The University of Reading

Approximate Gauss-Newton methods for nonlinear least squares problems

S. Gratton[1], A.S. Lawless[2] and N.K. Nichols[2]

NUMERICAL ANALYSIS REPORT 9/04

[1] *CERFACS*
42 Avenue Gustave Coriolis
31057 Toulouse
CEDEX,France

[2] *Department of Mathematics*
The University of Reading
Box 220, Whiteknights
Reading, Berkshire RG6 6AX, UK

Department of Mathematics

Abstract

The Gauss-Newton algorithm is an iterative method regularly used for solving nonlinear least squares problems. It is particularly well-suited to the treatment of very large scale variational data assimilation problems that arise in atmosphere and ocean forecasting. The procedure consists of a sequence of linear least squares approximations to the nonlinear problem, each of which is solved by an 'inner' direct or iterative process. In comparison with Newton's method and its variants, the algorithm is attractive because it does not require the evaluation of second-order derivatives in the Hessian of the objective function. In practice the exact Gauss-Newton method is too expensive to apply operationally in meteorological forecasting and various approximations are made in order to reduce computational costs and to solve the problems in real time. Here we investigate the effects on the convergence of the Gauss-Newton method of two types of approximation used commonly in data assimilation. Firstly, we examine 'truncated' Gauss-Newton methods where the 'inner' linear least squares problem is not solved exactly, and secondly, we examine 'perturbed' Gauss-Newton methods where the true linearized 'inner' problem is approximated by a simplified, or perturbed, linear least squares problem. We give conditions ensuring that the truncated and perturbed Gauss-Newton methods converge and also derive rates of convergence for the iterations. The results are illustrated by a simple numerical example.

Keywords Nonlinear least squares problems; approximate Gauss-Newton methods; variational data assimilation

1 Introduction

The Gauss-Newton method is a well-known iterative technique used regularly for solving the nonlinear least squares problem (NLSP)

$$\min_x \phi(x) = \frac{1}{2} \|f(x)\|_2^2 , \tag{1}$$

where x is an n-dimensional real vector and f is an m-dimensional real vector function of x [9].

Problems of this form arise commonly from applications in optimal control and filtering and in data fitting. As a simple example, if we are given m observed data (t_i, y_i), that we wish to fit with a model $S(x,t)$, determined by a vector x of n parameters, and if we define the i-th component of $f(x)$ to be $f_i(x) = S(x,t_i) - y_i$, then the solution to the NLSP (1) gives the best model fit to the data in the sense of the minimum sum of square errors. The choice of norm is often justified by statistical considerations [12]. Recently, very large inverse problems of this type arising in *data assimilation* for numerical weather, ocean and climate prediction and for other applications in the environmental sciences have attracted considerable attention [3, 8].

The Gauss-Newton method consists in solving a sequence of linearized least squares approximations to the nonlinear (NLSP) problem, each of which can be solved efficiently by an 'inner' direct or iterative process. In comparison with Newton's method and its variants, the Gauss-Newton method for solving the NLSP is attractive because it does not require computation or estimation of the second derivatives of the function $f(x)$ and hence is numerically more efficient.

In practice, particularly for the very large problems arising in data assimilation, approximations are made within the Gauss-Newton process in order to reduce computational costs. The effects of these approximations on the convergence of the method need to be understood. Here we investigate the effects of two types of approximation used commonly in data assimilation: firstly, we examine 'truncated' Gauss-Newton methods where the 'inner' linear least squares problem is not solved exactly, and secondly, we examine 'perturbed' Gauss-Newton methods where the true linearized 'inner' problem is approximated by a simplified, or perturbed, linear least squares problem. We give conditions ensuring that the truncated and perturbed Gauss-Newton methods converge and also derive rates of convergence for the iterations.

In the next section we state the problem in detail, together with our assumptions, and define the Gauss-Newton algorithm. We also present some basic theory for the exact method. The truncated and perturbed algorithms that are to be investigated are then defined. In the following sections theoretical convergence results are established for the approximate Gauss-Newton methods. Two different approaches are used to derive the theory. Firstly, we apply extensions of the results of [9, 4] for inexact Newton methods to the approximate Gauss-Newton methods in order to obtain general convergence theorems. We then derive more restricted results using the approach of [5]. The restricted results also provide estimates for the rates of convergence of the methods. Conditions for linear, super-linear and quadratic convergence are noted. Finally, in the remaining sections, numerical results demonstrating and validating the theory are presented and the conclusions are summarized.

2 Gauss-Newton Method

We begin by introducing the Gauss-Newton method and reviewing briefly some results on the convergence of the method. We then define the truncated and perturbed approximate Gauss-Newton methods that will be examined in subsequent sections.

2.1 Statement of the algorithm

We consider the nonlinear least squares problem (NLSP) defined in (1), where we assume that $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a nonlinear, twice continuously Fréchet differentiable function. We denote the Jacobian of the function f by $J(x) \equiv f'(x)$. The gradient and Hessian of $\phi(x)$ are then given by

$$\nabla\phi(x) = J^T(x)f(x), \tag{2}$$
$$\nabla^2\phi(x) = J^T(x)J(x) + Q(x), \tag{3}$$

where $Q(x)$ denotes the second order terms

$$Q(x) = \sum_{i=1}^{m} f_i(x)\nabla^2 f_i(x). \tag{4}$$

Finding the stationary points of ϕ is equivalent to solving the gradient equation

$$F(x) \equiv \nabla\phi(x) = J^T(x)f(x) = 0. \tag{5}$$

Techniques for treating the NLSP can thus be derived from methods for solving this nonlinear algebraic system.

A common method for solving nonlinear equations of form (5) and hence for solving the NLSP (1) is Newton's method [9]. This method requires the inversion of the full Hessian matrix (3) of function ϕ. For many large scale problems, the second order terms $Q(x)$ of the Hessian are, however, impracticable to calculate and, in order to make the procedure more efficient, Newton's method is approximated by ignoring these terms. The resulting iterative method is known as the Gauss-Newton algorithm [9] and is defined as follows.

Gauss-Newton Algorithm (GN)

Step 0 : Choose an initial $x_0 \in \mathbb{R}^n$

Step 1 : Repeat until convergence:

Step 1.1 : Solve $J(x_k)^T J(x_k)s_k = -J^T(x_k)f(x_k)$

Step 1.2 : Set $x_{k+1} = x_k + s_k$.

□

Remarks: We note that at each iteration, Step 1.1 of the method is equivalent to solving the linearized least squares problem

$$\min_s \frac{1}{2}\left\|J(x_k)s + f(x_k)\right\|_2^2. \tag{6}$$

We note also that the GN method can be written as a fixed-point iteration of the form

$$x_{k+1} = G(x_k) \tag{7}$$

where $G(x) \equiv x - J^+(x)f(x)$ and $J(x)^+ \equiv (J^T(x)J(x))^{-1}J^T(x)$ denotes the Moore-Penrose pseudo-inverse of $J(x)$.

2.2 Convergence of the exact Gauss-Newton method

Sufficient conditions for the convergence of the Gauss-Newton method are known in the case where the normal equations for the linearized least squares problem (6) are solved *exactly* in Step 1.1 at each iteration. We now recall some existing results. The following assumptions are made in order to establish the theory:

A1. there exists $x^* \in \mathbb{R}^n$ such that $J^T(x^*)f(x^*) = 0$;

A2. the Jacobian matrix $J(x^*)$ at x^* has full rank n.

We introduce the notation $\rho(A)$ to indicate the spectral radius of an $n \times n$ matrix A, and define

$$\varrho = \rho\left(\left(J(x^*)^T J(x^*)\right)^{-1} Q(x^*)\right). \tag{8}$$

The following theorem on local convergence of the Gauss-Newton method then holds.

Theorem 1 *[9] Let assumptions* **A1.** *and* **A2.** *hold. If $\varrho < 1$, then the Gauss-Newton iteration converges locally to x^*; that is, there exists $\varepsilon > 0$ such that the sequence $\{x_k\}$ generated by the Gauss-Newton algorithm converges to x^* for all $x_0 \in \mathcal{D} \equiv \{x \mid \|x - x^*\|_2 < \varepsilon\}$.*

Theorem 1 has a geometrical interpretation as described in [13]. (See also [1]). We denote by $\mathcal{S}$ the surface in $\mathbb{R}^m$ given by the parametric representation $y = f(x)$, $x \in \mathbb{R}^n$, and let M be the point on $\mathcal{S}$ with coordinates $f(x^*)$, taking O as the origin of the coordinate system. The vector OM is orthogonal to the plane tangent to the surface $\mathcal{S}$ through M.

Theorem 2 *[13] Suppose that the assumptions of Theorem 1 hold and that $f(x^*)$ is nonzero. Then*

$$\varrho = \|f(x^*)\|_2 \chi, \tag{9}$$

where χ is the maximal principal curvature of the surface $\mathcal{S}$ at point M with respect to the normal direction $w^* = f(x^*)/\|f(x^*)\|_2$.

In the *zero residual* case, where $f(x^*) = 0$, the relation (9) continues to hold. In this case the origin O lies on the surface $\mathcal{S}$ and χ denotes the maximal principle curvature of $\mathcal{S}$ with respect to the direction normal to the tangent surface at O. Since we then have $Q(x^*) = 0$ and hence $\varrho = 0$, the result still holds.

For the Gauss-Newton method to converge it is therefore sufficient for the maximal principal curvature χ of the surface $\mathcal{S}$ at the point $f(x^*)$ to satisfy $1/\chi > \|f(x^*)\|_2$. This condition holds if and only if $\nabla^2\phi(x^*)$ is positive definite at x^* and ensures that x^* is a local minimizer of the objective function ϕ [1]. The relation (9) implies that the convergence condition of Theorem 1 is invariant under transformation of the NLSP by a local diffeomorphism, since the quantity $\|f(x^*)\|_2 \chi$ has this property [13].

The proofs of these results depend on theory for stationary fixed point iteration processes [9]. The theory ensures local convergence at a linear rate. Additional, but more restrictive, conditions for local convergence are given in [5]. Conditions giving higher order rates of convergence can be deduced from this theory. The Gauss-Newton method can also be treated as an inexact Newton's method [11], [4], [2]. Results of these types will be discussed further in Sections 4 and 5.

3 Approximate Gauss-Newton Algorithms

A serious difficulty associated with the use of the Gauss-Newton method in large scale applications, such as data assimilation, is that the linearized least squares problem (6) is computationally too expensive to solve exactly in Step 1.1 of the algorithm at each iteration. The dimensions of the normal matrix equations to be solved in Step 1.1 are often so great that the system coefficients cannot be stored in core memory, even in factored form. Therefore, in order to solve the full nonlinear problem efficiently, in real forecasting time, approximations must be made within the Gauss-Newton procedure.

Two types of approximation are commonly applied. Firstly, the linearized least squares problem (6) is solved only approximately by an 'inner' iteration method that is truncated before full accuracy is reached. We refer to this approximate algorithm as the Truncated Gauss-Newton (TGN) method . Secondly, the linearized least squares problem in Step 1.1 is replaced by an approximate, simplified or perturbed, linear problem that can be solved more efficiently in the inner loop. We refer to this algorithm as the Perturbed Gauss-Newton (PGN) method . Here we examine both of these approximate Gauss-Newton methods and also the combined Truncated Perturbed Gauss-Newton (TPGN) method, where both approximations are applied. In the next subsections we define these procedures explicitly, and in Sections 4 and 5 we analyse the convergence of the approximate methods.

3.1 Truncated Gauss-Newton method

At each outer iteration k of the Gauss-Newton method, we solve the normal equations

$$J(x_k)^T J(x_k) s = -J(x_k)^T f(x_k) \tag{10}$$

for the linearized least squares problem (6) using an iterative procedure. Intuitively, when x_k is far from x^* and the function f is nonlinear, it is not worth solving (10) to high accuracy. A natural stopping criterion for the iterative process is where the relative residual satisfies

$$\left\| J(x_k)^T J(x_k) s_k + J(x_k)^T f(x_k) \right\|_2 / \left\| J(x_k)^T f(x_k) \right\|_2 \leq \beta_k. \tag{11}$$

Here s_k denotes the current estimate of the solution of (10) and β_k is a specified tolerance. For this reason we define the Truncated Gauss-Newton algorithm as follows.

Truncated Gauss-Newton Algorithm (TGN)

Step 0 : Choose an initial $x_0 \in \mathbb{R}^n$

Step 1 : Repeat until convergence:

Step 1.1 : Find s_k such that
$(J(x_k)^T J(x_k)) s_k = -J(x_k)^T f(x_k) + r_k$,
with $\|r_k\|_2 \leq \beta_k \left\| J(x_k)^T f(x_k) \right\|_2$

Step 1.2 : Update $x_{k+1} = x_k + s_k$.

□

The tolerances β_k, $k = 0, 1, 2 \ldots$, must be chosen to ensure overall convergence of the procedure to the optimal x^* of the nonlinear least squares problem (1). Conditions guaranteeing convergence of the TGN method are presented in Sections 4 and 5.

3.2 Perturbed Gauss-Newton method

For some applications it is desirable to apply a perturbed Gauss-Newton method in which the true Jacobian J is replaced by an approximation $\tilde{J}$; this is practical, for example, in cases where a perturbed Jacobian is much easier or computationally less expensive to calculate. We therefore define the perturbed Gauss-Newton method as follows.

Perturbed Gauss-Newton Algorithm (PGN)

Step 0 : Choose an initial $x_0 \in \mathbb{R}^n$

Step 1 : Repeat until convergence:

Step 1.1 : Solve $\tilde{J}(x_k)^T \tilde{J}(x_k) s_k = -\tilde{J}(x_k)^T f(x_k)$

Step 1.2 : Set $x_{k+1} = x_k + s_k$.

□

We emphasize that in Step 1.1 of the PGN algorithm only the Jacobian is approximated and not the nonlinear function $f(x_k)$. The approximate Jacobian, $\tilde{J}(x)$, is assumed to be continuously Fréchet differentiable.

In order to interpret this iteration, it is convenient to define the function

$$\tilde{F}(x) = \tilde{J}(x)^T f(x), \tag{12}$$

and to write its first derivative in the form

$$\tilde{F}'(x) = \tilde{J}(x)^T J(x) + \tilde{Q}(x), \tag{13}$$

where $J(x)$ is the Jacobian of the function $f(x)$ and $\tilde{Q}(x)$ represents second order terms arising from the derivative of $\tilde{J}(x)$. Then the PGN algorithm can be considered as a fixed point algorithm for finding a solution $\tilde{x}^*$ to the equation

$$\tilde{F}(x) = 0. \tag{14}$$

We remark that, just as the GN method can be regarded as an inexact Newton's method for solving the gradient equation (5), the PGN method can be treated as an inexact Newton's method for solving the perturbed gradient equation (14). In the PGN method, the second order term in the derivative $\tilde{F}'$ is ignored and the first order term is now approximated, allowing the fixed point iteration to be written as a sequence of linear least squares problems.

For the zero residual NLSP, where $f(x^*) = 0$, the solution x^* of the problem satisfies (14), and so the fixed point $\tilde{x}^* = x^*$ of the PGN procedure is also a fixed point of the exact GN iteration. Similarly, if $f(x^*)$ lies in the null space of $\tilde{J}(x^*)$, then (14) is satisfied by x^* and the fixed point of the PGN method is again a fixed point of the exact GN method. In general, the fixed point of the PGN method will not be the same as that of the GN algorithm. We might expect, however, that if $\tilde{J}$ is close to J, then the solution $\tilde{x}^*$ of (14) will be close to the solution x^* of the true gradient equation (5).

In Sections 4 and 5 we give conditions for the PGN method to converge locally, and in Section 4 we also examine the distance between the fixed points of the two algorithms.

3.3 Truncated Perturbed Gauss-Newton method

In the PGN method, we solve the normal equations in Step 1.1 of the algorithm at each outer iteration k by applying an iterative method to the perturbed linear least squares problem

$$\min_s \frac{1}{2} \left\| \tilde{J}(x_k)s + f(x_k) \right\|_2^2. \tag{15}$$

To improve the efficiency of the PGN procedure, the iteration is truncated before full accuracy is reached. The iterations are halted where the relative residual satisfies

$$\left\| \tilde{J}(x_k)^T \tilde{J}(x_k)s_k + \tilde{J}(x_k)^T f(x_k) \right\|_2 / \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2 \leq \beta_k. \tag{16}$$

Here s_k is the current estimate of the solution of (15) and β_k is a specified tolerance. This procedure is referred to as the Truncated Perturbed Gauss-Newton (TPGN) method and is defined as follows.

Truncated Perturbed Gauss-Newton Algorithm (TPGN)

Step 0 : Choose an initial $x_0 \in \mathbb{R}^n$

Step 1 : Repeat until convergence:

Step 1.1 : Find s_k such that
$\tilde{J}(x_k)^T \tilde{J}(x_k)s_k = -\tilde{J}(x_k)^T f(x_k) + r_k,$
with $\|r_k\|_2 \leq \beta_k \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2$

Step 1.2 : Update $x_{k+1} = x_k + s_k$.

□

The tolerances β_k, $k = 0, 1, 2 \ldots$, must be chosen to ensure overall convergence of the procedure to the optimal $\tilde{x}^*$ of the perturbed gradient equation (14). Conditions guaranteeing local convergence of the TPGN method are presented in Sections 4 and 5.

4 Convergence of Approximate Gauss-Newton Methods I

We now derive sufficient conditions for the convergence of the truncated and perturbed Gauss-Newton methods. The theory is based on two different approaches. In this section we present results based on theory for inexact Newton methods found in [4] and [2]. In the subsequent section we extend the arguments of [5] for exact Gauss-Newton methods to the approximate truncated and perturbed methods.

We begin by introducing the theory for inexact Newton methods. This theory is applied to the exact Gauss-Newton method to obtain a new convergence condition. Criteria for the convergence of the truncated and perturbed methods are then derived using these results.

4.1 Inexact Newton methods

The inexact Newton method for solving the NLSP problem (1), as defined in [4], is given as follows.

Inexact Newton Algorithm (IN)

Step 0 : Choose an initial $x_0 \in \mathbb{R}^n$

Step 1 : Repeat until convergence:

Step 1.1 : Solve $\nabla^2\phi(x_k)s_k = -\nabla\phi(x_k) + \tilde{r}_k$

Step 1.2 : Set $x_{k+1} = x_k + s_k$.

□

In Step 1.1 the residual errors $\tilde{r}_k$ measure the amount by which the calculated solution s_k fails to satisfy the exact Newton method at each iteration. It is assumed that the relative sizes of these residuals are bounded by a nonnegative forcing sequence $\{\eta_k\}$ such that for each iteration

$$\frac{\|\tilde{r}_k\|_2}{\|\nabla\phi(x_k)\|_2} \leq \eta_k. \tag{17}$$

Conditions for the convergence of IN are established in the following theorem.

Theorem 3 *[4] Let assumptions* **A1.** *and* **A2.** *hold and let* $\nabla^2\phi(x^*)$ *be nonsingular. Assume* $0 \leq \eta_k \leq \hat{\eta} < t < 1$. *Then there exists* $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 \leq \varepsilon$, *the sequence of inexact Newton iterates* $\{x_k\}$ *satisfying (17) converges to* x^*. *Moreover, the convergence is linear in the sense that*

$$\| x_{k+1} - x^* \|_* \leq t \| x_k - x^* \|_*, \tag{18}$$

where $\| y \|_* = \left\|\nabla^2\phi(x^*)y\right\|_2$.

In [2] Theorem 3 is applied to obtain more general results in which the Jacobian and Hessian matrices are perturbed on each iteration of the Newton method. Here we adopt similar techniques to derive results for the approximate Gauss-Newton methods based on theory for the inexact Newton methods.

4.2 Gauss-Newton as an inexact Newton method

We first establish novel sufficient conditions for the exact Gauss-Newton method to converge by treating it as an inexact Newton method.

Theorem 4 *Let assumptions* **A1.** *and* **A2.** *hold and let* $\nabla^2\phi(x^*)$ *be nonsingular. Assume* $0 \leq \hat{\eta} < 1$. *Then there exists* $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 \leq \varepsilon$ *and if*

$$\left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 \leq \eta_k \leq \hat{\eta}, \quad \textit{for} \quad k = 0, 1, \ldots, \tag{19}$$

the sequence of Gauss-Newton iterates $\{x_k\}$ *converges to* x^*.

Proof of Theorem 4 : We can write the GN method as an IN method by setting

$$\begin{aligned} \tilde{r}_k &= \nabla\phi(x_k) - \nabla^2\phi(x_k)(J^T(x_k)J(x_k))^{-1}\nabla\phi(x_k) \\ &= (I - \nabla^2\phi(x_k)(J^T(x_k)J(x_k))^{-1})\nabla\phi(x_k). \end{aligned} \tag{20}$$

Then, using (3), we have

$$\begin{aligned} \|\tilde{r}_k\|_2 &= \left\|(I - \nabla^2\phi(x_k)(J^T(x_k)J(x_k))^{-1})\nabla\phi(x_k)\right\|_2 \\ &\leq \left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 \|\nabla\phi(x_k)\|_2 . \end{aligned} \tag{21}$$

By Theorem 3, a sufficient condition for local convergence is therefore

$$\left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 \leq \eta_k \leq \hat{\eta}, \quad k = 0, 1, \ldots. \tag{22}$$

□

The convergence condition derived in this theorem is less general than that obtained in Theorem 1, which requires a bound only on the spectral radius of the matrix $Q(x)(J^T(x)J(x))^{-1}$ at the fixed point $x = x^*$, rather than on its norm at each iterate x_k. However, the technique used in the proof of Theorem 4 provides a practical test for convergence and is more readily extended to the case of the approximate Gauss-Newton iterations.

4.3 Convergence of the Truncated Gauss-Newton method (I)

We now give a theorem that provides sufficient conditions for the convergence of the truncated Gauss-Newton (TGN) method. It is assumed that the residuals in the TGN method are bounded such that

$$\|r_k\|_2 \leq \beta_k \|\nabla\phi(x_k)\|_2 , \tag{23}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence. The theorem is established by considering the algorithm as an inexact Newton method, as in the proof of Theorem 4.

Theorem 5 *Let assumptions* **A1.** *and* **A2.** *hold and let $\nabla^2\phi(x^*)$ be nonsingular. Assume that $0 \leq \hat{\beta} < 1$ and select β_k, $k = 0, 1, \ldots$ such that*

$$0 \leq \beta_k \leq \frac{\hat{\beta} - \left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2}{1 + \|Q(x_k)(J^T(x_k)J(x_k))^{-1}\|_2}, \quad k = 0, 1, \ldots . \tag{24}$$

Then there exists $\varepsilon > 0$ such that, if $\|x_0 - x^*\|_2 \leq \varepsilon$, the sequence of truncated Gauss-Newton iterates $\{x_k\}$ satisfying (23) converges to x^*.

Proof of Theorem 5 : We can write the TGN method as an IN method by setting

$$\tilde{r}_k = \nabla\phi(x_k) - \nabla^2\phi(x_k)(J^T(x_k)J(x_k))^{-1}\nabla\phi(x_k) + \nabla^2\phi(x_k)(J^T(x_k)J(x_k))^{-1}r_k. \tag{25}$$

Then we have

$$\begin{aligned}
\|\tilde{r}_k\|_2 &\leq \left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 \|\nabla\phi(x_k)\|_2 + \left\|I + Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 \|r_k\|_2 \\
&\leq (\left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 + \beta_k(1 + \left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2)) \|\nabla\phi(x_k)\|_2 \\
&\leq (\left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 + (\hat{\beta} - \left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2)) \|\nabla\phi(x_k)\|_2 \\
&\leq \hat{\beta} \|\nabla\phi(x_k)\|_2 .
\end{aligned} \tag{26}$$

Local convergence then follows from Theorem 3.

□

Since $\beta_k \geq 0$ is necessary, we require that $\left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2 \leq \hat{\beta} < 1$. This is just the sufficient condition given by Theorem 4 for the exact Gauss-Newton method to converge.

We remark also that when the problem is highly nonlinear, then $\left\|Q(x_k)(J^T(x_k)J(x_k))^{-1}\right\|_2$ will be large and hence the limit on β_k will be small. The 'inner' iteration of the TGN method must then be solved more accurately to ensure convergence of the algorithm.

4.4 Convergence of the Perturbed Gauss-Newton method (I)

Next we present sufficient conditions for the perturbed Gauss-Newton (PGN) method to converge. The theorem is established by considering the PGN method as an inexact Newton method for solving the perturbed gradient equation (14). We make the assumptions:

A1′. there exists $\tilde{x}^* \in \mathbb{R}^n$ such that $\tilde{F}(\tilde{x}^*) \equiv \tilde{J}^T(\tilde{x}^*) f(\tilde{x}^*) = 0$;

A2′. the matrix $\tilde{J}(\tilde{x}^*)$ at $\tilde{x}^*$ has full rank n.

We then obtain the following theorem.

Theorem 6 *Let assumptions* **A1′.** *and* **A2′.** *hold and let* $\tilde{F}'(\tilde{x}^*) \equiv \tilde{J}(\tilde{x}^*)^T J(\tilde{x}^*) + \tilde{Q}(\tilde{x}^*)$ *be nonsingular. Assume* $0 \leq \hat{\eta} < 1$. *Then there exists* $\varepsilon > 0$ *such that if* $\|x_0 - x^*\|_2 \leq \varepsilon$ *and if*

$$\left\| I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \right\|_2 \leq \eta_k \leq \hat{\eta}, \quad k = 0, 1, \ldots, \tag{27}$$

the sequence of perturbed Gauss-Newton iterates $\{x_k\}$ *converges to* $\tilde{x}^*$.

Proof of Theorem 6 : We can write the PGN method as an IN method by setting

$$\begin{aligned} \tilde{r}_k &= \tilde{J}(x_k)^T f(x_k) - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \tilde{J}(x_k)^T f(x_k) && (28) \\ &= (I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}) \tilde{J}(x_k)^T f(x_k). && (29) \end{aligned}$$

Then, provided the condition (27) holds, we have

$$\|\tilde{r}_k\|_2 \leq \hat{\eta} \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2, \tag{30}$$

and by Theorem 3 local convergence is guaranteed.

□

The theorem gives explicit conditions on the perturbed Jacobian $\tilde{J}$ that are sufficient to guarantee the convergence of the perturbed Gauss-Newton method. The requirement is that $\tilde{J}(x_k)^T \tilde{J}(x_k)$ should be a good approximation to the derivative $\tilde{F}'(x) = \tilde{J}(x)^T J(x) + \tilde{Q}(x)$ of the perturbed gradient equation (14).

4.5 Fixed point of the Perturbed Gauss-Newton method

We now consider how close the solution $\tilde{x}^*$ of the perturbed gradient equation (14) is to the solution x^* of the original NLSP. To answer this question we treat the GN method as a stationary fixed-point iteration of the form (7).

We assume that the GN iteration converges locally to x^* for all x_0 in an open convex set $\mathcal{D}$ containing x^* (defined as in Theorem 1) and that $G(x)$ satisfies

$$\|G(x) - G(x^*)\|_2 \leq \nu \|x - x^*\|_2, \quad \forall\, x \in \mathcal{D}, \quad \text{with} \quad \nu < 1. \tag{31}$$

Then we have the following theorem, which bounds the distance between the solutions of the exact and perturbed iterations.

Theorem 7 *Let assumptions* **A1.**, **A2.**, **A1′.** *and* **A2′.** *hold and assume* $\varrho < 1$. *Let (31) be satisfied and let* $\tilde{x}^* \in \mathcal{D}$. *Then*

$$\|\tilde{x}^* - x^*\|_2 \leq \frac{1}{1 - \nu} \left\| (\tilde{J}^+(\tilde{x}^*) - J^+(\tilde{x}^*)) f(\tilde{x}^*) \right\|_2. \tag{32}$$

Proof of Theorem 7 : We define $\tilde{G}(x) = x - \tilde{J}^+(x)f(x)$. Then $\tilde{x}^* = \tilde{G}(\tilde{x}^*)$ and we have

$$\begin{aligned} \|\tilde{x}^* - x^*\|_2 &= \left\|\tilde{G}(\tilde{x}^*) - G(x^*)\right\|_2 \\ &\leq \left\|\tilde{G}(\tilde{x}^*) - G(\tilde{x}^*)\right\|_2 + \|G(\tilde{x}^*) - G(x^*)\|_2 \\ &\leq \nu \|\tilde{x}^* - x^*\|_2 + \left\|\tilde{G}(\tilde{x}^*) - G(\tilde{x}^*)\right\|_2 . \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \|\tilde{x}^* - x^*\|_2 &\leq \frac{1}{1-\nu}\left\|\tilde{G}(\tilde{x}^*) - G(\tilde{x}^*)\right\|_2 \\ &\leq \frac{1}{1-\nu}\left\|(\tilde{J}^+(\tilde{x}^*) - J^+(\tilde{x}^*))f(\tilde{x}^*)\right\|_2 . \end{aligned}$$

□

The theorem shows that the distance between x^* and $\tilde{x}^*$ is bounded in terms of the distance between the pseudo-inverses of $\tilde{J}$ and J at $\tilde{x}^*$ and will be small if these are close together. The theorem also implies, from (14), that the bound given in (32) equals $\|J^+(\tilde{x}^*)f(\tilde{x}^*)\|_2/(1-\nu)$, which is proportional to the residual in the true gradient equation (5) evaluated at the solution $\tilde{x}^*$ of the perturbed gradient equation (14). (We remark that to ensure that the conditions of the theorem may be met, that is for $\tilde{x}^* \in \mathcal{D}$ to hold, it is sufficient that $\|J^+(\tilde{x}^*)f(\tilde{x}^*)\|_2/(1-\nu) < \varepsilon$, where ε is defined as in Theorem 1.)

A different approach to the convergence of the perturbed fixed point iteration can be found in [9]. This approach shows essentially that if the GN method converges, then the PGN iterates eventually lie in a small region around x^* of radius $\delta/(1-\nu)$, where δ bounds the distance $\left\|\tilde{G}(x) - G(x)\right\|_2$ over all $x \in \mathcal{D}$. This theory does not establish convergence of the perturbed method, but the theory for the distance between the fixed points of the GN and PGN methods presented here is consistent with these results.

4.6 Convergence of the Truncated Perturbed Gauss-Newton method (I)

We now examine the convergence of the approximate Gauss-Newton method where the Jacobian is perturbed and the inner linear least squares problem is not solved exactly. The residuals in the inner normal equations at each outer iteration are assumed to be bounded such that

$$\|r_k\|_2 \leq \beta_k \left\|\tilde{J}(x_k)^T f(x_k)\right\|_2 , \tag{33}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence. Sufficient conditions for the convergence of this truncated, perturbed Gauss-Newton method (TPGN) are given by the next theorem.

Theorem 8 *Let assumptions* **A1′.** *and* **A2′.** *hold and let* $\tilde{F}'(\tilde{x}^*) \equiv \tilde{J}(\tilde{x}^*)^T J(\tilde{x}^*) + \tilde{Q}(\tilde{x}^*)$ *be nonsingular. Assume that* $0 \leq \hat{\beta} < 1$ *and select* $\beta_k,\ k = 0, 1, \ldots$ *such that*

$$0 \leq \beta_k \leq \frac{\hat{\beta} - \left\|I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}\right\|_2}{\left\|(\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}\right\|_2}. \tag{34}$$

Then there exists $\varepsilon > 0$ *such that if* $\|x_0 - \tilde{x}^*\|_2 \leq \varepsilon$, *the sequence of perturbed Gauss-Newton iterates* $\{x_k\}$ *satisfying (33) converges to* $\tilde{x}^*$.

Proof of Theorem 8 : We can write TPGN in the same form as IN by setting

$$\begin{aligned} \tilde{r}_k &= (I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1})\tilde{J}(x_k)^T f(x_k) \\ &+ (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} r_k. \end{aligned} \tag{35}$$

Then, provided the condition (33) holds, we have

$$\|\tilde{r}_k\|_2 \leq \hat{\beta} \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2, \tag{36}$$

and by Theorem 3 local convergence is guaranteed.

□

We remark that in order to ensure $\beta_k \geq 0$ we also require that

$$\left\| I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \right\|_2 \leq \hat{\beta} < 1,$$

which is simply the sufficient condition found in Theorem 6 for the PGN method to converge.

4.7 Summary

In this section we have established theory ensuring local linear convergence of the Gauss-Newton, the truncated Gauss-Newton, the perturbed Gauss-Newton and the truncated perturbed Gauss-Newton methods based on the theory of [4] for inexact Newton methods. Numerical examples illustrating the results for the three approximate Gauss-Newton methods are shown in Section 6. In the next section we derive additional convergence conditions for these methods based on the theory of [5] for exact Gauss-Newton methods.

5 Convergence of Approximate Gauss-Newton Methods II

We now derive conditions for the convergence of the approximate Gauss-Newton methods by extending the results of [5] for the exact Gauss-Newton method. These results are more restrictive than those given in Section 4, but provide more precise estimates of the rates of convergence of the methods. Conditions for linear, super-linear and quadratic convergence are established.

5.1 Sufficient conditions for the exact Gauss-Newton method

We begin by recalling the sufficient conditions of [5] for local convergence of the Gauss-Newton iterates to a stationary point x^* of the nonlinear least squares problem (NLSP).

Theorem 9 *[5] Let assumptions* **A1.** *and* **A2.** *hold and let λ be the smallest eigenvalue of the matrix $J(x^*)^T J(x^*)$. Suppose that there exists an open convex set $\mathcal{D}$ containing x^* such that*

(i) $J(x)$ is Lipschitz continuous in $\mathcal{D}$ with a Lipschitz constant equal to γ;

(ii) $\|J(x)\|_2 \leq \alpha$ for all $x \in \mathcal{D}$;

(iii) there exists $\sigma \geq 0$ such that $\left\|J(x)^T f(x^*)\right\|_2 \leq \sigma \|x - x^*\|_2, \forall x \in \mathcal{D}$;

(iv) $\sigma < \lambda$.

Let c be such that $1 < c < \lambda/\sigma$. Then there exists $\varepsilon > 0$ such that, if $\|x_0 - x^*\|_2 < \varepsilon$, the iterates $\{x_k\}$ generated by the Gauss-Newton algorithm converge to x^*. Additionally, the following inequality holds

$$\|x_{k+1} - x^*\|_2 \leq \frac{c\sigma}{\lambda} \|x_k - x^*\|_2 + \frac{c\alpha\gamma}{2\lambda} \|x_k - x^*\|_2^2 . \tag{37}$$

□

The constant σ may be regarded as an approximation to the norm of the second-order terms $\|Q(x^*)\|_2$ and is a combined measure of the nonlinearity of the problem and the size of the residual [5]. The theorem shows that the convergence of the Gauss-Newton method is quadratic in the case $\sigma = 0$. This holds, for example, for the zero-residual problem where $f(x^*) = 0$.

The sufficient conditions given by Theorem 9 for the local convergence of the Gauss-Newton method are more restrictive than those given in Theorem 1. We demonstrate this as follows.

Theorem 10 *If the assumptions of Theorem 9 hold, then $\varrho < 1$.*

Proof of Theorem 10 : By differentiating the map $x \mapsto J(x)^T f(x^*)$ with respect to x, we find

$$J(x)^T f(x^*) = Q(x^*)(x - x^*) + \|x - x^*\|_2 \Theta(x - x^*), \tag{38}$$

with $\lim_{h\to 0} \Theta(h) = 0$. We denote $f(x^*)$, $J(x^*)$ and $Q(x^*)$ by f^*, J^* and Q^*, respectively. Then multiplying (38) by $(J^{*T}J^*)^{-1}$ on the left yields

$$(J^{*T}J^*)^{-1}J(x)^T f^* = (J^{*T}J^*)^{-1}Q^*(x - x^*) + \|x - x^*\|_2 \Theta_1(x - x^*), \tag{39}$$

with $\lim_{h\to 0} \Theta_1(h) = 0$. We let v be the right singular vector associated with the largest singular value of $(J^{*T}J^*)^{-1}Q^*$ and let $x_\epsilon = x^* + \epsilon v$ for $\epsilon > 0$. Substituting x_ϵ for x in (39) and rearranging the terms of the equality then gives us

$$\epsilon(J^{*T}J^*)^{-1}Q^*v = (J^{*T}J^*)^{-1}J(x_\epsilon)^T f^* - \epsilon\Theta_1(\epsilon v). \tag{40}$$

By the assumptions of Theorem 9, we have $\left\|J(x_\epsilon)^T f^*\right\|_2 \leq \sigma\epsilon$ for ϵ sufficiently small and therefore

$$\left\|(J^{*T}J^*)^{-1}J(x_\epsilon)^T f^*\right\|_2 \leq \left\|(J^{*T}J^*)^{-1}\right\|_2 \sigma\epsilon = \epsilon\sigma/\lambda. \tag{41}$$

Taking norms in (40) and letting ϵ tend to 0 then yields

$$\left\|(J^{*T}J^*)^{-1}Q^*\right\|_2 \leq \sigma/\lambda. \tag{42}$$

Since

$$\varrho \leq \left\|(J^{*T}J^*)^{-1}Q^*\right\|_2 , \tag{43}$$

we obtain $\varrho \leq \sigma/\lambda$. Therefore, if $\sigma < \lambda$, then $\varrho < 1$.

□

The conditions of Theorem 9 ensure that the conditions of Theorem 1 hold and that the exact Gauss-Newton method converges, but the conditions of Theorem 1 are weaker than those of Theorem 9 . Since the quantity $\sigma > \|Q(x^*)\|_2$ can be made arbitrarily close to $\|Q(x^*)\|_2$ in a sufficiently small neighbourhood of x^*, the condition $\sigma < \lambda < 1$ can be achieved only if $\|Q(x^*)\|_2 \left\|(J(x^*)^T J(x^*)^T)^{-1}\right\|_2 < 1$, which is a stronger requirement than that of Theorem 1 for convergence (see [6]).

We now extend the theory of Theorem 9 to the approximate Gauss-Newton methods. The results are not as general as those of Section 4, but allow the rates of convergence of the methods to be determined.

5.2 Convergence of the Truncated Gauss-Newton method (II)

By an extension of Theorem 9, we now establish alternative conditions for the truncated Gauss-Newton (TGN) method to converge. We assume, as previously, that the residuals in the TGN method are bounded such that

$$\|r_k\|_2 \leq \beta_k \left\|J(x_k)^T f(x_k)\right\|_2, \tag{44}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence.

Theorem 11 *Let the conditions of Theorem 9 hold and let c be such that $1 < c < \lambda/\sigma$. Select $\beta_k,\ k = 0, 1, \ldots$ to satisfy*

$$0 \leq \beta_k \leq \hat{\beta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)}, \quad k = 0, 1, \ldots. \tag{45}$$

Then there exists $\varepsilon > 0$ such that if $\|x_0 - x^*\|_2 < \varepsilon$, the sequence of truncated Gauss-Newton iterates $\{x_k\}$ satisfying (44) converges to x^*. Additionally, the following inequality holds :

$$\|x_{k+1} - x^*\|_2 \leq \frac{c}{\lambda}(\sigma + \beta_k(\sigma + \alpha^2)) \|x_k - x^*\|_2 + C \|x_k - x^*\|_2^2, \tag{46}$$

where $C = \dfrac{c\alpha\gamma}{2\lambda}(1 + \hat{\beta})$.

Proof of Theorem 11 : The proof is by induction. Let us denote by J_0, f_0, J^* and f^* the quantities $J(x_0)$, $f(x_0)$, $J(x^*)$ and $f(x^*)$. From the proof of Theorem 9 (see [5, Theorem 10.2.1]), there exists a positive quantity ε_1 such that, if $\|x_0 - x^*\|_2 < \varepsilon_1$, then $x_0 \in \mathcal{D}$, $J_0^T J_0$ is nonsingular, $\left\|(J_0^T J_0)^{-1}\right\|_2 < c/\lambda$, and

$$\left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 - x^*\right\|_2 \leq \frac{c\sigma}{\lambda} \|x_0 - x^*\|_2 + \frac{c\alpha\gamma}{2\lambda} \|x_0 - x^*\|_2^2. \tag{47}$$

Let

$$\varepsilon = \min\left\{\varepsilon_1, \frac{\lambda - c(\sigma + \hat{\beta}(\sigma + \alpha^2))}{c\alpha\gamma(1 + \hat{\beta})}\right\}, \tag{48}$$

where $\lambda - c(\sigma + \hat{\beta}(\sigma + \alpha^2)) > 0$ by (45).

We start from

$$\left\|J_0^T f_0\right\|_2 = \left\|J_0^T f^* + J_0^T(J_0(x_0 - x^*) + f_0 - f^*) - J_0^T J_0(x_0 - x^*)\right\|_2, \tag{49}$$

and bound successively each term in the norm. From the definitions of σ and α in Theorem 9, we have $\left\|J_0^T f^*\right\|_2 \leq \sigma \|x_0 - x^*\|_2$ and $\left\|J_0^T J_0(x_0 - x^*)\right\|_2 \leq \alpha^2 \|x_0 - x^*\|_2$. From [5, Lemma 4.1.12] and the Lipschitz continuity of J_0, we also have

$$\|J_0(x_0 - x^*) + f^* - f_0\|_2 \leq \frac{\gamma}{2} \|x_0 - x^*\|_2^2. \tag{50}$$

Using the triangular inequality then shows that

$$\left\|J_0^T f_0\right\|_2 \leq (\sigma + \alpha^2) \|x_0 - x^*\|_2 + \frac{\alpha\gamma}{2} \|x_0 - x^*\|_2^2. \tag{51}$$

Gathering the partial results (47) and (51), we obtain

$$\begin{aligned}\|x_1 - x^*\|_2 &= \left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 + (J_0^T J_0)^{-1} r_0 - x^*\right\|_2 \\ &\leq \left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 - x^*\right\|_2 + \|r_0\|_2 \left\|(J_0^T J_0)^{-1}\right\|_2 \\ &\leq \left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 - x^*\right\|_2 + \beta_0 \left\|(J_0^T J_0)^{-1}\right\|_2 \left\|J_0^T f_0\right\|_2 \\ &\leq \frac{c}{\lambda}(\sigma + \beta_0(\sigma + \alpha^2)) \|x_0 - x^*\|_2 + C \|x_0 - x^*\|_2^2 , \end{aligned} \tag{52}$$

where $C = c\alpha\gamma(1 + \hat{\beta})/(2\lambda)$, which proves (46) in the case $k = 0$. Since $\|x_0 - x^*\|_2 < \varepsilon$ is assumed initially, it follows from (48) that

$$\|x_1 - x^*\|_2 \leq \left(\frac{c}{\lambda}(\sigma + \hat{\beta}(\sigma + \alpha^2)) + C\varepsilon\right) \|x_0 - x^*\|_2 \leq K \|x_0 - x^*\|_2 < \|x_0 - x^*\|_2 , \tag{53}$$

where $K = (\lambda + c(\sigma + \hat{\beta}(\sigma + \alpha^2)))/(2\lambda) < 1$. The convergence is then established by repeating the argument for $k = 1, 2, \ldots$.

□

The theorem shows that to ensure the convergence of the TGN method, the relative residuals in the solution of the 'inner' linear least square problem must be bounded in terms of the parameters σ, λ and α. The theorem also establishes the rates of convergence of the method in various cases. These cases are discussed in Section 5.5.

5.3 Convergence of the Perturbed Gauss-Newton method (II)

In the next theorem we consider the perturbed Gauss-Newton iteration where an approximate Jacobian $\tilde{J}$ is used instead of J.

Theorem 12 *Let the conditions of Theorem 9 hold and let $\tilde{J}(x)$ be an approximation to $J(x)$. Let c be such that $1 < c < \lambda/\sigma$. Assume that*

$$0 \leq \hat{\eta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)}. \tag{54}$$

Then there exists $\varepsilon > 0$ such that if $\|x_0 - x^*\|_2 < \varepsilon$, and if

$$\left\|J(x_k)^T J(x_k) \left(J^+(x_k) - \tilde{J}^+(x_k)\right) f(x_k)\right\|_2 / \left\|J(x_k)^T f(x_k)\right\|_2 \leq \eta_k \leq \hat{\eta}, \quad k = 0, 1, \ldots, \tag{55}$$

the sequence of perturbed Gauss-Newton iterates $\{x_k\}$ converge to x^*. Additionally, the following inequality holds :

$$\|x_{k+1} - x^*\|_2 \leq \frac{c}{\lambda}(\sigma + \eta_k(\sigma + \alpha^2)) \|x_k - x^*\|_2 + C \|x_k - x^*\|_2^2 , \tag{56}$$

where $C = c\alpha\gamma(1 + \hat{\eta})/(2\lambda)$.

Proof of Theorem 12 : The perturbed Gauss-Newton iteration takes the form $x_{k+1} = x_k + s_k$, where $s_k = -\tilde{J}^+(x_k) f(x_k)$. Therefore, using the notation of Theorem 11, we may consider the PGN method as a truncated Gauss-Newton method with the residual defined by

$$r_k = J(x_k)^T J(x_k) s_k + J(x_k)^T f(x_k) = J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k)) f(x_k).$$

The conclusion then follows directly from Theorem 11.

□

We remark that Theorem 12 establishes the convergence of the PGN method to the fixed point x^* of the *exact* Gauss-Newton method. At the fixed point, the perturbed Jacobian $\tilde{J}$ must, therefore, be such that $\tilde{J}(x^*)^T f(x^*) = 0$ in order to be able to satisfy the conditions of the theorem; that is, at the fixed point x^*, the null space of $\tilde{J}(x^*)^T$ must contain $f(x^*)$. In contrast the convergence results of Theorem 6 only require that a point $\tilde{x}^*$ exists such that $\tilde{J}(\tilde{x}^*)^T f(\tilde{x}^*) = 0$ and $\tilde{J}(\tilde{x}^*)$ is full rank.

5.4 Convergence of the Truncated Perturbed Gauss-Newton method (II)

In the following theorem we consider the truncated perturbed Gauss-Newton iteration where an approximate Jacobian $\tilde{J}$ is used and the inner linear least squares problem (15) is not solved exactly on each outer step. The residuals in the inner normal equations at each outer iteration are assumed to be bounded such that

$$\|r_k\|_2 \leq \beta_k \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2, \tag{57}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence. Sufficient conditions for the TPGN method to converge are then given as follows.

Theorem 13 *Let the conditions of Theorem 9 hold and let $\tilde{J}(x)$ be an approximation to $J(x)$. Let c be such that $1 < c < \lambda/\sigma$. Assume that $\eta_k \leq \hat{\eta} < (\lambda - c\sigma)/(c(\sigma + \alpha^2))$ and select β_k, $k = 0, 1, \ldots$ such that*

$$\begin{aligned} 0 \;\leq\; \beta_k \;\leq\; & (\eta_k \left\| J(x_k)^T f(x_k) \right\|_2 - \left\| J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k)) f(x_k) \right\|_2) \\ & \cdot \left(\left\| J(x_k)^T J(x_k)(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \right\|_2 \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2 \right)^{-1}, \end{aligned} \tag{58}$$

for $k = 0, 1, \ldots$. Then there exists $\varepsilon > 0$ such that if $\|x_0 - x^*\|_2 < \varepsilon$, the sequence of perturbed Gauss-Newton iterates $\{x_k\}$ satisfying (57) converges to x^*. Additionally, the following inequality holds :

$$\|x_{k+1} - x^*\|_2 \leq \frac{c}{\lambda}(\sigma + \eta_k(\sigma + \alpha^2)) \|x_k - x^*\|_2 + C \|x_k - x^*\|_2^2, \tag{59}$$

where $C = c\alpha\gamma(1 + \hat{\eta})/(2\lambda)$.

Proof of Theorem 13 : The TPGN iteration takes the form $x_{k+1} = x_k + s_k$, where $s_k = -\tilde{J}^+(x_k) f(x_k) + (\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} r_k$. Therefore, using the notation of Theorem 11, we may consider the TPGN method as a truncated Gauss-Newton method with the residual defined as

$$\tilde{r}_k = J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k)) f(x_k) + J(x_k)^T J(x_k)(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} r_k. \tag{60}$$

Then, provided the condition (57) holds, we have

$$\begin{aligned} \|\tilde{r}_k\|_2 \;\leq\; & \left\| J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k)) f(x_k) \right\|_2 + \\ & \left\| J(x_k)^T J(x_k)(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \right\|_2 \beta_k \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2 \\ \leq\; & \eta_k \left\| J(x_k)^T f(x_k) \right\|_2. \end{aligned} \tag{61}$$

The conclusion then follows from Theorem 11.

□

We remark that to ensure $\beta_k \geq 0$, we require that the relation given by equation (55) holds. This is simply the condition of Theorem 12 that guarantees the convergence of the PGN method in the case where the inner loop is solved exactly without truncation.

Theorem 13 gives conditions for the truncated perturbed Gauss-Newton method to converge to the fixed point x^* of the *exact* Gauss-Newton method, and is therefore more restrictive than the theorem developed in Section 4. Here the allowable form of the perturbed Jacobian is constrained to satisfy $\tilde{J}(x^*)^T f(x^*) = J(x^*)^T f(x^*) = 0$ in order that the conditions of the theorem may be met. The theorem does, however, establish that the method converges with rates of convergence higher than linear in certain cases. These cases are discussed in the next subsection.

5.5 Rates of convergence of the approximate Gauss-Newton methods

¿From Theorems 11, 12 and 13, the expected convergence rates of the approximate Gauss-Newton methods may be established for various cases. The convergence rates are shown in (46), (56) and (59) for the truncated Gauss-Newton, the perturbed Gauss-Newton and the truncated perturbed Gauss-Newton methods, respectively. These rates are dependent on the parameters σ, λ and α, defined as in Theorem 9, and can be contrasted directly with the convergence rates of the exact Gauss-Newton method, given by (37). We observe the following.

1. *Linear convergence.* The theorems show that in general the GN, TGN, PGN and TPGN methods converge linearly. In comparison with the exact GN algorithm, we see that the price paid for the inaccurate solution of the linear least squares problem in the inner step of the approximate methods is a degradation of the local linear rate of convergence.

2. *Super-linear convergence.* As previously noted, if $\sigma = 0$, which holds, for example, in the zero-residual case where $f(x^*) = 0$, the convergence of the exact GN method is quadratic. In this same case, if $\sigma = 0$ and if the forcing sequence $\{\beta_k\}$ satisfies $\lim_{k\to+\infty} \beta_k = 0$, then the convergence rates of the TGN and TPGN methods are super-linear. For the PGN method to converge super-linearly in this case, the sequence $\{\eta_k\}$ must satisfy $\lim_{k\to+\infty} \eta_k = 0$.

3. *Quadratic convergence.* From the proof of Theorem 11, we see that the convergence of the TGN method is quadratic if $\sigma = 0$ and if the normal equation residual is such that

$$\|r_k\|_2 \equiv \left\|J(x_k)^T J(x_k) s_k + J(x_k)^T f(x_k)\right\|_2 \leq C_1 \left\|J(x_k)^T f(x_k)\right\|_2^2,$$

   for some positive constant C_1. Similarly, in the case $\sigma = 0$, the PGN method converges quadratically if

$$\left\|J(x_k)^T J(x_k) \left(J^+(x_k) - \tilde{J}^+(x_k)\right) f(x_k)\right\|_2 \leq C_2 \left\|J(x_k)^T f(x_k)\right\|_2^2,$$

   as does the TPGN method in this case if

$$\left\|(J(x_k)^T J(x_k))((J(x_k)^+ - \tilde{J}(x_k)^+) f(x_k) + (\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} r_k)\right\|_2 \leq C_3 \left\|J(x_k)^T f(x_k)\right\|_2^2,$$

   for positive constants C_2, C_3.

4. *Effect of nonlinearity.* Since $\lambda - c\sigma > 0$, we also see from the theorems that the allowable upper bound on the truncation decreases as σ increases. Since σ is a combined measure

of the nonlinearity and the residual size in the problem, we see therefore that, in order to guarantee convergence of the approximate methods, the inner linearized equation must be solved more accurately when the problem is highly nonlinear or when there is a large residual at the optimal.

In Section 6 we give numerical results demonstrating the convergence behaviour of the approximate Gauss-Newton methods. The rates of convergence of the approximate methods are also illustrated for various cases.

5.6 Summary

In this section we have established theory ensuring local convergence of the Gauss-Newton, the truncated Gauss-Newton, the perturbed Gauss-Newton and the truncated perturbed Gauss-Newton methods based on the theory of [5] for exact Gauss-Newton methods. The conditions for convergence derived in this section are more restrictive than those of Section 4, but enable the rates of convergence to be established. Numerical examples illustrating the results for the three approximate Gauss-Newton methods are shown in the next section.

6 Numerical example

We examine the theoretical results of Sections 4 and 5 using a simple initial value problem discretized by a second-order Runge-Kutta scheme. The example is based on that in [7, Chapter 4] and is used because it provides a clear way of producing a perturbed Jacobian. We consider the ordinary differential equation

$$\frac{dz}{dt} = z^2, \tag{62}$$

where $z = z(t)$ and $z(0) = z_0$ is given. Application of a second order Runge-Kutta scheme gives a discrete nonlinear model

$$x^{n+1} = x^n + (x^n)^2\Delta t + (x^n)^3\Delta t^2 + \frac{1}{2}(x^n)^4\Delta t^3, \tag{63}$$

where Δt denotes the model time step and $x^n \approx z(t_n)$ at time $t_n = n\Delta t$. We define a least squares problem

$$\min_{x^0} \phi(x) = \frac{1}{2}(x^0 - y^0)^2 + \frac{1}{2}(x^1 - y^1)^2 \tag{64}$$

subject to (63), where $y^0,\ y^1$ are values of observed data at times $t_0,\ t_1$. This is of the same form as (1), with

$$f = \begin{pmatrix} x^0 - y^0 \\ x^1 - y^1 \end{pmatrix}. \tag{65}$$

Then the Jacobian of f is given by

$$J(x^0) = \begin{pmatrix} 1 \\ 1 + 2x^0\Delta t + 3(x^0)^2\Delta t^2 + 2(x^0)^3\Delta t^3 \end{pmatrix} \tag{66}$$

and the second order terms of the Hessian are

$$Q(x^0) = (x^0 + (x^0)^2\Delta t + (x^0)^3\Delta t^2 + \frac{1}{2}(x^0)^4\Delta t^3 - y^1)(2\Delta t + 6x^0\Delta t^2 + 6(x^0)^2\Delta t^3). \tag{67}$$

Table 1: Perfect observations, exact Jacobian

ϵ	Iterations	Error	Gradient
0.00	5	0.000000e+00	0.000000e+00
0.25	20	9.015011e-14	1.364325e-13
0.50	37	7.207568e-13	1.092931e-12
0.75	84	2.246647e-12	3.407219e-12
0.90	210	8.292034e-12	1.257587e-11
0.95	401	1.857048e-11	2.816403e-11
1.00	1000	3.143301e-04	4.765072e-04
1.05	431	2.652062e-02	3.880614e-02
1.10	231	5.357142e-02	7.568952e-02
1.15	163	8.101821e-02	1.106474e-01
1.20	130	1.093852e-01	1.444877e-01
1.25	112	1.394250e-01	1.781241e-01

We now use this example to test some of the theorems we have derived in Section 4. For the experiments the true value of x^0 is set to be -2.5 and we begin with an initial estimate of -2.3. Observations are generated using the truth at the initial time t_0 and using the discrete numerical model (63) to calculate the 'truth' at time t_1. The time step is set to be $\Delta t = 0.5$. We begin by testing the convergence of the TGN algorithm.

6.1 Truncated Gauss-Newton method - numerical results

For this example the exact Gauss-Newton method is easy to compute, since we have only a scalar equation and so can obtain a direct solution of each inner iteration. In order to test the theory for the truncated Gauss-Newton algorithm we apply an error to this solution by solving the approximate equation

$$(J(x_k^0)^T J(x_k^0))s_k = -J(x_k^0)^T f(x_k^0) + r_k, \tag{68}$$

where on each iteration we select the size of the residual r_k. We choose

$$r_k = \epsilon \left(\frac{\hat{\beta} - |Q(x_k^0)(J^T(x_k^0)J(x_k^0))^{-1}|}{1 + |Q(x_k^0)(J^T(x_k^0)J(x_k^0))^{-1}|} \right) |\nabla \phi(x_k^0)|, \tag{69}$$

with ϵ a specified parameter and $\hat{\beta} = 0.999$. From Theorem 5 we expect the algorithm to converge to the correct solution for values of ϵ less than one. In Table 1 we show the results of the iterative process for various levels of truncation. In the first column we have the value of ϵ chosen for the truncation. The second column shows the number of iterations to convergence. The algorithm is considered to have converged when the difference between two successive iterates is less than 10^{-12} and we restrict the maximum number of iterations to 1000. The third column of the table shows the difference between the iterated solution and the true solution and the fourth column shows the gradient of the objective function at the iterated solution, which should be zero if the true minimum has been reached.

We see that for $\epsilon = 0$ (the exact Gauss-Newton method) the exact solution has been found in 5 iterations. As the value of ϵ is increased, the number of iterations to convergence also increases. Once we reach a value of $\epsilon = 0.95$ then the number of iterations to convergence is 401. However an examination of columns three and four of the table shows that even with this

Table 2: Imperfect observations, exact Jacobian

ϵ	Iterations	Error	Gradient
0.00	10	4.440892e-15	7.778500e-15
0.25	17	9.503509e-14	1.806853e-13
0.50	32	6.181722e-13	1.176347e-12
0.75	66	1.671552e-12	3.180605e-12
0.90	128	4.250822e-12	8.088735e-12
0.95	181	6.231016e-12	1.185694e-11
1.00	359	1.052936e-11	2.003732e-11
1.05	157	6.324736e-02	1.093406e-01
1.10	116	8.697037e-02	1.452842e-01
1.15	93	1.103473e-01	1.783861e-01
1.20	79	1.336149e-01	2.092708e-01
1.25	69	1.570351e-01	2.384890e-01

large amount of truncation, the correct solution is reached. For a value of $\epsilon = 1.0$ we find that the algorithm fails to converge within 1000 iterations. Interestingly for values of ϵ greater than one, the algorithm does converge, but to the wrong solution. If we examine the case of $\epsilon = 1.05$ for example, then we have convergence in 431 iterations. However the final gradient is of the order 10^{-2}, indicating that the true minimum has not been found. Thus from these results it would seem that the bound on the truncation proposed in Theorem 5 is precise. For truncations less than this bound we converge to the correct solution of the original nonlinear least squares problem, but this is not true for truncations above this bound.

We note that for this test case, when we have perfect observations, we have a zero residual problem. In order to test what happens when we do not have a zero residual we add an error of 10% to observation y^0 and subtract an error of 10% from observation y^1. The true solution is calculated by applying the full Newton method to the problem, which gives a fixed point of $x^0 = -2.5938$ in 7 iterations. The accuracy of this solution is tested by substituting it into the gradient equation and ensuring that the gradient is zero. The convergence results for this case are shown in Table 2, where the third column is now the difference between the iterated TGN solution and the solution calculated using the exact Newton method.

We see a very similar pattern of behaviour as for the perfect observation (zero residual) case. For all values of ϵ less than one the truncated Gauss-Newton algorithm converges to the same solution as the exact Newton method, but the number of iterations taken to converge increases as ϵ increases. We also find that for this case the method converges when $\epsilon = 1$. For values of ϵ greater than one, convergence is achieved, but the converged solution is not equal to that found by the Newton method. For these cases the gradient information indicates that a minimum has not been found.

6.2 Perturbed Gauss-Newton method - numerical results

In the application of data assimilation, a perturbed Jacobian may be derived by replacing the linearization of the discrete nonlinear model by a simplified discretization of the linearized continuous model. For this test case we produce a perturbed Jacobian in the same way, following the example of [7]. We note that the linearization of the continuous nonlinear equation (62) is given by

$$\frac{d(\delta z)}{dt} = 2z\delta z. \tag{70}$$

If we apply the second order Runge-Kutta scheme to this equation, we obtain

$$\delta x^{n+1} = (1 + 2x^n\Delta t + 3(x^n)^2\Delta t^2 + 3(x^n)^3\Delta t^3 + \frac{5}{2}(x^n)^4\Delta t^4 + (x^n)^5\Delta t^5)\delta x^n. \tag{71}$$

Thus for the example we obtain the perturbed Jacobian

$$\tilde{J}(x^0) = \begin{pmatrix} 1 \\ 1 + 2x^0\Delta t + 3(x^0)^2\Delta t^2 + 3(x^0)^3\Delta t^3 + \frac{5}{2}(x^0)^4\Delta t^4 + (x^0)^5\Delta t^5 \end{pmatrix}. \tag{72}$$

Using this perturbed Jacobian we apply the PGN algorithm on our example, where on each iteration we confirm that the sufficient condition (27) is satisfied. For this example we find that that the second order terms $\tilde{Q}$ are given by

$$\begin{aligned} \tilde{Q}(x^0) &= (x^0 + (x^0)^2\Delta t + (x^0)^3\Delta t^2 + \frac{1}{2}(x^0)^4\Delta t^3 - y^1) \cdot \\ &\quad (2\Delta t + 6x^0\Delta t^2 + 9(x^0)^2\Delta t^3 + 10(x^0)^3\Delta t^4 + 5(x^0)^4\Delta t^5). \end{aligned} \tag{73}$$

For the case in which we have perfect observations we find that (27) is satisfied on each iteration and the PGN method converges to the true solution in 18 iterations. When error is added on to the observations, as in the previous section, the PGN method converges in 9 iterations and again we find that the condition for convergence is always satisfied. This time the converged solution is not the same as that of the exact Gauss-Newton method. The solution differs from the true solution $x_0 = -2.5$ by approximately 0.01.

In order to examine a case in which the sufficient condition (27) is not satisfied on each iteration, we change the time step to $\Delta t = 0.6$, keeping all other parameters of the problem the same as before. For the case of perfect observations the PGN converges to the correct solution in 23 iterations, compared to 5 iterations for the exact GN and 6 iterations for the Newton method. We find that the condition for convergence is satisfied on each iteration, with the maximum value of the left hand side of (27) reaching 0.994. However, when error is present on the observed values, the convergence condition fails by the second iteration and we find that the PGN fails to converge in 1000 iterations. For this case the exact GN, using the true Jacobian, converges to the correct solution in 8 iterations.

6.3 Truncated Perturbed Gauss-Newton method - numerical results

Finally in this section we consider the case in which the perturbed Gauss-Newton method is also truncated. Following the same method as in the previous two sections, we solve on each iteration the approximate equation

$$\tilde{J}(x_k^0)^T\tilde{J}(x_k^0)s_k = -\tilde{J}(x_k^0)^T f(x_k^0) + r_k, \tag{74}$$

where we choose the residual r_k. We choose

$$r_k = \epsilon \left(\frac{\hat{\beta} - |1 - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T\tilde{J}(x_k))^{-1}|}{|(\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T\tilde{J}(x_k))^{-1}|} \right) |\nabla\phi(x_k^0)|, \tag{75}$$

where ϵ is a specified parameter and $\hat{\beta} = 0.999$.The other data are as before, with errors added to the observations. From Theorem 8 we expect the method to converge for values of $\epsilon < 1$. As previously the true solution is calculated by applying the exact Newton method, but this time to the perturbed problem. This gives a result in 5 iterations of $x^0 = -2.6477$. In Table 3 we

Table 3: Imperfect observations, inexact Jacobian

ϵ	Iterations	Error	Residual
0.00	21	8.215650e-14	6.693951e-14
0.25	33	4.911627e-13	4.007662e-13
0.50	56	1.217249e-12	9.930633e-13
0.75	121	3.732126e-12	3.044658e-12
0.90	306	1.105871e-11	9.021988e-12
0.95	596	2.444178e-11	1.993989e-11
1.00	1000	1.260007e-01	9.382085e-02
1.05	90	1.714365e+00	1.765471e+00
1.10	53	1.842029e+00	1.934063e+00
1.15	36	1.940084e+00	2.069636e+00
1.20	25	2.019233e+00	2.184031e+00
1.25	23	2.085381e+00	2.283791e+00

show the convergence results for the TPGN method using various levels of truncation. The third column now shows the difference between the TPGN solution and the exact Newton method applied to the perturbed problem, and the fourth column gives the residual in the perturbed equation (14). We find that as expected from the theory, the TPGN algorithm converges to the correct solution for values of $\epsilon < 1$. For values of $\epsilon > 1$ the algorithm converges to an incorrect solution. Thus it appears that the bound derived in Theorem 8 is robust.

6.4 Rates of convergence

Finally we test numerically the convergence rates derived in Section 5. We begin by running the numerical example with perfect observations, so that we have a zero residual problem. The convergence rates can be plotted by considering the norms of the residuals on each iteration. If the convergence is of order p then we have

$$\| x_{k+1} - x^* \|_2 = K \| x_k - x^* \|_2^p, \tag{76}$$

for some constant K. Thus a plot of $\log(\| x_{k+1} - x^* \|_2)$ against $\log(\| x_k - x^* \|_2)$ will have slope p. In Figure 1(a) we plot this slope for the case when the exact Jacobian is used. The absolute values of the logarithms are plotted for clarity. Three curves are shown, corresponding to the exact GN method, the TGN method with constant truncation and the TGN with $\beta_k \to 0$. From the theory of Section 5 we expect the rates of the convergence for these cases to be quadratic, linear and superlinear. We find that this is the case. For the exact GN method the slope of the line in the figure is 1.97 and for the TGN it is 0.98. For the case in which the truncation tends to zero the slope is 0.96 at the top of the line, which corresponds to the initial iterations, but steepens to 1.5 at the bottom of the line, demonstrating the super-linear nature of the convergence.

In Figure 1(b) we show the convergence for the PGN method with no truncation and the TPGN method with constant truncation. From our previous theory we expect both of these to show linear convergence. The numerical results show that this is the case, with both lines in the figure having a slope of one.

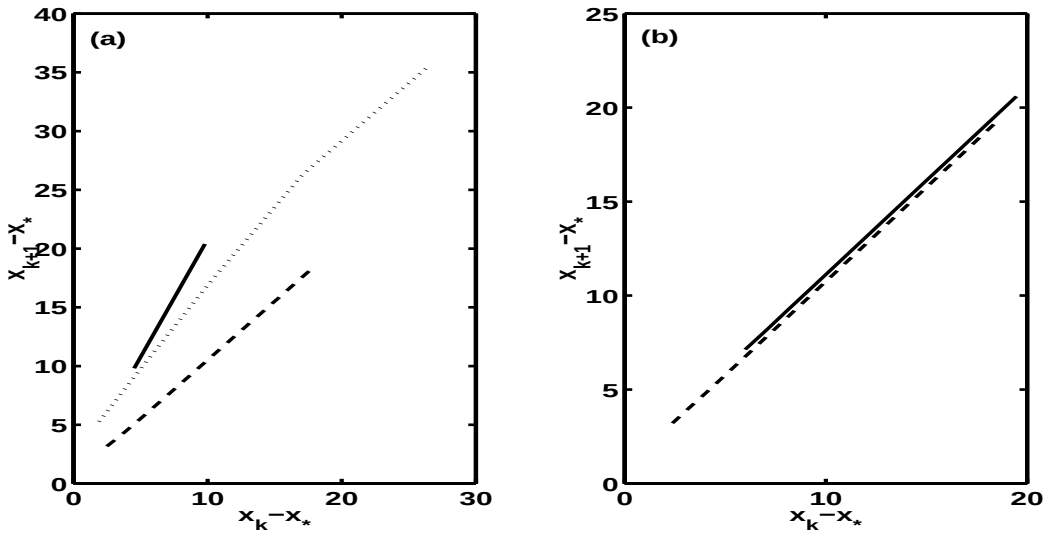

Figure 1: Convergence rates for the cases of (a) exact Jacobian and (b) perturbed Jacobian for the zero residual case. The solid line is for no truncation, the dashed line for constant truncation and the dotted line in plot (a) is for variable truncation.

7 Conclusions

We have described here three approximate Gauss-Newton methods, the truncated, the perturbed and the truncated perturbed Gauss-Newton methods, for solving the nonlinear least squares problem (NLSP). We have derived conditions for the convergence of these approximate methods by treating them as inexact Newton methods, following the theory of [4]. More restricted convergence results, including rates of convergence, have also been derived for the approximate methods by extending the theory of [5] for the exact Gauss-Newton method. In practice, the approximate Gauss-Newton methods are used to treat very large data assimilation problems arising in atmosphere and ocean modelling and prediction. The convergence properties of these algorithms have not previously been investigated. We show by a simple numerical example that the bounds established by the theory are precise, in a certain sense, and that the approximate methods are convergent if the conditions of the theory hold.

References

[1] Åke Björck. *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[2] E. Catinas. Inexact perturbed Newton methods and applications to a class of Krylov solvers. *Journal of Optimization Theory and Applications*, 3:543–571, 2001.

[3] P. Courtier, J.N. Thepaut, A. Hollingsworth. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120:1367–1387, 1994.

[4] R.S. Dembo, S.C. Eisenstat and T. Steihaug. Inexact Newton methods. *SIAM Journal on Numerical Analysis*, 19:400–408, 1982.

[5] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[6] S. Gratton. *Outils théoriques d'analyse du calcul à précision finie*, Institut National Polytechnique de Toulouse, PhD Dissertation, Note: TH/PA/98/30, 1998.

[7] A.S. Lawless. *Development of linear models for data assimilation in numerical weather prediction*, The University of Reading, Department of Mathematics, PhD Thesis, 2001.

[8] A.S. Lawless, S. Gratton, N.K. Nichols. An investigation of incremental 4D-Var using non-tangent linear models. *Quarterly Journal of the Royal Meteorological Society* (to appear).

[9] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[10] A. Ostrowsky. *Solution of Equations and Systems of Equations*, 2nd Edition, Academic Press, New York, 1966.

[11] V. Pereyra. Iterative methods for solving nonlinear least squares problems. *SIAM Journal on Numerical Analysis*, 4:27–36, 1967.

[12] J.S. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*, Springer, New York/Berlin, 1980.

[13] P-Å. Wedin. On the Gauss-Newton method for the nonlinear least-squares problems. Institute for Applied Mathematics, Stockolm, Sweden, Working Paper 24, 1974.