INEXACT NEWTON METHODS FOR SINGULAR PROBLEMS*

C. T. KELLEY and Z. Q. XUE

North Carolina State University, Center for Research in Scientific Computation and Department of Mathematics, Box 8205, Raleigh, N. C. 27695-8205, USA.

In this paper we describe the effects of an inexact implementation of Newton's method on the behavior of the iteration for certain nonlinear equations in Banach space for which the Fréchet derivative is singular at the solution. We give a termination criterion for the inner iteration that preserves not only the q-linear convergence of the Newton iterates but also the fine structure required for an acceleration method.

KEY WORDS: inexact Newton method, singular nonlinear equation, simple fold, acceleration of convergence

1 INTRODUCTION

In this paper we describe the effects of an inexact [12] implementation of Newton's method on the behavior of the iteration for certain nonlinear equations in Banach space for which the Fréchet derivative is singular at the solution. As a particular example we consider the Newton-GMRES iteration [2]. We give a termination criterion for the inner iteration that preserves not only the q-linear convergence of the Newton iterates but also the fine structure required for a generalization of the acceleration method given in [22].

Let F be a map from a Banach space E into itself. Assume that $F(x^*) = 0$ for some $x^* \in E$ and that F is twice Lipschitz continuously differentiable in a neighborhood of x^*. We consider a class of problems where the standard assumption in nonlinear equations that $F'(x^*)$ is nonsingular fails to hold. In several cases [3], [6], [7], [5] [24], [25], [26], [15], [14], [17], [18], the performance of Newton's method is different from that in the standard situation in two respects. First of all, the set of initial iterates from which the Newton iterates will converge is not a ball about x^* but rather a restricted region that avoids the set on which F' is singular. Also, the convergence is not quadratic, but q-linear with a q-factor that depends

* This research was supported by Air Force Office of Scientific Research grant #AFOSR-FQ8671-9101094 and National Science Foundation grant #DMS-9024622. Computing activity was also partially supported by an allocation of time from the North Carolina Supercomputing Center.

on the nature of the singularity. This behavior is shared by problems that are near such singular problems [11], [16], and the structure of the singularity affects the performance of various Newton-like methods such as the Shamanskii method [21], the chord method [9], and Broyden's method [10].

In this paper we consider the class of singular problems with simple quadratic folds [19] at the root. Our methods extend to more general situations including the higher order singularities with higher dimensional null spaces considered in [22], but the case of simple quadratic folds is sufficient to completely expose the ideas. We make the following assumptions on the singularity.

Assumption 1.1. *$F'(x^*)$ has a one dimensional null space N spanned by $\phi \in E$ and closed range X such that $E = N \oplus X$. For any projection P_N onto N parallel to X we have*

$$P_N F''(x^*)(\phi, \phi) \neq 0.$$

Assumption 1.1 can be modestly weakened by replacing X with any compliment of N in E and assuming that the range of $F'(x^*)$ has codimension one.

Following the notation in much of the literature on singular problems we let $\tilde{x} = x - x^*$ for $x \in E$ and $P_X = I - P_N$. For singularities satisfying Assumption 1.1 one region of admissible initial iterates for Newton's method is

$$W(\rho, \theta) = \{x \in E \,|\, 0 < \|\tilde{x}\| < \rho, \|P_X \tilde{x}\| \leq \theta \|P_N \tilde{x}\|\},$$

for ρ and θ sufficiently small. Other, larger, regions have been described in the literature [14], but iterates after the first lie in $W(\rho, \theta)$. The basic convergence and structure result is, [6], [25],

Theorem 1.1. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Then if ρ and θ are sufficiently small and $x_0 \in W(\rho, \theta)$ the Newton sequence*

$$x_{n+1} = x_n - F'(x_n)^{-1} F(x_n)$$

exists (i.e. *$F'(x_n)$ is nonsingular for all n), remains in $W(\rho, \theta)$, and converges to x^* with q-factor $1/2$,*

$$\lim_{n \to \infty} \frac{\|\tilde{x}_{n+1}\|}{\|\tilde{x}_n\|} = \frac{1}{2}.$$

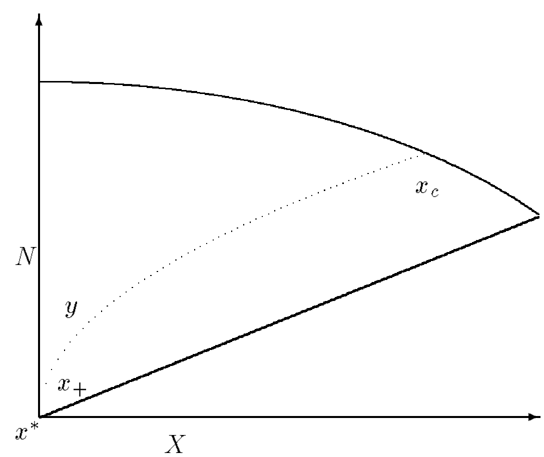
If $E = R$ then quadratic convergence can be recovered by the simple artifice of doubling the Newton step at each iterate. If E has dimension larger than one this trick will not work because the modified iterates may leave $W(\rho, \theta)$ and even diverge. Several methods have been proposed [8], [15], [5], [22] to recover superlinear convergence. The most generally applicable of these is the method from [22] which is described in the following theorem.

Theorem 1.2. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Let $\alpha \in (0, 1)$ and real $C \neq 0$ be given. Then if ρ and θ are sufficiently small, $x_0 \in W(\rho, \theta)$, and y_n, s_n, and x_{n+1} are given* *for* $n \geq 0$ *by*

$$\begin{aligned} y_n &= x_n - F'(x_n)^{-1}F(x_n), \\ s_n &= -F'(y_n)^{-1}F(y_n), \\ x_{n+1} &= y_n + (2 - C\|s_n\|^{\alpha})s_n, \end{aligned} \tag{1}$$

then $\{x_n\}$ *exists, remains in* $W(\rho, \theta)$ *and converges q-superlinearly to* x^* *with q-order* $1 + \alpha$.

FIGURE 1: $W(\rho, \theta)$



In Figure 1 we plot X on the horizontal axis and N on the vertical. The figure shows one quadrant of the convergence cone $W(\rho, \theta)$ and x_c is located near the boundary. We plot the locations of the iterations generated by one sweep of the iteration in (1) beginning with $x_c \in W(\rho, \theta)$. We plot the intermediate iterate y and the modified Newton iterate x_+. The dotted curve is the the boundary of the set

$$Q = \{x \in W(\rho, \theta) \,|\, \|P_X \tilde{x}\| \leq \|P_N \tilde{x}\|^2\}.$$

It is known, see Lemma 2.1 in § 2, that the Newton iterates described by Theorem 1.1 satisfy $x_n \in Q$ for $n \geq 1$ and ρ and θ sufficiently small. Hence the intermediate iterate y satisfies $y \in Q$ and $\|\tilde{y}\| \approx \|\tilde{x}_c\|/2$. The purpose of the intermediate iterate is to locate y deeply enough in $W(\rho, \theta)$ so that the modified iteration remains in $W(\rho, \theta)$. For this purpose the sign of C does not matter, $|C|$ need only be large enough to keep x_+ in $W(\rho, \theta)$ so that the iteration can continue. Smaller values of $|C|$ require smaller values of ρ and θ in order that $x_+ \in W(\rho, \theta)$. $C = 0$ will not suffice for this purpose for any values of ρ and θ. The parameters C and α can be adjusted to improve the convergence rate by increasing α or decreasing $|C|$. However, when this is done the bounds required in the proof of convergence on ρ

and θ must be decreased as well. $C = 1$ and $\alpha = .9$ was a reasonable choice in the numerical experiments reported [22]. In the inexact form of (1) we make different decisions and discuss those in § 2 and 3.

If the method of Theorem 1.2 is applied to a problem for which $F'(x^*)$ is nonsingular, the intermediate iterate y will reduce the error quadratically, since it is a true Newton iterate. However the modified iterate x_+ will increase the error in y by a factor of roughly 2. The overall rate, therefore, will be two-step quadratic, the modified step being wasted. One might try to detect singularity by monitoring the progress of an unmodified Newton iteration and turning on the acceleration procedure when the ratios of successive values of $\|F(x_n)\|$ seem to be converging to $1/4$. Similar ideas might be applicable to the inexact situation considered here.

Inexact Newton methods solve the equation for the Newton step at a current iterate x_c

$$F'(x_c)s = -F(x_c)$$

approximately via some method that returns a step that satisfies

$$\|F'(x_c)s + F(x_c)\| \leq \eta_c \|F(x_c)\|. \tag{2}$$

The tolerance η_c can vary as the iteration progresses and the results in [12] give bounds on η_c that are sufficient to maintain various convergence rates. We state the following special case of the results in [12] to have a basis for comparison with the results in this paper.

THEOREM 1.3. *Let F be Lipschitz continuously Fréchet differentiable near a root x^* and let $F'(x^*)$ be nonsingular. Let*

$$x_{n+1} = x_n + s_n$$

where

$$\|F'(x_n)s_n + F(x_n)\| \leq \eta_n \|F(x_n)\|.$$

Then for all $\bar{\eta} \in [0, 1)$, if x_0 is sufficiently near x^* and $\eta_n \leq \bar{\eta}$ then $\{x_n\}$ converges to x^* q-linearly in the norm $\|\cdot\|_*$ given by

$$\|w\|_* = \|F'(x^*)w\|.$$

Moreover

- *if $\eta_n \to 0$ then $\{x_n\}$ converges to x^* q-superlinearly, and*
- *if $\eta_n = O(\|F(x_n)\|^p)$ as $n \to \infty$ for some $p \in (0, 1]$ then $\{x_n\}$ converges to x^* q-superlinearly with q-order at least $1 + p$.*

The purpose of this paper is to describe the behavior of inexact Newton methods for the class of singular problems described by Assumption 1.1. We have recently reported on success with a nested iteration multilevel version of this algorithm [13] in the context of neutron transport theory and in this paper we focus on the theoretical aspects. In § 2 we describe conditions on the sequence $\{\eta_n\}$ that allow the conclusions of Theorem 1.1 and an extension of Theorem 1.2 to hold. In § 3 we present some numerical examples, using GMRES [27] as the iterative method, that illustrate the convergence theorems.

2 CONVERGENCE RESULTS

2.1 The Structural Lemma

As was the case in [22] we require a critical lemma on the structure of F and the Newton iterates. This lemma was stated in roughly the form that we use it here in [22], and follows from earlier estimates in [6], [25], and [26].

LEMMA 2.1. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Then there are $\theta > 0$, $\rho > 0$, and $K > 0$ such that for all $x \in W(\rho, \theta)$, $F'(x)$ is nonsingular,*

$$\|F(x)\| \le K\|\tilde{x}\|(\|\tilde{x}\| + \theta),\ \|P_N F'(x)^{-1}\| \le K\|\tilde{x}\|^{-1},\ \text{and}\ \|P_X F'(x)^{-1}\| \le K. \quad (3)$$

Moreover

$$y = x - F'(x)^{-1}F(x)$$

satisfies

$$\begin{array}{l} \|P_N\tilde{y} - \frac{1}{2}P_N\tilde{x}\| \le K(\|P_X\tilde{x}\| + \|P_N\tilde{x}\|^2) \\ \text{and} \\ \|P_X\tilde{y}\| \le K(\|P_X\tilde{x}\|\,\|P_N\tilde{x}\| + \|P_N\tilde{x}\|^3). \end{array} \quad (4)$$

We will require a more detailed inexact form of this lemma.

LEMMA 2.2. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Then there are $K_I > 0$, $\theta > 0$, $\rho > 0$, and $\eta > 0$ such that if $\eta_c \le \eta$, $x \in W(\rho, \theta)$, and s satisfies*

$$\|F'(x)s + F(x)\| \le \eta_c\|F(x)\| \quad (5)$$

then the inexact Newton iterate

$$x_+ = x + s \quad (6)$$

satisfies

$$\begin{array}{l} \|P_N\tilde{x}_+ - \frac{1}{2}P_N\tilde{x}\| \le K_I(\|P_X\tilde{x}\| + \|P_N\tilde{x}\|^2 + \eta_c\rho_c + \eta_c\theta_c)\ \text{and} \\ \\ \|P_X\tilde{x}_+\| \le K_I(\|P_X\tilde{x}\|\,\|P_N\tilde{x}\| + \|P_N\tilde{x}\|^3 + \eta_c\rho_c^2 + \eta_c\theta_c\rho_c), \end{array} \quad (7)$$

where

$$\rho_c = \|\tilde{x}\|\ \text{and}\ \|P_X\tilde{x}\| = \theta_c\|P_N\tilde{x}\|. \quad (8)$$

Proof. Let ρ and $\theta < 1$ be small enough so that the conclusions of Lemma 2.1 hold. Keep in mind that $\theta_c \le \theta$ and $\rho_c \le \rho$. Let

$$\xi = F'(x)s + F(x)$$

and let the Newton iterate from x be

$$x_+^N = x - F'(x)^{-1}F(x).$$

Since $x_+ - x_+^N = F'(x)^{-1}\xi$ we have by (3) that

$$\|P_N(x_+ - x_+^N)\| \le K^2\eta_c(\rho_c + \theta_c) \text{ and } \|P_X(x_+ - x_+^N)\| \le K^2\eta_c\rho_c(\rho_c + \theta_c). \quad (9)$$

Together with (4), this proves (7) with $K_I = \max(K, K^2)$. □

From Lemma 2.2 we can make estimates on $\rho_+ = \|\tilde{x}_+\|$ and $\theta_+ = \|P_X\tilde{x}_+\|/\|P_N\tilde{x}_+\|$. The inexactness adds complexity that is not present in the exact case (cf. [25]) in that the first linear iteration needs to be done particularly accurately as described in (10).

LEMMA 2.3. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Let $r \in (1/2, 1)$. Then there are $\bar{K}$, $\bar{M}$, $\theta > 0$, $\rho > 0$, and $\eta > 0$ such that if $\eta_c \le \eta$, $x \in W(\rho, \theta)$, x_+ is given by (5) and (6), (8) holds, and*

$$\theta_c\eta_c \le \bar{M}\rho_c \quad (10)$$

then $x_+ \in W(\rho_+, \theta_+) \subset W(\rho, \theta)$ with

$$\begin{aligned}(1-r)\rho_c &\le \rho_c/2 - \bar{K}((\rho_c + \theta_c + \eta_c)\rho_c + \eta_c\theta_c) \\ &\le \rho_+ \le \rho_c/2 + \bar{K}((\rho_c + \theta_c + \eta_c)\rho_c + \eta_c\theta_c) \le r\rho_c,\end{aligned} \quad (11)$$

$$\theta_+ \le \bar{K}\rho_c, \quad (12)$$

and

$$\eta_+\theta_+ \le \bar{M}\rho_+ \text{ for all } \eta_+ \le \eta. \quad (13)$$

Proof. We begin by letting ρ, θ, and η be such that the conclusions of Lemma 2.2 hold. We will reduce ρ, θ, and η as the proof progresses. By (8)

$$(1+\theta_c)^{-1}\rho_c \le \|P_N\tilde{x}\| \le (1-\theta_c)^{-1}\rho_c$$

and hence, decreasing θ if needed so that $\theta_c \le \theta \le 1/2$,

$$\|P_N\tilde{x}\| \le (1+2\theta_c)\rho_c \le 2\rho_c, \|P_X\tilde{x}\| \le 2\theta_c\rho_c.$$

Hence

$$\begin{aligned}\|P_N\tilde{x}_+ - \tfrac{1}{2}P_N\tilde{x}\| &\le \|P_N\tilde{x}_+^N - \tfrac{1}{2}P_N\tilde{x}\| + \|P_N\tilde{x}_+^N - P_N\tilde{x}_+\| \\ &= \|P_N\tilde{x}_+^N - \tfrac{1}{2}P_N\tilde{x}\| + \|P_N(x_+^N - x_+)\| \\ &\le K(2\theta_c\rho_c + 4\rho_c^2) + K^2\eta_c(\theta_c + \rho_c) \le 4K_I(\theta_c + \rho_c)(\eta_c + \rho_c),\end{aligned} \quad (14)$$

and

$$\begin{aligned}\|P_N\tilde{x}_+\| &\le \tfrac{1}{2}\|P_N\tilde{x}\| + 4K_I(\theta_c + \rho_c)(\eta_c + \rho_c) \\ &\le \rho_c/2 + \theta_c\rho_c + 4K_I(\theta_c + \rho_c)(\eta_c + \rho_c).\end{aligned}$$

Similarly

$$\|P_N\tilde{x}_+\| \ge \rho_c/2 - \theta_c\rho_c - 4K_I(\theta_c + \rho_c)(\eta_c + \rho_c).$$

Set $K_I' = 1 + 8K_I$. We have

$$\rho_c/2 - K_I'(\theta_c + \rho_c)(\eta_c + \rho_c) \le \|P_N \tilde{x}_+\| \le \rho_c/2 + K_I'(\theta_c + \rho_c)(\eta_c + \rho_c). \tag{15}$$

In the same way we have

$$\begin{aligned} \|P_X \tilde{x}_+\| &\le \|P_X \tilde{x}_+ - P_X \tilde{x}_+^N\| + \|P_X \tilde{x}_+^N\| \\ &\le K^2 \eta_c \rho_c (\rho_c + \theta_c) + K(4\theta_c + 8\rho_c)\rho_c^2. \end{aligned}$$

Therefore,

$$\|P_X \tilde{x}_+\| \le K_I' \rho_c (\theta_c + \rho_c)(\eta_c + \rho_c). \tag{16}$$

Combining (15) and (16) with

$$\|P_N \tilde{x}_+\| - \|P_X \tilde{x}_+\| \le \rho_+ \le \|P_N \tilde{x}_+\| + \|P_X \tilde{x}_+\|$$

yields

$$\rho_c/2 - K_I'(1+\rho)(\theta_c + \rho_c)(\eta_c + \rho_c) \le \rho_+ \le \rho_c/2 + K_I'(1+\rho)(\theta_c + \rho_c)(\eta_c + \rho_c). \tag{17}$$

This implies all but the first and last inequalities in (11) with any choice of $\bar{K}$ that satisfies

$$\bar{K} \ge K_I'(1+\rho).$$

Now let $r \in (1/2, 1)$ be given. If ρ, θ, and η are small enough so that

$$\bar{K}(\theta + \eta + \rho) < r/2 - 1/4$$

and

$$\eta_c \theta_c \le \rho_c (r/2 - 1/4)/\bar{K} \tag{18}$$

then

$$\bar{K}(\theta_c + \rho_c)(\eta_c + \rho_c) < (r - 1/2)\rho_c$$

which yields the last inequality in (11) with any choice of $\bar{M}$ that satisfies

$$\bar{M} \le (r/2 - 1/4)/\bar{K}$$

To obtain (12) we note that (16) and (15) imply

$$\begin{aligned} \theta_+ &= \frac{\|P_X \tilde{x}_+\|}{\|P_N \tilde{x}_+\|} \le \frac{K_I' \rho_c((\theta_c + \eta_c + \rho_c)\rho_c + \eta_c \theta_c)}{\rho_c/2 - K_I'((\theta_c + \eta_c + \rho_c)\rho_c + \eta_c \theta_c)} \\ &= \rho_c \frac{K_I'(\theta_c + \eta_c + \rho_c + \bar{M})}{1/2 - K_I'(\theta_c + \eta_c + \rho_c + \bar{M})}. \end{aligned}$$

Reduce $\bar{M}$ if needed so that $\bar{M} \le 1/(8K_I')$. Reducing θ, ρ, and η if necessary so that $K_I'(\theta_c + \eta_c + \rho_c) \le 1/8$ proves $\theta_+ \le \rho_c$ and therefore (12) holds with any choice of $\bar{K}$ that satisfies

$$\bar{K} \ge 1.$$

Setting $\bar{K} = \max(1, K_I'(1+\rho))$ completes the proofs of (11) and (12).

It remains to prove (13). Note that

$$\eta_+\theta_+ \leq \bar{K}\eta_+\rho_c \leq \frac{\bar{K}}{1-r}\eta\rho_+.$$

Hence if we reduce η so that $\eta \leq \bar{M}(1-r)/\bar{K}$ then the proof is complete. □

Note that enforcement of the condition $\eta_c\theta_c \leq \bar{M}\rho_c$, perhaps by reduction of η, is required only on the first iterate and is satisfied automatically on all following iterates. This requirement that the first inexact Newton step be computed particularly accurately never needed to be enforced in our computations, but is important in our convergence results.

2.2 Linear Convergence

The linear convergence result on inexact Newton methods parallels Theorem 1.3.

THEOREM 2.4. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Let $\bar{M}$ be from Lemma 2.3. Then there are ρ, θ, and η, such that if $x_0 \in W(\rho, \theta)$,*

$$\eta_0 \|P_X \tilde{x}_0\| \leq \bar{M}\|P_N \tilde{x}_0\|\|\tilde{x}_0\|, \tag{19}$$

and $\{\eta_n\}$ is such that $\eta_n \leq \eta$ for all n then the sequence given by

$$x_{n+1} = x_n + s_n,$$

where s_n satisfies

$$\|F'(x_n)s_n + F(x_n)\| \leq \eta_n \|F(x_n)\|,$$

remains in $W(\rho, \theta)$ and converges q-linearly to x^*. Moreover if $\eta_n \to 0$ as $n \to \infty$, then $x_n \to x^*$ with q-factor $1/2$.

Proof. Let $x_0 \in W(\rho_0, \theta_0)$ with $\|\tilde{x}_0\| = \rho_0 < \rho$ and $\|P_X\tilde{x}\|/\|P_N\tilde{x}\| = \theta_0 \leq \theta$. Let ρ, θ, and η be such that the conclusions of Lemma 2.3 hold for some $r \in (1/2, 1)$. We may do this because (19) implies that (10) holds.

By Lemma 2.2 we have

$$\rho_1 \leq r\rho_0$$

and

$$\theta_1 \leq \bar{K}\rho_0.$$

Hence if $\bar{K}\rho_0 \leq \theta$, $x_1 \in W(\rho, \theta)$. Since Lemma 2.3 also states that $\eta_1\theta_1 \leq \bar{M}\rho_1$, we may proceed with the iteration to find that $x_n \in W(\rho, \theta)$ for all n and that

$$\rho_{n+1} = \|\tilde{x}_{n+1}\| \leq r\rho_n.$$

Hence the convergence is q-linear.

To get a more detailed estimate we use Lemma 2.3 to obtain

$$\theta_{n+1} \leq \bar{K}\rho_n \to 0$$

as $n \to \infty$. Hence

$$\rho_{n+1} = \rho_n/2 + O(\eta_n\rho_n) + o(\rho_n),$$

and so if $\eta_n \to 0$ as $n \to \infty$ the q-factor is $1/2$. □

2.3 Acceleration of Convergence

We consider a modification of the algorithm given in [22]. For σ_c given (perhaps depending on x_c And η_c) and for $x_c \in W(\rho, \theta)$ and $\eta_c \leq \eta$, where ρ, θ, η are such that the hypotheses of Theorem 2.4 hold we compute x_+ through the following algorithm.

ALGORITHM 2.1.

1. *Compute s^x such that*

$$\|F'(x_c)s^x + F(x_c)\| \leq \eta_c \|F(x_c)\|.$$

2. *Set $y = x_c + s^x$.*
3. *Compute s^y such that*

$$\|F'(y)s^y + F(y)\| \leq \eta_c \|F(y)\|. \tag{20}$$

4. *Set $x_+ = y + (2 + \sigma_c)s^y$.*

We use Lemma 2.2, Lemma 2.3, and Theorem 2.4 to obtain the refined form of Lemma 2.3 that we will need for an acceleration result.

LEMMA 2.5. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Then there are $K_\mu > 0$, ρ, θ, and η such that if $x_c \in W(\rho, \theta)$, $\eta_c \leq \eta$,*

$$\|\tilde{x}_c\| = \rho_c \text{ and } \|P_X \tilde{x}_c\| = \theta_c \|P_N \tilde{x}_c\|, \tag{21}$$

(10) *holds, and s^y is given by* (20) *in step 3 of Algorithm 2.1 then*

$$s^y = -\frac{\tilde{y}}{2} + \mu_1,$$

where

$$\|P_N \mu_1\| \leq K_\mu \rho_c(\rho_c + \eta_c) \text{ and } \|P_X \mu_1\| \leq K_\mu \rho_c^2. \tag{22}$$

Proof. Let ρ, θ, and η be such that the conclusions of Lemma 2.2, Lemma 2.3, and Theorem 2.4 hold for some $r \in (1/2, 1)$. As in the proof of Lemma 2.3 we assume that $\theta \leq 1/2$. Since $\theta \leq 1/2$ we have for all $w \in W(\rho, \theta)$ that

$$\|\tilde{w}\|/2 \leq \|P_N \tilde{w}\| \leq 2\|\tilde{w}\|.$$

If y is given by step 2 of Algorithm 2.1 then (21) and Lemma 2.2 imply that $y \in W(\rho, \theta)$ and

$$\begin{aligned} \|P_X \tilde{y}\| &\leq K_I(\|P_X \tilde{x}_c\| \, \|P_N \tilde{x}_c\| + \|P_N \tilde{x}_c\|^3 + \eta_c \rho_c^2 + \eta_c \theta_c \rho_c) \\ &\leq K_I(\theta_c 4\rho_c^2 + 8\rho_c^3 + \eta_c \rho_c^2 + \eta_c \theta_c \rho_c) \\ &\leq K_I(4\theta_c + 8\rho_c + \eta_c + \bar{M})\rho_c^2 \leq K_A \rho_c^2. \end{aligned} \tag{23}$$

In (23), $K_A = K_I(4\theta + 8\rho + \bar{M})$.

Recall from (12) that

$$\theta_y = \frac{\|P_X\tilde{y}\|}{\|P_N\tilde{y}\|} \le \bar{K}\rho_c. \tag{24}$$

Let $z = y + s^y$. As pointed out above, Theorem 2.4 implies that $y \in W(\rho, \theta)$. Hence, we may apply Lemma 2.2 to z and conclude that

$$\tilde{z} = \frac{1}{2}P_N\tilde{y} + \nu_1,$$

where, since $\|\tilde{y}\| \le r\rho_c$,

$$\begin{aligned} \|P_N\nu_1\| &\le K_I(\|P_X\tilde{y}\| + \|P_N\tilde{y}\|^2 + \eta_c\|\tilde{y}\| + \eta_c\theta_y) \\ &\le K_I(K_A\rho_c^2 + 4r^2\rho_c^2 + \eta_c r\rho_c + \bar{K}\eta_c\rho_c) \\ &\le K_B\rho_c(\rho_c + \eta_c), \end{aligned} \tag{25}$$

where $K_B = K_I(K_A + 4r^2 + r + \bar{K})$.

Similarly there is $K_C > 0$ such that

$$\begin{aligned} \|P_X\nu_1\| &\le K_I(\|P_X\tilde{y}\|\|P_N\tilde{y}\| + \|P_N\tilde{y}\|^3 + \eta_c\|\tilde{y}\|^2 + \eta_c\theta_y\|\tilde{y}\|) \\ &\le K_C\rho_c^2(\rho_c + \eta_c). \end{aligned}$$

We may estimate μ_1 in terms of ν_1 since

$$\begin{aligned} \mu_1 &= s^y + \frac{\tilde{y}}{2} = (\tilde{z} - \tilde{y}) + \frac{\tilde{y}}{2} = \tilde{z} - \frac{\tilde{y}}{2} \\ &= \tilde{z} - \frac{1}{2}P_N\tilde{y} - \frac{1}{2}P_X\tilde{y} = \nu_1 - \frac{1}{2}P_X\tilde{y}. \end{aligned}$$

Hence

$$\|P_X\mu_1\| \le \frac{1}{2}\|P_X\tilde{y}\| + \|P_X\nu_1\| \le \frac{K_A}{2}\rho_c^2 + K_C\rho_c^2(\rho_c + \eta_c)$$

and $P_N\nu_1 = P_N\mu_1$. This completes the proof with $K_\mu = \max(K_B, K_A/2 + (\rho + \eta)K_C)$. □

If we select σ_c properly we can apply Lemma 2.5 to Algorithm 2.1 to obtain superlinear convergence. We state the main result in terms of the transition from x_c to x_+ with a fairly general choice of σ_c and then as a corollary discuss possible implementations. Note that we require a nonzero lower bound for $|\sigma_c|$ in (26). The parameters C and α play the same role in the inexact algorithm as they do in the algorithm described in Theorem 1.2.

THEOREM 2.6. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 and (10) hold. Let $r \in (1/2, 1)$, $C \ne 0$, $C \in R$, and $\alpha \in (0,1)$ be given. Then there are $\rho, \theta, \eta, \sigma, K_\sigma^+$, and K_σ^- such that if $x_c \in W(\rho, \theta)$, $\eta_c \le \eta$, (21) holds,*

$$C(\rho_c + \eta_c)^\alpha \le |\sigma_c| \le \sigma \le 1. \tag{26}$$

and x_+ is given by Algorithm 2.1 then $x_+ \in W(\rho, \theta)$,

$$\|\tilde{x}_+\| \leq r\rho_c, \text{ and } K_\sigma^- |\sigma_c|\rho_c \leq \|\tilde{x}_+\| \leq K_\sigma^+ |\sigma_c|\rho_c. \tag{27}$$

Moreover, there is K_η such that if

$$\eta_+ \leq K_\eta |\sigma_c|(\eta_c + \rho_c)^\alpha \tag{28}$$

then

$$\eta_+\theta_+ \leq \bar{M}\rho_+. \tag{29}$$

Proof. Let ρ, θ, and η be small enough so that the conclusions of Lemma 2.2, Lemma 2.3, Theorem 2.4, and Lemma 2.5 hold for the $r \in (1/2, 1)$ given in the statement of the theorem.

Let

$$\rho_+ = \|\tilde{x}_+\| \text{ and } \theta_+ = \|P_X \tilde{x}_+\|/\|P_N \tilde{x}_+\|.$$

Note that

$$\tilde{x}_+ = \tilde{y} + (2 + \sigma_c)s^y = \tilde{y} + (2 + \sigma_c)(-\tilde{y}/2 + \mu_1) = \frac{-\sigma_c \tilde{y}}{2} + (2 + \sigma_c)\mu_1. \tag{30}$$

Theorem 2.4 implies that

$$(1 - r)\rho_c \leq \|\tilde{y}\| \leq r\rho_c.$$

Therefore, since $|\sigma_c| \leq \sigma \leq 1$, Lemma 2.5 implies that

$$\frac{(1 - r)|\sigma_c|\rho_c}{2} - 6K_\mu(\rho_c + \eta_c)\rho_c \leq \rho_+ \leq \frac{r|\sigma_c|\rho_c}{2} + 6K_\mu(\rho_c + \eta_c)\rho_c. \tag{31}$$

Reducing ρ, η, and σ, if needed so that

$$r\sigma/2 + 6K_\mu(\rho + \eta) \leq r$$

proves the first inequality in (27). To prove the second, note that (26) implies that

$$6K_\mu(\rho_c + \eta_c) = 6K_\mu(\rho_c + \eta_c)^{1-\alpha}(\rho_c + \eta_c)^\alpha \leq 6K_\mu(\rho_c + \eta_c)^{1-\alpha}C^{-1}|\sigma_c|.$$

Hence

$$|\sigma_c|\rho_c((1 - r)/2 - 6K_\mu C^{-1}(\rho + \eta)^{1-\alpha}) \leq \rho_+ \leq |\sigma_c|\rho_c(r/2 + 6K_\mu C^{-1}(\rho + \eta)^{1-\alpha})$$

Reducing ρ and η if needed so that

$$(r/2 + 6K_\mu C^{-1}(\rho + \eta)^{1-\alpha}) < 1/2$$

proves the second inequality in (27) with

$$K_\sigma^- = (1 - r)/2 - 6K_\mu C^{-1}(\rho + \eta)^{1-\alpha} \text{ and } K_\sigma^+ = r/2 + 6K_\mu C^{-1}(\rho + \eta)^{1-\alpha}$$

It remains to show that $x_+ \in W(\rho, \theta)$ and that (29) holds. Decrease ρ and η if needed so that

$$C_N = \frac{(1 - r)C}{4} - 6K_\mu(\rho + \eta)^{1-\alpha} > 0, \tag{32}$$

where C and r are given as in the statement of the theorem. By (11)

$$\|P_N \tilde{y}\| \geq (1-r)\rho_c/2.$$

If we apply P_N to both sides of (30), take norms, and use Lemma 2.5 we have

$$\|P_N \tilde{x}_+\| \geq \frac{|\sigma_c|(1-r)\rho_c}{4} - 6K_\mu(\rho_c+\eta_c)\rho_c \geq C_N(\rho_c+\eta_c)^\alpha \rho_c.$$

If we apply P_X to both sides of (30) and use (22) and (23) we have

$$\|P_X \tilde{x}_+\| \leq \|P_X \tilde{y}\|/2 + 3K_\mu \rho_c^2 \leq (K_A/2 + 3K_\mu)\rho_c^2.$$

Setting $C_X = K_A/2 + 3K_\mu$ to obtain

$$\|P_X \tilde{x}_+\| \leq C_X \rho_c^2.$$

Hence

$$\theta_+ \leq \frac{C_X \rho_c^2}{C_N(\rho_c+\eta_c)^\alpha \rho_c} \leq \frac{C_X \rho_c}{C_N(\rho_c+\eta_c)^\alpha} \leq \frac{C_X}{C_N}\rho_c^{1-\alpha}. \tag{33}$$

Reduce ρ if needed so that

$$C_X \rho^{1-\alpha} \leq \theta C_N$$

to obtain, using the final inequality in (33),

$$\frac{C_X}{C_N}\rho_c^{1-\alpha} \leq \frac{C_X}{C_N}\rho^{1-\alpha} \leq \theta.$$

This shows that $x_+ \in W(\rho_+, \theta_+) \subset W(\rho, \theta)$.

We complete the proof by verifying (29). We use the first inequality in (33)

$$\eta_+\theta_+ \leq \eta_+ \frac{C_X \rho_c}{C_N(\rho_c+\eta_c)^\alpha}$$

and (27) to see that (29) holds if

$$\eta_+ \frac{C_X \rho_c}{C_N(\rho_c+\eta_c)^\alpha} \bar{M} K_\sigma^- |\sigma_c| \rho_c \leq \bar{M}\rho_+,$$

which holds if

$$\eta_+ \leq \frac{\bar{M} K_\sigma^- C_N}{C_X}(\rho_c+\eta_c)^\alpha |\sigma_c|.$$

Setting

$$K_\eta = \frac{\bar{M} K_\sigma^- C_N}{C_X}$$

completes the proof. □

Verification of (28) is simple in the case

$$\sigma_c \approx \bar{C}(\eta_c+\rho_c)^\alpha$$

for some $\bar{C}$. In view of (11) we may approximate ρ_c by $\|s_c^y\|$. The choice

$$\sigma_c = \bar{C}(\eta_c + \|s_c^y\|)^\alpha \tag{34}$$

will be made for the remainder of this paper. By Lemma 2.2 there are $c_-, c_+ > 0$, independent of $\bar{C}$ and σ so that the choice of σ_c given in (34) satisfies

$$\bar{C}c_-(\eta_c + \rho_c)^\alpha \leq \sigma_c \leq \bar{C}c_+(\eta_c + \rho_c)^\alpha.$$

Hence if

$$\eta_+ \leq K_\eta \bar{C} c_- \eta_c^{2\alpha}. \tag{35}$$

then

$$\eta_+ \leq K_\eta \bar{C} c_- (\eta_c + \rho_c)^{2\alpha} \leq K_\eta (\eta_c + \rho_c)^\alpha |\sigma_c|$$

and therefore (28) follows from (35). If we set

$$C_\eta = K_\eta c_-$$

if $\eta_n = \eta_0 \beta^n$, for some $\beta \in [0,1]$, and $\alpha \in [0,1]$, then (35) will follow from

$$\beta \leq C_\eta \beta^{n(2\alpha-1)} \bar{C} \eta_0^{2\alpha}.$$

Hence if $\bar{C}$ is large enough and $\alpha \in [0, 1/2]$, (28) will hold for any given η_0 and $\beta \in [0,1)$. If $\beta = 1$ and $\bar{C}$ is large enough then (28) will hold for any given η_0 and $\alpha \in [0,1)$. This control of θ_+ is, in fact, one of the roles of $\bar{C}$. As we will see from the numerical experiments an aggressive choice of $\bar{C}$ (*i. e.* a small value) is usually sufficient to keep the iteration in $W(\rho, \theta)$ and the aggressive choice makes the convergence more rapid.

We make a particular choice for definiteness in the statement of our next theorem, which is a trivial consequence of Theorem 2.6.

THEOREM 2.7. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Let $\beta \in [0,1]$ be given. Then there are ρ, θ, η such that if $x_0 \in W(\rho, \theta)$, and $\eta_n = \eta_0 \beta^n$, with $\eta_0 \leq \eta$, $\{x_n\}$ and $\{y_n\}$ are given by Algorithm 2.1, and $\{\sigma_n\}$ is given by (34) with $\alpha \in [0, 1/2]$, and $\bar{C}$ sufficiently large, then $x_n \to x^*$ q-superlinearly.*

Another approach is to set $\beta = 1$, fix $\eta_n = \eta_0 = \eta$ at a small value, and apply Algorithm 2.1 with a goal of obtaining q-linear convergence with a small q-factor. This approach would be especially appropriate in the context of Newton-iterative methods where the goal would be to reduce the number of inner iterations. Performance of this type of algorithm is described in the final theorem of this section.

THEOREM 2.8. *Let F be twice Lipschitz continuously differentiable in a neighborhood of x^* and let Assumption 1.1 hold. Let $\alpha \in [0,1)$ and $\bar{C} > 0$. Then there are ρ, θ, η such that if $x_0 \in W(\rho, \theta)$, $\bar{C}$ is sufficiently large, $\{x_n\}$ and $\{y_n\}$ are given by Algorithm 2.1, and*

$$\sigma_n = \bar{C}(\eta + \|s_n^y\|)^\alpha, \tag{36}$$

then $x_n \to x^*$ q-linearly with q-factor at most $K_\sigma^+ \bar{C} \eta^\alpha$.

3 NUMERICAL RESULTS

As an example we consider the H-equation of Chandrasekhar [4].

$$F(H)(\mu) = H(\mu) - \left(1 - \frac{1}{2}\int_0^1 \frac{\mu H(\nu)}{\mu + \nu}\, d\nu\right)^{-1} = 0. \qquad (37)$$

It is known [23], [20] that the assumptions of the theorems in § 2 hold for this equation in the space $E = L^2([0,1])$ and for discretized formulations, provided the quadrature rule integrates constant functions exactly.

We report on computations using a composite 20 point Gaussian quadrature with five subintervals to approximate all integrals. The L^2 inner product was also approximated using this quadrature rule. In all computations the initial iterate was the function identically one. GMRES [27] was used as the iterative method with the function identically zero used as the initial iterate for the inner iteration. We use a modification of the Brown-Hindmarsh GMRES code [1] with changes made in the inner product to allow the approximate L^2 inner product to be used instead of the R^N inner product. Products of $F'(x)$ with vectors were approximated with a forward difference with a difference step of 10^{-7}. The computations reported in the tables were done on a SUN SPARC 1+ running SUN OS version 4.1.1 and FORTRAN compiler f77 version 1.3.1.

In each of the tables that follow we tabulate the iteration counter n, the number of inner iterates i_g required to satisfy the termination criterion for the inner iteration, the norm of the function at the current point $\|F(x_n)\|$, and for $n \geq 1$, the ratio of successive function norms. For the unmodified Newton iteration, one would expect this ratio to tend to 1/4. We tabulate the inner iteration counter i_g in order to compare the various methods in terms of function evaluations. Each Newton-GMRES iterate requires a function evaluation for the difference approximation of the action of F' on a vector. In all cases the iteration was terminated when $\|F(x_n)\| < 10^{-12}$. All norms in the tables are L^2 norms.

Our first table illustrates the conclusions of Theorem 2.4. In the results reported in the left columns of Table 1 we set $\eta_n = .25$ for all n and in the right columns $\eta_n = 2^{-n-2}$. Note that while the number of Newton iterates is larger for the $\eta_n = .25$ case, the number of function evaluations is slightly smaller because fewer GMRES iterates are required for each Newton step. A total of 58 GMRES iterates were required for termination of the iteration in the first computation and 74 for the second. This easily offsets the additional outer iterations. The slightly irregular behavior of the iteration in the first computation is also worth noting.

To illustrate the modified method of Algorithm 2.1, Theorems 2.7 and 2.6 we set $\alpha = .25$, $\eta_n = 2^{-n-2}$ and compute σ_n by (34) with $\bar{C} = .01$. This is an aggressive (small) choice of $\bar{C}$. While such a choice was not appropriate for the exact algorithm described in [22], the inexactness had the effect of keeping σ large enough to remain in $W(\rho, \theta)$ without the need to keep $\bar{C}$ large.

In the left columns of Table 2 we do not tabulate information on $F(y_n)$ but do add the GMRES iterations required for computation of y_n to the total in i_g. Hence

TABLE 1: Newton-GMRES Iteration

$\eta_n = .25$			
n	i_g	$\|F(x_n)\|$	$\frac{\|F(x_n)\|}{\|F(x_{n-1})\|}$
0		0.37D+00	
1	1	0.12D+00	0.31D+00
2	1	0.28D-01	0.24D+00
3	2	0.11D-01	0.40D+00
4	2	0.28D-02	0.25D+00
5	2	0.68D-03	0.24D+00
6	2	0.20D-03	0.29D+00
7	2	0.47D-04	0.24D+00
8	2	0.12D-04	0.26D+00
9	3	0.60D-05	0.48D+00
10	3	0.15D-05	0.25D+00
11	3	0.37D-06	0.25D+00
12	3	0.92D-07	0.25D+00
13	3	0.23D-07	0.25D+00
14	3	0.56D-08	0.24D+00
15	4	0.15D-08	0.27D+00
16	3	0.36D-09	0.24D+00
17	4	0.11D-09	0.32D+00
18	3	0.30D-10	0.27D+00
19	4	0.11D-10	0.37D+00
20	4	0.27D-11	0.24D+00
21	4	0.65D-12	0.24D+00

$\eta_n = 2^{-n-2}$			
n	i_g	$\|F(x_n)\|$	$\frac{\|F(x_n)\|}{\|F(x_{n-1})\|}$
0		0.37D+00	
1	1	0.12D+00	0.31D+00
2	2	0.21D-01	0.18D+00
3	2	0.50D-02	0.24D+00
4	2	0.12D-02	0.24D+00
5	2	0.31D-03	0.26D+00
6	3	0.82D-04	0.26D+00
7	3	0.21D-04	0.25D+00
8	3	0.51D-05	0.25D+00
9	4	0.13D-05	0.25D+00
10	4	0.32D-06	0.25D+00
11	4	0.81D-07	0.25D+00
12	4	0.20D-07	0.25D+00
13	4	0.50D-08	0.25D+00
14	5	0.13D-08	0.25D+00
15	5	0.31D-09	0.25D+00
16	5	0.78D-10	0.25D+00
17	5	0.19D-10	0.25D+00
18	5	0.47D-11	0.25D+00
19	5	0.11D-11	0.24D+00
20	6	0.27D-12	0.23D+00

TABLE 2: Modified Method

$\eta_n = 2^{-n-2}, \alpha = .25, C = .01.$			
n	i_g	$\|F(x_n)\|$	$\frac{\|F(x_n)\|}{\|F(x_{n-1})\|}$
0		0.37D+00	
1	2	0.98D-01	0.26D+00
2	3	0.14D-01	0.14D+00
3	4	0.26D-04	0.19D-02
4	2	0.42D-06	0.16D-01
5	6	0.27D-08	0.64D-02
6	7	0.15D-12	0.55D-04

$\eta_n = .25, \alpha = .9, C = .01.$			
n	i_g	$\|F(x_n)\|$	$\frac{\|F(x_n)\|}{\|F(x_{n-1})\|}$
0		0.37D+00	
1	2	0.99D-01	0.26D+00
2	3	0.14D-01	0.14D+00
3	4	0.24D-04	0.18D-02
4	2	0.42D-06	0.17D-01
5	5	0.35D-08	0.84D-02
6	2	0.34D-09	0.97D-01
7	2	0.23D-11	0.68D-02
8	2	0.12D-12	0.51D-01

more GMRES iterations are required for each outer iterate x_n. The accelerated iteration is more efficient than the unaccelerated versions, requiring a total of 24 GMRES iterations. The superlinear convergence can be clearly seen in the tables. Small changes in the parameters α and $\bar{C}$ had modest effects. For example, with $\bar{C} = .01$ and $\alpha = .5$, 26 GRMES iterations were required and with $\bar{C} = .1$ and $\alpha = .25$ 27 GMRES iterations were needed. The number of outer iterations, six, was the same as in the tabulated results.

Finally, in the right columns of Table 2 we illustrate Theorem 2.8. Here we set $\eta_n = \eta = .25$ for all n, $\alpha = .9$, and $\bar{C} = .01$. 22 GMRES iterates were required. Small changes in $\bar{C}$ and α produced decreases in performance of at most 10% in our observations. Note that the number of functions required for 8 outer iterations and 22 inner iterations is the same as that required for 6 outer iterations and 24 inner iterations, so the costs of the two computations reported in Table 2 are the same.

REFERENCES

1. P. N. Brown and A. C. Hindmarsh, *Reduced storage matrix methods in stiff ode systems*, Tech. Report ucrl-95088, Lawrence Livermore National Laboratory, 1987.
2. P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Stat. Comp., 11 (1990), pp. 450–481.
3. R. Cavanaugh, *Difference Equations and Iterative Processes*, PhD thesis, University of Maryland, 1970.
4. S. Chandrasekhar, *Radiative Transfer*, Dover, New York, 1960.
5. D. W. Decker, H. B. Keller, and C. T. Kelley, *Convergence rates for Newton's method at singular points*, SIAM J. Numer. Anal., 20 (1983), pp. 296–314.
6. D. W. Decker and C. T. Kelley, *Newton's method at singular points I*, SIAM J. Numer. Anal., 17 (1980), pp. 66–70.
7. ———, *Newton's method at singular points II*, SIAM J. Numer. Anal., 17 (1980), pp. 465–471.
8. ———, *Convergence acceleration for Newton's method at singular points*, SIAM J. Numer. Anal., 19 (1982), pp. 219–229.

9. ———, *Sublinear convergence of the chord method at singular points*, Numer. Math., 42 (1983), pp. 147–154.
10. ———, *Broyden's method for a class of problems having singular Jacobian at the root*, SIAM J. Numer. Anal., 22 (1985), pp. 566–574.
11. ———, *Expanded convergence domains for Newton's method at nearly singular roots*, SIAM J. Sci. Stat. Comp., 6 (1985), pp. 951–966.
12. R. Dembo, S. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
13. B. D. Ganapol, C. T. Kelley, and G. C. Pomraning, *Asymptotically exact boundary conditions for the P-N equations.* Nuclear Science and Engineering, to appear.
14. A. Griewank, *Starlike domains of convergence for Newton's method at singularities*, Numer. Math., 35 (1980), pp. 95–111.
15. ———, *Analysis and modification of Newton's method at singularities*, PhD thesis, Australian National University, 1981.
16. ———, *On solving nonlinear equations with simple singularities or nearly singular solutions*, SIAM Review, 27 (1985), pp. 537–572.
17. A. Griewank and M. R. Osborne, *Newton's method for singular problems when the dimension of the null space is* > 1, SIAM J. Numer. Anal., 18 (1981), pp. 179–189.
18. ———, *Analysis of Newton's method at irregular singular points*, SIAM J. Numer. Anal., 20 (1983), pp. 747–773.
19. H. B. Keller, *Lectures on Numerical Methods in Bifurcation Theory*, Tata Institute of Fundamental Research, Lectures on Mathematics and Physics, Springer-Verlag, New York, 1987.
20. C. T. Kelley, *Solution of the Chandrasekhar H-equation by Newton's method*, J. Math. Phys., 21 (1980), pp. 1625–1628.
21. ———, *A Shamanskii-like acceleration scheme for nonlinear equations at singular roots*, Math. Comp., 47 (1986), pp. 609–623.
22. C. T. Kelley and R. Suresh, *A new acceleration method for Newton's method at singular points*, SIAM J. Numer. Anal., 20 (1983), pp. 1001–1009.
23. T. W. Mullikin, *Some probability distributions for neutron transport in a half space*, J. Appl. Prob., 5 (1968), pp. 357–374.
24. L. B. Rall, *Convergence of the Newton process to multiple solutions*, Numer. Math., 9 (1961), pp. 23–37.
25. G. W. Reddien, *On Newton's method for singular problems*, SIAM J. Numer. Anal., 15 (1978), pp. 993–986.
26. ———, *Newton's method and high order singularities*, Comput. Math. Appl., 5 (1980), pp. 79–86.
27. Y. Saad and M. Schultz, *GMRES a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comp., 7 (1986), pp. 856–869.