SIAM J. NUMER. ANAL.
Vol. 18, No. 5, October 1981

0036-1429/81/1805-0016 $01.00/0

ON STEPLENGTH ALGORITHMS FOR A CLASS OF CONTINUATION METHODS*

C. DEN HEIJER† AND W. C. RHEINBOLDT‡

Abstract. The continuation methods considered here are algorithms for the computational analysis of the regular parts of the solution field of equations of the form $Fx = b$, $F: D \subset R^{n+1} \to R^n$, for given $b \in R^n$. While these methods are similar in structure to those used for ODE-solvers, their errors are independent of the history of the process and are solely determined by the termination criterion of the corrector at the current step. This suggests the use of a posteriori estimates of the convergence radii of the corrector. It is proved here that such estimates cannot be obtained from the sequence of corrector iterates alone but that they require some global information about F. However, it is shown that a finite sequence of corrector iterates does allow for the computation of effective estimates of the convergence quality of certain types of correctors. This is used for the design of various step-algorithms for continuation processes; two of them are based on a Newton-corrector while the third one is applicable to any corrector. Some numerical results show the effectiveness of the three algorithms. Finally some asymptotic analysis of continuation steps is given.

1. Introduction.

Continuation methods have by now established themselves as a very valuable tool for the numerical solution of nonlinear equations in several variables, (see, e.g., the references given in [9], [11], [19], [21]). In line with [14], we consider here a class of such methods for the computational analysis of specified parts of the solution field of equations of the form

$$Fx = b, \qquad F: D \subset R^{n+1} \to R^n, \qquad b \in R^n. \tag{1.1}$$

This solution field consists of segments of smooth curves interspersed with certain singular points (see, e.g., [14] for some existence theory).

For a computational trace of a specific solution branch of (1.1) it appears to be most advantageous to parametrize the desired curve in terms of the arclength or some approximation of it and to apply predictor-corrector continuation algorithms. These methods resemble those used in the multistep ODE-solvers for the computer solution of initial value problems for ordinary differential equations. But there are several major differences which need to be noted.

Based on the information on the already computed portion of the continuation curve, a suitable extrapolation is calculated which approximates a further curve segment. Standard polynomial extrapolation of the curve data can be used here as well as any of the predictor formulas of the standard ODE-solvers. Now a point on this extrapolating curve is chosen as the starting point for a corrector iteration designed to converge to a point on the continuation curve. This corrector is a locally convergent iterative process for an $(n+1)\times(n+1)$ system of equations consisting of the n equations (1.1) coupled with a suitable $(n+1)$st equation that identifies the location of the point on the curve. Hence, if only the predicted starting point is sufficiently close, the corrector iteration is guaranteed to converge to a point on the exact curve. In contrast to this, the ODE solvers involve a corrector equation which is obtained by interpolation and hence for which the solutions are not, in general, on the exact curve.

* Received by the editors May 30, 1980, and in final form January 12, 1981.

† Philips Electronics, Eindhoven, the Netherlands. The research of this author was supported in part by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.) through a NATO Science Fellowship.

‡ Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261. The research of this author was supported in part by the National Science Foundation under grant MCS-78-05299.

This distinction has several fundamental consequences. It implies that the actual errors at each step of the ODE-solvers are functions of the history of the process up to that point while the errors of the continuation methods are independent of the history and are solely determined by the termination criterion of the corrector iteration at the current step as long as there is convergence. This in terms means that, in principle, any step along the predictor-curve is acceptable for which the resulting starting point is in the convergence domain of the corrector. The steplength criteria needed for ODE-solvers are by no means as easily characterizable. Here is also one of the reasons why—unlike for ODE-solvers—predictors of higher order are rarely very effective. In fact, the higher order improvements in the predicted point seldom produce a sufficiently large effect in the convergence behavior of the corrector to balance the added cost of their computation. As a consequence a simple Euler predictor is usually most effective and in § 2 below we summarize a class of continuation methods based upon it.

In view of the above comments, a principal aspect of a step-algorithm for a continuation method is certainly a technique for assessing whether the corrector may be expected to converge when started from the predicted point. This is the principal topic of this paper. Ideally, it would be desirable to determine a posteriori estimates of the convergence radii of the corrector at all computed points along the curve and to extrapolate these data to the next point that is to be obtained. In § 3 below it is proved that this is impossible if these estimates are to be based solely on the sequences of corrector iterates at the different points. Instead, such convergence radii require some global information which is available only at considerable expense. In order to overcome this difficulty, we show in § 4 that a finite sequence of corrector iterates does allow us to compute effective estimates of the convergence quality of certain types of corrector processes. This can be used in the design of step-algorithms for our continuation method, and in § 5 three such algorithms are described in detail. Two of them are based on the methods involving a Newton corrector while the third one is applicable to any corrector. In § 6 we present some numerical results for various problems which show the comparative effectiveness of the three algorithms. Finally, § 7 develops an asymptotic theory for the continuation steps and the relative stability of a step algorithm.

2. The basic continuation method. In this section we summarize briefly the principal aspects of the continuation method considered here. This discussion essentially follows [14]. As in the introduction the equations to be solved are written in the "underdetermined" form

$$Fx = b \tag{2.1}$$

involving a mapping F from R^{n+1} into R^n and a given vector $b \in R^n$. For ease of discussion we assume that F is defined and r-times continuously differentiable on all of R^{n+1} for some $r \geqq 2$.

A (nontrivial, parametric) C^r-solution of (2.1) is any r-times continuously differentiable function

$$x: J \subset R^1 \to R^{n+1}, \qquad x'(s) \neq 0 \quad \forall s \in J \tag{2.2}$$

on some open interval $J \subset R^1$ such that $Fx(s) = b$ for all s in J. Evidently, this concept is invariant under any C^r-parameter transformation with nonzero derivative. Hence there is no restriction to assume that in (2.2) the parameter s is the arc-length. This parameter choice has certain advantages and has been used by various authors. Without attempting any historical survey, we mention only the articles [4], [5], [14], [16].

The regularity set of F is defined as

$$\mathcal{R}(F) = \{x \in R^{n+1};\ \text{rank } F'(x) = n\}. \tag{2.3}$$

For any $x \in \mathcal{R}(F)$ there exists a unique vector $v \in R^{n+1}$ such that

$$F'(x)v = 0, \quad \|v\|_2 = 1, \quad \det\begin{pmatrix} F'(x) \\ v^T \end{pmatrix} > 0. \tag{2.4}$$

Hence the mapping

$$T: \mathcal{R}(F) \to R^{n+1}, \qquad Tx = v \tag{2.5}$$

is well defined and can be shown to be $(r-1)$-times continuously differentiable on $\mathcal{R}(F)$. Moreover, for given $x^0 \in \mathcal{R}(F)$ with $Fx^0 = b$, a solution of the autonomous initial value problem

$$x' = Tx, \qquad x(0) = x^0 \tag{2.6}$$

is a C^r-solution (2.2) of (2.1) with the arc-length as parameter. As usual such a solution (2.2) is called saturated if its interval of definition $\mathcal{J}$ is maximal with respect to set inclusion. Then, the following result holds (see [14]):

THEOREM 2.1. *For any $x^0 \in \mathcal{R}(F)$ with $Fx^0 = b$ there exists a unique saturated C^r-solution* (2.2) *of* (2.1) *in $\mathcal{R}(F)$ with the arc-length as parameter s and $x(0) = x^0$. Moreover, if $\alpha \in \partial\mathcal{J}$ is finite then $x(s) \to \partial\mathcal{R}(F)$ or $\|x(s)\|_2 \to \infty$ as $s \to \alpha$, $s \in \mathcal{J}$.*

From a given starting point $x^0 \in \mathcal{R}(F)$, $Fx^0 = b$, suppose that for some $p \geqq 0$ our continuation process has produced the points

$$x^j = x(s_j), \qquad j = 0, 1, \cdots, p, \quad 0 = s_0 < s_1 < \cdots < s_p, \tag{2.7}$$

approximating the C^r-solution (2.2) in $\mathcal{R}(F)$ through x^0 specified by the theorem. For the computation of the tangent vector Tx^p we use the system of equations

$$\begin{pmatrix} F'(x^p) \\ (e^i)^T \end{pmatrix} v = e^{n+1}, \tag{2.8}$$

where $e^1, \cdots, e^{n+1}$ denote the natural basis of R^{n+1}. The index i has to be chosen such that $(Tx^p)^T e^i \neq 0$. Generally, let

$$|(Tx^l)^T e^{i_l}| = \max\{|(Tx^l)^T e^j|, j = 1, \cdots, n+1\}, \qquad l = 0, 1, \cdots. \tag{2.9}$$

Then it is reasonable to use $i = i_{p-1}$ which will ensure that the matrix of (2.8) is nonsingular as long as the last step $x^p - x^{p-1}$ is not too large. Of course, this applies only for $p \geqq 1$; for $p = 0$ we have to assume that a suitable index i has been given.

Once the solution $v \neq 0$ of (2.8) has been computed the tangent vector at x^p is obtained as $Tx^p = \sigma v / \|v\|_2$. Here, by (2.4), the direction factor $\sigma = \pm 1$ is set to $\sigma = \sigma_0 \operatorname{sign} v^T e^i$, where σ_0 is the sign of the determinant of the matrix (2.8), computed easily during the solution of that system.

In line with the discussion in the introduction we use now some step $h = h_{p+1}$ along the Euler predictor

$$y_p(h) = x^p + hTx^p, \qquad h > 0. \tag{2.10}$$

Moreover, the corrector is chosen as a locally convergent iterative process for the solution of the augmented system

$$\begin{aligned} Fx &= b, \\ (e^{i_p})^T(x - y_p(h_{p+1})) &= \gamma, \end{aligned} \tag{2.11}$$

where i_p is defined by (2.9) and the scalar γ will be fixed shortly. If Newton's method is used as the corrector, then the linear systems of equations to be solved have exactly the same structure as (2.8).

The choice of the step h_{p+1} to be taken along the predictor depends on the selection of γ. This is equivalent with the selection of the point $x(s_p+\Delta s)$ on the solution curve which is to be approximated by the corrector. In fact, as long as h_{p+1} is not too large, the equation

$$(e^{i_p})^T x(s_p+\Delta s) = \gamma + (e^{i_p})^T y_p(h_{p+1}) \tag{2.12}$$

defines a one-to-one correspondence between Δs and γ for all sufficiently small Δs. Evidently, $\Delta s = h_{p+1}$ is an obvious but by no means necessary choice for Δs.

In order to compute γ for $\Delta s = h_{p+1}$ and to estimate the distance between the predicted point $y_p(h_{p+1})$ and the desired solution $x(s_p+h_{p+1})$ we proceed as in the case of the ODE-solvers. The quadratic polynomial

$$\begin{aligned} &q_p(s) = x^p + (s-s_p)Tx^p + (s-s_p)^2 w^p, \qquad s > s_p, \\ &w^p = \frac{1}{\Delta s_p}\left(Tx^p - \frac{1}{\Delta s_p}(x^p - x^{p-1})\right), \qquad \Delta s_p = \|x^p - x^{p-1}\|_2 \doteq s_p - s_{p-1} \end{aligned} \tag{2.13}$$

represents an approximation of $x(s)$ for which—in the case of $r \geqq 3$—the error is asymptotically of order $(\Delta s_{\max})^3$, where $\Delta s_{\max}$ is the maximal step in s. On the other hand, the error of the predictor-approximation is of the order $(\Delta s_{\max})^2$.

If $x(s)$ is replaced by $q_p(s)$ in (2.12) we obtain for $\Delta s = h_{p+1}$

$$\gamma = \frac{h_{p+1}^2}{\Delta s_p}\left[(e^{i_p})^T Tx^p - \frac{1}{\Delta s_p}(e^{i_p})^T (x^p - x^{p-1})\right]. \tag{2.14}$$

On the other hand, for any given tolerance $\rho_{p+1} > 0$ the relations

$$\|y_p(h_{p+1}) - x(s_p+h_{p+1})\| \doteq \|y_p(h_{p+1}) - q_p(s_p+h_{p+1})\| \leqq \rho_{p+1} \tag{2.15}$$

lead to the estimate

$$h_{p+1} \leqq \sqrt{\frac{\rho_{p+1}}{\|w^p\|_2}}. \tag{2.16}$$

Clearly ρ_{p+1} should be an estimate of a "safe" convergence distance between the starting point and desired limit point of the corrector iteration. The selection of ρ_{p+1} will be the topic of the subsequent sections.

In connection with the formulas (2.14) and (2.16) it should be noted that the vector w^p is numerically difficult to evaluate. Thus in (2.14) the difference in square brackets should be computed with care and preferably in double precision. At the same time, the norm of w^p is obtained best in the form

$$\|w^p\|_2 = \frac{2}{\Delta s_p}\left|\sin\frac{\alpha_p}{2}\right|, \qquad \alpha_p = \arccos\left((Tx^p)^T \frac{x^p - x^{p-1}}{\Delta s_p}\right) \tag{2.17}$$

which is numerically reliable as long as the last step Δs_p was not overly small.

3. Negative results about convergence radii.

3.1. General stationary processes. Let $P: \mathrm{Dom}\,(P) \subset R^m \to R^m$ be a given nonlinear mapping on the open domain $\mathrm{Dom}\,(P) \subset R^m$, and $z^* \in \mathrm{Dom}\,(P)$ a solution of the equation

$$Pz = 0 \tag{3.1}$$

which is to be approximated by means of a given iterative process $\mathcal{J}$. In our setting (3.1) represents any one of the equations to be solved by the corrector. As in [9] the domain of convergence of the process $\mathcal{J}$ at z^* is the set $C(\mathcal{J}, z^*) \subset \mathrm{Dom}\,(P)$ of all starting points $z^0 \in \mathrm{Dom}\,(P)$ for which the sequence $\{z^k\}$ generated by $\mathcal{J}$ exists, remains in Dom (P), and converges to z^*. If $\mathcal{J}$ is locally convergent at z^* then $C(\mathcal{J}, z^*)$ contains an open neighborhood of z^*. The radius $\rho > 0$ of any ball $B(z^*, \rho) \subset C(\mathcal{J}, z^*)$ with center z^* is a convergence radius of $\mathcal{J}$ at z^* and

$$r(\mathcal{J}, z^*) = \sup\{\rho > 0;\ \bar{B}(z^*, \rho) \subset C(\mathcal{J}, z^*)\} \tag{3.2}$$

is the maximal convergence radius. Clearly these radii depend on the norm used in the definition of B.

Ideally, in the setting of continuation methods, we would like to compute from a given iteration sequence $\{z^k\}$ with $z^0 \in C(\mathcal{J}, z^*)$ some convergence radius ρ, that is, some lower bound of the maximal radius $r(\mathcal{J}, z^*)$. Unfortunately, this is impossible.

A precise formulation of this statement, of course, depends on the process $\mathcal{J}$. Suppose that $\mathcal{J}$ is of the form

$$\mathcal{J}: z^{k+1} = Gz^k, \qquad k = 0, 1, \cdots \tag{3.3}$$

where $G: \mathrm{Dom}\,(G) \subset R^m \to R^m$ is a given continuous mapping on some open domain Dom $(G) \subset$ Dom (P) with the fixed point $z^* = Gz^* \in \mathrm{Dom}\,(G)$. Many processes used in practice have this form. Suppose that $\bar{B}(z^*, \varepsilon_0) \subset \mathrm{Dom}\,(G)$ and that the sequence $\{\bar{z}^k\}$ with $\bar{z}^0 \in C(\mathcal{J}, z^*)$, generated by $\mathcal{J}$, is the known information about our process. For any given $\varepsilon \in (0, \varepsilon_0)$ there are only finitely many $\bar{z}^k$ with $\|\bar{z}^k - z^*\| > \varepsilon/2$. Hence it is possible to find a vector $w \in R^m$, $\varepsilon/2 < \|w\| < \varepsilon$ and a number $\delta \in (0, \min(\varepsilon - \|w\|, \|w\| - \varepsilon/2))$ such that for the points $u_+ = z^* + w$, $u_- = z^* - w$ the balls $\bar{B}(u_+, \delta)$ and $\bar{B}(u_-, \delta)$ do not intersect the sequence $\{\bar{z}^k\}$, (see Fig. 1). Let

$$\psi: R^1 \to R^1, \quad \psi(t) = \begin{cases} 0 & \text{for } t \leqq 0, \\ (10 - 15t + 6t^2)t^3 & \text{for } 0 \leqq t \leqq 1, \\ 1 & \text{for } t \geqq 1, \end{cases} \tag{3.4}$$

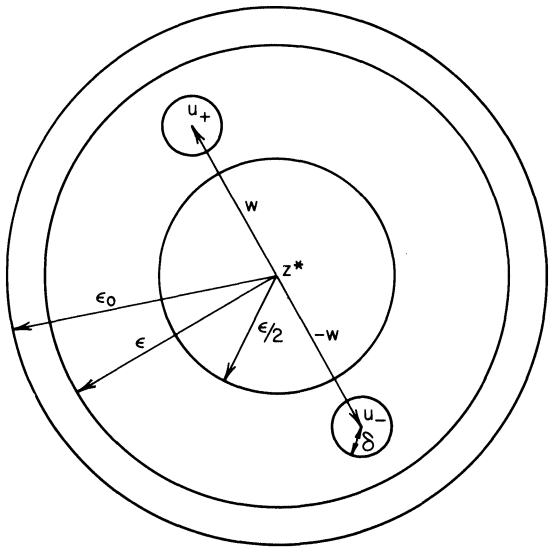

FIG. 1.

and define the function $\tilde{G}: \text{Dom}\,(G) \subset R^m \to R^m$ by

$$\tilde{G}z = \left(1 - \psi\left(\frac{2}{\delta}\|z - u_+\| - 1\right)\right)u_- + \left(1 - \psi\left(\frac{2}{\delta}\|z - u_-\| - 1\right)\right)u_+ + \min\left(\psi\left(\frac{2}{\delta}\|z - u_+\| - 1\right), \psi\left(\frac{2}{\delta}\|z - u_-\| - 1\right)\right)Gz. \tag{3.5}$$

Clearly $\tilde{G}$ is continuous and satisfies $\tilde{G}z = Gz$ for $z \in \text{Dom}\,(G) \setminus (\bar{B}(u_+, \delta) \cup \bar{B}(u_-, \delta))$. Since the sequence $\{\bar{z}^k\}$ does not intersect the balls, it follows that the process $\tilde{\mathcal{J}}$ (of the form (3.3)) defined by $\tilde{G}$ generates exactly the sequence $\{\bar{z}^k\}$ when the starting point $\bar{z}^0$ is used. On the other hand, for $z^0 = u_+$ as starting point, $\tilde{\mathcal{J}}$ generates $z^k = u_+$ for even k and $z^k = u_-$ for odd k. Thus this sequence $\{z^k\}$ does not converge to z^* and we have $r(\tilde{\mathcal{J}}, z^*) \leqq \|w\| < \varepsilon$.

Since we cannot distinguish whether our "information-sequence" $\{\bar{z}^k\}$ was generated by $\mathcal{J}$ or $\tilde{\mathcal{J}}$, and since ε was arbitrarily small, this result shows that on the basis of one sequence $\{\bar{z}^k\}$, $\bar{z}^0 \in C(\mathcal{J}, z^*)$ it is impossible to compute any lower bound of $r(\mathcal{J}, z^*)$, that is, any convergence radius ρ of $\mathcal{J}$ at z^*. Clearly, the same result holds when any finite number of such sequences are used.

Under slightly more restrictive conditions about G we may also show that no upper bound for $r(\mathcal{J}, z^*)$ can be computed. More specifically, suppose that G is locally attractive at $z^* = Gz^*$ in the sense that

$$\|Gz - z^*\| \leqq \alpha\|z - z^*\| \quad \forall z \in B_\varepsilon = \bar{B}(z^*, \varepsilon) \subset \text{Dom}\,(G) \subset \text{Dom}\,(P) \tag{3.6}$$

for some $\varepsilon > 0$ and $0 < \alpha < 1$. This holds for many processes of the form (3.3) under appropriate choices of the norm (see, e.g., [9], Theorem 10.1.3). Evidently, (3.6) ensures that $B_\varepsilon \subset C(\mathcal{J}, z^*)$ and hence that $r(\mathcal{J}, z^*) \geqq \varepsilon$.

Now consider the mapping

$$\tilde{G}: R^m \to R^m, \qquad \tilde{G}z = \begin{cases} Gz & \text{for } z \in B_\varepsilon, \\ G\left(z^* + \dfrac{\varepsilon}{\|z - z^*\|}(z - z^*)\right) & \text{for } z \notin B_\varepsilon. \end{cases} \tag{3.7}$$

Then $\tilde{G}$ maps R^m into B_ε and the process $\tilde{\mathcal{J}}$ defined by this function $\tilde{G}$ converges to z^* for any $z^0 \in R^m$; that is, we have $r(\tilde{\mathcal{J}}, z^*) = \infty$. Moreover, for any $z^0 \in B_\varepsilon$ the sequences generated by $\mathcal{J}$ and $\tilde{\mathcal{J}}$ are identical; and hence from our knowledge of $\mathcal{J}$ in B_ε alone we cannot deduce a finite upper bound for $r(\mathcal{J}, z^*)$.

3.2. Newton's method. In practice, the iteration function G of (3.3) is of a specific type and the above results do not guarantee that the modified functions $\tilde{G}$ of (3.5) or (3.7) are of the same type. But in many instances these functions $\tilde{G}$ are readily adjusted to belong to the same class as G itself. We shall discuss this here only in the case of the Newton iteration for the equation (3.1).

For this we assume from now on that our given mapping $P: \text{Dom}\,(P) \subset R^m \to R^m$ is continuously differentiable on the open domain Dom (P) and that the first derivative P' is Lipschitz continuous. We shall use always the Euclidean norm and denote the Lipschitz constant of P' on Dom (P) by γ. Moreover, the solution $z^* \in \text{Dom}\,(P)$ of (2.1) is supposed to be simple in the sense that $P'(z^*)$ is nonsingular which means that $\beta = \|P'(z^*)^{-1}\|$ is well-defined.

The following attraction theorem is given in [13]:

THEOREM 3.1. *If the open ball* $B^* = B(z^*, r_1^*)$ *with* $r_1^* = 2/(3\beta\gamma)$ *is contained in* Dom (P), *then* B^* *is in the domain* Dom (N) *of the Newton function*

$$N: \text{Dom}\,(N) \subset R^m \to R^m, \qquad Nz = z - P'(z)^{-1}Pz, \tag{3.8}$$

and

$$\|Nz - z^*\| \leqq \frac{1}{2}\frac{\beta\gamma\|z - z^*\|^2}{1 - \beta\gamma\|z - z^*\|} \quad \forall z \in \beta^*, \tag{3.9}$$

implies that for any $z^0 \in B^*$ *the Newton process*

$$\mathcal{N}: z^{k+1} = Nz^k, \qquad k = 0, 1, \cdots \tag{3.10}$$

converges to z^*.

In [3] it has been shown that the radius r_1^* of B^* is sharp, that is, that there are functions P which satisfy the stated properties and for which $r_1^* = r(\mathcal{N}, z^*)$.

In order to show that the modified iteration function $\tilde{G}$ of (3.5) may be defined as the Newton function of some operator, we will use the following simple result:

LEMMA 3.2. *Let* $\tilde{P}, Q: D \subset R^m \to R^m$ *both be continuously differentiable mappings on the open domain* D *with Lipschitz-continuous derivatives* $\tilde{P}'$ *and* Q'. *Then for any ball* $\bar{B}(u, \sigma) \subset D$, $\sigma > 0$, *there exists a continuously differentiable function* $H: D \subset R^m \to R^m$ *with Lipschitz-continuous derivatives* H' *such that* $Hz = Qz$ *for* $z \in \bar{B}(u, \sigma/2)$ *and* $Hz = \tilde{P}z$ *for* $z \in D \backslash \bar{B}(u, \sigma)$.

For the proof let

$$\xi: R^m \to R^1, \qquad \xi(z) = \frac{4}{\sigma^2}(z - u)^T(z - u) - 1 \tag{3.11}$$

and with ψ given by (3.4) set

$$Hz = \psi(\xi(z))\tilde{P}z + (1 - \psi(\xi(z)))Qz \quad \forall z \in D. \tag{3.12}$$

Then

$$\begin{aligned} H'(z) = {} & \frac{8}{\sigma^2}\psi'(\xi(z))(\tilde{P}z - Qz)(z - u)^T + \psi(\xi(z))\tilde{P}'(z) \\ & + (1 - \psi(\xi(z)))Q'(z) \quad \forall z \in D \end{aligned} \tag{3.13}$$

and it is easily checked that H satisfies the properties required in the lemma.

With this we can now proceed as in the previous section. Let $\{\bar{z}^k\}$ be a given Newton sequence for (3.1) which converges to z^*. By Theorem 3.1 all balls $\bar{B}(z^*, \varepsilon)$ with sufficiently small $\varepsilon > 0$ are in the domain of the Newton function (3.8). Given such an $\varepsilon > 0$ we may choose once again a vector $w \in R^m$, $\varepsilon/2 < \|w\| < \varepsilon$ and a number $\delta \in (0, \min(\varepsilon - \|w\|, \|w\| - \varepsilon/2$ such that the balls $\bar{B}(z^* \pm w, \delta)$ do not intersect $\{\bar{z}^k\}$.

In order to construct a map for which the Newton process sends $z^* + w$ into $z^* - w$ and vice versa, note first that Newton's method for the real equation $\phi(t) \equiv \arctan \lambda t = 0$ has a cycle point t_0/λ with $t_0 \doteq 1.391745$. Accordingly, we choose $\lambda = t_0/\|w\|$ and introduce the map

$$\Phi: R^m \to R^m, \qquad \Phi y = (\phi(y_1), \cdots, \phi(y_m))^T. \tag{3.14}$$

Now let $u^1 = w/\|w\|$, $u^2, \cdots, u^m$ be an orthonormal basis of R^m and with $U = ((u^i)^T e^j; i, j = 1, \cdots, m)$ introduce the affine transformation $y = \lambda U(z - z^*)$. Then the mapping

$$Q: R^m \to R^m, \qquad Qz = U^T\Phi(y) \tag{3.15}$$

is certainly continuously differentiable and has a Lipschitz continuous derivative

$$Q'(z) = \lambda U^T\Phi'(y)U, \tag{3.16}$$

which is nonsingular on all of R^m. For $z = z^* \pm w$ the transformed points are $y = \pm t_0 e^1$

and hence we obtain

$$z^* \pm w - Q'(z^* \pm w)^{-1} Q(z^* \pm w) = z^* + \frac{1}{\lambda} U^T [\pm t_0 e^1 - \Phi'(\pm t_0 e^1)^{-1} \Phi(\pm t_0 e^1)]$$

$$= z^* + \frac{1}{\lambda} U^T (\mp t_0 e^1) = z^* \mp w.$$

With $\tilde{P} = P$ and this mapping Q and the ball $\bar{B}(z^* + w, \delta)$ we obtain after application of Lemma 3.2 a mapping H_0. Then another application of the lemma with H_0, Q, and the ball $\bar{B}(z^* - w, \delta)$ gives a mapping H which is continuously differentiable on Dom (P), and has a Lipschitz-continuous derivative on that domain. Moreover, we have $Hz = Qz$ for $z \in \bar{B}(z^* - w, \delta/2) \cup \bar{B}(z^* + w, \delta/2)$ and $Hz = Pz$ for $z \in \text{Dom } (P) \setminus (\bar{B}(z^* - w, \delta) \cup \bar{B}(z^* + w, \delta))$. Hence, in particular, the Newton method $\mathcal{N}_H$ for $Hz = 0$ started from $\bar{z}^0$ generates exactly the original sequence $\{\bar{z}^k\}$, while, of course, for $z^* + w$ as starting point, $\mathcal{N}_H$ oscillates between $z^* + w$ and $z^* - w$. Thus it follows that $r(\mathcal{N}_H, z^*) \leqq \|w\| < \varepsilon$ and as in § 3.1 we see that on the basis of the given Newton-sequence $\{\bar{z}^k\}$ alone it is impossible to compute a nonzero lower bound for $r(\mathcal{N}, z^*)$, that is, any convergence radius of the Newton method (3.10) for the equation (3.1).

We turn to the proof that it is equally impossible to obtain an upper bound of $r(\mathcal{N}, z^*)$ on the basis of local information about P near z^*. For this we assume that P is twice-continuously differentiable on Dom (P) and that γ is a bound of $\|P''\|$ on that domain. Moreover, with the same conditions for z^* suppose that the ball $B(z^*, r_1^*)$, $r_1^* = 2/(3\beta\gamma)$, of Theorem 3.1 is contained in Dom (P).

Let $\varepsilon \in (0, 2/(5\beta\gamma))$ be given. Then the Newton process $\mathcal{N}$ for P is certainly convergent to z^* for any starting point in the closed ball $B_\varepsilon = \bar{B}(z^*, \varepsilon)$. Set

$$\Pi : R^m \setminus \{z^*\} \to R^m, \qquad \Pi z = z^* + \frac{\varepsilon}{\|z - z^*\|}(z - z^*) \quad \forall z \in R^m, \quad z \neq z^* \tag{3.17}$$

and

$$K : R^m \setminus \{z^*\} \to R^m, \quad Kz = P(\Pi(z)) + P'(\Pi(z))(z - \Pi(z)) \quad \forall z \in R^m, \quad z \neq z^*. \tag{3.18}$$

Then

$$\Pi'(z) = \frac{\varepsilon}{\|z - z^*\|} I - \frac{\varepsilon}{\|z - z^*\|^3}(z - z^*)(z - z^*)^T \tag{3.19}$$

has for fixed $z \neq z^*$ the eigenvector $z - z^*$ belonging to the eigenvalue zero and $m - 1$ eigenvectors orthogonal to $z - z^*$ belonging to the eigenvalue $\varepsilon/\|z - z^*\|$, which implies that

$$\|\Pi'(z)\| \leqq \frac{\varepsilon}{\|z - z^*\|}. \tag{3.20}$$

Moreover, we have

$$K'(z)h = P'(\Pi(z))h + P''(\Pi(z))(z - \Pi(z), \Pi'(z)h) \quad \forall z \in R^m, \quad z \neq z^*, \quad h \in R^m \tag{3.21}$$

which shows that K is continuously differentiable for all $z \neq z^*$.

Now define

$$Q : R^m \to R^m, \qquad Qz = \begin{cases} Pz, & z \in B_\varepsilon \\ Kz & \text{otherwise.} \end{cases} \tag{3.22}$$

On the boundary of B_ε, that is, for $\|z - z^*\| = \varepsilon$, we have $\Pi z = z$ and thus $Pz = Kz$, and it follows from (3.21) that also $K'(z) = P'(z)$. In other words, Q is continuously differentiable on all of R^m. Theorem 3.1 provides that for $z \in B_\varepsilon$ the derivative $Q'(z) = P'(z)$ is nonsingular with

$$\|Q'(z)^{-1}\| \leqq \frac{\beta}{1-\beta\gamma\varepsilon} \quad \forall z \in B_\varepsilon. \tag{3.23}$$

From (3.20) and (3.21) it follows that

$$\|Q'(z) - P'(\Pi(z))\| \leqq \gamma\varepsilon\left(1 - \frac{\varepsilon}{\|z - z^*\|}\right) \quad \forall z \in R^m \backslash \bar{B}(z^*, \varepsilon)$$

and the perturbation lemma ensures that $Q'(z)$ is nonsingular also outside the ball and, together with (3.23), that

$$\|Q'(z)^{-1}\| \leqq \frac{\beta}{1-\beta\gamma\varepsilon(2-\theta(z))} \quad \forall z \in R^m \tag{3.24}$$

where $\theta(z) = \min(\varepsilon/\|z - z^*\|, 1)$. Now a straightforward estimate shows that

$$\|Qz^* - Qz - Q'(z)(z^* - z)\| \leqq \tfrac{1}{2}\gamma\varepsilon^2 \quad \forall z \in R^m$$

whence

$$\|(z - Q'(z)^{-1}Qz) - z^*\| < \frac{\frac{1}{2}\beta\gamma\varepsilon^2}{1-2\beta\gamma\varepsilon} < \varepsilon.$$

In other words, for any starting point outside the ball B_ε one step of the Newton process $\mathcal{N}_Q$ for Q brings us into the ball where, of course, $\mathcal{N}_Q$ is identical with the Newton process $\mathcal{N}$ for P and hence converges. Therefore we see that, as desired, $r(\mathcal{N}_Q, z^*) = \infty$. Since we cannot distinguish $\mathcal{N}$ and $\mathcal{N}_Q$ inside B_ε it follows that no upper bound for $r(\mathcal{N}, z^*)$ can be obtained from information about P and hence $\mathcal{N}$ in B_ε alone.

4. Convergence assessment of an iterative sequence. The negative results of the previous section imply that from a given (convergent) sequence $\{z^k\}$ generated by an iterative process $\mathcal{J}$ we can infer at best some results about the convergence behavior of that particular sequence, but not of any other sequence produced by $\mathcal{J}$. We present in this section some approaches toward computing a posteriori estimates of the convergence quality of a single iterative sequence. In the next section these estimates will be applied to our continuation process.

There are many convergence measures for a sequence $\{z^k\}$. For theoretical purposes the R- and Q-convergence factors and related orders (see [9]) are probably most frequently used. But, except in the case of linear convergence, they often do not provide an adequate measure of a finite segment of an iterative sequence, and hence other measures will have to be considered as well. We begin with a discussion of the linear case in a form which exhibits the ideas used afterwards for quadratically convergent sequences.

Let $\{z^k\}$ be the given sequence generated by $\mathcal{J}$ with limit z^* and denote the errors by $e_k = \|z^k - z^*\|$, $k = 0, 1, \cdots$. The definition of any convergence measure is based on a hypothetic model of the behavior of the errors. These models usually take the form of a difference inequality for the errors derived from theoretical considerations. If $\{z^k\}$ converges linearly it is reasonable to assume that

$$(0 \leqq) e_{k+1} \leqq \lambda e_k, \qquad k = 0, 1, \cdots \tag{4.1}$$

with some constant $\lambda \in (0, 1)$ depending on $\{z^k\}$. In the case when $\mathcal{J}$ has the basic form

(3.3), this is certainly valid if the mapping G satisfies a local contraction condition

$$\|Gz - z^*\| \leqq \mu \|z - z^*\|, \quad \forall z \in \bar{B}(z^*, \rho), \quad \rho > 0, \quad 0 < \mu < 1. \tag{4.2}$$

In fact, (4.2) implies (4.1) with $\lambda = \mu$ for any sequence $\{z^k\}$ started from a point z^0 in $\bar{B}(z^*, \rho)$.

Clearly, the constant λ of (4.1) is an upper bound of the Q-factor $Q_1\{z^k\}$, and we want to compute an a posteriori estimate of λ. This can be accomplished in various ways. Suppose that the process was terminated with the iterate z^{k^*}. Then, hopefully, e_{k^*} is small and we may estimate e_k by $\|z^k - z^{k^*}\|$, $k = 0, 1, \cdots, k^* - 1$. This suggests the quantity

$$\tilde{\lambda} = \left[\frac{\|z^{k^*-1} - z^{k^*}\|}{\|z^0 - z^{k^*}\|}\right]^{1/(k^*-1)} \doteq \left[\frac{e_{k^*-1}}{e_0}\right]^{1/(k^*-1)} \leqq \lambda, \qquad k^* \geqq 2 \tag{4.3}$$

as a computable estimate of λ. Under the assumption (4.1) we obtain the following upper bound:

$$\hat{\lambda} = \lambda\left[\frac{1+\lambda}{1-\lambda^{k^*}}\right]^{1/(k^*-1)} \geqq \max(\tilde{\lambda}, \lambda), \qquad k^* \geqq 2. \tag{4.4}$$

In order to assure a relative error $e_{k^*}/e_0 \leqq \varepsilon$ at least $k^* \geqq \ln \varepsilon / \ln \lambda$ steps are required. Table 4.1 gives some data for k^* and the upper bound $\hat{\lambda}$ of (4.4) for different values of ε and λ. Obviously, the estimate (4.4) is fairly coarse and yet the table shows that $\tilde{\lambda}$ cannot overestimate λ too much. Of course, no lower bound for λ can be found if only the upper estimate (4.1) for the errors is given.

TABLE 4.1

	$\lambda = 0.1$		$\lambda = 0.5$		$\lambda = 0.9$	
ε	k^*	$\hat{\lambda}$	k^*	$\hat{\lambda}$	k^*	$\hat{\lambda}$
10^{-3}	4	0.1032	10	0.5231	66	0.9089
10^{-5}	6	0.1019	17	0.5128	110	0.9053
10^{-7}	8	0.1014	24	0.5089	153	0.9038

Suppose now that $\{z^k\}$ is generated by the Newton process (3.10) for the equation (3.1) and that the mapping P satisfies the conditions stated before Theorem 3.1. If the starting point z^0 is in the ball $B(z^*, r_1^*)$ with the radius $r_1^* = 2/(3\beta\gamma)$ specified in the theorem then the inequality (3.9) holds and hence the errors satisfy

$$(0 \leqq) e_{k+1} \leqq \frac{\frac{1}{2}\beta\gamma e_k^2}{1 - \beta\gamma e_k}, \qquad k = 0, 1, \cdots. \tag{4.5}$$

This difference inequality shall now be used as our model for the error behavior. Evidently (4.5) suggests that we consider the relative errors $\varepsilon_k = e_k / r_1^*$, $k = 0, 1, \cdots$, for which (4.5) reduces to

$$\begin{aligned} &(0 \leqq) \varepsilon_{k+1} \leqq \phi(\varepsilon_k), \qquad k = 0, 1, \cdots, \\ &\phi(t) = \frac{t^2}{3 - 2t}, \qquad 0 \leqq t \leqq 1. \end{aligned} \tag{4.6}$$

Clearly, (4.5) shows that $Q_2\{z^k\} \leqq \frac{1}{3}r_1^*$, and hence the Q-factor here is not a very informative quantity. We need a measure of the convergence quality which involves the number of steps taken to termination. Such a measure is suggested by the following result which provides an explicit upper bound for the relative errors ε_k:

LEMMA 4.1. *With the function ϕ of* (4.6) *let $\{\eta_k\}$ be defined by*

$$\eta_{k+1} = \phi(\eta_k), \qquad k = 0, 1, \cdots, \quad 0 < \eta_0 < 1. \tag{4.7}$$

Then

$$\eta_k = \frac{3}{1 + 2\cosh 2^k \alpha}, \qquad k = 0, 1, \cdots, \tag{4.8}$$

where α is the unique positive solution of

$$\frac{3}{1 + 2\cosh \alpha} = \eta_0. \tag{4.9}$$

Proof. Because of $0 < \eta_0 < 1$ we have $0 < \eta_k < 1$ for all $k \geqq 0$. Then under the substitution $\tau_k = -1 + 3/\eta_k$ the iteration (4.7) becomes $\tau_{k+1}^2 = \tau_k^2 - 2$, $k = 0, 1, \cdots$ and the quantity $\alpha > 0$ defined by (4.9) is the unique positive solution of $\cosh \alpha = \tau_0$. But then it follows immediately that $\tau_k = 2\cosh 2^k \alpha$, $k = 0, 1, \cdots$ which proves (4.8).

If $\{\varepsilon_k\}$ satisfies (4.6) and $0 < \varepsilon_0 < 1$ then it is readily verified that for $\eta_0 = \varepsilon_0$ the sequence $\{\eta_k\}$ defined by (4.7) majorizes $\{\varepsilon_k\}$, or, in other words, that $\eta_k \geqq \varepsilon_k$, $k = 0, 1, \cdots$. As we noted after Theorem 3.1, there are functions P which satisfy the conditions of that theorem and for which $r_1^* = r(\mathcal{N}, z^*)$. It is easily seen that for the particular function P given in [3] we have $\varepsilon_k = \eta_k$ for all $k \geqq 0$. In this sense, the majorizing sequence $\{\eta_k\}$ of the lemma is best possible.

Our results provides that $\varepsilon_0 \in (0, 1)$ or the equivalent solution $\alpha = \alpha(\varepsilon_0) > 0$ of (4.9) with $\eta_0 = \varepsilon_0$ represents a measure of the convergence quality of our sequence. Evidently $\varepsilon_0 = e_0/r_1^*$ indicates how far inside the ball $B(z^*, r_1^*)$ we started the process, and hence ε_0 is related to the complexity measures of Newton's method considered in [18].

In order to compute an a posteriori estimate of ε_0 and/or α note first that for any $\omega \in (0, 1)$ and $k^* \geqq 2$ the equation

$$\frac{\eta_{k^*-1}}{\eta_0} \equiv \frac{\phi^{k^*-1}(\eta_0)}{\eta_0} = \omega \tag{4.10}$$

has a unique solution $\eta_0 \in (0, 1)$. Here ϕ^k denotes the kth iterate of ϕ. In fact it is easily verified that for any fixed $k \geqq 1$ the function

$$\Psi_k(t) \equiv \begin{cases} \dfrac{1}{t}\phi^k(t) & \text{for } 0 < t \leqq 1, \\ 0 & \text{for } t = 0 \end{cases} \tag{4.11}$$

is continuous and strictly monotonically increasing on $[0, 1]$. Hence, since $\Psi_k(0) = 0$, $\Psi_k(1) = 1$, it follows that the equation $\Psi_{k^*-1}(\eta_0) = \omega$, that is, (4.10), has a unique solution $\eta_0 = \eta_0(\omega, k^*) \in (0, 1)$ for any $\omega \in (0, 1)$ and that $\eta_0(\omega, k^*)$ increases with ω. For the computation, note that by Lemma 4.1 equation (4.10) has the explicit form

$$\Psi_{k^*-1}(\eta_0(\alpha)) \equiv \frac{1 + 2\cosh \alpha}{1 + 2\cosh 2^{k^*-1}\alpha} = \omega. \tag{4.12}$$

Hence it is easiest to use Newton's method to solve (4.12) for α and then to compute η_0 from (4.9).

Now suppose that our basic iterative process terminated with z^{k^*}. Then, as in the linear case, we may use the approximation

$$\tilde{\omega} = \frac{\|z^{k^*-1} - z^{k^*}\|}{\|z^0 - z^{k^*}\|} \doteq \frac{e_{k^*-1}}{e_0} = \frac{\varepsilon_{k^*-1}}{\varepsilon_0} \leqq \frac{\eta_{k^*-1}}{\eta_0} = \omega^*, \tag{4.13}$$

and compute, as indicated, the solution $\tilde{\varepsilon}_0 = \eta_0(\tilde{\omega}, k^*) \leqq \eta_0(\omega^*, k^*) = \varepsilon_0$ which represents an a posteriori lower estimate of ε_0. Once again, of course, it is impossible to find an upper bound for ε_0 unless lower bounds for the errors e_k are known. Note also that $\tilde{\varepsilon}_0$ was obtained under the assumption that the errors satisfy (4.5); hence it makes little sense to derive from $\tilde{\varepsilon}_0$ an approximation of the radius r_1^*.

The error model (4.5) is by no means the only one we may use for the derivation of convergence measures for Newton's method. In order to illustrate this we discuss another model which was essentially introduced in [3]. Suppose again that P satisfies the condition stated before Theorem 3.1. Let $D_0 \subset \mathrm{Dom}\,(P)$ be a compact, convex subset with $z^* \in \mathrm{int}\,(D_0)$ such that $P'(z)$ is nonsingular and $\|P'(z)^{-1}\| \leqq \kappa$ for any $z \in D_0$. On D_0 the Newton-operator (3.8) is well-defined and for any $z \in D_0$ the standard estimate

$$\|Nz - z^*\| = \|-P'(z)^{-1}[P(z^*) - P(z) - P'(z)(z^* - z)]\| \leqq \tfrac{1}{2}\kappa\gamma\|z - z^*\|^2 \tag{4.14}$$

holds. For abbreviation set $\mu = \frac{1}{2}\kappa\gamma$. Moreover, if the Newton process converges to z^* then the convergence is at least quadratic with $Q_2\{z^k\} \leqq \mu$.

However, we noted before that $Q_2\{z^k\}$ is not a particularly informative quantity. By (4.14) the supremum

$$c = \sup_{z \in D_0, z \neq z^*} \frac{\|Nz - z^*\|^2}{\|z - z^*\|^4 + \|z - z^*\|^2\|Nz - z\|^2} \tag{4.15}$$

exists and we have $c \leqq \mu^2$. Evidently, c is closely related to the Q_2-factor but differs from it in the use of the supremum rather than the limit superior. In analogy to Theorem 3.1 we now obtain the following result:

THEOREM 4.2. *If* $\bar{B}(z^*, \rho_0) \subset D_0$ *and* $c^2\rho_0 < 1$, *then*

$$\|Nz - z^*\| \leqq \frac{c\|z - z^*\| + \sqrt{2c - c^2\|z - z^*\|^2}}{1 - c\|z - z^*\|^2}\|z - z^*\|^2, \qquad z \in \bar{B}(z^*, \rho_0). \tag{4.16}$$

Moreover, if $\bar{B}(z^*, r_2^*) \subset D_0$ *for* $r_2^* = (5c)^{-1/2}$ *then for any* $z^0 \in B(z^*, r_2^*)$ *the Newton sequence remains in* $B(z^*, r_2^*)$ *and converges to* z^*.

Proof. For any $z \in \bar{B}(z^*, \rho_0)$ it follows from (4.15) that

$$\|Nz - z^*\|^2 \leqq c[\|z - z^*\|^2 + \|Nz - z\|^2]\|z - z^*\|^2. \tag{4.17}$$

Hence using

$$\|Nz - z^*\|^2 \geqq \|z - z^*\|^2 + \|Nz - z\|^2 - 2\|z - z^*\|\|z - Nz\| \tag{4.18}$$

we obtain

$$\begin{aligned}\|Nz - z^*\|^2 &\leqq c[\|Nz - z^*\|^2 + 2\|z - z^*\|\|z - Nz\|]\|z - z^*\|^2 \\ &\leqq c[\|Nz - z^*\|^2 + 2\|z - z^*\|(\|z - z^*\| + \|Nz - z^*\|)]\|z - z^*\|^2\end{aligned}$$

which is equivalent with

$$(1 - c\|z - z^*\|^2)\|Nz - z^*\|^2 - 2c\|Nz - z^*\|\|z - z^*\|^3 - 2c\|z - z^*\|^4 \leqq 0.$$

For any $z \in B(z^*, \rho_0)$ the solution of this quadratic inequality in $\|Nz - z^*\|$ gives (4.16).

Now let $\zeta:[0, r_2^*]\to R^1$ be defined by

$$\zeta(t)=\frac{ct+\sqrt{2c-c^2t^2}}{1-ct^2}t. \tag{4.19}$$

Then (4.16) assumes the form

$$\|Nz-z^*\|\leqq\zeta(\|z-z^*\|)\|z-z^*\| \quad \forall z\in B(z^*, r_2^*). \tag{4.20}$$

Evidently ζ satisfies $\zeta(0)=0$, $\zeta(r_2^*)=1$ and is continuous and monotonically increasing. This implies the second part of the statement.

The result shows that for any starting point z^0 in the ball $B(z^*, r_2^*)$ the errors of the corresponding Newton sequence $\{z^k\}$ satisfy

$$(0\leqq)e_{k+1}\leqq\zeta(e_k)e_k, \qquad k=0, 1, \cdots. \tag{4.21}$$

This difference inequality is our new model for the error behavior of our Newton sequence. As before we introduce the relative errors $\varepsilon_k=e_k/r_2^*$, $r_2^*=(5c)^{-1/2}$, for which (4.21) reduces to

$$(0\leqq)\varepsilon_{k+1}\leqq\frac{\varepsilon_k+\sqrt{10-\varepsilon_k^2}}{5-\varepsilon_k^2}\varepsilon_k^2, \qquad k=0, 1, \cdots. \tag{4.22}$$

Evidently, the initial relative error $\varepsilon_0\in(0, 1)$ represents a measure of the convergence quality of our Newton sequence.

An a posteriori estimate of ε_0 can now be computed as in the previous case. Instead of the function ϕ of (4.6) we introduce

$$\phi(t)=\frac{t+\sqrt{10-t^2}}{5-t^2}t^2, \qquad 0\leqq t\leqq 1 \tag{4.23}$$

and consider the sequence

$$\eta_{k+1}=\phi(\eta_k), \qquad k=0, 1, \cdots. \tag{4.24}$$

Clearly, for $\eta_0=\varepsilon_0$ this sequence $\{\eta_k\}$ majorizes the sequence of relative errors $\{\varepsilon_k\}$. With ϕ defined by (4.23) the equation (4.10) has again a unique solution $\eta_0=\eta_0(\omega, k^*)\in(0, 1)$ for any $\omega\in(0, 1)$ and $k^*\geqq 2$. In fact, for fixed $k\geqq 1$ the corresponding function Ψ_k specified as in (4.11) is again continuous and strictly monotonically increasing on $[0, 1]$ and satisfies $\Psi_k(0)=0$ and $\Psi_k(1)=1$. Hence we see also that $\eta_0(\omega, k)$ increases with ω. The only difference is here that there is no explicit form for Ψ_{k^*-1} comparable to (4.12). Hence an evaluation of $\Psi_{k^*-1}(\eta_0)$ requires the computation of the iterates (4.24) up to the (k^*-1)th term. Various standard solution techniques may be applied to the solution of (4.10), such as bisection or even Newton's method. As in the previous case we now calculate the lower bound $\tilde{\omega}$ of ω^* given by (4.13) and then use it to compute the solution $\tilde{\varepsilon}_0=\eta_0(\tilde{\omega}, k^*)\geqq\eta_0(\omega^*, k^*)=\varepsilon_0$. This is the desired a posteriori lower estimate of ε_0.

As in the case of the linear error model (4.1) it is possible to assess the influence of the approximation (4.13) upon the computed value of ε_0, for both our models for Newton's method. In fact in each case we have

$$\hat{\omega}=\frac{\eta_{k^*-1}+\eta_{k^*}}{\eta_0-\eta_{k^*}}\geqq\max(\omega^*, \tilde{\omega}) \tag{4.25}$$

whence

$$\hat{\varepsilon}_0=\eta_0(\hat{\omega}, k^*)\geqq\max(\tilde{\varepsilon}_0, \varepsilon_0). \tag{4.26}$$

If $\varepsilon_0 \in (0, 1)$ and a tolerance $\delta \in (0, \varepsilon_0)$ are given then it is easy to determine from (4.8) or (4.24) the smallest index $k^* \geqq 2$ such that $\eta_{k^*} \leqq \delta$. With this k^* the quantities $\hat{\omega}$ of (4.25) and $\hat{\varepsilon}_0$ of (4.26) can be computed. Tables 4.2 and 4.3 gives some results for different values of ε_0 and δ for the case of the model (4.5) and (4.21), respectively. Clearly, the approximations are certainly very good.

TABLE 4.2

	$\varepsilon_0 = 0.1$		$\varepsilon_0 = 0.5$		$\varepsilon_0 = 0.9$	
ε	k^*	$\hat{\varepsilon}_0$	k^*	$\hat{\varepsilon}_0$	k^*	$\hat{\varepsilon}_0$
10^{-3}	2	0.10012	3	0.50023	4	0.90047
10^{-5}	2	0.10012	4	0.50000	5	0.90000
10^{-7}	3	0.10000	4	0.50000	5	0.90000

TABLE 4.3

	$\varepsilon_0 = 0.1$		$\varepsilon_0 = 0.5$		$\varepsilon_0 = 0.9$	
ε	k^*	$\hat{\varepsilon}_0$	k^*	$\hat{\varepsilon}_0$	k^*	$\hat{\varepsilon}_0$
10^{-3}	2	0.10043	3	0.50227	5	0.90053
10^{-5}	3	0.10000	4	0.50002	6	0.90000
10^{-7}	3	0.10000	4	0.50002	6	0.90000

5. Steplength algorithms. In this section the convergence measures introduced in § 4 will be used in the design of several practical steplength algorithms for our continuation process. Basically, the algorithms are based on the formulas (2.16), (2.17); that is, the predicted step is essentially given by

$$\tilde{h}_{p+1} = \tilde{\Lambda}_{p+1}\Delta s_p, \qquad \Delta s_p = \|x^p - x^{p-1}\|_2, \qquad \tilde{\Lambda}_{p+1} = \sqrt{\frac{\frac{\rho_{p+1}}{\Delta s_p}}{2\left|\sin\left(\frac{\alpha_p}{2}\right)\right|}},$$

(5.1)

$$\alpha_p = \arccos\left((Tx^p)^T \frac{x^p - x^{p-1}}{\Delta s_p}\right).$$

The notation $\tilde{h}_{p+1}$, $\tilde{\Lambda}_{p+1}$ is used here to indicate that these values still need to be adjusted to fall between prescribed bounds. We discuss first the question of the choice of ρ_{p+1}.

Already simple examples show that it is very disadvantageous to set ρ_{p+1} equal to an a priori constant; instead, the choice of ρ_{p+1} should reflect the expected convergence behavior of the corrector process. Ideally, we would like to obtain ρ_{p+1} by extrapolation from the convergence radii at prior points along the curve. But as we saw in § 3 computed estimates of these radii are impossible to obtain. On the other hand, at x^p

we can calculate the distance

$$\delta_p = \|z^0 - x^p\| \tag{5.2}$$

between the predicted starting point

$$z^0 = y_{p-1}(h_p) = x^{p-1} + h_p T x^{p-1} \tag{5.3}$$

and the accepted point x^p of the corrector process. It is then natural to consider the choice $\rho_{p+1} = \delta_p$. However, this leads to a poor step-algorithm. In fact, the smaller the prediction error δ_p the smaller will be $\tilde{h}_{p+1}$. Hence if the curvature does not change too quickly the result is likely to be an even smaller prediction error δ_{p+1}. In other words, the choice $\rho_{p+1} = \delta_p$ may be expected to lead to a decreasing sequence of steps and numerical experiments confirm readily that this is indeed the case.

Evidently, we should choose

$$\rho_{p+1} = \theta_{p+1}\delta_p, \tag{5.4}$$

where the factor θ_{p+1} is a function of some computable measure of the convergence of the corrector process at x^p. More specifically, θ_{p+1} should increase when the convergence quality increases. In order to illustrate our approach in designing the choice of θ_{p+1}, suppose that the corrector is a linearly convergent process to which the error model (4.1) applies. Let $z^0, z^1, \cdots, z^{k^*} = x^p$ be the corrector iterates at x_p. Then the estimate $\tilde{\lambda}$ of (4.3) of $Q_1\{z^k\}$ can be computed. A reasonable aim in the construction of the steps along the curve is to ensure that the number of corrector iterates remains approximately constant. In other words, we aim at taking always, say, m^* corrector steps. Hence, under the heuristic assumption that the error model (4.1) remains valid for some interval of starting errors $e_0 \doteq \delta_p$, we should have started with an error $\theta\delta_p$ such that

$$\tilde{\lambda}^{m^*}(\theta\delta_p) \doteq \tilde{\lambda}^{k^*}\delta_p. \tag{5.5}$$

In line with this it is natural to define the factor θ_{p+1} in (5.4) as

$$\theta_{p+1} = \tilde{\lambda}^{k^*-m^*}. \tag{5.6}$$

Hence for $k^* < m^*$ we have $\theta_{p+1} > 1$ and δ_p is indeed increased as expected.

This technique is also readily applicable in the case of the two convergence measures for Newton's method discussed in the previous section. Let $z^0, z^1, \cdots, z^{k^*} = x^p$ again be the corrector iterates at x^p and assume that this time either one of the models (4.5) or (4.21) applies. Then the quantity $\tilde{\omega}$ of (4.13) can be computed and with it the estimated convergence measure $\tilde{\varepsilon}_0 = \eta_0(\tilde{\omega}, k^*)$. By definition $\tilde{\varepsilon}_0 = e_0/r_i^*$ approximates the relative starting error with respect to the (unknown) convergence radius r_1^* of Theorem 3.1 or r_2^* of Theorem 4.2. On the other hand, for the desired number of m^* corrector steps the corresponding value $\tilde{\varepsilon}_0'$ satisfies

$$\phi^{m^*}(\tilde{\varepsilon}_0') = \phi^{k^*}(\tilde{\varepsilon}_0), \tag{5.7}$$

where ϕ is given by (4.6) or (4.23), respectively. Hence $\tilde{\varepsilon}_0'$ is the unique solution of

$$\tilde{\varepsilon}_0' \Psi_{m^*}(\tilde{\varepsilon}_0') = \delta, \qquad \delta = \phi^{k^*}(\tilde{\varepsilon}_0) \in (0, 1), \tag{5.8}$$

where Ψ_k is given by (4.11). By the properties of Ψ_k this equation obviously has a unique solution $\tilde{\varepsilon}_0' \in (0, 1)$. The computation of $\tilde{\varepsilon}_0'$ therefore is similar to that of $\tilde{\varepsilon}_0$. In line with our approach in the linear case, it is then natural to define the factor θ_{p+1} in (5.4) by

$$\theta_{p+1} = \frac{\tilde{\varepsilon}_0'}{\tilde{\varepsilon}_0}. \tag{5.9}$$

In all cases so far our steplength algorithm consists in the computation of a particular factor θ_{p+1} which by (5.4) gives us the tolerance ρ_{p+1} and with it the values $\tilde{\Lambda}_{p+1}$ and $\tilde{h}_{p+1}$ of (5.1). As indicated before, these values still need to be adjusted to fall between allowable minimal and maximal values. In other words, the actual values Λ_{p+1} and h_{p+1} are required to satisfy

$$\text{(5.10)} \qquad \begin{aligned} &\text{(i)} && \frac{1}{\kappa} \leqq \Lambda_{p+1} \leqq \kappa, \qquad \kappa > 1, \\ &\text{(ii)} && 0 < h_{\min} \leqq h_{p+1} \leqq h_{\max}. \end{aligned}$$

As usual, for any $\alpha < \beta$ define the piecewise linear function

$$\text{(5.11)} \qquad \chi(s; \alpha, \beta) = \begin{cases} \alpha & \text{if } s \leqq \alpha, \\ s & \text{if } \alpha \leqq s \leqq \beta \\ \beta & \text{if } s \geqq \beta. \end{cases} \quad \forall s \in R^1,$$

Then (5.10) holds if we set

$$\text{(5.12)} \qquad \begin{aligned} \Lambda_{p+1} &= \chi\left(\tilde{\Lambda}_{p+1}; \frac{1}{\kappa}, \kappa\right), \\ h_{p+1} &= \chi(\Lambda_{p+1} \Delta s_p; h_{\min}, h_{\max}). \end{aligned}$$

Generally, the constants κ, $h_{\min}$, $h_{\max}$ depend on the problem and on the computer used.

In the computation of the critical factor θ_{p+1} of (5.4) we assumed that the corrector errors satisfy a particular error model for which a measure of the convergence quality of the corrector iteration can be obtained. Several possible models were discussed here and, of course, others can be analyzed by similar techniques. On the other hand, if no satisfactory information about a possible error model is available it is possible to obtain a rough value for θ_{p+1} by a careful enforcement of the condition (5.10)(i). Evidently, the value of $\tilde{\Lambda}_{p+1}$ in (5.1) is strongly influenced by that of α_p. In particular, we need to avoid small values of $|\alpha_p|$. Let $\alpha \in (0, \pi/2)$ be a lower threshold of $|\alpha_p|$ and replace the denominator under the square root of $\tilde{\Lambda}_{p+1}$ in (5.1) by the quantity

$$\text{(5.13)} \qquad \omega_p = 2\left|\sin \frac{1}{2} \chi\left(\alpha_p; \alpha, \frac{\pi}{2}\right)\right|$$

where χ is again given by (5.11). In line with (5.10)(i) it is then natural to consider the following bounds for the relative tolerance $\sigma_{p+1} = \rho_{p+1}/\Delta s_p$:

$$\sigma_{\max} = 2\kappa^2 \sin \frac{\alpha}{2}, \qquad \sigma_{\min} = \frac{2}{\kappa^2} \sin \frac{\pi}{4} = \frac{\sqrt{2}}{\kappa^2}.$$

Accordingly the relative tolerance should be restricted by

$$\text{(5.14)} \qquad \sigma_{p+1} = \begin{cases} \sigma_{\max} & \text{if } \left|\sin \frac{\alpha_p}{2}\right| \leqq \sin \frac{\alpha}{2} \text{ or } \delta_p \geqq \sigma_{\max} \Delta s_p, \\ \sigma_{\min} & \text{if } \left|\sin \frac{\alpha_p}{2}\right| \geqq \sqrt{2} \text{ or } \delta_k \leqq \sigma_{\min} \Delta s_p, \\ \dfrac{\delta_p}{\Delta s_p} & \text{otherwise.} \end{cases}$$

Clearly, we need

(5.15) $\alpha \geqq \underline{\alpha} = 2 \arcsin((\sqrt{2}\,\kappa^4))^{-1})$ to ensure that $\sigma_{\min} < \sigma_{\max}$.

Then for $\underline{\alpha} < \alpha < \pi/2$ the tentative step

$$\tilde{h}_{p+1} = \Lambda_{p+1}\Delta s_p, \quad \Lambda_{p+1} = \sqrt{\frac{\sigma_{p+1}}{\omega_p}} \tag{5.16}$$

satisfies (5.10)(i). As before, the actual step is again given by (5.12). This approach has completely avoided the calculation of a quality measure for the corrector iterates. In other words, the steps computed in this way do not take any account of any changes in the convergence behavior of the correctors. Nevertheless for suitably chosen values of κ and α the resulting algorithm does perform surprisingly well. Obviously the choice of κ and α depends on the problem, but often $\kappa = 3$ and $\alpha = 0.05$ are fairly reasonable.

Once the predictor step h_{p+1} has been chosen, the corrector iteration is started from $x^p + h_{p+1}Tx^p$. This process has to incorporate provisions for monitoring the convergence and for aborting the iteration as soon as divergence is detected. These provisions depend of course on the type of corrector used. But in most cases it has been found to be satisfactory to declare nonconvergence if either one of the following three conditions are true:

$$\begin{aligned} &\text{(i)} && \|Pz^j\|_\infty \geqq \mu \|Pz^{j+1}\|_\infty && \text{for some } j \geqq 1, \\ &\text{(ii)} && \|z^j - z^{j-1}\|_\infty \geqq \mu \|z^{j-1} - z^{j-2}\|_\infty && \text{for some } j \geqq 2, \\ &\text{(iii)} && j \geqq j_{\max}. \end{aligned} \tag{5.17}$$

Here P denotes the operator defining the current corrector equation (2.11), μ is a suitable constant, say, $\mu = 1.05$, and $j_{\max}$ some integer, say $j_{\max} = 8$. On the other hand, the jth corrector iterate z^j is accepted as the next point x^{k+1} if either one of the two conditions holds

$$\begin{aligned} &\text{(i)} && \|Pz^0\|_\infty \leqq \delta_1, \\ &\text{(ii)} && (\|Pz^j\|_\infty \leqq \delta_2) \quad \text{and} \quad (\|z^j - z^{j-1}\|_\infty \leqq \delta_3 + \delta_4 \|z^j\|_\infty) \quad \text{for some } j \geqq 1, \end{aligned} \tag{5.18}$$

with given tolerances δ_1, δ_2, δ_3, δ_4 which depend on the machine and the problem.

In the case of nonconvergence the predictor step is halved and the corrector is restarted from $x^p + \frac{1}{2}h_{p+1}Tx^p$ (with the corresponding constant (2.14)). However, it is required here that $\frac{1}{2}h_{p+1} \geqq h_{\min}$, otherwise some user-dependent action needs to be taken to modify $h_{\min}$ or to stop the overall process. On the other hand, if the convergence is declared with $x^{p+1} = z^j$ as the last iterate then, with Δs_p given by (5.1), the arclength s_{p+1} corresponding to x^{p+1} is approximated by $s_{p+1} \doteq s_p + \Delta s_p$ and—if desired—a new predictor-corrector step is taken.

6. Some numerical results. In § 5 the formulas (5.9)/(4.6), (5.9)/(4.23), and (5.10) define three steplength algorithms which we shall identify from now on by the labels I, II, and III, respectively. Each one of these step-algorithms was incorporated in the general continuation process described in § 2 (see also [15] for some further details of this process). The resulting three procedures are again identified by the labels I, II, III; note that they are the same in all details except for the step-prediction. We present here some numerical results obtained with these three procedures.

Example 1. In order to illustrate the comparative performance of the three procedures we consider first a very small problem which was originally formulated in [8] and subsequently used as a test case by many authors. More specifically, the

mapping F has here the form

$$F: R^3 \to R^2, \quad Fx = Gy - (1-t)Gy^0, \quad x = \begin{pmatrix} y \\ t \end{pmatrix}, \qquad y \in R^2, \quad t \in R^1 \tag{6.1}$$

where

$$G: R^2 \to R^2, \qquad Gy = \begin{pmatrix} y_1 + 5y_2^2 - y_2^3 - 2y_2 - 13 \\ y_1 + y_2^2 + y_2^3 - 14y_2 - 29 \end{pmatrix}, \qquad y^0 = \begin{pmatrix} 15 \\ 2 \end{pmatrix}. \tag{6.2}$$

For the starting point $x^0 = (15, 2, 0)^T$ the solution of $Fx = 0$ passes through $x^* = (5, 4, 1)^T$ and this curve was traced from x^0 to x^*.

It may be noted that this problem represents an excellent example for the importance of choosing the proper parametrization of the solution curve. In fact, in this case we have

$$Fx \equiv \begin{pmatrix} 1 & 34 \\ 1 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} - \begin{pmatrix} 47 \\ 39 \end{pmatrix} - \begin{pmatrix} x_2^3 - 5x_2^2 + 2x_2 \\ -x_2^3 - x_2^2 + 14x_2 \end{pmatrix} = 0,$$

and hence

$$\begin{aligned} x_1 &= \frac{107}{3} - \frac{11}{6}x_2^3 + \frac{2}{3}x_2^2 + \frac{57}{3}x_2, \\ x_3 &= \frac{1}{3} + \frac{1}{12}x_2^3 - \frac{1}{6}x_2^2 - \frac{1}{2}x_2 \end{aligned} \tag{6.3}$$

is an explicit representation of the solution in terms of the parameter x_2. From this it follows readily that for $x_2 \doteq 1.741376$ and $x_2 \doteq 1.983801$ we have limit points with respect to x_1 and for $x_2 \doteq -0.896805$ and $x_2 \doteq 2.230139$ limit points with respect to x_3.

TABLE 6.1

x_2	$\tilde{\Lambda}_p^{\mathrm{I}}$	$\tilde{\Lambda}_p^{\mathrm{II}}$	$\tilde{\Lambda}_p^{\mathrm{III}}$	Λ_p^{I}	h_p	Δs_p
−1.94	3.43	3.91	3.	3.	0.91	0.23(3)
−1.00	1.99	2.10	1.54	1.99	0.46	0.23(4)
−1.76	0.55	0.51	0.59	0.55	0.15	0.15
−1.66(1)	17.02	17.27	0.97	3.	0.45	0.45
−1.53	1.60	1.60	1.40	1.60	0.62	0.62
−1.42	2.57	2.74	3.	2.57	1.58	1.58
−1.21	2.61	2.77	3.	2.61	4.09	4.09
−0.86(2)	2.78	2.94	3.	2.78	11.33	11.33
−0.18	3.84	3.98	3.	3.	33.98	16.99(4)

Table 6.1 shows comparative values of the step-factors Λ_p for procedure I. More specifically the following parameters were used:

$$\kappa = 3, \quad h_{\max} = 100, \quad h_{\min} = 0.001, \quad \alpha = 0.05, \quad \mu = 1.05, \quad j_{\max} = 8, \tag{6.4}$$

$$i_0 = 3, \quad h_0 = 0.3, \quad \delta_i = 10^{-5}, \quad i = 1, 2, 3, 4, \tag{6.5}$$

and the table gives some data for the first nine steps. The numbered notes refer to the following observations: (1) A limit point for x_1 has been passed; (2) a limit point for x_2 has been passed; (3) two step reductions were needed since the initial step $h_0 = 0.3$ was too large; and (4) one step reduction was taken. Evidently the $\tilde{\Lambda}_p$-values for the first two step-algorithms are essentially the same and with one exception those of the

third algorithm are also comparable. This is not an isolated case but represents our experience for numerous problems. Table 6.2 shows the total number of steps, step-reductions, and Jacobian evaluations for the three procedures to reach x^* from x^0. Once again the performance of all three procedures is comparable. The simple procedure III has here a slight edge over the others. This is due to the fact that the more conservative strategy of procedure III results in fewer step-reductions and with that in a slightly better overall performance. In connection with Table 6.2 it may be mentioned that the procedure discussed in [8] requires 25 s-steps and 157 Jacobian evaluations for the same task.

TABLE 6.2

	Procedure		
Number of	I	II	III
s-steps	24	23	22
step reductions	19	19	15
Jacobian evaluations	146	142	128

Example 2. The following two-point boundary value problem arises in the analysis of electrostatic potentials in semiconductors (see [10]):

$$\text{(6.6) (i)} \qquad \begin{aligned} V''(\xi)+f(V(\xi),\xi)&=0, \qquad a<\xi<b,\\ V(a)&=V_a, \qquad V(b)=V_b \end{aligned}$$

where

$$\text{(6.6) (ii)} \qquad \begin{aligned} f(V,\xi)&=C_a e^{\hat{\beta}(V_a-V)}-C_b e^{\hat{\beta}(V-V_b)}+D(\xi),\\ D(\xi)&=\begin{cases} -C_a & \text{for } \xi\leqq 0,\\ C_b & \text{for } \xi>0. \end{cases} \end{aligned}$$

We embed (6.6) into the family of problem defined by

$$\text{(6.7)} \qquad V_a=v_a t, \quad V_b=v_b t, \quad C_a=c_a t, \quad C_b=c_b t, \quad \hat{\beta}=\beta t, \quad 0\leqq t\leqq 1.$$

Clearly, for $t=0$ the solution is trivial.

For the numerical solution a finite element approximation of (6.6/7) was introduced based on piecewise linear elements on a fixed mesh

$$\text{(6.8)} \qquad a=\xi_0<\xi_1<\cdots<\xi_{n+1}=b.$$

Simpson's rule was used in the computation of the elemental stiffness matrices. The resulting system of nonlinear equations has the form

$$\text{(6.9)} \qquad Fx\equiv Ay-H(y,t)=0, \qquad x=\begin{pmatrix} y\\ t\end{pmatrix}, \qquad y\in R^n, \quad t\in R^1$$

with some nonsingular $n\times n$ matrix A and nonlinear mapping H from R^{n+1} into R^n. It is easily verified that for practically interesting values of the parameters the solution of $Fx=0$ through $x^0=0$ exists and that it passes through a point $x^*=(y^*, 1)$ for which y^* represents a finite element approximation of the original problem (6.6) with the parameters $V_a=v_a$, $V_b=v_b$, $C_a=c_a$, $C_b=c_b$, $\hat{\beta}=\beta$.

Practically interesting values of the parameters in (6.6) are

$$a = -9 \cdot 10^{-5}, \quad b = 10^{-5}, \quad v_a = 0, \quad v_b = 700, \quad \alpha = 40, \quad c_a = 10^{12}, \quad c_v = 10^{13}.$$

The solution of the corresponding problem (6.9) certainly requires double precision. Since our current codes were set up for single precision we used instead the parameter set

$$(6.10)\quad a = -9 \cdot 10^{-3}, \quad b = 10^{-3}, \quad v_a = 0, \quad v_b = 25, \quad \alpha = 20, \quad c_a = c_b = 10^r, \quad r = 6, 7$$

and corresponding to it the mesh (6.8) given by

i	0	1	2	3	4	5	6	7	8	9	10
$\xi_i \cdot 10^3$	−9	−7	−5	−4.5	−4.25	−4	−3.75	−3.5	−3	−1	1

The solution of $Fx = 0$ was traced from x^0 to x^* with the continuation parameters selected according to (6.4) and

$$(6.11)\qquad i_0 = 10, \quad \delta_i = 10^{-2}, \quad i = 1, 2, 3, 4.$$

Table 6.3 gives summary data for the three procedures to reach x^* from x^0.

TABLE 6.3

	$c_a = c_b = 10^6$, $h_0 = 3$			$c_a = c_b = 10^7$, $h_0 = 1$		
	I	II	III	I	II	III
Number of						
s-steps	4	4	4	8	8	9
step reductions	0	0	0	5	5	2
Jacobian evaluations	10	10	10	38	41	33

As before, the performance of the three procedures is fully comparable. The second case is near the limit of the single precision computation and, once again, the conservative strategy of procedure III proves to be slightly more advantageous. However, this conservative nature turns out to be disadvantageous when the stopping criteria is loosened up as Table 6.4 shows where the parameters of (6.11) were replaced by $i_0 = 10$, $\delta_i = 5.10^{-2}$, $i = 1, 2, 3, 4$.

TABLE 6.4

	I	II	III
Number of			
s-steps	5	5	9
step reductions	0	0	1
Jacobian evaluations	18	19	40

Example 3. Our final example concerns a finite element analysis of a clamped, thin, shallow, circular arch under uniform load. This problem has been used as a test case by various authors, see, e.g., [1], [6], [7], [17], [20]. We follow the specific formulation given in [7] and shall not repeat the details. Briefly, the nonlinear strain-displacement relationships are of the form

$$(6.12)\qquad \varepsilon_\theta = \frac{1}{R}\left(\frac{dw}{d\theta} - u\right) + \frac{1}{2R^2}\left(\frac{du}{d\theta}\right)^2, \qquad \kappa_\theta = \frac{1}{R^2}\frac{d^2u}{d\theta^2},$$

where u and w are the radial and axial displacements, respectively, and R is the arch radius. For uniform radial loading of intensity p the total potential energy of the system is then given by

$$\frac{ER}{2}\int_{-\theta_0}^{\theta_0}(A\varepsilon_\theta^2+I\kappa_\theta^2)\,d\theta-R\int_{-\theta_0}^{+\theta_0}pu\,d\theta. \tag{6.13}$$

In our case the following arch properties were used:

$$R=10\text{ in},\quad EA=2.056\cdot 10^6\text{ lb},\quad EI=796\text{ lb in}^2,\quad \theta_0=15^0. \tag{6.14}$$

The finite element approximation was based on the element-displacement functions of [20] and eight equal elements. In addition to the "perfect" structure the case of asymmetric imperfection was considered as specified in [7]. The solution curve was traced from the zero solution for $p=0$ through the limit point with respect to the load variable until a central deflection value of 1.8 was reached. In the case of the perfect structure this means that a bifurcation point has to be passed as well. Tables 6.5–6.7 show the summary data for this trace for the perfect structure and the imperfect structure with perturbation parameters $\gamma=0.5$ and $\gamma=1.5$, respectively (see [7]). The continuation parameters (6.4) and

$$i_0=46,\quad h_0=2,\quad \delta_i=0.01,\quad i=1,2,3,4$$

were used.

TABLE 6.5

	Perfect structure		
	I	II	III
Number of			
s-steps	7	7	10
step reductions	5	6	3
Jacobian evaluations	30	31	28

TABLE 6.6

	Inperfect structure $\gamma=0.5$		
	I	II	III
Number of			
s-steps	5	5	4
step reductions	2	2	0
Jacobian evaluations	21	21	15

TABLE 6.7

	Imperfect structure $\gamma=1.5$		
	I	II	III
Number of			
s-steps	5	5	5
step reductions	2	2	0
Jacobian evaluations	22	22	17

Once again the same observations apply, procedures I and II behave essentially in the same way, and the conservative procedure III is better whenever the solution curve shows larger changes of the curvature. It appears that this is essentially due to the relatively large value of the threshold α for the angles α_p used in (5.15). The procedures I and II could be made more conservative as well by introducing such a large threshold.

7. Asymptotic theory of optimal steps. The steplength estimation (2.16) used in § 2 for the design of our step-algorithms represents a first order approximation. The question arises what change would result if we take into account the next higher order term. This is the topic of this section.

In order to be able to consider a higher-order approximation for the step, suppose from now on that the solution curve $x: J \subset R^1 \to \mathcal{R}(F) \subset R^{n+1}$ of (2.2) under consideration is $r \geqq 4$ times continuously differentiable on its open interval if definition J. By Theorem 2.1 this is certainly true whenever F is r-times continuously differentiable. Suppose further that the points (2.7) computed so far are exactly on the curve; in other words that

$$x^j = x(s_j), \quad j = 0, 1, \cdots, p, \quad 0 = s_0 < s_1 < \cdots < s_p. \tag{7.1}$$

Then we have

$$Tx^j = \dot{x}(s_j), \qquad j = 0, 1, \cdots, p. \tag{7.2}$$

For ease of notation we suppress from now on the subscripts p and $p+1$ and hence write simply $s = s_p$. Moreover, without restriction of the generality, we assume that $\ddot{x}(s)(\equiv \ddot{x}(s_p)) \neq 0$. With a suitably chosen tolerance $\rho = \rho_{p+1}$ the "ideal"step $h = h_{p+1}$ then is the solution of the equation

$$\|x(s+h) - x(s) - h\dot{x}(s)\| = \rho. \tag{7.3}$$

For some $\hat{\delta} > 0$ the function

$$\psi(\tau) = \begin{cases} \operatorname{sgn}(\tau)\|x(s+\tau) - x(s) - \tau\dot{x}(s)\|^{1/2} & \text{for } |\tau| < \hat{\delta}, \quad \tau \neq 0 \\ 0 & \text{for } \tau = 0. \end{cases} \tag{7.4}$$

is well defined, and with $\xi = \rho^{1/2}$ (7.3) becomes

$$\psi(h) - \xi = 0. \tag{7.5}$$

Because of $\ddot{x}(s) \neq 0$ there exists a $\delta_0 \in (0, \hat{\delta})$ such that for $0 < |\tau| < \delta_0$ we have $\psi(\tau) \neq 0$ and ψ is three times continuously differentiable. More specifically it follows that

$$\psi'(\tau) = \frac{\operatorname{sgn}(\tau)}{\psi(\tau)^{3/2}}(\dot{x}(s+\tau) - \dot{x}(s))^T(x(s+\tau) - x(s) - \tau\dot{x}(s)), \qquad 0 < |\tau| < \delta_0 \tag{7.6}$$

and, by L'Hôpital's rule, that the limit for $\tau \to 0$ exists and equals

$$\psi'(0) = \frac{1}{\sqrt{2}}\|\ddot{x}(s)\|^{1/2}. \tag{7.7}$$

Since $\psi'(0) \neq 0$ the implicit function theorem now implies the existence of some small constants δ, $\varepsilon > 0$ and of a three-times continuously differentiable function $\hat{h}: (-\varepsilon, \varepsilon) \to (-\delta, \delta)$ such that $\hat{h}(0) = 0$, and for all $|\xi| < \varepsilon$, $\hat{h}(\xi)$ is the unique solution of (7.5) in $(-\delta, \delta)$. Since

$$\hat{h}(\xi) = a\xi + b\xi^2 + O(\xi^3) \quad \text{as } \xi \to 0, \tag{7.8}$$

it follows from (7.5) that

$$\begin{aligned}\xi^4 &= \|x(s+\hat{h}(\tau)) - x(s) - \hat{h}(\tau)\dot{x}(s)\|^2 \\ &= \|\tfrac{1}{2}\hat{h}(\tau)^2\ddot{x}(s) + \tfrac{1}{6}\hat{h}(\tau)^3\dddot{x}(s) + O(\xi^4)\|^2 \\ &= \tfrac{1}{4}\hat{h}(\tau)^4\{\|\ddot{x}(s)\|^2 + \tfrac{2}{3}a\xi\ddot{x}(s)^T\dddot{x}(s)\} + O(\xi^6) \quad \text{as } \xi \to 0\end{aligned}$$

whence

$$\tfrac{1}{4}(a^4 + 4a^3 b\xi)\{\|\ddot{x}(s)\|^2 + \tfrac{2}{3}a\xi\ddot{x}(s)^T\dddot{x}(s)\} + O(\xi^2) = 1 \quad \text{as } \xi \to 0. \tag{7.9}$$

Thus we have

$$\text{(i)} \qquad a = \sqrt{\frac{2}{\|\ddot{x}(s)\|}},$$

$$\text{(ii)} \qquad b = -\frac{1}{6}\kappa a^2, \qquad \kappa = \frac{\ddot{x}(s)^T\dddot{x}(s)}{\|\ddot{x}(s)\|^2}. \tag{7.10}$$

In the steplength estimate (2.16) the term $\|w^p\|$ is asymptotically equal to $\|\ddot{x}(s)\|$ as $\Delta s_p \to 0$. It is not unreasonable to assume that $\Delta s_p = O(\xi)$. Then the term $h_1 = a\xi$ in (7.8) is asymptotically equal to our steplength bound (2.16) as $\xi \to 0$.

Now consider the second order term $h_2 = b\xi^2$ given by (7.8)/(7.10)(ii). Clearly, if the third order term is neglected then $\xi > 0$ must be such that $a^4 + 4a^3 b\xi > 0$ for the approximation to be meaningful. Hence by (7.10) we have

$$1 - \tfrac{2}{3}a\xi\kappa > 0, \tag{7.11}$$

while (7.9) implies that

$$1 + \tfrac{2}{3}a\xi\kappa > 0. \tag{7.12}$$

In other words, if the third order term is negligible then

$$\left|\frac{h_2}{h_1}\right| < \frac{1}{4}, \tag{7.13}$$

which means that the use of the second order term h_2 in the approximation for the steplength is equivalent to multiplying the first order approximation by a factor between $\frac{3}{4}$ and $\frac{5}{4}$. Since we do not have a very accurate value for the tolerance ρ_{p+1} the use of the second order term appears to be unwarranted.

Acknowledgments. We would like to thank the sponsoring organizations for their support of this work, and Arthur J. Dadamo for his cooperation in the numerical experiments.

REFERENCES

[1] A. ENDO, S. KAWAMATA AND Y. HANGAI, *Post-bifurcation analysis of shallow spherical shells under uniform pressure*, Seisan-Kewkyu, 26 (1974), pp. 380–385.

[2] F. FREUDENSTEIN AND B. ROTH, *Numerical solution of systems of nonlinear equations*, J. Assoc. Comput. Mach., 10 (1963), pp. 550–556.

[3] C. DEN HEIJER, *The Numerical Solution of Nonlinear Operator Equations by Imbedding Methods*; *Thesis*, Mathematical Centre, Amsterdam, 1979.

[4] H. B. KELLER, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, Applications of Bifurcation Theory, P. Rabinowitz, ed., Academic Press, New York, 1977, pp. 359–384.

[5] ———, *Global homotopies and Newton methods*, Recent Advances in Numerical Analysis, C. deBoor, G. H. Golub, ed., Academic Press, New York, 1978, pp. 73–94.

[6] A. D. KERR AND M. T. SOIFER, *The linearization of the prebuckling state and its effect on the determined instability load*, Trans. ASME J. Appl. Mech., 36 (1969), pp. 775–783.
[7] S. T. MAU AND R. H. GALLAGHER, *A finite element procedure for nonlinear prebuckling and initial postbuckling analysis*, NASA Contractor Report, NASA CR-1936, January 1972.
[8] R. MENZEL AND H. SCHWETLICK, *Zur Lösung parameterabhängiger nichtlinearer Gleichungen mit singulären Jacobi-Matrizen*, Numer. Math., 30 (1978), pp. 65–79.
[9] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[10] S. J. POLAK, *et al.*, *A Continuation Method for the Calculation of Electrostatic Potentials in Semiconductors*, N. V. Philips' Gloeilampenfabrieken, Eindhoven, ISA-TIS/CARD, 1979.
[11] W. C. RHEINBOLDT, *Methods for Solving Systems of Nonlinear Equations*, CBMS Regional Conference Series in Applied Mathematics, 14, Society for Industral and Applied Mathematics, Philadelphia, 1974.
[12] ———, *Numerical continuation methods for finite element applications*, Formulations and Computational Algorithms in Finite Element Analysis, K. J. Bathe, J. T. Oden, W. Wunderlich, eds., MIT Press, Cambridge, MA, 1977, pp. 599–631.
[13] ———, *An adaptive continuation process for solving systems of nonlinear equations*, Polish Academy of Science, Banach Ctr. Publ., 3, 1977, pp. 129–142.
[14] ———, *Solution fields of nonlinear equations and continuation methods*, SIAM J. Numer. Anal., 17 (1980), pp. 221–237.
[15] ———, *Numerical analysis of continuation methods for nonlinear structural problems*, University of Pittsburgh, Inst. for Comp. Math. and Appl., Techn. Rept. ICAM-80-15, 1980.
[16] E. RIKS, *A unified method for the computation of critical equilibrium states of nonlinear elastic systems*, Techn. Rept., NLR-MP-77041U, National Aerospace Laboratory NLR, the Netherlands, 1977.
[17] H. L. SCHREYER AND E. F. MASUR, *Buckling of shallow arches*, Proc. A.S.C.E., J. of the Engineering Mechanics Div., 92, No. EM4, August 1966, pp. 1–20.
[18] J. F. TRAUB AND H. WOŹNIAKOWSKI, *Convergence and complexity of Newton iteration for operator equations*, J. Assoc. Comput. Mach., 26 (1979), pp. 250–258.
[19] H. J. WACKER, ed., *Continuation Methods*, Academic Press, New York, 1978.
[20] A. C. WALKER, *A non-linear finite element analysis of shallow circular arches*, Int. J. Solids Structures, 5 (1969), pp. 97–107.
[21] E. WASSERSTROM, *Numerical solutions by the continuation method*, SIAM Rev., 15 (1973, pp. 89–119.