GLOBAL APPROXIMATE NEWTON METHODS *

RANDOLPH E. BANK† AND DONALD J. ROSE‡

Abstract. We derive a class of globally convergent and quadratically converging algorithms for a system of nonlinear equations $g(u) = 0$, where g is a sufficiently smooth homeomorphism. Particular attention is directed to key parameters which control the iteration. Several examples are given that have successful in solving the coupled nonlinear PDEs which arise in semiconductor device modelling.

AMS subject classifications. 65H10

1. Introduction. In this paper we derive an algorithm for solving the nonlinear system

$$g(u) = 0 \tag{1.1}$$

where $g = (g_1, g_2, \ldots, g_n)^T$ is a sufficiently smooth homeomorphism from $I\!R^n$ to $I\!R^n$. Recall that a homeomorphism is a bijection (1-1, onto) with both g and the inverse map, g^{-1}, continuous. Physically, a homeomorphism means that the process modelled by g has a unique solution x for any set of input conditions y, i.e., $g(x) = y$, and that the solution x varies continuously with input y. Sometimes this notion is referred to as a "well-posed" process g. Actually, the requirement that g be a homeomorphism is a special case of our assumptions, but we defer a more detailed and general discussion to Section 2.

In its generic form the algorithm we propose is well known. Starting at some initial guess u_0, we solve and $n \times n$ linear system

$$M_k x_k = -g(u_k) \equiv -g_k \tag{1.2}$$

and then set

$$u_{k+1} = u_k + t_k x_k. \tag{1.3}$$

We call the method an *approximate Newton method* because M_k will be chosen to be related to the Jacobian $g'(u_k) \equiv g'_k$, in such a manner that x_k approximates the Newton step $w_k = -\{g(u_k)'\}^{-1} g(u_k)$, and because usually $t_k \neq 1$. In many applications, (1.2) can be interpreted as an "inner" iterative method for solving the linear system

$$g'_k w_k = -g_k. \tag{1.4}$$

When g is a smooth homeomorphism, we will show how to choose the damping parameters t_k and the approximate Newton steps x_k such that the u_k *converge* to u^* with $g(u^*) = 0$ *quadratically* for *any* initial u_0 (see Section 2 for the notions of quadratic and more general higher order convergence). The choice of $x_k = w_k$ in (1.4), the damped Newton method, in an important special case.

In Section 2 we show that, for any choice of norm $\|\cdot\|$, the choice $t_k = (1+\mathcal{K}\|g_k\|)^{-1}$ for some sequence $0 \leq \mathcal{K}_k \leq \mathcal{K}_0$ produces the convergence mentioned above. More

*Received by the editors September 17, 1980.
†Department of Mathematics, University of California at San Diego, La Jolla, California 92093.
‡Bell Laboratories, Murray Hill, New Jersey 07974.

precisely, for $\mathcal{K}_0$ sufficiently large, the sequence $\|g_k\|$ decreases monotonically and quadratically to zero. While it is possible in theory to take $\mathcal{K}_k = \mathcal{K}_0$ for all k, such a strategy often leads to the quagmire of slow initial convergence and can prove disastrous in practice (see Sections 3-4). As we shall see, this rule for choosing t_k is motivated by the requirement $t_k \to 1$ such that $1 - t_k = O(\|g_k\|)$. By specifying a formula for picking t_k, we attempt to avoid most of the searching common to other damping strategies.

We also show in Section 2 how to choose the x_k (or M_k) in (1.2) such that the x_k approximates w_k of (1.4). In this setting our analysis continues and extends investigations of approximate Newton methods initiated by Dennis and Moré (see [5, 6]). Motivated by problems in optimization where it may be difficult or undesirable to deal with the Jacobian, g'_k, they choose M_k, for example, such that

$$\|\{M_k - g(u^*)\}(u_{k+1} - u_k)\|\|u_{k+1} - u_k\|^{-1} \to 0 \tag{1.5}$$

to obtain superlinear convergence. Equation (1.5) may not be immediately useful in contexts where M_k represents an iterative process for solving (1.4); contexts, for example, arising from nonlinear PDEs where it is often possible to evaluate g'_k but it may not be easy to solve (1.4) exactly. Sherman [12] discusses such Newton-iterative methods, showing that to obtain quadratic convergence it suffices to take $m_k = O(2^k)$ inner iterations as $k \to \infty$.

Computationally, it is more convenient to measure the extent to which x_k approximates w_k by monitoring the quantity $\|g'_k x_k + g_k\|$, that is, checking the residual of (1.4) when x_k replaces w_k. To obtain quadratic convergence, for example, we choose x_k such that

$$\alpha_k \equiv \frac{\|g'_k x_k + g_k\|}{\|g_k\|} \leq c\|g_k\|, \tag{1.6}$$

$c > 0$, with $x_k \neq 0$, a suggestion also made by Dembo, Eisenstat and Steihaug [4].

The discussions by the above named researchers all deal with local convergence; that is, they examine convergence in a local region containing a root u^* such that the choice $t_k = 1$ is appropriate. We derive (1.6) within a global framework consistent with our choice of t_k; that is we require $\alpha_k = O(\|g_k\|)$ to balance the quantity $1 - t_k = O(\|g_k\|)$.

We also show that conditions such as (1.6) are equivalent to the original conditions imposed by Dennis and Moré. The whole key to our analysis is the judicious use of the Taylor expansion:

$$\begin{aligned} g_{k+1} = g_k &+ g'_k\{u_{k+1} - u_k\} \\ &+ \int_0^1 \{g'(u_k + s(u_{k+1} - u_k)) - g'_k\}\{u_{k+1} - u_k\}\, ds \end{aligned} \tag{1.7}$$

or, using (1.3) and the notation of Section 2

$$\begin{aligned} g_{k+1} = (1 - t_k)g_k &+ t_k\|g_k\|(g'_k x_x + g_k)\|g_k\|^{-1} \\ &+ \int_0^1 G(s; u_{k+1}, u_k) t_k x_k\, ds. \end{aligned} \tag{1.8}$$

Taking norms will lead to

$$\|g_{k+1}\| \leq \|g_k\|\{(1 - t_k) + t_k\alpha_k + t_k^2\beta_k\|g_k\|\} \tag{1.9}$$

under appropriate conditions on the smoothness of g and the sequence M_k^{-1}. Since β_k will be bounded, (1.9) shows it is possible to insure that the $\|g_k\| \to 0$ monotonically and quadratically by forcing each term in braces to be $O(\|g_k\|)$. Finally, note that (1.7) can be interpreted in a Banach space, and the extension of our results to such a setting is immediate (see [13], Section 12.1).

Section 2 contains a detailed discussion of our assumptions and global convergence analysis. Section 3 encodes the analysis of Section 2 into a general algorithm. Section 4 presents a further algorithmic discussion and discusses several important examples. We conclude in Section 5 with some numerical results relevant to the solution of semiconductor device partial differential equations.

The authors acknowledge discussions with S. Eisenstat, I. Sandberg (see [10, 11]), and W. Fichtner and appreciate the support and encouragement of J. McKenna.

2. Parameter Selection and Convergence. Given an arbitrary initial iteration u_0, we consider here the convergence of the iteration (1.2)-(1.3) where the parameters t_k are chosen by the rule

$$t_k = \frac{1}{1+\mathcal{K}_k\|g_k\|}. \tag{2.1}$$

We make the following assumptions on the mapping $g(u)$ and the sequence M_k.

Assumption A1: The closed level set

$$S_0 = \{u|\, \|g(u)\| \le \|g_0\|\} \tag{2.2}$$

is bounded.

Assumption A2: g is differentiable and the Jacobian $g'(u)$ is a continuous and nonsingular on S_0, and the sequence $\|M_k^{-1}\|$ is uniformly bounded, i.e.,

$$\|M_k^{-1}\| \le k_1 \ \text{ on } S_0 \qquad \text{for all } k \ge 0. \tag{2.3}$$

We embed S_0 in the closed convex ball

$$S_1 = \left\{ u|\, \|u\| \le \sup_{v\in S_0} \|v\| + k_1\|g_0\| \right\}. \tag{2.4}$$

Assumption A3: The Jacobian g' is Lipshitz; i.e.,

$$\|g'(u) - g'(v)\| \le k_2\|u-v\|; \qquad u, v \in S_1. \tag{2.5}$$

Without loss suppose $g_k \ne 0$ for all k, and let the quantities $\mathcal{A}_k$, α_k, $\mathcal{B}_k$, β_k be defined as follows:

$$\begin{aligned} \mathcal{A}_k &\equiv \frac{g_k + g_k' x_k}{\|g_k\|}; \\ \alpha_k &\equiv \|\mathcal{A}_k\|; \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} \mathcal{B}_k &\equiv \frac{\int_0^1 \{g'(u_k + st_k x_k) - g_k'\} t_k x_k \, ds}{(t_k\|g_k\|)^2} \\ \beta_k &\equiv \|\mathcal{B}_k\|; \end{aligned} \tag{2.7}$$

for t_k, u_k, and x_k as in (1.2)-(1.3).

The parameters α_k measure the extent to which the x_k of (1.2) differ from the Newton correction ($\alpha_k \equiv 0$). For $u_k \in S_0$ note that

$$\alpha_k \leq k_1 \|g_k' - M_k\|. \tag{2.8}$$

Typically, given the sequence $\|g_k\|$ and $\alpha_0 \in (0,1)$, we will consider the convergence process when all $\alpha_k \leq \alpha_0$ and

$$\alpha_k \leq c\|g_k\|^p \qquad \text{for } p \in (0,1],\ c > 0: \tag{2.9}$$

for example,

$$\alpha_k \leq \alpha_0 \left(\frac{\|g_k\|}{\|g_0\|}\right)^p = \alpha_{k-1}\left(\frac{\|g_k\|}{\|g_{k-1}\|}\right)^p. \tag{2.10}$$

For many Newton-like methods the α_k can be easily computed, and α_0 and p can be specified a priori.

The parameters β_k reflect the size of higher (second) order terms which are ignored in the derivation of a Newton-like method. For $u_{k+1} = u_k + t_k x_k \in \mathcal{S}_1$ and $u_k \in \mathcal{S}_0$, (1.2), (2.3), (2.5), and (2.7) imply

$$\beta_k \leq \frac{k_2}{2}\left(\frac{\|x_k\|}{\|g_k\|}\right)^2 \leq \frac{k_1^2 k_2}{2} \tag{2.11}$$

Suppose $u_{k+1} \in S_1$ and $u_k \in S_0$. Taylor's Theorem ([9], Section 3.2) implies

$$\begin{aligned} g_{k+1} &= g_k + g_k'\{u_{k+1} - u_k\} + \int_0^1 \{g'(u_k + s(u_{k+1} - u_k)) - g_k'\}(u_{k+1} - u_k)\, ds \\ &= (1 - t_k)g_k + \mathcal{A}_k t_k \|g_k\| + \mathcal{B}_k t_k^2 \|g_k\|^2, \end{aligned} \tag{2.12}$$

$\mathcal{A}_k$ and $\mathcal{B}_k$ as above. Equation (2.12) immediately yields the Taylor inequality

$$\|g_{k+1}\| \leq \|g_k\| \left\{(1 - t_k) + \alpha_k t_k + \beta_k t_k^2 \|g_k\|\right\}. \tag{2.13}$$

We will show the sequence $\|g_k\| \to 0$ by analyzing the term in braces.

Proposition 2.1. *Let $\delta \in (0, 1-\alpha_0)$, $\alpha_0 \in (0,1)$ and t_k chosen as in (2.1) where*

$$0 \leq \mathcal{K}_k \leq \mathcal{K}_0 \tag{2.14}$$

and

$$\mathcal{K}_k \geq \frac{k_1^2 k_2}{2(1 - \alpha_k - \delta)} - \frac{1}{\|g_k\|}. \tag{2.15}$$

Assume A1-A3 and all $\alpha_k \leq \alpha_0$. Then

(i) all $u_k \in S_0$, the sequence $\|g_k\|$ is strictly decreasing and $\|g_k\| \to 0$; furthermore,

(ii) $\|g_{k+1}\|/\|g_k\| \to 0$ if and only if $\alpha_k \to 0$, and for any fixed $p \in (0,1]$,

$$\|g_{k+1}\| \leq c_1 \|g_k\|^{1+p} \tag{2.16}$$

if and only if

$$\alpha_k \leq c_2 \|g_k\|^p \tag{2.17}$$

for positive constants c_1 and c_2.

Proof. To show (i), suppose $u_j \in S_0$ and $\|g_j\| < \|g_{j-1}\|$ for $1 \leq j \leq k$. Since $u_{k+1} = u_k + t_k M_k^{-1} g_k$, $\|u_{k+1}\| \leq \|u_k\| + k_1\|g_k\|$ so $u_{k+1} \in S_1$. Thus (2.11) and (2.15) imply

$$\mathcal{K}_k \geq \frac{\beta_k}{1-\alpha_k-\delta} - \frac{1}{\|g_k\|} \tag{2.18}$$

Rearranging (2.18) and using (2.1) shows

$$(1-t_k) + \alpha_k t_k + \beta_k t_k^2 \|g_k\| \leq 1 - \delta t_k, \tag{2.19}$$

hence

$$\|g_{k+1}\| \leq (1-\delta t_k)\|g_k\| \leq (1-\delta t_0)\|g_k\| \tag{2.20}$$

recalling (2.13). Equation (2.20) implies the conclusion (i).

Part (ii) follows from the pair of inequalities

$$\frac{\|g_{k+1}\|}{\|g_k\|} \leq (\mathcal{K}_k + k_1^2 k_2/2)\|g_k\| + \alpha_k \tag{2.21}$$

and

$$\alpha_k \leq (1 + \mathcal{K}_k\|g_k\|)\frac{\|g_{k+1}\|}{\|g_k\|} + (\mathcal{K}_k + k_1^2 k_2/2)\|g_k\| \tag{2.22}$$

since $\|g_k\| \to 0$. Recalling $t_k \leq 1$, Equations (2.21)-(2.22) are immediate from (2.13) and the analogous inequality derived by transposing (2.12). □

Note that (2.15) is satisfied for the constant sequence

$$\mathcal{K}_k = \mathcal{K}_0 \qquad \text{for all } k. \tag{2.23}$$

Furthermore (2.15) allows the choice $\mathcal{K}_k = 0$ when

$$\|g_k\| \leq \frac{2(1-\alpha_k-\delta)}{k_1^2 k_2}. \tag{2.24}$$

However, as noted in Section 1, (2.23) can be quite unsatisfactory, and we have chosen to force $t_k \to 1$ by using (2.1) rather than using a test to determine whether the choice $t_k = 1$ is satisfactory as eventually guaranteed by (2.24); see Sections 3-4.

We will show later that the sequence u_k converges to the root u^* with $g(u^*) = 0$. Recall that convergence is *superlinear* if

$$\|u_{k+1} - u^*\| \leq \eta_k \|u_k - u^*\| \qquad \text{and } \eta_k \to 0; \tag{2.25}$$

it is order $Q-(p+1)$, $(p \in (0,1])$ if

$$\|u_{k+1} - u^*\| \leq c_p \|u_k - u^*\|^{p+1} \qquad c_p > 0. \tag{2.26}$$

Convergence is *R-linear* if

$$\|u_{k+1} - u^*\| \leq \eta_{k+1} \tag{2.27}$$

and if

$$\eta_{k+1} \leq c\eta_k, \qquad c \in (0,1). \tag{2.28}$$

To examine the nature of the convergence of $\{u_k\}$ we consider the relationship between $\|g_k\| \to 0$ and $\|u_k - u^*\| \to 0$. Consider the Taylor expansion

$$0 = g(u^*) = g_k + g_k'\{u^* - u_k\} + \int_0^1 \{g'(u_k + s(u^* - u_k)) - g_k'\}\{u^* - u_k\}\, ds$$

$$= g_k + g_k' t_k x_k + g_k'\{u^* - u_{k+1}\} + \int_0^1 \{g'(u_k + s(u^* - u_k)) - g_k'\}\{u^* - u_k\}\, ds \tag{2.29}$$

Sine g' is continuous and invertible on S_0, $\|(g_k')^{-1}\| \leq k_3$; rearranging the second inequality of (2.29) implies

$$\|u_{k+1} - u^*\| \leq k_3\{(1 - t_k)\|g_k\| + t_k\alpha_k\|g_k\| + (k_2/2)\|u_k - u^*\|^2\}. \tag{2.30}$$

Letting $k_5 = \sup_{u \in S_1} \|g'(u)\|$, note

$$\|g_k\| \leq k_5\|u_k - u^*\|; \tag{2.31}$$

hence,

$$\|u_{k+1} - u^*\| \leq \|u_k - u^*\| k_3\{k_5\alpha_k + (k_5^2\mathcal{K}_0 + k_2/2)\|u_k - u^*\|\} \tag{2.32}$$

using (2.1) and $t_k \leq 1$. Equation (2.32) shows that the convergence of $\{u_k\}$ to u^* is superliner if $\alpha_k \to 0$ and is of order $Q - (p+1)$ if $\alpha_k \leq c_2\|g_k\|^p$, again using (2.31).

Suppose that in some set $S \subseteq S_0$ (2.31) can be extended to

$$k_4\|u_k - u^*\| \leq \|g_k\| \leq k_5\|u_k - u^*\|. \tag{2.33}$$

Then under the conditions of Proposition 2.1, it is immediate from (2.20),, (2.21), and (2.33) that the convergence of $\{u_k\}$ tp u^* is:

(i) R-linear, and furthermore
(ii) superlinear if and only if $\|g_{k+1}\|/\|g_k\| \to 0$, and;
(iii) order $Q - (p+1)$ if and only if $\|g_{k+1}\| \leq c_1\|g_k\|^{p+1}$, $p \in (0,1]$.

In general (2.33) may not be valid on the entire set S_0. However (2.29) implies (2.33) for u_k sufficiently close to u^* as follows. The first inequality in (2.29) leads to

$$\|u_k - u^*\|\{1 - (k_2k_3/2)\|u_k - u^*\|\} \leq k_3\|g_k\|. \tag{2.34}$$

Thus for any $\rho \in (0,1)$ (say $\rho = 1/2$), (2.34) and (2.31) imply (2.33) for $S = S_0 \cap S_\rho$ where

$$S_\rho = \left\{ u |\, \|u - u^*\| \leq \frac{2(1-\rho)}{k_2k_3} \right\} \tag{2.35}$$

and k_4 of (2.33) is $k_4 = \rho k_3^{-1}$. Summarizing Proposition2.1 and the discussion involving (2.25)-(2.35) we have

Theorem 2.2. *Under the conditions of Proposition 2.1*

(i) there exists a $u^* \in S_0$ with $u^* = \lim u_k$ and $g(u^*) = 0$;
(ii) on S_0 the convergence of $\{u_k\}$ to u^* is superlinear or order $O - (p+1)$ if $\alpha_k \to 0$ or $\alpha_k \leq c_2\|g_k\|^p$, respectively;

(iii) *on any set $S = S_0 \cap S_\rho$ as in (2.35), the convergence of $\{u_k\}$ to u^* is at least R-linear; it is superlinear or order $Q-(p+1)$ if and only if $\alpha_k \to 0$ or $\alpha_k \le c_2\|g_k\|^p$, respectively.*

Proof. It remains only to show (i), which we establish by showing that $\{u_k\}$ is a Cauchy sequence. But since

$$\|u_{k+j} - u_k\| = \left\| \sum_{i=k}^{k+j-1} t_i x_i \right\| \le k_1 \sum_i \|g_i\|, \tag{2.36}$$

and the $\|g_k\| \to 0$ with $\|g_{k+1}\| \le c\|g_k\|$, $c < 1$, $\{u_k\}$ is clearly Cauchy with a limit u^* in the closed set S_0. Continuity of g implies that $g(u^*) = \lim g(u_k) = 0$. □

We now have the following global result.

THEOREM 2.3. *Let $G : I\!R^n \to I\!R^n$ be a homeomorphism. Suppose g' is Lipschitz on closed bounded sets and A2 is satisfied. Then, given any u_0, the sequence u_k of (1.2)-(1.3) with t_k as in (2.1) and $\mathcal{K}_k$ as in (2.15) converges to u^* as in Theorem 2.2.*

Proof. Since g' is Lipschitz on closed bounded sets, g' is continuous on $I\!R^n$. Thus $\|g(u)\| \to \infty$ as $\|u\| \to \infty$ since g is a homeomorphism ([9], page 137). Hence S_0 of (2.2) is bounded for any u_0 and A1 and A3 are satisfied. The result now follows from Theorem 2.2. □

As mentioned in Section 1, early investigations by Dennis and Moré examined higher order convergence of approximate Newton methods. In our notation, they characterized convergence by studying the quantity $\|(M_k - g'(u^*))x_k\|/\|x_k\|$ where they chose $x_k = u_{k+1} - u_k$ $(t_k = 1)$. Their results can be recast in the framework of Theorem 2.2 as the following

THEOREM 2.4. *(cf. [6], pages 51-52). In addition to the conditions of Proposition2.1, let $\|M_k\| \le k_6$. Then*

(i) *on any set $S = S_0 \cap S_\rho$, S_ρ as in (2.35), convergence of $\{u_k\}$ to u^* is superlinear if and only if*

$$\frac{\|(M_k - g'(u^*))x_k\|}{\|x_k\|} \to 0 \tag{2.37}$$

while on S_0, (2.37) implies superlinear convergence;

(ii) *on S convergence is $Q-(p+1)$ if and only if*

$$\frac{\|(M_k - g'(u^*))x_k\|}{\|x_k\|} \le \mu_p \|x_k\|^p, \qquad p \in (0,1]. \tag{2.38}$$

Proof. The conclusion follows from the pair of inequalities

$$\alpha_k \le k_1 \frac{\|(M_k - g'(u^*))x_k\|}{\|x_k\|} + k_1 k_2 \|u_k - u^*\|, \tag{2.39}$$

and

$$\frac{\|(M_k - g'(u^*))x_k\|}{\|x_k\|} \le \alpha_k \frac{\|g_k\|}{\|x_k\|} + \frac{k_2}{k_4}\|g_k\|, \tag{2.40}$$

which can be derived from the definitions of α_k and the constants k_i. For example, to show the "only if" part of (ii), we first note that $Q-(p+1)$ convergence implies

(Theorem 2.2) that $\alpha_k \leq c_2\|g_k\|^p \leq c_2 k_6^p \|x_k\|^p$. Hence by (2.40)

$$\begin{aligned}\frac{\|(M_k - g'(u^*))x_k\|}{\|x_k\|} &\leq c_2 k_6^{p+1}\|x_k\|^p + \frac{k_2 k_6}{k_4}\|x_k\| \\ &\leq \mu_p \|x_k\|^p.\end{aligned}$$

The other conclusions follow similarly. □

We conclude this section with some remarks concerning the generality of the analysis presented here.

Remark R1: If Hölder continuity, i.e.,

$$\|g'(u) - g'(v)\| \leq k_e \|u - v\|^e, \qquad v, u \in S_1, \qquad e \in (0,1) \tag{2.41}$$

replaces Lipschitz continuity in A3, the above analysis remains valid with minor modifications including the restriction of the order exponent p to $p \in (0, e]$. In fact, if g' is only uniformly continuous on S_1 (continuous in $\mathbb{R}^n$), then

$$\|g'(u) - g'(v)\| \leq w(\|u - v\|) \tag{2.42}$$

where $w(t)$ is the modulus of continuity for g' on S_1 (see [9], page 64). Again much of the analysis remains valid; however, it is now only possible to obtain superlinear convergence. The restriction on $\mathcal{K}_k$ analogous to (2.15) in this case is

$$w\left(\frac{k_1\|g_k\|}{1 + \mathcal{K}_k\|g_k\|}\right) \leq \frac{1 - \alpha_k - \delta}{k_1}, \tag{2.43}$$

showing it is possible to choose $0 \leq \mathcal{K}_k \leq \mathcal{K}_0$ since w is an isotone continuous function with $w(0) = 0$. See Daniel ([3], Section 4.2, Chapter 8) and Sanberg [11] for discussions of the role of uniform continuity in similar contexts.

Remark R2: Note that (1.2) and (1.3) can be replaced by any procedure for determining x_k such that

$$0 < \|x_k\| \leq k_1\|g_k\|. \tag{2.44}$$

In all our applications, however, x_k can be shown to derive from g_k by a linear relationship of the form (1.2). Furthermore, the bound on $\|M_k^{-1}\|$ usually follows from the continuity and invertibility of g' on S_0 in addition to convergence assumptions on the inner process which determines x_k (see Section 4).

In the spirit of R1, it is also possible to generalize (2.44) to

$$0 < \|x_k\| \leq k_1\|g_k\|^s, \qquad s \in (1/2, 1] \tag{2.45}$$

for g' Lipschitz on S_1. If g' is less smooth, s must be suitably restricted.

Remark R3: As we have seen in Theorems 2.2 and 2.4, the inability to extend (2.33), in general, to the entire set S_0 leads to somewhat disquieting technicalities concerning the necessary conditions on the convergence to zero of the sequence $\{\alpha_k\}$. In the important special case that $g(u)$ is *uniformly monotine* on S_0, i.e.,

$$(g(u) - g(v))^T(u - v) \geq k_7(u - v)^T(u - v), \tag{2.46}$$

the Cauchy-Schwarz inequality implies (in the 2-norm) that

$$k_7\|u - u^*\|_2 \leq \|g(u)\|_2 \qquad \text{on } S_0. \tag{2.47}$$

Hence (2.33) is valid on S_0 and the appropriate statements in Theorems 2.2 and 2.4 can be simplified. In an algorithmic setting, however, note that statements (i) and (ii) of Theorem 2.2 are the real content of the result.

3. Algorithm. We now turn to the computational aspects of the analysis described in Section 2. In particular, we consider the problem of determining the $\mathcal{K}_k$ of (2.1) such that (2.15), or more importantly (2.18), is satisfied.

Note that inequality (2.20) can be rewritten as

$$\delta \leq \left(1 - \frac{\|g_{k+1}\|}{\|g_k\|}\right)\frac{1}{t_k}. \tag{3.1}$$

Since the right-hand side of (3.1) is easily computed, and since we may choose $\delta \in (0, 1-\alpha_0)$, equation (3.1) is a convenient test. Failure to satisfy (3.1) implies $\mathcal{K}_k$ fails to satisfy (2.18). We then increase $\mathcal{K}_k$ and compute new values for t_k and g_k. These increases will eventually lead to $\mathcal{K}_k$ satisfying (2.15), (2.18), and (3.1), and this process leads to convergence as in Proposition 2.1.

Consider the choice of $\mathcal{K}_0$. Given a guess, say $\mathcal{K}_0 = 0$, each failure of the test (3.1) requires a function evaluation of $g(u)$ to compute a new g_1. This aspect of the procedure has the flavor of a line search, but with one important difference. Once a value of $\mathcal{K}_0$ has been accepted, one might reasonable expect to pass (3.1) for $\mathcal{K}_k = \mathcal{K}_0$ on almost all subsequent iterations k. However, note that as $\|g_k\|$ decreases, the right-hand side of (2.18) decreases, suggesting the possibility of taking $\mathcal{K}_k \leq \mathcal{K}_{k-1}$. If the $\mathcal{K}_k$ decrease in a orderly manner (for example $\mathcal{K}_k = \mathcal{K}_{k-1}/10$), we anticipate a process which uses only one function evaluation on most steps. In fact decreasing the $\mathcal{K}_k$ can be important; we have found than an excessively large value of $\mathcal{K}_0$ will cause the convergence of $t_k \to 1$ to be much slower than necessary, delaying the onset of the observed superlinear convergence, and possibly resulting in many iterations.

The above discussion motivates the following algorithm.

Algorithm Global

(1) input u_0, $\delta \in (0, 1-\alpha_0)$
(2) $\mathcal{K} \leftarrow 0$, $k \leftarrow 0$; compute g_0, $\|g_0\|$
(3) compute x_k
(4) $t_k \leftarrow (1 + \mathcal{K}\|g_k\|)^{-1}$
(5) compute u_{k+1}, g_{k+1}, $\|g_{k+1}\|$
(6) if $(1 - \|g_{k+1}\|/\|g_k\|)t_k^{-1} < \delta$
(7) then {if $\mathcal{K} = 0$, then $\mathcal{K} \leftarrow 1$; else $\mathcal{K} \leftarrow 10\mathcal{K}$}; GOTO (4)
(8) else {$\mathcal{K} \leftarrow \mathcal{K}/10$; $k \leftarrow k+1$}
(9) if converge, then return; else GOTO (3)

In Global, failure to satisfy (3.1) causes $\mathcal{K}$ to be increased in line (7). Each failure requires on additional function evaluation on line (5). On line (8), we take $\mathcal{K}/10$ as the initial estimate for $\mathcal{K}_{k+1}$. Alternatively, we have considered $\mathcal{K}_{k+1} = \mathcal{K}_k 4^{j-k-1}$ where j is the last index resulting in a failure of the test on line (6). In practice, we have found these methods for decreasing $\mathcal{K}$ to be a reasonable compromise between the (possibly) conflicting goals of having $t_k \to 1$ quickly and having (3.1) satisfied on the first function evaluation for most steps.

The procedure for increasing $\mathcal{K}$ is also important, and we have found a procedure other than the relatively simple one given on line (7) to be advantageous. In this scheme, one specifies a priori the maximum number of function evaluations to be allowed on a given step, say ℓ (typically $\ell = 10$), The trial values of $\mathcal{K}$ denoted $\mathcal{K}_{k,j}$,

$1 \leq j \leq \ell$, satisfy

$$\mathcal{K}_{k,j} = \left(\frac{1}{\|g_k\|} + \frac{\mathcal{K}_{k-1}}{10}\right)\left(\frac{\|x_k\|}{\mu\|u_k\|}\right)^{((j-1)/(\ell-1))^2} - \frac{1}{\|g_k\|}, \quad \text{for } \frac{\|x_k\|}{\|u_k\|} > \mu. \tag{3.2}$$

This corresponds to the easily implemented formulae

$$\begin{aligned} t_{k,1} &= (1 + \mathcal{K}\|g_k\|/10)^{-1}; \\ t_{k,j} &= t_{k,1}(\mu\|u_k\|\|x_k\|)^{((j-1)/(\ell-1))^2}, \qquad 2 \leq j \leq \ell. \end{aligned} \tag{3.3}$$

We take μ to be a constant on the order of the machine epsilon $\times 10^3$. For small values of $j \geq 2$, $t_{k,j}$ represents a modest decrease of $t_{k,j-1}$. As j increases, $t_{k,j}$ decreases more rapidly until $t_{k,\ell}\|x_k\| = \|u_k\|\mu t_{k,1}$. If (3.1) fails for $t_{k,\ell}$, the calculation is terminated and an error flag set. Equations (3.3) represent a compromise between the conflicting goals of increasing $\mathcal{K}$ slowly (so as not to accept a value which is excessively large) and of finding an acceptable value in few function evaluations.

On line (3) we have not detailed the computation involving (1.2). If $M_k = g'_k$ in (1.2), Global is a damped Newton method and $\alpha_k = 0$ for all k (disregarding round off). Alternatively, (1.2) may represent an iterative process for solving $g'_k x_k + g_k = 0$ terminated when α_k satisfies some tolerance such as (2.9)-(2.10). Such damped Newton methods and other approximate Newton methods are outlined in the following section.

4. Applications. In this section we present several applications for the results in the previous sections. In particular, we show how Newton-iterative methods, Newton-approximate Jacobian methods, and other Newton-like methods fit within our global approximate Newton framework.

4.1. Newton-Iterative Methods. Suppose that x_k in line (3) of algorithm Global is computed by using an iterative method to solve the Newton equations

$$g'_k w_k = -g_k. \tag{4.1}$$

For example, we might use a standard iterative method such as SOR or a Newton-Richardson method where g'_k in (4.1) is replaced by a previous Jacobian $g'_{k'}$. The Newton-Richardson choice is useful when a (possibly sparse) LU factorization of the Jacobian is relatively expensive. Hence the Jacobian is factored infrequently, and in outer iterations where the factorization is not computed, we iterate to approximately solve (4.1) using the last computed factorization. (see [12]).

In all such Newton-iterative methods ([9], Section 7.4), we suppose g'_k has a uniformly convergent splitting on S_0; i.e.,

$$g'_k = A_k - B_k \tag{4.2}$$

with $\|H_k\| \leq \rho_0 < 1$ for all k, where

$$H_k = A_k^{-1} B_k = I - A_k^{-1} g'_k. \tag{4.3}$$

We then compute x_k by computing the *inner iteration*

$$A_k x_{k,m} = B_k x_{k,m-1} - g_k \tag{4.4}$$

until $m = m_k$, taking $x_{k,0} = 0$ and setting $x_k = x_{k,m_k}$. Note that (4.4) can be rewritten as

$$A_k(x_{k,m} - x_{k,m-1}) = -(g'_k x_{k,m-1} + g_k). \tag{4.5}$$

Using induction and (4.3), it can be shown that the $x_{k,m}$ in (4.4) satisfy

$$g'_k(I - H_k^m)^{-1} x_{k,m} = -g_k, \qquad m \geq 1. \tag{4.6}$$

Hence we may identify M_k of (1.2) with $M_k = g'_k(I - H_k^{m_k})^{-1}$, and these M_k satisfy (2.3) since $\|(g'_k)^{-1}\|$ is bounded on S_0 and $\|I - H_k^{m_k}\| \leq 2$.

Notice that the right hand side of (4.5) contains the Newton residual which suggest defining the quantities

$$\alpha_{k,m} \equiv \frac{\|g_k + g'_k x_{k,m}\|}{\|g_k\|} \tag{4.7}$$

in analogy with (2.6) and the quantities $x_{k,m}$. The $\alpha_{k,m}$ are easily computed, certainly when the iteration proceeds as in (4.5) rather then (4.4). Since we have assumed that the A_k and B_k are a convergent splitting, $\alpha_{k,m} \to 0$ as $m \to \infty$. Thus to obtain convergence as discussed in Section 2, we stop the inner iteration when $\alpha_{k,m}$ attains the desired tolerance $\alpha_k \equiv \alpha_{k,m_k}$. For example, to obtain orer $Q-(p+1)$ superlinear convergence, $p \in (0,1]$, we stop after m_k iterations where

$$\alpha_{k,m} \leq \alpha_0 \left(\frac{\|g_k\|}{\|g_0\|}\right)^p, \qquad \alpha_0 \in (0,1), \tag{4.8}$$

as in (2.9)-(2.10).

Note that

$$g_k + g'_{k,m} x_{k,m} = \hat{H}_k^m g_k \tag{4.9}$$

where $H_k = g'_k \hat{H}_k (g'_k)^{-1}$; this implies

$$\alpha_{k,m} \leq \|\hat{H}_k\|^m \tag{4.10}$$

Assuming (4.1) is an equality with $\|\hat{H}_k\| < 1$ and that equation (2.16)-(2.17) are equalities, we see that

$$m_k = \frac{\log c_2 \|g_k\|^p}{\log \|\hat{H}_k\|}. \tag{4.11}$$

Asymptotically (as $k \to \infty$) we expect $\|\hat{H}_{k+1}\| \approx \|\hat{H}_k\|$; again assuming equality and using (2.16)-(2.17) and (4.11) shows

$$m_{k+1} \sim (1+p) m_k. \tag{4.12}$$

Hence for k sufficiently large we can expect the number of inner iterations per outer iteration to approximately increase by a factor of $(1+p)$.

In the preceding general analysis of Newton-iterative methods all $\alpha_k = \alpha_{k,m}$ are possibly nonzero. For the special case of a Newton-Richardson method a decision is made at the beginning of the k-th outer iteration whether to factor g'_k thus doing an "exact" damped Newton iteration. Such a factorization implies $\alpha_k = 0$; otherwise the inner iteration corresponds to a splitting with $A_k = g'_{k'}$, $k' < k$, where A_k has been previously factored.

It is not difficult to decide when to refactor: one should refactor when the total cost of inner iterations using the factored $g'_{k'}$ just surpasses the cost of a new factorization. For example, using nested dissection on an $n \times n$ mesh cost approximately $10n^3$ operations for a factorization and $5n^2 \log_2 n$ operations for a backsolution. Thus approximately $2n/\log_2 n$ inner iterations compared with a new factorization. Note however that initially these $2n/\log_2 n$ inner iterations will be part of several outer iterations monitored by the $\alpha_{k,m}$; each time a new such outer iteration is started, g'_k is computed (but not factored) for use in the right hand side of (4.5). The relative time T_1 corresponding to $10n^3$ and T_2 corresponding to $5n^2 \log_2 n$ can often be timed dynamically using a "clock routine" and need not be known a priori. As a final remark, note that superlinear convergence will require an increasing number of inner iterations, and ultimately, the inner iteration time will surpass the cost of a new factorization. However, in practice when only a modest overall accuracy is required Newton-Richardson methods can prove to be highly effective, and we have found such cases.

4.2. Newton-Approximate Jacobian Methods. When the partial derivatives required for the computation of g'_k are unavailable or expensive to compute, it is common to approximate g'_k, perhaps using finite differences ([6], page 49, [9], pages 185-186). We denote such an approximation to g'_k by $\tilde{g}'_k$.

We assume that the $\tilde{g}'_k$ satisfy

$$\|g'_k - \tilde{g}'_k\| \le \frac{\delta_k}{k_1} \tag{4.13}$$

for $\delta_k < 1$. Let $\mathcal{A}_k$ of (2.6) be written as

$$\mathcal{A}_k = \tilde{\mathcal{A}}_k + \Delta_k \tag{4.14}$$

where

$$\tilde{\mathcal{A}}_k = \frac{g_k + \tilde{g}'_k x_k}{\|g_k\|}; \qquad \Delta_k = \frac{(g'_k - \tilde{g}'_k) x_k}{\|g_k\|}. \tag{4.15}$$

Following (2.6), let $\tilde{\alpha}_k = \|\tilde{\mathcal{A}}_k\|$ and note that $\|\Delta_k\| \le \delta_k$; hence

$$\alpha_k \le \tilde{\alpha}_k + \delta_k. \tag{4.16}$$

If x_k is obtained by the linear system

$$\tilde{g}'_k x_k = -g_k, \tag{4.17}$$

then $\tilde{\alpha}_k = 0$, corresponding to $M_k = \tilde{g}'_k$ in (1.2). Alternatively (4.17) can be solved approximately, perhaps by a Newton-iterative method as in Section 4.1; then $\tilde{\alpha}_k \ne 0$ in general.

If all $\tilde{\alpha}_k = 0$ the δ_k play the role of α_k in Section2. For example, if all $\tilde{\alpha}_k = 0$ and

$$\delta_k \le \delta_0 \left(\frac{\|g_k\|}{\|g_0\|} \right)^p; \qquad p \in (0, 1],\ \delta_0 < 1, \tag{4.18}$$

then the Newton-approximate Jacobian scheme will converge with order $Q-(p+1)$. More generally, let $\tilde{\alpha}_k \neq 0$ and δ_k satisfy (4.18) with

$$\tilde{\alpha}_k \leq \max\left\{\tilde{\alpha}_0\left(\frac{\|g_k\|}{\|g_o\|}\right)^q, \delta_k\right\}, \qquad \tilde{\alpha}_0+\delta_0<1. \tag{4.19}$$

In a typical situation we might have $p=0$, $q=1$, and δ_0 small; i.e., we compute the approximation to g'_k to a fixed accuracy and use an iterative method to solve (4.17) approximately. For the first few outer iterations, relatively few inner iterations will be required. The inner iterations will increase until δ_k becomes the larger of the two terms on the right-hand side of (4.19). From this point onward, approximately a constant number of inner iterations will be used, and the asymptotic outer convergence will be R-linear (Theorem 2.2). It may not be easy to estimate or compute δ_k; although, if we are computing $\tilde{g}'_k$ to a fixed accuracy, we expect all δ_k to be approximately equal. Since δ_k enters the computation only through (4.19), we see that its main purpose is to prevent useless inner iterations. If nothing is known about δ_k, we can set all $\delta_k=\epsilon$ in (4.19), ϵ a sufficiently small number. Convergence of the outer iteration can loosely be described as superlinear at the beginning and ultimately linear depending on the actual values of δ_k. The choice of ϵ could have a significant effect on the total computation cost and may require some experimentation.

4.3. Two Parameter Damping. In an earlier investigation, [1], we studied the Newton-like method

$$(I/s_k+g'_k)x_k=-g_k, \tag{4.20}$$

$$u_{k+1}=u_k+x_k, \tag{4.21}$$

motivating the method by considering Euler integration on the autonomous system of ODEs

$$\frac{du}{dt}+g(u)=0; \qquad u(0)=u_0. \tag{4.22}$$

Under appropriate conditions, including the uniform monotonicity of $g(u)$ on $I\!R^n$ in the form

$$x^T g'(u)x \geq k_7 x^T x, \tag{4.23}$$

we showed that it is possible to obtain global quadratic convergence by forcing $s_k\|g_k\|$ to be a sufficiently small constant for all k. We used a norm reducing argument similar to the analysis of Section 2. Here we sketch a more general treatment using the results of Section 2.

Consider the iteration (1.2)-(1.3) with t_k as in (2.1), $\lambda_k \geq 0$, and

$$M_k=\lambda_k\|g_k\|I+g'_k. \tag{4.24}$$

We call such a method *two parameter damping* because the λ_k as well as the t_k limit the change $(u_{k+1}-u_k)$. It is immediate from (4.24) that the α_k of (2.6) satisfy

$$\alpha_k=\lambda_k\|x\|_k. \tag{4.25}$$

If the $\|M_k\|$ are uniformly bounded (for all λ_k) as in (2.3), then

$$\alpha_k \leq k_1\lambda_k\|g\|_k. \tag{4.26}$$

This shows that the λ_k can be chosen such that $\alpha_0 < 1$ and $\alpha_k \leq \alpha_0$ as in Proposition 2.1. Furthermore, the Q-quadratic convergence (2.16) is also a consequence of (4.26).

The above discussion can be made more precise by seeking a relation between the uniform bound k_1 and λ_k. Suppose that $g(u)$ is uniformly monotone as in (4.23) on S_0. Then, in the 2-norm,

$$\|M_k^{-1}\|_2 \leq (\lambda_k \|g_k\|_2 + k_7)^{-1} \tag{4.27}$$

and

$$\alpha_k = \lambda_k \|x_k\| \leq \frac{\lambda_k \|g_k\|_2}{\lambda_k \|g_k\|_2 + k_7} < \min(1, k_7^{-1} \lambda_k \|g_k\|_2). \tag{4.28}$$

Consider the sequence of λ_k with $0 \leq \lambda_k \leq \lambda_0$. Note that the α_k are easily computable and all $\alpha_k < 1$, although it may not be the case that all $\alpha_k \leq \alpha_0$. This requires a minor modification in Proposition 2.1. We will assume that for each $k \geq 0$, $\delta \in (0, 1 - \hat{\alpha}_k)$ where $\hat{\alpha}_k = \max_{j \leq k} \alpha_j$. Note the $\sup \hat{\alpha}_k < 1$ by (4.28) and the induction argument leading to (2.20). Since we do not know $\hat{\alpha}_k$ a priori, we may be required to change (decrease) δ dynamically as the iteration proceeds; that is, if for some α_k, $\delta \geq 1 - \alpha_k$. Such decreases in δ cause no convergence problems since (3.1) continues to hold if δ is decreased in subsequent iterations.

It is possible to show that for $\lambda_k = \lambda_0$, $k \geq 0$, with λ_0 sufficiently large, t_k can be chosen as $t_k = 1$ for all k. One starts with (2.13) and uses (4.28) and $\beta_k \leq (k_2/2)(\lambda_0 \|g_k\|_2 + k_7)^{-2}$ as in [1], Section 3. However, such an analysis (without the explicit damping parameter t_k) is more existential than the two parameter analysis sketched above, mainly since the sufficient decrease parameter analogues to δ of (2.20) depends on the usually unknown constant k_7. In practice, there seems to be no advantage in using only λ damping, whereas two parameter damping may be advantageous when the g'_k themselves are numerically ill-conditioned.

5. Numerical Remarks. The methods described in this work and our earlier presentation [1], are part of a larger study aimed at solving effectively the coupled partial differential equations arising in semiconductor device modelling. These equations often take the form

$$-\Delta u + e^{u-v} - e^{w-u} = k(x, y) \tag{5.1}$$

$$-\nabla \cdot (\mu_n e^{u-v} \nabla v) = 0 \tag{5.2}$$

$$-\nabla \cdot (\mu_p e^{w-u} \nabla w) = 0 \tag{5.3}$$

Here u, v, and w are functions of $(x, y) \in D \subseteq \mathbb{R}^2$, as are the known functions μ_n, μ_p and $k(x, y)$, and D is a union of rectangles. The function $k(x, y)$ is the doping profile of the device; (5.1) is a nonlinear Poisson equation and (5.2)-(5.3) are continuity equations.

Equations (5.1)-(5.3) have be attacked numerically on two discretization fronts. Finite differences are used in W. Fichtner's simulation package; the now routine solution of the coupled equations is reported in [8, 7]. In this package, Newton-Richardson and Newton-block SOR methods (as in Section 4) have proved to be particularly effective.

The device equations, especially (5.1), have also been attacked by a nonlinear multilevel iteration package using piecewise linear elements on triangles. This package has been designed concurrently with developing analysis presented in this paper and is an extension of the linear package described in [2]. The package presently solves a single nonlinear PDE of the form

$$-\nabla \cdot (a(x, y)\nabla u) + f(u, u_x, u_y) = 0 \tag{5.4}$$

on a connected region Ω in $\mathbb{R}^2$ with standard elliptic boundary conditions; the formal generalization to the case

$$-\nabla \cdot (a(x, y, u, u_x, u_y)\nabla u) + f(u, u_x, u_y) = 0 \tag{5.5}$$

is straightforward. A special Newton-multilevel iterative method, along the lines discussed in Section 4, is used to solve the discrete equations. Details will be presented elsewhere.

To illustrate the use of the Newton-multilevel iteration package and Algorithm Global, consider, as in [1], Section 4, the $p - n$ junction problem of the form (5.1) above. The functions v, w, and k are given, and the domain and boundary conditions are shown in Figure 5.1. Recall that the doping profile $k(x, y)$, and the solution gradient, $\nabla u(x, y)$, vary over several orders of magnitude in a small region near the junction, and there is a notable singularity due to the change in boundary conditions along the upper boundary.

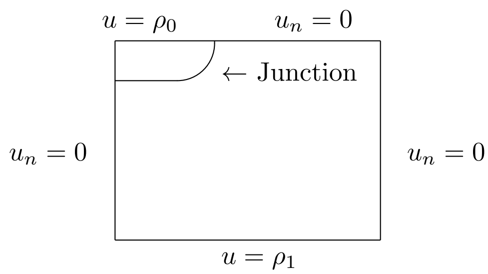

FIG. 5.1. *$p - n$ junction problem with boundary conditions*

We consider only the level-one nonuniform grid with $n = 25$ vertices (unknowns) and a (very poor) initial guess, $u_0 = 0$. (The higher levels are less interesting.) We use Algorithm Global as in Section 4 with the modification (3.2)-(3.3) in line (7). The x_k are computed by a sparse LU factorization of g'_k. The convergence trace is presented in Table 5.1.

The relatively large number of iterations necessary in this experiment compared with the experiments reported in [1], Section 4, is a direct consequence of taking $u_0 = 0$ rather than attempting to even roughly interpolate the boundary values ρ_0 and ρ_1 as we did before. This also leads to more searching (evals) than might otherwise be expected.

As a cautionary remark, we report that failing to dynamically change $\mathcal{K}$ (i.e., taking $\mathcal{K}_k = \mathcal{K}_0$ for all k) led to time overrun termination after $k = 320$, $t_k = 1.106(-4)$ and $\|g_k\| = 4.57(6)$ in the same experiment.

REFERENCES

[1] R. E. Bank and D. J. Rose, *Parameter selection for Newton-like methods applicable to nonlinear partial differential equations*, SIAM J. Numer. Anal., 17 (1980), pp. 806–822.

$\lVert g_0 \rVert = 4.738 \cdot 10^6$; evals $\equiv$ evaluations of $g(u)$				
k	t_k	$\lVert g_k \rVert$	$\lVert u_u - u_{k-1} \rVert / \lVert u_k \rVert$	Evals
1	1.07(-4)	4.73(6)	9.28(3)	6
2	7.30(-4)	4.70(6)	8.11(2)	2
3	7.31(-3)	4.56(6)	9.16(1)	1
4	7.05(-2)	3.41(6)	1.12(1)	1
5	2.81(-2)	3.31(6)	1.33(1)	4
6	2.30(-1)	2.21(6)	6.26(-1)	1
7	2.44(-1)	5.27(5)	1.33	3
8	9.31(-1)	5.88(4)	4.37(-1)	1
9	6.15(-2)	3.64(4)	2.33	4
10	1.52(-1)	3.11(4)	1.56	3
11	6.78(-1)	1.29(4)	1.41	1
12	9.81(-1)	5.73(3)	1.39(-1)	1
13	9.99(-1)	4.08(3)	1.28(-1)	1
14	7.39(-2)	3.79(3)	4.39(-1)	4
15	3.48(-1)	3.00(3)	2.67(-1)	2
16	8.71(-1)	1.75(3)	1.36(-1)	1
17	3.18(-1)	9.51(2)	2.83(-1)	3
18	8.95(-1)	5.18(2)	3.49(-2)	1
19	7.73(-1)	3.58(2)	1.98(-2)	2
20	7.66(-1)	1.03(2)	1.41(-2)	2
21	9.83(-1)	1.37(1)	3.26(-3)	1
22	1.[a]	1.44	2.24(-4)	1
23	1.[a]	1.57(-1)	2.36(-5)	1
24	1.[a]	7.67(-6)	2.90(-6)	1
25	1.[a]	4.27(-7)	1.40(-10)	1

[a] 3 place rounding

Table 5.1

[2] R. E. Bank and A. H. Sherman, *PLTMG user's guide*, Tech. Rep. CNA-152, Center for Numerical Analysis, University of Texas at Austin, 1979.

[3] J. W. Daniel, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[4] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact-Newton methods*, Tech. Rep. SOM Working Paper Series number 47, Yale University, 1981.

[5] J. E. Dennis and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[6] ———, *Quasi-Newton methods: motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[7] W. Fichtner and D. J. Rose, *On the numerical solution of nonlinear PDEs arising from semiconductor device modelling*, in Elliptic Problem Solvers, (M. H. Schultz, ed.), Academic Press, New York, 1980.

[8] ———, *Numerical semiconductor device simulation*, Tech. Rep. Technical memo, Bell Laboratories, 1981.

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

[10] I. W. Sandberg, *One Newton-direction algorithms and diffeomorphisms*, Tech. Rep. Technical memo, Bell Laboratories, 1980.

[11] ———, *Diffeomorphisms and Newton-direction algorithms*, Bell Sys. Tech. J., (1981).

[12] A. H. Sherman, *On Newton-iterative methods for the solution of systems of nonlinear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 755–771.

[13] A. Wouk, *A Course of Applied Functional Analysis*, J. Wiley and Sons, New York, 1979.