov (2007). The new variant includes the provision for approximate solutions of the subproblem and also features an additional regularisation which might be advantageous in practice. We have developed a complete global convergence theory for this new variant, and have also shown that convergence to zero residual solution is quadratic under reasonable assumptions.

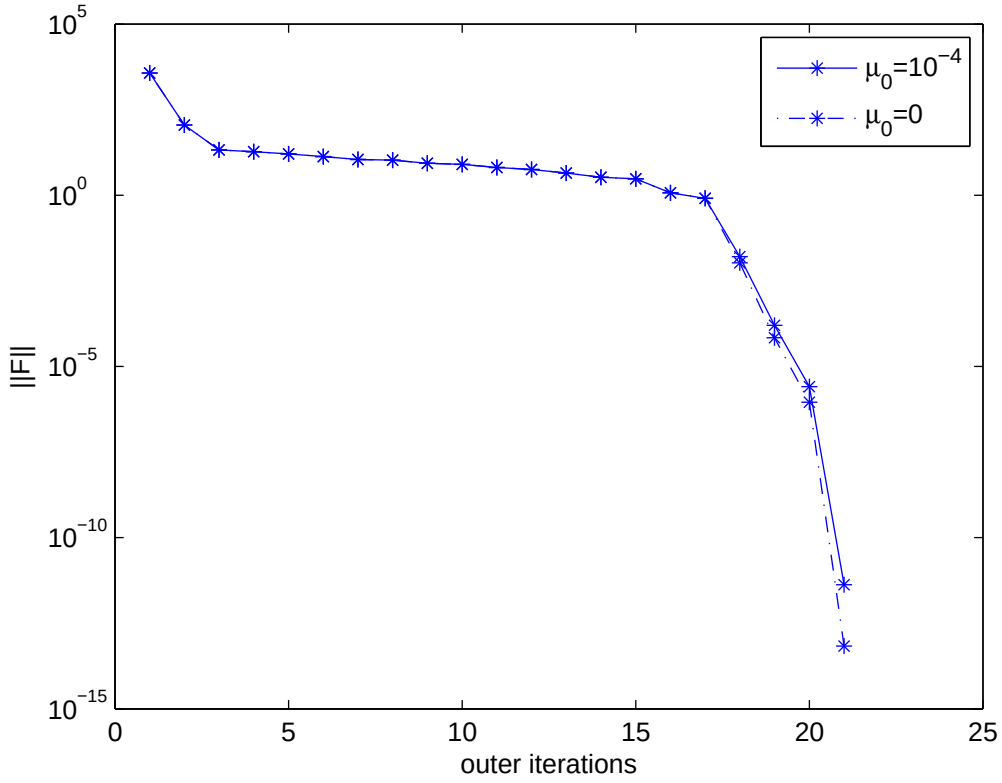

Figure 1: Convergence history for problem YATP1SQ.

Several extensions of the present work are possible. It seems in particular desirable to develop a variant of the full Newton's method (as opposed to the Gauss-Newton algorithm) which would be based on the regularized Euclidean residual and yet could handle negative curvature and nonzero residuals. However, this extension does not seem obvious at this stage. It is also of direct interest to investigate whether, as is the case for the adaptive cubic overestimation (ACO) method, the complexity results obtained by Nesterov could be extended to the case where the subproblems are solved inexactly.

Finally, the true potential of the new variant has yet to be compared to competing techniques in extensive numerical tests, which are currently under way and will be reported on separately.

Acknowledgements

The authors are indebted to Margherita Porcelli for providing the numerical examples presented in the paper.

References

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004.

R. H. Byrd, R. B. Schnabel, and G. A. Shultz. A trust region algorithm for nonlinearly constrained optimization. *SIAM Journal on Numerical Analysis*, **24**, 1152–1170, 1987.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming, Series A*, 2009*a*. DOI: 10.1007/s10107-009-0286-5, 51 pages.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Trust-region and other regularisation of linear least-squares problems. *BIT*, **49**(1), 21–53, 2009*b*.

M. R. Celis. A trust region strategy for nonlinear equality constrained optimization. Technical Report TR85-4, Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA, 1985.

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.

J. Fan and Y. Yuan. On the quadratic convergence of the Levenberg-Marquardt method without nonsingularity assumption. *Computing*, **74**, 23–39, 2005.

N. I. M. Gould, D. Orban, and Ph. L. Toint. GALAHAD—a library of thread-safe Fortran 90 packages for large-scale nonlinear optimization. *ACM Transactions on Mathematical Software*, **29**(4), 353–372, 2003.

A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom, 1981.

J. J. Moré. Recent developments in algorithms and software for trust region methods. *in* A. Bachem, M. Grötschel and B. Korte, eds, 'Mathematical Programming: The State of the Art', pp. 258–287, Heidelberg, Berlin, New York, 1983. Springer Verlag.

J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, **4**(3), 553–572, 1983.

Yu. Nesterov. Modified Gauss-Newton scheme with worst-case guarantees for global performance. *Optimization Methods and Software*, **22**(3), 469–483, 2007.

Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, **108**(1), 177–205, 2006.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Series in Operations Research. Springer Verlag, Heidelberg, Berlin, New York, 1999.

E. O. Omojokun. *Trust region algorithms for optimization with nonlinear equality and inequality constraints*. PhD thesis, University of Colorado, Boulder, Colorado, USA, 1989.

M. J. D. Powell and Y. Yuan. A trust region algorithm for equality constrained optimization. *Mathematical Programming*, **49**(2), 189–213, 1990.

R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, **1**, 97–116, 1976.

T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, **20**(3), 626–637, 1983.

S. Thomas. *Sequential estimation techniques for quasi-Newton algorithms*. PhD thesis, Cornell University, Ithaca, New York, USA, 1975.

Ph. L. Toint. Towards an efficient sparsity exploiting Newton method for minimization. *in* I. S. Duff, ed., 'Sparse Matrices and Their Uses', pp. 57–88, London, 1981. Academic Press.

A. Vardi. A trust region algorithm for equality constrained minimization: convergence properties and implementation. *SIAM Journal on Numerical Analysis*, **22**(3), 575–591, 1985.

M. Weiser, P. Deuflhard, and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, **22**(3), 413–431, 2007.