Technical Report TR-142
GJ-231 and NGL-21-002-008

January 1971

CONTINUATION METHODS FOR NONLINEAR EQUATIONS

by

John H. Avila

UNIVERSITY OF MARYLAND
COMPUTER SCIENCE CENTER
COLLEGE PARK, MARYLAND

Technical Report TR-142 January 1971
GJ-231 and NGL-21-002-008

CONTINUATION METHODS FOR NONLINEAR EQUATIONS

by

John H. Avila

This research was supported in part by Grant GJ-231 from the National Science Foundation and by Grant NGL-21-002-008 from the National Aeronautics and Space Administration to the Computer Science Center of the University of Maryland.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

Introduction and Background

Section 1. Introduction

Many iterative techniques for the solution of nonlinear equations have the drawback that convergence depends on a good initial approximation to the solution. Correspondingly most convergence results only guarantee the existence of a well defined convergent sequence of iterates for very restricted sets of starting vectors. This leads to the following well-known concepts of a domain and a point-of-attraction.

1.1 Definition. Let $G: D \subset X \to X$ be a mapping on some metric space X. Then any non-empty set $D_0 \subset D$ is a domain of attraction of the iterative process

$$x^{n+1} = G(x^n) \ , \ n = 0,1,... \tag{1}$$

with respect to the point x^* if for any $x^0 \in D_0$ we have $\{x^n\} \subset D$ and $\lim_{n\to\infty} x^n = x^*$.

If $x^* \in \text{int}\,(D_0)$ for some domain of attraction D_0, then x^* is a point of attraction of (1).

Note that the domain of attraction need not contain all subsequent iterates after the first.

A basic theorem which gives a sufficient condition for some x^* to be a point of attraction of (1) is the following well-known result of Ostrowski [1966].

1.2 Theorem. Let X be a normed linear space and assume that $G: D \subset X \to X$ has a Frechet derivative at the fixed point $x^* \in \text{int } D$ of G.

If

(2) $\quad \|G'(x^*)\| < 1,$

then x^* is a point of attraction of (1) and, more precisely, there is an open ball $S(x^*,r)$ with center x^* and radius $r > 0$ which is a domain of attraction of (1) with respect to x^*.

Ostrowski states the theorem for the case when $X = R^n$. Then (2) can be weakened to the condition

(3) $\quad \rho(G'(x^*)) < 1$

where $\rho(\cdot)$ denotes the spectral radius of $G'(x^*)$, (see, e.g. Ortega and Rheinboldt [1970]).

This paper is concerned with one approach in overcoming the local convergence nature of iterative processes, namely with the so-called continuation method. Basically, this method can be described as follows: let $F: D \subset X \to X$ be a given mapping on the metric space X and consider the problem of solving the equation

(4) $\quad F(x) = 0.$

This problem is imbedded into a family of problems of the form

(5) $\quad H(t,x) = 0.$

where $t \in [0,1] \equiv J$ is a parameter. The imbedding is chosen so that at $t = 0$, the solution of (5) is a known point x_0, while at $t = 1$, the solution x_1 of (5) also solves (4). For example H might have the form

(6) $H(t,x) = F(x) + (t-1)\,F(x_0)$.

Suppose now that there exists a continuous solution curve $x: J \to D$ of (5) starting at x_0. The idea behind the continuation method then involves proceeding in some as yet unspecified manner along or near this curve $x = x(t)$ from the initial point $x_0 = x(0)$ to the terminal point $x_1 = x(1)$.

A basic method we will consider in this paper is the following: Suppose that $G: J \times D \subset J \times X \to X$ is a mapping with the property that

(7) $x(t) = G(t,x(t)), \quad \forall\, t \in J,$

where $x: J \to D$ denotes again the solution of (5). Then we can consider, for each t, the iterative process

(8) $x^{n+1} = G(t,x^n)$.

In general, for fixed t this process will converge to x(t) only for starting values near that point. In particular, we cannot hope that we have convergence for the process with t = 1 starting from x(0). This leads to the following numerical continuation process.

A partition of J

(9) $0 = t_0 < t_1 < \ldots < t_N = 1,$

and a sequence of integers $\{j_k\}$, $k = 1, \ldots, N-1$, is chosen such that the points

(10a)
$$x_k^{j+1} = G(t_k, x_k^j),\ j = 0,\ldots,j_k-1,\ k = 1,\ldots,N-1,$$
$$x_{k+1}^0 = x_k^{j_k},\ x_1^0 = x_0$$

are well defined and that

(10b) $\quad x_N^{j+1} = G(1, x_N^j)$, $j = 0,1,\ldots$

converges to $x(1)$ as $j \to \infty$.

The principal problem is, of course, to choose the partition (9) so that $x(t_k)$ lies in some domain of attraction $D_{t_{k+1}}$ of $x(t_{k+1})$ for each k, $1 \le k < N$. Then, if $x_k^o \in D_{t_k}$ the sequence generated by (7) for $t = t_k$ must ultimately produce an iterate $x_k^{j_k} \in D_{t_{k+1}}$ which in turn can be taken as the starting point $x_{k+1}^o = x_k^{j_k}$ for the next iteration involving t_{k+1}. Thus with $x_1^o = x_o$ as initial point the entire process can be carried out until finally $t_k = t_N = 1$ is reached. For $t = 1$, $x_N^o = x_N^{j_{N-1}}$ is then in D_1 which ensures that (10b) converges to $x(1)$ as $j \to \infty$.

There is obviously a trade-off between the number of iterative steps j_k taken at $t = t_k$ and the allowable length of the parameter step $t_{k+1} - t_k$. Clearly, for certain choices of the partition (9) it may happen that (10) simply cannot be carried out since a proper overlap of the domains of attraction D_{t_k} is not guaranteed.

1.3 Definition. If a partition (9) exists so that with some sequence of integers $\{j_k\}$ the entire process (10) is well-defined and that (10b) converges to $x(1)$, then the numerical continuation process (10a,b) is called feasible.

In Chapter II we shall examine the feasibility of numerical continuation processes. In particular we will establish a point-of-attraction theorem which extends Ostrowski's result to the process (10).

This general result is then applied to obtain specific theorems for various iterative processes. As a practical application of the results to a particular example, we also introduce in Chapter II a model problem to which we return in each of the later chapters.

The question of the existence and continuity of the solution curve $x = x(t)$ satisfying (5) is considered in Chapter III. Since the literature in this area is extensive, we confine ourselves here to a survey of some of the typical and important results. Among these results is a theorem due to Yakovlev [1965] for which the original proof was unfortunately incomplete, and for which we present a completed proof. Here, too, we apply the results to the model problem.

In Chapter IV we take up the problem of estimating the size of the parameter step $t_{k+1} - t_k$ in (10a). The results show the difficulty in obtaining good steplength estimates. This is not surprising since we already saw above that the steplength question is intimately bound up with the question of the determination of domains of attraction, a problem generally acknowledge to be very difficult. The results which we obtain are applied to a class of mildly nonlinear elliptic boundary value problems. These examples show clearly the conservative nature of the types of steplength estimates derivable from standard convergence theorems.

In Chapter IV we also consider a generalization of the basic continuation process in which the initial phase (10a) is essentially uncoupled from the terminal phase (10b). Broadly speaking, (10a) is replaced by any process which generates after finitely many steps

a point in the convergence domain of the terminal process, but for which we need not be able to guarantee any longer that the intermediate points are close to the solution curve. These generalized numerical continuation methods include a number of the processes obtained from a numerical treatment of a differential equation describing the underlying homotopy.

Finally in Chapter V we present a new approach to numerical continuation. This approach involves basically the creation of a related minimization problem and the application of the techniques of unconstrained minimization such as, for example, described in Ortega and Rheinboldt [1970, Chapters 8 and 14]. The information available in this setting allows the development of algorithms which appear to be well suited to the numerical continuation problem and which by nature generate in essence their own steplength.

Included in Chapter V are also computational results both for the model problem and for some additional examples studied previously by other authors.

Section 2: Historical Background

The use of continuation, or imbedding, methods as a theoretical tool for the study of operator equations goes back to the last century; see Ficken [1951] for an excellent historical summary and for many references. As a numerical tool one of the earliest uses of the basic numerical continuation process appears to be due to Lahaye [1934],

[1935] for a single equation and later, [1948], for systems of equations. Subsequently, many authors, including Sidlovskaya [1958], Anselone and Moore [1966], Davis [1966], and Deist and Sefor [1967], have used the numerical continuation approach.

Concurrent with the work of these authors, another approach (which in essence is included in our generalized numerical continuation processes) has been investigated extensively in the literature. This approach is based on the differential equation underlying the homotopy and integrates numerically the subsequent initial value problem describing the continuation problem. The first in this area appears to have been Davidenko [1953a,b] although he himself refers to some earlier work by other Soviet authors. Davidenko has since applied the method to a variety of problems and, at least in the Soviet Union, the technique is generally called the method of differentiation with respect to a parameter.

Among those who have examined the application of these techniques to Banach spaces are Yakovlev [1965] and Meyer [1968]. Many other authors have also employed these techniques. For example, Bittner [1967], Kleinmichel [1968], and Bosarge [1968] have each investigated the use of higher order numerical integration methods in solving the initial value problem.

Finally Davis [1966] and Meyer [1968] also considered the problem arising when $H_x(t,x)$ is singular at some point along the solution curve.

So far, little attention appears to have been paid to the theoretical foundation of the basic numerical continuation process and, in particular, to its feasibility. This is one of the topics

of this paper. As far as can be determined there also has been no attempt made of using minimization techniques in connection with continuation problems. The algorithms of this type developed here show that in such a minimization setting many of the serious difficulties encountered in estimating the steplength are greatly reduced, and that in their overall efficiency these new algorithms compare favorably with the known methods.

CHAPTER II

The Feasibility of Numerical Continuation

Section 1. An Example

From the definition of numerical continuation in Chapter I, one might conclude that if every point $x(t)$ of the solution curve x is a point of attraction of

$$x_t^{j+1} = G(t,x_t^j) \quad , \tag{1}$$

then a compactness argument for the curve x automatically insures feasibility of the process. In this section we will consider an example which shows that this is not the case. A second conjecture might be that if $G:J \times D \subset J \times X \to X$ is Frechet differentiable with respect to x and if $\|G_x(t,x(t))\| < 1$, for all $t \in J$, (that is, the Ostrowski Theorem applies) then continuity of $G_x(\cdot,x(\cdot))$ would guarantee feasibility of numerical continuation with respect to the iterative process defined by G. This too is incorrect as our example will show.

2.1 Example. For $t \in J$, $x \in R^1$, define $G:J \times [0,1] \to R^1$ by

$$G(t,x) = t \text{ for } 0 \leq t \leq \frac{1}{2},\ x \in [0,1], \tag{2a}$$

and for $\frac{1}{2} < t \leq 1$, by

$$(2b)\quad G(t,x) = \begin{cases} 6t - \frac{5}{2} & \text{for } x < -2t + \frac{3}{2} \\ -2x + 2t + \frac{1}{2} & \text{for } -2t + \frac{3}{2} \leq x < \frac{1}{2}t + \frac{1}{4} \\ t & \text{for } \frac{1}{2}t + \frac{1}{4} \leq x \leq \frac{3}{2}t - \frac{1}{4} \\ -2x + 4t - \frac{1}{2} & \text{for } \frac{3}{2}t - \frac{1}{4} < x \leq 4t - \frac{3}{2} \\ -4t + \frac{5}{2} & \text{for } 4t - \frac{3}{2} < x \end{cases}$$

For $t \leq \frac{1}{2}$ it is clear that the iterative process (1) converges in one step to the solution $x(t) = t$ for any starting value $x_t^0 \in [0,1]$. For $t > \frac{1}{2}$ we have the following:

$$(3a)\qquad G(t,x) \geq 2t - \frac{1}{2} \text{ for } x \leq \frac{1}{2}$$

$$(3b)\qquad G(t,x) \leq \frac{1}{2} \text{ for } x \geq 2t - \frac{1}{2}$$

and

$$(4a)\qquad G(t,x) \leq 2t - \frac{1}{2} - 2\varepsilon \text{ for } x \geq \frac{1}{2} + \varepsilon,\ \varepsilon > 0$$

$$(4b)\qquad G(t,x) \geq \frac{1}{2} + 2\varepsilon \text{ for } x \leq 2t - \frac{1}{2} - \varepsilon,\ \varepsilon > 0.$$

From the inequalities (3a,b) it follows that for $t > \frac{1}{2}$ any $x_t^0 \notin (\frac{1}{2}, 2t - \frac{1}{2})$ cannot be in a domain of attraction of $x(t) = t$. The inequalities (4a,b) show that $x_t^0 \in (\frac{1}{2}, 2t - \frac{1}{2})$ implies that (1) eventually produces an iterate $x_t^{j'} \in [\frac{1}{2}t + \frac{1}{4}, \frac{3}{2}t - \frac{1}{4}]$ and, since

then $x_t^{j'+1} = x(t)$, that x_t^0 lies in a domain of attraction.

In short, the domains of attraction for $x(t)$ and the iterative process (1) are precisely

$$D_t = \begin{cases} [0,1] & \text{for } t \leq \frac{1}{2} \\ (\frac{1}{2}, 2t - \frac{1}{2}) \cap [0,1] & \text{for } t > \frac{1}{2} \end{cases}$$

Assume that numerical continuation were feasible and let $\{t_k\}$, $k = 0,1,\ldots,N$ be the corresponding partition of J and suppose that t_m is the first t_k greater than $\frac{1}{2}$. Then, in the sequence of iterates

$$(5) \qquad x_{m-1}^{j+1} = G(t_{m-1}, x_{m-1}^j), \; j = 0,1,\ldots,$$

there must be one which is contained in D_{t_m}. However, since $t_{m-1} \leq \frac{1}{2}$ and $x_{m-1}^j = t_{m-1}$ for all $j > 0$, we clearly have $t_{m-1} \notin D_{t_m}$. This contradiction shows that numerical continuation cannot be feasible.

Section 2. The Basic Numerical Continuation Theorem

We recall from Chapter I the equations which define the process of numerical continuation. Throughout this chapter, X denotes a Banach space unless otherwise stated. Assume that a mapping $G: J \times D \subset J \times X \to X$ exists together with a curve $x: J \to D$ such that

(6) $\qquad x(t) = G(t,x(t)), \quad \forall t \in J.$

For a partition of J:

(7) $\qquad 0 = t_0 < t_1 < \ldots < t_N = 1$

and a sequence of finite integers j_k, $k = 1,\ldots,N-1$, define

(8a)
$$\begin{cases} x_k^{j+1} = G(t_k, x_k^j), \; j = 0,\ldots,j_k-1, \; k = 1,\ldots,N-1 \\ x_{k+1}^0 = x_k^{j_k}, \; x_1^0 = x(0) \end{cases}$$

and

(8b) $\qquad x_N^{j+1} = G(1,x_N^j), \; x_N^0 = x_{N-1}^{j_{N-1}}, \; j = 0,1,\ldots .$

Then the numerical continuation process is feasible if there exists a partition (7) and integers $\{j_k\}$ so that (8a) and (8b) are well-defined and x_N^j in (8b) converges to $x(1)$ as $j \to \infty$.

In this section we state and prove a theorem which will serve as a basis for most of our results concerning the feasibility of numerical continuation.

2.2 Theorem. Consider the mapping $G:J \times D \subset J \times X \to X$ where $J = [0,1]$ and D is an open subset of the metric space X. Suppose that a continuous curve $x:J \to D$ exists and satisfies (6).

If there is a fixed radius $r > 0$ such that $D_t = S(x(t)),r) \subset D$ is a domain of attraction of

(9) $$x_t^{j+1} = G(t, x_t^j),$$

then numerical continuation is feasible.

Proof. Let $d_X(\cdot,\cdot)$ denote the metric on X. Since J is compact and x is continuous, for some $\delta > 0$, we have that

(10) $$d_X(x(t'),x(t")) < \frac{r}{2}, \quad \forall t', t" \in J, \ |t'-t"| < \delta.$$

Choose t_1 so that $\frac{\delta}{2} < t_1 - t_0 < \delta$. Then by (10), $x(t_0) \in S(x(t_1),r)$ and the sequence generated by (8a) for $k = 1$ is well-defined and converges to $x(t_1)$. Hence we can find j_1 so that $d_X(x_1^{j_1},x(t_1)) < \frac{r}{2}$.

In general, if $d_X(x_m^{j_m},x(t_m)) < \frac{r}{2}$, for some $m \geq 1$, we choose t_{m+1} so that $\frac{\delta}{2} < t_{m+1} - t_m < \delta$ or $t_{m+1} = 1$ if $1 - t_m < \delta$.

Then by the triangle inequality, $d_X(x_m^{j_m},x(t_{m+1})) < r$ which implies that (8a) is well-defined for $k = m+1$ and $\lim_{j\to\infty} x_{m+1}^j = x(t_{m+1})$ and hence that j_{m+1} can be selected so that

$$d_X(x_{m+1}^{j_{m+1}},x(t_{m+1})) < \frac{r}{2}.$$

It is clear from our construction of $\{t_k\}$ that for some N, $t_N = 1$ and that $t_N - t_{N-1} < \delta$. Then with $x_N^0 = x_{N-1}^{j_{N-1}}$, (8b) converges to $x(1)$ as $j\to\infty$, and this completes the proof.

As an example of the use of Theorem 2.2 we present a result which, in a slightly different setting, is due to Friedrichs [1950] (see also Ficken [1951]).

2.3 Example. Consider a family $\{L(t)\}$, $t \in J$, of bounded linear operators from X into itself so that $L(0) = L_0$ has a known bounded inverse L_0^{-1} and the inverse of $L(1) = L^*$ is to be determined.

More specifically, suppose that the following conditions are satisfied:

i) $L: J \to L(X,X)$ is continuous in the strong operator topology on the space of bounded linear operators $L(X,X)$ from X into itself, i.e., for every $\varepsilon > 0$ there exists $\delta(\varepsilon,t) > 0$ so that

$$\|L(t')x - L(t)x\| \leq \varepsilon\|x\|, \quad \forall x \in X, \ |t'-t| < \delta(\varepsilon,t).$$

ii) $L(t)$ is bounded below for each $t \in J$, i.e., there exists a $\mu(t) > 0$ so that

$$\|L(t)x\| \geq \frac{1}{\mu(t)}\|x\|, \quad \forall x \in X.$$

Then for $t \in J$, the inverse $L(t)^{-1}$ of $L(t)$--which exists by ii)--is uniformly bounded and $L(1)^{-1}$ can be found from L_0^{-1} by numerical continuation using as the local iterative process, a technique due to Hotelling for linear mappings on R^n (see Faddeev and Faddeeva [1960, p.159]). More precisely, with $B^{(k)} \in L(X,X)$, this process is defined by

$$B^{(k+1)} = B^{(k)}(2I - L(t)B^{(k)}). \tag{11}$$

If we let $E^{(k)}(t) = I - L(t)B^{(k)}$, it follows readily that

$$E^{(k+1)}(t) = [E^{(k)}(t)]^2 \tag{12}$$

and hence that for $\|E^{(0)}(t)\| < 1$, $\lim_{k\to\infty} B^{(k)} = L(t)^{-1}$.

If $\|L(t)^{-1}-B^{(0)}\| < \|L(t)\|^{-1}$, then

$$\|E^{(0)}(t)\| = \|L(t)[L(t)^{-1}-B^{(0)}]\| \leq \|L(t)\|\|L(t)^{-1}-B^{(0)}\| < 1$$

which implies that (11) converges starting from $B^{(0)}$.

As a result, $S(L(t)^{-1},\|L(t)\|^{-1})$ is a domain of attraction of $L(t)^{-1}$ and the process defined by (11).

For arbitrary but fixed $x \in X$, with $\|x\| = 1$, let $\phi_x(t) = \|L(t)x\|$. Using i) and ii) we find that given $\varepsilon > 0$, there exists a $\delta(t,\varepsilon)$, independent of x, so that for $|t-t'| < \delta(t,\varepsilon)$,

$$\|L(t')x\| \geq \|L(t)x\| - \|L(t)x-L(t')x\| \geq \frac{1}{\mu(t)} - \varepsilon.$$

Thus, $\phi_x(t)$ is a lower semi-continuous function of t and, since $\mu(t) > 0$ on J, for some $\mu > 0$, it follows that $\phi_x(t) \geq \mu^{-1} > 0$ for all $t \in J$ and for all $x \in X$. But then $\|L(t)^{-1}\| \leq \mu$ for all $t \in J$.

Therefore,

$$\|[L(t_1)^{-1}-L(t_2)^{-1}]x\| \leq \|L(t_1)^{-1}\|\|L(t_1)-L(t_2)\|\|L(t_2)^{-1}\|\|x\|$$

$$\leq \mu^2\|L(t_1)-L(t_2)\|\|x\|.$$

From this it follows that $L(t)^{-1}$ is continuous in the strong operator topology.

By the continuity of $L(t)$, $\|L(t)\| \leq \alpha$ for all $t \in J$ and some

$\alpha > 0$. Hence $\|L(t)\|^{-1} \geq \frac{1}{\alpha}$. Now by Theorem 2.2 it follows that numerical continuation is feasible along the curve $L(t)^{-1}$.

The setting for the preceding result in the cited work of Friedrichs and Ficken involved a continuation process in a functional analytic framework. First, a constant δ is shown to exist such that $L(t)^{-1}$ exists for $0 \leq t \leq \delta$. Then knowing $L(\delta)^{-1}$ one can obtain the existence of $L(t)^{-1}$ on $(\delta, 2\delta]$, etc. Here, the operator $L(t)^{-1}$, except at $t = 0$ as a starting value, is not needed explicitly in order to obtain in the limit the existence of $(L^*)^{-1}$.

The condition that $x(t)$ is a point of attraction of the process (9) will become our key point in the proof of the feasibility of numerical continuation. Recall that a point of attraction is by Definition 1.1 an interior point of some domain of attraction and hence that this condition insures that for some $r > 0$, $S(x(t),r)$ must be a domain of attraction. Motivated by this fact, we consider the function $r:J \to [0,\infty)$ defined by

(13) $r(t) = \sup \{s \geq 0 \mid S(x(t),s) \text{ is a domain of attraction}\}$

where $x:J \to X$ is a solution of the fixed point equation (6) and the iterative process for each t is given by (9).

If we are now able to show that $r(t) \geq r > 0$ on J, Theorem 2.2 can be used to prove the feasibility of numerical continuation. In general, it is too much to hope for continuity of $r(t)$ as in the Example 2.3. However, $r(t)$ is frequently bounded below by a positive

lower semi-continuous function on J and this insures again the existence of a suitable constant $r > 0$.

The following simple lemma summarizes the equivalence of various properties for positive real-valued functions on J, all of which will provide a sufficient condition for the existence of the desired lower bound on $r(t)$.

2.4 Lemma. For any function $r:J \to R^1$, the following statements are equivalent:

(1) $r(t) \geqslant r > 0$ for all $t \in J$;

(2) For each $t_0 \in J$ there exists a constant $\eta = \eta(t_0) > 0$ and a $\delta = \delta(t_0) > 0$ such that $r(t) \geqslant \eta$ for $|t-t_0| < \delta$;

(3) $r(t) \geqslant \xi(t) > 0$ for all $t \in J$, where $\xi:J \to R^1$ is lower semi-continuous.

Proof. Clearly, (1) implies (2). To see that (2) implies (3) let $U_s \equiv (s-\delta(s),s+\delta(s))$; then $r(t) \geqslant \eta(s)$ for $t \in U_s \cap J$. By compactness of J we can choose a finite cover $\{U_{t_i}\}$ of J and define ξ by $\xi(t) \equiv \eta$, $t \in J$, where $\eta = \inf \eta_{t_i} > 0$. Clearly then $r(t) \geqslant \eta > 0$ for $t \in J$ and ξ is lower semi-continuous.

Since a lower semi-continuous function $\xi:J \to R^1$ assumes its minimum, we have $r(t) \geqslant r = \min_{t \in J} \xi(t)$ and $r = 0$ is excluded since otherwise $\xi(t_0) = 0$ for some $t_0 \in J$ against assumption. Thus (3) implies (1).

2.5 Remark. We will often refer to property (2) of Lemma 2.4 by saying that r is locally positively bounded below.

Section 3. Point of Attraction Theorems and Numerical Continuation

As we have seen in Section 1, even if along the solution curve x the hypotheses of Ostrowski's Theorem are satisfied for the iterative process

$$x_t^{j+1} = G(t,x_t^j), \tag{14}$$

we cannot guarantee feasibility of numerical continuation. In this section we shall show that if, in addition to the Ostrowski hypotheses, the derivative G_x at $(t,x(t))$ has an additional property, then numerical continuation is feasible.

This property involves the concept of a strong derivative first introduced by Ortega and Rheinboldt [1970].

2.6 Definition. Let $G:J \times D \subset J \times X \to X$ where X is a normed linear space and D is open in X. Assume that for some fixed $t_0 \in J$, $x_0 \in D$, the partial Frechet derivative $G_x(t_0,x)$ of G with respect to x, exists at x_0. Then $G_x(t_0,x_0)$ is a strong Frechet derivative at (t_0,x_0) if, for any $\varepsilon > 0$, there exists a $\delta > 0$ depending on t_0,x_0, and ε such that

$$\|G(t,y+h)-G(t,y+k)-G_x(t_0,x_0)(h-k)\| \leq \varepsilon\|h-k\|$$

for $|t-t_0| + \|x_0-y\| \leq \delta$, $\|h\| \leq \delta$, and $\|k\| \leq \delta$.

The next lemma shows that continuity of G_x jointly in t and x at (t_0,x_0) implies the strongness of G_x (see again Ortega and Rheinboldt [1970]).

2.7 Lemma. Let $G:J \times D \subset J \times X \to X$ where D,X are as specified in Definition 2.6. If G_x exists in an open neighborhood of $(t_0,x_0) \in J \times D$ and is continuous at (t_0,x_0), then $G_x(t_0,x_0)$ is strong.

Proof. Let $\varepsilon > 0$ be given and choose δ_1,δ_2 so that $G_x(t,x)$ exists for $t \in S_1 = S(t_0,\delta_1) \cap J$, $x \in S_2 = S(x_0,\delta_2)$ and that

$$\|G_x(t,x)-G_x(t_0,x_0)\| \leq \varepsilon \text{ for } t \in S_1,\ x \in S_2.$$

For given $t \in S_1$ let $\phi:S_2 \to X$ be defined by

$$\phi(x) = G(t,x) - G_x(t_0,x_0)x.$$

Then

$$\|\phi'(x)\| = \|G_x(t,x)-G_x(t_0,x_0)\| \leq \varepsilon,$$

independent of x in S_2. Thus for $\delta = \delta_2/2$, $y \in S_2$, $\|y-x_0\| \leq \delta$, $h,k \in X$, $\|h\|$, $\|k\| \leq \delta$,

$$\|G(t,y+h)-G(t,y+k)-G_x(t_0,x_0)(h-k)\|$$

$$= \|\phi(y+h)-\phi(y+k)\|$$

$$\leq \sup_{0\leq s\leq 1} \|\phi'(y+k+s(h-k))\|\|h-k\|$$

$$\leq \varepsilon\|h-k\|;$$

hence the conditions of Definition 2.6 are satisfied.

With the concept of a strong derivative, we can now state the fundamental result of this chapter.

2.8 Theorem. Let $G:J \times D \subset J \times X \to X$ where X is a normed linear space and $D \subset X$ is open. Assume that $x:J \to D$ is continuous and satisfies $x(t) = G(t,x(t))$. Suppose further that the partial Frechet derivative G_x of G with respect to x at $(t,x(t))$ exists and is strong for all $t \in J$. Then if $\|G_x(t,x(t))\| < 1$, numerical continuation is feasible.

Proof. Let $t_0 \in J$ and $d \equiv \|G_x(t_0,x(t_0))\|$. Since $d < 1$, there exists an $\eta > 0$ such that $d + \eta < 1$. Write $x = x(t_0) + (x-x(t_0))$ and $x(t) = x(t_0) + (x(t)-x(t_0))$. By the definition of a strong derivative, there exists a $\delta_1 > 0$ such that if $\|x-x(t_0)\| \leq \delta_1$, $x \in D$, $\|x(t)-x(t_0)\| \leq \delta_1$ and $|t-t_0| \leq \delta_1$, then

$$(15) \quad \|G(t,x)-G(t,x(t))-G_x(t_0,x(t_0))(x-x(t))\| \leq \eta\|x-x(t)\|.$$

Choose δ_2 sufficiently small so that $S(x(t_0),\delta_2) \subset D$. Finally, choose δ_3 so that

$$(16) \quad \|x(t)-x(t_0)\| \leq \frac{\delta_1}{2} \text{ for } |t-t_0| \leq \delta_3.$$

Let $\delta = \min\{\delta_1,\delta_2,\delta_3\}$. Then if $|t-t_0| < \delta$, $x \in S(x(t),\frac{\delta}{2})$, we have by (16), (15), and the triangle inequality:

$$\|G(t,x)-G(t,x(t))\| \leq \|G(t,x)-G(t,x(t))-G_x(t_0,x(t_0))(x-x(t))\|$$

$$+ \|G_x(t_0,x(t_0))(x-x(t))\|$$

$$\leq (d+\eta)\|x-x(t)\|.$$

Hence $|t-t_0| < \delta$ implies $r(t) \geq \frac{\delta}{2}$ where $r:J \to R^1$ is defined by (13). Since t_0 was arbitrary, Lemma 2.4, part (2), and Theorem 2.2 complete the proof.

Recall from Chapter I that for finite dimensional spaces, Ostrowski's Theorem relates the point of attraction result to the spectral radius of the Frechet derivative. Correspondingly we can also rephrase Theorem 2.8 in the finite dimensional case on the basis of a condition upon the spectral radius of $G_x(t,x(t))$.

2.9 Theorem. Let $G:J \times D \subset J \times R^n \to R^n$, where D is open and assume that $x:J \to D$ is continuous and satisfies $x(t) = G(t,x(t))$. Let G have a strong partial Frechet derivative with respect to x at $(t,x(t))$ for every $t \in J$.

If $\rho(G_x(t,x(t))) < 1$ for all $t \in J$, then numerical continuation is feasible.

Proof. Let $\|\cdot\|$ denote the given norm on R^n. By the Ostrowski theorem the function $r:J \to R^1$ given by (13) is well-defined. We wish to show that r is locally positively bounded below.

Let $t_0 \in J$ be fixed. Since $\rho = \rho(G_x(t_0,x(t_0))) < 1$, there exists an $\eta > 0$ such that $\rho + 2\eta < 1$. It is well-known (see, e.g.,

Householder [1964; p. 46]) that there exists a norm $\|\cdot\|_{t_0}$ on R^n so that the induced matrix norm on $L(R^n,R^n)$ satisfies $\|G_x(t_0,x(t_0))\|_{t_0} \leq \rho + \eta$. For this norm it follows exactly as in the proof of Theorem 2.8 that there exists a δ so that for $|t-t_0| < \delta$ and $\|x-x(t)\|_{t_0} < \frac{\delta}{2}$, we have

$$\|G(t,x)-G(t,x(t))\|_{t_0} \leq \eta\|x-x(t)\|_{t_0}$$

$$+ \|G_x(t_0,x(t_0))\|_{t_0}\|x-x(t)\|_{t_0} \leq (\rho+2\eta)\|x-x(t)\|_{t_0},$$

and hence that x lies in a domain of attraction of $x(t)$ for the iterative process (9).

By the equivalence of norms on R^n, there exists a constant c so that

$$\|y\|_{t_0} \leq c\|y\| \text{ for all } y \in R^n.$$

Therefore if $|t-t_0| < \delta$ and $\|x-x(t)\| < \frac{1}{c}\frac{\delta}{2}$, then

$$\|x-x(t)\|_{t_0} \leq c\|x-x(t)\| < \frac{\delta}{2}.$$

Hence $S(x(t),\frac{1}{c}\frac{\delta}{2})$ is a domain of attraction of $x(t)$ for $|t-t_0| < \delta$. As a result, since t_0 was arbitrary, r as defined by (13) is locally positively bounded below. Now apply Lemma 2.4 and Theorem 2.2 and the proof is complete.

Section 4. Model Problem

We have seen so far that the problem of numerical continuation poses some interesting questions when one seeks to determine even the feasibility of the process. As an example of the practical resolution of some of these questions, we consider what will serve as a model problem throughout this paper.

The problem arises in the study of diffusion processes, for instance, when oxygen diffuses into a cell in which an enzyme-catalyzed reaction occurs. It has been studied by several authors, for example, see Murray [1968], and Keller [1968, p. 162 ff]. In its general setting this problem can be formulated as the two-point boundary value problem:

$$(17a) \qquad (s^2 d(x)x')' = s^2 f(x),\ 0 < s < 1$$

$$(17b) \qquad x'(0) = 0;\ x(1) = 1.$$

Here the function f has the form

$$(18) \qquad f(x) = \frac{1}{\varepsilon}\frac{x}{x+k},\ \varepsilon, k > 0,\ x > -k/2,$$

and d is the molecular diffusion coefficient which, in this nonlinear example, depends on x. For small values of the parameter ε, singular perturbation techniques, as developed by Murray, provide an analytic approach. For larger values of ε, Keller states a result that for $6\varepsilon k > 1$ and $d(x) \equiv 1$, the iterative process

$$(s^2x'_{n+1})' = s^2f(x_n),\ n = 0,1,\ldots;\ x_0 \equiv 0,$$

(19)

$$x'_n(0) = 0;\ x_n(1) = 1,\ \text{for } n > 0,$$

converges to the unique positive solution of (17a,b) and that

$$0 \equiv x_0(s) \leqslant x_2(s) \leqslant \ldots \leqslant x(s) \leqslant \ldots \leqslant x_3(s) \leqslant x_1(s) \equiv 1.$$

For our purposes we will phrase (17a,b) as an integral equation, and we begin with the following theorem. Here $C[0,1]$ is the usual Banach space of continuous functions on $[0,1]$ with the norm

$$\|x\| = \sup_{s\in[0,1]} |x(s)|,\ x \in C[0,1])$$

and

(20) $$D_k = \{x \in C[0,1] \mid x(s) > -\frac{k}{2},\ 0 \leqslant s \leqslant 1\}.$$

2.10 Theorem. Let $f:(-k/2,\infty) \to R^1$ be defined by (18) and assume that $d:(-k/2,\infty) \to R^1$ is continuous and satisfies $d(s) \geqslant \alpha > 0$ for $s \in (-k/2,\infty)$. Then for any fixed $t \in J$ every non-negative solution $x \in C[0,1]$ of

(21) $$x(s) = 1 - t\int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi))d\xi du,$$

is a continuously differentiable solution of

(22a) $$(s^2 d(x)x')' = ts^2 f(x),\ 0 < s < 1,$$

(22b) $$x'(0) = 0;\ x(1) = 1,$$

and conversely, every continuously differentiable non-negative solution of (22a,b) is a solution of (21).

For the sake of the continuity in the present discussion we have placed the proof of this theorem in the appendix to this chapter.

Theorem 2.10 already suggests the inbedding which we will consider for the solution of (17a,b). More specifically, let $G:J \times D_k \to J \times C[0,1]$ be defined by

(23) $$G(t,x)(s) = 1 - t\int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi))d\xi du.$$

We then consider the fixed point equation

(24) $$x = G(t,x),\ t \in J,\ x \in D_k.$$

The existence of a solution curve $x:J \to D_k$ will be based on the following theorem which we shall prove in Chapter 3.

2.11 Theorem. Let $G:J \times D_k \to C[0,1]$ be defined by (23) with the functions d and f as given in Theorem 2.10. Moreover, assume that

(25) $$6\varepsilon\alpha \geq 1,\ \alpha = \inf\{d(s) | -k/2 < s < \infty\} > 0.$$

Then, for every fixed $t \in J$, a solution $x = x(t) \in D_k$ of (24) exists.

If d is continuously differentiable on its domain, then, in addition, the partial Frechet derivative G_x exists for $(t,x) \in J \times D_k$ and is continuous on $J \times D_k$. Moreover, in this case for k and $\max_{0\le x\le 1} |d'(x)|/d^2(x)$ sufficiently small, $\|G_x(t,x(t))\| < 1$.

The last portion of the theorem concerning the existence of G_x is proved in the appendix. In fact, we shall show there that for $h \in C[0,1]$

$$(26)\qquad [G_x(t,x)h](s) = -t\int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f'(x(\xi))h(\xi)\,d\xi\,du$$

$$+ t\int_s^1 \frac{1}{u^2} \frac{d'(x(u))h(u)}{d^2(x(u))} \int_0^u \xi^2 f(x(\xi))\,d\xi\,du.$$

From (26) we can estimate the norm of G_x as is again shown in the appendix. More precisely, we obtain that

$$(27)\qquad \|G_x(t,x(t))\| \le \frac{tk}{2b(a+k)\alpha\varepsilon}\left[1-\left(\frac{b}{a+k}\right)^{-1/2} \arctan\left(\frac{b}{a+k}\right)^{1/2}\right]$$

$$+ \frac{\beta t}{\varepsilon}\left\{\frac{1}{6} - \frac{k}{2(1-c)} - \left(\frac{1-c}{c+k}\right)\right.$$

$$\left. + \frac{k(c+k)(1+k)}{(1-c)^3} \ln\left(\frac{1+k}{c+k}\right)\right\}$$

where $0 < \alpha \le \min_{0\le x\le 1} d(x)$, $\beta \ge \max_{0\le x\le 1} \frac{|d'(x)|}{d^2(x)}$, $b = \frac{t}{6\alpha\varepsilon}$, $a = 1 - b$,

$$c = 1 - \frac{t}{\gamma\varepsilon}\left\{\frac{1}{6} - \frac{k}{2b}\ \ell n\ (1 + \frac{b}{a+k}) + \frac{k}{b}\left(\frac{b}{a+k}\right)^{-1/2} \arctan\left(\frac{b}{a+k}\right)^{1/2} - \frac{k}{b}\right\},$$

$\gamma \geq \max_{0<x<1} d(x)$. The purpose of this somewhat formidable expression is to indicate, first of all, that estimates for the norm of G_x at $(t,x(t))$ are possible. More importantly, however, this expression gives us a means of estimating the relationship between ε and k. In particular, since one can see from (27) that $\|G_x(t,x(t))\| \to 0$ as $k\to 0$ and $\beta \to 0$, it follows that for ε as specified by (25) and for sufficiently small k and β numerical continuation is feasible by Theorem 2.8 and Lemma 2.7.

In comparison to the estimate $6\varepsilon k > 1$ given by Keller for $d \equiv 1$, we can use (27) to obtain that $\|G_x(t,x(t))\| < 1$ for $\varepsilon = \frac{1}{6}$, $k > .4749$. Furthermore, for $\varepsilon > .178$, $k > 0$ is already sufficient to ensure that $\|G_x(t,x(t))\| < 1$.

Section 5. Continuation with Specific Locally Convergent Iterative Processes

In the preceding sections we have dealt with the feasibility of numerical continuation in the general case. In this section we consider the question of feasibility as applied to the use of specific locally convergent processes in the numerical continuation process.

Most authors mentioned in Chapter I as having employed some form of a continuation approach have used Newton's method as the local iterative process. We shall consider Newton's method and several of the so-called generalized linear iterations.

We begin by recalling the following well-known lemmas. As before, X represents a given Banach space.

2.12 Lemma. Let $A \in L(X,X)$ have the property that $\|A\| < 1$. Then $I - A$ has a bounded linear inverse and $\|(I-A)^{-1}\| \leq (1-\|A\|)^{-1}$.

The next lemma is sometimes referred to as the perturbation lemma.

2.13 Lemma. Let $A \in L(X,X)$ have a bounded linear inverse and assume that $B \in L(X,X)$ satisfies $\|A-B\| < \|A^{-1}\|^{-1}$. Then B has a bounded inverse and

$$\|B^{-1}\| \leq \|A^{-1}\|/(1-\|A^{-1}\|\|A-B\|). \tag{28}$$

For a proof of both of these lemmas, see Taylor [1958; p. 164f.].

Next, we prove a lemma due to Ortega and Rheinboldt [1970] which will enable us to obtain the Frechet derivative for operators of a certain form. We have adapted the lemma to the case of the homotopy equation which concerns us here, and we added an extension concerning the strongness of the derivative.

2.14 Lemma. Suppose that $H:J \times D \subset J \times X \to X$, D open, has a partial Frechet derivative with respect to x at a point $(t^*,x^*) \in J \times D$ for which $H(t^*,x^*) = 0$. Let $S_0 \subset D$ be an open neighborhood of x^* and $I_0 \subset J$ a relatively open neighborhood of t^*, and assume that the mapping $A:I_0 \times S_0 \to L(X,X)$ is continuous at (t^*,x^*) and that $A(t^*,x^*)$ has a bounded inverse. Then there exists a relatively open interval $I \subset I_0$ and a ball $S = \bar{S}(x^*,\delta) \subset S_0$, $\delta > 0$, on which the mapping

$$G:I \times S \to X,\ G(t,x) = x - A(t,x)^{-1}H(t,x), \quad \forall (t,x) \in I \times S$$

is well defined and G has a partial Frechet derivative with respect to x at (t^*,x^*) given by

$$G_x(t^*,x^*) = I - A(t^*,x^*)^{-1}H_x(t^*,x^*). \tag{29}$$

If A has the additional property that there exists a constant K such that

$$\|A(t,x)-A(t,y)\| \leq K\|x-y\|, \forall (t,x),(t,y) \in I_0 \times S_0,$$

then continuity of H at (t^*,x^*) and strongness of $H_x(t^*,x^*)$ implies the strongness of $G_x(t^*,x^*)$.

Proof. Let $\beta = \|A(t^*,x^*)^{-1}\|$ and choose ε and δ so that $0 < \varepsilon < (2\beta)^{-1}$, $S = \bar{S}(x^*,\delta) \subset S_0$, $I = \{t \in J \mid |t-t^*| < \delta\} \subset I_0$, and

$$\|A(t,x)-A(t^*,x^*)\| \leq \varepsilon, \quad \forall (t,x) \in I \times S. \tag{30}$$

By the perturbation lemma 2.13, $A(t,x)^{-1}$ exists for all $(t,x) \in I \times S$ and

$$\|A(t,x)^{-1}\| \leq \beta/(1-\beta\varepsilon) < 2\beta. \tag{31}$$

Therefore, G is well-defined on S.

Since H has a partial Frechet derivative with respect to x at (t^*,x^*), we may restrict δ if necessary so that

$$\|H(t^*,x)-H(t^*,x^*)-H_x(t^*,x^*)(x-x^*)\| \leq \varepsilon\|x-x^*\|, \quad \forall x \in S. \tag{32}$$

Since $x^* = G(t^*,x^*)$ we combine (30)-(32) to obtain

$$\begin{aligned}
&\|G(t^*,x)-G(t^*,x^*)-[I-A(t^*,x^*)^{-1}H_x(t^*,x^*)](x-x^*)\| \\
&\quad = \|A(t^*,x^*)^{-1}H_x(t^*,x^*)(x-x^*)-A(t^*,x)^{-1}H(t^*,x)\| \\
&\quad \leq \|-A(t^*,x)^{-1}[H(t^*,x)-H(t^*,x^*)-H_x(t^*,x^*)(x-x^*)]\| \\
&\qquad + \|A(t^*,x)^{-1}[A(t^*,x^*)-A(t^*,x)]A(t^*,x^*)^{-1}H_x(t^*,x^*)(x-x^*)\| \\
&\quad \leq [2\beta\varepsilon+2\beta^2\varepsilon\|H_x(t^*,x^*)\|]\|x-x^*\|, \quad \forall x \in S.
\end{aligned} \tag{33}$$

Since ε is arbitrary and β as well as $\|H_x(t^*,x^*)\|$ are fixed, (33) shows that G has at (t^*,x^*) a partial Frechet derivative with

respect to x given by (29).

Now suppose that A satisfies in $I_0 \times S_0$ a Lipschitz condition with respect to x and that H is continuous at (t^*,x^*) and $H_x(t^*,x^*)$ is strong. Then, given $\varepsilon > 0$, we can restrict $\delta > 0$ if necessary so that

$$\begin{aligned}\|H(t,x)-H(t,y)\| &\leq \|H(t,x)-H(t,y)-H_x(t^*,x^*)(x-y)\| \\ &\quad + \|H_x(t^*,x^*)(x-y)\| \\ &\leq (\varepsilon+\|H_x(t^*,x^*)\|)\|x-y\|, \quad \forall (t,x),(t,y) \in I \times S,\end{aligned}$$

and $\|H(t,y)\| \leq \varepsilon$, $\forall (t,y) \in I \times S$. Hence, similarly as in (33),

$$\begin{aligned}&\|G(t,x)-G(t,y)-G_x(t^*,x^*)(x-y)\| \\ &\quad \leq \|A(t^*,x^*)^{-1}[H(t,x)-H(t,y)-H_x(t^*,x^*)(x-y)]\| \\ &\qquad + \|[A(t,x)^{-1}-A(t^*,x^*)^{-1}]F(x)+[A(t^*,x^*)^{-1}-A(t,y)^{-1}]F(y)\| \\ &\quad \leq 2\beta\varepsilon\|x-y\| \\ &\qquad + \|[A(t,x)^{-1}-A(t^*,x^*)^{-1}][H(t,x)-H(t,y)]\| \\ &\qquad\quad + \|[A(t,x)^{-1}-A(t,y)^{-1}]H(t,y)\| \\ &\quad \leq 2\beta\varepsilon\|x-y\| + 2\beta^2\varepsilon(\varepsilon+\|H_x(t^*,x^*)\|)\|x-y\| \\ &\qquad\quad + 4\beta^2K\|H(t,y)\|\|x-y\| \\ &\quad \leq \varepsilon[2\beta+2\beta^2(\varepsilon+\|H_x(t^*,x^*)\|)+4\beta^2K]\|x-y\|\end{aligned} \tag{34}$$

which proves that $G_x(t^*,x^*)$ is strong.

We turn now our attention to Newton's method as the local process in numerical continuation. Consider the homotopy $H:J \times D \subset J \times X \to X$ and the curve $x:J \to D$ such that $H(t,x(t)) = 0$. If H has a Frechet derivative H_x with respect to x on $J \times D$ and $H_x(t,x)^{-1}$ exists and is bounded in some neighborhood of the curve x, the Newton numerical continuation process has the following form.

For some partition $\{t_k\}$, $k = 0,1,\ldots,N$, of J and finite integers $\{j_k\}$, $k = 1,\ldots,N-1$, the process is defined by the equations:

$$(35a)\quad \begin{cases} x_k^{j+1} = x_k^j - H_x(t_k,x_k^j)^{-1}H(t_k,x_k^j), & k=1,\ldots,N-1,\ j=0,1,\ldots,j_k-1, \\ x_{k+1}^0 = x_k^{j_k}, \quad x_1^0 = x(0), \end{cases}$$

followed by

$$(35b)\quad \begin{cases} x_N^{j+1} = x_N^j - H_x(1,x_N^j)^{-1}H(1,x_N^j), & j = 0,1,\ldots \\ x_N^0 = x_{N-1}^{j_{N-1}}. \end{cases}$$

Consider the mapping $G:J \times D \subset J \times X \to X$ given by

$$(36)\qquad G(t,x) = x - H_x(t,x)^{-1}H(t,x).$$

By using Lemma 2.14 we see that

$$G_x(t,x(t)) = I - H_x(t,x(t))^{-1}H_x(t,x(t)) = 0.$$

Hence, Ostrowski's Theorem applies. By insuring that $G_x(t,x(t))$ is strong we can use Theorem 2.8 to prove the feasibility of (35a,b).

2.15 Theorem. Consider the mapping $H:J \times D \subset J \times X \to X$ where D is open and assume that H has a jointly continuous partial Frechet derivative H_x which, for every $(t_0,x_0) \in J \times D$ and some $\delta(t_0,x_0)$, satisfies

$$\|H_x(t,x)-H_x(t,y)\| \leq K(t_0,x_0)\|x-y\|, \quad \forall |t-t_0| < \delta, \quad \forall x,y \in S(x_0,\delta). \tag{37}$$

Assume that a continuous function $x:J \to D$ exists and satisfies $H(t,x(t)) = 0$. If $H_x(t,x(t))^{-1}$ exists and is bounded for each $t \in J$, then the numerical continuation process (35a,b) is feasible.

Proof. From the joint continuity of H_x it follows by Lemma 2.7 that $H_x(t,x(t))$ is strong; hence (37) and Lemma 2.14 imply that also $G_x(t,x(t))$ is strong. Now apply Theorem 2.8 and the proof is complete.

2.16 Remark. Theorem 2.15 can be applied to the model problem in Section 4. Let $H:J \times D_k \subset J \times X \to X$ be defined by

$$H(t,x) = x - G(t,x),$$

where D_k and G are given by (20) and (23) respectively. Then

by Theorem 2.11, $\|G_x(t,x(t))\| < 1$ for k and $\max_{0\leq s\leq 1} |d'(s)|/d^2(s)$ sufficently small.
For those values of k, Lemma 2.12 implies the existence and boundedness of $H_x(t,x(t))^{-1}$. Then if the diffusion coefficient $d(x)$ has a Lipschitz continuous first derivative, Theorem 2.15 guarantees the feasibility of numerical continuation with the process (35a,b).

We next turn to a class of local processes referred to as generalized linear iterations and studied in detail by Ortega and Rheinboldt [1968].

Consider a mapping $F:D \subset R^n \to R^n$ and the equation

$$F(x) = 0. \tag{38}$$

In many of the standard iterative methods for the solution of (38) a system of linear equations must be solved at each step. For example, in Newton's method, the following system has to be solved at each step:

$$F'(x_k)(x-x_k) = -F(x_k). \tag{39}$$

If n is large, then it is appropriate to employ a linear iterative technique to solve (39). With the splitting $F'(x_k) = B(x_k) - C(x_k)$ a large class of linear iterative processes for solving (39) is given by

$$z^{j+1} = E(x_k)z^j + B(x_k)^{-1}b_k$$

where $E(x_k) = B(x_k)^{-1}C(x_k)$, $b_k = F'(x_k)x_k - F(x_k)$.

For $z^0 = x_k$ and $x_{k+1} = z^m$ and with the help of the identity

$$(I+E+\ldots+E^{m-1})(I-E) = I - E^m, \quad m \geq 1,$$

we have the "combined" process

$$x_{k+1} = x_k - (I+E(x_k)+\ldots+E(x_k)^{m-1})B^{-1}(x_k)F(x_k). \tag{40}$$

Some of the processes contained in the formulation (40) are, for example, Newton-(m-step)SOR, Newton-(m-step)Jacobi, etc.

We now consider the process defined by (40) as the local iterative process in numerical continuation. That is, assume $H:J \times D \subset J \times R^n \to R^n$ has a Frechet derivative H_x jointly continuous on $J \times D$ and suppose that $x:J \to D$ satisfies $H(t,x(t)) = 0$.

Assume

$$H_x(t,x) = B(t,x) - C(t,x) \tag{41}$$

is a given splitting where $B(t,x)$ is invertible in some neighborhood D_0 D of the curve x. Finally, let G be defined by

$$G(t,x) = x - [\sum_{i=0}^{m-1} E(t,x)^i]B(t,x)^{-1}H(t,x) \tag{42}$$

where

$$E(t,x) = B(t,x)^{-1}C(t,x). \tag{43}$$

Then G is defined on $J \times D_0 \subset J \times R^n$.

With these notations we are now led to the following numerical continuation result.

2.17 Theorem. Consider $H:J \times D \subset J \times R^n \to R^n$, where D is open, and a continuous curve $x:J \to D$ such that $H(t,x(t)) = 0$. Assume that H has a Frechet derivative H_x which is jointly continuous on $J \times D$ and satisfies

$$(44) \qquad \|H_x(t,x)-H_x(t,y)\| \leq K\|x-y\|, \quad \forall t \in J,\ x,y \in D.$$

Let the splitting (41) of H_x be given where B and C are jointly continuous on $J \times D$ and satisfy there the same Lipschitz condition as H_x in (44) with the same Lipschitz constant K.

Also suppose that for each $t \in J$, $H_x(t,x(t))$ is strong and that $B(t,x(t))^{-1}$ exists, and finally that

$$(45) \qquad \rho(B(t,x(t))^{-1}C(t,x(t))) < 1, \quad t \in J.$$

Then the numerical continuation process (8a,b) with G defined by (42)-(43) is feasible.

Proof. Since $B(t,x)$ is continuous and $B(t,x(t))^{-1}$ exists, a standard compactness argument on the curve x and an application of the perturbation lemma as in the proof of Theorem 2.15 gives a constant $r_0 > 0$ such that $B(t,x)^{-1}$ exists and $|B^0(t,x)^{-1}| \leq \beta$ for $(t,x) \in J \times D_0$ where

$$D_0 = \{x \mid \|x-x(t)\| \leq r_0 \text{ for some } t \in J\}.$$

Consider the identity

$$I - E(t,x)^m = [I-E(t,x)][I+\ldots+E(t,x)^{m-1}].$$

The condition (45) guarantees that $I - E(t,x)^m$ is nonsingular for $m \geq 1$. Hence

$$A(t,x) = B(t,x)[I+E(t,x)+\ldots+E(t,x)^{m-1}]^{-1}$$

is well-defined, continuous, and nonsingular on $J \times D_0$. Thus the mapping G defined by (42), which can be written

$$G(t,x) = x - A(t,x)^{-1}H(t,x)$$

is defined on $J \times D_0$ and by Lemma 2.14

$$G_x(t,x(t)) = I - A(t,x(t))^{-1}H_x(t,x(t))$$

$$= I - [I+E(t,x(t))+\ldots+E(t,x(t))^{m-1}]B(t,x(t))^{-1}H_x(t,x(t))$$

$$= E(t,x(t))^m.$$

Thus $\rho(G_x(t,x(t))) < 1$.

Since $B(t,x)^{-1}$ is bounded on $J \times D_0$ and B and C satisfy the Lipschitz condition corresponding to (44), A satisfies

$$\|A(t,x_1)-A(t,x_2)\| \leq K_1\|x_1-x_2\| \tag{46}$$

for some appropriate K_1. Thus Lemma 2.14 applies and $G_x(t,x(t))$ is strong. Thus by Theorem 2.8 the proof is complete.

So far we have considered only single-step local iterative processes for which x_{k+1} is determined as a function of x_k alone. It is similarly possible to consider numerical continuation

for multi-step processes such as, for instance, the interpolatory secant methods. The concept of a domain of attraction extends to these methods and one can phrase a basic feasibility theorem for numerical continuation analogous to Theorem 2.2. In the absence of a general point of attraction result for some of the multi-step methods such as the sequential secant method, the resulting theorems add relatively little to the present discussion, and we shall therefore not pursue here the multi-step question any further.

Appendix to Chapter II

In this appendix we prove the theorems quoted in Section 2.4.

2.10 Theorem. Define D_k by

$$(A.1) \qquad D_k = \{x \in C[0,1] \mid x(s) > -\frac{k}{2}, \ 0 \leq s \leq 1\},$$

let $f:(-\frac{k}{2},\infty) \subset R^1 \to R^1$ be given by

$$(A.2) \qquad f(s) = \frac{1}{\varepsilon}\frac{s}{s+k}, \ \varepsilon,k > 0,$$

and assume that $d:(-\frac{k}{2},\infty) \to R^1$ is a continuous function satisfying $d(s) \geq \alpha > 0$, $s \in [0,\infty)$. Then, for any fixed $t \in J$, every non-negative solution $x \in D_k$ of

$$(A.3) \quad x(s) = 1-t\int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi))d\xi du, \ s \in [0,1]$$

is a continuously differentiable solution of

$$(A.4) \qquad \begin{cases} (s^2 d(x)x')' = ts^2 f(x), \ 0 < s < 1 \\ x'(0) = 0, \ x(1) = 1 \end{cases}$$

and conversely, every continuously differentiable non-negative solution of (A.4) is a solution of (A.3).

Proof: Let x be a continuously differentiable non-negative solution of (A.4). Then integrating (A.4) from 0 to $u \in (0,1)$ and employing the boundary condition at 0 we obtain that

$$(A.5) \qquad u^2 d(x(u))x' = t\int_0^u \xi^2 f(x(\xi))d\xi,$$

and hence by integration for $s > 0$

$$(A.6) \qquad x(s) = 1-t\int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi))d\xi du.$$

Let $\bar{x} = \max_{s\in[0,1]} x(s)$ and $\bar{f} = \max_{s\in[0,\bar{x}]} f(s)$. Then, because of the existence of the lower bound on d, it follows that

$$0 \leq \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi))d\xi \leq \frac{\bar{f}}{\alpha}\,\frac{u}{3}.$$

Hence (A.6) is defined for $s = 0$ by taking the right-hand limit.

Conversely let $x \in C[0,1]$ be a non-negative solution of (A.3). Since the right-hand side is differentiable on $(0,1)$, it follows that

$$x'(s) = \frac{t}{s^2 d(x(s))} \int_0^s \xi^2 f(x(\xi))d\xi. \tag{A.7}$$

Then the limit for $s \to 0$ on the right exists and

$$0 \leq \lim_{s\to 0} \frac{1}{s^2 d(x(s))} \int_0^s \xi^2 f(x(\xi))d\xi \leq \lim_{s\to 0} \frac{1}{sd(x(s))}\, s^2 f(x(s)) = 0,$$

where we have used the fact that $\xi^2 f(x(\xi))$ is monotone increasing in ξ. Since the limit is bounded below by 0, it follows that $x'(0) = 0$.

From (A.7) now,

$$s^2 d(x(s))x'(s) = t \int_0^s \xi^2 f(x(\xi))d\xi \tag{A.8}$$

and since the right side is differentiable, so is the left and

$$(s^2 d(x(s))x'(s))' = ts^2 f(x(s)),\ s \in (0,1). \tag{A.9}$$

The remaining boundary condition $x(1) = 1$ is an immediate consequence of (A.3). This concludes the proof.

The existence theorem 2.11 is proved in Chapter IV. Here we discuss several facts about any given solution $x \in D_k$ of (A.3). As in Section 2.4 we write

$$(A.10) \qquad G(t,x) = 1 - t \int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi)) d\xi du, \quad t \in J.$$

Since $f(x(\xi)) \leq \frac{1}{\varepsilon}$ for $x \geq 0$, $x \in C[0,1]$, it follows that

$$t \int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi)) d\xi du \leq \frac{t}{\varepsilon} \int_s^1 \frac{1}{u^2 d(x(u))} \frac{u^3}{3} du$$

$$\leq \frac{t}{6\varepsilon \min\{d(x(u))\}} \leq \frac{t}{6\varepsilon\alpha},$$

and hence, for $6\varepsilon\alpha \geq 1$, we have $1 \geq G(t,x) \geq 0$, since the integrands are positive.

Therefore if $A = \{x \in C[0,1] \mid 0 \leq x(s) \leq 1, \ \forall s \in [0,1]\}$, then for each fixed $t \in J$, $G(t,A) \subset A$.

In order to show that (26) is indeed the Frechet derivative $G_x(t,x)$ of G, let $B: J \times D_k \subset J \times C[0,1] \to L(C[0,1],C[0,1])$, for any $h \in C[0,1]$, be defined by

$$(A.11) \qquad B(t,x)h = -t \int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f'(x(\xi))h(\xi) d\xi du$$

$$+ t \int_s^1 \frac{d'(x(u))h(u)}{u^2 d^2(x(u))} \int_0^u \xi^2 f(x(\xi)) d\xi du.$$

Then

$$G(t,x+h) - G(t,x) - B(t,x)h$$

$$= -t \int_s^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 [f(x(\xi)+h(\xi)) - f(x(\xi)) - f'(x(\xi))h(\xi)] d\xi du$$

$$- t \int_s^1 \frac{1}{u^2} \left[\frac{1}{d(x(u)+h(u))} - \frac{1}{d(x(u))} + \frac{d'(x(u))h(u)}{d(x(u)+h(u))d(x(u))} \right] \int_0^u \xi^2 f(x(\xi)+h(\xi)) d\xi du$$

$$+ t \int_s^1 \frac{d'(x(u))h(u)}{u^2 d(x(u))^2 d(x(u)+h(u))} [d(x(u)) - d(x(u)+h(u)] \int_0^u \xi^2 f(x(\xi)+h(\xi)) d\xi du$$

$$+ t \int_s^1 \frac{d'(x(u))h(u)}{u^2 d^2(x(u))} \int_0^u \xi^2 [f(x(\xi)+h(\xi)) - f(x(\xi))] d\xi du.$$

Hence, using the differentiability of f and d and the Lipschitz continuity of f, we obtain with certain constants $c_1,\ldots,c_4$,

$$\|G(t,x+h)-G(t,x)-B(t,x)h\|_\infty \leq c_1 \|f(x+h)-f(x)-f'(x)h\|_\infty$$

$$+ c_2 \|d(x+h)-d(x)-d'(x)h\|_\infty$$

$$+ c_3 \|d(x)-d(x+h)-d'(x)h\|_\infty + c_3 \|d'(x)\| \|h\|$$

$$+ c_4 \|f(x+h)-f(x)\|_\infty \leq \theta(\|h\|_\infty)\|h\|_\infty$$

where $\lim_{\|h\|_\infty \to 0} \theta(\|h\|_\infty) = 0$. Therefore

$$G_x(t,x) = B(t,x). \tag{A.12}$$

In order to bound $\|G_x(t,x(t))h\|_\infty$, we proceed as follows. Using the definition of f, (A.11) and (A.12), for $x \in D_k$, we clearly have

$$(A.13)\qquad |G_x(t,x)h(s)| \leq \left(\frac{tk}{\alpha\varepsilon}\right)|h|_\infty \int_s^1 \frac{1}{u^2} \int_0^u \xi^2 \frac{1}{(x(\xi)+k)^2}\, d\xi du$$

$$+ \frac{t\beta}{\varepsilon}|h|_\infty \int_s^1 \frac{1}{u^2} \int_0^u \xi^2 \frac{x(\xi)}{x(\xi)+k}\, d\xi du,$$

where $\beta \geq \max_{s\in[0,1]} |d'(s)/d^2(s)|$.

For fixed $t \in J$ let $x \in C[0,1]$ denote any non-negative solution of (A.6). Then, since $G(t,x) = x$, and we have shown that $0 \leq G(t,x) \leq 1$, it follows that also $0 \leq x \leq 1$.

From (A.7) now,

$$(A.14)\qquad x'(s) \geq 0,$$

which implies that

$$(A.15)\qquad x(s) \leq (1-c_t)s + c_t \equiv m_t(s)$$

where

$$c_t \geq x(0).$$

Also from (A.6), since $0 \leq f(x) \leq \frac{1}{\varepsilon}$

$$(A.16)\qquad x(s) \geq 1 - \frac{t}{\alpha\varepsilon} \int_s^1 \frac{1}{u^2} \int_0^u \xi^2 d\xi du$$

$$= 1 - \frac{t}{6\alpha\varepsilon}(1-s^2) \equiv \mu_t(s).$$

Since f is isotone and f' is antitone for positive values of their arguments, and since

(A.17) $$\mu_t(s) \leqslant x(s) \leqslant m_t(s),$$

it follows that

(A.18) $$f(\mu_t(s)) \leqslant f(x(s)) \leqslant f(m_t(s))$$

and analogously that

(A.19) $$f'(\mu_t(s)) \geqslant f'(x(s)) \geqslant f'(m_t(s)).$$

From (A.6), (A.18) and (A.15) we have

(A.20) $$x(s) \leqslant 1 - \frac{t}{\gamma\varepsilon} \int_s^1 \frac{1}{u^2} \int_0^u \xi^2 \frac{a+b\xi^2}{a+k+b\xi^2}\, d\xi du$$

where $b = \frac{t}{6\alpha\varepsilon}$, $a = 1 - b$. The double integral in (A.20) is equal to

$$\frac{t}{\gamma\varepsilon} \int_s^1 \frac{1}{u^2} \int_0^u \{ \xi^2 - \frac{k}{b} + \frac{(a+k)k}{b} \frac{1}{(a+k)+b\xi^2} \}\, d\xi du$$

$$= \frac{t}{\gamma\varepsilon} \int_s^1 \frac{1}{u^2} [\frac{u^3}{3} - \frac{ku}{b} + \frac{(a+k)k}{b} ((a+k)b)^{-1/2} \arctan u(\frac{b}{a+k})^{1/2}]du$$

$$= \frac{t}{\gamma\varepsilon} [\frac{1}{6} - \frac{s^2}{6} - \frac{k}{b} \ell n\, (1/s) + \frac{k}{b} \ell n\, (1/s)$$

$$- \frac{k}{2b} \ell n (\frac{a+k+b}{a+k+bs^2}) + \frac{k}{b} (\frac{b}{a+k})^{-1/2} \arctan (\frac{b}{a+k})^{1/2}$$

$$- \frac{k}{b} (\frac{b}{a+k})^{-1/2} s^{-1} \arctan (\frac{b}{a+k})^{1/2} s].$$

Thus,

$$(A.21)\qquad x(s) \leq 1 - \frac{t}{\gamma\varepsilon}\left\{\frac{1}{6} - \frac{s^2}{6} - \frac{k}{2b}\,\ell n\left(\frac{a+k+b}{a+k+bs^2}\right)\right.$$

$$+ \frac{k}{b}\left(\frac{b}{a+k}\right)^{-1/2}\arctan\left(\frac{b}{a+k}\right)^{1/2}$$

$$\left. - \frac{k}{b}\left(\frac{b}{a+k}\right)^{-1/2} s^{-1}\arctan\left(\frac{b}{a+k}\right)^{1/2} s\right\} .$$

Taking the limit as $s \to 0$, we see that

$$(A.22)\qquad x(0) \leq 1 - \frac{t}{\gamma\varepsilon}\left\{\frac{1}{6} - \frac{k}{2b}\,\ell n\left(1+\frac{b}{a+k}\right)\right.$$

$$\left. + \frac{k}{b}\left(\frac{b}{a+k}\right)^{-1/2}\arctan\left(\frac{b}{a+k}\right)^{1/2} - \frac{k}{b}\right\} .$$

Therefore we can take the bound in (A.22) for the constant c_t in (A.15).

Returning to (A.13) and employing (A.18), we obtain

$$\int_s^1 \frac{1}{u^2}\int_0^u \xi^2\,\frac{d\xi}{(x+k)^2}\,du \leq \int_s^1 \frac{1}{u^2}\int_0^u \frac{\xi^2}{[(a+k)+b\xi^2]^2}\,d\xi du$$

where a,b are as above. The last integral is equal to

$$(A.23)\qquad \int_s^1 \frac{1}{u^2}\left[-\frac{u}{2b(a+k+bu^2)} + \frac{1}{2b}\,(b(a+k))^{-1/2}\arctan u\left(\frac{b}{a+k}\right)^{1/2}\right]du$$

$$= -\frac{1}{2b(a+k)}\int_s^1 \frac{du}{u} + \frac{1}{2(a+k)}\int_s^1 \frac{udu}{a+k+bu^2}$$

$$+ \frac{1}{2b}\,((a+k)b)^{-1/2}\int_s^1 \frac{1}{u^2}\arctan u\left(\frac{b}{a+k}\right)^{1/2} du$$

$$= -\frac{1}{2b(a+k)}\left(\frac{b}{a+k}\right)^{-1/2}\arctan\left(\frac{b}{a+k}\right)^{1/2}$$

$$+ \frac{1}{2b(a+k)}\left(\frac{b}{a+k}\right)^{-1/2} s^{-1}\arctan s\left(\frac{b}{a+k}\right)^{1/2} .$$

Similarly,

$$(A.24)\qquad \int_s^1 \frac{1}{u^2}\int_0^u \xi^2 \frac{x}{x+k}\,d\xi du \le \int_s^1 \frac{1}{u^2}\int_0^u \xi^2 \frac{(1-c)\xi+c}{(1-c)\xi+c+k}\,d\xi du$$

$$= \int_s^1 \frac{1}{u^2}\left[\frac{u^3}{3} - \frac{k}{(1-c)}\frac{u^2}{2} + \frac{k(c+k)}{(1-c)^2}u - \frac{k(c+k)^2}{(1-c)^3}\ln[(1-c)u+c+k] + \frac{k(c+k)^2}{(1-c)^3}\ln(c+k)\right.$$

$$= \frac{1}{6} - \frac{s^2}{6} - \frac{k}{2(1-c)} + \frac{ks}{2(1-c)} - \frac{k(c+k)}{(1-c)^2}\ln s + \frac{k(c+k)^2}{(1-c)^3}\ln(1+k) - \frac{k(c+k)^2}{(1-c)^3}\frac{\ln[(1-c)s+c+k]}{s}$$

$$+ \frac{k(c+k)}{(1-c)^2}\ln s + \frac{k(c+k)}{(1-c)^2}\ln(1+k) - \frac{k(c+k)}{(1-c)^2}\ln[(1-c)s+c+k] - \frac{k(c+k)^2}{(1-c)^3}\ln(c+k) + \frac{k(c+k)^2}{(1-c)^3}\frac{\ln(c+k)}{s}.$$

Now since the integrands in (A.23) and (A.24) are positive, the maximum over s is taken at $s = 0$.

In (A.24) we must consider the limit

$$\lim_{s\to 0} \frac{1}{s}\ln(c+k) - \frac{1}{s}\ln[(1-c)s+c+k]$$

which, by L'hospital's rule is equal to

$$\lim_{s\to 0} \frac{1}{s}\ln\frac{c+k}{(1-c)s+c+k} = -\frac{1-c}{c+k}.$$

Thus (A.24) becomes in the limit as $s \to 0$

$$\frac{1}{6} - \frac{k}{2(1-c)} + \frac{k(c+k)^2}{(1-c)^3} \ln (1+k) + \frac{k(c+k)}{(1-c)^2} \ln (1+k)$$

$$- \frac{k(c+k)}{(1-c)^2} \ln (c+k) - \frac{k(c+k)^2}{(1-c)^3} \ln (c+k)$$

$$= \frac{1}{6} - \frac{k}{2(1-c)} - \left(\frac{1-c}{c+k} \right) + \frac{k(c+k)(1+k)}{(1-c)^3} \ln \left(\frac{1+k}{c+k} \right).$$

Taking the limit as $s \to 0$ in (A.23) and recalling (A.13) we finally have

$$\|G_x(t,x)h\|_\infty \leq \left(\frac{tk}{\alpha\varepsilon} \right) \|h\|_\infty \frac{1}{2b(a+k)} \left[1 - \left(\frac{b}{a+k} \right)^{-1/2} \arctan \left(\frac{b}{a+k} \right)^{1/2}\right] \tag{A.25}$$

$$+ \frac{\beta t}{\varepsilon} \|h\|_\infty \left\{\frac{1}{6} - \frac{k}{2(1-c)} - \left(\frac{1-c}{c+k} \right) \right.$$

$$\left. + \frac{k(c+k)(1+k)}{(1-c)^3} \ln \left(\frac{1+k}{c+k} \right)\right\}$$

which yields the norm bound for $\|G_x(t,x)\|_\infty$ given in equation (2.27).

CHAPTER III

Path Existence Theorems Using Continuation Methods

Section 1. The Continuation Property and Path Existence

In Chapter II we separated the question of the existence of the curve x along which numerical continuation takes place from the feasibility of the numerical continuation process itself. In this chapter we direct our attention to the existence of the curve. Clearly, it would be impossible to cover here to any meaningful degree the multitude of possible existence theorems; the literature in that area is simply too voluminous. Instead, we present a survey of some of the typical and important results from various sources that relate to the existence of a solution curve in the setting we are considering here. At the same time some of the proofs appear to have novel aspects or represent improvements of the original proofs. More importantly, by a consistent use of a continuation property introduced by Rheinboldt [1969] we are able to provide a certain unification of several approaches that previously seemed to be rather unrelated.

As we discussed in Chapter I, historically, the numerical techniques for proceeding from one endpoint of the solution curve to the other have either been the numerical continuation process, as we considered it in the preceding chapter, or the numerical integration of some underlying differential equation. For the existence proof of the curve there are basically the same two approaches.

The most widely used method for proving existence of a solution

curve appears to have been to consider an underlying differential equation for the curve--provided such an equation exists--and to utilize the standard existence and continuation results of the theory of ordinary differential equations in Banach space.

The other method for guaranteeing existence of the curve is again based on a continuation argument not unlike that used in the processes of Chapter II. Let $H(t,x) = 0$ be the homotopy equation and x_0 an initial point for which $H(0,x_0) = 0$. In the domain under consideration $H(t,\cdot)$ is often known to possess some local solvability property such as, for instance, that given by the usual implicit function theorem. Thus a continuous solution curve x of $H(t,x) = 0$ exists for t in some small interval $[0,t_1)$ with $t_1 \in (0,1]$. Then again using the local solvability property at $(t_1-\varepsilon,x(t_1-\varepsilon))$ we can ensure the existence of x on a larger interval $[0,t_2)$, $t_2 > t_1$, and so on. The problem is now that the local solvability domains may get progressively smaller and hence that continuation up to $t = 1$ might not be possible. For a treatment of this continuation approach in a general setting, see Ficken [1951] and the references mentioned there as well as Rheinboldt [1969].

As we have already mentioned above, underlying the continuation approach is always a local solvability condition which can often be obtained by the use of the implicit function theorem. The following version of this theorem, involving strong derivatives, was given by Ortega and Rheinboldt [1970]. As before, X denotes a Banach space unless otherwise indicated.

3.1 Theorem. Suppose that $H:J \times D \subset J \times X \to X$ is continuous on a neighborhood $I_0 \times D_0$ of a point (t_0,x_0), where I_0 is relatively open in J and D_0 is open in X, and that $H(t_0,x_0) = 0$. Assume that H_x exists and is strong at (t_0,x_0) and that $H_x(t_0,x_0)$ has a bounded inverse. Then there exist a relatively open neighborhood $I \subset I_0$ of t_0 and an open neighborhood $S \subset D_0$ of x_0, such that for any $t \in \bar{I}$, the equation $H(t,x) = 0$ has a unique solution $x = \phi(t) \in \bar{S}$ and the mapping $\phi:I \to X$ is continuous. Moreover, if H_t exists at (t_0,x_0), then ϕ is differentiable at t_0 and

$$\phi'(t_0) = -H_x(t_0,x_0)^{-1}H_t(t_0,x_0).$$

In Ortega and Rheinboldt the theorem was phrased in finite dimensional spaces and for an open neighborhood of (t_0,x_0) in $J \times X$. However, the same proof carries over to Banach spaces and for the relatively open neighborhood which we have used.

Let the mapping $H:J \times D \subset J \times X \to X$ and the homotopy equation

$$H(t,x) = 0 \tag{1}$$

be given with the initial point $(0,x_0)$ such that $H(0,x_0) = 0$. As described above the continuation approach for H need not yield a solution curve for the entire t-interval J; this is shown by the following simple example.

3.2 Example. Let $H:J \times R^1 \to R^1$ be defined by

$$H(t,x) = [x \cos \tfrac{\pi}{2}(t-1) - \sin \tfrac{\pi}{2}(t-1)][x \cos \tfrac{\pi}{2} t - \sin \tfrac{\pi}{2} t].$$

For $x_0 = (0,0)^T$ as initial point, the solution curve $x:[0,1) \to R^1$ is then defined by

$$x(t) = \tan \tfrac{\pi}{2} t.$$

Obviously x cannot be continued to the closed unit interval, although locally a continuous solution exists for each $t \in J$.

The following concept, introduced by Rheinboldt [1969] in a more general form, will enable us to prove for certain mappings H the existence of the solution curve x on all of J.

3.3 Definition. The mapping $H:J \times D \subset J \times X \to X$ has the continuation property with respect to $x \in D$ if the existence of a continuous function $x:[0,t^*) \subset J \to D$, such that $H(t,x(t)) = 0$ for all $t \in [0,t^*)$ implies that $\lim_{t \to t^*-} x(t) = \hat{x}$ exists, $\hat{x} \in D$ and $H(t^*,\hat{x}) = 0$.

With this definition, we can prove the following special form of a theorem of Rheinboldt [1969].

3.4 Theorem. Let $H:J \times D \subset J \times X \to X$ be continuous on $J \times D$ where D is open, and assume that H has a strong partial Frechet derivative H_x on $J \times D$ (or that H_x is continuous on $J \times D$). If $H_x(t,x)^{-1}$ exists on $J \times D$ and H has the continuation property with respect to x, then for any initial point $(0,x_0)$ such that $H(0,x_0) = 0$, there

exists a unique and continuous curve $x:J \to D$ satisfying (1) for each $t \in J$.

Proof. Since $H_x(0,x_0)$ exists and is strong (if H_x is continuous, the strongness follows by Lemma 2.7), Theorem 3.1 ensures the existence of a continuous solution curve for $t \in [0,t_1]$ with $x(0) = x_0$. If $t_1 < 1$, then, as discussed above, the same argument can be applied at $(t_1,x(t_1))$ and the curve can be extended to some $[0,t_2] \subset J$ with $t_2 > t_1$, etc. Let $[0,t^*) \subset J$ be the largest interval up to which the solution curve x starting from x_0 can be continuously extended in this way. Clearly then, $t^* > 0$ and by the continuation property, $\lim_{t \to t^*-} x(t) = \hat{x}$ exists, $\hat{x} \in D$ and $H(t^*,\hat{x}) = 0$. Thus we can set $x(t^*) = \hat{x}$ and x is well-defined on $[0,t^*] \subset J$.

If $t^* = 1$, then the existence proof is complete. Suppose therefore that $t^* < 1$. Then by Theorem 3.1 there exists a $\delta > 0$ and a neighborhood $S \subset D$ of $x(t^*)$ such that on $I = \{t \in J \mid |t-t^*| < \delta\}$ a unique continuous function $\phi:I \to S$ exists and satisfies $\phi(t^*) = x(t^*)$. By uniqueness, ϕ must agree with x on $(t^*-\delta,t^*]$. But then $\bar{x}:[0,t^*+\delta) \to D$ defined by

$$\bar{x}(t) = \begin{cases} x(t) \text{ for } t \in [0,t^*] \\ \\ \phi(t) \text{ for } t \in (t^*,t^*+\delta) \end{cases}$$

is a continuous extension of x which contradicts the maximality of t^* unless $t^* = 1$.

To prove uniqueness let $x:J \to D$, $\hat{x}:J \to D$ be continuous solutions of (1) such that $x(0) = \hat{x}(0) = x_0$. Then $\hat{t} = \sup \{t \in J \mid x(s)=\hat{x}(s) \text{ for } s \in [0,t]\} > 0$ and by continuity $x(\hat{t}) = \hat{x}(\hat{t})$. If $\hat{t} < 1$, then by Theorem 3.1 applied to the point $(\hat{t},x(\hat{t}))$, the solution x of (1) through $x(\hat{t})$ is unique on some interval $(\hat{t}-\delta,\hat{t}+\delta)$, $\delta > 0$ which contradicts the definition of $\hat{t}$. Thus the solution of (1) starting from x_0 is unique on J and the proof is complete.

Next we prove a result of Ostrowski [1966, p. 177] which forms the basis of a theorem due to Yakovlev to be considered in the next section. By utilizing Theorem 3.4 we can shorten Ostrowski's proof somewhat although the proof remains essentially the same.

3.5 Corollary. Let $F:D \subset X \to X$ be continuously Frechet differentiable on the ball $S = \bar{S}(x_0,r)$ in the open set D and assume that

i) $F'(x)^{-1}$ exists and $\|F'(x)^{-1}\| \leq d$, $\forall x \in S$;

ii) $r > d\|F(x_0)\|$.

Then $F(x) = 0$ has at least one solution $x^* \in \bar{S}(x_0,d\|F(x_0)\|)$.

Proof. Consider $H:J \times D \to X$ defined by

$$H(t,x) = F(x) + (t-1)F(x_0). \tag{2}$$

Clearly, H_x exists and is continuous on $J \times S$. Therefore, in order to apply Theorem 3.4 we need to prove the continuation property for H. Suppose that $x:[0,t^*) \subset J \to D$, $t^* > 0$, exists such that $H(t,x(t)) = 0$ for $t \in [0,t^*)$ and $x(0) = x_0$. Then by Theorem 3.1,

x is differentiable on $[0,t^*)$ and

$$x'(t) = -F'(x(t))^{-1}F(x_0), \quad \forall t \in [0,t^*). \tag{3}$$

Hence by the integral mean value theorem, for $0 < t' \leq t'' < t^*$,

$$\begin{aligned} \|x(t')-x(t'')\| &\leq \int_{t'}^{t''} \|F(x(s))^{-1}F(x_0)\|ds \\ &\leq d\|F(x_0)\|\,|t'-t''|. \end{aligned} \tag{4}$$

Thus for any increasing sequence $\{t_k\}$ with $t_k < t^*$ and $\lim_{k\to\infty} t_k = t^*$, (4) shows that $\{x(t_k)\}$ is a Cauchy sequence and hence has a limit $\hat{x}$. Moreover, again by (4),

$$\|\hat{x}-x_0\| \leq d\|F(x_0)\|t^* \leq d\|F(x_0)\| < r, \tag{5}$$

and hence $\hat{x} \in S$. By continuity of H, therefore, $H(t^*,\hat{x}) = 0$.

We must still show that $\hat{x}$ is independent of the sequence $\{t_k\}$. Let $\{s_k\}$ be any other sequence $\lim_{k\to\infty} s_k = t^*$, $s_k < t^*$. Then we have

$$\|\hat{x}-x(s_k)\| \leq \|\hat{x}-x(t_k)\| + \|x(t_k)-x(s_k)\|. \tag{6}$$

Given $\varepsilon > 0$, choose j_0 and k_0 so that $|s_k-t_0|$ and $|t_k-t_0|$ are each less than $\varepsilon/(2d\|F(x_0)\|)$ and $k_1 \geq k_0$ so that $\|\hat{x}-x(t_k)\| < \varepsilon/2$ for $k \geq k_1$. Then by (4) and (6), $\|\hat{x}-x(s_k)\| < \varepsilon$ for $j \geq j_0$. Since ε was arbitrary, $\hat{x} = \lim_{k\to\infty} x(s_k)$, and hence $\lim_{t\to t^*-} x(t) = \hat{x}$; that is, H has the continuation property. Theorem 3.4 therefore applies and the proof is complete.

In the case when $H(t,\cdot) = 0$ has a unique solution on its domain

then the entire problem simplifies considerably. In fact, we shall see that this is the case for the model problem introduced in Chapter II.

3.6 Theorem. Assume that $H:J \times D \subset J \times X \to X$ is continuous on $J \times D$ where D is open and that for each $t \in J$ a solution $x(t)$ of (1) exists and is unique. If for any $t \in J$ the strong derivative H_x exists at $(t,x(t))$ and has a bounded inverse, then $x:J \to D$ is continuous.

Proof. By Theorem 3.1, x is continuous on $[0,t_0) \subset J$ for some $t_0 > 0$. Assume that x is continuous on $[0,t^*) \subset J$ where t^* has been chosen maximal. Then applying Theorem 3.1 at $(t^*,x(t^*))$ we obtain for some $\delta > 0$ a unique function $\phi:(t^*-\delta,t^*+\delta) \cap J \to D$, such that $\phi(t^*) = x(t^*)$ and $H(t,\phi(t)) = 0$ for $t \in (t^*-\delta, t^*+\delta)$. Then by the uniqueness of x, $\phi(t) = x(t)$ for $t \in (t^*-\delta,t^*)$ and

$$\lim_{t \to t^*-} x(t) = \lim_{t \to t^*-} \phi(t) = \phi(t^*) = x(t^*).$$

If $t^* < 1$, then $x(t) = \phi(t)$ for $t \in [t^*,t^*+\delta)$ defines a continuous extension of x which contradicts the maximality of t^*. Then $t^* = 1$ and the proof is complete.

Section 2. The Yakovlev Theorem and Some of Its Applications

On the basis of Corollary 3.5, Yakovlev [1965] phrased an existence theorem for a solution curve x satisfying (1). As we mentioned in Chapter I, Yakovlev's theorem is correct, but his proof is incomplete since he concluded from the local unique existence and continuity of a

solution for each $t \in J$ that this automatically implies the existence of a unique continuous solution curve, through $x(0) = x_0$, on all of J. Example 3.2 shows that this need not be the case.

We phrase Yakovlev's result in a Banach space setting and accordingly assume H_x to be uniformly continuous rather than only continuous.

3.7 Theorem. Let $H:J \times D \subset J \times X \to X$ be continuous on the ball $S_0 = \bar{S}(x_0,r) \subset D$ and assume that H has a Frechet derivative with respect to x on $J \times S$ and that H_x is uniformly continuous there. Suppose further that

i) $H_x(t,x)^{-1}$ exists and $\|H_x(t,x)^{-1}\| \leq d$, $\forall (t,x) \in J \times S$;

ii) $r > r_0 = \max \{d\|H(t,x_0)\| \mid t \in J\}$;

iii) $H(0,x_0) = 0$.

Then there exists a unique and continuous curve $x:J \to \bar{S}(x_0,r_0)$ such that $H(t,x(t)) = 0$, $\forall t \in J$ and $x(0) = x_0$.

Furthermore, if the Frechet derivative H_t exists on $J \times S$, then x is differentiable and satisfies

$$x'(t) = -H_x(t,x(t))^{-1}H_t(t,x(t)),\ x(0) = x_0,\ t \in J. \tag{7}$$

Proof. Clearly, Corollary 3.5 applies to $H(t,\cdot): S = \bar{S}(x_0,r) \to X$ and hence for any $\hat{t} \in J$ there exists an $\hat{x} \in S$ such that $H(\hat{t},\hat{x}) = 0$. Set $\hat{A} = H_x(\hat{t},\hat{x})$ and $G:J \times S \to X$, $G(t,x) = x - \hat{A}^{-1}H(t,x)$.

By the uniform continuity of H_x on $J \times S$ there exist $\delta > 0$, $\eta > 0$ for which

$$\|H_x(t,x)-H_x(s,y)\| < \frac{1}{2d}\ ,\ \forall s,t \in J,\ |s-t| \leq \delta,\quad \forall x,y \in S_0,$$

$$\|x-y\| \leq \eta.$$

Let $\hat{\eta} \in (0,\frac{\eta}{2})$ be such that $\hat{S} = \bar{S}(\hat{x},\hat{y}) \subset S$. Then for $x,y \in \hat{S}$ and all $t \in \hat{I} = \{s \in J \mid |s-\hat{t}| \leq \delta\}$

$$\|G(t,y)-G(t,x)\| \leq \|\hat{A}^{-1}\| \left\| \int_0^1 [\hat{A}^{-1}-H_x(t,x+s(y-x))](y-x)ds \right|$$

(8)

$$\leq \|\hat{A}^{-1}\|[\frac{1}{2d}\|y-x\|] \leq \frac{1}{2}\|y-x\|$$

and, in particular,

$$\|G(t,y)-\hat{x}\| = \|G(t,y)-G(\hat{t},\hat{x})\| \leq \frac{1}{2}\|y-\hat{x}\|,\quad \forall y \in \hat{S},\ t \in \hat{I}.$$

Thus by the contraction mapping theorem, $G(t,\cdot)$ has a unique fixed point $x(t) \in \hat{S}$ for any $t \in \hat{I}$. Moreover, the resulting curve $x:\hat{I} \to \hat{S}$ is continuous; this follows from the continuity of G since by (8), for $t,\ t_0 \in \hat{I}$,

$$\|x(t)-x(t_0)\| \leq \|G(t,x(t))-G(t,x(t_0))\| + \|G(t,x(t_0))-x(t_0)\|$$

or

$$\|x(t)-x(t_0)\| \leq 2\|G(t,x(t_0))-x(t_0)\|.$$

Finally, if $y:\hat{I} \to S$, $H(t,y(t)) = 0$, $y(\hat{t}) = x(\hat{t}) = \hat{x}$ is any other continuous solution, then $x(t) = y(t)$ for $t \in \hat{I}$. This is a direct consequence of the contraction condition (8) since $y(t) \in \hat{S}$ for $t \in \hat{I}$.

In fact, y can never reach the boundary of $\hat{S}$, for $\|y(s)-\hat{x}\| = \eta$ for some $s \in \hat{I}$ would by (8) lead to the contradiction $\hat{\eta} \leq \hat{\eta}/2$.

Now the continuation condition follows readily. Suppose that a continuous curve $x:[0,t^*) \subset J \to S$, $H(t,x(t)) = 0$, $t \in [0,t^*)$, $x(0) = x_0$ has been given. By Theorem 3.1, clearly, $t^* > 0$, and hence we can choose $\hat{t} \in [0,t^*)$ with $|\hat{t}-t^*| \leq \frac{\delta}{2}$. From the first part of the proof now follows the existence of a continuous $\phi:\hat{I} \to \bar{S}(x(\hat{t}),\hat{\eta})$ for which $H(t,\phi(t)) = 0$ for $t \in \hat{I}$ and $\phi(\hat{t}) = x(\hat{t})$. Moreover, the uniqueness statement proved above ensures that $x^* = \lim_{t\to t^*-} x(t) = \lim_{t\to t^*-} \phi(t) = \phi(t^*) \subset \bar{S}(x(\hat{t}),\hat{\eta}) \subset S$ exists with $H(t^*,x^*) = H(t^*,\phi(t^*)) = 0$. This is the continuation property and the first part of the theorem follows from Theorem 3.4. The second part is a direct consequence of the last statement of Theorem 3.1.

The proof of Theorem 3.7 essentially shows that the local solvability domains guaranteed by the Implicit Function Theorem 3.1 have a uniform lower bound and hence that they cannot "shrink to zero" as $t \to 1$. This is a special case of a "covering" property considered by Rheinboldt [1969] in a much more general setting.

We next consider a result due originally to Hadamard for which proofs have already been obtained by both the continuation argument considered here (see Rheinboldt [1969]) as well as by the differential equations approach mentioned at the beginning of Section 1 (see Meyer [1968]). We shall apply this result later to guarantee the existence of solution curves in special cases. The proof here is based on Yakovlev's theorem and appears to have some new aspects.

Although the theorem also holds in a general Banach space setting, we have phrased it only in R^n since this is the setting in which we will apply it. Similarly, R^n is used in the results of the remainder of this section.

3.8 Corollary. Let $F:R^n \to R^n$ be continuously Frechet differentiable on R^n and suppose that $\|F'(x)^{-1}\| \leq d$ for all $x \in R^n$. Then F is a homeomorphism from R^n onto R^n.

Proof. Let $z \in R^n$ be given and, with arbitrary $q \in R^n$, consider the homotopy $H:J \times R^n \to R^n$ and the resulting equation (1) defined by

$$H(t,x) \equiv F(x) - z + (t-1)[F(q)-z] = 0. \tag{9}$$

By Theorem 3.7 a continuous solution curve $x:J \to R^n$, $x(0) = q$ exists and, from (9), $F(x(1)) = z$. Since z was arbitrary, this shows that F is onto. The uniqueness is more difficult to prove.

Assume that x_1 and x_2 are each solutions of $F(x) = z$ and let $q(s) = sx_1 + (1-s)x_2$. We consider the homotopy equation (9) with q replaced by $q(s)$, $s \in J$. By Theorem 3.7 there exists for any fixed $s \in J$ a continuously differentiable solution curve $x_s = x(\cdot,s):J \to R^n$ such that $x(0,s) = q(s)$ and

$$\frac{d}{dt}\, x(t,s) = -F'(x(t,s))^{-1}[F(q(s))-z], \quad \forall t \in J. \tag{10}$$

Moreover,

$$x(t,s) \in \bar{S}(q(s),d\|F(q(s))-z\|), \quad \forall t,s \in J. \tag{11}$$

From (11) we find that $x(\cdot,s)$ is uniformly bounded and from (17) that $x(\cdot,s)$ satisfies a uniform Lipschitz condition for all $s \in [0,1]$. Hence by Ascoli's theorem the set $A = \{x(\cdot,s) \mid s\in[0,1]\}$ is compact in $C[0,1]$.

Let $\{s_k\} \subset [0,1]$ be any sequence such that $\lim_{k\to\infty} s_k = \hat{s} \in J$. The sequence $\{x(\cdot,s_k)\} \subset A$ has a uniformly convergent subsequence $\{x(\cdot,s_{k_j})\}$; let $\hat{x} = \hat{x}(t)$ denote its limit. By continuity of H, $H(t,\hat{x}(t)) = 0$ and by continuity of q, $\hat{x}(0) = q(\hat{s})$. But by Theorem 3.7, $x(\cdot,\hat{s})$ is the unique continuous solution of (9) such that $x(0,\hat{s}) = q(\hat{s})$, hence $\hat{x}(t) = x(t,\hat{s})$. By the same argument, every subsequence of $\{x(\cdot,s_k)\}$ has a uniformly convergent subsequence with limit $x(t,\hat{s})$, hence $\lim_{k\to\infty} x(t,s_k) = x(t,\hat{s})$ for all $t \in J$, which implies that for each fixed t, $x(t,\cdot):J \to R^n$ is continuous as a function of s.

From (9) we obtain

$$F(x(t,s)) = z + (1-t)[F(q(s))-z] \tag{12}$$

and in particular for $s = 0$, $x(t,0)$ is a solution of

$$F(x) = z \tag{13}$$

for all $t \in J$. By the standard inverse function theorem, $x(0,0) = x_2$ is the unique solution of (13) in a neighborhood of x_2. The continuity of $x(\cdot,0)$ then implies that $x(t,0) \equiv x_2$ for all $t \in J$. Similarly for $s = 1$ we find that $x(t,1) \equiv x_1$. Finally, for $t = 1$, $x(1,\cdot)$ is a

continuous solution of (20) with $x(1,0) = x_2$ and again the inverse function theorem implies that $x(1,t) \equiv x_2$.

Now by continuity

$$x_2 = \lim_{t\to 1} x(1,t) = \lim_{s\to 1} x(s,1) = x_1$$

and we have uniqueness, and the proof is complete.

As an example of the application of Corollary 3.8, we give a simple result about certain mildly nonlinear equations to be considered in Chapter IV.

3.9 Corollary. Let $A \in L(R^n,R^n)$ be an M-matrix and $B:R^n \to R^n$ a continuously differentiable mapping for which $B'(x)$ is a non-negative diagonal matrix for each $x \in R^n$. Then for $H:J \times R^n \to R^n$, $H(t,x) = Ax + tB(x) - b$, with fixed $b \in R^n$, there exists a unique continuous solution curve x of the equation $H(t,x) = 0$ such that $x(0) = x_0 = A^{-1}b$.

Proof. Clearly H is continuously differentiable and for each $t \in J$, $H_x(t,x) = A + tB'(x) \geq A$ is again an M-matrix (see Varga [1962; p. 84]). Thus, $0 \leq H_x(t,x)^{-1} \leq A^{-1}$ for all $x \in R^n$ and hence $\|H_x(t,x)^{-1}\| \leq \|A^{-1}\|$ under any monotonic norm. By Corollary 3.8, $H(t,\cdot)$ is a homeomorphism from R^n onto R^n. Now apply Theorem 3.6 and the proof is complete.

We next consider a result which was originally obtained from an underlying differential equation for the solution curve. Our proof shows that the theorem can be obtained as well from the continua-

tion arguments utilized here. The result is due to Gavurin [1958] and is representative for a class of results which arise in considering continuous analogs of iterative processes in the following sense:

Given an iterative process of the form

$$x_{n+1} = x_n - G(x_n),\ x_0 \text{ given}, \tag{14}$$

Gavurin and other authors thereafter (see Polyak [1964], Bittner [1967], Kleinmichel [1968], and Bosarge [1968]) regard (14) as the finite difference approximation of

$$x'(t) = -G(x(t)),\ x(0) = x_0,\ t \quad [0,\infty), \tag{15}$$

which arises when at $t_n = n$, $n = 0,1,\ldots$, $x'(t)$ is replaced by $\frac{1}{h}[x_{n+1}-x_n]$. Correspondingly these authors consider (15) as the continuous analog of the iterative process (14). The topic of these continuous processes is, of course, outside the scope of this paper, but the existence of a solution curve $x:[0,\infty) \to R^n$ of (15) is of some interest to us as another type of existence theorem for solution curves.

Gavurin's result considers the continuous analog of Newton's method.

3.10 Theorem. Let $F:D \subset R^n \to R^n$ be continuously differentiable on the ball $S = \bar{S}(x_0,r)$ contained in the open set D and suppose that

$$\|F'(x)^{-1}\| < d, \quad \forall x \in \bar{S}(x_0,r) \tag{16}$$

and

$$r > d\|F(x_0)\|. \tag{17}$$

Then the equation

$$x' = -F'(x)^{-1}F(x),\ x(0) = x_0 \tag{18}$$

has a solution $x = x(t) \in \bar{S}(x_0, d\|F(x_0)\|)$, $t \in [0,\infty)$, and $\lim_{t\to\infty} x(t) = x^*$ exists and is a solution of $F(x) = 0$.

Proof. By Theorem 3.7 there exists a unique continuous solution $y: J \to \bar{S}(x_0, d\|F(x_0)\|)$ of

$$H(t,y) = F(y) + (t-1)F(x_0) = 0 \tag{19}$$

with $y(0) = x_0$. Moreover, y is differentiable and $F'(y(t))y'(t) + F(x_0) = 0$. Let $(1-t) = e^{-s}$ and set $x(s) = y(\ell n(1-t)^{-1})$, then (19) is equivalent to

$$H(s,x(s)) = F(x(s)) - e^{-s}F(x_0),\ s \in [0,\infty),$$

and by the differentiability of y and hence of x,

$$\frac{d}{ds}\ x(s) = -e^{-s}F'(x(s))^{-1}F(x_0)$$

$$= -F'(x(s))^{-1}F(x(s)).$$

Thus x satisfies (18) and the proof is complete.

Section 3. Application to the Model Problem

Recall the model problem:

$$x = G(t,x) \tag{20}$$

where $x \in D_k = \{x \in C[0,1] \mid x(s) \geq -\frac{k}{2},\ 0 \leq s \leq 1\}$ and $G: J \times D_k \subset J \times C[0,1] \to C[0,1]$ is defined by

$$G(t,x)(s) = 1 - t\int_0^1 \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi))\,d\xi\,du \tag{21}$$

with

$$f:(-\tfrac{k}{2}, \infty) \to R^1,\ f(s) = \frac{1}{\varepsilon}\,\frac{s}{s+k},\ \varepsilon,k > 0 \tag{22}$$

and a differentiable function $d:(-\frac{k}{2}, \infty) \to R^1$ for $d(s) \geq \alpha > 0$, for all $s \in (-\frac{k}{2}, \infty)$.

We state the following existence theorem.

3.11 Theorem. Consider the problem (20) where G is given by (21), f by (22), and $d:(-\frac{k}{2}, \infty) \to R^1$ has the stated properties. If $6\varepsilon\alpha \geq 1$, then for each $t \in J$ a solution $x(t)$ of (11) with $x(0) = 1$ exists. Furthermore, for $\max_{0 \leq s \leq 1} |d'(s)|/d^2(s)$ and k sufficiently small, the curve $x: J \to D_k$, with $x(0) \equiv 1$, is unique and continuous.

Proof. Let $A = \{x \in C[0,1] \mid 0 \leq x(s) \leq 1 \text{ for } s \in [0,1]\}$. In the Appendix to Chapter II we have seen that if $6\varepsilon\alpha \geq 1$ then for each $t \in J$, $G(t,\cdot)$ maps A into itself. For fixed $t \in J$, $G(t,A)$ is compact. In fact, since $G(t,A) \subset A$ and A is bounded, $G(t,A)$ is a uniformly bounded family of functions in $C[0,1]$. Moreover, since

$0 \leq f(x(s)) \leq \frac{1}{\varepsilon}$ and $0 \leq \frac{1}{d(x(s))} \leq \frac{1}{\alpha}$, for $s \in J$, we have for $s_1, s_2 \in J$, $s_1 < s_2$, and any $x \in A$,

$$|G(t,x)(s_1)-G(t,x)(s_2)| = t \int_{s_1}^{s_2} \frac{1}{u^2 d(x(u))} \int_0^u \xi^2 f(x(\xi))\,d\xi\,du$$

$$\leq \frac{t}{\alpha\varepsilon} \int_{s_1}^{s_2} \frac{1}{u^2} \int_0^u \xi^2\,d\xi\,du$$

$$\leq \frac{t}{6\alpha\varepsilon} \; |s_2-s_1|$$

which shows that $G(t,A)$ is an equi-continuous family of functions. Thus by Ascoli's theorem, $G(t,A)$ is indeed compact and $G(t,\cdot)$ maps the closed bounded convex set A into a compact subset of A. Now by a well-known form of Schauder's fixed point theorem (see Dunford and Schwartz [1958; p. 456]) $G(t,\cdot)$ has a fixed point $x(t) \in A$ for any $t \in J$.

We will use Theorem 3.6 to prove that the mapping $x:J \to A$ is continuous. In the appendix to Chapter II we showed that for each fixed $t \in J$, any solution $x(t)$ of (20) belongs to the set

$$S_t = \{x \in D_k \,|\, m_t(s) \leq x(s) \leq \mu_t(s)\}$$

where m_t and μ_t are defined by equations (A.15) and (A.16), respectively. Furthermore, from (A.25) it follows that for fixed t and for $\max_{0 \leq s \leq 1} |d'(s)|/d^2(s)$ and k sufficiently small such that $\|G_x(t,x)\|_\infty < 1$ for all $x \in S_t$. Therefore under these conditions, G is non-expansive on S_t and for each t the solution $x(t)$ is unique.

Moreover, if now $H: J \times D_k \to C[0,1]$ is defined by

$$H(t,x) = x - G(t,x), \tag{23}$$

then for the same conditions on d and k, it follows by Lemma 2.12 that for each $t \in J$, $H_x(t,x(t))^{-1}$ exists. Since H_x is continuous on $J \times D_k$ by the continuity of G_x, Theorem 3.6 now proves that $x: J \to A \subset D_k$ is continuous.

CHAPTER IV

Steplength Estimation

In Chapter II we dealt with the question of the feasibility of numerical continuation. The theorems proved have primarily theoretical value; they are not in general useful as a means of estimating steplengths in the parameter variable or the size of the convergence domains. In this chapter we examine some aspects of the difficult problem of estimating these steplengths and convergence domains.

Section 1. Estimation from Local Convergence Theorems

In the process of numerical continuation described in equations (2.8a,b) there are two quantities that must be estimated. The first of these is the parameter steplength $t_k - t_{k-1}$ for $k = 1,\ldots,N$, and the second is the choice of the integers j_k which determine how many iterates are to be taken at $t = t_k$. The following theorem shows that we can restrict our attention to the first of these quantities provided we retain the basic assumptions under which we could guarantee feasibility.

Throughout this chapter X denotes a Banach space.

4.1 Theorem. Consider the mapping $G:J \times D \subset J \times X \to X$, where D is open, and let $x:J \to D$ be a continuous solution of the fixed point equation

$$x = G(t,x).$$

If, for each $t \in J$, the partial derivative G_x of G exists at $(t,x(t))$ and is strong and if $\|G_x(t,x(t))\| < 1$ for all $t \in J$, then numerical continuation is feasible with fixed steplength $t_k - t_{k-1} = \Delta t$ and $j_k = 1$, for all $k = 1,\ldots,N$.

Proof. The proof represents a certain refinement of that of Theorem 2.8. Observe first that by the openness of D there exists an $r_0 > 0$ such that

$$(1) \qquad D_0 = \{x \in X \mid \|x-x(t)\| \leq r_0 \text{ for some } t \in J\} \subset D.$$

Given $t' \in J$ let $d_{t'} = \|G_x(t',x(t'))\|$ and choose $\eta_{t'}$ such that $d_{t'} + 2\eta_{t'} < 1$. Pick $\delta_1 \leq r_0$ so that for $|t-t'| < \delta_1$ and $\|x-x(t')\| < \delta_1$, we have

$$\begin{aligned}\|G(t,x)-G(t,x(t))\| &\leq \|G(t,x)-G(t,x(t))-G_x(t'x(t'))(x-x(t))\| \\ &\quad + \|G_x(t',x(t'))\|\|x-x(t)\| \\ &\leq \eta_{t'}\|x-x(t)\| + d_{t'}\|x-x(t)\|.\end{aligned}$$

By continuity of x there exists a $\delta_{t'} \leq \delta_1$ so that $\|x(t)-x(t')\| \leq \delta_1/2$ for $|t-t'| \leq \delta_{t'}$. Thus for $|t-t'| < \delta_{t'}$ and $\|x-x(t)\| \leq \frac{1}{2}\delta_{t'}$,

$$(2) \qquad \|G(t,x)-x(t)\| \leq (d_{t'}+\eta_{t'})\|x-x(t)\|.$$

By the compactness of J there exist a finite number of points $t_i' \in J$ and intervals

$$J_i = \{t \in J \mid |t-t_i'| < \delta_{t_i'}\},\ i = 1,\ldots,K$$

covering J so that for $t \in J_i$, $x \in S(x(t),\frac{1}{2}\delta_{t_i'})$,

$$\|G(t,x)-x(t)\| \leq (d_{t_i'}+\eta_{t_i'})\|x-x(t)\|.$$

Let $d = \min_i \{d_{t_i'}\}$, $\eta = \min_i \{\eta_{t_i'}\}$, and $\delta = \min_i \{\delta_{t_i'}\}$. Choose $\frac{1}{N} = \Delta t > 0$ so that $\|x(t+\Delta t)-x(t)\| < \frac{1}{2}\eta\delta$ for all $t \in J$. Then for any $t \in J$ and $x \in S(x(t),\frac{1}{2}\delta)$

$$\|G(t,x)-x(t+\Delta t)\| \leq \|G(t,x)-x(t)\|+\|x(t)-x(t+\Delta t)\|$$

$$\leq (d+\eta)\frac{1}{2}\delta + \frac{1}{2}\delta\eta \tag{3}$$

$$< \frac{1}{2}\delta.$$

Equations (2) and (3) show that, for $x^0 \in S(x(t),\frac{1}{2}\delta)$, $x^1 = G(t,x^0)$ lies in $S(x(t+\Delta t),\frac{1}{2}\delta)$. By induction for $t_k = k\Delta t$, it follows that the process (2.8a) is well-defined with $j_k = 1$ and that (2.8b) converges to $x(1)$ as $j \to \infty$. This completes the proof.

This theorem allows us to restrict our attention to the case where $j_k = 1$ in (2.8a) for all k and to examine only estimates for Δt. Unfortunately, the a priori estimation of Δt appears to provide only rather unsatisfactory results since it depends on estimates which are numerically hard to obtain.

We first prove a simple lemma which will enable us to estimate the steplength for Newton's method.

4.2 Lemma. Consider the recursive relation in R^1

$$(4) \qquad s_{k+1} = as_k^2 + c,\ k = 0,1,\ldots,\ s_0 = 0,\ a,c > 0.$$

If $1 - 4ac \geq 0$, then

$$(5) \qquad s_k \leq \frac{1}{2a}(1 - \sqrt{1-4ac}),\ \forall k \geq 0.$$

If $1 - 4ac < 0$, then $\lim_{k\to\infty} s_k = \infty$.

Proof. Let $r = \frac{1}{2a}(1 - \sqrt{1-4ac})$, then $r = ar^2 + c$ and, if $0 \leq s_k \leq r$, we obtain, since $as^2 + c$ is monotone increasing on $[0,\infty)$,

$$0 \leq as_k^2 + c \leq ar^2 + c = r.$$

For $1 - 4ac < 0$ the parabola $f(s) = as^2 - s + c$ has no zeros, hence $f(s) \geq \alpha > 0$ on $[0,\infty)$ and

$$s_{k+1} - s_k = as_k^2 - s_k + c \geq \alpha > 0$$

and thus $\lim_{k\to\infty} s_k = \infty$. This completes the proof.

As an example of the kinds of estimates obtainable from the local convergence theorems for certain iterative processes, we present the following result for Newton's method.

4.3 Theorem. Consider $H:J \times D \subset J \times X \to Y$, where X and Y are normed linear spaces, D is open, and H has partial Frechet derivatives both with respect to t and x. Suppose that $x:J \to D$ is a continuous solution of the equation $H(t,x) = 0$, $t \in J$. By the

openness of D there is an $r_0 > 0$ so that (1) holds. Assume that

1) H_t is continuous on $J \times D_0$ and

$$\|H_t(t,x(t))\| \leq \alpha, \quad \forall t \in J;$$

2) $\|H_x(t,x)-H_x(t,y)\| \leq K\|x-y\|, \quad \forall t \in J, \quad \forall x,y \in D_0;$

3) H_x has a bounded linear inverse on $J \times D_0$ and

$$\|H_x(t,x)^{-1}\| \leq \beta, \quad \forall (t,x) \in J \times D_0.$$

Set

$$r_1(h) = \frac{1}{\beta K}\left[1-\sqrt{1-2\alpha\beta^2 Kh}\right] \tag{6}$$

and let h_0 be the largest value in $[0,\frac{1}{2\alpha\beta^2 K}]$ such that

$$r_1(h_0) \leq r_0. \tag{7}$$

Then, if the integer N is chosen such that

$$0 < \Delta t = \frac{1}{N} \leq h_0,$$

the numerical continuation process

$$x^k = x^{k-1}-H_x(k\Delta t,x^{k-1})^{-1}H(k\Delta t,x^{k-1}), \quad k = 1,\ldots,N-1, \quad x^0 = x(0); \tag{8a}$$

$$x^k = x^{k-1}-H_x(1,x^{k-1})^{-1}H(1,x^{k-1}), \quad k = N,N+1,\ldots \tag{8b}$$

is feasible.

Proof. For $0 < k < N$, $x^k \in D_0$, we have

$$\|x^k - x((k+1)\Delta t)\| \leq \|x^{k-1} - x(k\Delta t) - H_x(k\Delta t, x^{k-1})^{-1} H(k\Delta t, x^{k-1})\|$$

(9) $$+ \|x(k\Delta t) - x((k+1)\Delta t)\|$$

$$\leq \beta \|H(k\Delta t, x(k\Delta t)) - H(k\Delta t, x^{k-1}) - H_x(k\Delta t, x^{k-1})(x(k\Delta t) - x^{k-1})\|$$

$$+ \|\int_{k\Delta t}^{(k+1)\Delta t} H_x(s, x(s))^{-1} H_t(s, x(s)) ds\|$$

$$\leq \frac{\beta K}{2} \|x^{k-1} - x(k\Delta t)\|^2 + \beta\alpha\Delta t.$$

With $e_k = \|x^{k-1} - x(k\Delta t)\|$, (9) assumes the form

(10) $$e_{k+1} \leq \frac{\beta K}{2} e_k^2 + \beta\alpha\Delta t.$$

For $k \geq N$, the analogous estimate gives

(11) $$\|x^{k+1} - x(1)\| \leq \frac{\beta K}{2} \|x^k - x(1)\|^2.$$

Thus, if $e_k \leq r_0$ and $e_N < \frac{2}{\beta K}$, numerical continuation is feasible.

By Lemma 3.2, $e_k \leq r_1(h)$ for $0 < h \leq (2\alpha\beta^2 K)^{-1}$. If h_0 is chosen as stated, we have $r_1(h) \leq r_0$ for all $h \in [0, h_0]$ and, since $r_1(1/(2\alpha\beta^2 K)) < \frac{2}{\beta K}$, it follows that $e_N < \frac{2}{\beta K}$, and the proof is complete.

4.4 Remark. In Theorem 4.3 if r_0 is not restricted, as in the case when D is the entire space X, then the only condition for the feasibility proof is that $e_N < \frac{2}{\beta K}$. Thus even if $4ac > 1$, (i.e., $\Delta t > (2\alpha\beta^2 K)^{-1}$) and hence if e_k is divergent, it may happen that $e_N < \frac{2}{\beta K}$. However, there appears to be no simple

explicit condition relating Δt to the restriction $e_N < \frac{2}{\beta K}$ and in fact numerical experiments indicate that significantly larger Δt estimates are not likely.

Theorem 4.3 is similar to a result of Šidlovskaya [1958]. There, instead of the point-of-attraction type of estimates used here, the Newton-Kantorovich Theorem (see, e.g., Ortega and Rheinboldt [1970; p. 421]) was used in an attempt to prove at the same time the existence of the solution curve x. However, the same criticism that applied to the Yakovlev result in Chapter III applies to Šidlovskaya's result as well since she showed only local existence and uniqueness without proving the continuation property necessary to guarantee existence and continuity of x for all of J.

Theorem 4.3 for Newton's method is typical of the types of results obtainable for the estimation of Δt using other locally convergent iterative processes. Numerically, the Lipschitz constant K and the bounds α and β of the theorem are difficult to determine. We examine this problem in detail in the following section.

Section 2. An Application to Mildly Nonlinear Equations

Consider the Dirichlet problem

$$\Delta u(s,t) = u^{2M+1}(s,t), \quad \forall (s,t) \in D \subset R^2 \tag{12a}$$

$$u(s,t) = v(s,t), \quad \forall (s,t) \in \partial D, \tag{12b}$$

where $M > 0$ is some integer and D is the interior of the unit square $[0,1] \times [0,1]$. We introduce the uniform grid on $\bar{D}$ with the grid points

$$\bar{\Omega} = \{(s_i,t_i) = (ih,jh) \mid h = \frac{1}{m+1},\ 0 \leq i,j \leq m+1\}$$

and set

$$\Omega = \{(s_i,t_j) \mid i \leq i,j \leq M\},\ \Omega' = \bar{\Omega} \setminus \Omega .$$

For a point $P = (s_i,t_j) \in \Omega$, let $N(P)$ be the four adjacent points in $\bar{\Omega}$: $(s_i \pm h,t_j),(s_i,t_j \pm h)$. Then using the standard 5-point difference approximation to the Laplace operator, we obtain the standard finite difference form of (12a,b)

$$(13a) \qquad -\frac{4}{h^2}\,u(P) + \frac{1}{h^2} \sum_{Q \in N(P)} u(Q) = u^{2M+1}(P),\ P \in \Omega$$

$$(13b) \qquad u(P) = v(P),\ P \in \Omega'.$$

Assume the points $P \in \Omega$ are numbered row-wise from top to bottom, left to right. Then the equations (13a,b) can be phrased in matrix notation in R^n, where $n = m^2$. With $x_i = u(P_i)$, $i = 1,\ldots,n$, we have

$$(14) \qquad -\frac{1}{h^2}\,Ax = \frac{1}{h^2}\,B(x) - b,\ x,b \in R^n$$

where A is a block-tridiagonal matrix

$$(15) \qquad A = \begin{bmatrix} A_{11} & -I & & 0 \\ -I & A_{22} & \ddots & \\ & \ddots & \ddots & -I \\ 0 & & -I & A_{nn} \end{bmatrix}$$

with

$$(16) \qquad A_{ii} = \begin{bmatrix} 4 & -1 & & 0 \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 4 \end{bmatrix},$$

$B: R^n \to R^n$ has the components

$$(17) \qquad B_i(x) = h^2 x_i^{2M+1}$$

and b has the components

$$(18) \qquad b_i = \begin{cases} 0, \text{ for } P_i \text{ such that } N(P_i) \subset \Omega \\ v_i, \text{ for all other } P_i \end{cases}$$

where

$$(19) \qquad v_i = \sum_{Q \in N(P_i) \cap \Omega'} v(Q).$$

We imbed the problem (14) in the homotopy equation

$$(20) \qquad H(t,x) \equiv Ax + tB(x) - b = 0, \; t \in J$$

and let $x_0 \in R^n$ be the solution of $Ax = b$.

The properties of A are well-known. We enumerate those that concern us here (see Varga [1962, p. 202f]).

4.5 Lemma. The matrix A of (15/16) has the following properties:

1) $A = (a_{ij})$ is a Stieltjes matrix, that is $a_{ij} \leq 0$ for all $i \neq j$ and A is symmetric and nonsingular with $A^{-1} \geq 0$.

2) The minimum eigenvalue of A is

(21) $$\mu = 4(1-\cos \pi h).$$

We next prove a lemma about some of the resulting properties of H as defined by (20).

4.6 Lemma. Consider $H:J \times R^n \to R^n$ as given by (20) where A and B are as defined by (15)-(17). Then H has the following properties:

1) H_x exists and is continuous on $J \times R^n$.

2) H_x has a bounded inverse for all $(t,x) \in J \times R^n$ and

(22) $$\|H_x(t,x)^{-1}\|_2 \leq \frac{1}{\mu}$$

where μ is given by (21).

3) If $x:J \to R^n$ is a solution of (20), then

(23) $$\|x(t)\|_2 \leq \frac{\|b\|_2}{\mu}, \quad \forall t \in J.$$

4) $H(t,\cdot)$ is a homeomorphism from R^n onto R^n for each $t \in J$.

Proof. The derivative $B'(x)$ exists and is given by the diagonal matrix

(24) $$B'(x) = (2M+1)h^2 \operatorname{diag}(x_1^{2M},\ldots,x_n^{2M}).$$

Hence also

(25) $$H_x(t,x) = A + tB'(x)$$

exists.

Since A is symmetric and B' diagonal and positive, H_x is symmetric and

$$(26) \qquad \langle H_x(t,x)\xi,\xi\rangle = \langle A\xi,\xi\rangle + t\langle B'(x)\xi,\xi\rangle$$

$$\geq \mu\|\xi\|_2^2, \quad \forall \xi \in R^n.$$

Therefore, H_x is positive definite and $H_x(t,x)^{-1}$ exists on $J \times R^n$ and

$$\|H_x(t,x)^{-1}\|_2 \leq \frac{1}{\mu}, \quad \forall (t,x) \in J \times R^n.$$

By Corollary 3.10 this implies now that $H(t,\cdot)$ is a homeomorphism from R^n onto R^n. Moreover, by the inverse function theorem, the inverse mapping $H(t,\cdot)^{-1}$ of $H(t,\cdot)$ has a partial Frechet derivative with respect to x given by $H_x(t,\cdot)^{-1}$. Therefore, by the integral mean value theorem

$$\|x-x(t)\|_2 \leq \int_0^1 \|H_x(t,sx+(1-s)x(t))^{-1}[H(t,x)-H(t,x(t))]\|_2 ds$$

$$\leq \frac{1}{\mu}\|H(t,x)\|_2$$

and, letting $x = 0$, we have

$$\|x(t)\|_2 \leq \frac{1}{\mu}\|H(t,0)\|_2 = \frac{\|b\|_2}{\mu}.$$

This completes the proof.

On the basis of these lemmas, we can apply the results of Theorem 4.3 to the problem (20).

4.7 Theorem. Let $H:J \times R^n \to R^n$ be given by (20) where A and B are defined by (15)-(17). Then a solution curve $x:J \to R^n$ satisfying (20) exists. Furthermore, numerical continuation with the Newton process, as given by (8a,b) is feasible.

Moreover, if r_0 is the first positive solution of

$$(27) \qquad r(\mu r+\|b\|_2)^{2M-1} = \frac{\mu^{2M}}{h^2(4M^2-1)}$$

then

$$(28) \qquad \Delta t = \frac{1}{N} \leqslant \frac{\mu^{4M^2+1}}{2mh^4\|b\|_2^{2M+1}(4M^2-1)(\mu r_0+\|b\|_2)^{2M-1}}$$

is an admissible steplength.

Proof. The existence and continuity of $x:J \to R^n$ satisfying (20) follows from Lemma 4.6, part 4, and Theorem 3.6.

As we pointed out in Remark 4.4 since the domain is not restricted, we need only require that e_N as given by (10) insures that $\lim_{k\to\infty} x^k = x(1)$ in (8b).

Define D_0 again by (1) where $r_0 > 0$ is left unspecified for the time being. Since $H_t(t,x(t)) = B(x(t))$, by Lemma 4.6, part 3, and (17), we have

$$(29) \qquad \|H_t(t,x(t))\|_2 \leqslant mh^2\left(\frac{\|b\|_2}{\mu}\right)^{2M+1} \equiv \alpha.$$

In order to obtain a Lipschitz condition for H_x we consider

$$\|H_x(t,x)-H_x(t,y)\|_2 = t\|B'(x)-B'(y)\|_2$$

$$(30) \qquad \leq (2M+1)h^2 \|\text{diag}\,(x_1^{2M}-y_1^{2M},\ldots,x_n^{2M}-y_n^{2M})\|_2$$

$$\leq (2M+1)h^2 \max_j |x_j^{2M}-y_j^{2M}|.$$

Since

$$\|x-y\|_2 = \Big(\sum_j (x_j-y_j)^2\Big)^{1/2} \geq \max_j |x_j-y_j|,$$

we may continue from (30) to obtain for $x \neq y$, $x,y \in D_0$,

$$\frac{\|H_x(t,x)-H_x(t,y)\|_2}{\|x-y\|_2} \leq (2M+1)h^2(2M-1)\left(r_0 + \frac{\|b\|_2}{\mu}\right)^{2M-1},$$

where we have used the identity

$$(a^{2M}-b^{2M}) = (a-b)(a^{2M-1}+a^{2M-2}b+\ldots+b^{2M-1}).$$

Therefore, if

$$K = (4M^2-1)h^2\left(r_0 + \frac{\|b\|_2}{\mu}\right)^{2M-1},$$

then H_x satisfies

$$\|H_x(t,x)-H_x(t,y)\|_2 \leq K\|x-y\|_2. \quad x,y \in D_0.$$

With $\beta = \frac{1}{\mu}$ and α and K as specified, the hypotheses of Theorem 4.3 are satisfied. Thus for fixed r_0, it suffices to take

$$(31) \qquad \Delta t = \frac{1}{N} \leq \frac{1}{2\alpha\beta^2 K}.$$

We can pick r_0 so that D_0 contains all the iterates; hence by Lemma 4.2, with e_k as defined by (10), we have

$$e_k \leq r_0 \leq \frac{1}{\beta K}$$

or

$$r_0 k \leq \frac{1}{\beta} . \tag{32}$$

We use (32) to define r_0 by choosing r_0 to be the smallest positive solution of $rk = \frac{1}{\beta}$. By substitution, this corresponds to equation (22) and (31) corresponds to (28). Hence the proof is complete.

The following table gives the values of Δt obtained from (28) for the problem (13a,b) with $v \equiv 1$ on Ω'.

	$\Delta u = u^3$	$\Delta u = u^5$
$h = \frac{1}{6}$	1.229×10^{-3}	2.745×10^{-8}
$h = \frac{1}{8}$	5.899×10^{-5}	7.953×10^{-11}
$h = \frac{1}{11}$	1.910×10^{-6}	1.147×10^{-13}

As we can see, even for a reasonably straightforward example, the a priori estimates for Δt obtainable in this way are numerically of doubtful value.

Section 3. Generalized Continuation Processes

We have seen that in a numerical continuation process our attention can be directed, in general, to the case of taking only one step of the local iterative process for each parameter step. This produces a sequence of N points $x^0 = x(0), x^1, \ldots, x^{N-1}$ which may not even remain close to the curve $x = x(t)$, $t \in J$. Besides the actual existence of the points $\{x^k\}$, the critical requirement for feasibility of the numerical continuation process is the fact that the last point x^{N-1} is actually in the domain of attraction of the local iterative process at the endpoint $x(1)$ of the curve.

This remark represents in effect an uncoupling of the initial and terminal phases of the numerical continuation process which was already suggested by the separate equations (2.8a) and (2.8b). It is now natural to uncouple the two phases even more formally by assuming that instead of the single function G in (2.8a,b), two mappings $G^1: J \times D_1 \subset R^1 \times X \to X$ and $G^2: D_2 \subset X \to X$ are given. With these mappings and some uniform partition $\{k\Delta t, k=0,1,\ldots,N\}$ of J, $N\Delta t = 1$, we then consider the process

$$x^k = x^{k-1} - G^1(k\Delta t, x^{k-1}), \quad k = 1,\ldots,N-1, \tag{33a}$$

starting at a given x^0, followed by

$$x^k = x^{k-1} - G^2(x^{k-1}), \quad k = N, N+1, \ldots \ . \tag{33b}$$

Note that the mapping G^1 need no longer define a locally convergent

iterative process as G did in (2.8a). Moreover, G^2 might have no direct relationship to G^1. For example, in the case when the curve $x:J \to X$ satisfies a differential equation, G^1 might arise as a finite difference approximation to the derivative x' and G^2 might be a locally convergent iterative process at $x(1)$.

The basic definitions for the combined process (33a,b) are now stated as follows.

4.8 Definition. The mapping $G^1:J \times D_1 \subset R^1 \times X \to X$ is admissible with respect to the curve $x:J \to X$ if for $x^0 = x(0)$ and all $\Delta t \leq h_0$, (33a) produces a well-defined sequence x^k, $k = 0,1,\ldots,N-1$, in D_1.

The combined process (33a,b) is feasible if the following conditions hold: (i) $G^2:D_2 \subset X \to X$ defines an iterative process (33b) which has $x(1)$ as a point of attraction; (ii) G^1 is admissible with respect to the curve x; and (iii) the last point x^{N-1} is in a domain of attraction of $x(1)$ for the processes (33b).

We shall refer to (33a,b) as a generalized numerical continuation process and call (33a) and (33b) the initial and terminal phases of this generalized process, respectively.

Note that if $G^1 \equiv G:J \times D \subset J \times X \to X$ and $G^2 \equiv G(1,\cdot):D \subset X \to X$ and for any $t \in J$, $x(t)$ is a point of attraction for the iterative process defined by $G(t,\cdot)$, then by Theorem 2.2 the existence of a uniform $r > 0$, such that $S(x(t),r) \subset D$ is a domain of attraction for $x(t)$, implies the feasibility of the process (33a,b). Thus (33a,b) includes the numerical continuation processes considered earlier.

4.9 Definition. Let the mapping $G^1: J \times D \subset J \times X \to X$ be admissible with respect to the curve $x: J \to X$. If $\|x^{N-1} - x(1)\| \leq C(\Delta t)^P$ for $\Delta t = \frac{1}{N} \leq h_0$ and some constant C independent of Δt, then (33a) is called a pth order initial numerical continuation process.

4.10 Theorem. Let the mapping $G^1: J \times D_1 \subset J \times X \to X$ be admissible with respect to the curve $x: J \to X$ and assume that $G^2: D_2 \subset X \to X$ defines a locally convergent iterative process (33b) with $x(1)$ as a point of attraction. If G^1 is a pth order initial numerical continuation process, then the entire process (33a,b) is feasible.

Proof. Since $x(1)$ is a point of attraction of the iterative process (33b), there exists a constant $r > 0$ such that $S(x(1),r)$ is a domain of attraction for $x(1)$. Choose

$$\Delta t < \min \{(r/C)^{1/P}, (1/N_0)\} \tag{34}$$

where N_0 and C are given by Definition 4.9. Then $x^{N-1} \in S(x(1),r)$ which proves the feasibility.

In Theorem 4.3 let us assume that r_0 can be taken larger than $(\beta K)^{-1}$ where β, K and α are as stated in the theorem. Then for $\Delta t \leq (2\alpha\beta^2 K)^{-1}$,

$$e_k = \|x^{k-1} - x(k\Delta t)\|$$

satisfies (10), and, in particular, $e_1 = \alpha\beta\Delta t$. Assume now that $e_k \leq a_k \alpha\beta\Delta t$ with some constant a_k; then

$$e_{k+1} \leq a_k^2(2\alpha\beta^2K\Delta t)\,\frac{\alpha\beta}{4}\,\Delta t + \alpha\beta\Delta t \leq \left(\frac{1}{4}\,a_k^2 + 1\right)\alpha\beta\Delta t,$$

shows that $\{a_k\}$ may be specified by

$$a_{k+1} \leq \frac{1}{4}\,a_k^2 + 1,\ a_0 = 0.$$

Hence, by Lemma 4.2, $a_k \leq 2$ for all k, and

$$e_N \leq 2\alpha\beta\Delta t.$$

Therefore, if $G^1 \equiv G:J \times D \subset J \times X \to X$ is defined by

$$G(t,x) = x - H_x(t,x)^{-1}H(t,x) \tag{35}$$

where H is the homotopy in Theorem 4.3, i.e., if G^1 represents for each $t \in J$ the Newton iteration function, then under the hypotheses of Theorem 4.3, (33a) is an initial numerical continuation process of order 1.

The following is an example of the application of Theorem 4.10 with Newton's method as the terminal iterative process. For clarity we have phrased all subsequent results of this section for mappings in R^n.

4.11 Corollary. Let $H:J \times D \subset J \times R^n$ be continuously Frechet differentiable with respect to x on $J \times D$, where D is open. Assume that there exists a continuous solution curve $x:J \to D$ of $H(t,x) = 0$.

Let $G^1:J \times D_1 \subset J \times R^n \to R^n$ be a pth order initial numerical continuation process with respect to the curve x, and consider

$G^2: D \subset R^n \to R^n$, $G^2(x) = G(1,x)$, where G is defined by (35) and H has the following properties:

i) $\|H_x(t,x)-H_x(t,y)\| \leq K\|x-y\|$, $\forall x,y \in D$, $t \in J$;

ii) $H_x(t,x)$ has a bounded inverse on $J \times D$ and $\|H_x(t,x)^{-1}\| \leq \beta$, $\forall (t,x) \in J \times D$.

Then (33a,b) is feasible with

$$\Delta t < \min \left\{ \left(\frac{2}{\beta K C} \right)^{1/P}, \frac{1}{N_0} \right\} \tag{36}$$

where C and N_0 are the constants in Definition 4.9.

Proof. Using the bound on H_x^{-1}, the Lipschitz condition on H_x and the integral mean value theorem, we obtain

$$\begin{aligned} \|x^{j+1}-x(1)\| &= \|x^j-x(1)-H_x(1,x^j)^{-1}H(1,x^j)\| \\ &\leq \beta\|H_x(1,x^j)(x^j-x(1))-[H(1,x^j)-H(1,x(1))]\| \\ &\leq \frac{\beta K}{2}\|x^j-x(1)\|^2. \end{aligned}$$

Therefore, if $\|x^{N-1}-x(1)\| < (\frac{2}{\beta K}) \equiv r$, it follows that $\lim_{j\to\infty} x^j = x(1)$. Now apply Theorem 4.10 and in particular equation (34) to get (36).

We next prove a result which gives some conditions under which we have a numerical continuation process of higher order.

4.12 Theorem. Let $x: J \to D \subset R^n$ be a continuous function from J into the open set D, and choose $r_0 > 0$ such that (1) holds. For $h \in (0,h_0]$, $h_0 > 0$, let $Q_h: J \times D_0 \to R^n$ be continuous and such that

(37) $\|Q_h(t,x)-Q_h(t,y)\| \leq a\|x-y\|, \quad \forall x,y \in D_0, \ t \in J,$

where a is a constant.

Finally assume that for any $t \in J$, $0 < h \leq h_0$,

(38) $\|x(t+h)-x(t)-hQ_h(t,x(t))\| \leq bh^p, \ p \geq 2.$

Then for

(39) $$\Delta t \leq \min\left\{\left[\frac{ar_0}{b(e^a-1)}\right]^{1/(p-1)}, h_0\right\},$$

$G^1: J \times D_0 \to R^n$, $G^1(t,x) = \Delta t Q_h(t,x)$, is admissible with respect to the curve x and (33a) is an initial numerical continuation process of order $p - 1$.

Proof. Assume that $x^j \in D_0$ is well-defined for $j = 0,1,\ldots,k$, and set $\varepsilon_j = x^j - x(j\Delta t)$, $0 \leq j \leq k$.

Then

$$\begin{aligned}\varepsilon_j - \varepsilon_{j-1} &= x^j - x^{j-1} - [x(j\Delta t)-x((j-1)\Delta t)] \\ &= \Delta t[Q_h((j-1)\Delta t,x^{j-1})-Q_h((j-1)\Delta t,x((j-1)\Delta t))] \\ &\quad - [x(j\Delta t)-x((j-1)\Delta t)-\Delta t Q_h((j-1)\Delta t,x((j-1)\Delta t))]\end{aligned}$$

implies that

$$\|\varepsilon_j-\varepsilon_{j-1}\| \leq a\Delta t\|\varepsilon_{j-1}\| + b(\Delta t)^p,$$

and hence that

$$\|\varepsilon_j\| \leq \|\varepsilon_j - \varepsilon_{j-1}\| + \|\varepsilon_{j-1}\|$$

$$\leq (1+a\Delta t)\|\varepsilon_{j-1}\| + b(\Delta t)^p.$$

Therefore since $\|\varepsilon_0\| = 0$, $\|\varepsilon_j\|$ satisfies

$$(40) \qquad \|\varepsilon_j\| \leq \left(\frac{b}{a}\right)(\Delta t)^{p-1}(e^{ja\Delta t}-1),\ j = 0,1,\ldots,k.$$

Since x^{k+1} is defined, so is $\|\varepsilon_{k+1}\|$; and (40) is satisfied for j=k+1. From (39) it then follows that $\|\varepsilon_{k+1}\| \leq r_0$ which shows that $x^{k+1} \in D_0$. Now by induction x^j is defined and remains in D_0 for $j = 0,1,\ldots,N-1$. Furthermore (40) also proves that (33a) is of order $p - 1$. This completes the proof.

Theorem 4.12 includes those methods for which the curve x is differentiable and $Q_h(t,x)$ is an approximation to x' in the sense of (38). In fact, if Q_h does not depend on h, then by (38) x is differentiable with $x'(t) = Q(t,x(t))$. In this case, (33a) becomes the Euler method for solving numerically the initial value problem

$$(41) \qquad x' = Q(t,x),\ t \in J,\ x \in D \subset R^n,\ x(0) = x_0,$$

and the estimate (40) for $p = 2$ is a well-known result establishing the Euler method as a first order integration method.

This approach of considering (41) as the determining equation for the curve $x:J \to R^n$ and employing numerical integration techniques to estimate $x(1)$ is essentially the approach used by Davidenko and other authors mentioned in this connection in Chapter I.

As an example of the application of the preceding theorems to a specific homotopy, we give the following result of Meyer [1968].

4.13 Corollary. Let the mapping $F:R^n \to R^n$ be continuously differentiable on all of R^n and assume that

$$\|F'(x)-F'(y)\| \leq K\|x-y\|, \quad x,y \in R^n.$$

Furthermore, suppose that F' has an inverse for each $x \in R^n$ and that $\|[F'(x)]^{-1}\| \leq \beta$. For some $x_0 \in D$ let $H:J \times R^n \to R^n$ be defined by

$$H(t,x) = F(x) + (t-1)F(x_0). \tag{42}$$

Then a solution curve $x:J \to R^n$ satisfying $H(t,x(t)) = 0$, $x(0) = x_0$, exists and is continuous.

Furthermore, if

$$G^1(t,x) \equiv G^1(x) = \Delta t[F'(x)]^{-1}F(x_0) \tag{43}$$

and

$$G^2(t,x) = -[F'(x)]^{-1}F(x), \tag{44}$$

then G^1 is admissible with respect to the curve x and (33a,b) is feasible for

$$\Delta t = \frac{1}{N} < \frac{4}{\beta^2 K\|F(x_0)\|[\exp(\beta^2 K\|F(x_0)\|)-1]} \quad . \tag{45}$$

Proof. The existence of the curve x follows from Hadamard's Corollary 3.8.

Let $Q_h(t,x) \equiv Q(x) = -[F'(x)]^{-1}F(x_0)$. For $x,y \in R^n$, we have

$$\|Q(x)-Q(y)\| = \|\{[F'(x)]^{-1}-[F'(y)]^{-1}\}F(x_0)\|$$

$$\leq \beta^2 K\|F(x_0)\|\|x-y\|.$$

Furthermore, since x satisfies $x' = Q(x)$, we see that

$$\|x(t+h)-x(t)-hQ(x(t))\| = \| \int_0^1 [Q(x(t+sh))-Q(x(t))]hds \|$$

$$\leq \beta^2 K\|F(x_0)\| \int_0^1 \|x(t+sh)-x(t)\| hds$$

$$\leq \frac{1}{2} \beta^3 K\|F(x_0)\|^2 h^2.$$

Now apply Theorem 4.12 and (40), in particular, to obtain that for G^1 given by (43)

$$\|x^{N-1}-x(1)\| \leq \frac{\beta\|F(x_0)\|\Delta t}{2} (e^{\beta^2 K\|F(x_0)\|}-1). \tag{46}$$

Taking (46) to define the constant C of Definition 4.8, we can apply Corollary 4.11 to obtain that for G^2 given by (44) the numerical continuation process (33a,b) is feasible with Δt satisfying (45).

Theorem 4.12 also covers the case where higher order integration methods are used to solve the initial value problem (41). For example, assume that (41) is given and that $Q(t,x)$ denotes the

right-hand side of the differential equation. Then if $Q_h(t,x) = \frac{1}{6}[k_1+2k_2+2k_3+k_4]$, where

$$k_1 = Q(t,x) \quad , \quad k_3 = Q(t+\tfrac{1}{2}h, x+\tfrac{1}{2}hk_2),$$

$$k_2 = Q(t+\tfrac{1}{2}h, x+\tfrac{1}{2}hk_1), \quad k_4 = Q(t+h, x+hk_3) \quad ,$$

we have the classical Runge-Kutta method and (38) is satisfied with $p = 4$ for all sufficiently small h (see Henrici [1962; p. 120f.]). This is the method used by Kleinmichel [1968] and Bosarge [1968].

We can also phrase a theorem which proves that numerical continuation is feasible for an initial process analogous to the predictor-corrector techniques in solving initial value problems.

4.14 Theorem. Let F and H satisfy the conditions of Corollary 4.13 and for each fixed t let $G:J \times R^n \to R^n$ be the iteration function of a locally convergent iterative process at $x(t) = G(t,x(t))$. Finally, for each $t \in J$ let $r_0 > 0$ be such that with some $\delta \in (0,1)$

$$(47) \qquad \|G(t,x)-x(t)\| \leq (1-\delta)\|x-x(t)\|, \quad \forall x \in S(x(t),r_0).$$

Consider the initial numerical continuation process

$$(48) \qquad y^k = x^{k-1} - \Delta t\, F'(x^{k-1})^{-1}F(x^0), \quad x^0 = x(0),$$

$$(49) \qquad x^k = G(k\Delta t, y^k), \quad k = 1,2,\ldots,N-1,$$

and the terminal process

$$x^{k+1} = G(1,x^k),\ k = N,N+1,\ldots . \tag{50}$$

Then (48)-(50) is a feasible generalized numerical continuation process with $\Delta t < h_0$ for h_0 sufficiently small.

Proof. The initial process can be related back to the previous results but the discussion is considerably simpler if we prove the result over again following the same proof procedure.

Let $\varepsilon_k = x^k - x(k\Delta t)$ and $\eta_k = y^k - x(k\Delta t)$. Then from the identity

$$x(k\Delta t) = x((k-1)\Delta t) - \int_0^1 \Delta t F'(x((k-1+s)\Delta t))^{-1}F(x^0)ds$$

and the equations (48,49) we find that

$$\eta_k = \varepsilon_{k-1} - \Delta t \int_0^1 [F'(x((k-1)\Delta t))^{-1} - F'(x((k-1+s)\Delta t))^{-1}]F(x^0)ds$$

$$- \Delta t \int_0^1 [F'(x^{k-1})^{-1} - F'(x((k-1)\Delta t))^{-1}]F(x^0)ds.$$

Therefore as in the proof of Corollary 4.13

$$\|\eta_k\| \leq \|\varepsilon_{k-1}\| + \tfrac{1}{2}\beta^3 K\|F(x^0)\|^2(\Delta t)^2$$

$$+ \beta^2 K\|F(x^0)\|\|\varepsilon_{k-1}\|\Delta t$$

or

$$\|\eta_k\| \leq (1+\beta^2 K\|F(x^0)\|\Delta t)\|\varepsilon_{k-1}\| + \frac{1}{2}\beta^3 K\|F(x^0)\|^2(\Delta t)^2.$$

From (47) it follows that if $y^k \in S(x(k\Delta t),r_0)$ then

$$\|\varepsilon_k\| < (1-\delta)r_0.$$

Thus, in order to ensure that $y^{k+1} \in S(x((k+1)\Delta t),r_0)$, it suffices to choose t such that

$$(1+\beta^2 K\|F(x^0)\|\Delta t)(1-\delta)r_0 + \frac{1}{2}\beta^3 K\|F(x^0)\|^2(\Delta t)^2 < r_0.$$

A straightforward calculation now shows that we may take

$$(51) \qquad \Delta t < \frac{\sqrt{\beta^2K^2(1-\delta)^2r_0^2+2\beta K\delta r_0}}{\beta^2 K\|F(x^0)\|} - \frac{(1-\delta)r_0}{\beta\|F(x^0)\|} \equiv h_0.$$

Then it follows by induction that (48,49) produces a well-defined sequence with $x^{N-1} \in S(x(1),r_0)$. Since this ball is a domain of attraction for the process (50) at $x(1)$, the feasibility of numerical continuation is proved for $\Delta t < h_0$.

4.15 Remark. In the case when G corresponds to Newton's method we can take $\delta = \frac{1}{2}$, $r_0 = \frac{1}{\beta K}$ in equation (11) of the proof of Theorem 4.3. Then it is readily verified that (51) yields

$$(52) \qquad h_0 = \frac{5}{8}\,\frac{1}{\beta^2 K\|F(x_0)\|}\,.$$

A comparison of the initial processes (43) and (48,49), and with G as the Newton iteration function, shows that the latter requires somewhat over twice as much work for each parameter step. To be

precise, the first process requires the solution of one set of n linear equations, and the second requires the solution of two sets of n linear equations and a function evaluation. However, (52) is in general larger than the estimate given in (45). Assuming that the two methods are rated equal if the estimate (52) is three times the estimate in (45), one finds that (48,49) is "better" if $\beta^2 K\|F(x_0)\| \geq 3.01$. This would indicate a preference for (48)-(50) in most instances.

We conclude this section with a brief one-dimensional example which shows that even the steplength obtainable by the use of these methods, and, in particular, by (45) and (52), are not only numerically difficult to calculate but generally too conservative to be useful.

4.16 Example. Consider $f:R \to R$, $f(x) = \tan^{-1}x$ and the homotopy $H:J \times R \to R$, $H(t,x) \equiv f(x) + (t-1)f(x_0)$, where $x_0 = 2$.

Clearly a solution curve $x:J \to R$ exists satisfying $H(t,x) = 0$. If Newton's method is applied directly to the problem $f(x) = 0$, it is readily verified that the method converges to $x = 0$ if $|x_0| < y_0$, and diverges if $|x_0| \geq y_0$ where y_0 is the positive solution of

$$\tan [2y/(1+y^2)] = y. \tag{53}$$

Since $f'(x) = (1+x^2)^{-1}$ and $f''(x) = -2x(1+x^2)^{-2}$, it follows that for $D = [-x_0,x_0]$,

$$[f'(x)]^{-1} \leq 5, \; x \in D$$

and

$$|f''(x)| \leq 1.48, \; x \in D.$$

Let $\beta = 5$, $K = 1.48$, $|F(x_0)| = 1.108$. Then by (52) we may take

$$\Delta t \leq .0152,$$

while from (45) we find the estimate

$$\Delta t \leq .02 \; e^{-41}.$$

Here the second estimate is obviously useless, and the first is not particularly good. However, $x_0 = 2$ is still relatively close to the solution $x = 0$; and if we choose a somewhat worse initial point $x_0 = 5$, we find that the estimate (52) yields the overly conservative estimate of

$$\Delta t \leq 5.6 \times 10^{-4}.$$

CHAPTER V

A New Approach to Numerical Continuation

Section 1. Numerical Continuation by Minimization

In Chapter IV we saw that good stepsize estimates under which numerical continuation is guaranteed to be feasible are difficult to obtain. Even when obtainable, our examples indicate that these estimates tend to be overly conservative. For the basic continuation process as defined in Chapter I, the step size depends critically on the generally unknown size of the domains of attraction for the local process at each point $x(t)$ of the solution curve. In fact the maximal choice $\Delta t = 1$ is only possible if $x(0)$ is in the domain of attraction of $x(1)$. Since the domains of attraction are rarely under our control, other means for finding better stepsize estimates are needed. One such approach was in effect our introduction of the generalized continuation process in the last section of Chapter IV. For those generalized processes the intermediate points need not remain close to the curve as long as they are well-defined and meet the critical condition that at the end an iterate lies in the domain of attraction of the terminal process at $x(1)$. In this chapter we consider a second somewhat related approach to the problem. It is based on unconstrained minimization techniques widely used in conjunction with the solution of nonlinear equations.

We begin by summarizing some basic facts about such minimization processes. Consider the problem

$$F(x) = 0 \tag{1}$$

where $F:D \subset R^n \to R^n$ is a given mapping. If F is a potential operator, that is, if $F(x) = g'(x)^T$ for some differentiable functional $g:D \subset R^n \to R^1$, then every minimizer of g in the interior of D is a solution of (1). Even if F is not a potential operator, (1) can be converted to a minimization problem by considering a functional $f:R^n \to R^1$ which has $x = 0$ as its unique global minimizer and by setting

$$g:D \subset R^n \to R^1,\ g(x) = f(F(x)). \tag{2}$$

For example, $f(x) = \|x\|$ with some norm on R^n, or $f(x) = x^TAx$ with a symmetric, positive definite $A \in L(R^n,R^n)$ are possible choices for f.

A descent method for finding a minimizer of a functional $g:D \subset R^n \to R^1$ has the general form

$$x_{k+1} = x_k - \omega_k\alpha_k p_k,\ g(x_{k+1}) < g(x_k),\ k = 0,1,\ldots,$$

where p_k is a direction vector, α_k a suitable steplength along this direction, and ω_k some relaxation parameter. In the case of most convergence theorems for descent methods, the iterates x_k, $k = 1,2,\ldots$, remain in the same path-connected component L^0 of the level set $L = \{x \in D \mid g(x) \leq g(x_0)\}$ which contains x_0; that is, only minimizers in L^0 are obtained. This is a basic observation in our development of minimization methods in the continuation setting.

We turn now to this continuation setting and consider again a homotopy $H:\hat{J} \times D \subset R^1 \times R^n \to R^n$ where $\hat{J} \subset R^1$ is an open interval

containing $J = [0,1]$. As before, suppose that there exists a continuous curve $x:J \to D$ with $x(0) = x_0$ which solves

$$H(t,x) = 0,\ t \in J. \tag{3}$$

In analogy to (2), let $f:R^1 \times R^n \to R^1$ be a functional which has $(1,0) \in R^1 \times R^n$ as its unique global minimizer and set

$$g:\hat{J} \times D \subset R^1 \times R^n \to R^1,\ g(t,x) = f(t,H(t,x)). \tag{4}$$

Then clearly, $(1,x(1)) \in R^1 \times R^n$ must be a global minimizer of g on $\hat{J} \times D$, and in order to minimize g we can employ some descent method. But, as mentioned above, if $(0,x(0))$ is used as the starting point, convergence to $(1,x(1))$ will, in general, necessitate that $(1,x(1))$ be in the path-connected component L^0 of the level set

$$L = \{(t,x) \in \hat{J} \times D \,|\, g(t,x) \leq g(0,x_0)\},\ x_0 = x(0), \tag{5}$$

containing $(0,x(0))$. Usually this is not the case. However, in our setting we have an additional piece of information since we know that $x:J \to D$ connects $x(0)$ with $x(1)$. Hence, if we place upon the functional f the additional condition that

$$\phi:J \to D,\ \phi(t) = g(t,x(t)) = f(t,0) \tag{6}$$

is antitone in t, then automatically $(0,x(0))$ as well as $(1,x(1))$ and in fact all points $(t,x(t))$, $t \in J$, must be in L^0. This condition on ϕ is easily satisfied; for example, we can use

$f(t,y) = c_1\|y\| + c_2(t-1)^2$ or $f(t,y) = c_1 y^T A y + c_2(t-1)^2$ with some norm on R^n or some positive definite symmetric $A \in L(R^n,R^n)$ respectively, and with arbitrary positive constants c_1 and c_2.

Under the antitonicity condition for ϕ, the curve

$$z:J \to J \times D,\ z(t) = (t,x(t))$$

represents a "continuous descent method" leading from the known initial point $(0,x(0))$ to the desired minimizer $(1,x(1))$. In practice, we are, of course, not able to move along the curve itself, and approximations to z must be considered instead. Here we have essentially the same choices as before in numerical continuation.

In general, the domain D is assumed to be open and the set L^0 closed. Hence, if ϕ is in fact strictly antitone, then $z(t) \in \text{int } L^0$ for all $t \in (0,1]$. This implies that for any approximation process which produces points sufficiently close to z, the approximating points will remain in L^0 and will end close to $(1,x(1))$. In some sense this is not unlike the basic numerical continuation process of Chapter I and Chapter II where our concern was directed strongly at keeping all intermediate points close to the curve and hence at getting a good approximation of the entire curve and not only of the final point.

In analogy to the generalized continuation process, we may again disregard this requirement of staying close to the curve. In other words, we may consider any initial process which produces a finite sequence of points subject only to the condition that these points

remain in L^0, and hence are well-defined, and that the final point is near $(1,x(1))$. As in Chapter IV, any locally convergent iterative process can then be used as terminal process to reach $(1,x(1))$ with a desired accuracy. Evidently, a natural initial process is then any general descent method applied to the functional g. In fact, as discussed above, such a descent method will in general yield only points from L^0. If it converges, it will ultimately produce points near the global minimizer $(1,x(1))$ of g which is guaranteed to be in L^0 by the existence of the descent curve z. This is the approach we shall pursue here.

As mentioned before, any descent method for g has the general form

$$(7)\qquad \begin{cases} (t_{k+1},x_{k+1}) = (t_k,x_k) - \omega_k\alpha_k p_k, \\ \qquad\qquad\qquad\qquad\qquad\qquad k = 0,1,\ldots, \\ g(t_{k+1},x_{k+1}) < g(t_k,x_k), \end{cases}$$

and the method is specified once appropriate algorithms for the computation of the direction p_k and of the (relaxed) steplength $\omega_k\alpha_k$ are given. We shall discuss some standard steplength algorithms in the next section. For the choice of the direction algorithm we can utilize some additional information provided by the existence of the curve z. For this we note first the following lemma:

5.1 Lemma. If the functional $g:\hat{J} \times D \subset R^1 \times R^n \to R^1$ is differentiable at the interior point (t,x) of $\hat{J} \times D$, and if $g'(t,x)p > 0$ for some

$p \in R^1 \times R^n$ then there is a $\delta > 0$ so that $g(t_\alpha,x_\alpha) < g(t,x)$ for $(t_\alpha,x_\alpha) = (t,x) - \alpha p$ and all $\alpha \in (0,\delta)$.

The well-known proof is simple and follows directly from the differentiability definition (see, e.g., Ortega and Rheinboldt [1970]).

If we assume again that the function ϕ of (6) is strictly antitone and if g as well as z are differentiable, then evidently

$$g'(z(t))z'(t) = \frac{d}{dt}\, g(t,x(t)) = \phi'(t) < 0$$

which means that $p_t = -z'(t)$ is a suitable descent direction at $z(t)$.

In practice we will consider the functional $f:R^1 \times R^n \to R^1$, $f(t,x) = c_1 x^T x + c_2(t-1)^2$ and hence the functional

$$(8)\quad g:\hat{J} \times D \subset \hat{J} \times R^n \to R^1,\ g(t,x) = c_1 H(t,x)^T H(t,x) + c_2(t-1)^2$$

where $H:\hat{J} \times D \to R^n$ is our given homotopy and c_1,c_2 are suitable positive constants. If H has a continuous partial Frechet derivative with respect to both t and x and if H_x is nonsingular on $\hat{J} \times D$, then $z'(t)$ is given by

$$(9)\qquad z'(t) = (1,-H_x(t,x(t))^{-1}H_t(t,x(t))).$$

Hence we may expect that under the conditions considered above the direction given by

$$q(t,x) = (-1,\ H_x(t,x)^{-1}H_t(t,x)) \tag{10}$$

is a suitable direction vector if the point (t,x) is reasonably close to $z(t)$. For points farther away from z a correction may become necessary, and it is natural to use the gradient direction for this purpose. Consequently for $(t,x) \in \hat{J} \times D$ we will consider the direction given by

$$p(t,x) = \beta\tau q(t,x) + (1-\beta)g'(t,x)^T \tag{11}$$

where $\beta \in [0,1)$ and $\tau = \operatorname{sgn}(1-t)$. Theorem 5.8 below shows when p is a descent direction. The significance of τ is that for $t > 1$ the direction $z'(t)$ points away from $z(1)$. In the next section we shall combine this choice of direction vector for the functional (8) with several steplength algorithms and prove that the resulting process converges to $z(1)$ under appropriate conditions for H.

Section 2. Steplength Algorithms and Convergence Results

In this section we consider the initial phase of our new numerical continuation process in the minimization setting, namely, the use of a descent method of the form (7) to reach a neighborhood of the terminal point $(1,x(1))$ from $(0,x(0))$ as starting point.

For our discussion of the steplength algorithms below, we make the following assumptions about the functional g.

5.2 Basic Conditions for g: The functional $g:\hat{J} \times D \subset R^1 \times R^n \to R^1$ is continuously differentiable on the open set $\hat{J} \times D$ where $[0,1] \subset \hat{J}$. Moreover, $(0,x_0) \in \hat{J} \times D$ is given such that the path-connected component L^0 of the level set L defined by (5) containing $(0,x_0)$ is compact.

For the steplength analysis needed for our process, we use the convergence theory of Elkin [1968] (see also the presentation in Chapter 14 of Ortega and Rheinboldt [1970]). For the sake of brevity only three steplength algorithms are considered although other ones could have been used as well. For any such algorithms there are two conditions which must be met in determining (t_{k+1},x_{k+1}) by (7) for given (t_k,x_k) and descent-direction vector p_k. First of all we must have $(t_{k+1},x_{k+1}) \in L^0$ or at least in L. In this way the iterates remain well-defined; and if L^0 (or L) is compact, then the sequence must have an accumulation point. The second condition is that the value of g must be decreased sufficiently at each step. This principle of "sufficient decrease" is embodied in the condition that

$$(12) \qquad g(t_k,x_k) - g(t_{k+1},x_{k+1}) > \sigma(g'(t_k,x_k)\frac{p_k}{\|p_k\|})$$

where $\sigma:[0,\infty) \to [0,\infty)$ is a function independent of k with the property that for any sequence $\{s_j\} \subset [0,\infty)$, $\lim_{j\to\infty} s_j = 0$, it follows that $\lim_{j\to\infty} \sigma(s_j) = 0$. Such a function σ will be called a forcing function in accord with Elkin's terminology.

We consider first the Curry steplength algorithm. Although the proof can be found, for instance, in Ortega and Rheinboldt [1970], we include it here for the sake of a complete discussion of the present method.

5.3 Theorem. Under the basic condition 5.2 for the functional g, let $(t,x) \in L^0$ and $p \in R^{n+1}$ be given such that $g'(t,x)p \geq 0$, $p \neq 0$. Then with the Curry steplength

$$\alpha = \min \{\beta \geq 0 \,|\, g'((t,x)-\beta p)p = 0\}, \tag{13}$$

it follows that $(\hat{t},\hat{x}) = (t,x) - \alpha p \in L^0$, and

$$g(t,x) - g(\hat{t},\hat{x}) \geq \sigma\left(g'(t,x)\frac{p}{\|p\|}\right) \tag{14}$$

with some fixed forcing function σ which depends only on g and L^0.

Proof. By Lemma 5.1 there exists a $\delta > 0$ such that

$$g((t,x)-\beta p) < g(t,x), \quad \forall\, \beta \in (0,\delta),$$

and hence that for these $\beta \in (0,\delta)$

$$[(t,x),(t,x)-\beta p] = \{(s,y) \in R^1 \times R^n \,|\, (s,y)=(t,x)-\tau p, 0 \leq \tau \leq \beta\} \subset L^0.$$

By the compactness of L^0 now

$$\beta^* = \sup \{\beta \geq 0 \,|\, [(t,x),(t,x)-\beta p] \subset L^0\} \subset (0,\infty),$$

and it follows from the continuity of g on the open set $\hat{J} \times D$ that

necessarily $g((t,x)-\beta^*p) = g(t,x)$. By the mean-value theorem there exists therefore an $\alpha \in (0,\beta^*)$ for which (13) holds, and which is such that $[(t,x),(\hat{t},\hat{x})] \subset L^0$. In addition, clearly

$$\bar{\alpha} = \min \{\beta \geq 0 \,|\, g'((t,x)-\beta p)p = \tfrac{1}{2} g'(t,x)p\} \in (0,\alpha)$$

as well as

$$(15) \qquad g(\hat{t},\hat{x}) \leq g((t,x)-\bar{\alpha}p) < g(t,x).$$

With the induced norm on $L(R^n,R^1)$, let

$$\gamma = \sup \{\|g'(t,x)-g'(s,y)\| \,|\, (t,x),(s,y) \in L^0\}$$

and define $\delta:[0,\infty) \to [0,\infty)$ by

$$\delta(u) = \begin{cases} \inf \{\|(t,x)-(s,y)\| \,|\, (t,x),(s,y) \in L^0, \|g'(t,x)-g'(s,y)\| \geq u\} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } u \in [0,\beta) \\ \lim\limits_{v \to \beta-} \delta(v), \text{ for } u \in [\beta,\infty) \end{cases}$$

Since g' is uniformly continuous on the compact set L^0, we have $\delta(0) = 0$ and $\delta(u) > 0$ for $u > 0$ unless g' is identically zero, an excludable case. Thus, δ is a forcing function.

Now evidently

$$\tfrac{1}{2} g'(t,x) \frac{p}{\|p\|} = g'(t,x) \frac{p}{\|p\|} - g'((t,x)-\bar{\alpha}p) \frac{p}{\|p\|}$$

$$\leq \|g'(t,x) - g'((t,x)-\bar{\alpha}p)\|$$

or

$$\bar{\alpha}\|p\| \geq \delta\left(\frac{1}{2}\, g'(t,x)\, \frac{p}{\|p\|}\right).$$

Therefore by (15) and the mean-value theorem with some $\beta \in (0,\bar{\alpha}]$

$$g(t,x) - g(\hat{t},\hat{x}) \geq g(t,x) - g((t,x)-\bar{\alpha}p)$$

$$= \bar{\alpha} g'((t,x)-\beta p)p \geq \bar{\alpha}\|p\| \frac{1}{2}\, g'(t,x)\, \frac{p}{\|p\|}$$

$$\geq \sigma\left(g'(t,x)\, \frac{p}{\|p\|}\right)$$

where $\sigma(u) = \frac{1}{2}\, u\delta(u)$ is evidently a forcing function.

It is easily seen that underrelaxation with any $\omega \in (\varepsilon,1]$ and fixed $\varepsilon > 0$ could have been used in conjunction with the Curry step (13).

As a corollary of this result, we obtain the following theorem about two minimization steplength algorithms. The proof follows again that of Ortega and Rheinboldt [1970].

5.4 Theorem. Under the basic condition 5.2 for the functional g, let $(t,x) \in L^0$ and $p \in R^{n+1}$ be given such that $g'(t,x)p \geq 0$, $p \neq 0$. Then with $\alpha \geq 0$ such that either

$$(16) \qquad g((t,x)-\alpha p) = \min \{g((t,x)-\beta p) \,|\, [(t,x),(t,x)-\beta p] \subset L^0\}$$

or

$$(17) \qquad g((t,x)-\alpha p) = \min \{g((t,x)-\beta p) \,|\, (t,x)-\beta p \in L^0\}$$

it follows that in both cases $(\hat{t},\hat{x}) = (t,x) - \alpha p \in L^0$ and that (14) holds with the same forcing function σ as in Theorem 5.3.

Proof. Let α^1,α^2 be the steplengths given by (16) and (17), respectively; then by definition $(t,x) - \alpha^i p \in L^0$. Moreover, if α is defined by (13), then evidently

$$g(t,x) \geq g((t,x)-\alpha p) \geq g((t,x)-\alpha^1 p) \geq g((t,x)-\alpha^2 p),$$

and hence the inequality (14) for α^1 or α^2 follows directly from that for α.

As mentioned before, there are various other steplength algorithms which could be considered here as well. For a detailed discussion we refer again to Ortega and Rheinboldt [1970].

For practical purposes, particularly, the Goldstein-Armijo algorithm appears to be useful. In that case, a fixed forcing function $\sigma_0:[0,\infty) \to [0,\infty)$ and constants $\mu \in (0,1)$, $q > 1$ are used, and for any $(t,x) \in L$ and $p \in R^{n+1}$ such that $g'(t,x)p \geq 0$, $p \neq 0$, the steplength α and the relaxation factor ω are obtained as follows:

(i) If $g'(t,x)p = 0$, then $\hat{\alpha} = 0$; otherwise, let $\hat{\alpha} \geq 0$ be any real number such that

$$\hat{\alpha}\|p\| \geq \sigma_0(g'(t,x)\frac{p}{\|p\|})$$

(ii) If $(t,x) - \hat{\alpha}p \in \hat{J} \times D$ and

(18) $$g(t,x) - g((t,x)-\alpha p) \geq \mu\alpha g'(t,x)p$$

holds for $\alpha = \hat{\alpha}$, set $\omega = 1$. Otherwise, let ω be the largest number in the sequence $\{q^{-j} \mid j=1,2,\ldots\}$ such that $(t,x) - \omega\hat{\alpha}p \in \hat{J} \times D$ and (18) holds with $\alpha = \omega\hat{\alpha}$.

For this algorithm Ortega and Rheinboldt [1970] prove a theorem corresponding to Theorems 5.3 and 5.4 above. However, in this case L itself is assumed to be compact; and it can only be guaranteed that $(t,x) - \omega\hat{\alpha}p \in L$. Thus, unless the additional assumption that $L^0 = L$ is made, we can no longer guarantee that the iterates remain in the same path-connected component of L as the desired minimizer $(1,x(1))$. We shall not go into further details here.

We turn now to the convergence for our descent methods (7) in the setting of the previous section. Before using the specific direction vectors given by (11), we will first consider the more general class of gradient-related directions introduced by Elkin [1968]. These are defined by the property that for each step (t_k,x_k)

$$g'(t_k,x_k)\frac{p_k}{\|p_k\|} \geq \bar{\sigma}(\|g'(t_k,x_k)^T\|) \tag{19}$$

where $\bar{\sigma}$ is again a fixed forcing function.

The following theorem gives the basic result for the use of gradient-related directions in our setting.

5.5 Theorem. Assume that g satisfies the basic condition 5.2 and that, in addition, g has a unique critical point $(1,x^*) \in L^0$. Consider the descent method (7) where $\omega_k = 1$, and α_k is chosen either by (13), (16), or (17), and where $p_k \in R^{n+1}$, $p_k \neq 0$, is

such that (19) holds. Then $(t_k,x_k) \in L^0$, $k = 0,1,\ldots$, and $\lim_{k\to\infty} (t_k,x_k) = (1,x^*)$.

Proof. By Theorems 5.3 and 5.4, $(t_k,x_k) \in L^0$ for all $k = 0,1,\ldots$, and

$$(20) \qquad g(t_k,x_k) - g(t_{k+1},x_{k+1}) \geq \sigma(g'(t_k,x_k) \frac{p_k}{\|p_k\|}), \quad k = 0,1,\ldots .$$

Since L^0 is compact, g is bounded below on L^0 and hence $\lim_{k\to\infty} g(t_k,x_k)$ exists. But then (20) together with (19) and the definition of a forcing function imply that $\lim_{k\to\infty} g'(t_k,x_k)^T = 0$. For any convergent subsequence $\{(t_{k_i},x_{k_i})\}$ with, say, $\lim_{i\to\infty} (t_{k_i},x_{k_i}) = (\hat{t},\hat{x})$, it then follows by the continuity of g' that $g'(\hat{t},\hat{x})^T = 0$ and hence by the uniqueness of the critical point $(1,x^*)$ that $(\hat{t},\hat{x}) = (1,x^*)$. Since any subsequence of $\{(t_k,x_k)\}$ has a convergent sub-subsequence with this same limit, this proves that $\lim_{k\to\infty} (t_k,x_k) = (1,x^*)$.

Now consider the basic homotopy $H:\hat{J} \times D \subset R^1 \times R^n \to R^n$ which defines the descent curve x in the setting of the previous section. We will assume that H has the following properties.

5.6 Basic Conditions for H and x.

(i) The mapping $H:\hat{J} \times D \subset R^1 \times R^n \to R^n$ is continuously differentiable on the open set $\hat{J} \times D$ where $\hat{J} \supset [0,1]$, and $H_x(t,x)$ is nonsingular for all $(t,x) \in \hat{J} \times D$.

(ii) For a given $x_0 \in D$ there exists a continuous curve

$x:J \to D$ with $x(0) = x_0$ such that $H(t,x(t)) = 0$ for $t \in J$, and that $x(1)$ is the unique solution of $H(1,x) = 0$ in D.

In line with Section 1 we now choose a functional $f:\hat{J} \times R^n \to R^1$ and set

$$(21) \qquad g:\hat{J} \times D \subset R^1 \times R^n \to R^1,\ g(t,x) = f(t,H(t,x)).$$

Then we obtain as a corollary from Theorem 5.5:

5.7 Corollary. Suppose that the basic condition 5.6 holds for H and x, and that $f:\hat{J} \times R^n \to R^1$ is continuously differentiable, $\phi:J \to R^1$, $\phi(t) = f(t,0)$ is strictly antitone, and f has the unique critical point $(1,0)$. Moreover, assume that for the functional g of (21) the path-connected component L^0 of the level set (5) which contains $(0,x_0)$ is compact. Then the statement of Theorem 5.5 for the descent method (7) holds.

Proof. We only need to verify the conditions of Theorem 5.5. Clearly, the basic conditions 5.2 for the functional g are satisfied. Moreover, we have

$$g'(t,x) = f'(t,H(t,x))A(t,x)$$

where $A(t,x)$ has the matrix representation

$$A(t,x) = \begin{pmatrix} 1 & H_t(t,x) \\ 0 & \\ \vdots & H_x(t,x) \\ 0 & \end{pmatrix}.$$

Clearly, $A(t,x)$ is nonsingular and hence $g'(t,x) = 0$ implies that $f'(t,H(t,x)) = 0$. But then, by assumption, $(t,H(t,x)) = (1,0)$ and therefore we have $H(1,x) = 0$ which in turn implies that $x = x(1)$. Therefore $(1,x(1))$ is the unique critical point of g in $\hat{J} \times D$, and Theorem 5.5 applies.

We remark that the nonsingularity of H_x assumed in Condition 5.6 appears only quite weakly in the preceding result, namely, in proving that the critical points of g are the critical points of f. If the curve x is known to exist, and if H_x has a singularity for some $t < 1$ along the curve x, then the resulting critical point of g is only a saddle point and there exist well-defined descent directions since g decreases along the curve for increasing t. This appears to offer a possible new approach to overcome singularities of H_x which is different from previous attempts in dealing with this problem (see Davis [1966] and Meyer [1968]).

The conditions on f in this theorem are satisfied for the functional

$$(22) \qquad f:R^1 \times R^n \to R^n,\ f(t,x) = c_1 x^T x + c_2(t-1)^2,\ c_1,c_2 > 0,$$

where c_1,c_2 are given constants. In fact, f is continuously differentiable and $f'(t,x) = 0$ if and only if $(t,x) = (1,0)$. Moreover, $\phi:J \to R^1$, $\phi(t) = f(t,0)$, is strictly antitone for $t \in J$.

The next theorem shows that under our conditions for H the direction p defined by (11) is gradient-related.

5.8 Theorem. Suppose that H satisfies part (i) of the basic condition 5.6, and let g be given by (21) and (22). Then for any (t,x) in a compact subset $\hat{L}$ of $\hat{J} \times D$ the direction vector p defined by (10) and (11) satisfies $p(t,x) \neq 0$, unless $g'(t,x)^T = 0$, and

$$g'(t,x)\,\frac{p(t,x)}{\|p(t,x)\|_2} \geqslant \frac{1-\beta}{M}\,\|g'(t,x)^T\|_2^2, \tag{23}$$

with some constant M dependent only on g and $\hat{L}$.

Proof. Evidently

$$g_t(t,x) = 2c_1 H(t,x)^T H_t(t,x) + 2c_2(t-1)$$

and

$$g_x(t,x) = 2c_1 H(t,x)^T H_x(t,x).$$

Hence for $t \leqslant 1$

$$\begin{aligned} g'(t,x)p(t,x) &= 2\beta[-c_1 H(t,x)^T H_t(t,x) + 2c_2(1-t) \\ &\qquad + c_1 H(t,x)^T H_x(t,x) H_x(t,x)^{-1} H_t(t,x)] \\ &\quad + (1-\beta)\|g'(t,x)^T\|_2^2 \\ &= 2\beta c_2(1-t) + (1-\beta)\|g'(t,x)^T\|_2^2 \geqslant (1-\beta)\|g'(t,x)^T\|_2^2. \end{aligned} \tag{24}$$

For $t > 1$ we get the analogous expression with $(1-t)$ replaced by $(t-1)$. Since p is uniformly continuous on the compact set $\hat{L}$, we have $\|p(t,x)\| \leqslant M$ on $\hat{L}$ with some constant M. From (24) we see that

$p(t,x) = 0$ implies that $g'(t,x)^T = 0$, and hence for $p(t,x) \neq 0$, (23) holds.

Under the conditions of Corollary 5.7, the only critical point of g is $(1,x(1))$ and hence we always are assured of $p(t,x) \neq 0$ unless we actually reach the desired point.

Altogether now we have obtained the following result.

5.9 Theorem. Suppose that the basic condition 5.6 holds for H and x and that g is given by (21) and (22). Assume further that for g the component L^0 of the level set (5) which contains $(0,x_0)$ is compact. Consider the descent method (7) where $\omega_k = 1$ and α_k is chosen either by (13), (16), or (17) and where for $(t_k,x_k) \in L^0$, p_k is defined by $p_k = p(t_k,x_k)$ with p given by (10), (11). Then $(t_k,x_k) \in L^0$, $k = 0,1,\ldots$, and $\lim_{k\to\infty} (t_k,x_k) = (1,x^*)$.

Proof. We proceed by induction to prove $(t_k,x_k) \in L^0$. If this holds for $k = 0,1,\ldots,m$, then by Theorem 5.8, $g'(t_m,x_m)p(t_m,x_m) \geq 0$, and $p(t_m,x_m) \neq 0$ unless $(t_m,x_m) = (1,x(1))$ in which case the process stops. Hence by Theorem 5.3 or 5.4, $(t_{m+1},x_{m+1}) \in L^0$. Now we can apply Corollary 5.7 since p remains gradient-related as long as the process does not reach $(1,x(1))$ in finitely many steps.

The case $\beta = 1$ is excluded in the preceding theorem since in that case we can no longer guarantee that the direction $p(t,x)$ is gradient-related. Nevertheless, our numerical results in Sections 3 and 4 show that when the minimization algorithm converges, the convergence is best for $\beta = 1$. This is not unexpected since

this direction points in the downward direction of the "valley" whose "spine" is the curve z.

In the basic condition 5.6 the restriction that x(1) be the unique solution of $H(1,x) = 0$ in D can be relaxed in the following way.

5.10 Theorem. Assume that part (i) of the basic condition 5.6 holds for the homotopy H and that x satisfies:

(ii)' For given $x_0 \in D$ there exists a continuous curve $x:J \to D$ with $x(0) = x_0$ such that for each $t \in J$, $H(t,x(t)) = 0$ and $x(t)$ is the unique solution of $H(t,x) = 0$ in $\bar{S}(x(t),r) \subset D$ where r is independent of t.

Then for g given by (21) and (22) there exists a choice of c_1 and c_2 such that $(1,x(1))$ is the unique critical point of g in the path-connected component L_J^0 of $L_J = \{(t,x) \in L^0 \mid t \in J\}$.

Proof. Define $B = \{(t,x) \mid t \in J, x \in \bar{S}(x(t),r)\}$, $B' = \{(t,x) \mid t \in J, x \in \partial\bar{S}(x(t)r)\}$. Since B' is compact and $x(t)$ is unique in $\bar{S}(x(t),r)$, there exists a constant $d > 0$ such that

$$(25) \qquad d = \min \{\|H(t,x)\|_2^2 \mid (t,x) \in B'\}.$$

Let $c_1 \geq 1$ and choose $c_2 > 0$ such that $(c_2/c_1) < d$. Then we claim that $L_J^0 \subset B$. Assume that this is not the case; then there exists a point $(\hat{t},\hat{x}) \in L_J^0 \setminus B$. Let $q:[0,1] \to L_J^0$ be a path from $(0,x(0))$ to $(\hat{t},\hat{x}) \in L_J^0$ and let π_1 and π_2 denote the natural projections

from R^{n+1} onto R^1 and R^n, respectively. Then, since $q(s) \in L_J^0$ for all $s \in [0,1]$, it follows that

$$c_2 = g(0,x(0)) \geq g(\pi_1 q(s),\pi_2 q(s))$$

$$= c_1 \|H(\pi_1 q(s),\pi_2 q(s))\|_2^2 + c_2(\pi_1 q(s)-1)^2$$

$$\geq c_1 \|H(\pi_1 q(s),\pi_2 q(s))\|_2^2$$

from which it follows that

$$\|H(\pi_1 q(s),\pi_2 q(s))\|_2^2 \leq \frac{c_2}{c_1} < d. \tag{26}$$

Since $\hat{t} \in J$, $(\hat{t},\hat{x}) \notin B$ implies that $\hat{x} \notin \bar{S}(x(\hat{t}),r)$, and thus $q(\hat{s}) \in B'$ for some $\hat{s} \in (0,1)$. By (26) we therefore have $\|H(\pi_1 q(s),\pi_2 q(s))\|_2^2 < d$, which contradicts (25). Hence $L_J^0 \subset B$, and in particular $\{(1,x) \in L_J^0\} \subset \bar{S}(x(1),r)$. Since all critical points of g lie on the hyperplane $t = 1$ and $x(1)$ is the unique root of $H(1,x)$ in $\bar{S}(x(1),r)$, the theorem is proved.

With this theorem it is natural in practical applications to vary the ratio c_2/c_1 so that when the solutions of $H(1,x) = 0$ are isolated in D the point $(1,x(1))$ is forced to be the unique solution in L^0. The restriction to $t \in J$ in Theorem 5.10 can be overcome by introducing into the functional g a penalty function for $t > 1$. For example, we might redefine $g:\hat{J} \times D \subset R^1 \times R^n \to R^1$ as follows:

$$g(t,x) = \begin{cases} c_1 H(t,x)^T H(t,x) + c_2(t-1)^2, & t \leq 1 \\ c_1 H(t,x)^T H(t,x) + \hat{c}(t-1)^2, & t > 1 \end{cases}$$

where $\hat{c}$ is large.

Section 3. Numerical Solution of the Model Problem.

We consider again our model problem in the computational framework of this chapter. Recall that with the parameter $t \in J$ we introduce the family of two-point boundary value problems given by

$$(27) \qquad \begin{cases} (s^2 d(x)x')' = ts^2 f(x) & s \in (0,1),\ t \in J \\ x'(0) = 0 \quad , \quad x(1) = 1 \end{cases}$$

where for fixed $k > 0$,

$$(28) \qquad f:\ D_k = (-\frac{k}{2}\ ,\ \infty\ ) \subset R^1 \to R^1,\ f(x) = \frac{1}{\varepsilon}\frac{x}{x+k}$$

and $d:\ D_k \to R^1$ represents a continuously differentiable diffusion-coefficient function. As suggested by Keller [1968], we will use, in particular, d given by

$$(29) \qquad d(x) = 1 + \frac{\lambda}{(x+c)^2} \quad , \quad \lambda\ ,\ c > 0,$$

where we assume $c \geq k$ so that d and f are defined on the same domain.

In order to solve (27) numerically we consider a finite difference approximation based on a standard central difference scheme as follows: For given $n > 1$, let $h = 1/n$ and $s_j = (j-1)h$, $j = 1, \ldots, n+1$. If x_j denotes the approximation to $x(s_j)$, we approximate $s_j^2\, d(x(s_j))x'(s_j)$ by

(30) $\frac{1}{h}\ s_j^2\ d(x_j)(x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}) \equiv v_j$

and $(s_j^2\ d(x(s_j))x'(s_j))'$ by

$$\frac{1}{h}\,(v_{j+\frac{1}{2}} - v_{j-\frac{1}{2}})$$

where

$$s_{j\pm\frac{1}{2}} = (j-1\pm\tfrac{1}{2})h\ ;\ d(x_{j\pm\frac{1}{2}}) = d\Big(\frac{x_{j\pm 1} + x_j}{2}\Big).$$

Then we have to solve the non-linear system of equations

(31) $$\begin{cases} \frac{1}{h^2}\,[s^2_{j+\frac{1}{2}}\ d(x_{j+\frac{1}{2}})(x_{j+1}-x_j) - s^2_{j-\frac{1}{2}} d(x_{j-\frac{1}{2}})(x_j-x_{j-1})] = ts_j^2 f(x_j), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1,2,\ldots,n \\ \text{where } x_o = x_1,\ x_{n+1} = 1 \end{cases}$$

Let

(32) $$a_j^{\pm}(x) = s^2_{j\pm\frac{1}{2}}\ d(x_{j\pm\frac{1}{2}})$$

and

$$D_k^n = \{x \in R^n \ \mid x_i \geq -\tfrac{k}{2}\ ,\ 1 \leq i \leq n\}\ .$$

Then define the matrix function $A: D_k^n \to L(R^n,R^n)$ by

(33) $$A(x) = \frac{1}{h^2}\begin{pmatrix} a_1^+(x) & -a_1^+(x) & 0 & \cdots & 0 \\ -a_2^-(x) & a_2^-(x)+a_2^+(x) & -a_2^+(x) & & \vdots \\ 0 & -a_3^-(x) & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & a_{n-1}^-(x) \\ 0 & \cdots & 0 & -a_n^-(x) & a_n^-(x)+a_n^+(x) \end{pmatrix}$$

and set

$$(34) \qquad F: D_k^n \to R^n, \quad F(x) = (s_1^2 f(x_1),\ldots,s_n^2 f(x_n))^T \quad ,$$

$$(35) \qquad b: D_k^n \to R^n, \quad b(x) = \frac{1}{h^2} (0,\ldots,0,a_n^+(x))^T \ .$$

With this notation the problem (31) has the form

$$(36) \qquad -A(x)x = t\, F(x) - b(x).$$

As in Chapter II, we assume that $d(x) \geq \alpha > 0$ for $x \geq 0$. Then, for any $x \geq 0$, the matrix $A(x)$ is irreducibly diagonally dominant and thus has a positive inverse (see Varga [1962;p85]); in fact, $A(x)$ is even a Stieltjes matrix. Therefore any solution $x \geq 0$ of (36) must also satisfy

$$(37) \qquad x = A(x)^{-1}b(x) - tA(x)^{-1}F(x).$$

This is now in the fixed point form which, in the case of the continuous problem (27), allowed us to obtain existence and uniqueness theorems. In an analogous manner as before in the Banach space setting, we shall establish the corresponding existence and uniqueness results for the finite dimensional problem (37).

From (33) it follows immediately that $A(x)e = b(x)$ for $e = (1,\ldots,1)^T \in R^n$, and hence that (37) can be written as

$$(38) \qquad x = e - t\, A(x)^{-1}F(x).$$

Since $A(x)^{-1} \geq 0$ and $F(x) \geq 0$ for all $x \geq 0$, any solution $x^* \geq 0$ of (38) must satisfy $x^* \leq e$.

We next show that for certain values of k and ε, $x \geq 0$ implies that

$$(39) \qquad 0 \leq t\, A(x)^{-1}F(x) \leq e.$$

For any $x \geq 0$, set

$$(40) \qquad u(x) = A(x)^{-1}F(x),\ u(x) = (u_1,\ldots,u_n)^T \ .$$

Then

$$(41) \qquad \begin{cases} s_1^2 f(x_1) = \dfrac{1}{h^2}\ a_1^+(x)(u_1-u_2) \\ s_i^2 f(x_i) = \dfrac{1}{h^2}\ [a_i^-(x)(u_i-u_{i-1})-a_i^+(x)(u_{i+1}-u_i)] \qquad 2 \leq i \leq n-1 \\ s_n^2 f(x_n) = \dfrac{1}{h^2}\ [a_n^-(x)(u_n-u_{n-1})+a_n^+(x)u_n]. \end{cases}$$

Let $\beta_j(x) = h^2 s_j^2\ f(x_j)$ and note that $a_i^+(x) = a_{i+1}^-(x)$.

Then

$$\sum_{i=1}^{j} \beta_i(x) = -a_j^+(x)(u_{j+1}-u_j), \qquad 1 \leq j \leq n-1$$

$$\sum_{i=1}^{n} \beta_i(x) = a_n^+(x)u_n,$$

and hence back-substitution gives,

$$(42) \qquad u_{n-j}(x) = \sum_{m=0}^{j} \left(\frac{\sum_{i=1}^{n-m} \beta_i(x)}{a_{n-m}^+(x)} \right), \quad j = 0,\ldots,n-1.$$

From this we readily obtain the bound

$$(43)\qquad |u_{n-j}| \le \frac{h^2}{6\alpha\varepsilon} \sum_{m=0}^{j} \frac{(n-m-1)(2n-2m-1)(n-m)}{(n-m-\frac{1}{2})(n-m-\frac{1}{2})}$$

$$= \frac{h^2}{6\alpha\varepsilon}\, 2 \sum_{m=0}^{j} (n-m)\left(\frac{n-m-1}{n-m-\frac{1}{2}}\right), \quad j=0,\ldots,n-1,$$

where we used again that $d(x) \ge \alpha > 0$ for $x \ge 0$.

With the help of the inequality

$$(44)\qquad \frac{n-m-1}{n-m-\frac{1}{2}} = 1 - \frac{1}{2(n-m-\frac{1}{2})} \le 1 - \frac{1}{2n-1}, \quad 0 \le m < n,$$

(43) can be simplified to

$$(45)\qquad |u_{n-j}| \le \frac{h^2}{6\alpha\varepsilon}\,[2nj + 2n - j^2 - j]\,(1 - \frac{1}{2n-1}), \quad j=0,\ldots,n-1,$$

and hence it follows that

$$(46)\qquad \|u\|_\infty \le \frac{h^2}{6\alpha\varepsilon}(n+\tfrac{1}{2})^2(1-\frac{1}{2n-1}) = \frac{1}{6\alpha\varepsilon}(1+\frac{h}{2})^2(1-\frac{1}{2n-1})$$

$$< \frac{1}{6\alpha\varepsilon}(1+\frac{h}{2})^2$$

We state the existence result as a theorem.

5.11 Theorem. Consider the problem (38) with A and F defined by (33) and (34). Assume that $6\alpha\varepsilon \ge (1+\frac{h}{2})^2$, where $h = 1/n$ is the mesh size in the discretization of (27). Then (38) has a solution for each fixed $t \in J$.

Proof: Define $G: J \times D_k^n \subset J \times R^n \to R^n$ by

$$(47)\qquad G(t,x) = e - t\,A(x)^{-1}F(x).$$

Clearly $G(t,x) < e$ for $x \geq 0$ and $t \in J$. Moreover, with u defined by (40) it follows trom (46) that $G(t,x) \geq 0$ for $6\alpha\varepsilon \geq (1+\frac{h}{2})^2$. Thus $G(t,\cdot)$ maps the compact convex set $B = \{x \in R^n \mid 0 \leq x \leq e\}$ into itself for each fixed t, and by Brouwer's fixed point theorem, $G(t,.)$ must have a fixed point which is clearly a solution of (38). This completes the proof.

Note that the estimate in Theorem 5.11 compares favorably to the estimate in Theorem 3.11.

We can also establish uniqueness of the solution under conditions similar to those used in Chapters II and III. In particular, if $G(t,.)$ is strictly non-expansive on B, for each t, then the uniqueness is immediate. The condition of strict-non-expansivity is evidently equivalent with

$$(48) \qquad \|A(x)^{-1}F(x) - A(y)^{-1}F(y)\|_\infty < \|x-y\|_\infty, \ \forall x,y \in B$$

and using (40), we see that (48) is satisfied if

$$(49) \qquad |u_{n-j}(x) - u_{n-j}(y)| < \max_{1<i<n} |x_i - y_i|, j=0,\ldots,n-1.$$

Now for $0 \leq m \leq j \leq n-1$, $n > 1$,

$$\left| \frac{\sum_{i=1}^{n-m} \beta_i(x)}{a^+_{n-m}(x)} - \frac{\sum_{i=1}^{n-m} \beta_i(y)}{a^+_{n-m}(y)} \right|$$

$$= \frac{h^4}{s^2_{n-m+\frac{1}{2}}} \left| \frac{\sum_{i=1}^{n-m} (i-1)^2 f(x_i)}{d(x_{n-m+\frac{1}{2}})} - \frac{\sum_{i=1}^{n-m} (i-1)^2 f(y_i)}{d(y_{n-m+\frac{1}{2}})} \right|$$

$$= \frac{h^4}{\varepsilon s^2_{n-m+\frac{1}{2}}} \left| \sum_{i=1}^{n-m} (i-1)^2 \left[\frac{x_i}{x_i+k} \Big/ d(x_{n-m+\frac{1}{2}}) - \frac{y_i}{y_i+k} \Big/ d(y_{n-m+\frac{1}{2}}) \right] \right|$$

$$= \frac{h^4}{\varepsilon s^2_{n-m+\frac{1}{2}}} \left| \left[\frac{1}{d(x_{n-m+\frac{1}{2}})} - \frac{1}{d(y_{n-m+\frac{1}{2}})} \right] \sum_{i=1}^{n-m} (i-1)^2 \frac{x_i}{x_i+k} \right.$$

$$\left. - \frac{1}{d(y_{n-m+\frac{1}{2}})} \sum_{i=1}^{n-m} (i-1)^2 \left[\frac{y_i}{y_i+k} - \frac{x_i}{x_i+k} \right] \right|$$

$$\leq \frac{h^4}{s^2_{n-m+\frac{1}{2}}} \left\{ \sup_{0<s<1} \left| \frac{d'(s)}{d^2(s)} \right| \left[\max_{1<i<n} |x_i - y_i| \right] \left[\frac{1}{1+k} \frac{(n-m-1)(n-m)(2n-2m-1)}{6} \right] \right.$$

$$\left. + \frac{1}{k\alpha} \left[\max_{1<i<n} |x_i - y_i| \right] \frac{(n-m-1)(n-m)(2n-2m-1)}{6} \right\} .$$

Therefore, letting $\beta \geq \sup_{0<s<1} \left| \frac{d'(s)}{d^2(s)} \right|$ as in Chapter II, we obtain using (44),

$$|u_{n-j}(x) - u_{n-j}(y)| < \frac{h^2}{6\varepsilon} \sum_{m=0}^{j} \left[\frac{\beta}{1+k} \frac{(n-m-1)(n-m)(2n-2m-1)}{(n-m-\frac{1}{2})(n-m-\frac{1}{2})} \right.$$

$$\left. + \frac{1}{\alpha k} \frac{(n-m-1)(n-m)(2n-2m-1)}{(n-m-\frac{1}{2})(n-m-\frac{1}{2})} \right] \|x-y\|_\infty .$$

$$< \frac{h^2}{6\varepsilon} \left[\frac{\beta}{1+k} + \frac{1}{\alpha k} \right] [2nj + 2n - j^2 - j] \left[1 - \frac{1}{2n-1} \right] \|x-y\|_\infty .$$

and hence

$$(50) \quad \|u(x) - u(y)\|_\infty \leq \frac{1}{6\varepsilon}\{[\frac{\beta}{1+k} + \frac{1}{\alpha k}](1+\frac{h}{2})^2(1-\frac{1}{2n-1})\}\|x-y\|_\infty.$$

Thus if $6\varepsilon\alpha \geq (1+\frac{h}{2})^2$ as in Theorem 5.11, $G(t,\cdot)$ will even be contractive provided that

$$(51) \quad \frac{\beta\alpha}{1+k} + \frac{1}{k} = \gamma < 1,$$

Accordingly, we can state the following theorem.

5.12 Theorem. Under the conditions of Theorem 5.11, let (51) be valid with $\beta \geq \sup_{0<s<1} \left| \frac{d'(s)}{d^2(s)} \right|$. Then for each fixed $t \in J$, the solution of (38) is unique in the set B. Furthermore, the solution curve $x: J \to B$ is continuous.

Proof: The uniqueness for each $t \in J$ follows from the strict nonexpansivity of $G(t,\cdot)$.

Define

$$(52) \quad H: J \times D_k^n \to R^n, \; H(t,x) = e-x-t\,A(x)^{-1}F(x),$$

then the equation

$$(53) \quad H(t,x) = 0$$

is equivalent to (36). With the notation (40) we have

$$(54) \quad H(t,x) = e - x - t\,u(x).$$

and from (42) we see that $u'(x)$ and hence also $H_x(t,x)$ exists.

Moreover, it follows from the Lipschitz condition (50) that for any $z \in R^n$ and $\rho > 0$ there exists a $\sigma = \sigma(z,\rho) > 0$ such that

$$\| u'(x)\sigma z \|_\infty < \|u(x+\sigma z) - u(x) - u'(x)\sigma z \|_\infty + \|u(x+\sigma z) - u(x)\|_\infty$$

$$< (\gamma + \rho)\| \sigma z\|_\infty$$

Since $\rho > 0$ is arbitrary and $\gamma < 1$ we see that $\| u' \|_\infty < 1$ from which it follows by Lemma 2.12 that $H(t,x)^{-1}$ exists on $J \times B$. Now by Theorem 3.6, the solution curve $x: J \to B$ is continuous.

In the numerical computations we employed the numerically simpler homotopy $\hat{H}(t,x) = A(x)H(t,x)$ to carry out the numerical continuation process based on minimization as described in Sections 1 and 2 of this chapter. The invertibility of $\hat{H}_x(t,x(t))$ is not guaranteed; however, since the path $x: J \to R^n$ exists, the remark following Corollary 5.7 applies and descent directions exist, allowing the algorithm to continue at each step.

The steplength algorithm employed in the numerical computations was an application of the Goldstein-Armijo idea, described following Theorem 5.4, to the minimization process (17) with L^0 replaced by L. To be precise, with the functional g defined by (21) and (22) and with a descent direction $-p \in R^{n+1}$, $\| p\| = 1$, at the point $(t,x) \in J \times R^n \subset R^{n+1}$, we set $\alpha = 1$. Then ω_0 is chosen as the first element in the sequence 2^{-j}, $j = 0,1,\ldots$, such that

$$g((t,x) - \omega_0 p) < g(t,x),$$

and with $\omega_1 = \omega_0/2$, the relaxation factor ω was selected so that

$$g((t,x) - \omega p) = \min \{ g((t,x) - \omega_0 p),\ g((t,x) - \omega_1 p) \} .$$

Finally, we used the additional information that $x(t) \geq 0$ and added a "penalty function" to g and considered the functional $\hat{g}: J \times D_k^n \to R^1$ given by

$$\hat{g}(t,x) = g(t,x) + 1000 \langle v, v \rangle$$

where $v \in R^n$ is such that $v_i = \min \{x_i, 0\}$.

The computations were done using the descent algorithm until $|t-1| < .05$. At this point Newton's method with $t = 1$, was used until successive iterates agreed to 6 significant digits. The starting vector for all cases was $(t^0, x^0) = (0,1,\ldots,1)^T \in R^{n+1}$. With $x^0 = (1,\ldots,1)^T \in R^n$, Newton's method does not converge to the desired positive solution for any of the cases tried.

The computer program was written in BASIC and run on the G.E. Mark II Timesharing system. The conversational mode allowed us to experiment freely with various parameters and to interrupt the computations when non-convergence or overly slow convergence was apparent. With minor modifications, however, the same program could have been run in a batch environment.

Table 5.1 shows five groups of test cases for the model problem (38). Group 1 shows the results for values of the parameters k and ε for which our estimates (of Chapter II) guaranteed existence and continuity of the solution curve x(t). Groups 1 through 4 are for the case when $\lambda = 0$, i.e., when the function d(x) is the constant 1. Groups 2 through 4 show values of the parameters ε and k outside the range from which existence of the solution curve x is guaranteed.

Case No.	Constants for f		Constants for d		Constants for g			Iterates initial	Iterates terminal	Comments
	ε	k	λ	c	c_1	c_2	β			
1	.167	.5	0	-	1	1	1	-	-	t = 1.144
2	.167	.5	0	-	1	1	.95	3	3	
3	.167	.5	0	-	1	1	.75	3	3	
4	.17	.001	0	-	1	1	1	3	3	
5	.17	.001	0	-	1	1	.8	3	3	
6	.1	.1	0	-	1	1	1	-	-	t = .9024
7	.1	.1	0	-	1	10	1	-	-	t = .9193
8	.1	.1	0	-	10	100	1	5	5	
9	.1	.1	0	-	10	100	.5	-	-	t = .1791
10	.05	.1	0	-	1	1	1	-	-	t = .4135
11	.05	.1	0	-	10	10	1	10	5	
12	.05	.1	0	-	10	10	.9	-	-	t = .8276
13	.01	.01	0	-	1	100	1	-	-	t = .1346
14	.01	.01	0	-	100	100	1	-	-	t = .1346
15	.01	.01	0	-	10	1	1	-	-	t = .0673
16	.01	.01	0	-	100	10	1	27	6	
17	.1	.1	.01	.01	1	1	1	5	5	
18	.01	.05	.01	.01	1	1	1	12	5	
19	.01	.01	.01	.01	1	1	1	17	5	
20	.01	.01	.1	.1	1	1	1	20	6	

Table 5.1 Computational Results for the Model Problem.

Group 5 contains cases with $\lambda > 0$. Whenever a value of t is indicated in the final column, this indicates that the algorithm failed to determine a steplength parameter $\omega > 2^{-15}$ at the value of the parameter t indicated.

Group 1 shows generally that within the range where the existence of the curve x is known, the computations were well-behaved with the exception of case 1. This case, however, points out the problem discussed in the remarks following Theorem 5.9, namely, that near t = 1, the direction given by (10) and (11) with $\beta = 1$ pointing along the "spine of the valley" is nearly orthogonal to the direction pointing toward x(1).

The Groups 2 through 4 show the importance of the proper choice of the constants c_1 and c_2 in (22). Cases 9 and 12 show that even when c_1 and c_2 are properly chosen so that convergence occurs with $\beta = 1$, a decrease in the value of β may prove fatal. These three groups also show the ever increasing number of iterates required as the parameters ε and k become small. This behavior can be explained by comparing Table 5.2 and noticing that as ε and k become small, the solution vector approaches a function which is nearly zero on (0,1) and has a steep slope near s = 1. Since the starting vector was constant for all cases, as ε and k decrease x^0 is further and further away from the solution vector.

Finally, Group 5 shows a similar behavior for $\lambda > 0$ except that the sensitivity to the choice of parameters c_1 and c_2 seemed to be less pronounced.

parameters			
ε	.167	.1	.01
k	.5	.1	.01
λ	0	0	0
c	-	-	-

solutions			
x_1	.467150	2.97454×10^{-2}	1.08747×10^{-16}
x_2	.467150	2.97454×10^{-2}	1.08747×10^{-16}
x_3	.480005	3.99347×10^{-2}	4.94197×10^{-15}
x_4	.503404	6.18672×10^{-2}	3.22968×10^{-13}
x_5	.537413	.101138	2.42135×10^{-11}
x_6	.582496	.164624	1.95183×10^{-9}
x_7	.639305	.258537	1.64550×10^{-7}
x_8	.708610	.387218	1.43016×10^{-5}
x_9	.791248	.553103	1.26897×10^{-3}
x_{10}	.888089	.757271	.101995
	Cases 1-3	Case 8	Case 16

Table 5.2 Sample Numerical Solutions
to the Model Problem for h = .1.

On the average, 4 to 6 functional evaluations were required per iterate in the steplength algorithm. This number can vary widely, however, with different steplength algorithms.

Table 5.3 shows the convergence of the intermediate iterates for Case 2 and was typical of the rapid convergence for most of the cases tried.

Keller [1968] used a technique similar to the basic numerical continuation process described in Chapter II with Newton's method as the local iterative process. For example, to obtain the solution for our case 16, Keller used the parameter ε as the continuation parameter with 19 intermediate values from $\varepsilon = .1$ (taking that solution as the initial vector) to $\varepsilon = .01$ and with at least 5 Newton steps per parameter step. Thus at least 95 iterates were computed as compared with our 33. Keller, however, used a much finer mesh spacing with $h = 10^{-3}$ which we could not attempt with our timesharing program.

	(t',x')	(t^2,x^2)	(t^3,x^3)	(t^4,x^4)*	(t^5,x^5)*	(t^6,x^6)*
t	.543191	.859111	1.03507	1.00000	1.00000	1.00000
x_1	.663990	.507430	.434776	.466971	.467150	.467150
x_2	.663990	.507430	.434776	.466971	.467150	.467150
x_3	.673145	.520306	.449075	.479840	.480005	.480005
x_4	.689626	.543570	.474991	.503261	.503404	.503404
x_5	.713168	.576995	.512409	.537297	.537413	.537413
x_6	.743687	.620649	.561584	.582407	.582496	.582496
x_7	.781142	.674709	.622921	.639241	.639305	.639305
x_8	.825511	.739404	.696902	.708567	.708610	.708610
x_9	.876782	.814997	.784050	.791223	.791248	.791248
x_{10}	.934947	.901796	.884961	.888078	.888089	.888089

$t_o = 0$, $x^o = (1,1,\ldots,1)^T$

* terminal iterates with t = 1 using Newton's Method.

Table 5.3 Iterates Showing Convergence (Case 2)

Section 4. Other Numerical Experiments.

Several other examples were investigated using numerical continuation by minimization. These results are summarized in Table 5.4. The first example

$$x_1^2 + x_2^2 + x_3^2 = 5$$

$$x_1 + x_2 = 1$$

$$x_1 + x_3 = 3$$

is due to Deist and Sefor [1967], and we used the imbedding

$$x_1^2 + tx_2^2 + tx_3^2 = 5$$

$$x_1 + tx_2 = 1$$

$$x_1 + x_3 = 3$$

and the initial vector $(\sqrt{5} , 1- \sqrt{5}, 3- \sqrt{5})^T$. The table shows the dependence on β ; in fact for $\beta = .5$ or $\beta = .9$ the minimization process failed to find a decrease in the functional g of (21/22) in the direction given by the algorithm (7) for $\alpha_k = 1$, $\omega_k > 2^{-15}$ and the values of t indicated. A comparison of the first case and the last shows the effect of altering the constants c_1 and c_2 in (22).

Deist and Sefor used a generalized numerical continuation process by using Newton's method both as a predictor and a corrector in the integration of the differential equation underlying the homotopy.

	Constants for g c_1	c_2	β	Iterates initial	terminal	Comments
	1	1	1	15	6	
Deist &	1	1	.5	-	-	t = .0026
Sefor	1	1	.9	-	-	t = .706
	1	10	.98	8	4	
	1	1	1	1	6	
	1	1	0	2	5	
Brown	1	10	.8	5	5	
	10	1	.8	-	-	t = .8133
	1	5	.95	3	7	

Table 5.4 Computational results for two additional examples.

Their procedure is contained in Theorem 4.14 where the function G is the Newton iteration function. Our total number of iterates compares favorably with their requirement of 19 iterates needed to obtain the same solution with the same degree of accuracy.

The example of Brown [1966] is taken from a paper of Broyden [1969] and is given by:

$$h_1(t,x_1,x_2) = \frac{1}{2}\sin(x_1x_2) - x_2/(4\pi) - x_1/2 + (t-1)c_1 = 0$$

$$h_2(t,x_1,x_2) = (1 - 1/(4\pi))(e^{2x_1}-e) + ex_2/\pi - 2ex_1+(t-1)c_2 = 0$$

where $c_i = h_i(1,.6,3.0)$, $i = 1,2$. Brown did not use an imbedding technique and the homotopy is that of Broyden. The number of function evaluations required by Brown and Broyden to reduce the norm of $(h_1(1,x_1,x_2), h_2(1,x_1,x_2))^T$ to 10^{-6} was in their best case equal to 10 for Brown and to 14 for Broyden. This compares favorably with our first case which required only 7 iterates to obtain a difference between successive iterates of 10^{-6} in each component.

Notice in this example that the dependence on β seems minimal, Case 2 shows that even with the simple gradient directions ($\beta = 0$) convergence is still quite rapid. This seems to be due to the fact that the initial approximation is already sufficiently good.

Finally, Table 5.5 shows some numerical experiments for these same two problems using the basic numerical continuation process (2.35a,b) described in Chapter II with Newton's method as the local iterative process.

	ε	c	H Evaluations	Iterates	min Δt
Deist & Sefor	1.0	2	37	28	.0804
	.1	10	69	40	.0731
Brown	.1	10	6	7	1.0

Table 5.5 Computational results using numerical continuation with Newton as the local process.

For each k, j_k was determined by requiring $\| H(t_k, x_k^{j_k}) \| < \varepsilon$, and the parameter steplength $(t_{k+1} - t_k)$ was determined by the requirement that $\| H(t_{k+1}, x_k^{j_k}) \| < c\varepsilon$ for some $c > 1$. For $t = 1, \varepsilon$ was chosen as 10^{-6}.

For Brown's example, an initial steplength of 1 was possible so that the numerical continuation method simply reduced to Newton's method. Accordingly, no further testing was undertaken.

Another example due initially to Freudenstein and Roth [1963] was attempted. It consists in the following equations.

$$h_1(t,x_1,x_2) = t[-13 + x_1 - 2x_2 + 5x_2^2 - x_2^3] + (1-t)[-71 + x_1 - 50x_2 - 13x_2^2 - x_2^3] = 0$$

$$h_2(t,x_1,x_2) = t[-29 + x_1 - 14x_2 + x_2^2 + x_2^3] + (1-t) [129 + x_1 + 106x_2 + 19x_2^2 + x_2^3] = 0.$$

Both the numerical continuation process discussed in Chapter II and the process of numerical continuation by minimization failed to produce a sequence of iterates converging to the solution $(5,4)^T$ at $t = 1$ starting with the initial vector of $(15,-2)^T$ at $t = 0$. Subsequent analysis shows, however, that $H_x(t,x)$ is singular for $t \doteq .93$, $x_2 \doteq .564$ and all values of x_1. Thus the theory also breaks down.

One other example which failed involved the often tested derivative of the Rosebrock functional (see Rosebrock [1960]). One explanation might be that the continuation-by-minimization procedure we used was that described earlier in this section for the functional g defined by (21) and (22). In particular, the direction algorithm (10), (11)

is still rather unsophisticated, even if it utilizes the important direction of the solution curve $x(t)$.

In summary, several observations can be made together with some conjectures. First of all, the consistent indication that $\beta = 1$ in (10),(11) gives the best direction algorithm, indicates that the use of the gradient direction $(g')^T$ in (11) added little in helping to determine the best descent direction. Perhaps the use of a conjugate gradient direction in place of $(g')^T$ might produce better results. In addition, the fact that each iterate (t_k,x_k) is an approximation to $(t,x(t))$ in the minimization algorithm suggests that a simple interpolating function might be used to take into account the curvature of the "spine of the valley" and not only its direction.

REFERENCES

Anselone, P., and Moore, R. [1966]. An extension of the Newton-Kantorovich method for solving nonlinear equations with an application to elasticity, J. Math. Anal. Appl. 13, 476-501.

Bittner, L. [1967]. Einige Kontinuierliche Analogien von Iterationsverfahren, in Funktionalanalysis, Approximationstheorie, Numerische Mathematik, ISNM 7, pp. 114-135. Birkhauser-Verlag. Basel.

Bosarge, W. [1968]. Infinite dimensional iterative methods and applications, IBM Houston Sci. Center Rept. 320.2347. Houston, Texas.

Brown, K. [1966]. A Quadratically Convergent Method for Solving Simultaneous Nonlinear Equations. Ph.D. Diss., Purdue Univ., Lafayette, Indiana.

Broyden, C. [1969]. A new method of solving nonlinear simultaneous equations, Comput. J. 12, 94-99.

Davidenko, D. [1953a]. On a new method of numerically integrating a system of nonlinear equations (Russian), Dokl. Akad. Nank SSSR 88,601-604.

Davidenko, D. [1953b]. On the approximate solution of a system of nonlinear equations (Russian), Ukrain. Mat. Z. 5, 196-206.

Davis, J. [1966]. The Solution of Nonlinear Operator Equations with Critical Points. Ph.D. Diss., Oregon State Univ., Corvallis, Oregon.

Deist, F., and Sefor, L. [1967]. Solution of systems of nonlinear equations by parameter variation, Comput. J. 10, 78-82.

Dunford, N., and Schwartz, J. [1958]. Linear Operators I. Wiley (Interscience), New York.

Elkin, R. [1968]. Convergence Theorems for Gauss-Seidel and Other Minimization Algorithms. Ph.D. Diss., Univ. of Maryland, College Park, Maryland.

Faddeev, D., and Faddeeva, V. [1960]. Computational Methods of Linear Algebra, Fizmatgiz, Moscow; transl. by R. Williams, Freeman, San Francisco, California, 1963.

Ficken, F. [1951]. The continuation method for functional equations, Comm. Pure Appl. Math. 4,435-456.

Freudenstein, F., and Roth, B. [1963]. Numerical solution of systems of nonlinear equations, J. Assoc. Comput. Mach. 10, 550-556.

Friedrichs, K. O. [1950]. Functional Analysis (mimeographed lecture notes), Institute for Mathematics and Mechanics, New York Univ.

Gavurin, M. K. [1958]. Nonlinear functional equations and continuous analogs of iterative methods (Russian), Izv. Vyš. Učebn. Zaved. Matematica 6, 18-31. English translation by Mirijana Kocho, edited by J. H. Avila, TR 68-70, Univ. of Maryland, June 1968.

Henrici, P. [1962]. Discrete Variable Methods in Ordinary Differential Equations, John Wiley and Sons.

Householder, A. [1964]. The Theory of Matrices in Numerical Analysis Ginn(Blaisdell) Boston, Massachusetts.

Keller, H. B. [1968]. Numerical Methods for Two-Point Boundary Value Problems, Blaisdell, New York.

Kleinmichel, H. [1968]. Stetige Analoga and Iterationsverfahren für nichtlineare Gleichungen in Banachräumen, Math. Nachr. 37,313-344.

Lahaye, E. [1934]. Une méthode de résolution d'une catégorie d'équations transcendantes, C. R. Acad. Sci. Paris 198, 1840-1842.

Lahaye, E. [1935]. Sur la représentation des racines systèmes d'équations transcendantes, Deuxième Congrès National des Sciences 1,141-146.

Lahaye, E. [1948]. Solution of systems of transcendental equations, Acad. Roy. Belg. Bull. Cl. Sci. 5,805-822.

Meyer, G. [1968]. On solving nonlinear equations with a one-parameter operator imbedding, SIAM J. Numer. Anal. 5, 739-752.

Murray, J. D. [1968]. A Simple Method for Obtaining Approximate Solutions for a Class of Diffusion-Kinetics Enzyme Problems: I. General Class and Illustrative Examples, Math. Biosciences 2, 379-411.

Ortega, J. M., and Rheinboldt, W. C. [1968]. Local and Global Convergence of Generalized Linear Iterations, Studies in Num. Anal. 2, Soc. Ind. Appl. Math., Philadelphia, Pennsylvania.

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

Ostrowski, A. [1966], [1960]. Solution of Equations and Systems of Equations, Academic Press, New York, 1960, second edition, 1966.

Polyak, B. T. [1964]. Some methods of speeding up the convergence of iterative methods (Russian), Ž Vyčisl. Mat. i Mat. Fiz. 4, 791-803. English translation by H. F. Cleaves in U.S.S.R. Comput. Math. and Math. Phys. 4, 1964.

Rheinboldt, W. C. [1969]. Local mapping relations and global implicit function theorems, Trans. Amer. Math. Soc. 138,183-198.

Rosenbrock, H. [1960]. An automatic method for finding the greatest or least value of a function, Comput. J. 3,175-184.

Šidlovskaya, N. [1958]. Application of the method of differentiation with respect to a parameter to the solution of nonlinear equations in Banach spaces (Russian), Leningrad Gos. Univ. Učen. Zap. Ser. Mat. Nauk 33,3-17.

Taylor, A. [1958]. Introduction to Functional Analysis, Wiley, New York.

Varga, R. S. [1962]. Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey.

Yakovlev, M. N. [1965]. On Some Methods of Solving Nonlinear Equations, (Russian) Trudy Mat. Inst. Steklov. 84,8-40. English translation by Mirijana Kocho, edited by J. H. Avila, TR 68-75, Univ. of Maryland, July, 1968.