BANACH CENTER PUBLICATIONS
VOLUME 3

AN ADAPTIVE CONTINUATION PROCESS FOR SOLVING SYSTEMS OF NONLINEAR EQUATIONS

WERNER C. RHEINBOLDT*

Computer Science Center, University of Maryland, College Park, Md. 20742, USA

1. Introduction

In general, iterative methods for solving a nonlinear equation in R^n depend strongly on the selection of the initial data. In order to reduce this dependence, the continuation processes use a family of equations

$$H(x, t) = 0, \quad 0 \leqslant t \leqslant 1, \quad x \in R^n \tag{1.1}$$

which for $t = 1$ contains the given equation. If for each $t \in [0, 1]$ a solution $x(t)$ of (1.1) exists that varies continuously with t, then the function $x\colon [0, 1] \to R^n$ constitutes a curve in R^n between the—assumed to be given—point $x^0 = x(0)$ and the unknown solution $x^* = x(1)$ of the original equation. Hence iterative processes may be considered which use the curve as a guide and channel their iterates in its proximity from x^0 to the intended limit x^*.

While the continuation idea itself has a long history (see, e.g., [4]), its first use as a numerical tool is often attributed to E. Lahaye ([6], [7]) and D. F. Davidenko (see, e.g., [11] for some translations and a lengthy bibliography). The Davidenko methods are based on the observation that—under suitable differentiability conditions—the unknown continuation curve is a solution of the initial value problem

$$\left[\frac{\partial}{\partial x} H(x, t)\right] \frac{dx}{dt} + \frac{\partial}{\partial t} H(x, t) = 0, \quad 0 \leqslant t \leqslant 1, \quad x(0) = x^0. \tag{1.2}$$

Accordingly, we may approximate this curve by applying some discrete variable method to (1.2). On the other hand, Lahaye's iterative continuation approach uses a locally convergent iterative method for solving (1.1) at an increasing sequence of discrete parameter values $t_0 = 0 < t_1 < \ldots < t_m = 1$. A step $t_{k+1} - t_k$ may

* This work was in part supported by the National Science Foundation under Grant GJ-35568X.

be chosen such that, for instance, the last iterate at t_k is a permissible initial approximation for the iteration at t_{k+1}. More generally, extrapolation may be applied to obtain the next starting point.

In considering these approaches we distinguish two objectives: (a) The end point of the continuation curve is to be found as quickly as possible while the curve itself is of lesser interest; (b) the entire curve is of interest and is to be approximated closely.

By nature, numerical methods for solving initial value problems—such as (1.2)—are designed to approximate the entire solution curve within a given potentially small, error tolerance. This points to the Davidenko approach if (b) is our aim and suggests also that in connection with (a) the approach may waste effort since more information is obtained than required.

In iterative continuation, the main delimiter of the t-steps is the dependence of the local method on the initial data; that is, larger steps are possible when the convergence domains increase. Hence, iterative continuation might well be a suitable choice for (a) provided only that a step-algorithm is used which adjusts to the convergence behavior of the local iterations. Unfortunately, such algorithms do not appear to be available. In fact, typically the steps are simply kept small enough to ensure once again a close approximation of the entire curve; see, e.g., [5].

In this paper we present an adaptive step-algorithm for iterative continuation in the case of objective (a). It is designed to adjust the t-steps according to estimates of the local convergence domains and hence allows the process to move more rapidly along the curve whenever the local method is less sensitive to the choice of the initial values. Since effective, computable estimates for convergence radii are rarely available, it is natural that here heuristic procedures play an important role. As in various computer science problems, such heuristic techniques are proving to be increasingly valuable tools in the design of complex numerical processes.

In Section 2 below the general process is described; then in Section 3 estimates for convergence radii of several methods are discussed; and finally Section 4 contains results of a number of numerical experiments. On the basis of our experience the process performs very well in a wide variety of applications. In this connection, my sincere thanks are given to Mr. Charles K. Mesztenyi for the implementation of the procedures and for all his help in the computational experiments.

2. Description of the process

Let $H\colon D\times[0,1]\in R^{n+1}\to R^n$ be a continuous mapping with the property that for each $t\in[0,1]$ the equation (1.1) has a solution and that the curve

$$x\colon [0,1]\subset D\to R^n \tag{2.1}$$

has the known starting point $x(0)=x^0$ and is at least q ($\geqslant 0$) times continuously differentiable. We apply some iterative continuation process and assume that it has progressed through the parameter values $0=t_0<t_1<\ldots<t_k<1$, $k\geqslant 0$.

For $k\geqslant 1$, let $x^1,\ldots,x^k$ denote the terminal iterates of the local method at these t-values. In preparation for the prediction of the next step, we construct, with some as yet unspecified $p=p_k\in[0,q)$, the interpolation polynomial

$$L(k,p;t)=\sum_{j=0}^{p}\varphi_j(k,p;t)x^{k-j},\quad k\geqslant p\geqslant 0, \tag{2.2}$$

$$\varphi_j(k,p;t)=\prod_{\substack{l=0\\ l\neq j}}^{p}\frac{t-t_{k-l}}{t_{k-j}-t_{k-l}},\quad \varphi_0(k,0;t)\equiv 1,\quad j=0,\ldots,p,$$

for which

$$L(k,p;t_i)=x^i,\quad i=k,k-1,\ldots,k-p. \tag{2.3}$$

Then the prediction error may be estimated as

$$\|x(t)-L(k,p;t)\|\leqslant\|L(k,p+1;t)-L(k,p;t)\|+O(h_{\max}^{p+1}) \tag{2.4}$$

where $h_{\max}$ denotes the maximal step taken so far. This type of estimate is frequently used in the design of discrete variable methods for initial value problems. In [2] a formula is given for the difference

$$E(k,p;t)=L(k,p+1;t)-L(k,p;t),\quad k\geqslant p+1.$$

However, this formula appears to be incomplete, and we give here a new derivation. From

$$\delta_j=\varphi_j(k,p+1;t)-\varphi_j(k,p;t)=\Big[\prod_{l=0}^{p}(t-t_{k-l})\Big]\Big/\Big[\prod_{\substack{l=0\\ l\neq j}}^{p+1}(t_{k-j}-t_{k-l})\Big],\quad j=0,1,\ldots,p,$$

it easily follows that

$$\delta_0=\frac{t-t_k}{t_k-t_{k-p-1}}\prod_{l=1}^{p}\frac{t-t_{k-l}}{t_k-t_{k-l}}=\frac{t-t_k}{t_k-t_{k-p-1}}\varphi_0(k,p;t)$$

and

$$\delta_j=-\frac{t-t_k}{t_k-t_{k-p-1}}\prod_{l=1}^{p}\frac{t-t_{k-l}}{t_k-t_{k-l}}\prod_{\substack{l=1\\ l\neq j}}^{p+1}\frac{t_k-t_{k-l}}{t_{k-j}-t_{k-l}}$$

$$=-\frac{t-t_k}{t_k-t_{k-p-1}}\varphi_0(k,p;t)\varphi_{j-1}(k-1,p;t_k),\quad j=1,\ldots,p,$$

as well as

$$\varphi_{p+1}(k,p+1;t)=-\frac{t-t_k}{t_k-t_{k-p-1}}\varphi_0(k,p;t)\varphi_p(k-1,p;t_k).$$

Hence we have

$$E(k,p;t)=\sum_{j=0}^{p}\delta_j x^{k-j}+\varphi_{p+1}(k,p+1;t)=$$

$$= \frac{t-t_k}{t_k-t_{k-p-1}}\varphi_0(k,p;t)\Big[x^k-\sum_{j=1}^{p+1}\varphi_{j-1}(k-1,p;t_k)x^{k-j}\Big];$$

that is,

$$L(k,p+1;t)-L(k,p;t) = \frac{t-t_k}{t_k-t_{k-p-1}}\varphi_0(k,p;t)[x^k-L(k-1,p;t_k)]. \tag{2.5}$$

The φ_0-term on the right is missing in the formula given in [2].

Assume now that the local method used for solving the equation (1.1) with fixed t has a convergence ball

$$\bar{B}\big(x(t),r(t)\big) = \{x\in R^n \mid \|x-x(t)\| \leqslant r(t)\}, \quad r(t)>0. \tag{2.6}$$

In other words, for any starting point in this ball, the method generates a well-defined sequence of iterates which converges to $x(t)$. Moreover, suppose that $\hat{r}_k(t)$ is an approximation of $r(t)$ for small $t-t_k \geqslant 0$. For example, if values $r_j \doteq r(t_j)$, $j=0,1,\ldots,k$, are predicted during the local iterations, polynomial extrapolation could be used to obtain $\hat{r}_k(t)$. Since, in practice, the radii r_j are relatively coarse estimates of the $r(t_j)$, it is advisable to apply only low order interpolation formulas or to use numerical approximation. The simplest choice for $\hat{r}_k(t)$, is, of course, the constant r_k.

For continuing along the curve (2.1) we wish to choose some step $h_k = t_{k+1}-t_k$ and apply the local process to the equation (1.1) with $t = t_{k+1}$ starting from

$$x^{k+1,0} = L(k,p;t_{k+1}). \tag{2.7}$$

In [5], h_k is chosen such that $\|x^{k+1,0}-x(t_{k+1})\|$ is (asymptotically) bounded by a given, potentially small, error tolerance. This provides for a close approximation of the entire continuation curve in line with objective (b) of the Introduction.

Here, in line with objective (a), we choose h_k subject only to the condition that $x^{k+1,0}$ is in the predicted convergence ball $\bar{B}\big(x(t_{k+1}),\hat{r}_k(t_{k+1})\big)$. The (asymptotic) estimate (2.4)–(2.5) of $\|x^{k+1,0}-x(t_{k+1})\|$ then leads to the equation

$$\theta\hat{r}_k(t_k+h) = \begin{cases} \dfrac{h}{t_k-t_{k-1}}\|x^k-x^{k-1}\| & \text{for } p=0, \\ \dfrac{h}{t_k-t_{k-p-1}}\displaystyle\prod_{l=1}^{p}\Big(1+\frac{h}{t_k-t_{k-l}}\Big)\|x^k-L(k-1,p;t_k)\| & \text{for } p\geqslant 1, \end{cases} \tag{2.8}$$

for $h=h_k$, with some factor $\theta>0$ reflecting the uncertainty in the various estimates.

For constant $\hat{r}_k(t)\equiv r_k>0$, (2.8) is a polynomial in h which by Descartes' rule of signs has exactly one positive solution. In practice, $\hat{r}_k(t_k+h)$ is a low-order polynomial in h which is positive for $h=0$ and does not vary too much near zero. Then, in general, there is still only one positive root. For instance, this is certainly true if

$$\hat{r}_k(t_k+h) = r_k-\varrho_k h, \quad r_k\geqslant 0, \quad \varrho_k > -\frac{\|x^k-L(k-1,p;t_k)\|}{\theta(t_k-t_{k-p-1})}.$$

In all cases, the desired positive solution $h=h_k>0$ of (2.8) may be found iteratively starting from the previous step h_{k-1}.

For $p=p_{k-1}$ we see from (2.7) that the norm term in (2.8) is exactly the distance between the starting point $x^{k,0}$ and the terminal point x^k of the local iteration at t_k. More generally, the—assumed to be unique—positive solution of (2.8) depends on p, and we may choose $p=p_k$ such that h is maximal. In general, it is inadvisable to change p at each step. Moreover if there is to be a change, only the values of h corresponding to $p=p_{k-1}-\delta$, $\delta=-1,0,+1$, should be tested.

It remains to be discussed how the convergence radii may be estimated for various local methods; this will be the topic of the next section. Here we summarize the general procedure as an informal program:

1. $k:=k_0:=p_k:=0$, $t_k:=0$ [Initialization]
2. If $(k=0)$ then $h:=h_{\min}$, go to 10
3. Solve (2.8) for $h=h(p_k)$ [Normal h prediction]
4. If $(k-k_0<p_k)$ then go to 9 [Test for p-update]
5. $k_0:=k$
6. Solve (2.8) for $h=h(p_k-1)$ and $h=h(p_k+1)$
7. $h:=\max\{h(p);\ p=p_k-1,p_k,p_k+1\}$
8. $p_k:=p$-value at maximum in step 7 [Change p]
9. $h_k:=\max\big(h_{\min},\min(h_{\max},h)\big)$ [Next step h]
10. $t_{k+1}:=t_k+h_k$ [Next t-value]
11. $x^{k+1,0}:=L(k,p_k,t_{k+1})$ [Initial iterate]
12. Apply local iteration to (1.1) with $t=t_{k+1}$ starting from $x^{k+1,0}$ and determine estimated convergence radius r_{k+1}. [Local method]
    Normal return: Convergence criterion is met for final iterate $x^{k+1}:=x^{k+1,j}$, go to 14
    Error return: No satisfactory convergence, go to 13
13. if $(h_k=h_{\min})$ then error stop [No convergence]
    else $h:=\vartheta h_k$, go to 9 [Reduce step]
14. From local radii determine $\hat{r}_{k+1}(t)$
15. If $t_{k+1}<1$ then $k:=k+1$, go to 3 [Test for $t=1$]

The parameters $h_{\max}>h_{\min}>0$, $\vartheta\in(0,1)$ and $\theta\in(0,1]$ in (2.8) are given. Step 4 is an update criterion for p of a type frequently used with initial value problems. The convergence criterion and the error return in Step 11 vary with the particular method. However, the popular convergence test

$$\|y^j-y^{j-1}\| \leqslant \varepsilon_1\|y^{j-1}-y^{j-2}\|+\varepsilon_2, \quad \varepsilon_1>0, \quad \varepsilon_2>0,$$

for the sequence $y^j=x^{k+1,j}$ is often satisfactory.

3. Estimation of attraction radii

Let $\mathcal{J}$ denote the locally convergent process used in approximating the solution $x^* = x(t)$ of the equation

$$Fx \equiv H(x,t) = 0 \tag{3.1}$$

for some fixed t. We have to determine $r = r(t) > 0$ such that (2.6) is a convergence ball of $\mathcal{J}$. Only few *a posteriori* estimates for such radii are known (see, e.g., [9], [10]) and as mentioned before we have to rely heavily on heuristic approaches.

Suppose that $\mathcal{J}$ has the form

$$x^{j+1} = Gx^j, \quad j = 0, 1, \ldots \tag{3.2}$$

where x^* is a fixed point of G: $D_G \subset R^n \to R^n$ and that

$$\|Gx - x^*\| \leqslant \alpha\|x - x^*\|, \quad \forall x \in U, \quad \alpha < 1 \tag{3.3}$$

holds in some open neighborhood $U \subset D$ of x^*. Then any

$$\bar{B}(x^*, r) \subset U, \quad r > 0 \tag{3.4}$$

is an attraction ball of $\mathcal{J}$, and it follows by the triangle inequality that

$$r \geqslant \|x - x^*\| \geqslant \frac{1}{1+\alpha}\|Gx - x\| \geqslant \frac{1}{2}\|Gx - x\|, \quad \forall x \in \bar{B}(x^*, r). \tag{3.5}$$

(See also [9].) Under the heuristic assumption that the initial iterate x^0 of (3.2) is in the ball (3.4), the inequalities (3.5) provide simple lower bounds for r. Estimates for α require, of course, additional assumptions about G. As a simple example, suppose that (3.3) is strengthened to

$$\|Gy - Gx\| \leqslant \alpha\|y - x\|, \quad \forall x, y \in U, \quad \alpha < 1. \tag{3.6}$$

Then

$$q(y) = \frac{\|G^2y - Gy\|}{\|Gy - y\|} \leqslant \alpha, \quad \forall y \in U, \quad y \neq x^*, \tag{3.7}$$

together with (3.5) provides the bound

$$r \geqslant \max\left(\frac{1-q(y)}{1+q(y)}, \frac{1}{2}\right)\|Gx - x\|, \quad \forall x, y \in U, \quad y \neq x^*. \tag{3.8}$$

In practice, we may choose y as the iterate for which q is maximal.

As an application of these estimates let $\mathcal{J}$ be the modified Newton method for (3.1). For this assume that F: $D \subset R^n \to R^n$ is continuously differentiable and

$$\|F'(y) - F'(x)\| \leqslant \gamma\|y - x\|, \quad \forall x, y \in D. \tag{3.9}$$

For given $x^0 \in D$ such that $F'(x^0)$ is nonsingular, the iteration function now has the form

$$Gx = x - F'(x^0)^{-1}Fx, \quad \forall x \in D. \tag{3.10}$$

Hence from

$$Gy - Gx = -F'(x^0)^{-1}\int_0^1 [F'(x + t(y-x)) - F'(x^0)](y-x)\,dt$$

it follows that

$$\|Gy - Gx\| \leqslant \beta\gamma\int_0^1 \|(1-t)(x-x^0) + t(y-x^0)\|\|y-x\|\,dt$$

$$\leqslant \frac{1}{2}\beta\gamma[\|x-x^0\| + \|y-x^0\|]\|y-x\|, \tag{3.11}$$

where $\|F'(x^0)^{-1}\| \leqslant \beta$. Therefore, for $x^0 \in \bar{B}(x^*, r) \subset D$ we require $2\beta\gamma r < 1$ to ensure the contractivity condition (3.6) and hence the applicability of the bound (3.8). On the other hand, (3.3) holds for $\frac{3}{2}\beta\gamma r < 1$ and then (3.5) leads to the inequality

$$\tfrac{3}{2}\beta\gamma r^2 + r \geqslant \|Gx - x\|,$$

whence

$$r \geqslant \frac{2}{1 + \sqrt{1 + 6\beta\gamma\|Gx - x\|}}\|Gx - x\|, \quad \forall x \in \bar{B}(x^*, r). \tag{3.12}$$

From

$$\|FGx\| = \|FGx - Fx - F'(x^0)^{-1}(Gx - x)\| \leqslant \tfrac{1}{2}\gamma[\|Gx - x^0\| + \|x - x^0\|]\|Gx - x\|,$$

it follows that

$$2\frac{\|FGx\|}{\|Fx\|}\frac{1}{\|Gx - x^0\| + \|x - x^0\|} \leqslant \beta\gamma, \quad \forall x \in D. \tag{3.13}$$

Thus with (3.5), we obtain the necessary condition for x to be in the attraction ball:

$$\frac{\|FGx\|}{\|Fx\|}\frac{\|Gx - x\|}{\|Gx - x^0\| + \|x - x^0\|} < \frac{2}{3}, \quad \forall x \in \bar{B}(x^*, r). \tag{3.14}$$

The lower bound in (3.13)—multiplied by a suitable factor—often represents a good heuristic value for $\beta\gamma$ in (3.12).

If Newton's method itself is used, then the following local convergence result is valid.

THEOREM 3.1. *Let* (3.9) *hold for the continuously differentiable mapping* F: $D \subset R^n \to R^n$. *If* $x^* \in D$ *is a solution of* $Fx = 0$ *for which* $F'(x^*)$ *is nonsingular then any*

$$B_r \equiv \bar{B}(x^*, r) \subset D, \quad r < (2/3)(\beta\gamma)^{-1}, \quad \beta = \|F'(x^*)^{-1}\| \tag{3.15}$$

is an attraction ball of Newton's method.

Proof. By the standard perturbation lemma it follows that $F'(x)$ is nonsingular in B_r and

$$\|F'(x)^{-1}\| \leqslant \frac{\beta}{1 - \beta\gamma\|x - x^*\|} < 3\beta, \quad \forall x \in B_r. \tag{3.16}$$

Hence the Newton iteration function

$$Gx = x - F'(x)^{-1}Fx, \quad \forall x \in B_r$$

is well-defined on B_r and from

$$\|Gx-x^*\| \leqslant \|F'(x)^{-1}\|\|Fx^*-Fx-F'(x)(x^*-x)\| \tag{3.17}$$
$$\leqslant \frac{\beta}{1-\beta\gamma\|x-x^*\|}\frac{1}{2}\gamma\|x-x^*\|^2 \leqslant \alpha\|x-x^*\|, \quad \forall x \in B_r$$

and

$$\alpha = \frac{1}{2}\frac{\sigma}{1-\sigma} < 1, \quad \sigma = \beta\gamma r < \frac{2}{3}, \tag{3.18}$$

we obtain the result.

It may be noted that in [12] a corresponding result with the smaller radius $r < (1-1/\sqrt{2})(\beta\gamma)^{-1}$ was proved.

Because of (3.17) the simple bounds (3.5) apply and take here the form

$$r \geqslant 2\frac{1-\sigma}{2-\sigma}\|x-Gx\| \geqslant \frac{1}{2}\|x-Gx\|, \quad \forall x \in B_r. \tag{3.19}$$

In order to improve this we introduce the quantity

$$\tau(x) = \gamma\|F'(x)^{-1}\|\|x-Gx\|, \quad \forall x \in B_r, \tag{3.20}$$

which plays a central role in Kantorovich's convergence proof of Newton's method (see, e.g., [8]). From

$$\|x-Gx\| \leqslant \|x-x^*\|+\|F'(x)^{-1}\|\|Fx^*-Fx-F'(x)(x^*-x)\|$$
$$\leqslant \|x-x^*\|+\tfrac{1}{2}\gamma\|F'(x)^{-1}\|\|x-x^*\|^2, \quad \forall x \in B_r$$

it follows then that

$$r \geqslant \|x-x^*\| \geqslant \frac{2}{1+\sqrt{1+2\tau(x)}}\|x-Gx\|, \quad \forall x \in B_r. \tag{3.21}$$

The analogous estimate

$$\|x-x^*\| \leqslant \|x-Gx\|+\|Gx-x^*\| \leqslant \|x-Gx\|+\tfrac{1}{2}\gamma\|F'(x)^{-1}\|\|x-x^*\|^2, \quad \forall x \in B_r$$

leads to the error bound of the Kantorovich theorem

$$\|x-x^*\| \leqslant \frac{2}{1+\sqrt{1-2\tau(x)}}\|x-Gx\|, \quad \forall x \in B_r, \quad 0 \leqslant \tau(x) \leqslant \frac{1}{2}. \tag{3.22}$$

In order to obtain bounds for r itself, observe that

$$|\|F'(x)^{-1}\|-\|F'(y)^{-1}\|| \leqslant \|F'(x)^{-1}-F'(y)^{-1}\|$$
$$\leqslant \gamma\|F'(y)^{-1}\|\|F'(x)^{-1}\|\|x-y\|,$$

that is,

$$\left|\frac{1}{\gamma\|F'(x)^{-1}\|}-\frac{1}{\gamma\|F'(y)^{-1}\|}\right| \leqslant \|x-y\|, \quad \forall x, y \in B_r, \tag{3.23}$$

and, in particular,

$$\left|\sigma\frac{\|x-Gx\|}{\tau(x)}-r\right| \leqslant \sigma\|x-x^*\|, \quad \forall x \in B_r, \quad x \neq x^*. \tag{3.24}$$

This provides the bounds

$$\frac{\sigma}{1+\sigma}\frac{1}{\tau(x)}\|x-Gx\| \leqslant r \leqslant \frac{\sigma}{1-\sigma}\frac{1}{\tau(x)}\|x-Gx\|, \quad \forall x \in B_r, \quad x \neq x^* \tag{3.25}$$

and alternately, if (3.23) is used,

$$\sigma\frac{\sqrt{1-2\tau(x)}}{\tau(x)}\|x-Gx\| \leqslant r \leqslant \sigma\frac{2-\sqrt{1-2\tau(x)}}{\tau(x)}\|x-Gx\|, \tag{3.26}$$
$$\forall x \in B_r, \quad 0 < \tau(x) \leqslant \tfrac{1}{2}.$$

From

$$\|FGx\| = \|FGx-Fx-F'(x)(Gx-x)\| \leqslant \tfrac{1}{2}\gamma\|Gx-x\|^2 \leqslant \tfrac{1}{2}\tau(x)\|Fx\| \tag{3.27}$$

and (3.16) we obtain

$$\|Gx-G^2x\| = \|F'(Gx)^{-1}FGx\| \leqslant \frac{1}{2}\frac{\beta\gamma\|x-Gx^*\|}{1-\beta\gamma\|Gx-x^*\|}\|x-Gx\|, \quad \forall x \in B_r \tag{3.28}$$

and hence the additional upper bound

$$r \leqslant \frac{1}{2}\frac{\sigma}{1-\sigma}\frac{1}{q(x)}\|x-Gx\| \leqslant \frac{1}{q(x)}\|x-Gx\|, \quad \forall x \in B_r \tag{3.29}$$

with q given by (3.7). A comparison of (3.29) and (3.19) shows that

$$\frac{1}{2}\frac{\sigma}{1-\sigma}\frac{1}{q(x)} \geqslant 2\frac{1-\sigma}{2-\sigma},$$

that is,

$$\sigma \geqslant 1-\big(1+4q(x)\big)^{-1/2}, \quad \forall x \in B_r \tag{3.30}$$

and thus $q(x) < 2$. A similar comparison between (3.21) and (3.25) gives

$$\sigma \geqslant 1-\big(1+2\tau(x)\big)^{-1/2}.$$

Since by (3.27) we have

$$\tau(x) \geqslant 2\frac{\|FGx\|}{\|Fx\|} \equiv \bar{\tau}(x), \quad \forall x \in B_r, \tag{3.31}$$

this implies that $\|FGx\|/\|Fx\| < 2$. These estimates for q and the ratio of the function norms indicate that B_r may have to be restricted in order to avoid slow convergence. For example, we might choose $\sigma = 1-1/\sqrt{5}$ in which case $\tau(x) \leqslant 2$, $q(x) \leqslant 1$ and $\|FGx\| \leqslant \|Fx\|$ for all x in B_r.

The upper estimates in (3.25) and (3.26) remain valid if the lower bound $\bar{\tau}(x)$ in (3.31) for $\tau(x)$ is used. For the lower estimates of r we need an upper bound of

$\tau(x)$. In a variety of problems such bounds can be found. On the other hand, practical experience has shown that the use of $\bar{\tau}(x)$ instead of $\tau(x)$ in the lower estimates of (3.25) and (3.26) provides, in general, good and typically rather conservative estimates of r. The reason for this rests with the fact that obviously $\tau(x)-\bar{\tau}(x) = O(\|x-x^*\|)$, and by (3.24), $|\sigma\|x-Gx\|/\tau(x)-r| = O(\|x-x^*\|)$ as $x \to x^*$. It may also be noted that for quadratic functions in R^1 we always have $\tau(x) = \bar{\tau}(x)$. In view of the conservative nature of the estimates it was usually found advantageous to "overrelax" by working with some linear combination of the lower and upper bounds for r.

4. Some numerical experiments

The adaptive continuation process of the previous sections has been applied to a variety of problems. We present here a few selected results. All computations were performed on a Univac 1108 system in single precision.

As a first example we consider the problem

$$Fx = \begin{pmatrix} \frac{1}{2}\left[\sin(x_1 x_2) - \frac{1}{2\pi}x_2 - x_1\right] \\ \left(1-\frac{1}{4\pi}\right)(e^{2x_1}-e) + \frac{e}{\pi}x_2 - 2ex_1 \end{pmatrix}, \tag{4.1}$$

$$H(x,t) = Fx-(1-t)Fx^0, \quad x^0 = (0.3, 4)^T$$

found in [3]. The solution curve x: $[0, 1] \to R^2$ of $H(x, t) = 0$ terminates at the solution $x^* = x(1) \doteq (0.299449, 2.83693)^T$ of $Fx = 0$. This equation also has the solution $(0.5, \pi)^T$ and, moreover, Newton's method starting at x^0 converges to the further solution $(-0.26, 0.62)^T$.

The results with the continuation process are given in Table 1 below (rounded to four digits):

Table 1

Function (4.1), Local Newton Process, $0 \leqslant t \leqslant 1$

Step		Solution		Predicted Radii		Newton	Degree
t_k	h_k	x_1^k	x_2^k	r_{low}^k	r_{upper}^k	steps	p_k
.01250	.01168	.3933	2.990	.009432	.009432	5	1
.02418	.02165	.3893	2.984	.005880	.01308	4	1
.04584	.04216	.3837	2.975	.01147	.01418	3	2
.08799	.08158	.3755	2.963	.01767	.02184	3	2
.1696	.1595	.3637	2.944	.02620	.03239	3	2
.3290	.3133	.3468	2.917	.03786	.04680	3	2
.6423	.3577*)	.3220	2.876	.05336	.06596	3	2
1.0000	—	.2994	2.837	—	—	3	2

*) step adjusted to reach $t = 1$

In comparison Boggs [1] used a Davidenko technique with Euler's method as predictor and the trapezoidal rule as corrector. His basic method required 29 steps and an average of 6 evaluations of the Jacobian per step. In view of the objective (b) pursued there, this is as expected.

As a second problem, we consider a quasi-linear equation of the form

$$H(x, t) \equiv A(x)x - tb = 0, \quad 0 \leqslant t \leqslant 1, \tag{4.2}$$

where A: $D \subset R^n \to L(R^n)$ is a given matrix-valued mapping. Equations of this type occur naturally as discrete analogs of quasi-linear elliptic boundary value problems. They also arise, for instance, in the finite displacement analysis of mechanical structures.

A linearly convergent process for solving (4.1) with fixed t is given by

$$A(x^j)x^{j+1} = tb. \tag{4.3}$$

An attraction result for this iteration may be phrased as follows:

THEOREM 4.1. *Suppose that* A: $D \subset R^n \to L(R^n)$ *satisfies*

$$\|A(x)-A(y)\| \leqslant \beta\|x-y\|, \quad \forall x, y \in D,$$

and that $x^* \in \operatorname{int}(D)$ *is a solution of* (4.2) *(for fixed* t*) such that* $A(x^*)$ *is nonsingular and*

$$\beta\eta^2\|tb\| < 1, \quad \eta = \|A(x^*)^{-1}\|. \tag{4.4}$$

Then any

$$\bar{B}(x^*, r) \subset D, \quad r < \frac{1-\beta\eta^2\|tb\|}{\beta\eta} \tag{4.5}$$

is an attraction ball of (4.2).

Proof. From (4.4) it follows that $A(x)$ is nonsingular and

$$\|A(x)^{-1}\| \leqslant \frac{\eta}{1-\beta\eta r}, \quad \forall x \in \bar{B}(x^*, r).$$

Hence $Gx = tA(x)^{-1}b$ is well-defined on the ball and the result follows from

$$\|Gx-x^*\| = \|A(x^*)^{-1}\big(A(x^*)-A(x)\big)A(x)^{-1}tb\| \tag{4.6}$$

$$\leqslant \frac{\beta\eta^2\|tb\|}{1-\beta\eta r}\|x-x^*\|, \quad \forall x \in \bar{B}(x^*, r)$$

and $\alpha = \beta\eta^2\|tb\|/(1-\beta\eta r) < 1$.

The result indicates that for increasing t the radius r should decrease while the convergence constant increases. This is a widely known phenomenon in connection with equations (4.2) arising in structural mechanics.

As an example, we consider the torsion of an infinite rod with square cross-section described by the boundary value problem

$$\begin{cases} \frac{\partial}{\partial\xi}\left[q(u_\xi^2+u_\eta^2)u_\xi\right] + \frac{\partial}{\partial\eta}\left[q(u_\xi^2+u_\eta^2)u_\eta\right] = w, & (\xi, \eta) \in \Omega = [0, 1]\times[0, 1], \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{4.7}$$

We assume w to be constant. With a uniform subdivision of Ω of meshlength $h = 1/(m+1)$ and with the standard piecewise linear hill functions as trial functions, the component equations of (4.2) in this case have the form

$$(4.8a)\quad (Q_N+Q_W+Q_S+Q_E)x_{ij}-Q_N x_{i,j+1}-Q_W x_{i-1,j}-Q_S x_{i,j-1}-Q_E x_{i+1,j} = h^2 w,$$
$$i,j = 1, 2, \ldots, m,\ x_{ik} = 0 \text{ for } (ih, kh) \in \partial\Omega$$

where

$$(4.8b)\quad \begin{aligned} 2Q_N &= q\big(\Delta_\xi(i,j+1)^2+\Delta_\eta(i,j)^2\big)+q\big(\Delta_\xi(i-1,j)^2+\Delta_\eta(i,j)^2\big),\\ 2Q_W &= q\big(\Delta_\xi(i-1,j)^2+\Delta_\eta(i,j)^2\big)+q\big(\Delta_\xi(i-1,j)^2+\Delta_\eta(i-1,j-1)^2\big),\\ 2Q_S &= q\big(\Delta_\xi(i-1,j-1)^2+\Delta_\eta(i,j-1)^2\big)+q\big(\Delta_\xi(i,j)^2+\Delta_\eta(i,i-1)^2\big),\\ 2Q_E &= q\big(\Delta_\xi(i,j)^2+\Delta_\eta(i,j-1)^2\big)+q\big(\Delta_\xi(i,j)^2+\Delta_\eta(i+1,j)^2\big) \end{aligned}$$

and

$$(4.8c)\quad \Delta_\xi(l,k) = \frac{1}{h}(x_{l+1,k}-x_{lk}),\qquad \Delta_\eta(l,k) = \frac{1}{h}(x_{l,k+1}-x_{lk}).$$

Table 2 below gives some results for this function and the linear process (4.3):

Table 2

$q(s) = e^{5s}$, $w = 5$, $n = 49$, Linear Process (4.2)

t_k	$t_{k+1}-t_k$	No. of local steps	Observed α_k	Pred. Radius	Degree p_k
.1000	.0701	6	.115	.364(−1)	0
.1701	.1042	10	.325	.245(−1)	1
.2743	.0323	20	.628	.867(−2)	1
.3066	.0177	24	.718	.178(−2)	1
.3243	.0252	23	.767	.304(−3)	2
.3495	.0100*	23	.835	.360(−4)	2
.3595	.0100*	15	.861	.604(−5)	2
.3695	.0100*	13	.886	.316(−5)	2
.3795	.0101	8	.914	.162(−5)	3
.3896	.0100	31	.935	.330(−5)	3
.3996	—	66	.989	.592(−5)	3
		239			

* min step = .1

As expected by Theorem 4.1, we have approximately a linearly increasing convergence constant α.

The situation does not arise with Newton's method. In fact, for the same function q and parameters w, n we then reach $t = 1$ in four t-steps with a total of 20 local Newton steps.

As a more stringent test for the behavior of Newton's method as local process, the following smoothed step-function q was selected, representing a model for some elastic-plastic material:

$$(4.9)\quad q(s) = \begin{cases} q_0 & \text{for } s \leqslant 0.15,\\ \frac{1}{2}(q_0+q_1)+\frac{1}{4}(q_1-q_0)(3\bar{s}-\bar{s}^3) & \text{for } 0.15 \leqslant s \leqslant 0.5,\ \bar{s} = \dfrac{40s-13}{7},\\ q_1 & \text{for } s \geqslant 0.5. \end{cases}$$

A typical run with Newton's method as local process is given in Table 3.

Table 3

q of (4.9), $q_0 = 1$, $q_1 = 25$, $w = 10$, $n = 121$, $0 \leqslant t \leqslant 1.5$

t_k	Δt_k	No. of local steps	Pred. Radius	
.04	.4(−1)	2	—	No predictions made
.08	.48(−1)	2	—	Essentially linear range
.128	.576(−1)	2	—	
.1856	.543(−1)	7	.115(−1)	
.2399	.293(−1)	6	.472(−2)	
.2692	.398(−1)	5	.382(−2)	
.3090	.429(−1)	4	.214(−2)	
.3519	.524(−1)	5	.335(−2)	
.4044	.603(−1)	6	.375(−2)	
.4647	.737(−1)	6	.439(−2)	
(.5384	—	2	—)	Newton failure, Δt = .589(−1)
.5236	.581(−1)	4	.108(−2)	
.5818	.643(−1)	4	.129(−2)	
.6461	.1343	3	.204(−2)	
(.7804	—	3	—)	Newton failure, Δt = .1075
.7535	.1786	6	.310(−2)	
.9321	.1841	4	.320(−2)	
1.116	.2509	4	.304(−2)	
1.367	.1329	3	.268(−2)	Last step adjusted to t_{max}
1.500	—	4	—	

It turns out that for fixed q the total number of t-steps and local Newton steps is essentially dimension-independent. However, in our case if, say q_1-q_0 in (4.9) increases, then the conditioning of $F'(x^*)$ worsens; and we should expect a slowdown of the overall process. This was born out in all experiments. Some results along this line may be found in [13].

References

[1] P. T. Boggs, *The solution of nonlinear systems of equations by A-stable integration techniques*, SIAM J. Num. Anal. 8 (1971), pp. 767–785.

[2] R. K. Brayton, F. G. Gustavson, and G. D. Hachtel, *A new efficient algorithm*

for solving differential-algebraic systems using implicit backward differentiation formulas, Proc. of the IEEE 60 (1972), pp. 98–108.

[3] C. G. Broyden, *A new method of solving nonlinear simultaneous equations*, Comput. J. 12 (1969), pp. 94–99.

[4] F. Ficken, *The continuation method for functional equations*, Comm. Pure Appl. Math. 4 (1951), pp. 435–456.

[5] G. Hachtel and M. Mack, *A pseudo dynamic method for solving nonlinear algebraic equations*, in R. A. Willoughby (Ed.), *Stiff differential systems*, Plenum Press, New York 1974, pp. 135–150.

[6] E. Lahaye, *Une méthode de résolution d'une catégorie d'équations transcendantes*, C.R. Acad. Sci. Paris 198 (1934), pp. 1840–1842.

[7] —, *Solution of systems of transcendental equations*, Acad. Roy. Belg. Bull. Cl. Sci. 5 (1948), pp. 805–822.

[8] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York 1970.

[9] A. M. Ostrowski, *On error estimates a posteriori in iterative procedures*, in *Spline functions and approximation theory*, A. Meir, A. Sharma (Eds.), Birkhauser Verlag, Basel 1973.

[10] —, *A posteriori error estimates in iterative procedures*, SIAM J. Num. Anal. 10 (1973), pp. 290–298.

[11] L. B. Rall, *Davidenko's method for the solution of nonlinear operator equations*, The University of Wisconsin, Mathematics Research Center, MRC Tech. Summary Rept. 948, October 1968.

[12] —, *A note on the convergence of Newton's method*, SIAM J. Num. Anal. 11 (1974), pp. 34–36.

[13] W. C. Rheinboldt, *On the solution of some nonlinear equations arising in the application of finite element methods*, in J. Whiteman (Ed.), *Mathematics of Finite Elements and Applications*, Academic Press, London 1976, pp. 465–482.

Presented to the Semester
Mathematical Models and Numerical Methods
(February 3–June 14, 1975)

BANACH CENTER PUBLICATIONS
VOLUME 3

A FINITE ELEMENT METHOD FOR A TWO POINT BOUNDARY VALUE PROBLEM WITH A SMALL PARAMETER AFFECTING THE HIGHEST DERIVATIVE

JOHN J. H. MILLER

School of Mathematics, Trinity College, Dublin 2, Ireland

We consider the following two point boundary value problem on the open interval $\Omega =]0, 1[$:

(1) Given $f_0 \in L^2(\Omega)$ find $u \in H^2(\Omega)$ such that

$$-\varepsilon u'' + a_1 u' + a_0 u = f_0 \quad \text{in } \Omega,$$
$$u(0) = u(1) = 0.$$

Here the parameter ε is assumed to satisfy $0 < \varepsilon \ll 1$.

In the interests of clarity we restrict our attention in what follows to the (trivial) case where $a_0 \geqslant 0$ and $a_1 > 0$ are constants. However the ideas may be extended without difficulty to the (non-trivial) variable coefficient case.

It is known that under the above assumptions as $\varepsilon \to 0$ the solution of (1) converges weakly in $L^2(\Omega)$ to the solution of the initial value problem.

(2) Given $f_0 \in L^2(\Omega)$ find $u \in H^1(\Omega)$ such that

$$a_1 u' + a_0 u = f_0 \quad \text{in } \Omega,$$
$$u(0) = 0.$$

We put $V = H_0^1(\Omega)$ and we define the continuous bilinear and linear forms

$$a(v, w) = \int_\Omega (\varepsilon v'w' + a_1 v'w + a_0 vw) \qquad \forall v, w \in V,$$

$$f(v) = \int_\Omega f_0 v \qquad \forall v \in V.$$

The variational formulation of (1) is then:

(3) Find $u \in V$ such that

$$a(u, v) = f(v) \qquad \forall v \in V.$$