RIACS

Research Institute for Advanced Computer Science
NASA Ames Research Center

Globally Convergent Techniques in Nonlinear Newton-Krylov

Peter N. Brown and Youcef Saad

November 1989

Research Institute for Advanced Computer Science
NASA Ames Research Center

RIACS Technical Report 89.57

NASA Cooperative Agreement Number NCC 2-387

Globally Convergent Techniques in Nonlinear Newton-Krylov

Peter N. Brown and Youcef Saad

November 1989

Research Institute for Advanced Computer Science
NASA Ames Research Center

RIACS Technical Report 89.57

NASA Cooperative Agreement Number NCC 2-387

Globally Convergent Techniques in Nonlinear Newton-Krylov

Peter N. Brown and Youcef Saad

Research Institute for Advanced Computer Science
NASA Ames Research Center - MS: 230-5
Moffett Field, CA 94035

RIACS Technical Report 89.57

November 1989

The Research Institute of Advanced Computer Science is operated by Universities Space Research Association, The American City Building, Suite 311, Columbia, MD 244, (301)730-2656

Work reported herein was supported in part by Cooperative Agreements NCC 2-387 between the National Aeronautics and Space Administration (NASA) and the Universities Space Research Association (USRA).

Globally convergent techniques in nonlinear Newton-Krylov algorithms

Peter N. Brown *
Computing & Mathematics Research Division, L-316
Lawrence Livermore National Laboratory
Livermore, CA 94550

Youcef Saad †
RIACS
MS 230-5, NASA Ames Research Center
Moffett Field, CA 94035

November 27, 1989

Abstract

This paper presents some convergence theory for nonlinear Krylov subspace methods. The basic idea of these methods, which have been described by the authors in an earlier paper, is to use variants of Newton's iteration in conjunction with a Krylov subspace method for solving the Jacobian linear systems. These methods are variants of inexact Newton methods where the approximate Newton direction is taken from a subspace of small dimension. The main focus of this paper is to analyze these methods when they are combined with global strategies such as linesearch techniques and model trust region algorithms. Most of the convergence results are formulated for projection onto general subspaces rather than just Krylov subspaces.

Keywords: Nonlinear systems; Nonlinear Projection methods; Krylov subspace methods; Inexact Newton methods; Trust region techniques; Conjugate gradient methods.
AMS (MOS) Subject Classification: 65H10.

*This work was performed under the auspices of the U.S Department of Energy by the Lawrence Livermore National Laboratory under contract W-7405-Eng-48, and supported by the DOE Office of Energy Research, Applied Mathematical Sciences Research Program.

†Work supported by Cooperative Agreement NCC 2-387 between the National Aeronautics and Space Administration (NASA) and the Universities Space Research Association (USRA).

1 Introduction

In a previous paper [5] we have proposed several basic methods based upon the idea of employing a Newton iteration in which the Jacobian equations are solved approximately by a Krylov subspace method. Several theoretical issues raised in [5] were left unanswered. The purpose of this paper is to fill this gap by laying down the theoretical foundation of nonlinear Krylov subspace methods and by providing convergence results for them. In fact we will not limit ourselves to Krylov subspace methods. Rather, we discuss inexact Newton methods based on general projection techniques.

When defining algorithms for solving nonlinear systems of equations there are two possible options. First, one can use one of the globally convergent modifications of Newton's iteration [11]. The linear systems that arise in the course of the Newton iteration can be solved by either a direct solver or they may be solved approximately by an iterative method. The class of methods based on the latter approach is a particular case of *inexact Newton methods* and several such methods were considered in [1, 3, 4, 5]. Newton's method is essentially a linearization procedure. The mapping F is locally approximated by a linear function and the resulting linear equations are solved to yield the next point. The second approach to solving nonlinear equations does not rely on linearization. Thus, fixed point iterations are inherently nonlinear as are descent methods with accurate line searches. Another well-known example is that of the nonlinear conjugate gradient iteration.

We restrict our attention here to the first approach. In particular, we use linear Krylov methods to solve approximately the Newton equations. The Krylov methods considered are *Arnoldi's Method* [24], and the *Generalized Minimum Residual Method* (GMRES) [25]. In general, these methods have the virtue of requiring virtually no matrix storage, and as such have a distinct advantage over direct methods.

To be more specific, consider finding a solution u_* of the nonlinear system of equations

$$F(u) = 0, \tag{1.1}$$

where F is a nonlinear function from $\mathbf{R}^N$ to $\mathbf{R}^N$. Newton's method applied to (1.1) results in the iteration

1. Set $u_0 =$ an initial guess.
2. For $n = 0, 1, 2, \cdots$ until convergence do:

$$\begin{aligned} &\text{Solve } J(u_n)\delta_n = -F(u_n), \\ &\text{Set } u_{n+1} = u_n + \delta_n, \end{aligned} \tag{1.2}$$

where $J(u_n) = F'(u_n)$ is the system Jacobian. For the problems under consideration, N is large, and as a result the so-called *inexact Newton* methods [9] solve (1.2) only approximately.

One of the main advantages of the Krylov methods is that only the action of the Jacobian matrix J times a vector v is required, not J explicitly. In the current setting, this action can be well approximated by a difference quotient of the form

$$J(u)v \approx \frac{F(u + \sigma v) - F(u)}{\sigma},$$

where u is an approximation to a solution of (1.1), and σ is a scalar. Here, we will address the convergence behavior of the above algorithms when combined with a global linesearch backtracking

strategy or model trust region approach. We should emphasize that our convergence results are not restricted to the use of Krylov subspace methods when solving (1.2). Our theory is formulated in terms of projection techniques wherein the approximation to the linear system (1.2) is taken from a small dimensional subspace.

Inexact Newton algorithms have been studied by several authors in recent years, see for example the references in [1] and the recent report [6]. Several authors have considered using Krylov methods inside a Newton iteration in the context of systems of ordinary differential equations [3, 4, 7, 14]. Steihaug [27] and O'Leary [20] have used the *Conjugate Gradient* (CG) method in the unconstrained optimization of a real-valued function of several variables. Nash [18, 19] has also used a Newton-CG algorithm in unconstrained optimization. Wigton et al. [29], and more recently Kerkhoven and Saad [16] have accelerated nonlinear fixed point iterations of the form $u_{n+1} = M(u_n)$ by applying this approach to solving the nonlinear system of equations $u - M(u) = 0$. Note that as was observed by Chan and Jackson [7], the new system of equations $u - M(u) = 0$ can be viewed as a nonlinearly preconditioned version of the original system of equations.

In Section 2, we review inexact Newton algorithms, and present versions of the Newton-Arnoldi and Newton-GMRES methods. In Section 3 we give a convergence theory for inexact Newton methods when combined with a linesearch backtracking global strategy, and then in Section 4 we present a convergence theory for inexact Newton methods combined with model trust region strategies. In Section 5, we discuss applications of the basic results in the previous two sections to the Newton-Krylov methods, and then make some concluding remarks in Section 6.

2 Newton-Krylov methods

In this section we review some of the basic ideas of *inexact Newton methods* and *Newton-Krylov algorithms*. We begin with a discussion of the relevant results from Dembo, Eisenstat and Steihaug [9] on inexact Newton methods, and then present the two inexact Newton methods we considered in [5], namely the Newton-Arnoldi and Newton-GMRES methods. Note that a Newton-Krylov method is one example of an inexact Newton algorithm.

2.1 Inexact Newton methods

From [9], an *inexact Newton method* for (1.1) has the following general form:

1. Choose u_0 an initial guess for u_*.
2. For $n = 0, 1, \cdots$ until convergence, do:
   - Choose $\eta_n \in [0, 1)$.
   - Find (in some unspecified manner) a vector δ_n satisfying

$$J(u_n)\delta_n = -F(u_n) + r_n, \text{ with } \frac{\|r_n\|}{\|F(u_n)\|} \leq \eta_n \tag{2.1}$$

   - Set $u_{n+1} = u_n + \delta_n$.

The residual r_n represents the amount by which δ_n fails to satisfy the Newton equation (1.2). It is not generally known in advance, being the result of some inner algorithm which produces only an approximate solution to (1.2) (e.g., an iterative method). The forcing sequence $\eta_n \in (0, 1)$ is used to control the level of accuracy. Also, $\|\cdot\|$ represents any norm on $\mathbf{R}^N$.

We will make the following assumptions on F:

$$\begin{cases} \text{There exists a } u_* \in \mathbf{R}^N \text{ with } F(u_*) = 0. \\ F \text{ is } C^1 \text{ in a neighborhood of } u_*. \\ J(u_*) = F'(u_*) \text{ is nonsingular.} \end{cases} \tag{2.2}$$

The next theorem is shown in [9]:

Theorem 2.1 *Assume that F satisfies (2.2), and that $\eta_n \leq \eta_{\max} < t < 1$. There exists $\epsilon > 0$ such that if $\|u_0 - u_*\| \leq \epsilon$, then the sequence of inexact newton iterates $\{u_n\}$ converges to u_*. Moreover, the convergence is linear in the sense that*

$$\|u_{n+1} - u_*\|_* \leq t\|u_n - u_*\|_*,$$

where $\|y\|_* \equiv \|J(u_*)y\|$. If in addition,

$$\eta_n \to 0, \tag{2.3}$$

then the sequence $\{u_n\}$ converges to u_* superlinearly. Also, if F' is Lipschitz continuous near u_* and $\eta_n = O(\|F(u_n)\|)$, then the convergence is quadratic.

In the above theorem, $\|\cdot\|$ again represents any norm on $\mathbf{R}^N$.

For the case when N is large, a δ_n satisfying the residual condition (2.1) is often obtained by using an iterative procedure for the linear system. In [5], we considered using the Arnoldi and GMRES algorithms for nonsymmetric linear systems to obtain δ_n's satisfying the residual condition (2.1). For the convergence theory presented in this paper, the actual method which produces a δ_n satisfying (2.1) will be left unspecified. All that will be required is the existence of such a δ_n. This is easily guaranteed by assuming that J_n is nonsingular for all n.

2.2 Newton-Arnoldi and Newton-GMRES

At each iteration of the inexact Newton algorithm, we must obtain an approximate solution of the linear system (1.2) which we rewrite as

$$J\delta = -F, \tag{2.4}$$

where F and its Jacobian J are evaluated at the current iterate. If $\delta^{(0)}$ is an initial guess for the true solution of (2.4), then letting $\delta = \delta^{(0)} + z$, we have the equivalent system

$$Jz = r^{(0)}, \tag{2.5}$$

where $r^{(0)} = -F - J\delta^{(0)}$ is the initial residual. For a general $N \times N$ matrix A and vector v, define the *Krylov subpsace* $K(A, v, m)$ by

$$K(A, v, m) = \text{span}\{v, Av, \cdots, A^{m-1}v\}.$$

Let K^m denote

$$K^m \equiv K(J, r^{(0)}, m).$$

Arnoldi's method and GMRES both find an approximate solution

$$\delta^{(m)} = \delta^{(0)} + z^{(m)}, \text{with } z^{(m)} \in K^m,$$

such that either

$$(-F - J\delta^{(m)}) \perp K^m \text{ (equivalently } (r^{(0)} - Jz^{(m)}) \perp K^m) \tag{2.6}$$

for Arnoldi's method, or

$$\|F + J\delta^{(m)}\|_2 = \min_{\delta \in \delta^{(0)} + K^m} \|F + J\delta\|_2 \ \ (= \min_{z \in K^m} \|r^{(0)} - Jz\|_2) \tag{2.7}$$

for GMRES. Note that this condition is equivalent to demanding that the residual $r^{(m)} = -F - J\delta^{(m)}$ be orthogonal to JK^m. Here, $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbf{R}^N$ and orthogonality is meant in the usual Euclidean sense.

The following algorithm is a nonlinear version of the Arnoldi (GMRES) algorithm, which at every outer iteration generates an orthonormal system of vectors v_i $(i = 1, 2, \cdots, m)$ of the subspace K^m and then builds the vector $\delta^{(m)}$ that satisfies (2.6) (or (2.7) for GMRES). In both algorithms, v_1 is obtained by normalizing $r^{(0)}$.

Algorithm : Newton-Arnoldi (Newton-GMRES)

1. *Start:* Choose u_0 and compute $F(u_0)$. Set $n = 0$. Choose a tolerance ϵ_0.

2. *Arnoldi process:*

   - For an initial guess $\delta^{(0)}$, form $r^{(0)} = -F - J\delta^{(0)}$, where $F = F(u_n)$ and $J = J(u_n)$.
   - Compute $\beta = \|r^{(0)}\|_2$ and $v_1 = r^{(0)}/\beta$.
   - For $j = 1, 2, \cdots$, do:

     (a) Form Jv_j and orthogonalize it against the previous $v_1, \cdots, v_j$ via

$$\begin{aligned} h_{i,j} &= (Jv_j, v_i), \quad i = 1, 2, \cdots, j, \\ \hat{v}_{j+1} &= Jv_j - \sum_{i=1}^{j} h_{i,j} v_i \\ h_{j+1,j} &= \|\hat{v}_{j+1}\|_2, \quad \text{and} \\ v_{j+1} &= \hat{v}_{j+1}/h_{j+1,j}. \end{aligned} \tag{2.8}$$

     (b) Compute the residual norm $\rho_j = \|F + J\delta^{(j)}\|_2$, of the solution $\delta^{(j)}$ that would be obtained if we stopped at this step.

     (c) If $\rho_j \leq \epsilon_n$ set $m = j$ and go to (3).

3. *Form the approximate solution:*

   Arnoldi: Define H_m to be the $m \times m$ (Hessenberg) matrix whose (possibly) nonzero entries are the coefficients h_{ij}, $1 \leq i \leq j$, $1 \leq j \leq m$ and define $V_m \equiv [v_1, v_2, \cdots, v_m]$.

   - Find the vector y_m which solves the linear system $H_m y = \beta e_1$, where $e_1 = [1, 0, \cdots, 0]^T$.
   - Compute $\delta^{(m)} = \delta^{(0)} + z^{(m)}$, where $z^{(m)} = V_m y_m$, and $u_{n+1} = u_n + \delta^{(m)}$.

   GMRES: Define $\bar{H}_m$ to be the $(m+1) \times m$ (Hessenberg) matrix whose nonzero entries are the coefficients h_{ij}, $1 \leq i \leq j+1$, $1 \leq j \leq m$ and define $V_m \equiv [v_1, v_2, \cdots, v_m]$.

   - Find the vector y_m which minimizes $\|\beta e_1 - \bar{H}_m y\|_2$, where $e_1 = [1, 0, \ldots, 0]^T$, over all vectors y in $\mathbf{R}^m$.
   - Compute $\delta^{(m)} = \delta^{(0)} + z^{(m)}$ where $z^{(m)} = V_m y_m$, and $u_{n+1} = u_n + \delta^{(m)}$.

4. *Stopping test:* If u_{n+1} is determined to be a good enough approximation to a root of (1.1), then stop, else set $u_n \leftarrow u_{n+1}$, $n \leftarrow n + 1$, choose a new tolerance ϵ_n, and go to (2).

Therefore, in both Arnoldi and GMRES the outer iteration is of the form $u_{n+1} = u_n + \delta^{(m)}$ where $\delta^{(m)} = \delta^{(0)} + z^{(m)}$, with

$$z^{(m)} = V_m y_m,$$

and y_m is either the solution of an $m \times m$ linear system, for Arnoldi, or the solution of an $(m+1) \times m$ least squares problem for GMRES.

For simplicity, we have omitted several details of the practical implementation of the above linear and nonlinear methods, which are discussed at length in [5], [24], [4] and [25]. For example, the residual norm ρ_j referred to in step 2 of the algorithms does not require the computation of the approximate solution $\delta^{(j)}$ at every step. Instead an inexpensive formula, which evaluates ρ_j, is updated at each step while the factorization of the Hessenberg matrix H_m or $\bar{H}_m$ is updated.

An important observation that has been very useful in practice is that there is no need to explicitly compute the Jacobian matrix $J(u_n)$. This is due to the fact that the above algorithm only requires the product of this Jacobian times a vector and this can be well approximated by the difference formula:

$$J(u)v \approx \frac{F(u+\sigma v) - F(u)}{\sigma}. \tag{2.9}$$

In [1], Brown has given an analysis of the resulting inexact Newton/finite-difference Krylov algorithms when using (2.9) to approximate $J(u)v$. Sufficient conditions are given in [1] on the size of the $\sigma's$ in the finite-difference versions of Arnoldi and GMRES which guarantee the local convergence of the Newton-Krylov iteration. These results have been extended in [4] to include a finite-difference version of the Conjugate Gradient iteration.

One final aspect worth noting is the ability to use restarting in the linear Krylov methods. Typically, a maximum value of m is dictated by storage considerations. If we let $m_{\max}$ be this value, then it is possible that $m = m_{\max}$ in the Arnoldi process, and yet ρ_m is still greater than ϵ_n. In this case, one can set $\delta^{(0)}$ equal to $\delta^{(m)}$ and restart the Arnoldi process, effectively restarting the Krylov method. The convergence of such a procedure is not always guaranteed, but the idea seems to work well in practice. We note that for lack of a better initial guess we use $\delta^{(0)} = 0$ on the first (and possibly only) pass through the Arnoldi process at each stage of the Newton iteration. It is only when restarting that $\delta^{(0)}$ will be nonzero. We will refer to the restarted algorithms as Arnoldi(m) and GMRES(m), where m is the maximum subspace dimension. As will be seen below, it will also be important to choose the tolerance ϵ_n at each step of the Newton iteration.

3 Global convergence results for linesearch methods

We will be concerned with the convergence properties of the inexact Newton algorithms outlined in the previous section when combined with global strategies. In this section we will analyze the global convergence of inexact Newton algorithms when combined with linesearch backtracking strategies. The results given below are independent of the particular inexact Newton method used.

To begin, let $f(u) = \frac{1}{2}\|F(u)\|_2^2$. An easy calculation gives

$$\nabla f(u) = J(u)^T F(u),$$

where $J(u) = F'(u)$, the Jacobian matrix of F evaluated at u. Typically, convergence of a sequence of iterates $\{u_n\}$ is studied in terms of the scalar sequence

$$\epsilon_n \equiv \nabla f_n^T \frac{\delta_n}{\|\delta_n\|_2} \tag{3.10}$$

where

$$\delta_n \equiv u_{n+1} - u_n \text{ and } \nabla f_n = \nabla f(u_n).$$

When $\lim_{n\to\infty} \epsilon_n = 0$, the sequence u_n will converge to a solution u_* under fairly mild conditions. First, let us assume that the acute angle between δ_n and the gradient ∇f_n is bounded away from $\pi/2$, i.e., that at every step we have

$$\cos\theta(\nabla f_n, \delta_n) \geq \epsilon > 0, \tag{3.11}$$

(where we define $\cos\theta(u,v) = \frac{|u^T v|}{\|u\|_2 \cdot \|v\|_2}$). Then from the definition of ϵ_n in (3.10), the gradient ∇f_n will converge to zero whenever ϵ_n converges to zero.

We now recall the following two important results from Ortega and Rheinboldt [21], pp. 475-476.

Theorem 3.1 *Let $f : \mathbf{R}^N \to \mathbf{R}$ be continuously differentiable on a compact subset $D_0 \subset \mathbf{R}^N$ and suppose that $\{u_n\} \subset D_0$ is any sequence which satisfies $\lim_{n\to\infty} \nabla f(u_n) = 0$. Then the set $\Omega = \{u \in D_0 | \nabla f(u) = 0\}$ of critical points of f is not empty and,*

$$\lim_{n\to\infty} [\inf_{u\in\Omega} \|u - u_n\|] = 0 \tag{3.12}$$

In particular, if Ω consists of a single point u_* then $\lim_{n\to\infty} u_n = u_*$ and $\nabla f(u_*) = 0$.

Theorem 3.2 *Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable on a compact subset $D_0 \subset \mathbf{R}^N$ and suppose that the set $\Omega = \{u \in D_0 | \nabla f(u) = 0\}$ of critical points of f in D_0 is finite. Let $\{u_n\} \subset D_0$ be any sequence for which $\lim_{n\to\infty} \nabla f(u_n) = 0$ and $\lim_{n\to\infty}(u_{n+1} - u_n) = 0$. Then u_n converges to a certain u_* in Ω and $\nabla f(u_*) = 0$.*

Thus, we will often attempt to establish conditions under which ϵ_n as defined by (3.10) converges to zero and for which (3.11) holds.

To guarantee that the current iterate will make progress towards the solution in one step of the algorithm we must know that the inexact Newton step δ is a descent direction for f at the current approximation u. A *descent direction* p at u is one for which there exists a $\lambda_0 > 0$ such

that $f(u+\lambda p) < f(u)$ for all $\lambda < \lambda_0$. As is well-known, when f is differentiable this is equivalent to the condition that

$$\nabla f(u)^T p < 0,$$

where $\nabla f(u) = (\frac{\partial f}{\partial u_1}(u), \cdots, \frac{\partial f}{\partial u_N}(u))^T$. As noted above, $\nabla f(u) = J(u)^T F(u)$, and so p is a descent direction for f at u if

$$F(u)^T J(u) p < 0.$$

If $\bar{\delta}$ is an approximate solution of the Newton equations

$$J\delta = -F,$$

with $F = F(u)$ and $J = J(u)$, then

$$F^T J\bar{\delta} = -F^T F - F^T \bar{r}, \tag{3.13}$$

where $\bar{r} = -F - J\bar{\delta}$ is the residual associated with $\bar{\delta}$. Thus, $\bar{\delta}$ will be a descent direction for f at u whenever $|F^T \bar{r}| < F^T F$. In particular, if $\|\bar{r}\|_2 < \|F\|_2$, then $\bar{\delta}$ is a descent direction. This result was also given in [5] and is restated in the following proposition.

Proposition 3.3 *A sufficient condition for $p \in \mathbf{R}^N$ to be a descent direction for f at u is that*

$$\|F(u) + J(u)p\|_2 < \|F(u)\|_2. \tag{3.14}$$

As was seen earlier, it is also important to be able to guarantee that the angle between the gradient of f and the step δ_n is bounded away from $\pi/2$. The next lemma gives a lower bound for $|\epsilon_n|$ under an additional assumption on the step direction p.

Lemma 3.4 *Let $F : \mathbf{R}^N \to \mathbf{R}^N$ be continuously differentiable on $\mathbf{R}^N$. Let $u \in \mathbf{R}^N$ be given with $F(u) \neq 0$ and $J(u) = F'(u)$ nonsingular. Consider $p \in \mathbf{R}^N$ satisfying*

$$\|F(u) + J(u)p\|_2 \leq \eta \|F(u)\|_2,$$

with $\eta \in (0,1)$. Then

$$\frac{|\nabla f(u)^T p|}{\|p\|_2} \geq \frac{1-\eta}{(1+\eta)M} \|\nabla f(u)\|_2 > 0, \tag{3.15}$$

where $M = cond_2(J(u))$ and $f(u) \equiv \frac{1}{2} F(u)^T F(u)$.

Proof: For notational convenience, let $F = F(u)$, $f = f(u)$, $\nabla f = \nabla f(u)$, and $J = J(u) \equiv F'(u)$. Note that $F \neq 0$ implies $\nabla f \neq 0$. Let r be the residual associated with p so that $r = F + Jp$. Then $\|r\|_2 \leq \eta \|F\|_2$ and $p = -J^{-1}(F - r)$. So,

$$\begin{aligned} \nabla f^T p &= (J^T F)^T (-J^{-1}(F - r)) \\ &= -F^T F + F^T r. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{|\nabla f^T p|}{\|p\|_2} &= \frac{|F^T F - F^T r|}{\|J^{-1}(F - r)\|_2} \\ &\geq \frac{F^T F - |F^T r|}{\|J^{-1}(F - r)\|_2}. \end{aligned}$$

$\eta\|F\|_2$ implies $|F^T r| \leq \eta\|F\|_2^2$, which then gives

$$F^T F - |F^T r| \geq (1-\eta)\|F\|_2^2.$$

$$\begin{aligned} \|J^{-1}(F-r)\|_2 &\leq \|J^{-1}\|_2 \cdot \|F\|_2 + \|J^{-1}r\|_2 \\ &\leq (1+\eta)\|J^{-1}\|_2 \cdot \|F\|_2. \end{aligned} \tag{3.16}$$

Thus,

$$\frac{|\nabla f^T p|}{\|p\|_2} \geq \frac{(1-\eta)F^T F}{(1+\eta)\|J^{-1}\|_2 \cdot \|F\|_2} = \frac{(1-\eta)\|F\|_2}{(1+\eta)\|J^{-1}\|_2} \tag{3.17}$$

and as a result, using the fact that $\|\nabla f\|_2 = \|J^T F\|_2 \leq \|J\|_2\|F\|_2$, we get

$$\frac{|\nabla f^T p|}{\|\nabla f\|_2 \cdot \|p\|_2} \geq \frac{(1-\eta)}{(1+\eta)M}, \tag{3.18}$$

where $M = \text{cond}_2(J)$. □

Condition (3.15) can also be recast as

$$\cos\theta(\nabla f, p) \geq \frac{1-\eta}{(1+\eta)M}. \tag{3.19}$$

At every step of the inexact Newton method, we require that a condition of the form

$$\|F(u_n) + J(u_n)p_n\|_2 \leq \eta_n\|F(u_n)\|_2 \tag{3.20}$$

$$\eta_n \leq \eta < 1 \tag{3.21}$$

holds, and if we assume that the condition numbers $M_n = \text{cond}_2(J(u_n))$ are bounded from above by M, then (3.15) shows that

$$\cos\theta(p_n, \nabla f(u_n)) \geq \frac{1}{M}\frac{1-\eta}{1+\eta}. \tag{3.22}$$

This implies that a sufficient condition to guarantee both p_n being a descent direction and the validity of relation (3.11) is that the residual condition (3.20)-(3.21) holds.

A simple consequence of the above lemma which will be useful in the section on trust region techniques, is that the residual norm assumption (3.20)-(3.21) implies that the cosine of the angle between the gradient and the Krylov subspace is bounded from below. More specifically,

Corollary 3.5 *Let $F : \mathbf{R}^N \to \mathbf{R}^N$ be continuously differentiable on $\mathbf{R}^N$. Let $u \in \mathbf{R}^N$ be given with $F(u) \neq 0$ and $J(u) = F'(u)$ nonsingular. Consider the subspace $K = span\{V\}$ where the columns of V form an orthonormal set of vectors, and assume that there exists one vector p in K satisfying*

$$\|F(u) + J(u)p\|_2 \leq \eta\|F(u)\|_2, \tag{3.23}$$

with $\eta \in (0,1)$. Then

$$\|V^T \nabla f(u)\|_2 \geq \frac{1-\eta}{(1+\eta)M}\|\nabla f(u)\|_2, \tag{3.24}$$

where $M = cond_2(J(u))$ and $f(u) \equiv \frac{1}{2}F(u)^T F(u)$.

Proof: Let $p_0 = Vy_0$ be a vector of K that satisfies (3.23). We then have from the lemma

$$\frac{|\nabla f(u)^T V y_0|}{\|y_0\|_2} \geq \frac{1-\eta}{(1+\eta)M}\|\nabla f(u)\|_2, \tag{3.25}$$

where we have used the fact that $\|p_0\|_2 = \|Vy_0\|_2 = \|y_0\|_2$. Using the Cauchy-Schwartz inequality,

$$\|V^T \nabla f(u)\|_2 \geq \frac{|\nabla f(u)^T V y_0|}{\|y_0\|_2} \geq \frac{1-\eta}{(1+\eta)M}\|\nabla f(u)\|_2, \tag{3.26}$$

which proves the result. □

3.1 Convergence of inexact Newton sequences

In this subsection we consider the case of linearly converging inexact Newton sequences. That is, the sequence $\{\eta_n\}$ in the inexact Newton method is only required to satisfy $\eta_n \leq \eta_{\max} < t < 1$. Superlinearly converging inexact Newton sequences will be examined in the following subsection.

An important condition to guarantee global convergence, is the so-called α-condition in the Armijo and Goldstein principle [11, 21] wherein δ_n must satisfy

$$f(u_n + \delta_n) < f(u_n) + \alpha \nabla f(u_n)^T \delta_n. \tag{3.27}$$

We can show the following remarkably simple result if we require that the direction δ_n solve the linear system $J(u_n)\delta = -F(u_n)$ with a certain accuracy.

Theorem 3.6 *Let $f \equiv \frac{1}{2}\|F\|_2^2$ be given, where F is differentiable, and α, η two scalars such that $0 < \alpha < \frac{1}{2}$, $0 \leq \eta < 1$. Assume that the iterates u_n are defined by $u_{n+1} = u_n + \delta_n$ where δ_n satisfies (3.27) and*

$$\|F(u_n) + J(u_n)\delta_n\|_2 \leq \eta \|F(u_n)\|_2. \tag{3.28}$$

Then

$$\lim_{n\to\infty} f(u_n) = 0.$$

Proof: In this proof we let $J_n \equiv J(u_n)$, $F_n \equiv F(u_n)$. From the condition (3.27) and the expression for the gradient of f we get

$$f(u_{n+1}) < f(u_n) + \alpha \nabla f_n^T \delta_n = f(u_n) + \alpha F_n^T J_n \delta_n \tag{3.29}$$

Writing $J_n\delta_n = -F_n + r_n$ this gives

$$\begin{aligned} f(u_{n+1}) &< f(u_n) + \alpha F_n^T(-F_n + r_n) \\ &= (1-2\alpha) f(u_n) + \alpha F_n^T r_n \\ &\leq (1-2\alpha) f(u_n) + \alpha \|F_n\|_2 \|r_n\|_2. \end{aligned}$$

From (3.28) we have $\|r_n\|_2 \leq \eta \|F_n\|_2$ which yields the following inequality,

$$f(u_{n+1}) < (1-2\alpha) f(u_n) + 2\alpha\eta f(u_n) = [(1-2\alpha) + 2\alpha\eta] f(u_n) \tag{3.30}$$

Notice that the scalar in the brackets is a convex combination of 1 and η and is therefore always less than one under the conditions on α and η. The result follows immediately. □

Note that we have made virtually no assumption on the function F apart from differentiability, and so the result is very general. However, we cannot guarantee in general that one can indeed select a vector δ_n that satisfies condition (3.27) and (3.28) at the same time, but we do know that near a solution u_* for which $J(u_*)$ is nonsingular, a sufficiently good approximation to the Newton step will satisfy these two conditions simultaneously.

A more explicit result extending the above theorem is now shown. For this next theorem we assume that a general backtracking strategy is used. This means that the next iterate is of the form $u_n + \lambda p_n$, where p_n is any descent direction and λ is selected by the procedure described below. In the procedure the two parameters $\theta_{\min}, \theta_{\max}$ are such that $0 < \theta_{\min} \leq \theta_{\max} < 1$, a typical choice being $\theta_{\min} = \theta_{\max} = 1/2$. The procedure requires another parameter $\epsilon^* > 0$ which is used to essentially rescale the *starting step* in the process in order to prevent it from from being too small.

Algorithm 3.1: General Backtracking Procedure

1. Set $\lambda = \max\{1, \epsilon^* \frac{|\nabla f(u_n)^T p_n|}{\|p_n\|_2^2}\}$.
2. If $f(u_n + \lambda p_n) \leq f(u_n) + \alpha\lambda \nabla f(u_n)^T p_n$, then set $\lambda_n = \lambda$, and exit. Else:
3. Choose $\hat{\lambda} \in [\theta_{\min}\lambda, \theta_{\max}\lambda]$; set $\lambda \leftarrow \hat{\lambda}$. Go to (2).

As is shown next, the sequence is well defined in that under a mild condition on the gradient of f the procedure will deliver a nonzero λ_n in a finite number of steps. Moreover, the resulting λ_n can be bounded from below.

Lemma 3.7 *Let f be differentiable and assume that its gradient is such that*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \gamma \|x - y\|_2, \text{ for all } x, y \in \mathbf{R}^N. \tag{3.31}$$

Let $\alpha < 1$ and p_n be any descent direction. Then Algorithm 3.1 will produce an iterate $u_{n+1} = u_n + \lambda_n p_n$ in a finite number of backtracking steps and λ_n satisfies the inequality

$$\lambda_n \|p_n\|_2 \geq -\frac{\nabla f(u_n)^T p_n}{\|p_n\|_2} \min\{\epsilon^* \ , \ \frac{(1-\alpha)}{\gamma}\theta_{\min}\}. \tag{3.32}$$

Proof: The subscript n is dropped from this proof. Using the mean value theorem we have the equality:

$$f(u + \lambda p) = f(u) + \lambda \nabla f(u + \theta\lambda p)^T p, \tag{3.33}$$

where $0 \leq \theta \leq 1$. We rewrite the above equation as

$$\begin{aligned} f(u + \lambda p) &= f(u) + \lambda \nabla f(u)^T p + \lambda[\nabla f(u + \theta\lambda p)^T p - \nabla f(u)^T p] \\ &= f(u) + \alpha\lambda \nabla f(u)^T p + \lambda[(1-\alpha)\nabla f(u)^T p + (\nabla f(u + \theta\lambda p)^T p - \nabla f(u)^T p)] \\ &= f(u) + \alpha\lambda \nabla f(u)^T p + \lambda[(1-\alpha)\nabla f(u)^T p + \lambda\|p_n\|_2 \zeta], \end{aligned} \tag{3.34}$$

where for convenience we have set

$$\zeta \equiv \frac{\nabla f(u + \theta\lambda p)^T p - \nabla f(u)^T p}{\lambda \|p\|_2}.$$

Note that from the assumptions we have

$$|\zeta| = |\left(\nabla f(u+\theta\lambda p) - \nabla f(u)\right)^T \frac{p}{\lambda\|p\|_2}| \leq \gamma\theta\|p\|_2 \leq \gamma\|p\|_2. \tag{3.35}$$

If the test in step 2 is passed at the first step, then the first λ is accepted and, in this situation, the inequality (3.32) is trivially satisfied. If the test in step 2 fails for the first step, then λ is reduced according to the rule in step 3. Moreover, after a finite number of reductions, the term in brackets in the right-hand-side of (3.34) will become negative and the corresponding λ will be accepted. This will occur as soon as $\lambda\gamma\|p\|_2^2 < -(1-\alpha)\nabla f(u)^T p$. The first λ which is accepted will be such that

$$\lambda\|p\|_2 \geq -\theta_{\min}\frac{(1-\alpha)}{\gamma}\frac{\nabla f(u)^T p}{\|p\|_2},$$

and the inequality (3.32) is again satisfied. □

We note that the usual β condition of Armijo and Goldstein also guarantees that a lower bound on λ similar to (3.32) is satisfied. Indeed, the relation (3.34) is still valid with α replaced by β. Moreover, the β condition:

$$f(u+\lambda p) \geq f(u) + \beta\lambda\nabla f(u)^T p$$

implies that

$$(1-\beta)\nabla f(u_n)^T p_n + \lambda_n\|p_n\|_2\zeta \geq 0.$$

With the inequality (3.35) this immediately yields

$$\lambda_n\|p_n\|_2 \geq -\frac{(1-\beta)\nabla f(u_n)^T p_n}{\gamma\|p_n\|_2}. \tag{3.36}$$

Both (3.32) and (3.36) imply that the step length from u_n is bounded from below with respect to $\nabla f(u_n)^T p_n/\|p_n\|_2$.

We emphasize the importance of the initial λ in the procedure. There is no reason why one should always start the process with $\lambda = 1$ since $\|p_n\|_2$ can be arbitrarily small. As was explained before the choice of the initial λ in step 1 is essentially equivalent to a rescaling of the vector p_n. If we always start with $\lambda = 1$ and p_n happens to be very small at every step then the test in step 2 may be passed immediately and there is a danger that δ_n becomes too small for the iterates to make any progress towards the solution.

Note, however, that if p_n solves the linear system $Jp = -F$ approximately, then we may have additional information that will ensure that $\|p_n\|_2$ is bounded from below. Indeed, the following lemma shown by Walker, [28] is just one such result.

Lemma 3.8 *Suppose that $J(u)p = -F(u) + r$ and $\|r\|_2 \leq \eta\|F(u)\|_2$, with $0 \leq \eta < 1$. Then,*

$$\|p\|_2 \geq \frac{(1-\eta)}{\|J(u)\|_2^2}\frac{\nabla f(u)^T p}{\|p\|_2}. \tag{3.37}$$

Proof: We have

$$|\nabla f(u)^T p| = |F(u)^T J(u)p| \leq \|F(u)\|_2\|J(u)\|_2\|p\|_2. \tag{3.38}$$

Moreover, from $\|F(u)\|_2 = \|r - J(u)p\|_2 \leq \eta\|F(u)\|_2 + \|J(U)\|_2\|p\|_2$ we get

$$\|F(u)\|_2 \leq \frac{\|J(u)\|_2}{(1-\eta)}\|p\|_2. \tag{3.39}$$

The result follows from combining (3.38) and (3.39). □

A consequence of the above lemma is that the backtracking procedure will always start with $\lambda = 1$ in the first step if ϵ^* is small enough, or to be more accurate as long as

$$\epsilon^* \leq \frac{(1-\eta)}{\|J(u)\|_2^2}. \tag{3.40}$$

This may provide for a rational way of choosing ϵ^* since $\|J(u)\|_2$ may often be roughly estimated in the course of the algorithm.

We can now prove the following theorem.

Theorem 3.9 *Let $f \equiv \frac{1}{2}\|F\|_2^2$ satisfy the conditions of the previous Lemma and let p_n be such that $\|F_n + J_n p_n\|_2 \leq \eta\|F_n\|_2$ for all n, with $\eta < 1$. Further, let each iterate be chosen by Algorithm 3.1. Then, either*

$$\lim_{n\to\infty} f(u_n) = 0 \tag{3.41}$$

or

$$\lim_{n\to\infty} \|p_n\|_2 = \infty. \tag{3.42}$$

Proof: Letting as before $r = F(u_n) + J(u_n)p_n$, and dropping the subscript n we have

$$\nabla f^T p = F^T(-F + r) \leq -\|F\|_2^2(1-\eta) = -2(1-\eta)f, \tag{3.43}$$

and as a result, (with $u_{n+1} = u_n + \lambda_n p_n = u + \lambda p$)

$$f(u_{n+1}) \leq f(u) + \lambda\alpha\nabla f^T p \leq f(u) - 2\lambda\alpha(1-\eta)f(u) = f(u)[1 - 2\lambda\alpha(1-\eta)]. \tag{3.44}$$

From the result of the Lemma we have

$$-\lambda\|p\|_2 \leq \kappa\frac{\nabla f^T p}{\|p\|_2} \tag{3.45}$$

where we define

$$\kappa \equiv \min\{\epsilon^*\ ,\ \frac{(1-\alpha)}{\gamma}\theta_{\min}\}$$

and therefore, (3.44) becomes

$$f(u_{n+1}) \leq f(u)\left[1 + 2\alpha\kappa(1-\eta)\frac{\nabla f^T p}{\|p\|_2^2}\right]. \tag{3.46}$$

Denoting by t_n the quantity $\nabla f_n^T p_n / \|p_n\|_2^2$ and by c the constant $2\alpha\kappa(1-\eta)$ this relation can be rewritten as

$$f(u_{n+1}) \leq f(u_n)[1 + ct_n]. \tag{3.47}$$

Since $f(u_n)$ is bounded from below and nonincreasing, it converges to a certain limit ϕ. If this limit is zero the result of the theorem holds. If it is different from zero then by dividing both members of equation (3.47) by $f(u_n)$, we see that $1 + ct_n$ which is bounded from above by 1 and from below by a sequence converging to 1, has 1 as its limit. Equivalently, t_n converges to 0. Going back to the relation (3.43) which we rewrite as $2f(u_n)(1-\eta) \leq |t_n| \cdot \|p_n\|_2^2$, we see immediately that in this situation we must have $\|p_n\|_2 \to \infty$. □

We mention that Eisenstat and Walker [12] have recently established an extension to this result. More precisely, they show that in addition to the conclusion of the above theorem, one of the following holds:

(i) $\lim_{n\to\infty} \|u_n\|_2 = \infty$

(ii) The sequence u_n has finite limit-points, and F' is singular at each of them.

(iii) The sequence u_n has a limit point u_* such that $F(u_*) = 0$.

We can show a result that is more explicit than that of Theorem 3.9 if we make a few additional assumptions on $J(u_n)$.

Corollary 3.10 *Let $f \equiv \frac{1}{2}\|F\|_2^2$ satisfy the conditions of the previous Lemma and let p_n be such that $\|F_n + J(u_n)p_n\|_2 \leq \eta\|F_n\|_2$ for each n, with $\eta < 1$. Further, let each iterate be chosen by Algorithm 3.1, and assume that $J(u_n)^{-1}$ exists and its norm is bounded from above for all n. Then*

$$\lim_{n\to\infty} f(u_n) = 0 \tag{3.48}$$

Proof: From the relation (3.16), and the fact that $J(u_n)^{-1}$ is bounded from above, the norm of the vector $p_n = J(u_n)^{-1}(F(u_n) - r)$ is bounded from above. Therefore, from the previous theorem, we must have $\lim_{n\to\infty} f(u_n) = 0$. □

The following additional results do not require the use of the backtracking procedure described in Algorithm 3.1. They are based upon the ideas presented by Dennis and Schnabel [11]. Given the current Newton iterate $u = u_n$ and a descent direction p, we want to take a step in the direction of p that yields an acceptable u_{n+1}. We will define a step $\delta = \lambda p$ to be *acceptable* if both of the Goldstein-Armijo [11] conditions are met, namely

$$f(u + \lambda p) \leq f(u) + \alpha\lambda\nabla f(u)^T p, \tag{3.49}$$

and

$$f(u + \lambda p) \geq f(u) + \beta\lambda\nabla f(u)^T p, \tag{3.50}$$

for given scalars α, β satisfying $0 < \alpha < \beta < 1$. Again, we will refer to these two conditions as the α- and β-conditions, respectively. For a given descent direction p, the next result shows that there exist points $u + \lambda p$ satisfying (3.49) and (3.50).

Theorem 3.11 *Let $f : \mathbf{R}^N \to \mathbf{R}$ be continuously differentiable on $\mathbf{R}^N$ with $f(z) \geq 0$ for all $z \in \mathbf{R}^N$. Let $u, p \in \mathbf{R}^N$ be such that $\nabla f(u)^T p < 0$. Then given $0 < \alpha < \beta < 1$, there exist $\lambda_u > \lambda_\ell > 0$ such that $u + \lambda p$ satisfies* (3.49) *and* (3.50) *for any $\lambda \in (\lambda_\ell, \lambda_u)$.*

This is essentially Theorem 6.3.2, page 120, in Dennis and Schnabel [11], and so the proof is omitted.

Theorem 3.12 *Let $f : \mathbf{R}^N \to \mathbf{R}$ be continuously differentiable on $\mathbf{R}^N$ with $f(z) \geq 0$ for all $z \in \mathbf{R}^N$, and assume there exists a constant $\gamma \geq 0$ such that*

$$\|\nabla f(z) - \nabla f(u)\|_2 \leq \gamma \|z - u\|_2 \tag{3.51}$$

for every $u, z \in \mathbf{R}^N$. Then given any $u_0 \in \mathbf{R}^N$, there exists a sequence $\{u_n\}$ $(n = 0, 1, \cdots)$ satisfying conditions (3.49) *and* (3.50)*, and either*

$$\nabla f(u_n)^T \delta_n < 0$$

or

$$\nabla f(u_n) = 0 \text{ and } \delta_n = 0,$$

for each $n > 0$, where $\delta_n \equiv u_{n+1} - u_n$. Furthermore, for any such sequence, either

$$\text{(a)} \quad \nabla f(u_n) = 0 \text{ for some } n \geq 0, \text{ or}$$

$$\text{(b)} \quad \lim_{n \to \infty} \frac{\nabla f(u_n)^T \delta_n}{\|\delta_n\|_2} = 0.$$

Proof: This is essentially Theorem 6.3.3 in [11] (p. 122), except that condition (3.50) is slightly different and f is assumed to be bounded from below. For each n, if $\nabla f(u_n) = 0$, then (a) holds and the sequence is constant from then on. If $\nabla f(u_n) \neq 0$, then there exists a p_n such that $\nabla f(u_n)^T p_n < 0$ (e.g., take $p_n = -\nabla f(u_n)$). By Theorem 3.11, there exists $\lambda_n > 0$ such that $u_{n+1} = u_n + \lambda_n p_n$ satisfies (3.49) and (3.50). Let $\delta_n = \lambda_n p_n$. We must now show that if no term of $\{\delta_n\}$ is zero, then (b) must hold.

First, define $\omega_n \equiv \|\delta_n\|_2$ and

$$\sigma_n \equiv \frac{\nabla f(u_n)^T \delta_n}{\omega_n}.$$

By (3.49) and $\omega_i \sigma_i < 0$ for every i, we have for any $j > 0$,

$$\begin{aligned} f(u_j) - f(u_0) &= \sum_{n=0}^{j-1} f(u_{n+1}) - f(u_n) \\ &\leq \sum_{i=0}^{j-1} \alpha \nabla f(u_i)^T \delta_i \\ &= \alpha \sum_{i=0}^{j-1} \omega_i \sigma_i < 0. \end{aligned}$$

Hence, $f \geq 0$ on $\mathbf{R}^N$ implies that the series

$$\sum_{i=0}^{\infty} \omega_i \sigma_i < \infty.$$

Thus, $\omega_n \sigma_n \to 0$ as $n \to \infty$. To conclude that $\sigma_n \to 0$ we must use condition (3.50) whose purpose was to guarantee that the steps do not get too small.

By the Mean Value Theorem, there exists a $\bar{\lambda} \in (0, \lambda_n)$ such that

$$f(u_{n+1}) = f(u_n) + \nabla f(u_n + \bar{\lambda} p_n)^T (u_{n+1} - u_n) \tag{3.52}$$

which, using condition (3.50), yields,

$$f(u_{n+1}) - f(u_n) = \nabla f(u_n + \bar{\lambda} p_n)^T (u_{n+1} - u_n) \geq \beta \nabla f(u_n)^T \delta_n. \tag{3.53}$$

This implies that

$$\left[\nabla f(u_n + \bar{\lambda} p_n) - \nabla f(u_n)\right]^T \delta_n \geq (\beta - 1) \nabla f(u_n)^T \delta_n > 0.$$

Therefore,

$$\begin{aligned} 0 < (\beta - 1)\omega_n \sigma_n &\leq \omega_n \|\nabla f(u_n + \bar{\lambda} p_n) - \nabla f(u_n)\|_2 \\ &\leq \gamma \bar{\lambda} \|p_n\|_2 \omega_n \leq \gamma \omega_n^2. \end{aligned}$$

So,

$$\omega_n \geq \frac{\beta - 1}{\gamma} \sigma_n > 0$$

and

$$\omega_n \sigma_n \leq \frac{\beta - 1}{\gamma} \sigma_n^2 < 0.$$

Hence, $\omega_n \sigma_n \to 0$ implies $\sigma_n \to 0$ as $n \to \infty$. □

Note that $\sigma_n \equiv \nabla f(u_n)^T \delta_n / \omega_n \to 0$ does not imply that $\nabla f(u_n) \to 0$ as $n \to \infty$. However, it will as long as the angle between $\nabla f(u_n)$ and δ_n is bounded away from 90°. It is possible to guarantee this is the case in the inexact Newton setting as was shown by Lemma 3.4. Note also that $\nabla f_n \to 0$ does not imply $F(u_n) \to 0$, without some additional assumptions, e.g., as in Corollary 3.10.

If conclusion (b) holds in Theorem 3.12 with $\|F(u_n) + J(u_n)\delta_n\|_2 \leq \eta \|F(u_n)\|_2$ for all n, where $\eta \in (0, 1)$ and if we assume that the condition numbers $M_n = \text{cond}_2(J(u_n))$ are uniformly bounded from above, then (3.15) shows that $\sigma_n \to 0$ as $n \to \infty$ does imply $\nabla f(u_n) \to 0$. However, the conditions are too weak to imply that $\{u_n\}$ converges.

We should also point point out that the conclusions of the above theorem hold for a sequence generated by Algorithm 3.1.

3.2 Superlinear convergence of inexact Newton sequences

In this subsection we will require that $\eta_n \to 0$ as $n \to \infty$. As noted in Theorem 2.1, given that the sequence of inexact Newton iterates converges, this additional assumption on the η_n's implies that the convergence is at least superlinear. The main result of this subsection is a modification of a theorem obtained by Dennis and Moré [10], and shows that the global strategy based on the above $\alpha-$ and $\beta-$ conditions will permit full inexact Newton steps when close to a minimizer of f, provided that $\alpha < \frac{1}{2}$ and $\beta > \frac{1}{2}$.

Theorem 3.13 *Let $F : \mathbf{R}^N \to \mathbf{R}^N$ be twice continuously differentiable in an open convex set $D \subset \mathbf{R}^N$, and for $f \equiv \frac{1}{2} F^T F$ assume $\nabla^2 f \in Lip_\gamma(D)$. Consider the sequence $\{u_n\}$ generated by $u_{n+1} = u_n + \lambda_n p_n$, where $\|F(u_n) + J(u_n)p_n\|_2 \leq \eta_n \|F(u_n)\|_2$ for all n with $0 < \eta_n \leq \eta < 1$ for all n, and λ_n chosen so that (3.49) and (3.50) hold with $\alpha < \frac{1}{2}$ and $\beta > \frac{1}{2}$.*

If $u_n \to u_* \in D$ with $J(u_*)$ nonsingular, then $F(u_*) = 0$. If in addition, $\eta_n \to 0$ as $n \to \infty$, then there exists an $n_0 \geq 0$ such that for all $n \geq n_0$, $\lambda_n = 1$ is admissible (i.e., satisfies conditions (3.49) and (3.50)). Furthermore, if $\lambda_n = 1$ for all $n \geq n_0$, then $u_n \to u_*$ superlinearly. If also $\eta_n = O(\|F(u_n)\|_2)$, then the convergence is quadratic.

Proof: Since $f(u) = \frac{1}{2}F(u)^T F(u) \geq 0$, and since $\|F_n + J_n p_n\|_2 \leq \eta \|F_n\|_2$ (where $F_n = F(u_n)$, $J_n = J(u_n)$, etc.) implies $\nabla f_n^T p_n < 0$ for all n, with $\eta \in (0,1)$, we have by Theorem 3.12 that

$$\lim_{n\to\infty} \frac{\nabla f_n^T p_n}{\|p_n\|_2} = 0. \tag{3.54}$$

If $u_n \to u_* \in D$ with $J_* = J(u_*)$ nonsingular, then by continuity $M_n = \text{cond}_2(J_n) \to M_* = \text{cond}_2(J_*)$, and so the sequence $\{M_n\}$ is uniformly bounded from above. Thus, the discussion after the proof of Lemma 3.4 implies that $\nabla f(u_n) \to 0$ as $n \to \infty$. Hence, again by continuity

$$0 = \nabla f(u_*) = J_*^T F_*,$$

which gives $F(u_*) = 0$.

Next, we show that $\|p_n\|_2 \to 0$ as $n \to \infty$. Since

$$p_n = -J_n^{-1} F_n + J_n^{-1}(F_n + J_n p_n),$$

we have

$$\begin{aligned} \|p_n\|_2 &\leq \|J_n^{-1}\|_2 \cdot \|F_n\|_2 + \|J_n^{-1}\|_2 \cdot \|F_n + J_n p_n\|_2 \\ &\leq (1+\eta)\|J_n^{-1}\|_2 \cdot \|F_n\|_2 \end{aligned} \tag{3.55}$$

Thus, $\|F_n\|_2 \to 0$ implies $\|p_n\|_2 \to 0$ as $n \to \infty$. Also,

$$\|\nabla f_n\|_2 = \|J_n^T F_n\|_2 \geq \|J_n^T\|_2^{-1} \cdot \|F_n\|_2 = \|J_n\|_2^{-1} \cdot \|F_n\|_2.$$

Hence, by Lemma 3.4 we have

$$\frac{-\nabla f_n^T p_n}{\|p_n\|_2} \geq \frac{1-\eta}{(1+\eta)M_n} \|J_n\|_2^{-1} \cdot \|F_n\|_2,$$

where $M_n = \text{cond}_2(J_n)$. Thus,

$$\|F_n\|_2 \leq \frac{(1+\eta)M_n}{1-\eta} \|J_n\|_2 \cdot \frac{-\nabla f_n^T p_n}{\|p_n\|_2},$$

and so

$$\|F_n\|_2 \cdot \|p_n\|_2 \leq \frac{(1+\eta)M_n}{1-\eta} \|J_n\|_2 \cdot (-\nabla f_n^T p_n) = -a_n \nabla f_n^T p_n, \tag{3.56}$$

letting $a_n = \frac{1+\eta}{1-\eta} M_n \cdot \|J_n\|_2$. In addition, it immediately follows from (3.55) that

$$\|p_n\|_2^2 \leq \frac{(1+\eta)^2}{1-\eta} M_n^2 (-\nabla f_n^T p_n) = -b_n \nabla f_n^T p_n, \tag{3.57}$$

letting $b_n = \frac{(1+\eta)^2}{1-\eta} M_n^2$.

Using (3.56) and (3.57), and the fact that $\|p_n\|_2 \to 0$, we next show that $\lambda_n = 1$ satisfies conditions (3.49) and (3.50) for n large. First note that if $F = (F_1, \cdots, F_N)^T$, then

$$\begin{aligned} \nabla^2 f(u) &= J(u)^T J(u) + \sum_{i=1}^{N} F_i(u) \nabla^2 F_i(u) \\ &\equiv J(u)^T J(u) + S(u). \end{aligned}$$

Also note that $\|S(u_n)\|_2 \to 0$ as $n \to \infty$ since $u_n \to u_*$ and $F(u_*) = 0$. For each n, by the mean value theorem there exists a $\bar{u}_n$ on the line segment between u_n and $u_n + p_n$ such that

$$f(u_n + p_n) - f(u_n) - \frac{1}{2}\nabla f(u_n)^T p_n = \frac{1}{2}\left(\nabla f_n + \nabla^2 f(\bar{u}_n)p_n\right)^T p_n.$$

This then gives

$$\begin{aligned}
\left|f(u_n + p_n) - f(u_n) - \frac{1}{2}\nabla f(u_n)^T p_n\right| &= \frac{1}{2}\left|\left(\nabla f_n + \nabla^2 f(\bar{u}_n)p_n\right)^T p_n\right| \\
&= \left|\frac{1}{2}\left(\nabla f_n + \nabla^2 f_n p_n\right)^T p_n + \frac{1}{2}p_n^T\left(\nabla^2 f(\bar{u}_n) - \nabla^2 f_n\right)p_n\right| \\
&\le \frac{1}{2}\left(\|J_n^T(F_n + J_n p_n)\|_2 \cdot \|p_n\|_2 + (\|S_n\|_2 + \gamma\|p_n\|_2)\|p_n\|_2^2\right) \\
&\le \frac{1}{2}\left(\eta_n\|J_n\|_2 \cdot \|F_n\|_2 \cdot \|p_n\|_2 + (\|S_n\|_2 + \gamma\|p_n\|_2)\|p_n\|_2^2\right) \\
&\le -\frac{1}{2}\left(a_n\eta_n\|J_n\|_2 + b_n(\|S_n\|_2 + \gamma\|p_n\|_2)\right)\nabla f_n^T p_n \\
&= -\frac{1}{2}\epsilon_n \nabla f_n^T p_n,
\end{aligned}$$

where we have used (3.56) and (3.57), and defined $\epsilon_n \equiv a_n\eta_n\|J_n\|_2 + b_n(\|S_n\|_2 + \gamma\|p_n\|_2)$. Therefore,

$$\frac{1}{2}(1 + \epsilon_n)\nabla f_n^T p_n \le f(u_n + p_n) - f_n \le \frac{1}{2}(1 - \epsilon_n)\nabla f_n^T p_n.$$

Next, note that since a_n, b_n and $\|J_n\|_2$ are all bounded from above, and since η_n, $\|S_n\|_2$ and $\|p_n\|_2$ all converge to 0 as $n \to \infty$, we have that $\epsilon_n \to 0$. So, choose $n_0 \ge 0$ so that for all $n \ge n_0$ we have

$$\epsilon_n \le \min\{1 - 2\alpha, 2\beta - 1\}.$$

It then follows that for all $n \ge n_0$

$$\beta\nabla f_n^T p_n \le f(u_n + p_n) - f_n \le \alpha\nabla f_n^T p_n.$$

Thus, $\lambda_n = 1$ is admissible for all $n \ge n_0$.The superlinear (quadratic) convergence of the sequence follows from Corollary 3.5 in [9] or Theorem 2.1 above. □

One can relax the condition that $\eta_n \to 0$ in the above theorem somewhat, although the resulting condition on η is not computationally feasible in general.

Corollary 3.14 *Let $F : \mathbf{R}^N \to \mathbf{R}^N$, and let $f \equiv \frac{1}{2}F^T F$ be twice continuously differentiable in an open convex set $D \subset \mathbf{R}^N$ with $\nabla^2 f \in Lip_\gamma(D)$. Assume that $M = \sup_{u\in D}\{cond_2(J(u))\} < \infty$ and that $K = \sup_{u\in D}\{\|J(u)\|_2\} < \infty$. Consider the sequence $\{u_n\}$ generated by $u_{n+1} = u_n + \lambda_n p_n$, where $\|F(u_n) + J(u_n)p_n\|_2 \le \eta\|F(u_n)\|_2$ for all n with $0 < \eta < 1$ for all n, and λ_n chosen so that (3.49) and (3.50) hold with $\alpha < \frac{1}{2}$ and $\beta > \frac{1}{2}$.*

If $u_n \to u_* \in D$ with $J(u_*)$ nonsingular, then $F(u_*) = 0$. If in addition, η satisfies

$$\eta\frac{1+\eta}{1-\eta} \cdot M \cdot K < \min\{1 - 2\alpha, 2\beta - 1\}, \tag{3.58}$$

then there exists an $n_0 \ge 0$ such that for all $n \ge n_0$, $\lambda_n = 1$ is admissible (i.e., satisfies conditions (3.49) and (3.50)). Furthermore, if $\lambda_n = 1$ for all $n \ge n_0$, then $u_n \to u_*$ linearly.

Proof: From the proof of Theorem 3.13, the corollary will be true if we have

$$a_n\eta\|J_n\|_2 < \min\{1-2\alpha, 2\beta-1\},$$

for all n. But this follows immediately from the definitions of M and K, and condition (3.58). □

One may wonder whether or not the conditions required in the theorem are too strong if we want to ensure that $\lambda_n = 1$ is admissible. More precisely, does the weaker condition $\|F_n + J_n p_n\|_2 \leq \eta\|F_n\|_2$ for all n, where $\eta \in (0,1)$, allow the existence of sequences $u_{n+1} = u_n + \lambda_n p_n$ converging to a u_* for which $\lambda_n = 1$ is admissible for all large n. The answer is no as is illustrated in the following example.

Example: Consider the one-dimensional function $F(u) = u \in \mathbf{R}$. Choose $\eta \in (0,1)$ and $0 < \alpha < \frac{1}{2} < \beta < 1$ so that

$$\max\{1-2\alpha, 2\beta-1\} < \eta.$$

Choose θ so that $0 < \theta < 1$ and

$$\max\{1-2\alpha, 2\beta-1\} < 1-\theta < \eta.$$

Consider a sequence $\{u_n\}$ generated by $u_{n+1} = u_n + \lambda_n p_n$, where $u_0 = 1$, $p_n = -\theta u_n$, and $\lambda_n = (2-\alpha-\beta)/\theta$ for all n. Then $f(u) = \frac{1}{2}u^2$ with $\nabla f(u_n)^T p_n = -\theta u_n^2$, and so p_n is a descent direction for all n. Also, $\|F_n + J_n p_n\|_2 = |(1-\theta)u_n| \leq \eta\|F(u_n)\|_2 = \eta|u_n|$ for all n. We next show that λ_n is admissble for all n. We have

$$f(u_n + \lambda p_n) = \frac{1}{2}(u_n + \lambda_n p_n)^2 = \frac{1}{2}u_n^2 - \lambda\theta u_n^2 + \frac{1}{2}\lambda^2\theta^2 u_n^2.$$

Next,

$$f(u_n) + \lambda\alpha\nabla f(u_n)^T p_n = \frac{1}{2}u_n^2 + \lambda\alpha(-\theta u_n^2),$$

and

$$f(u_n) + \lambda\beta\nabla f(u_n)^T p_n = \frac{1}{2}u_n^2 + \lambda\beta(-\theta u_n^2).$$

Thus, λ is admissible if

$$\frac{1}{2}u_n^2 + \lambda\beta(-\theta u_n^2) < \frac{1}{2}u_n^2 - \lambda\theta u_n^2 + \frac{1}{2}\lambda^2\theta^2 u_n^2 < \frac{1}{2}u_n^2 + \lambda\alpha(-\theta u_n^2),$$

or if

$$\frac{2(1-\beta)}{\theta} < \lambda < \frac{2(1-\alpha)}{\theta}.$$

Clearly, the λ_n defined above is admissble for all n. However, for the parameters given above we have $1 < \frac{2(1-\beta)}{\theta}$, and so $\lambda = 1$ can never satisfy conditions (3.49) and (3.50). Notice that $u_{n+1} = (\alpha+\beta-1)u_n$, and so $u_n = (\alpha+\beta-1)^n u_0$ with $|\alpha+\beta-1| < 1$. Hence, $u_n \to 0$ as $n \to 0$ linearly.

Note that the convergence results of this section are similar to others in the literature. However the emphasis was put on the additional residual norm condition (3.28). As was seen, slightly different, and somewhat stronger, results can be shown in the situation where δ_n satisfies a residual norm condition.

For the purpose of illustration, we end this section by describing the particular backtracking algorithm we have used in [5]. The selection procedure for λ is modelled after that in [11].

Algorithm 3.2: Backtracking Procedure #2

1. Choose $\alpha \in (0, \frac{1}{2})$ and $\beta \in (\frac{1}{2}, 1)$.
2. Given u_n the current Newton iterate, find in some unspecified manner p with $\|F_n + J_n p\|_2 \leq \eta_n \|F_n\|_2$.
3. Find an acceptable new iterate $u_{n+1} = u_n + \lambda p$. First, set $\lambda = 1$. Define $u(\lambda) = u_n + \lambda p$.
   a. If $u(\lambda)$ satisfies (3.49) and (3.50), then exit. If not, then continue.
   b. If $u(\lambda)$ satisfies (3.49), but not (3.50), and $\lambda \geq 1$, set $\lambda \leftarrow 2\lambda$ and go to (a).
   c. If $u(\lambda)$ satisfies (3.49) only and $\lambda < 1$, or $u(\lambda)$ does not satisfy (3.49) and $\lambda > 1$, then
      c.1 If $\lambda < 1$, define $\lambda_{lo} = \lambda$ and λ_{hi} = last previously attempted value of λ. If $\lambda > 1$, define λ_{lo} = last previously attempted value of λ and $\lambda_{hi} = \lambda$. In both cases, $u(\lambda_{lo})$ satisfies (3.49) but not (3.50), $u(\lambda_{hi})$ does not satisfy (3.49), and $\lambda_{lo} < \lambda_{hi}$.
      c.2 Find $\lambda \in (\lambda_{lo}, \lambda_{hi})$ such that $u(\lambda)$ satisfies (3.49) and (3.50) using successive linear interpolation.
   d. Otherwise ($u(\lambda)$ does not satisfy (3.49) and $\lambda \leq 1$), decrease λ by a factor between 0.1 and 0.5 as follows:
      d.1 Select the new λ such that $u(\lambda)$ is the minimizer of the one dimensional quadratic interpolant passing through $f(u_n)$, $f'(u_n) = \nabla f(u_n)^T p$ and $f(u_n + \lambda p)$. Then take the maximum of this new λ and 0.1 as the actual value used. (One can show theoretically that the new λ value so chosen will be less than or equal to one-half the previous value.)
      d.2 Go to step (b).

4 Global convergence results for model trust region techniques

In [5], we presented a trust region algorithm based on a Newton-GMRES iteration. We give here a convergence theory for this algorithm and other methods based on projection principles. We start by describing in Section 4.1 general trust region methods and give some background on their theory. Our basic approach will be modelled after the work of Schultz, Schnabel and Byrd [26]. In Section 4.2 we will adapt this theory to the particular case that is of interest to us, namely the case in which a projection method onto a lower dimensional subspace is used.

4.1 General trust region techniques for nonlinear optimization

The model trust region algorithm generates a sequence of points u_n, and at the nth stage of the iteration a quadratic model of $f : \mathbf{R}^N \rightarrow \mathbf{R}$ near the current iterate u_n is used which has the form

$$\psi_n(w) = f_n + g_n^T w + \frac{1}{2} w^T B_n w,$$

where $f_n = f(u_n)$, $g_n = \nabla f(u_n)$, and $B_n \approx \nabla^2 f(u_n)$. (A projected analogue of the above function based on approximations from the subspace K will be developed later.) At this stage an initial value for the trust region size τ_n is also available. An inner iteration is then performed which consists of using the current trust region size τ_n and the information contained in the quadratic model to compute a step

$$p_n(\tau_n) = p(g_n, B_n, \tau_n).$$

Then a comparison of the actual reduction of the objective function

$$\text{ared}_n(\tau_n) = f_n - f(u_n + p_n(\tau_n))$$

and the reduction predicted by the quadratic model

$$\text{pred}_n(\tau_n) = f_n - \psi_n(p_n(\tau_n)),$$

is performed. If there is a satisfactory reduction, then the step can be taken, or a possibly larger trust region used. If not, then the trust region is reduced and the inner iteration is repeated.

For now we leave unspecified what algorithm is used to form the step computing function $p(g, B, \tau)$, and how the trust region size or radius τ_n is changed. We also leave unspecified the selection of B_n except to restrict it to be symmetric positive definite. Details on these options will be given later. Shultz, Schnabel and Byrd [26] describe an abstract trust region algorithm as follows:

Algorithm 4.1: General Trust Region Algorithm

1. Choose $\gamma_1, \alpha_1, \alpha_2 \in (0, 1)$, $u_1 \in \mathbf{R}^N$, $\tau_1 > 0$, and let $n = 1$.
2. Compute $f_n = f(u_n)$, $g_n = g(u_n) \equiv \nabla f(u_n)$, and $B_n \in \mathbf{R}^{N \times N}$ symmetric and positive definite.
3. Find τ_n and compute $p_n = p_n(\tau_n)$ satisfying: $\|p_n\|_2 \leq \tau_n$ and

(a) $\frac{\text{ared}_n(\tau_n)}{\text{pred}_n(\tau_n)} \geq \alpha_1$ and

(b) either $\tau_n \geq \tau_{n-1}$, or

- $\tau_n \geq \|B_{n-1}^{-1} g_{n-1}\|_2$, or
- for some $\tau \leq \tau_n/\gamma_1$, $\frac{\text{ared}_n(\tau)}{\text{pred}_n(\tau)} < \alpha_2$ or $\frac{\text{ared}_{n-1}(\tau)}{\text{pred}_{n-1}(\tau)} < \alpha_2$.

4. Let $u_{n+1} = u_n + p_n$ and $n = n + 1$.

5. Go to Step 2.

The conditions that the step selection function $p(g, B, \tau)$ must satisfy will now be considered. Defining

$$\text{pred}(g, B, \tau) = -g^T p(g, B, \tau) - \frac{1}{2} p(g, B, \tau)^T B p(g, B, \tau),$$

the conditions that we consider are:

Condition 1 There exist two scalars $\bar{c}_1, \sigma_1 > 0$ such that for all $g \in \mathbf{R}^N$, for all symmetric positive definite $B \in \mathbf{R}^{N \times N}$, and for all $\tau > 0$, $\text{pred}(g, B, \tau) \geq \bar{c}_1 \|g\|_2 \min\{\tau, \sigma_1 \|g\|_2 / \|B\|_2\}$.

Condition 2 If B is symmetric positive definite and $\|B^{-1} g\|_2 \leq \tau$, then $p(g, B, \tau) = -B^{-1} g$.

The first condition requires that the predicted decrease be at least as large as a given multiple of the minimun decrease that would be provided by a quadratic search along the steepest descent direction. The second condition forces the direction p to be equal to the Newton direction whenever the next point $u_n + p$ lies in the trust region. The following result is given in [26].

Theorem 4.1 *Let $f : \mathbf{R}^N \to \mathbf{R}$ be twice continuously differentiable and bounded below, and let $\nabla^2 f(u)$ satisfy $\|\nabla^2 f(u)\|_2 \leq \beta_1$ for all $u \in \mathbf{R}^N$. Suppose that an algorithm satisfying the conditions of Algorithm 4.1 above is applied to $f(u)$, starting from some $u_1 \in \mathbf{R}^N$, generating a sequence $\{u_n\}$, $n = 1, 2, \cdots$. Then*

(i) If $p(g, B, \tau)$ satisfies Condition 1 with $\|B_n\|_2 \leq \beta_2$ for all n, then $g_n \to 0$.

(ii) If $p(g, B, \tau)$ satisfies Conditions 1 and 2, $B_n = \nabla^2 f(u_n)$ for all n, $\nabla^2 f(u)$ is Lipschitz continuous with constant L, and u_* is a limit point of $\{u_n\}$ with $\nabla^2 f(u_*)$ positive definite, then u_n converges to u_* q-quadratically.

The first result in the theorem gives the first order stationary point convergence of the sequence of iterates while the first and second results taken together give second order stationary point convergence.

We next discuss the procedures normally used to form the step computing function $p(g, B, \tau)$. One way in which this can be done is to take $p(g, B, \tau)$ to be the solution of the minimization problem

$$\min_{\|w\|_2 \leq \tau} \psi(w), \text{ where } \psi(w) = f + g^T w + \frac{1}{2} w^T B w.$$

Assuming B is symmetric and positive definite, the solution to this problem is given by

$$\bar{p}(\tau) = \begin{cases} -(B + \mu I)^{-1} g, & \text{with } \|(B + \mu I)^{-1} g\|_2 = \tau \text{ when } \|B^{-1} g\|_2 > \tau, \text{ and} \\ -B^{-1} g, & \text{when } \|B^{-1} g\|_2 \leq \tau. \end{cases}$$

We note that the μ in the first part of the definition of $\bar{p}$ is unique. It is well-known that for $p(g, B, \tau) = \bar{p}(\tau)$ the following inequality is true:

$$f - \psi(\bar{p}(\tau)) \geq \frac{1}{2}\|g\|_2 \min\{\tau, \frac{\|g\|_2}{\|B\|_2}\}.$$

For a proof of this result see [17]. With this ideal choice of the direction $p(g, B, \tau)$, Condition 1 is trivially satisfied.

Since there is no finite method of determining μ such that $\|(B + \mu I)^{-1}g\|_2 = \tau$ when $\tau < \|B^{-1}g\|_2$, frequently a piecewise linear approximation to $\bar{p}(\mu)$ is used. The *dogleg strategy* of Powell [22] is an example of such a procedure. (See [11] for a discussion of this and other dogleg strategies.) If we denote Powell's dogleg solution by $\hat{p}(\tau)$, then it is also well-known that for $p(g, B, \tau) = \hat{p}(\tau)$ the following lemma is true.

Lemma 4.2 *Let $\hat{p}(\tau)$ be the dogleg solution to the minimization problem*

$$\min_{\|w\|_2 \leq \tau} \psi(w), \tag{4.59}$$

where $\psi(w) = f + g^T w + \frac{1}{2}w^T B w$ with B symmetric positive definite. Then

$$f - \psi(\hat{p}(\tau)) \geq \frac{1}{2}\|g\|_2 \min\{\tau, \frac{\|g\|_2}{\|B\|_2}\}. \tag{4.60}$$

For a proof of this lemma see Powell [23]. Again, a consequence of the above lemma is that Condition 1 is trivially satisfied, and as a result of Theorem 4.1 global convergence will take place under the mild condition that $\|B_n\|_2$ remains bounded.

4.2 Application to projection methods for nonlinear equations

In the context of nonlinear equations, one typically bases the global strategy on the related function $f(u) \equiv \frac{1}{2}F(u)^T F(u)$. Letting u be the current approximate solution, the local quadratic model now has the form

$$\psi(w) = f + g^T w + \frac{1}{2}w^T B w,$$

where $f = f(u)$, $g = \nabla f(u) = J(u)^T F(u)$, and B approximates $J(u)^T J(u)$. Note that $\nabla^2 f(u) = J(u)^T J(u) + \sum_{i=1}^{N} F_i(u)\nabla^2 F_i(u)$, and so in general $\nabla^2 f(u) = J(u)^T J(u)$ only when $F(u) = 0$. When using projection methods to solve the nonlinear system, the full quadratic model $\psi(w)$ is replaced by a quadratic model on a lower dimensional subspace K. Letting the columns of the $N \times m$ matrix V form an orthonormal basis for K (with m the dimension of K), we have

$$\begin{aligned}
\phi(y) &\equiv \psi(Vy) \\
&= f + g^T V y + \frac{1}{2}(Vy)^T B V y \\
&= f + (V^T g)^T y + \frac{1}{2}y^T (V^T B V) y.
\end{aligned}$$

Thus, $\nabla\phi(0) = V^T g$. Note that since V has orthonormal columns, the matrix $V^T BV$ is symmetric positive definite whenever B is. In the current setting, we will take $B = J(u)^T J(u)$. If $\hat{q}(\tau)$ is the dogleg solution to the minimization problem

$$\min_{\|y\|_2 \leq \tau} \phi(y),$$

then Lemma 4.2 implies

$$f - \phi(\hat{q}(\tau)) \geq \frac{1}{2}\|V^T g\|_2 \min\{\tau, \frac{\|V^T g\|_2}{\|V^T BV\|_2}\}.$$

We then have

$$f - \phi(\hat{q}(\tau)) \geq \frac{1}{2}\|V^T g\|_2 \min\{\tau, \frac{\|V^T g\|_2}{\|B\|_2}\}, \tag{4.61}$$

using the fact that $\|V^T BV\|_2 \leq \|B\|_2$.

In order to be able to apply the results of Shultz, Schnabel and Byrd [26], we must convert the lower bound in the above inequality to one involving $\|g\|_2$, and not $\|V^T g\|_2$. As indicated above, we have $f = \frac{1}{2}F^T F$, and so $\psi(w)$ has the form

$$\psi(w) = f + (J^T F)^T w + \frac{1}{2} w^T (J^T J) w,$$

which gives $g = J^T F$ and $B = J^T J$. Thus, we need a lower bound for $\|V^T g\|_2 = \|V^T J^T F\|_2$. Such a lower bound was derived in Section 3. Indeed, Corollary 3.5 states that

$$\|V^T \nabla f\|_2 = \|V^T g\|_2 \geq \frac{1-\eta}{(1+\eta)M}\|g\|_2, \tag{4.62}$$

provided that there exists at least one vector p in K such that

$$\|F(u) + J(u)p\|_2 \leq \eta \|F(u)\|_2. \tag{4.63}$$

Therefore, we immediately have the following lemma.

Lemma 4.3 *Let $J \in \mathbf{R}^{N\times N}$ be nonsingular and $F \in \mathbf{R}^N$ be given. Let $\eta \in [0,1)$ be chosen and let K be a subspace of dimension m in $\mathbf{R}^N$ such that*

$$\min_{p\in K} \|F + Jp\|_2 \leq \eta \|F\|_2.$$

Choose the $N \times m$ matrix V so that its columns form an orthonormal basis for K. Let $\hat{q}(\tau)$ be the dogleg solution to

$$\min_{\|y\|_2 \leq \tau} \phi(y) \text{ where } \phi(y) = f + (V^T g)^T y + \frac{1}{2} y^T (V^T BV) y,$$

with $f = \frac{1}{2}F^T F$, $g = \nabla f = J^T F$ and $B = J^T J$. Then

$$\|V^T g\|_2 \geq \sigma \|g\|_2, \tag{4.64}$$

and

$$f - \phi(\hat{q}(\tau)) \geq \frac{1}{2}\sigma\|g\|_2 \min\{\tau, \sigma\frac{\|g\|_2}{\|B\|_2}\}, \tag{4.65}$$

where

$$\sigma \equiv \frac{1-\eta}{(1+\eta)M}. \tag{4.66}$$

Given this lemma it is now possible to state a model trust region algorithm appropriate for use with a general projection method. The algorithm is stated in terms of a sequence of general subspaces K_n.

Algorithm 4.2: Inexact Newton Trust Region Algorithm

1. Choose an $\eta_{\max} \in (0,1)$.
2. Choose $\gamma_1, \alpha_1, \alpha_2 \in (0,1)$, $u_1 \in \mathbf{R}^N$, $\tau_1 > 0$, and let $n = 1$.
3. Compute F_n, J_n and choose $\eta_n \in [0, \eta_{\max})$. Then choose a subspace $K_n \subset \mathbf{R}^N$ satisfying
   - $\min_{p \in K_n} \|F_n + J_n p\|_2 \leq \eta_n \|F_n\|_2$.

   Let m_n be the dimension of K_n, and build $V_n \in \mathbf{R}^{N \times m_n}$ whose columns form an orthonormal basis for K_n.
4. Compute $f_n = \frac{1}{2}F_n^T F_n$, $V_n^T g_n = V_n^T J_n^T F_n$, and $\bar{B}_n = V_n^T J_n^T J_n V_n$ with J_n nonsingular.
5. Find τ_n and compute $p_n = V_n q_n = V_n q_n(\tau_n)$ satisfying: $\|p_n\|_2 \leq \tau_n$ and
   (a) $\frac{\text{ared}_n(\tau_n)}{\text{pred}_n(\tau_n)} \geq \alpha_1$ and
   (b) either $\tau_n \geq \tau_{n-1}$, or
      - $\tau_n \geq \|\bar{B}_{n-1}^{-1} V_{n-1}^T g_{n-1}\|_2$, or
      - for some $\tau \leq \tau_n/\gamma_1$, $\frac{\text{ared}_n(\tau)}{\text{pred}_n(\tau)} < \alpha_2$ or $\frac{\text{ared}_{n-1}(\tau)}{\text{pred}_{n-1}(\tau)} < \alpha_2$.
6. Let $u_{n+1} = u_n + p_n$ and $n = n + 1$.
7. Go to Step 3.

In the above algorithm, the step selection function $q_n(V_n^T g_n, \bar{B}_n, \tau_n)$ is given by

$$q_n(V_n^T g_n, \bar{B}_n, \tau_n) = \hat{q}_n(\tau_n)$$

where $\hat{q}_n(\tau_n)$ is the dogleg solution to the minimization problem

$$\min_{\|y\|_2 \leq \tau_n} \phi_n(y) \text{ with } \phi_n(y) \equiv \psi_n(V_n y) = f_n + (V_n^T g_n)^T y + \frac{1}{2} y^T \bar{B}_n y.$$

Then by Lemma 4.3 we have

$$f_n - \phi_n(q_n) \geq \frac{1}{2}\sigma_n\|g_n\|_2 \min\{\tau_n, \frac{\sigma_n\|g_n\|_2}{\|B_n\|_2}\}, \tag{4.67}$$

where

$$\sigma_n = \frac{1-\eta_n}{(1+\eta_n)M_n}$$

with $M_n = \text{cond}_2(J_n)$. We now state the main result of this section.

Theorem 4.4 *Let $F : \mathbf{R}^N \to \mathbf{R}^N$ be twice continuously differentiable. Define $f = \frac{1}{2}F^TF$ and assume $\|\nabla^2 f(u)\|_2 \leq \beta_1$ for all $u \in \mathbf{R}^N$. Suppose that an algorithm satisfying the conditions of Algorithm 4.2 above is applied to $f(u)$, starting from some $u_1 \in \mathbf{R}^N$, generating a sequence $\{u_n\}$, and assume that $\|\bar{B}_n\|_2 \leq \beta_2$ and $cond_2(J_n) \leq M$ for all n. Then*

(i) $g_n \to 0$.

(ii) If $\nabla^2 f(u)$ is Lipschitz continuous with constant L, and u_* is a limit point of $\{u_n\}$ with $J(u_*)$ nonsingular, then $F(u_*) = 0$. If in addition, the sequence $\eta_n \to 0$ as $n \to \infty$, then u_n converges to u_* superlinearly. Also, if $\eta_n = O(\|F(u_n)\|_2)$, then the convergence is quadratic.

This theorem is a direct adaptation of Theorem 4.1 above. The proof uses the previous lemma, and due to its length is deferred to an appendix.

The actual dogleg algorithm we have used in [5] is modelled after that in [11], and is described below. The condition for accepting u_{n+1} is the α-condition in §3, namely

$$f(u_n + \delta) \leq f(u_n) + \alpha \nabla f(u_n)^T \delta,$$

where $\delta = V\hat{q}$ and the columns of V form an orthonormal basis for the Krylov subspace K. The vector $\hat{q}$ will be that point on the dogleg curve for $\phi(y)$ such that $\|\hat{q}\|_2 = \tau$, where τ is the current trust region radius. The algorithm is then as follows.

Algorithm: Dogleg

1. Choose $\alpha \in (0, \frac{1}{2})$.
2. Given u_n, the current Newton iterate, calculate $\delta^{GM} = Vy_{GM}$. Here, y_{GM} is calculated using the GMRES method (without restarting) with initial guess $\delta^{(0)} = 0$, and it is assumed m is large enough so that $\|F_n + J_n\delta^{GM}\|_2 \leq \eta_n\|F_n\|_2$.
3. Given τ, the current trust region size, calculate $\hat{q}$, the point on the dogleg curve for $\phi(y)$ with $\|\hat{q}\|_2 = \tau$. Then calculate $u_{n+1} = u_n + V\hat{q}$. If u_{n+1} is acceptable, then go to step(5).
4. If u_{n+1} is not acceptable, then do one of the following:
   (a) If τ has been doubled during this iteration, then set u_{n+1} equal to its last accepted value and set $\tau \leftarrow \tau/2$. Then continue to the next Newton iteration. If not:
   (b) Determine a new τ by using the minimizer of the one dimensional quadratic interpolating $f(u_n)$, $f(u_{n+1})$, and the directional derivative of f at u_n in the direction $\delta = V\hat{q}$. Letting λ be the value for which $u_n + \lambda\delta$ is this minimizer, set $\tau \leftarrow \lambda\|\delta\|_2$, but constraining it to be between 0.1 and 0.5 of the old τ. Then go to Step 3.
5. For an acceptable u_{n+1}, calculate $\text{ared}_n(\tau)$ and $\text{pred}_n(\tau)$. Then do one of the following:

(a) If $\text{ared}_n(\tau)$ and $\text{pred}_n(\tau)$ agree to within relative error 0.1, and τ has not been decreased during this iteration, set $\tau \leftarrow 2 * \tau$, and go to step (3). If not:

(b) If $\text{ared}_n(\tau) < 0.1 * \text{pred}_n(\tau)$ set $\tau = \tau/2$, or if $\text{ared}_n(\tau) > 0.75 * \text{pred}_n(\tau)$, set $\tau = 2 * \tau$. Otherwise, do not change τ. Then continue to the next Newton iteration.

Note that the α-condition is equivalent to the condition in step 5(a) of Algorithm 4.2, and that the conditions for decreasing the size of the trust region in 5(b) are met. For more details on permissible trust region updating strategies, see §3 in [26].

5 Applications to Newton-Krylov methods

In this section we show how the theory outlined in the previous sections applies to the Newton-Krylov methods discussed in §2. The details of the implementation of the Newton-Krylov algorithms when combined with either a linesearch backtracking strategy or model trust region are discussed at length in our earlier paper [5], and so we will not go into them here. As was seen above, the most important condition to guarantee convergence is that the residual norm be reduced by a certain amount. This was crucial in both the linesearch methods and the trust region methods. Unfortunately, as we indicate below, there is always the possibility of stagnation when using either linear Arnoldi or GMRES, and as a result there remains the possibility of a breakdown in the nonlinear iteration. In these situations the basic residual condition (2.1) may not be satisfied using a single subspace. We begin this section by giving sufficient conditions under which stagnation of the linear iteration never occurs.

The simplest condition is simply to ensure that the steepest descent direction $-g = -J^T F$ belongs to the subspace K. Then the minimum of $\|F + Jp\|_2$ for $p \in K$, is reduced from its value of $\|F\|_2$ when $p = 0$ to an amount not less than that obtained by a steepest descent step, i.e.,

$$\min_{p \in K} \|F + Jp\|_2 \leq \min_{\lambda \in \mathbf{R}} \|F + \lambda J J^T F\|_2 \leq \frac{\mathrm{cond}_2(J)^2 - 1}{\mathrm{cond}_2(J)^2 + 1} \|F\|_2. \tag{5.1}$$

As a result, if the steepest descent direction is known, or computable, it suffices to add it to the subspace to guarantee the existence of an $\eta \in (0, 1)$ for which the residual condition is always satisfied. The resulting modification of the underlying algorithms is very simple.

A particular case of the above situation is when $v_1 = \pm F / \|F\|_2$ and the Jacobian J is symmetric. Then the Krylov subspace K^m will in fact contain the steepest descent direction for the function $f = \frac{1}{2} F^T F$ for any $m \geq 2$. This is because

$$\nabla f = J^T F = JF = \pm \|F\|_2 J v_1$$

which, apart from a scaling factor, is the second vector of the Krylov sequence. As a result, stagnation can be avoided if J is symmetric and a Krylov subspace of dimension $m \geq 2$ is used. Analogous arguments also guarantee that the steepest descent direction lies in K whenever J is skew-symmetric or, whenever $J = I + \alpha S$, with S skew-symmetric, and α real. More generally, if there is a polynomial q of degree m such that $J^T = q(J)$ then again the steepest descent direction lies in K^m. However, this is equivalent to the property that J is normal, see [13] for details.

A third case for which there is no difficulty is when the Jacobian at every point is positive real, i.e., $J + J^T$ is symmetric positive definite. It has been shown by several authors that, in this situation, linear Krylov subspace methods will converge. Thus, if one required the Jacobian matrix $J(u)$ to be positive definite for all $u \in \mathbf{R}^N$, then the residual conditions in the main results of the previous sections can be satisfied, and so convergence of the sequence of nonlinear iterates will follow. This will be the case for some problems, but it is clear that requiring $J(u)$ to be positive definite everywhere is very restrictive.

For problems whose coefficient matrices are indefinite, the convergence theory for the linear Arnoldi and GMRES algorithms is lacking in one very important area at the present time in that it is impossible to predict how well either algorithm will perform on the problem

$$Ax = b.$$

Some partial answers to this question have been given by several authors. For example, in [2], several results are shown which basically show that either both algorithms will perform well on a particular problem or both will perform poorly. Some numerical studies performed by Huang and van der Vorst [15] suggest that the convex hull of the upper Hessenberg matrix H_m in the Arnoldi process must well approximate that of the matrix A in order for the GMRES solution to well approximate the true solution of the above linear system.

There also are generalizations of the above results for the linear Krylov methods when the coefficient matrix is positive definite. We now show a result which gives necessary and sufficient conditions for the first k steps of the linear GMRES algorithm to stagnate.

Theorem 5.1 *Consider the solution of $Ax = b$ using GMRES with initial guess x_0 and residual $r_0 = b - Ax_0$. Let x_i be the i-th GMRES iterate for $i = 1, \cdots, k$. Assume A is nonsingular and let μ be the minimal degree of A with respect to r_0. Then for any $k \leq \mu$, we have $x_k = x_{k-1} = \cdots = x_0$ if and only if $r_0^T A^i r_0 = 0$ for $i = 1, \cdots, k$.*

Proof: Let $r_j = b - Ax_j$ $(j = 1, \cdots, k)$ with $k \leq \mu$. Because $j \leq k \leq \mu$, the $N \times j$ matrix $Z_j \equiv [r_0, Ar_0, ..., A^{k-1}r_0]$ is of full rank. It follows directly from the remark after (2.7) that $x_j = x_0 + Z_j y_j$, where $y_j \in \mathbf{R}^j$ solves the $j \times j$ system

$$(AZ_j)^T(r_0 - AZ_j y_j) = 0.$$

Since Z_j is of full rank and A is nonsingular, the above system has a unique solution. Observe that $x_j = x_0$ if and only if this unique solution is $y_j = 0$. This is true if and only if the right hand side $(AZ_j)^T r_0$ is zero, i.e., if and only if $r_0^T A^i r_0 = 0$ for $i = 1, \cdots, j$. The result follows immediately. □

Using this result, it follows immediately that if A^k is either positive or negative real, then GMRES(k') converges for any $k' \geq k$. Although interesting theoretically, this result is of limited practical application. However, it and the ones referred to above indicate some of the subtleties involved in analyzing the convergence behavior of the linear Krylov methods.

An important issue somewhat related to the question of stagnation is that of preconditioning. We mentioned above that adding the steepest descent direction to each subspace K guarantees convergence. In fact, using the normal equation approach, i.e., solving the linear systems by using the conjugate gradient algorithm on the normal equations, also guarantees convergence. However, the convergence can be very slow and this can be just as problematic as divergence. The usual way to improve convergence of the linear solvers is to precondition the systems. The key solution to avoiding stagnation is to use a good preconditioning technique. The importance of preconditioners over the Krylov subspace methods themselves has been illustrated in the tests in our previous paper [5]. There are a number of standard preconditioners that can be used when the Jacobian is explicitly available, the simplest and often the most inexpensive being the incomplete ILU factorization. Unfortunately when the Jacobian matrix is not available, i.e., in matrix-free methods, this is not feasible. Preconditionings that do not require the Jacobian explicitly but only its action on a given vector could be extremely useful. We should mention that any nonlinear fixed-point iteration can be considered as a matrix-free preconditioner [7]. We believe that much research remains to be done in this direction. For now, we can say that the best successes of nonlinear Krylov subspace methods have been in cases where the particular knowledge of the physics of the problem allows one to derive suitable preconditioners [29].

6 Conclusion

We have provided some theory for nonlinear projection methods with emphasis on those methods based on Krylov subspaces. The main results are similar to others in the literature, from which they have been adapted.

One of the main restrictions of most of the schemes used is that the subspace onto which a given Newton step is projected must solve the Newton equations with a certan accuracy which is dictated by the residual condition (2.1). This, as we have shown, is enough to essentially guarantee convergence of the standard linesearch and trust region algorithms. On the practical side, the main difficulty is that one does not know in advance if the subspace chosen for projection will be good enough to guarantee this residual condition. Techniques which use restarting of the linear iteration can be very useful in this context. Moreover, preconditioning is essential in the successful application of these methods.

Finally, there are generalizations of the Newton-Krylov methods considered in this paper which may be more effective on certain problems. For example, once a subspace K has been constructed along with an orthonormal basis V, one could consider solving the nonlinear least squares problem

$$\min_{y \in \mathbf{R}^m} f(u + Vy),$$

where $f(u) = \frac{1}{2} F(u)^T F(u)$ and u is the current approximate solution. This would be in lieu of the quadratic models considered in the trust region algorithms. Preliminary testing of such an algorithm has been encouraging. We will consider this and other generalizations in future work.

Acknowledgements. The authors are indebted to Homer Walker for the very valuable help he provided them by carefully reading the manuscript and pointing out a few errors. In particular the modification of the first step in Algorithm 3.1 was spurred by his remarks. Also, as was mentioned earlier, the result in Lemma 3.8 is due to Walker.

Appendix

The proof below follows closely that of Theorem 4.1 given in [26]. The major differences arise from the fact that a lower dimensional quadratic model is used, rather than the full N-dimensional model assumed in [26].

Proof of Theorem 4.4: By Taylor's theorem, for any n and any $\tau > 0$,

$$\begin{aligned} |\text{ared}_n(\tau) - \text{pred}_n(\tau)| &= \left| f_n - f(u_n + p_n(\tau)) - \left(f_n - f_n - g_n^T p_n(\tau) - \frac{1}{2} p_n(\tau)^T B_n p_n(\tau) \right) \right| \\ &= \left| \frac{1}{2} p_n(\tau)^T B_n p_n(\tau) - \int_0^1 p_n(\tau)^T \nabla^2 f(u_n + \xi p_n(\tau)) p_n(\tau) \cdot (1-\xi) d\xi \right| \\ &\leq \|p_n(\tau)\|_2^2 \int_0^1 \|B_n - \nabla^2 f(u_n + \xi p_n(\tau))\|_2 (1-\xi) d\xi. \end{aligned}$$

So,

$$\left| \frac{\text{ared}_n(\tau)}{\text{pred}_n(\tau)} - 1 \right| \leq \frac{\|p_n(\tau)\|_2^2 \int_0^1 \|B_n - \nabla^2 f(u_n + \xi p_n(\tau))\|_2 (1-\xi) d\xi}{|\text{pred}_n(\tau)|}. \tag{A.1}$$

Also, note that for any sequence $\{u_n\}$ generated by an algorithm satisfying the conditions of Algorithm 4.2, the related sequence $\{f_n\}$ is monotonically decreasing and bounded from below. Hence, f_n converges to an f_* as n goes to infinity. This fact will be used in the remainder of the proof.

Proof of (i): Since $\eta_n \leq \eta_{\max} < 1$ for all n, we have

$$\sigma_n \geq \frac{1 - \eta_{\max}}{(1 + \eta_{\max})M} \equiv \bar{\sigma} > 0 \text{ for all } n.$$

¿From (4.67) it follows that

$$f_n - \phi_n(q_n(\tau)) \geq \frac{1}{2} \bar{\sigma} \|g_n\|_2 \min\{\tau, \bar{\sigma} \frac{\|g_n\|_2}{\|B_n\|_2}\} \text{ for all } n. \tag{A.2}$$

Next, consider any k with $\|g_k\|_2 \neq 0$. For any u, $\|g(u) - g_k\|_2 \leq \beta_1 \|u - u_k\|_2$. So, if $\|u - u_k\|_2 < \|g_k\|_2/(2\beta_1)$, then

$$\|g(u)\|_2 \geq \|g_k\|_2 - \|g(u) - g_k\|_2 \geq \frac{\|g_k\|_2}{2}.$$

Let $R = \|g_k\|_2/2$ and $D_R = \{u : \|u - u_k\|_2 < R\}$.

At this point there are two possibilities: either $u_n \in D_R$ for all $n \geq k$ or eventually $\{u_n\}$ leaves the ball. We show the latter is true by contradiction. Suppose $u_n \in D_R$ for all $n \geq k$. Then for all $n \geq k$, $\|g_n\|_2 \geq \|g_k\|_2/2 (\equiv \bar{\epsilon})$. Thus, by what was shown above,

$$\begin{aligned} \text{pred}_n(\tau) &\geq \frac{1}{2} \bar{\sigma} \|g_n\|_2 \min\{\tau, \bar{\sigma} \frac{\|g_n\|_2}{\|B_n\|_2}\} \\ &\geq \bar{\sigma} \bar{\epsilon} \min\{\tau, \bar{\sigma} \frac{\bar{\epsilon}}{\beta_2}\}, \end{aligned}$$

for all $n \geq k$, since $\|B_n\|_2 \leq \beta_2$ and $\|g_n\|_2 \geq \bar{\epsilon}$ implies $\|g_n\|_2/\|B_n\|_2 \geq \bar{\epsilon}/\beta_2$. To simplify notation let $\delta = \bar{\sigma}\bar{\epsilon}$. Then, using the above inequality we have

$$\begin{aligned}\left|\frac{\text{ared}_n(\tau)}{\text{pred}_n(\tau)} - 1\right| &\leq \frac{\|p_n(\tau)\|_2^2 \int_0^1 \|B_n - \nabla^2 f(u_n + \xi p_n(\tau))\|_2 (1-\xi) d\xi}{\delta \min\{\tau, \delta/\beta_2\}} \\ &\leq \frac{\tau^2(\beta_1+\beta_2)}{\delta \min\{\tau, \delta/\beta_2\}} \\ &\leq \frac{\tau(\beta_1+\beta_2)}{\delta}\end{aligned}$$

for all $n \geq k$ and $\tau \leq \delta/\beta_2$. This gives for τ sufficiently small and $n \geq k$ that

$$\frac{\text{ared}_n(\tau)}{\text{pred}_n(\tau)} > \alpha_2.$$

In addition, we have

$$\|(V_n^T B_n V_n)^{-1} V_n^T g_n\|_2 \geq \frac{\|V_n^T g_n\|_2}{\|V_n^T B_n V_n\|_2} \geq \bar{\sigma}\frac{\|g_n\|_2}{\|B_n\|_2} \geq \frac{\delta}{\beta_2},$$

so that for τ_n sufficiently small none of the conditions allowing decrease of τ_n in 5(b) of Algorithm 4.2 above can hold. It follows that τ_n is bounded away from 0. But, since

$$f_n - f_{n+1} = \text{ared}_n(\tau_n) \geq \alpha_1 \text{pred}_n(\tau_n) \geq \alpha_1 \delta \min\{\tau_n, \frac{\delta}{\beta_2}\}, \tag{A.3}$$

and since f is bounded from below, we must have $\tau_n \to 0$, which is a contradiction. Therefore, $\{u_n\}$ must eventually be outside D_R for some $n > k$.

Let $l+1$ be the first index after k with u_{l+1} not in D_R. Then

$$\begin{aligned}f(u_k) - f(u_{l+1}) &= \sum_{n=k}^{l} f(u_n) - f(u_{n+1}) \\ &\geq \sum_{n=k}^{l} \alpha_1 \text{pred}_n(\tau_n) \\ &\geq \sum_{n=k}^{l} \alpha_1 \delta \min\{\tau_n, \frac{\delta}{\beta_2}\}.\end{aligned}$$

Now, if $\tau_n \leq \delta/\beta_2$ for $k \leq n \leq l$, we have that

$$f(u_k) - f(u_{l+1}) \geq \alpha_1 \delta \sum_{n=k}^{l} \tau_n \geq \alpha_1 \delta R.$$

Otherwise, we have that $f(u_k) - f(u_{l+1}) \geq \alpha_1 \delta^2/\beta_2$. (That is, there exists at least one n with $\tau_n > \delta/\beta_2$.) In either case,

$$\begin{aligned}f(u_k) - f(u_{l+1}) &\geq \alpha_1 \delta \min\{R, \frac{\delta}{\beta_2}\} \\ &= \alpha_1 \bar{\sigma}\frac{\|g_k\|_2}{2} \min\{\frac{\bar{\sigma}\|g_k\|_2}{2\beta_1}, \frac{\bar{\sigma}\|g_k\|_2}{2\beta_2}\} \\ &\geq \|g_k\|_2^2 \alpha_1 \bar{\sigma}^2 \frac{1}{4} \min\{\frac{1}{\beta_1}, \frac{1}{\beta_2}\}.\end{aligned}$$

By assumption, f is bounded below, and by construction f_n is monotonically decreasing. These imply that $f_n \to f_*$. Then by the preceding inequality

$$\|g_n\|_2^2 \leq \left[\alpha_1 \bar{\sigma}^2 \frac{1}{4} \min\{\frac{1}{\beta_1}, \frac{1}{\beta_2}\}\right]^{-1} (f_n - f_*).$$

Therefore, $g_n \to 0$ as $n \to \infty$.

Proof of (ii): By assumption, u_* is a limit point of $\{u_n\}$. Let $\{u_{n_j}\}$ be a subsequence converging to u_*. We show first that u_n converges to u_*. By (i), $g(u_*) = 0$. Since $J(u_*)$ is nonsingular and $0 = g(u_*) = J(u_*)^T F(u_*)$, it follows that $F(u_*) = 0$. We then have $\nabla^2 f(u_*) = J(u_*)^T J(u_*)$, and hence is positive definite. Since $\nabla^2 f$ is continuous, there exists a $\delta_1 > 0$ such that if $\|u - u_*\|_2 < \delta_1$, then $\nabla^2 f(u)$ is positive definite, and if $u \neq u_*$ then $g(u_*) \neq 0$. Let $D_1 = \{u : \|u - u_*\|_2 < \delta_1\}$.

Since $g(u_*) = 0$, we can find $\delta_2 > 0$ with $\delta_2 < \delta_1/4$ and $\|\nabla^2 f(u)^{-1}\|_2 \cdot \|g(u)\|_2 < \delta_1/2$ for all $u \in D_2 = \{u : \|u - u_*\|_2 < \delta_2\}$.

Find j_0 such that $f(u_{n_{j_0}}) < \inf\{f(u) : u \in D_1 - D_2\}$ and $u_{n_{j_0}} \in D_2$. Consider any u_l with $l \geq n_{j_0}$, $u_l \in D_2$. We claim that $u_{l+1} \in D_2$, which implies that the entire sequence beyond $u_{n_{j_0}}$ is in D_2. Suppose that u_{l+1} is not in D_2. Since $f_{l+1} < f_{n_{j_0}}$, u_{l+1} is not in D_1 either. So,

$$\begin{aligned} \tau_l \geq \|u_{l+1} - u_l\|_2 &\geq \|u_{l+1} - u_*\|_2 - \|u_l - u_*\|_2 \geq \delta_1 - \frac{\delta_1}{4} = \frac{3}{4}\delta_1 \\ &> \frac{\delta_1}{2} \geq \|B(u_l)^{-1}\|_2 \cdot \|g(u_l)\|_2 \\ &\geq \|(V_l^T B_l V_l)^{-1} V_l^T g_l\|_2. \end{aligned}$$

But, since the inexact-Newton step is within the trust region, we have

$$p_l(\tau_l) = -(V_l^T B_l V_l)^{-1} V_l^T g(u_l).$$

Since $\|p_l(\tau_l)\|_2 < \delta_1$, it follows that $u_{l+1} \in D_1$, which is a contradiction. Hence, for all $n \geq n_{j_0}$, $u_n \in D_2$, and so since $f(u_n)$ is a strictly decreasing sequence and u_* is the unique minimizer of f in D_2, we have that u_n converges to u_*.

Next, we show that the convergence rate is superlinear. This is done by showing that eventually $\|(V_n^T B_n V_n)^{-1} V_n^T g_n\|_2$ will always be less than τ_n, and hence inexact-Newton step will always be taken. Since $J(u_*)$ is nonsingular, it follows from the results in [9] that the convergence rate of u_n to u_* is superlinear.

To show that eventually the inexact-Newton step is always shorter than the trust radius, we need a particular lower bound on $\text{pred}_n(\tau)$. By the assumptions of (ii), for all n large enough, $B_n = \nabla^2 f(u_n)$ is positive definite. Hence, either the inexact-Newton step is longer than the trust radius, or $p_n(\tau)$ is the inexact-Newton step. In either case,

$$\|p_n(\tau)\|_2 \leq \|(V_n^T B_n V_n)^{-1} V_n^T g_n\|_2 \leq \|(V_n^T B_n V_n)^{-1}\|_2 \|V_n^T g_n\|_2,$$

and so it follows that $\|V_n^T g_n\|_2 \geq \|p_n(\tau)\|_2 / \|(V_n^T B_n V_n)^{-1}\|_2$. By what was shown in the proof of (i), for all n large enough we have

$$\text{pred}_n(\tau) \geq \frac{1}{2}\bar{\sigma} \frac{\|p_n(\tau)\|_2}{\|B_n^{-1}\|_2} \min\{\tau, \bar{\sigma} \frac{\|g_n\|_2}{\|B_n\|_2}\}$$

$$\geq \frac{1}{2}\bar{\sigma}\|g_n\|_2 \min\{\|p_n(\tau)\|_2, \bar{\sigma}\frac{\|p_n(\tau)\|_2}{\|B_n\|_2\|B_n^{-1}\|_2}\}$$
$$\geq \frac{1}{2}\frac{\bar{\sigma}}{M^2}\frac{\|p_n(\tau)\|_2^2}{\|B_n^{-1}\|_2}.$$

So, by the continuity of $\nabla^2 f$, for all n large enough,

$$\text{pred}_n(\tau) \geq \frac{\bar{\sigma}}{4}\frac{\|p_n(\tau)\|_2^2}{\|\nabla^2 f(u_*)^{-1}\|_2}.$$

Finally, by the argument leading up to (A.1) and Lipschitz continuity,

$$|\text{ared}_n(\tau) - \text{pred}_n(\tau)| \leq \|p_n(\tau)\|_2^3\frac{L}{2}.$$

Thus, for any $\tau > 0$ and n large enough,

$$\left|\frac{\text{ared}_n(\tau)}{\text{pred}_n(\tau)} - 1\right| \leq \frac{\|p_n(\tau)\|_2^3\frac{L}{2}}{\bar{\sigma}\|p_n(\tau)\|_2^2}4\|\nabla^2 f(u_*)^{-1}\|_2$$
$$= \frac{2L\|\nabla^2 f(u_*)^{-1}\|_2}{\bar{\sigma}}\|p_n(\tau)\|_2$$
$$\leq \frac{2L\|\nabla^2 f(u_*)^{-1}\|_2}{\bar{\sigma}}\tau.$$

Thus, by step 5(b) of Algorithm 4.2, there is a $\bar{\tau}$ such that if $\tau_{n-1} < \bar{\tau}$, then τ_n will be less than τ_{n-1} only if $\tau_n \geq \|(V_{n-1}^T B_{n-1} V_{n-1})^{-1} V_{n-1}^T g_{n-1}\|_2$. It follows from the superlinear convergence of the inexact-Newton method that for u_{n-1} close enough to u_* and n large enough, $\|(V_n^T B_n V_n)^{-1} V_n^T g_n\|_2 < \|(V_{n-1}^T B_{n-1} V_{n-1})^{-1} V_n^T g_{n-1}\|_2$. Now, if τ_n is bounded away from 0 for all large n, then we are done. Otherwise, if for an arbitrarily large n τ_n is reduced, i.e., $\tau_n < \tau_{n-1}$, then we have

$$\tau_n \geq \|(V_{n-1}^T B_{n-1} V_{n-1})^{-1} V_{n-1}^T g_{n-1}\|_2 > \|(V_n^T B_n V_n)^{-1} V_n^T g_n\|_2,$$

and so the full inexact-Newton step is taken. Inductively, this occurs for all subsequence iterates and superlinear convergence follows. □

References

[1] P. N. Brown, *A local convergence theory for combined inexact-Newton/finite-difference projection methods,* SIAM J. Numer. Anal., 24(1987),pp. 407-434.

[2] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms,* to appear in SIAM J. Sci. and Stat. Comp.

[3] P. N. Brown and A. C. Hindmarsh, *Matrix-free methods for stiff systems of ODEs,* SIAM J. Numer. Anal.,24(1987),pp. 610-638.

[4] _____, *Reduced storage methods in stiff ODE systems,* Lawrence Livermore National Laboratory UCRL-95088 report, Rev. 1, June 1987, to appear in Journal of Applied Math. and Computation.

[5] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations,* SIAM J. Stat. Scient. Comput., to appear.

[6] R. G. Carter, *Numerical Optimization in Hilbert Spaces Using Inexact Function and Gradient Evaluations,* Technical Report, ICASE, Hampton VA, 1989.

[7] T. F. Chan and K. R. Jackson, *The use of iterative linear equation solvers in codes for large systems of stiff IVPs for ODEs,* SIAM J. Sci. Stat. Comp.,7(1986),pp. 378-417.

[8] I. L. Chern and W. L. Miranker, *Dichotomy and conjugate gradients in the stiff initial value problem,* Technical Report 8032-34917, IBM, 1980.

[9] R. S. Dembo, S. C. Eisenstat and T. Steihaug, *Inexact Newton methods,* SIAM J. on Numer. Anal., 19 (1982), pp. 400-408.

[10] J. E. Dennis and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods,* Math. Comp. 28, pp. 549-560.

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

[12] S. C. Eisenstat and H. F. Walker, *Private communication,* 1989.

[13] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method,* SIAM J. Numer. Anal. 21 (1984), pp. 352-361.

[14] W. C. Gear and Y. Saad, *Iterative solution of linear equations in ODE codes,* SIAM J. Sci. Stat. Comp., 4(1983), pp. 583-601.

[15] Y. Huang and H. van der Vorst, *Some observations on the convergence behavior of GMRES,* preprint.

[16] T. Kerkhoven and Y. Saad, *Acceleration techniques for decoupling algorithms in semiconductor simulation,* CSRD report number 684, University of Illinois at Urbana Champaign, 1987.

[17] H. Mukai and E. Polak, *A second order methods for unconstrained optimization* J. Opt. Theory Appl., 26 (1978), pp.1-20.

[18] S. G. Nash, *Newton-like minimization via the Lanczos method,* SIAM J. Num. Anal., 21 (1984), pp. 770-788.

[19] _____, *Preconditioning of truncated-Newton methods,* SIAM J. Sci. Stat. Comp., 6 (1985), pp. 599-616.

[20] D. P. O'Leary, *A discrete Newton algorithm for minimizing a function of many variables,* Math. Programming, 23 (1982), pp. 20-33.

[21] J.M. Ortega and W.C. Rheinboldt, *Iterative solution of nonlinear equations in several variables,* Academic Press, New-York, 1970.

[22] M. J. D. Powell, A hybrid method for nonlinear equations, P. Rabinowitz ed., *Numerical Methods for Nonlinear Equations,* Gordon-Breach, 1970.

[23] _____, Convergence properties of a class of minimization algorithms, in *Nonlinear Programming* 2, O. Mangasarian, R. Meyer, and S. Robinson, eds. Academic Press, New York, pp. 1-27, 1975.

[24] Y. Saad, *Krylov subspace methods for solving unsymmetric linear systems,* Mathematics of Computation, 37(1981), pp. 105-126

[25] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems,* SIAM J. Sci. Stat. Comp., 7(1986), pp. 856-869.

[26] G. A. Shultz, R. B. Schnabel and R. H. Byrd, *A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties,* SIAM J. Numer. Anal., 22 (1985), pp. 47-67.

[27] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization,* SIAM J. on Numer. Anal., 20 (1983), pp. 626-637.

[28] H. F. Walker, *Lecture notes, CS 745, Yale University, Computer Science Dept,* 1989.

[29] L. B. Wigton, D. P. Yu and N. J. Young, GMRES acceleration of computational fluid dynamics codes, *Proceedings of the 1985 AIAA conference, Denver 1985,* AIAA, Denver, 1985.