On Newton-iterative methods for the solution of systems of nonlinear equations / by Andrew H. Sherman.
Sherman, Andrew H.
Urbana : Dept. of Computer Science, University of Illinois at Urbana Champaign, 1975.
https://hdl.handle.net/2027/uiuo.ark:/13960/t3nw0zv7n
The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

*Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.*

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

L161—O-1096
On Newton-Iterative Methods for the Solution of Systems of Nonlinear Equations

by

Andrew H. Sherman

December 1975
On Newton-Iterative Methods for the Solution of Systems of Nonlinear Equations

by

Andrew H. Sherman*

December 1975

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

* This work supported in part by the Office of Naval Research under grant N-0014-67-A-0097-0016 and by the U. S. Energy Research and Development Administration under grant US AEC AT(11-1) 2383.
1. Introduction

Let $R$ be a subset of $\mathbb{R}^N$ and let $F: \mathbb{R} \to \mathbb{R}^N$ be a nonlinear function. We wish to consider computing a solution $\bar{x}$ of the system of nonlinear equations

$$F(x) = 0,$$

(1.1)

where the Jacobian matrix

$$F'(\bar{x}) = \frac{\partial F}{\partial x}(\bar{x})$$

is sparse, symmetric, and positive definite. Although our results apply in general, nonlinear systems arising in the numerical solution of partial differential equations are of particular interest.

A very important method for the solution of (1.1) is Newton's method (cf. Ortega and Rheinboldt [11]). For a given initial guess $x^{(0)}$, we generate a sequence of iterates $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, ... by

$$x^{(k+1)} = x^{(k)} - \delta^{(k)},$$

(1.2)

where $\delta^{(k)}$ is the solution to the sparse linear system

$$F'(x^{(k)}) \delta^{(k)} = F(x^{(k)}).$$

(1.3)

Under suitable assumptions on $x^{(0)}$ and $F$, the sequence is well defined and converges quadratically to $\bar{x}$, i.e.,

$$\exists c < 1 \text{ such that } ||x^{(k+1)} - \bar{x}|| \leq c ||x^{(k)} - \bar{x}||.$$

A straightforward implementation of Newton's method could use sparse symmetric Gaussian elimination (cf. Eisenstat, Schultz, and Sherman).

† Throughout this paper we use $|| \cdot ||$ to denote the Euclidean norm.

‡‡ In the notation of Ortega and Rheinboldt [11], this condition defines $R$-quadratic convergence. This is slightly weaker than the more standard definition of $Q$-quadratic convergence, i.e.,

$$\exists c < 1 \text{ such that } ||x^{(k+1)} - \bar{x}|| \leq c ||x^{(k)} - \bar{x}||.$$
[8], Sherman [13]) to solve the linear systems which arise. Unfortunately, this would be inefficient computationally because at each step it would require the computation of $F(x^{(k)})$ and $F'(x^{(k)})$ and the exact solution of (1.3). Also, a large amount of storage would be required for the factorization of $F'(x^{(k)})$.

A number of modifications have been proposed to reduce some of the costs of Newton's method, notably the simplified Newton's or chord method and a variety of Newton-iterative methods (cf. Ortega and Rheinboldt [11]). For certain problems, these methods may require less storage or work than Newton's method. However, computational difficulties are typical in the solution of nonlinear systems, and each of these methods has its disadvantages. The choice of method for a specific nonlinear system (1.1) depends quite heavily on the properties of $F$ and on the storage and time constraints imposed by the computational environment.

The simplified Newton's method replaces $F'(x^{(k)})$ with $F'(x^{(0)})$ in (1.3). This avoids the computation of $F'(x^{(k)})$ at each step and reduces the cost of solving (1.3) for $k > 1$, since, if $F'(x^{(0)})$ is factored into the product $U^TDU$ at the first step, each succeeding step only requires forward and backsolving three triangular systems. However, the chord method converges only linearly $\dagger$ and requires the same amount of storage as Newton's method.

The Newton-iterative methods use a linear iterative method (such as SOR, cf. Ortega and Rheinboldt [11], p. 215) to approximate the solution of (1.3) instead of solving it exactly. The sequence of iterates generated by

\[ ||x^{(k+1)} - \bar{x}|| < c ||x^{(k)} - \bar{x}|| \]

$\dagger$ The iterates $x^{(k)}$ converge linearly to $\bar{x}$ if there exists a $c < 1$ such that $\left||x^{(k+1)} - \bar{x}\right|| < c \left||x^{(k)} - \bar{x}\right||$. 

\[ \sum_{i=0}^{\infty} T_i < \infty \]
such methods depends upon the particular iterative method chosen and the
criteria used to stop the inner iteration. The advantages of Newton-
iterative methods are that they require only a small amount of storage and
that, when $x^{(k)}$ is near $x^*$, they can take advantage of the fact that zero
is a good initial approximation to $\xi^{(k)}$. The problem with Newton-iterative
methods is that their rate of convergence is essentially determined by the
rate of convergence of the inner (linear) iterative method applied to the
linearized problem (1.3). Thus they encounter all of the usual difficulties
with linear iterative methods relating to parameter estimation, starting
guesses, and stopping criteria.

In this paper we present the Newton-Richardson methods (cf.
Eisenstat, Schultz, and Sherman [7], Sherman [13]), a class of Newton-iterative
methods which use a Richardson-D'Jakonov iteration (cf. Richardson
[12], D'Jakonov [4]) to approximate the solution of (1.3) at each step.
The Newton-Richardson methods require the same amount of storage as Newton's
method or the chord method, but if $2^k$ steps of the inner iteration are taken
at the $k$-th step, they will converge quadratically with less work per step
(on average) than with Newton's method. And as compared to most other
Newton-iterative methods, the Newton-Richardson methods have the advantage
that their rate of convergence is essentially independent of $N$.

In the next section we describe a model semilinear partial
differential equation which will be used to motivate the Newton-Richardson
methods. In Section 3 we discuss the implementation of Newton's method and
show how to derive Newton-Richardson methods for the model problem. In
Section 4, we examine the general theory of Newton-iterative methods and
state several local convergence results. In Section 5 we consider the
ramifications of the theory for the model problem, and finally, in the appendix we prove the results stated in Section 4.
2. A Model Problem

In this section we describe a model problem arising in the numerical solution of semilinear partial differential equations and show how to apply Newton's method to it. In the next section, we will motivate the Newton-Richardson methods by analyzing the costs of Newton's method and certain of its modifications for this model problem.

Let $D$ be the unit square $(0,1) \times (0,1)$ and let $\partial D$ denote the boundary of $D$. The model problem we wish to solve is the nonlinear Poisson equation

$$-\Delta v = g(v) \quad \text{in } D$$

with $v = 0$ on $\partial D$,

where $g$ satisfies

$$g_v(v) \leq \lambda < \Lambda$$

(where $\lambda$ is the fundamental eigenvalue of $-\Delta$ on $D$), so that $v(x,y)$ exists and is unique (cf. Ortega and Rheinboldt [11], pp. 110, 116).

To obtain an approximate solution to (2.1), the continuous problem must be reduced to a discrete form. Several techniques are available to do this, and we consider the use of five-point finite difference approximations. We superimpose on $D$ a uniform square $n \times n$ mesh $D_h$, with boundary $\partial D_h$, and replace the differential operator $-\Delta$ in (2.1) with the standard five-point difference approximation on $D_h$. (Here $h = \frac{1}{n+1}$ is the vertical or horizontal distance between adjacent mesh points.) Letting $V_{ij}$ be the approximation to $v(ih, jh)$, we obtain the system of equations:

$$-V_{i-1,j} - V_{i,j-1} - V_{i,j+1} + V_{i+1,j} + h^2 g(V_{ij}) = 0,$$

for $(ih, jh) \in D_h$.

$$V_{ij} = 0 \quad \text{for } (ih, jh) \in \partial D_h.$$
ordering, we may write these equations more concisely as an $N \times N$ system of nonlinear equations

$$A \hat{V} = G(\hat{V}) = G(V),$$

(2.3)

in which

$$\hat{V} = \{V_{ij}: (ih, jh) \in D_h\},$$

$$G(\hat{V}) = \{h^2g(V_{ij}): (ih, jh) \in D_h\},$$

and $A$ is the $n \times n$ block tridiagonal matrix

$$A = \begin{bmatrix}
B & -I & & & \\
-I & B & -I & & \\
& -I & B & -I & \\
& & -I & B & \\
& & & & B
\end{bmatrix},$$

where $I$ is the $n \times n$ identity matrix and $B$ is the $n \times n$ tridiagonal matrix

$$B = \begin{bmatrix}
4 & -1 & & & \\
-1 & 4 & -1 & & \\
& -1 & 4 & -1 & \\
& & -1 & 4 & \\
& & & & 4
\end{bmatrix}.$$

Under the conditions imposed on (2.1), this system of equations has a unique solution $\hat{V}$ which is a close approximation to $v(ih, jh)$ (cf. Ortega and Rheinboldt [11], p. 112).

Newton's method may be applied in a straightforward way to solve (2.3) for $\hat{V}$. Given an initial guess $\hat{V}^{(0)}$, we generate a sequence of successive approximations $\hat{V}^{(1)}, \hat{V}^{(2)}, \hat{V}^{(3)}, \ldots$ to $\hat{V}$ by

$$\hat{V}^{(k+1)} = \hat{V}^{(k)} - \hat{\delta}^{(k)},$$

where $\hat{\delta}^{(k)}$ is the solution of the system of linear equations
\[ J(\mathbf{y}(k)) \delta(k) = A\mathbf{y}(k) - G(\mathbf{y}(k)) \] (2.4)

and

\[ J(\mathbf{y}(k)) = A - \frac{\partial G}{\partial \mathbf{y}}(\mathbf{y}(k)). \]

Under the assumptions made above, \( J(\mathbf{y}(k)) \) is an \( N \times N \) sparse, symmetric, positive definite matrix having the same zero structure as \( A \) (i.e., \( (J(\mathbf{y}(k)))_{ij} \neq 0 \) if and only if \( A_{ij} \neq 0 \)). If \( \mathbf{y}(0) \) is sufficiently close to \( \mathbf{y} \), then Newton's method converges quadratically to the solution of (2.3) independent of \( h \) (cf. Greenspan and Parter [10]), and only \( O(\log \log N) \) Newton iterations are required to reduce the initial error \( ||\mathbf{y}(0) - \mathbf{y}|| \) by a factor of \( h^2 = \frac{1}{N} \), the discretization error.

\[ ^* \text{Throughout this paper we use } \log(x) \text{ to denote } \log_2(x). \]
3. Newton-Richardson Methods for the Model Problem

As we saw in the last section, each step of Newton's method for the model problem requires the solution of a sparse symmetric linear system. Thus Newton's method really amounts to the solution of a sequence of related linear systems, and its costs are directly related to the manner in which these systems are solved. In this section we consider a number of approaches to solving the systems (2.4), and we introduce Newton-Richardson methods for the model problem (cf. Eisenstat, Schultz, and Sherman [7], Sherman [13]).

In the standard implementation of Newton's method for the model problem, the linear systems which arise can be solved with sparse symmetric Gaussian elimination (cf. Eisenstat, Schultz, and Sherman [7, 8], Sherman [13]). Each solution requires \(O(N \log N)\) storage locations for the factors of \(J(\bar{y}(k))\), \(O(N^{3/2})\) arithmetic operations to factor \(J(\bar{y}(k))\), and \(O(N \log N)\) arithmetic operations to solve the resulting triangular systems (cf. Sherman [13]). For \(\bar{y}(0)\) sufficiently close to \(\bar{y}\), Newton's method converges quadratically, and the initial error \(||\bar{y}(0) - \bar{y}|||\) can be reduced by a factor of \(h^2 = \frac{1}{N}\) in \(O(\log \log N)\) steps at a total cost of \(O(\log \log N)\) storage locations and \(O(N^{3/2} \log \log N)\) arithmetic operations.

One of the reasons that Newton's method requires so many arithmetic operations is that it does not exploit the fact that the iterates \(\bar{y}(k)\) are converging quadratically to \(\bar{y}\). The sequence of systems (2.4) is solved as if the systems were unrelated, when, in fact, their solutions rapidly approach zero as \(\bar{y}(k)\) converges to \(\bar{y}\). To exploit this situation, we could use an iterative method to approximate the solution of (2.4) at each step. Most Newton-iterative methods require only \(O(N)\) storage locations for the model problem, and some require fewer arithmetic operations than Newton's method. Unfortunately, however, for the model problem the rate of
convergence of most Newton-iterative methods depend on the mesh spacing \( h \) (see Section 5) in such a way that they have higher asymptotic arithmetic costs than Newton's method. For example, even though Newton-SOR converges linearly, the rate of convergence depends on \( h \), and the method requires \( O(N^{1/2} \log N) \) iterations to reduce the error by a factor of \( h^2 \) (see Section 5). If the rate of convergence were independent of \( h \), then only \( O(\log N) \) iterations would be required.

We now present the Newton-Richardson methods for the model problem. These methods require more storage than most other Newton-iterative methods (\( O(N \log N) \) locations), but they converge at rates which are independent of \( h \). The methods are based on the principles embodied in the Strongly Implicit Procedures developed by Stone [14], Dupont [6], Diamond [3], and others. The basic idea is to use sparse Gaussian elimination to obtain \( \hat{\delta}'(0) \) at the first step and to compute \( \hat{\delta}'(k) \) at succeeding steps with a Richardson-D'Jakovov iteration (cf. Richardson [12], D'Jakovov [4]) involving the factorization of \( J(\hat{y}'(0)) \).

In particular, suppose that at the \( k+1 \)-st step \( (k \geq 1) \) we wish to solve the system

\[
J(\hat{y}'(k)) \hat{\delta}'(k) = \hat{\beta}'(k).
\]

Since we have already factored \( J(\hat{y}'(0)) \) at the first step, the system

\[
J(\hat{y}'(0)) \hat{\theta} = \hat{\beta}
\]

can be solved with \( O(N \log N) \) arithmetic operations. We can then solve (3.1) by using a Richardson-D'Jakovov iteration to generate a sequence of iterates \( \hat{\delta}'_0, \hat{\delta}'_1, \hat{\delta}'_2, \ldots \) which converges to \( \hat{\delta}'(k) \). (The idea here is

\[\text{†}
\]

It is also possible to use a scaled conjugate gradient iteration here (cf. Douglas and Dupont [5]), but the analysis is more complicated, and, in any case, the results are improved by at most a constant factor.
similar to that of iterative improvement, cf. Forsythe and Moler [9], pp. 49-54). For \( \delta_0^{(k)} = 0 \) and a fixed real constant \( \gamma \), the Richardson-D'Jakonov iterates satisfy

\[
\delta_i^{(k+1)} = \delta_i^{(k)} - \gamma \xi_i^{(k)},
\]

where \( \xi_i^{(k)} \) is the solution to the linear system

\[
J(z^{(0)}) \xi_i^{(k)} = J(z^{(k)}) \delta_i^{(k)} - b^{(k)}.
\]

Each step of the iteration requires \( O(N \log N) \) arithmetic operations, and the iterates will converge linearly to \( \delta^{(k)} \) if

\[
\rho = \left\| I - \gamma J(z^{(0)})^{-1} J(z^{(k)}) \right\| < 1
\]

(cf. Young [15], p. 77).

To maximize the rate of convergence, we choose \( \gamma \) to minimize \( \rho \). If all the eigenvalues of \( J(z^{(0)})^{-1} J(z^{(k)}) \) lie in the interval \([\alpha, \beta]\), this is accomplished by choosing

\[
\gamma = \frac{2}{\alpha + \beta},
\]

yielding

\[
\rho = (\beta - \alpha)/(\alpha + \beta).
\]

For the model problem, \( \rho \) can be made independent of \( h \) by choosing \( z^{(0)} \) near enough to \( z \) (see Section 5), so the Richardson-D'Jakonov iteration will converge linearly, independent of the mesh size.†

Different Newton-Richardson methods are obtained by varying the number of Richardson-D'Jakonov iterates computed at each step of the outer iteration. Of particular interest are the case in which \( 2^k \) iterates are

† It has been shown by several authors (cf. Young [15]) that the rate of convergence is increased by a constant factor by using Tchebychev acceleration to choose a sequence of parameters \( \{\gamma_k\} \) instead of the single parameter \( \gamma \).
computed at the k-th step and the case in which just one iterate is computed at each step. In Section 5 we will show that if $y^{(0)}$ is sufficiently close to $y$, then the first of these methods converges quadratically to $y$, while the second, which reduces to a relaxed chord method, converges linearly. However, in either case the solution of the model problem requires a total of $O(N \log N)^2$ arithmetic operations in addition to the cost of the initial factorization of $J(y^{(0)})$ (which we treat as preprocessing).
4. Convergence Results

In this section we present an analysis of the local convergence properties of Newton-iterative methods. We will show that it is possible to adjust the rate of convergence of such methods by varying the number of inner iterations taken at each step and that by taking $2^k$ inner iterations at the $k$-th step, quadratic convergence is obtained. Since the Newton-Richardson methods described in the last section are a particular instance of Newton-iterative methods, the convergence results for them follow immediately. The proofs of the results described here are contained in the appendix.

We begin by restating our problem in a more precise and formal way. If $R$ is a bounded subset of $\mathbb{R}^N$ and $F: R \to \mathbb{R}^N$ is a nonlinear function, we consider the solution of

$$F(x) = 0,$$  \hspace{1cm} (4.1)

where we assume that a solution $x^*$ exists. We also assume that there is an $r_0 > 0$ such that if $S_0 = \{x: ||x - x^*|| < r_0\}$, then

(i) $F$ is differentiable in $S_0$;

(ii) $F'(x) = \frac{\partial F}{\partial x}(x)$ is nonsingular at $x^*$;

(iii) There is an $L < +\infty$ such that for $x \in S_0$,

$$||F'(x) - F'(x^*)|| \leq L||x - x^*||.$$

It can be shown that the discretized model problem of Section 2 satisfies (i) - (iii) (cf. Greenspan and Parter [10]), so that we may directly apply the theory developed here. In the next section we will consider the model problem as a practical illustration of the theory.

From assumptions (i) - (iii), it follows that there is a $0 < r_1 < r_0$ such that $F'$ is continuous and nonsingular in $S_1 = \{x: ||x - x^*|| < r_1\}$.
Thus Newton's method can be used to solve (4.1) for \( \mathbf{x}^* \). For a given \( \mathbf{x}^{(0)} \in S_1 \), we generate the sequence of iterates \( \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots \) by

\[
\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \delta^{(k)},
\]

where

\[
F'(\mathbf{x}^{(k)}) \delta^{(k)} = \mathbf{F}(\mathbf{x}^{(k)}).
\]

It can be shown that for some \( 0 < r_2 \leq r_1 \), Newton's method converges quadratically to \( \mathbf{x}^* \) for \( \mathbf{x}^{(0)} \in S_2 = \{ \mathbf{x} : ||\mathbf{x} - \mathbf{x}^*|| < r_2 \} \) (cf. Ortega and Rheinboldt [11], p. 312).

Newton-Richardson methods are a specific instance of general Newton-iterative methods in which a linear iterative method is used to approximate the solution of (4.3) at each step. A common way of generating Newton-iterative methods involves splitting \( F'(\mathbf{x}) \) by writing it as

\[
F'(\mathbf{x}) = \mathbf{B}(\mathbf{x}) - \mathbf{C}(\mathbf{x}),
\]

where

1. \( \mathbf{B}(\mathbf{x}) \) is continuous at \( \mathbf{x}^* \) and
2. \( \mathbf{B}(\mathbf{x}^*) \) is nonsingular.

Furthermore, if the splitting (4.4) is to lead to a computationally efficient method, it must be possible to quickly solve the system

\[
\mathbf{B}(\mathbf{x}) \mathbf{y} = \mathbf{d}.
\]

It follows from (iv) - (v) that there is a \( 0 < r_3 \leq r_2 \) such that \( \mathbf{B} \) is continuous and nonsingular in \( S_3 = \{ \mathbf{x} : ||\mathbf{x} - \mathbf{x}^*|| < r_3 \} \) (cf. Ortega and Rheinboldt [11], p. 46), and if we let

\[
\mathbf{H}(\mathbf{x}) = \mathbf{B}(\mathbf{x})^{-1} \mathbf{C}(\mathbf{x}) = \mathbf{I} - \mathbf{B}(\mathbf{x})^{-1} \mathbf{F}(\mathbf{x}),
\]

it then follows that \( \mathbf{H} \) is continuous in \( S_3 \).
The Newton-Richardson methods are obtained with the splitting \((4.4)\) in which \(B\) and \(C\) are defined by

\[
B(x) = \frac{1}{\gamma} F'(x(0)), \quad C(x) = B(x) - F'(x),
\]

where \(\gamma\) is a real constant satisfying \(0 < \gamma < 2\) and \(x(0) \in S_1\). Clearly this splitting satisfies (iv) and (v), and once \(F'(x(0))\) has been factored at the first step, the system \((4.5)\) can be solved quickly.

At each step of a Newton-iterative method, we use the linear iteration to generate a sequence of inner iterates \(\delta_0, \delta_1, \delta_2, \ldots\) in which \(\delta_0 = 0\) and

\[
\delta_{i+1} = \delta_i - \frac{\delta_i}{\delta_i - f_i(k)},
\]

where

\[
B(x(k)) \delta_i(k) = F'(x(k)) \delta_i(k) - F(x(k)).
\]

For any positive integer \(m\), let

\[
A_m(x) = (I + H(x) + H(x)^2 + \ldots + H(x)^{m-1}) B(x)^{-1}
\]

\[
= (I - H(x)^m)(I - H(x))^{-1} B(x)^{-1}
\]

\[
= (I - H(x)^m) F'(x)^{-1}.
\]

Then letting

\[
B_k = B(x(k))
\]

and

\[
H_k = H(x(k)),
\]

we have
\[ \delta_{i+1} = \delta_i - B_k^{-1} (F'(x_i) \delta_i - F(x_i)) \]

\[ = H_k \delta_i + B_k^{-1} F(x_i) \]

\[ = H_k \left[ H_k \delta_{i-1} + B_k^{-1} F(x_i) \right] + B_k^{-1} F(x_i) \]

\[ = H_k^{i+1} \delta_0 + (I + H_k + \ldots + H_k^i) B_k^{-1} F(x_i) \]

\[ = (I + H_k + \ldots + H_k^i) B_k^{-1} F(x_i) \]

\[ = A_{i+1} (x_i) F(x_i) . \]

At the \( k \)-th Newton-iterative step, we define

\[ \delta_i = \delta_m, \]

for some positive integer \( m \). Letting

\[ G_m (x) = x - A_m (x) F(x), \]

we have

\[ x_{i+1} = x_i - \delta_i \]

\[ = G_m (x_i). \] (4.6)

If \( \| H(x^*) \| < 1 \), then under assumptions (i) - (iii) it can be shown that for some \( 0 < r_4 < r_3 \), the outer iterates \( \{ x_i \} \) defined by (4.6) exist and converge to \( x^* \) for \( x^*(0) \in S_4 = \{ x : \| x - x^* \| < r_4 \} \) (cf. Ortega and Rheinboldt [11], pp. 350-351).†

† It actually suffices that the spectral radius of \( H(x^*) \) satisfy

\[ \rho(H(x^*)) < 1 \]

but in order to avoid problems in relating norms to the spectral radius, we give all our results in terms of the Euclidean norm.

(If \( H(x^*) \) is symmetric, then of course we have

\[ \| H(x^*) \| = \rho(H(x^*)). \]
The rate at which the Newton-iterative iterates converge to \( \bar{x} \) depends mainly on \( ||H(\bar{x})|| \) and the sequence \( \{m_k\} \). To obtain our main convergence result, we make two further assumptions on the splitting (4.4):

(vi) For some constant \( \lambda, 0 < \lambda < 1 \), \( ||H(\bar{x})|| < \lambda \);

(vii) There are an \( L_1 < +\infty \) and a \( 0 < r_5 < r_4 \) such that for \( x \in S_5 = \{x: ||x - \bar{x}|| < r_5\} \),

\[ ||B(x) - B(\bar{x})|| < L_1||x - \bar{x}||. \]

We now state our main convergence result, deferring its proof to the appendix.

**Theorem 4.1:** Let \( F \) satisfy (i) - (iii) and let \( F'(x) = B(x) - C(x) \) be a splitting which satisfies (iv) - (vii). Let \( m_0, m_1, m_2, \ldots \) be a sequence of positive integers, and define

\[ m = \max \{1, m_0\} \cup \{(m_k - \sum_{\ell=0}^{k-1} m_\ell): k = 1, 2, 3, \ldots\}. \]

Then, if \( m < +\infty \), there are a \( 0 < r < r_5 \) (depending on \( \lambda, H, \) and \( F \)) and a positive constant \( c \) (depending on \( \lambda \) and \( m \)) with \( \lambda < c < 1 \) such that for \( x(0) \in S = \{x: ||x - \bar{x}|| < r\} \), the sequence of iterates defined by (4.6) satisfies

\[ ||x^{(k+1)} - \bar{x}|| \leq c^k ||x^{(k)} - \bar{x}||. \]

While the restriction on \( m \) in Theorem 4.1 may seem contrived, it simply reflects the fact that Newton-iterative methods cannot be expected to exhibit super-quadratic convergence. Hence choosing \( m_k > 2^k \) is pointless. If we choose \( m_k = 1 \) or \( m_k = 2^k \), then \( m = 1 \) in the theorem, and we have the following corollaries.
Corollary 4.2: Let $F$ satisfy (i) - (iii) and let $F'(x) = B(x) - C(x)$ be a splitting which satisfies (iv) - (vii). Define the sequence 
{$m_k: k = 0, 1, 2, \ldots$} by $m_k = 1$. Then there are a $0 < r < r_5$ (depending on $\lambda, H, \text{ and } F$) and a positive constant $c$ (depending on $\lambda$) with $\lambda < c < 1$ such that for $x^{(0)} \in S = \{ x: \| x - x^* \| < r \}$, the sequence of iterates defined by (4.6) converges linearly, i.e.,
\[
\| x^{(k+1)} - x^* \| \leq c \| x^{(k)} - x^* \|.
\]

Corollary 4.3: Let $F$ satisfy (i) - (iii) and let $F'(x) = B(x) - C(x)$ be a splitting which satisfies (iv) - (vii). Define the sequence 
{$m_k: k = 0, 1, 2, \ldots$} by $m_k = 2^k$. Then there are a $0 < r < r_5$ (depending on $\lambda, H, \text{ and } F$) and a positive constant $c$ (depending on $\lambda$) with $\lambda < c < 1$ such that for $x^{(0)} \in S = \{ x: \| x - x^* \| < r \}$, the sequence of iterates defined by (4.6) converges quadratically, i.e.,
\[
\| x^{(k+1)} - x^* \| \leq c^{2^k} \| x^{(k)} - x^* \|.
\]
5. Ramifications of the Theory for the Model Problem

Corollary 4.2 expresses the previously known result that a Newton-iterative method converges linearly for $m_k = 1$ (cf. Ortega and Rheinboldt [11], p. 350). The real import of Theorem 4.1 is the rather surprising result of Corollary 4.3, i.e., that we can obtain $R$-quadratic convergence with a Newton-iterative method by choosing $m_k = 2^k$. Since Newton's method itself is only quadratically convergent (cf. Ortega and Rheinboldt [11], p. 312) this is the fastest convergence that can be expected for any Newton-iterative method.

A problem with the result of Theorem 4.1 is that $c$ depends on $\lambda$: specifically, $\lambda < c < 1$. To reduce the initial error $||v^{(0)} - v||$ by a factor of $K < 1$ requires $l$ outer iterations, where

$$K = \prod_{k=0}^{\lambda-1} c^k.$$  \hspace{1cm} (5.1)

Taking logarithms, we obtain

$$\log K = (\sum_{k=0}^{\lambda-1} m_k)(\log c),$$

so that $l$ satisfies

$$\sum_{k=0}^{\lambda-1} m_k = \log K / \log c \geq \log K / \log \lambda.$$  \hspace{1cm} (5.1)

As an illustration, we consider the model problem of Section 2. For Newton-SOR methods it can be shown that even with the optimal SOR parameter $\omega$, $\log \lambda = -2\pi \theta$ (cf. Young [15], p. 190). Reducing the initial error by a factor of the discretization error $h^2$ requires $l$ iterations, where $l$ satisfies
\[ \sum_{k=0}^{l-1} m_k \approx \log h^2 / \log \lambda \]

\[ = (-\log N) / (-2\pi h) \]

\[ = O(N^{1/2} \log N). \]

Thus if \( m_k = 2^k \), we have

\[ \sum_{k=0}^{l-1} m_k \approx 2^l = O(N^{1/2} \log N), \]

so that \( l = O(\log N) \). And if \( m_k = 1 \), then we have

\[ \sum_{k=0}^{l-1} m_k = l = O(N^{1/2} \log N). \]

In either case, though, since each inner SOR iteration requires \( O(N) \) operations (cf. Young [15]), a total of \( O(N^{3/2} \log N) \) iteration time is required, as compared to the \( O(N^{3/2} \log \log N) \) time required by Newton's method.

For Newton-Richardson methods, however, the situation is quite different. We have

\[ ||H(x^*)|| = ||B(x^*)^{-1} C(x^*)|| \]

\[ = ||\gamma F'(x(0))^{-1} [\frac{1}{\gamma} F'(x(0)) - F'(x^*)]|| \]

\[ = ||I - \gamma F'(x(0))^{-1} F'(x^*)||. \]

Since \( F'(x)^{-1} \) is continuous in \( S_3 \), we may make \( ||H(x^*)|| \) arbitrarily close to \( |1 - \gamma| \) by choosing \( ||x(0) - x^*|| \) small enough. In particular, we can always make \( ||H(x^*)|| < (1 + |1 - \gamma|)/2 < 1 \). For the model problem, then, the constants \( \lambda \) and \( c \) of Theorem 4.1 are independent of \( h \) for Newton-Richardson methods. Thus to reduce the initial error by a factor of \( h^2 \) requires \( l \) outer iterations, where

\[ \sum_{k=0}^{l-1} m_k \approx \log h^2 / \log c = -\log N / \log c. \]
and log c is a constant independent of N. Hence if $m_k = 2^k$, we have

$$\sum_{k=0}^{l-1} m_k = 2^l - \log N / \log c,$$

so that $l = O(\log \log N)$. And if $m_k = 1$, then we have

$$\sum_{k=0}^{l-1} m_k = l = -\log N / \log c = O(\log N).$$

This leads to the following result on the cost of solving the model problem with Newton-Richardson methods (for a proof, see Eisenstat, Schultz, and Sherman [11, Sherman [13]]).

**Theorem 5.1:** Consider the use of Newton-Richardson methods to solve the model semilinear problem presented in Section 2, and assume that $y^{(0)}$ and $y$ are chosen so that Theorem 4.1 holds for $c$ independent of $h$. Then ignoring the costs of preprocessing and the factorization at the first step, we may reduce the initial error $\|y^{(0)} - y\|$ by a factor of $h^2$ in $O(N(\log N)^2)$ iteration time by choosing either $m_k = 2^k$ or $m_k = 1$.

From Theorems 4.1 and 5.1 we conclude that the choice of $m_k$ for Newton-Richardson methods should be based on the relative costs of function and derivative evaluations. For the model problem, choosing $m_k = 2^k$ requires $O(\log \log N)$ function evaluations and $O(\log \log N)$ derivative evaluations, while choosing $m_k = 1$ requires $O(\log N)$ function evaluations and only one derivative evaluation (i.e., $F'(y^{(0)})$). Since function evaluations and derivative evaluations are equally costly for the model problem ($\frac{\partial G}{\partial y} (y^{(k)})$ is a diagonal matrix), it seems best to choose $m_k = 2^k$. However, we have observed experimentally that for other problems it may be more efficient computationally to choose $m_k = 1$ (cf. Sherman [13]). The reason for this is that derivative evaluations are often far more...
costly than function evaluations for problems more general than the model problem.
6. Conclusion

In this paper we have given a general analysis of the local rates of convergence of Newton-iterative methods for the solution of systems of nonlinear equations. In addition, we have presented the Newton-Richardson methods, a particular class of Newton-iterative methods which appear to have computational advantages over both Newton's method and other Newton-iterative methods for sparse nonlinear systems. In practice these methods have performed quite well (cf. Eisenstat, Schultz and Sherman [7], Sherman [13]), although they are by no means good methods for all problems. It is clear that the choice of a method for a particular nonlinear system depends heavily on the nonlinear operator involved and the context in which it arises.

In closing, it is worth pointing out that the ideas underlying the Newton-Richardson methods have important applications in many other settings. Already, a number of authors (e.g., Bank [1], Concus and Golub [2], Douglas and Dupont [5]) have used SIP-like iterations to solve linear systems arising in the numerical solution of partial differential equations, and in the future similar techniques will no doubt be applied to a wide variety of nonlinear and time-dependent problems. In fact, it would seem that the idea of solving difficult problems by iterating on the solutions of nearby, easily-solved problems is one which should prove fruitful whenever it can be applied in a natural way.
Appendix: Proofs of Convergence Results

In this appendix we shall prove Theorem 4.1. We will make frequent use of the assumptions and results of Section 4, and the reader is advised to refer to that section, particularly for the statements of assumptions (i) - (vii). Much of our analysis follows that of Ortega and Rheinboldt [11], pp. 350-352.

Lemma A.1: Let $F$ satisfy (i) - (iii) and let $F'(x) = B(x) - C(x)$ be a splitting which satisfies (iv) - (vi). Then there is a $0 < r_6 < r_4$ such that for $\bar{x} \in S_6 = \{x: \|x - x^*\| < r_6\}$,

$$\|H(x)\| < \lambda$$

and for any positive integer $m$,

$$\|H(x)^m - H(x^*)^m\| < m\lambda^{m-1}\|H(x) - H(x^*)\|.$$  

Proof: Since $H$ is continuous in $S_4$, we can choose $r_6 < r_4$ so that

$$\|H(x)\| < \lambda \text{ for } x \in S_6.$$  

The second part of the lemma follows by induction. It holds trivially for $m = 1$; and if it holds for $m = k$, then

$$\|H(x)^{k+1} - H(x^*)^{k+1}\| = \|H(x)^k (H(x) - H(x^*)) + (H(x)^k - H(x^*)^k)h(x^*)\|$$

$$\leq \lambda^k \|H(x) - H(x^*)\| + k\lambda^{k-1}\|H(x) - H(x^*)\|\|H(x^*)\|$$

$$\leq (k+1)\lambda^k\|H(x) - H(x^*)\|.$$  

Lemma A.2: Let $F$ satisfy (i) - (iii) and let $F'(x) = B(x) - C(x)$ be a splitting which satisfies (iv) - (vi). Then there are positive constants $c_1 < +\infty$ and $c_2 < +\infty$ and a $0 < r_7 < r_6$ such that for any positive integer $m$ and $x \in S_7 = \{x: \|x - x^*\| < r_7\}$,

$$\|G_m(x) - x^*\| < c_1 \|x - x^*\|^2 + c_2\|H(x) - H(x^*)\| \|x - x^*\| + \lambda^m\|x - x^*\|.$$  

Proof: From (iv) - (vi) and Lemma A.1, it follows that we may choose $r_7$. 

Digitized by
INTERNET ARCHIVE
so that there is a $\beta < + \infty$ such that for $\bar{x} \in S_7$, $||H(\bar{x})|| \leq \lambda$ and $||B(\bar{x})^{-1}|| \leq \beta$ (cf. Ortega and Rheinboldt [11], p. 351). Then for $\bar{x} \in S_7$ we have

$$||G_m(\bar{x}) - x^*|| \leq ||\bar{x} - A_m(\bar{x}) F(\bar{x}) - x^* - H(x^*)^m(\bar{x} - x^*)|| + ||H(x^*)^m(\bar{x} - x^*)||$$

$$= ||-A_m(\bar{x}) F(\bar{x}) + (I - H(x^*)^m)(\bar{x} - x^*)|| + ||H(x^*)^m(\bar{x} - x^*)||$$

$$= ||-A_m(\bar{x}) F(\bar{x}) + A_m(\bar{x}) F'(x^*)(\bar{x} - x^*)|| + ||H(x^*)^m(\bar{x} - x^*)||$$

$$\leq ||-A_m(\bar{x})[F(\bar{x}) - F(x^*) - F'(x^*)(\bar{x} - x^*)]|| +$$

$$||A_m(\bar{x})[F'(x) - F'(x^*)](\bar{x} - x^*)|| +$$

$$||[A_m(\bar{x}) F'(x) - A_m(\bar{x}) F'(x^*)](\bar{x} - x^*)|| + \lambda^m ||x - x^*||.$$

Using

$$||A_m(\bar{x})|| = ||(I + H(x) + H(x)^2 + \ldots + H(x)^m)^{-1}B(\bar{x})||$$

$$\leq (1 + \lambda + \lambda^2 + \ldots + \lambda^{m-1})\beta$$

$$\leq \frac{\beta}{1 - \lambda},$$

we can bound the terms of (A.1) individually. From the mean value theorem we obtain

$$||-A_m(\bar{x})[F(\bar{x}) - F(x^*) - F'(x^*)(\bar{x} - x^*)]||$$

$$\leq ||A_m(\bar{x})|| ||x - x^*|| \sup_{0 \leq t \leq 1} ||F'(x^* + t(x - x^*)) - F'(x^*)||$$

$$\leq ||A_m(\bar{x})|| ||x - x^*|| \sup_{0 \leq t \leq 1} ||t(x - x^*)||$$

$$\leq \beta L/(1 - \lambda) ||x - x^*||^2.$$

The second term of (A.1) is bounded by
\[ \left| A_m(x) F'(x) - F'(x^*) \right| (x - x^*) \leq \left| A_m(x) \right| \left| F'(x) - F'(x^*) \right| \left| x - x^* \right| \]
\[ \leq \beta L / (1 - \lambda) \left| x - x^* \right|^2. \]

Lemma A.1 now yields
\[ \left| (A_m(x) F'(x) - A_m(x^*) F'(x^*)) \right| (x - x^*) \right| = \left| (H(x^*)^m - H(x)^m) (x - x^*) \right| \]
\[ \leq m \lambda^{m-1} \left| H(x) - H(x^*) \right| \left| x - x^* \right|. \]

Since \( \lambda < 1 \), we can choose \( c_2 = c_2(\lambda) < +\infty \) so that the set \( \{m \lambda^{m-1}\} \) is bounded by \( c_2 \). Letting \( c_1 = c_1(\lambda, r_7, \beta, L) = 2 \beta L / (1 - \lambda) \) and combining terms, we obtain
\[ \left| G_m(x) - x^* \right| \leq c_1 \left| x - x^* \right| + c_2 \left| H(x) - H(x^*) \right| \left| x - x^* \right| + \lambda^m \left| x - x^* \right|. \]

Proof of Theorem 4.1: Let \( \beta, c_1, c_2, \) and \( r_7 \) be defined as in Lemma A.2 and define
\[ c(t) = ((c_1 + c_2 L_1)t + \lambda_1)^{1/m}. \]

We select \( r < r_8 \) so that \( c(r) \leq c = (\lambda + 1)/2 < 1 \) and choose \( x^{(0)} \in S \). We can now prove the theorem by induction using Lemma A.2. For \( k = 0 \) we have
\[ ||\hat{x}(1) - x^*|| = ||G_{m_0}(\hat{x}(0)) - x^*|| \]

\[ \leq c_1 ||\hat{x}(0) - x^*||^2 + c_2 ||H(\hat{x}(0)) - H(x^*)|| ||\hat{x}(0) - x^*|| + \lambda m_0 ||\hat{x}(0) - x^*|| \]

\[ \leq ((c_1 + c_2 L_1)r + \lambda m_0) ||\hat{x}(0) - x^*|| \]

\[ \leq (c^m - \lambda + \lambda m_0) ||\hat{x}(0) - x^*||. \]

But \( 0 < \lambda m_0 < \lambda < c^m \), so

\[ 0 < c^m - \lambda + \lambda m_0 < c^m < c^m, \]

and we have

\[ ||\hat{x}(1) - x^*|| \leq c^m ||\hat{x}(0) - x^*||. \]

If we assume that the theorem holds for \( 0 \leq k < j \), then for \( k = j \) we have
\[ ||x^{(j+1)} - x^*|| = ||G_{m_j}(x(j)) - x^*|| \]

\[ \leq c_1 ||x(j) - x^*||^2 + c_2 ||H(x(j)) - H(x^*)|| + \lambda ||x(j) - x^*|| \]

\[ \leq (c_1 + c_2 L_1^j) \eta \frac{\beta_0}{m_j} \lambda \leq c_1 + c_2 L_1 \]

\[ \leq [(c_1 + c_2 L_1^j) \eta \frac{\beta_0}{m_j} \lambda \leq c_1 + c_2 L_1 \]

\[ \leq [c^m - \lambda c^{m-1} + \lambda] \leq c^m - \lambda c^{m-1} + \lambda \]

\[ = c^m [1 - \lambda c^{m-1} + \lambda c^{m-1}] \leq c^m [1 - \lambda c^{m-1} + \lambda c^{m-1}] \]

\[ = c^m [1 - \lambda c^{m-1} + \lambda c^{m-1}] \leq c^m [1 - \lambda c^{m-1} + \lambda c^{m-1}] \]

\[ = c^m [1 - \lambda c^{m-1} + \lambda c^{m-1}] \leq c^m [1 - \lambda c^{m-1} + \lambda c^{m-1}] \]

But \( 0 < \lambda < c^m < 1 \), so \( \lambda c^{m-1} < 1 \) and

\[ 0 < \lambda c^{m-1} + \lambda c^{m-1} < \lambda c^{m-1} + \lambda c^{m-1} \]

Hence

\[ -1 < \lambda c^{m-1} + \lambda c^{m-1} < 0, \]

and we have

\[ ||x^{(j+1)} - x^*|| \leq c^m ||x(j) - x^*||. \]
Acknowledgments

This work was performed as part of my doctoral dissertation research in the Department of Computer Science at Yale University. I am most grateful to my advisors, Professors Martin H. Schultz and Stanley C. Eisenstat, for their assistance and helpful comments.
References

[1] Bank, R. E. Marching Algorithms for Elliptic Boundary Value Problems. Ph.D. Dissertation, Division of Engineering and Applied Physics, Harvard University, 1975.

[2] Concus, P. and G. H. Golub. Use of fast direct methods for the efficient numerical solution of non-separable elliptic equations. SIAM Journal on Numerical Analysis 10: 1103-1120, 1973.

[3] Diamond, M. An Economical Algorithm for the Solution of Finite Difference Equations. Ph.D. dissertation, Department of Computer Science, University of Illinois at Urbana-Champaign, 1971.

[4] D'Jakonov, E. G. On certain iterative methods for solving nonlinear differential equations. Proceedings of the Conference on the Numerical Solution of Differential Equations (Scotland, 1969), Springer-Verlag, 1969, pp. 7-22.

[5] Douglas, J. and T. Dupont. Preconditioned conjugate gradient iteration applied to Galerkin methods for a mildly nonlinear Dirichlet problem. Proceedings of the Symposium on Sparse Matrix Computations, Argonne, IL, 1975.

[6] Dupont, T. A factorization procedure for the solution of elliptic difference equations. SIAM Journal on Numerical Analysis 5: 753-782, 1968.

[7] Eisenstat, S. C., M. H. Schultz, and A. H. Sherman. The application of sparse matrix methods to the numerical solution of nonlinear elliptic partial differential equations. In A. Dold and B. Eckmann, editors, Proceedings of the Symposium on Constructive and Computational Methods for Differential Equations, Springer-Verlag, 1974, pp. 131-153.

[8] Eisenstat, S. C., M. H. Schultz, and A. H. Sherman. Efficient implementation of sparse symmetric Gaussian elimination. Proceedings of the AICA International Symposium on Computer Methods for Partial Differential Equations, Bethlehem, Pennsylvania, 1975, pp. 33-39.

[9] Forsythe, G. E. and C. B. Moler. Computer Solution of Linear Algebraic Systems, Prentice-Hall, 1967.

[10] Greenspan, D. and S. V. Parter. Mildly nonlinear elliptic partial differential equations and their numerical solution, II. Numerische Mathematik 7: 129-146, 1965.

[11] Ortega, J. M. and W. C. Rheinboldt. Iterative solution of Nonlinear Equations in Several Variables, Academic Press, 1970.

[12] Richardson, L. F. The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stresses in a masonry dam. Philosophical Transactions of the Royal Society, Series A(210), London, pp. 307-357, 1910.
[13] Sherman, A. H. On the Efficient Solution of Sparse Systems of Linear and Nonlinear Equations. Ph.D. Dissertation, Department of Computer Science, Yale University, 1975.

[14] Stone, H. L. Iterative solution of implicit approximations of multidimensional partial differential equations. SIAM Journal on Numerical Analysis 5: 530-558, 1968.

[15] Young, D. M. Iterative Solution of Large Linear Systems, Academic Press, 1971.
1. **AEC REPORT NO.**
   
   000-2383-0027

2. **TITLE**
   
   On Newton-Iterative Methods for the Solution of Systems of Nonlinear Equations

3. **TYPE OF DOCUMENT** (Check one):
   
   - a. Scientific and technical report
   - b. Conference paper not to be published in a journal:
     - Title of conference
     - Date of conference
     - Exact location of conference
     - Sponsoring organization
   - c. Other (Specify)

4. **RECOMMENDED ANNOUNCEMENT AND DISTRIBUTION** (Check one):
   
   - a. AEC's normal announcement and distribution procedures may be followed.
   - b. Make available only within AEC and to AEC contractors and other U.S. Government agencies and their contractors.
   - c. Make no announcement or distribution.

5. **REASON FOR RECOMMENDED RESTRICTIONS:**

6. **SUBMITTED BY: NAME AND POSITION (Please print or type)**

   C. W. Gear
   Professor and Principal Investigator

   Organization
   
   Department of Computer Science
   University of Illinois at Urbana-Champaign
   Urbana, IL 61801

   Signature
   [Signature]
   Date
   December, 1975

7. **AEC CONTRACT ADMINISTRATOR’S COMMENTS, IF ANY, ON ABOVE ANNOUNCEMENT AND DISTRIBUTION RECOMMENDATION:**

8. **PATENT CLEARANCE:**
   
   - a. AEC patent clearance has been granted by responsible AEC patent group.
   - b. Report has been sent to responsible AEC patent group for clearance.
   - c. Patent clearance not required.
In this paper we consider the local rates of convergence of Newton-iterative methods for the solution of systems of nonlinear equations. We show that under certain conditions on the inner, linear iterative method, Newton-iterative methods can be made to converge quadratically by computing a sufficient number of inner iterates at each step. As an example of this phenomenon, we examine the Newton-Richardson methods which use an inner Richardson-D’Jakonov iteration.
