Choosing the Forcing Terms in an Inexact Newton Method

Stanley C. Eisenstat
Homer F. Walker

CRPC-TR94463
May, 1994

Center for Research on Parallel Computation
Rice University
P.O. Box 1892
Houston, TX 77251-1892

Submitted to *SIAM Journal on Scientific Computing*,
May, 1994.

CHOOSING THE FORCING TERMS IN AN INEXACT NEWTON METHOD *

STANLEY C. EISENSTAT† AND HOMER F. WALKER‡

Abstract. An inexact Newton method is a generalization of Newton's method for solving $F(x) = 0$, $F : R^n \rightarrow R^n$, in which, at the kth iteration, the step s_k from the current approximate solution x_k is required to satisfy a condition $\|F(x_k) + F'(x_k)\, s_k\| \leq \eta_k \|F(x_k)\|$ for a "forcing term" $\eta_k \in [0, 1)$. In typical applications, the choice of the forcing terms is critical to the efficiency of the method and can affect robustness as well. Promising choices of the forcing terms are given, their local convergence properties are analyzed, and their practical performance is shown on a representative set of test problems.

Key words. forcing terms, inexact Newton methods, Newton iterative methods, truncated Newton methods, Newton's method, iterative linear algebra methods, GMRES

AMS(MOS) subject classifications. 65H10, 65F10

1. Introduction.

Suppose that $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable in a neighborhood of $x_* \in \mathbb{R}^n$ for which $F(x_*) = 0$ and $F'(x_*)$ is nonsingular. Suppose further that F' is Lipschitz continuous at x_* with constant λ, i.e.,

$$\|F'(x) - F'(x_*)\| \leq \lambda \|x - x_*\| \tag{1.1}$$

for x near x_*, where $\|\cdot\|$ denotes some norm on $\mathbb{R}^n$ and the induced norm on $\mathbb{R}^{n\times n}$.

An *inexact Newton method* (Dembo, Eisenstat, and Steihaug [4]) is an extension of classical Newton's method for approximating x_* formulated as follows:

Algorithm IN: Inexact Newton Method [4]

```
LET $x_0$ BE GIVEN.
FOR $k = 0$ STEP 1 UNTIL "CONVERGENCE" DO:
    FIND some $\eta_k \in [0,1)$ AND $s_k$ THAT SATISFY
(1.2)        $\|F(x_k) + F'(x_k)\, s_k\| \leq \eta_k \|F(x_k)\|$.
    SET $x_{k+1} = x_k + s_k$.
```

Note that (1.2) expresses both a certain reduction in the norm of $F(x_k) + F'(x_k)\, s$, the local linear model of F, and a certain accuracy in solving the *Newton equation* $F'(x_k)s = -F(x_k)$, the exact solution of which is the *Newton step*. In many applications, notably Newton iterative or truncated Newton methods[1], each η_k is specified first, and then an s_k is determined so that (1.2) holds. The role of η_k is, then, to force

* Submitted to *SIAM Journal on Scientific Computing*, May, 1994.

† Department of Computer Science, Yale University, P. O. Box 208285, New Haven, CT 06520-8285 (eisenstat-stan@cs.yale.edu). The research of this author was supported in part by U. S. Army Research Office contract DAAL03-91-G-0032.

‡ Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900 (walker@math.usu.edu). The work of this author was supported in part by United States Air Force Office of Scientific Research Grant AFOSR-91-0294 and United States Department of Energy Grant DE-FG02-92ER25136, both with Utah State University. It was done in part during a visit to the Computational and Applied Mathematics Department and the Center for Research on Parallel Computation, Rice University.

[1] These are implementations of Newton's method in which iterative linear algebra methods are used to solve the Newton equation approximately.

$\|F(x_k) + F'(x_k)\, s_k\|$ to be small in a particular way; accordingly, η_k is often called a *forcing term.*

The local convergence of an inexact Newton method is controlled by the forcing terms. Some specific illustrative results are the following (see Dembo, Eisenstat, and Steihaug [4]): Under the present assumptions, if x_0 is sufficiently close to x_* and $0 \le \eta_k \le \eta_{\max} < 1$ for each k, then $\{x_k\}$ converges to x_* q-linearly in the norm $\|\cdot\|_*$, defined by $\|v\|_* \equiv \|F'(x_*)v\|$ for $v \in I\!\!R^n$, with asymptotic rate constant no greater than $\eta_{\max}$. Furthermore, if $\lim_{k\to\infty} \eta_k = 0$, then the convergence is q-superlinear, and if $\eta_k = O(\|F(x_k)\|)$, then the convergence is q-quadratic.

In addition to controlling local convergence, there is another important issue associated with the forcing terms. Away from a solution, F and its local linear model may disagree considerably at a step that closely approximates the Newton step. Thus choosing η_k too small may lead to *oversolving* the Newton equation, by which we mean imposing an accuracy on an approximation of the Newton step that leads to significant disagreement between F and its local linear model. Oversolving may result in little or no decrease in $\|F\|$ and, therefore, little or no progress toward a solution. Moreover, in applications such as Newton iterative or truncated Newton methods, in which additional accuracy in solving the Newton equation requires additional expense, it may entail pointless costs; a less accurate approximation of the Newton step may be both cheaper and more effective.

Our purpose is to propose choices of the forcing terms that achieve desirably fast local convergence and also tend to avoid oversolving. All of the proposed choices incorporate information about F but are scale independent in that they do not change if F is multiplied by a constant.

In §2, we outline the proposed choices and analyze the local convergence of Algorithm IN that results from them; we also note some practical safeguards that improve performance. In §3, we discuss numerical experiments. The algorithm used in the experiments is a special case of Algorithm IN and is outlined in §3.1. The test problems are described in §3.2. An example of oversolving is given in §3.3, with additional observations and examples in §3.4. Summary test results are shown in §3.5. Conclusions are given in §4.

Preliminaries. We define some useful constants and formulate several elementary results. Set $M \equiv \max\{\|F'(x_*)\|, \|F'(x_*)^{-1}\|\}$. For $\delta > 0$, define

$$N_\delta(x_*) \equiv \{x \in I\!\!R^n : \|x - x_*\| < \delta\},$$

and let $\delta_* > 0$ be sufficiently small that

1. F is continuously differentiable and F' is nonsingular on $N_{\delta_*}(x_*)$,
2. $\|F'(x)^{-1}\| \le 2M$ for $x \in N_{\delta_*}(x_*)$,
3. inequality (1.1) holds for $x \in N_{\delta_*}(x_*)$,
4. $\delta_* < 2/(\lambda M)$.

LEMMA 1.1. *If $x \in N_{\delta_*}(x_*)$ and if s is such that $x_+ \equiv x + s \in N_{\delta_*}(x_*)$, then*

$$\|F(x_+) - F(x) - F'(x)\, s\| \le \lambda \left(2\|x - x_*\| + \frac{\|s\|}{2}\right) \|s\|.$$

Proof. Setting $x(t) \equiv x + ts$ for $0 \leq t \leq 1$, we have

$$\begin{aligned} \|F(x_+) - F(x) - F'(x)\,s\| &= \| \int_0^1 F'(x(t))\,s\,dt - F'(x)\,s\| \\ &\leq \| \int_0^1 \left[F'(x(t)) - F'(x_*)\right] dt - \left[F'(x) - F'(x_*)\right] \| \, \|s\| \\ &\leq \left(\int_0^1 \lambda \left[\|x - x_*\| + t\|s\|\right] dt + \lambda \|x - x_*\| \right) \|s\| \\ &= \lambda \left(2\|x - x_*\| + \frac{\|s\|}{2} \right) \|s\|. \end{aligned}$$

□

LEMMA 1.2. *There is a $\mu > 0$ such that*

$$\frac{1}{\mu}\|x - x_*\| \leq \|F(x)\| \leq \mu \|x - x_*\|.$$

whenever $x \in N_{\delta_*}(x_*)$.

Proof. With Lemma 1.1, we have

$$\begin{aligned} \|F(x)\| &\leq \|F'(x_*)(x - x_*)\| + \|F(x) - F(x_*) - F'(x_*)(x - x_*)\| \\ &\leq M\|x - x_*\| + \frac{\lambda}{2}\|x - x_*\|^2 \leq \left(M + \frac{\lambda\delta_*}{2}\right) \|x - x_*\| \end{aligned}$$

and

$$\begin{aligned} \|F(x)\| &\geq \|F'(x_*)(x - x_*)\| - \|F(x) - F(x_*) - F'(x_*)(x - x_*)\| \\ &\geq \frac{1}{M}\|x - x_*\| - \frac{\lambda}{2}\|x - x_*\|^2 \geq \left(\frac{1}{M} - \frac{\lambda\delta_*}{2}\right) \|x - x_*\|. \end{aligned}$$

The lemma follows with $\mu \equiv \max\left\{M + \lambda\delta_*/2, (1/M - \lambda\delta_*/2)^{-1}\right\}$. □

LEMMA 1.3. *If $x \in N_{\delta_*}(x_*)$ and $\|F(x) + F'(x)\,s\| \leq \eta\|F(x)\|$ for some s and $\eta \in [0, 1)$, then $\|s\| \leq 4M\|F(x)\|$.*

Proof. We have

$$\begin{aligned} \|s\| &\leq \|F'(x)^{-1}\| \|F'(x)\,s\| \\ &\leq 2M(\|F(x)\| + \|F(x) + F'(x)\,s\|) \\ &\leq 2M(1 + \eta)\|F(x)\| \leq 4M\|F(x)\|. \end{aligned}$$

□

LEMMA 1.4. *There is a $B > 0$ such that if $x \in N_{\delta_*}(x_*)$ and if s and $\eta \in [0, 1)$ are such that $\|F(x) + F'(x)\,s\| \leq \eta\|F(x)\|$ and $x_+ \equiv x + s \in N_{\delta_*}(x_*)$, then*

$$\|F(x_+)\| \leq (\eta + B\|F(x)\|)\|F(x)\|.$$

Proof. With Lemmas 1.1–1.3, we have that

$$\begin{aligned} \|F(x_+)\| &\leq \|F(x) + F'(x)\,s\| + \|F(x_+) - F(x) - F'(x)\,s\| \\ &\leq \eta\|F(x)\| + \lambda\left(2\|x - x_*\| + \frac{\|s\|}{2}\right) \|s\| \\ &\leq \eta\|F(x)\| + \lambda(2\mu\|F(x)\| + 2M\|F(x)\|) \cdot 4M\|F(x)\| \\ &= (\eta + B\|F(x)\|)\|F(x)\|, \end{aligned}$$

where $B \equiv 8\lambda M(\mu + M)$. □

2. The proposed choices. In the analysis in this section, we use the Lipschitz constant λ in (1.1) and the constants M, δ_*, μ, and B introduced in the preliminaries in §1. We also let δ be such that $0 < \delta \leq \delta_*/(1+4\mu M)$ and note that, by Lemmas 1.2 and 1.3, if $x \in N_\delta(x_*)$ and $\|F(x)+F'(x)\,s\| \leq \eta\|F(x)\|$ for some s and $\eta \in [0,1)$, then $x+s \in N_{\delta_*}(x_*)$. We assume for convenience that Algorithm IN continues indefinitely without termination and that $F(x_k) \neq 0$ for all k. Note that if $x_k \in N_{\delta_*}(x_*)$, then $F'(x_k)$ is nonsingular and suitable s_k and x_{k+1} exist for any $\eta_k \in [0,1)$. Our standing assumptions on F and x_* are those made in the first paragraph of §1.

Our first choice is the following:

Choice 1: Given $\eta_0 \in [0,1)$, choose

$$\eta_k = \frac{\|F(x_k) - F(x_{k-1}) - F'(x_{k-1})\,s_{k-1}\|}{\|F(x_{k-1})\|}, \qquad k=1,2,\ldots, \tag{2.1}$$

or

$$\eta_k = \frac{\Big|\|F(x_k)\| - \|F(x_{k-1}) + F'(x_{k-1})\,s_{k-1}\|\Big|}{\|F(x_{k-1})\|}, \qquad k=1,2,\ldots. \tag{2.2}$$

Note that η_k given by either (2.1) or (2.2) directly reflects the agreement between F and its local linear model at the previous step. The choice (2.2) may be more convenient to evaluate than (2.1) in some circumstances. Since it is at least as small, local convergence will be at least as fast as with (2.1); however, if it is significantly smaller, then it may be more difficult to find a suitable step in some applications and perhaps risk greater oversolving as well.

THEOREM 2.1. *Under the standing assumptions on F and x_*, if x_0 is sufficiently near x_*, then $\{x_k\}$ produced by Algorithm IN with $\{\eta_k\}$ given by Choice 1 remains in $N_{\delta_*}(x_*)$ and converges to x_* with*

$$\|x_{k+1} - x_*\| \leq \beta\|x_k - x_*\|\|x_{k-1} - x_*\|, \qquad k=1,2,\ldots, \tag{2.3}$$

for a constant β independent of k.

Remark: It follows immediately from (2.3) that the convergence is q-superlinear and two-step q-quadratic. As in the case of the classical secant method, it also follows that the convergence is of r-order $(1+\sqrt{5})/2$; see, e.g., [13, p. 293] for the argument.

Proof. It suffices to prove the theorem with $\{\eta_k\}$ given by (2.1).

Suppose that $\eta_0 \in [0,1)$ is given. Let τ be such that $\eta_0 < \tau < 1$, and let $\epsilon > 0$ be sufficiently small that $\eta_0 + B\epsilon \leq \tau$, $[8\lambda M(\mu+M)+B]\epsilon \leq \tau$, and $\epsilon < \delta/\mu$. Note that if $x \in N_{\delta_*}(x_*)$ and $\|F(x)\| \leq \epsilon$, then $x \in N_\delta(x_*)$ by Lemma 1.2.

Let $x_0 \in N_\delta(x_*)$ be sufficiently near x_* that $\|F(x_0)\| \leq \epsilon$. Since $x_0 \in N_\delta(x_*)$, we have $x_1 \in N_{\delta_*}(x_*)$. Also, by Lemma 1.4,

$$\begin{aligned} \|F(x_1)\| &\leq (\eta_0 + B\|F(x_0)\|)\|F(x_0)\| \leq (\eta_0 + B\epsilon)\|F(x_0)\| \\ &\leq \tau\|F(x_0)\| \leq \|F(x_0)\| \leq \epsilon, \end{aligned} \tag{2.4}$$

and, hence, $x_1 \in N_\delta(x_*)$.

As an inductive hypothesis, suppose that, for some $k \geq 1$, we have $x_k \in N_\delta(x_*)$, $x_{k-1} \in N_\delta(x_*)$, $\|F(x_k)\| \leq \epsilon$, and $\|F(x_{k-1})\| \leq \epsilon$. Lemmas 1.1–1.3 give

$$\begin{aligned} \eta_k &= \frac{\|F(x_k) - F(x_{k-1}) - F'(x_{k-1})\,s_{k-1}\|}{\|F(x_{k-1})\|} \\ &\leq \frac{\lambda(2\|x_{k-1} - x_*\| + \|s_{k-1}\|/2)\|s_{k-1}\|}{\|F(x_{k-1})\|} \\ &\leq \frac{\lambda(2\mu\|F(x_{k-1})\| + 2M\|F(x_{k-1})\|)\cdot 4M\|F(x_{k-1})\|}{\|F(x_{k-1})\|} \\ &\leq 8\lambda M(\mu + M)\|F(x_{k-1})\|. \end{aligned}$$

Then Lemma 1.4 implies

$$\begin{aligned} \|F(x_{k+1})\| &\leq (\eta_k + B\|F(x_k)\|)\|F(x_k)\| \\ &\leq \Big[8\lambda M(\mu + M)\|F(x_{k-1})\| + B\|F(x_k)\|\Big]\|F(x_k)\| \\ &\leq [8\lambda M(\mu + M) + B]\epsilon\|F(x_k)\| \leq \tau\|F(x_k)\| \\ &\leq \|F(x_k)\| \leq \epsilon. \end{aligned} \tag{2.5}$$

Thus $\|F(x_{k+1})\| \leq \epsilon$ and, hence, $x_{k+1} \in N_\delta(x_*)$.

It follows from this induction that $\{x_k\} \subset N_\delta(x_*) \subset N_{\delta_*}(x_*)$. Furthermore, (2.4) and (2.5) give $\|F(x_{k+1})\| \leq \tau\|F(x_k)\|$ for each $k \geq 0$; hence, $F(x_k) \rightarrow 0$ and, by Lemma 1.2, $x_k \rightarrow x_*$ as well.

To show (2.3), we note that (2.4) and (2.5) give, for $k \geq 1$, $\|F(x_k)\| \leq \|F(x_{k-1})\|$ and

$$\begin{aligned} \|F(x_{k+1})\| &\leq \Big[8\lambda M(\mu + M)\|F(x_{k-1})\| + B\|F(x_k)\|\Big]\|F(x_k)\| \\ &\leq [8\lambda M(\mu + M) + B]\|F(x_{k-1})\|\|F(x_k)\|. \end{aligned}$$

With Lemma 1.2, this implies (2.3) with $\beta \equiv \mu^3[8\lambda M(\mu + M) + B]$. □

A possible way to obtain faster local convergence while retaining the potential advantages of (2.1) and (2.2) is to raise those expressions to powers greater than one. A particular possibility that we considered in our numerical experiments is the following:

Choice 1^2*:* Given $\eta_0 \in [0, 1)$, choose

$$\eta_k = \left(\frac{\|F(x_k) - F(x_{k-1}) - F'(x_{k-1})\,s_{k-1}\|}{\|F(x_{k-1})\|}\right)^2, \qquad k = 1, 2, \ldots, \tag{2.6}$$

or

$$\eta_k = \left(\frac{\big|\|F(x_k)\| - \|F(x_{k-1}) + F'(x_{k-1})\,s_{k-1}\|\big|}{\|F(x_{k-1})\|}\right)^2, \qquad k = 1, 2, \ldots. \tag{2.7}$$

This choice was not as successful in our experiments as other choices proposed here; see §3. For completeness, we state without proof the following local convergence theorem.

THEOREM 2.2. *Under the standing assumptions on F and x_*, if x_0 is sufficiently near x_*, then $\{x_k\}$ produced by Algorithm IN with $\{\eta_k\}$ given by Choice* 1^2 *remains in $N_{\delta_*}(x_*)$ and converges to x_* with*

$$(2.8)\quad \|x_{k+1} - x_*\| \leq \beta \max\left\{\|x_{k-1} - x_*\|^2, \|x_k - x_*\|\right\} \|x_k - x_*\|, \quad k = 1, 2, \ldots,$$

for a constant β independent of k.

Remark: It follows from (2.8) that the convergence is r-quadratic.

Our second choice is the following:

Choice 2: Given $\gamma \in [0,1]$ and $\eta_0 \in [0,1)$, choose

$$(2.9)\qquad \eta_k = \gamma \left(\frac{\|F(x_k)\|}{\|F(x_{k-1})\|}\right)^2, \qquad k = 1, 2, \ldots.$$

The choice (2.9) does not directly reflect the agreement between F and its local linear model, as does Choice 1. However, the experiments in §3 show that it results in little oversolving in practice, and the following theorem shows that it gives faster guaranteed local convergence than Choice 1.

THEOREM 2.3. *Under the standing assumptions on F and x_*, if x_0 is sufficiently near x_*, then $\{x_k\}$ produced by Algorithm IN with $\{\eta_k\}$ given by Choice 2 remains in $N_{\delta_*}(x_*)$ and converges to x_*. If $\gamma < 1$, then the convergence is q-quadratic. If $\gamma = 1$, then the convergence is r-quadratic and of q-order p for every $p \in [1, 2)$.*

Proof. Suppose that $\eta_0 \in [0,1)$ is given and let $\epsilon > 0$ be sufficiently small that $\eta_0 + B\epsilon \leq \sqrt{\eta_0}$ and $\epsilon < \delta/\mu$. Note that if $x \in N_{\delta_*}(x_*)$ and $\|F(x)\| \leq \epsilon$, then $x \in N_\delta(x_*)$ by Lemma 1.2.

Let $x_0 \in N_\delta(x_*)$ be sufficiently near x_* that $\|F(x_0)\| \leq \epsilon$. As an inductive hypothesis, suppose that, for some $k \geq 0$, we have $x_k \in N_\delta(x_*)$, $\|F(x_k)\| \leq \epsilon$, and $\eta_k \leq \eta_0$. Since $x_k \in N_\delta(x_*)$, we have $x_{k+1} \in N_{\delta_*}(x_*)$. Also, by Lemma 1.4,

$$\begin{aligned} \|F(x_{k+1})\| &\leq (\eta_k + B\|F(x_k)\|)\|F(x_k)\| \\ (2.10)\qquad &\leq (\eta_0 + B\epsilon)\|F(x_k)\| \leq \sqrt{\eta_0}\,\|F(x_k)\| \\ &\leq \|F(x_k)\| \leq \epsilon. \end{aligned}$$

Then $\|F(x_{k+1})\| \leq \epsilon$, and it follows that $x_{k+1} \in N_\delta(x_*)$. Furthermore, (2.10) gives

$$\eta_{k+1} = \gamma(\|F(x_{k+1})\|/\|F(x_k)\|)^2 \leq \gamma\eta_0 \leq \eta_0.$$

It follows from this induction that $\{x_k\} \subset N_\delta(x_*) \subset N_{\delta_*}(x_*)$. Furthermore, (2.10) gives $\|F(x_{k+1})\| \leq \sqrt{\eta_0}\|F(x_k)\|$ for each $k \geq 0$; hence, $F(x_k) \to 0$ and, by Lemma 1.2, $x_k \to x_*$ as well.

It remains to show the desired rates of convergence. Note that, for $k > 0$, (2.10) and (2.9) give

$$(2.11)\qquad \|F(x_{k+1})\| \leq \left[\gamma \left(\frac{\|F(x_k)\|}{\|F(x_{k-1})\|}\right)^2 + B\|F(x_k)\|\right] \|F(x_k)\|.$$

First, suppose that $\gamma < 1$ and set $\rho_k \equiv \|F(x_k)\|/\|F(x_{k-1})\|^2$ for $k > 0$. From (2.11), we have $\rho_{k+1} \leq \gamma\rho_k + B$ for $k > 0$, and it follows inductively that

$$\rho_{k+1} \leq \gamma^k \rho_1 + \left(\sum_{j=0}^{k-1} \gamma^j\right) B \leq \rho_1 + \frac{B}{1-\gamma}.$$

Thus $\{\rho_k\}$ is uniformly bounded. Consequently, $F(x_k) \to 0$ q-quadratically, and it follows from Lemma 1.2 that $x_k \to x_*$ q-quadratically as well.

Now, suppose that $\gamma = 1$. We first show that the convergence is of q-order p for $p \in [1, 2)$. For $k > 0$, (2.11) gives

$$\begin{aligned} \|F(x_{k+1})\| &\le \left[\left(\frac{\|F(x_k)\|}{\|F(x_{k-1})\|}\right)^2 + B\|F(x_k)\|\right] \|F(x_k)\| \\ &= \left[\left(\frac{\|F(x_k)\|}{\|F(x_{k-1})\|}\right)^{2-p} \frac{\|F(x_k)\|}{\|F(x_{k-1})\|^p} + B\|F(x_k)\|^{2-p}\right] \|F(x_k)\|^p. \end{aligned} \tag{2.12}$$

For each $k > 0$, set $\sigma_k \equiv \|F(x_k)\|/\|F(x_{k-1})\|^p$ and recall that (2.10) gives $\|F(x_k)\| \le \sqrt{\eta_0}\,\|F(x_{k-1})\|$, whence $\|F(x_k)\| \le (\eta_0)^{k/2}\,\|F(x_0)\|$. Then for $k > 0$, (2.12) implies

$$\sigma_{k+1} \le \eta_0^{1-p/2}\sigma_k + B\eta_0^{k(1-p/2)}\|F(x_0)\|^{2-p} = \xi\sigma_k + \xi^k C,$$

where $\xi \equiv \eta_0^{1-p/2}$ and $C \equiv B\|F(x_0)\|^{2-p}$. It follows inductively that

$$\sigma_{k+1} \le \xi^k\left(\sigma_1 + kC\right),$$

and, hence,

$$\|F(x_{k+1})\| \le \xi^k\left(\sigma_1 + kC\right)\|F(x_k)\|^p.$$

Since $\xi^k\left(\sigma_1 + kC\right) \to 0$ as $k \to \infty$, we conclude that $F(x_k) \to 0$ with q-order p and, by Lemma 1.2, $x_k \to x_*$ with q-order p as well.

Still assuming $\gamma = 1$, we now show that $x_k \to x_*$ r-quadratically. By Lemma 1.2, it suffices to show that $\|F(x_k)\| \to 0$ r-quadratically; we shall prove the somewhat stronger result that $\alpha_k \equiv \|F(x_k)\|/\|F(x_{k-1})\| \to 0$ r-quadratically.

It follows from the results above that $\alpha_k \to 0$. Then there is a k_0 such that $2\alpha_{k_0+1} + 2B\|F(x_{k_0})\| \le 1$. For convenience, we re-index if necessary so that $k_0 = 0$. Then $2\alpha_1 + 2B\|F(x_0)\| \le 1$, which implies $D \equiv 1/(2\alpha_1) > 1$. Set $\beta_k \equiv D\alpha_k$ for $k \ge 0$. Note that $\beta_1 = 1/2$. It suffices to show that $\beta_k \to 0$ r-quadratically.

We claim that $\beta_k \le \beta_1^{2^{k-1}}$ for $k = 1, 2, \ldots$, from which it follows that $\beta_k \to 0$ r-quadratically. The claim clearly holds for $k = 1$. Suppose that it holds up to some $k \ge 1$. Then Lemma 1.4 implies

$$\|F(x_{k+1})\| \le \left(\alpha_k^2 + B\|F(x_k)\|\right)\|F(x_k)\|,$$

whence

$$\alpha_{k+1} \le \alpha_k^2 + B\alpha_k \ldots \alpha_1\|F(x_0)\|.$$

From this we obtain

$$\begin{aligned} \beta_{k+1} &\le \frac{1}{D}\beta_k^2 + \frac{B\|F(x_0)\|}{D^{k-1}}\beta_k \ldots \beta_1 \\ &\le \frac{1}{D}\left(\beta_1^{2^{k-1}}\right)^2 + B\|F(x_0)\|\beta_1^{(2^{k-1}+\ldots+1)} \\ &= \frac{1}{D}\beta_1^{2^k} + B\|F(x_0)\|\beta_1^{(2^k-1)} = \left(\frac{1}{D} + B\|F(x_0)\|/\beta_1\right)\beta_1^{2^k} \\ &= \left(2\alpha_1 + 2B\|F(x_0)\|\right)\beta_1^{2^k} \le \beta_1^{2^k}, \end{aligned}$$

and the proof is complete. □

Practical safeguards. Although the forcing term choices given above are usually effective in avoiding oversolving, we have observed in experiments that they occasionally become too small far away from a solution. There is a particular danger of the Choice 1 and 1^2 forcing terms becoming too small. Indeed, an η_k given by (2.1), (2.2), (2.6), or (2.7) can be undesirably small because of either a very small step or coincidental very good agreement between F and its local linear model. We have found the following safeguard to be practically effective for Choice 1:

Choice 1 safeguard: Modify η_k by $\eta_k \leftarrow \max\{\eta_k, \eta_{k-1}^2\}$.

The stringency of this safeguard depends on the size of η_{k-1}, which reflects previous agreement between F and its local linear model. Note that, with this safeguard, we have $\eta_k \geq \eta_{k-1}^2$ for all $k > 0$. It is possible for this safeguard to remain active and modify η_k given by (2.1) or (2.2) for arbitrarily large k, and so the convergence of (2.3) may no longer hold. (Indeed, if F is linear, then this safeguard gives $\eta_k = \eta_{k-1}^2$ for *all* k.) However, in almost all of our experiments, this safeguard eventually became inactive. Furthermore, the following lemma shows that the convergence will still be fast, even if the safeguard remains active for all k. (For perspective, recall from the remark after Theorem 2.1 that the convergence of (2.3) implies convergence of r-order $(1+\sqrt{5})/2$.)

LEMMA 2.4. *Suppose that $\{x_k\}$ is produced by Algorithm IN with $\eta_k = \eta_{k-1}^2$ for all $k > 0$. Under the standing assumptions on F and x_*, if x_0 is sufficiently near x_*, then $x_k \rightarrow x_*$ r-quadratically.*

Proof. If $\eta_0 = 0$, then the local convergence is q-quadratic, so assume $\eta_0 \in (0,1)$. Let k_0 be such that $\tau \equiv 2\eta_{k_0} < 1$. Set $\beta_k \equiv B\|F(x_k)\|$ for each k. By the results of Dembo, Eisenstat, and Steihaug [4] mentioned in §1, we can assume that x_0 is sufficiently near x_* that $x_k \rightarrow x_*$ and that $\eta_{k_0}\beta_{k_0} + (\beta_{k_0})^2 \leq 2\eta_{k_0}^2$. For convenience, we re-index if necessary so that $k_0 = 0$; then

$$\tau \equiv 2\eta_0 < 1 \quad \text{and} \quad \eta_0\beta_0 + \beta_0^2 \leq 2\eta_0^2. \tag{2.13}$$

We claim that, for $k > 0$,

$$\beta_k \leq \frac{1}{2}(2\eta_0)^{2^k} = \frac{1}{2}\tau^{2^k},$$

from which it follows that $\|F(x_k)\| \rightarrow 0$ r-quadratically and, with Lemma 1.2, that $x_k \rightarrow x_*$ r-quadratically as well. From Lemma 1.4, we have

$$\|F(x_{k+1})\| \leq (\eta_k + B\|F(x_k)\|)\,\|F(x_k)\|,$$

whence

$$\beta_{k+1} \leq \eta_k\beta_k + \beta_k^2 = \eta_0^{2^k}\beta_k + \beta_k^2. \tag{2.14}$$

It follows from (2.13) and (2.14) that

$$\beta_1 \leq \eta_0\beta_0 + \beta_0^2 \leq 2\eta_0^2 = \frac{1}{2}(2\eta_0)^2,$$

and the claim holds for $k = 1$. If the claim holds for some $k \geq 1$, then (2.14) implies

$$\begin{aligned} \beta_{k+1} &\leq \eta_0^{2^k}\beta_k + \beta_k^2 \leq \eta_0^{2^{k+1}}\left[2^{2^k-1} + \left(2^{2^k-1}\right)^2\right] \\ &\leq \eta_0^{2^{k+1}}\left[2\cdot\left(2^{2^k-1}\right)^2\right] = \frac{1}{2}(2\eta_0)^{2^{k+1}}, \end{aligned}$$

and the claim follows. □

For Choice 1^2, the safeguard below was effective in the experiments reported in §3. This safeguard is the same as that for Choice 1 when η_{k-1} is large but is more relaxed when η_{k-1} is small, which is appropriate in view of the faster convergence of Choice 1^2.

Choice 1^2 safeguard: Modify η_k by $\eta_k \leftarrow \max\{\eta_k, \eta_{k-1}^2\}$ whenever $\eta_{k-1}^2 > .1$ and by $\eta_k \leftarrow \max\{\eta_k, \eta_{k-1}^{2.5}\}$ whenever $\eta_{k-1}^2 \leq .1$.

In our experiments, we observed fewer occasions on which the Choice 2 forcing terms became undesirably small. However, the following safeguard resulted in improved performance.

Choice 2 safeguard: Modify η_k by $\eta_k \leftarrow \max\{\eta_k, \gamma\eta_{k-1}^2\}$ whenever $\gamma\eta_{k-1}^2 > .1$.

This safeguard results in no modification of η_k whenever $\gamma\eta_{k-1}^2 \leq .1$. Consequently, it eventually becomes inactive and does not alter the local convergence results given by Theorem 2.3.

Finally, we note that, away from a solution, it may be possible for each of the proposed choices to be greater than one. Accordingly, it may be necessary in practice to impose an additional safeguard, as in the algorithm in §3.1 below, to make sure that $\eta_k \in [0, 1)$ for each k.

3. Numerical experiments.

In this section, we report on numerical experiments with the forcing term choices outlined in §2, modified with the given safeguards. (For computational convenience, we always used η_k given by (2.2) for Choice 1 and (2.7) for Choice 1^2.) For a broader comparison, we also include the following representative choices that have appeared in the literature: (1) The choice $\eta_k = 10^{-4}$ used by Cai, Gropp, Keyes, and Tidriri [3]. This choice requires uniformly close approximations of Newton steps for all k and results in fast local linear convergence in the norm $\|\cdot\|_*$. (2) The choice $\eta_k = 1/2^{k+1}$ of Brown and Saad [2]. This choice results in local q-superlinear convergence and allows relatively inaccurate approximations of Newton steps for small k, when x_k may not be near x_*; however, it incorporates no information about F. (3) The choice $\eta_k = \min\{1/(k+2), \|F(x_k)\|\}$ of Dembo and Steihaug [5]. This choice results in q-quadratic local convergence and also may allow relatively inaccurate approximations of Newton steps for small k. It incorporates some information about F; however, it does not reflect the agreement of F and its local linear model and, in addition, depends on the scale of F.

3.1. The algorithm.

A globalized inexact Newton algorithm was necessary because initial approximate solutions were not always near a solution. We used Algorithm INB of Eisenstat and Walker [6, §6]. This is an inexact Newton method globalized by backtracking, which we write here as follows:

Algorithm INB: Inexact Newton Backtracking Method [6]

```
LET $x_0$, $\eta_{\max} \in [0,1)$, $t \in (0,1)$, AND $0 < \theta_{\min} < \theta_{\max} < 1$ BE GIVEN.
FOR $k = 0$ STEP 1 UNTIL "CONVERGENCE" DO:
    CHOOSE AN **initial** $\eta_k \in [0, \eta_{\max}]$ AND $s_k$ SUCH THAT
            $\|F(x_k) + F'(x_k)\, s_k\| \leq \eta_k \|F(x_k)\|$.
    WHILE $\|F(x_k + s_k)\| > [1 - t(1 - \eta_k)]\|F(x_k)\|$ DO:
        CHOOSE $\theta \in [\theta_{\min}, \theta_{\max}]$.
        UPDATE $s_k \longleftarrow \theta s_k$ AND $\eta_k \longleftarrow 1 - \theta(1 - \eta_k)$.
    SET $x_{k+1} = x_k + s_k$.
```

Note that Algorithm INB requires $\eta_k \in [0, \eta_{\max}]$ for each initial η_k. For the safeguarded choices in §2, this necessitates the additional safeguard $\eta_k \leftarrow \min\{\eta_k, \eta_{\max}\}$.

Theorem 6.1 of Eisenstat and Walker [6] states that if $\{x_k\}$ generated by Algorithm INB has a limit point x_* such that $F'(x_*)$ is invertible, then $F(x_*) = 0$ and $x_k \to x_*$. Furthermore, in this case, the initial η_k and s_k are accepted without modification for all sufficiently large k; it follows in particular that the asymptotic convergence to x_* is determined by the initial η_k's.

In implementing Algorithm INB, we first chose each initial η_k and then determined an initial s_k by approximately solving the Newton equation using GMRES(m), the restarted GMRES method of Saad and Schultz [11], with restart value $m = 20$. Products of $F'(x_k)$ with vectors were evaluated analytically in some cases and approximated by finite differences of F-values in others; see §3.2. When finite-difference approximations were used, a second-order central difference was used to evaluate the initial residual at the beginning of each cycle of 20 GMRES steps, and subsequently first-order forward differences were used within the cycle. This selective second-order differencing gave essentially the same accuracy as if central differences had been used throughout (see Turner and Walker [15]).

The parameters used were $\eta_{\max} = 1 - 10^{-4}$, $t = 10^{-4}$, $\theta_{\min} = 1/10$, and $\theta_{\max} = 1/2$. The norm was the Euclidean norm $\|\cdot\|_2$. In the while-loop, each θ was chosen to minimize over $[\theta_{\min}, \theta_{\max}]$ the quadratic $p(\theta)$ for which $p(0) = g(0)$, $p'(0) = g'(0)$, and $p(1) = g(1)$, where $g(\theta) \equiv \|F(x_k + \theta s_k)\|_2^2$. Convergence was declared when either $\|F(x_k)\|_2 \leq 10^{-12}\|F(x_0)\|_2$ or $\|s_k\|_2 \leq 10^{-12}$. These tight stopping tolerances allowed asymptotic convergence behavior to become evident. Failure was declared when one of the following occurred: (1) k reached 200 without convergence, (2) an initial s_k was not found in 1000 GMRES(20) iterations, or (3) ten iterations of the while-loop failed to produce an acceptable step. All computing was done in double precision on Sun Microsystems workstations using the Sun Fortran compiler.

3.2. The test problems.

The test set consists of four PDE problems and two integral equation problems. The PDE problems are all elliptic boundary value problems posed on $\Omega \equiv [0,1] \times [0,1] \subseteq \mathbb{R}^2$.

3.2.1. A PDE problem.

The problem is

$$\Delta u + u^3 = 0 \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega.$$

This problem has multiple solutions, but only one that is positive everywhere (McKenna [9], Schaaf [12]). These properties appear to be shared by the discretized problem, and finding the everywhere-positive solution can be difficult without a good initial approximate solution. Discretization was by the usual centered differences on a 100×100 uniform grid, so that $n = 10^4$. The discretized problem was preconditioned on the right using a fast Poisson solver from FISHPACK (Swartztrauber and Sweet [14]). Products of F' with vectors were evaluated analytically. The initial approximate solution was a discretization of $u_0(x) \equiv \alpha x_1(1 - x_1)x_2(1 - x_2)$, which should lead to the everywhere-positive solution for large α. Two test cases were considered: $\alpha = 100$ and $\alpha = 1000$. For the latter value, the initial approximate solution is farther from the solution and the problem is harder.

3.2.2. The (modified) Bratu problem. The problem is

$$\Delta u + \alpha \frac{\partial u}{\partial x_1} + \lambda e^u = 0 \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega.$$

The actual Bratu (or Gelfand) problem has $\alpha = 0$; see, e.g., Glowinski, Keller, and Reinhart [7] or the description by Glowinski and Keller in the collection of nonlinear model problems assembled by Moré [10, pp. 733-737]. As α and λ grow, solving the Newton equations for the discretized problem becomes harder for GMRES(20). Discretization and preconditioning were as in §3.2.1. Products of F' with vectors were evaluated analytically. The initial approximate solution was zero. Two test cases were considered: $\alpha = \lambda = 10$ and $\alpha = \lambda = 20$.

3.2.3. The driven cavity problem. The problem is

$$\begin{aligned}(1/Re)\Delta^2\psi + \frac{\partial\psi}{\partial x_1}\frac{\partial}{\partial x_2}\Delta\psi - \frac{\partial\psi}{\partial x_2}\frac{\partial}{\partial x_1}\Delta\psi &= 0 \quad \text{in } \Omega,\\ \psi = 0 \quad \text{and} \quad \frac{\partial\psi}{\partial n} &= g \quad \text{on } \partial\Omega,\end{aligned}$$

where $g(x_1, x_2) = 1$ if $x_2 = 1$ and $g(x_1, x_2) = 0$ if $0 \leq x_2 < 1$. This is a widely used test problem; see, e.g., Brown and Saad [2] or Glowinski, Keller, and Reinhart [7]. The numerical problem becomes harder as the Reynolds number Re increases. Discretization was by piecewise-linear finite elements on a uniform 63×63 grid[2], so that $n = 3969$. The discretized problem was preconditioned on the right using a fast biharmonic solver of Bjørstad [1]. Products of F' with vectors were approximated with finite differences. The initial approximate solution was zero. Two test cases were considered: $Re = 100$ and $Re = 500$.

3.2.4. The porous medium equation. The problem considered here is

$$\Delta\left(u^2\right) + d\frac{\partial}{\partial x_1}(u^3) + f = 0 \text{ in } \Omega,$$

with $u = 1$ on the bottom and left sides of Ω and $u = 0$ on the top and right sides. This is more or less a steady-state special case of a general problem considered by van Duijn and de Graaf [16]. Discretization was by the usual centered differences on 64×64 uniform grid, so that $n = 4096$. The discretized problem was preconditioned on the right using the tridiagonal part of the Jacobian. Products of F' with vectors were evaluated analytically. The function f was a point source of magnitude 50 at the lower left grid point. The initial approximate solution was a discretization of $u_0(x) \equiv 1 - x_1 x_2$, which tended to require more backtracking for negative d and to cause more oversolving for positive d. Two test cases were considered: $d = 50$ and $d = -50$.

3.2.5. An integral equation. The problem, from Kelley and Northrup [8], is

$$cu(x)^2 - \frac{1}{2}\int_0^1 \cos(yu(x))u(y)\,dy + \frac{1}{2}\sin 1 - c = 0, \quad x \in [0, 1].$$

[2] We thank P. N. Brown for providing the code for this.

Clearly, $u(x) \equiv 1$ is always a solution, and there exist other solutions for at least some values of c. The discretized problem was determined by approximating integrals using 20-point Gaussian quadrature over 20 subintervals of $[0,1]^3$, so that $n = 400$. No preconditioning was necessary. Products of F' with vectors were approximated with finite differences. The initial approximate solution was a discretization of $u_0(x) \equiv 1 + \alpha \cos 9\pi x$. One test case was considered: $c = \alpha = 1.25$.

3.2.6. The Chandrasekhar H-equation. The problem is

$$u(x) - \frac{1}{1 - Lu(x)} = 0, \quad x \in [0,1],$$

where

$$Lu(x) \equiv \frac{c}{2} \int_0^1 \frac{xu(\xi)}{x+\xi} \, d\xi.$$

This problem arises in radiative transfer problems; see, e.g., the description by Kelley in the Moré problem collection [10, pp. 737-739]. The continuous problem is singular at $c = 1$, and so is the discretized problem considered here with discretization as in §3.2.5. The discretized problem becomes more difficult to solve as $c \to 1$ but is still tractable at $c = 1$. As in §3.2.5, no preconditioning was necessary. Products of F' with vectors were approximated with finite differences. The initial approximate solution was zero. Three test cases were considered: $c = .5$, $c = .999$, and $c = 1$.

3.3. An example of oversolving. Algorithm INB with the Dembo–Steihaug [5] choice $\eta_k = \min\{1/(k+2), \|F(x_k)\|_2\}$ was applied to the driven cavity problem with $Re = 500$. The results are shown in Figure 3.1, in which the (base 10) logarithms of the norms of F and its local linear model are plotted as dotted and solid curves, respectively, versus the numbers of GMRES(20) iterations. Triangles indicate the start of new inexact Newton steps. In this example, $\eta_k = \|F(x_k)\|_2$ for each $k > 0$; the safeguard value $\eta_k = 1/(k+2)$ was never chosen for $k > 0$.

In Figure 3.1, gaps between the solid and dotted curves indicate oversolving. Note that once oversolving begins, there is virtually no further reduction in $\|F\|_2$ until the beginning of the next inexact Newton step; thus further GMRES(20) iterations represent wasted effort. Note also the vertical discontinuity in the dotted curve at the end of the fourth inexact Newton step (after 45 GMRES(20) iterations); this indicates a reduction of the initial inexact Newton step through backtracking.

To show the benefits gained by reducing oversolving, we applied Algorithm INB with η_k given by the safeguarded Choice 1 to the driven cavity problem with $Re = 500$. The results are shown in Figure 3.2. Note that oversolving is almost eliminated and there are no step reductions through backtracking. Also, the total number of GMRES(20) iterations is 221, compared to 327 in the previous case. However, the number of inexact Newton steps is 12, compared to 10 previously.

3.4. Additional observations and examples. In an algorithm such as the implementation of Algorithm INB used here, choosing a very small forcing term may risk more than needless expense in obtaining an unnecessarily accurate solution of the Newton equation. First, if oversolving results, then disagreement between F and its

[3] We thank C. T. Kelley for providing the code for this.

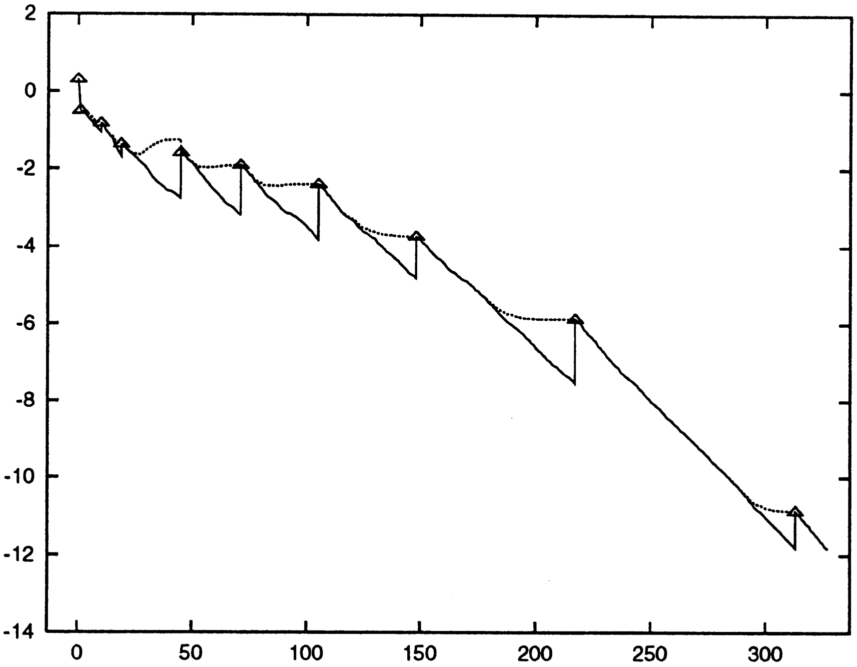

FIG. 3.1. *Illustration of oversolving with* $\eta_k = \min\{1/(k+2), \|F(x_k)\|_2\}$ *on the driven cavity problem with* $Re = 500$. *The horizontal axis indicates the number of* GMRES(20) *iterations. The solid curve is* $\log_{10} \|F + F's\|_2$; *the dotted curve is* $\log_{10} \|F\|_2$. *Triangles indicate new inexact Newton steps.*

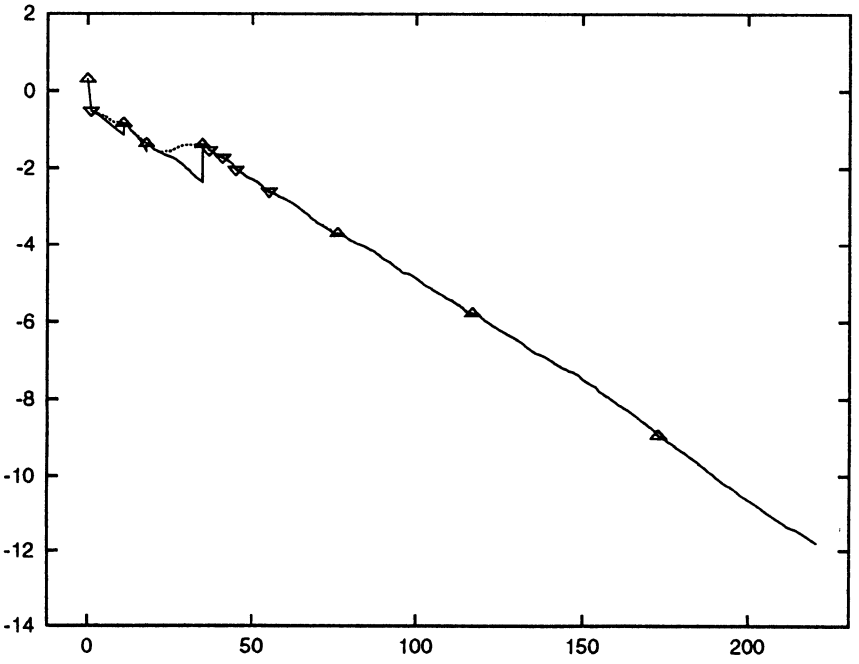

FIG. 3.2. *Illustration of reduction of oversolving with the safeguarded Choice 1 forcing terms on the driven cavity problem with* $Re = 500$. *The horizontal axis indicates the number of* GMRES(20) *iterations. The solid curve is* $\log_{10} \|F + F's\|_2$; *the dotted curve is* $\log_{10} \|F\|_2$. *Triangles indicate new inexact Newton steps: "Δ" indicates* η_k *given by Choice 1; "∇" indicates* $\eta_k = \eta_{k-1}^2$.

local linear model may require significant work from the globalization procedure or even cause it to fail. In the example in §3.3, the choice $\eta_k = \min\{1/(k+2), \|F(x_k)\|_2\}$ required one backtracking, while the safeguarded Choice 1 did not. We observed a more dramatic example involving the PDE problem of §3.2.1 with $\alpha = 1000$. With the safeguarded Choice 1, the iterates from Algorithm INB converged to the everywhere-positive solution in 40 GMRES(20) iterations; two backtracks were required. With the choice $\eta_k = \min\{1/(k+2), \|F(x_k)\|_2\}$, 164 GMRES(20) iterations and 11 backtracks were necessary; furthermore, convergence was to a solution other than the everywhere-positive solution. Such convergence to a "wrong" solution may or may not be undesirable per se, but it does indicate the potentially serious effects of disagreement between F and its local linear model.

Second, unless special care is taken, a very small forcing term may risk inaccuracy in an iterative linear solver such as GMRES, especially when products of F' with vectors are approximated with finite differences. Recall from §3.1 that our implementation of Algorithm INB uses selective second-order differencing to obtain essentially the same accuracy as if second-order differences were used throughout. Using the safeguarded Choice 2 forcing terms with $\gamma = .9$, we applied this implementation to the driven cavity problem with $Re = 500$; the results are shown in Figure 3.3. There is no evidence of inaccuracy in GMRES(20), and 218 iterations were required for successful termination. However, when the implementation was changed to use only first-order forward differences throughout, we obtained the results in Figure 3.4. Note the increase in the linear residual norm curve (the solid curve) just after iteration 200. The linear residual norm values used for this curve were evaluated directly at the beginning of each GMRES(20) cycle and then maintained recursively within the cycle; the observed increase occurs after the direct evaluation at iteration 200 and indicates that the recursively maintained values have become inaccurate. We note also that the number of GMRES(20) iterations required for termination has increased to 232.

3.5. Summary test results. In Table 3.1, we summarize the results of applying Algorithm INB to all test problem cases described in §3.2. In Table 3.2, we summarize the results over the PDE problem cases only. The results for the PDE problems are broken out in a separate table not only because these problems constitute an important problem class but also because the characteristic performance of Algorithm INB on these problems differed from that on the integral equations. On the integral equations, and on the H-equation in particular, GMRES(20) was so effective that the effects of different forcing term choices tended to be obscured. In most cases, only 1–3 GMRES(20) iterations were required for each inexact Newton step, and the linear residual norm was often reduced by several orders of magnitude in a single iteration. On the PDE problems, many more GMRES(20) iterations were typically required for each inexact Newton step, with only modest linear residual norm reduction per GMRES(20) iteration. Thus the PDE problems were better suited for showing the effects of different forcing term choices.

The first three columns of Tables 3.1 and 3.2 give geometric means of the numbers of linear iterations (GMRES(20) iterations), inexact Newton steps, and "function evaluation equivalents", where, for each test case, we define the number of "function evaluation equivalents" to be the sum of the numbers of linear iterations, backtracks, and inexact Newton steps. The number of linear iterations is the same as the number of products of F' with vectors; if these products were always approximated by first-

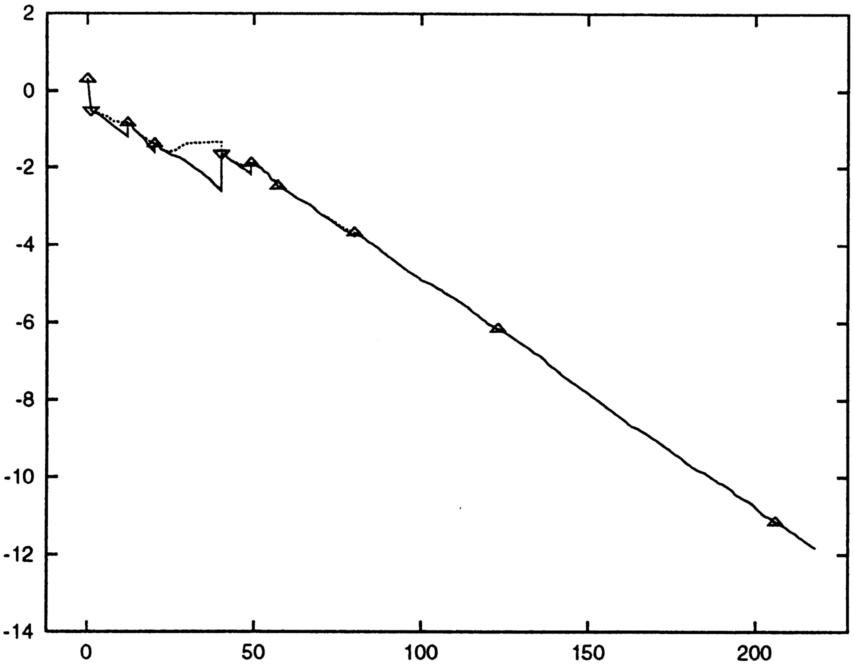

FIG. 3.3. *Illustration of the performance of Algorithm INB with selective second-order differencing and safeguarded Choice 2 forcing terms,* $\gamma = .9$*, on the driven cavity problem with Re* $= 500$*. The horizontal axis indicates the number of* GMRES(20) *iterations. The solid curve is* $\log_{10} \|F + F's\|_2$*; the dotted curve is* $\log_{10} \|F\|_2$*. Triangles indicate new inexact Newton steps: "*Δ*" indicates* η_k *given by Choice 2; "*∇*" indicates* $\eta_k = \eta_{k-1}^2$*.*

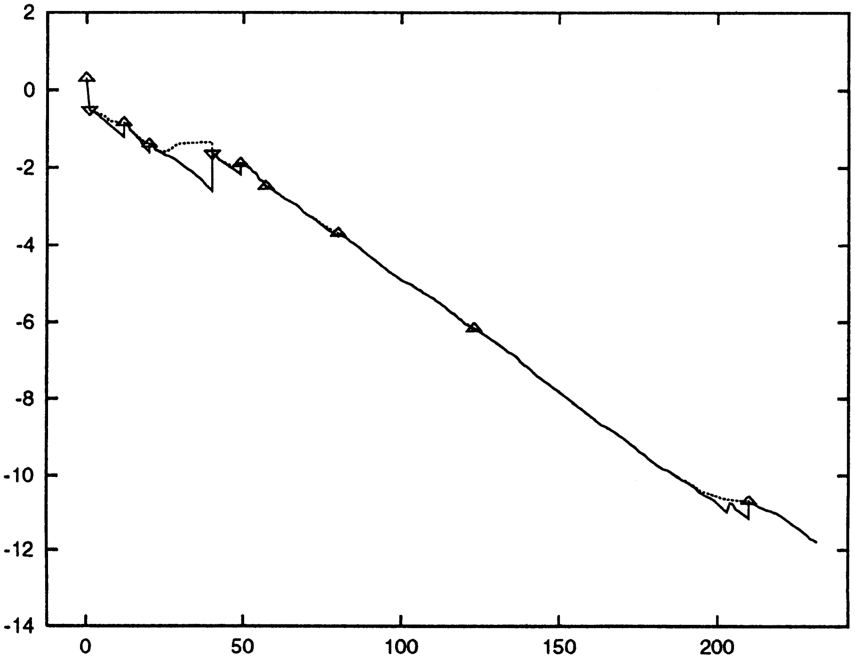

FIG. 3.4. *Illustration of the performance of Algorithm INB with first-order differencing throughout and safeguarded Choice 2 forcing terms,* $\gamma = .9$*, on the driven cavity problem with Re* $= 500$*. The horizontal axis indicates the number of* GMRES(20) *iterations. The solid curve is* $\log_{10} \|F + F's\|_2$*; the dotted curve is* $\log_{10} \|F\|_2$*. Triangles indicate new inexact Newton steps: "*Δ*" indicates* η_k *given by Choice 2; "*∇*" indicates* $\eta_k = \eta_{k-1}^2$*.*

order forward differences, then the number of "function evaluation equivalents" would be just the number of function evaluations. This number provides a rough relative measure of overall work for these test problems. It would be a less suitable measure, e.g., if there were additional costs associated with beginning a new inexact Newton step, such as initializing a new preconditioner. The fourth column gives numbers of backtracks over all test cases, i.e., numbers of step-reductions in the while-loop in Algorithm INB. The fifth column gives numbers of instances of convergence to a "wrong" solution, i.e., convergence to a solution other than the everywhere-positive solution in the PDE problem of §3.2.1 or to a solution other than $u \equiv 1$ in the integral equation problem of §3.2.5. As noted previously, convergence to a "wrong" solution illustrates the potentially serious effects of disagreement between F and its local linear model. The sixth column gives the number of failures over all test cases. If failure occurred in a test case, then that case was not included in the statistics for columns 1–5.

TABLE 3.1
Summary test results over all problems. GMLI, GMINS, *and* GMFEE *are geometric means of the numbers of linear iterations, inexact Newton steps, and "function evaluation equivalents", respectively.* NB, NW, *and* NFAIL *are the total numbers of backtracks, instances of convergence to a "wrong" solution, and failures, respectively. Results marked "*" were over succesful runs only.*

η_k choice	GMLI	GMINS	GMFEE	NB	NW	NFAIL
10^{-4}	90.2*	7.21*	103.3*	1*	0*	2
$1/2^{k+1}$	70.3*	9.24*	85.4*	6*	1*	1
$\min\{1/(k+2), \|F(x_k)\|_2\}$	72.3	8.72	86.6	18	2	0
Choice 1	53.5	9.33	67.1	5	0	0
Choice 1^2	55.1	8.90	69.7	13	0	0
Choice 2, $\gamma = 1$	52.4	8.82	65.9	8	0	0
Choice 2, $\gamma = .9$	52.5	7.89	64.7	8	0	0
Choice 2, $\gamma = .5$	66.8	7.93	79.4	13	1	0

TABLE 3.2
Summary test results over the PDE problems. GMLI, GMINS, *and* GMFEE *are geometric means of the numbers of linear iterations, inexact Newton steps, and "function evaluation equivalents", respectively.* NB, NW, *and* NFAIL *are the total numbers of backtracks, instances of convergence to a "wrong" solution, and failures, respectively. Results marked "*" were over succesful runs only.*

η_k choice	GMLI	GMINS	GMFEE	NB	NW	NFAIL
10^{-4}	152.4*	6.68*	163.7*	1*	0*	1
$1/2^{k+1}$	104.2*	8.95*	118.4*	3*	0*	1
$\min\{1/(k+2), \|F(x_k)\|_2\}$	117.9	8.22	130.5	15	1	0
Choice 1	82.5	8.87	95.3	3	0	0
Choice 1^2	89.1	8.48	103.8	11	0	0
Choice 2, $\gamma = 1$	82.7	8.60	96.3	6	0	0
Choice 2, $\gamma = .9$	83.3	7.57	95.2	6	0	0
Choice 2, $\gamma = .5$	98.4	7.57	110.4	10	0	0

One sees from Table 3.1 that, in terms of "function evaluation equivalents", the best performances over all problem cases were from, in order, Choice 2 with $\gamma = .9$, Choice 2 with $\gamma = 1$, and Choice 1. Choice 2 with $\gamma = .9$ also gave the smallest mean number of inexact Newton steps and essentially tied Choice 2 with $\gamma = 1$ for the smallest mean number of linear iterations. Thus, in terms of overall effort, Choice 2 with $\gamma = .9$ seems to be the winner, followed closely by Choice 2 with $\gamma = 1$ and Choice 1. However, note that Choice 1 required significantly fewer backtracks, although at the expense of more inexact Newton steps. Requiring fewer backtracks indicates that better agreement was maintained between F and its local linear model and, therefore, suggests greater robustness.

Table 3.2 shows that, over the PDE problem cases, Choice 2 with $\gamma = .9$ and Choice 1 were essentially tied for the lowest mean number of "function evaluation equivalents", followed closely by Choice 2 with $\gamma = 1$. Choice 1 and Choice 2 with $\gamma = 1$ were essentially tied for the lowest mean number of linear iterations, followed very closely by Choice 2 with $\gamma = .9$. Choice 2 with $\gamma = .9$ also tied with Choice 2 with $\gamma = .5$ for the smallest mean number of inexact Newton steps. Choice 2 with $\gamma = .9$ could be judged a very slight winner over Choice 1 and Choice 2 with $\gamma = 1$ in terms of overall effort, but note that Choice 1 again required significantly fewer backtracks, which suggests greater robustness.

In summary, the best performances were from Choice 2 with $\gamma = .9$ and $\gamma = 1$ and from Choice 1. In terms of overall effort, Choice 2 with $\gamma = .9$ seemed best by a small margin; however, Choice 1 required significantly fewer backtracks and, therefore, seems likely to result in greater robustness, albeit at the probable cost of more inexact Newton steps.

Choice 1^2 and Choice 2 with $\gamma = .5$ were notably less efficient. In addition, they required much larger numbers of backtracks and, therefore, seem likely to reduce robustness. The representative choices from the literature that were included in the tests were significantly less effective than the choices proposed here.

4. Conclusions. We have outlined forcing term choices that result in desirably fast local convergence and also tend to avoid oversolving the Newton equation, i.e., imposing an accuracy on an approximation of the Newton step that leads to significant disagreement between F and its local linear model. The choices, along with theoretical support and practical safeguards, are given in §2. Practical performance on a representative set of test problems is discussed in §3.

Choice 1 directly reflects the agreement between F and its local linear model at the previous step. It results in fast, although not q-quadratic, local convergence; see Theorem 2.1 for the precise statement. Choice 2 does not directly reflect the agreement between F and its local linear model; however, it has faster (up to q-quadratic) guaranteed local convergence (see Theorem 2.3) and was effective in our tests.

The best performances in our tests were from Choice 2 with $\gamma = .9$ and $\gamma = 1$ and from Choice 1. Choice 2 with $\gamma = .9$ seemed the most efficient overall by a slight margin and also gave the smallest mean numbers of inexact Newton steps. However, Choice 1 was almost as efficient and resulted in significantly fewer backtracks, which suggests greater robustness.

The numerical experiments and theoretical results suggest that Choice 1 might be preferred for good efficiency and superior robustness in general use, while Choice 2 with $\gamma = .9$ might offer more efficiency with good robustness on many problems, especially mildly nonlinear problems for which there is significant cost associated with beginning a new inexact Newton step.

In conclusion, we recall from §3.4 that, in a globalized Newton iterative or truncated Newton method such as the implementation of Algorithm INB used here, oversolving resulting from a small forcing term may place significant demands on the globalization and even cause it to fail, as well as incurring unnecessary expense in solving the Newton equation. In addition, unless special care is taken, a very small forcing term may risk inaccuracy in the iterative linear solver, especially when finite differences are used to approximate products of F' with vectors.

REFERENCES

[1] P. Bjørstad, *Fast numerical solution of the biharmonic Dirichlet problem on rectangles*, SIAM J. Numer. Anal., 20 (1983), pp. 59–71.

[2] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 450–481.

[3] X.-C. Cai, W. D. Gropp, D. E. Keyes, and M. D. Tidriri, *Newton-Krylov-Schwarz methods in* CFD, in Proceedings of the International Workshop on the Navier–Stokes Equations, R. Rannacher, ed., Notes in Numerical Fluid Mechanics, Braunschwieg, 1994 (to appear), Vieweg Verlag.

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

[5] R. S. Dembo and T. Steihaug, *Truncated Newton algorithms for large-scale optimization*, Math. Prog., 26 (1983), pp. 190–212.

[6] S. C. Eisenstat and H. F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optimization, 4 (1994), pp. 393–422.

[7] R. Glowinski, H. B. Keller, and L. Reinhart, *Continuation-conjugate gradient methods for the least squares solution of nonlinear boundary value problems*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 793–832.

[8] C. T. Kelley and J. I. Northrup, *A pointwise quasi-Newton method for integral equations*, SIAM J. Numer. Anal., 25 (1988), pp. 1138–1155.

[9] P. J. McKenna, 1992. Private communication.

[10] J. J. Moré, *A collection of nonlinear model problems*, in Computational Solution of Nonlinear Systems of Equations, E. L. Allgower and K. Georg, eds., Lectures in Applied Mathematics Vol. 26, American Mathematical Society, 1990, pp. 723–762.

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

[12] R. Schaaf, 1994. Private communication.

[13] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer Verlag, New York, 1980.

[14] P. N. Swartztrauber and R. A. Sweet, *Algorithm* 541: *Efficient Fortran subprograms for the solution of separable elliptic partial differential equations*, ACM Trans. Math. Software, 5 (1979), pp. 352–364.

[15] K. Turner and H. F. Walker, *Efficient high accuracy solutions with* GMRES(m), SIAM J. Sci. Stat. Comput., 13 (1992), pp. 815–825.

[16] C. J. van Duijn and J. M. de Graaf, *Large time behaviour of solutions of the porous medium equation with convection*, J. Differential Equations, 84 (1990), pp. 183–203.