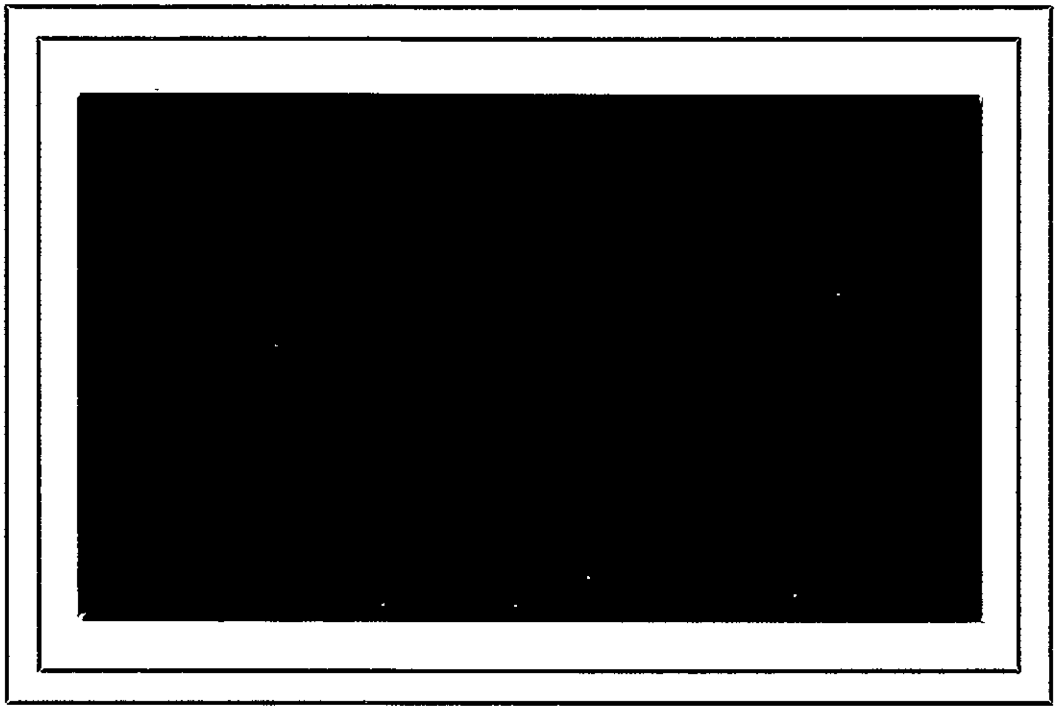

UNIVERSITY OF MARYLAND
COMPUTER SCIENCE CENTER

COLLEGE PARK, MARYLAND

TABLE OF CONTENTS

Technical Report 70-108 February 1970
NGL-21-002-008
GJ-231

DIFFERENCE EQUATIONS AND ITERATIVE PROCESSES

by

Raymond C. Cavanagh

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

INTRODUCTION

The aim of the work presented here is to unify, clarify, and extend the various techniques used in proving convergence of iterative processes for solving systems of (nonlinear) equations in $\mathbb{R}^p$. For a large class of these proofs the common approach appears to be the generation of some related, and preferably simpler, difference equation in one or several variables which has the property that convergence criteria and rate of convergence estimates for the iterates can be deduced from the behavior of its solutions.

Hence, the topics to be discussed here are (1) the principle leading to the generation of suitable one- or multidimensional difference equations for proving convergence of iterative methods, (2) the analysis of certain types of solutions of the difference equations arising from (1), and (3) the application of the results from (2) to convergence proofs for various typical classes of iterative processes. We emphasize, however, that the primary goal here is not to add new convergence results for specific iterative methods to the large number already found in the literature, but rather to present a general theory which includes these special cases and which promises to lead to new results.

When considering the ways in which the difference equation in (1) can be obtained, we note first that, technically speaking, the equations of the iterative process itself can always be considered as a p-dimensional difference equation. To prove convergence we must then analyze the iterates directly. Although this is rarely possible, when successful it usually leads to global convergence results. Critical to these results are the special form of the process equations and the use of particular structural properties of the space. This appears to preclude the development of a general theory for such an approach, and we shall not attempt to include it here.

This leaves essentially two different means of generating the difference equation in (1). The first of these begins again with the difference equation defined by the process itself, but then considers this equation locally, in the neighborhood of a specific solution (usually the trivial solution). This leads to perturbed linear difference equations and, by nature, to convergence results for the process which are always local theorems. The second of the two approaches consists in using norm estimates to majorize the process equations by some difference equation or system of equations. In this case, both local and semi-local, but rarely global, convergence results are obtained.

For the sake of illustrating all of these approaches, consider the standard Newton process

$$F'(x(n))(x(n+1)-x(n)) + Fx(n) = 0, \; n = 0,1,\ldots$$

for solving $Fx = 0$ where $F: \mathbb{R}^p \to \mathbb{R}^p$ and $F'(x)$ is the Frechet (F-)

derivative of F at x. The problem is to find conditions under which $\{x(n)\}$ can be constructed for each n and $x(n)$ converges to a solution of $Fx = 0$.

A typical treatment of the process equation itself is provided by the following global convergence result of Baluev [1952]: Let F be convex and continuously F-differentiable, and suppose that $(F'(x))^{-1}$ exists and is nonnegative on $\mathbb{R}^p$. If $Fx = 0$ has a solution x^*, then the Newton iterates $\{x(n)\}$ exist and converge to x^* from any initial point $x(0)$.

If F is twice continuously F-differentiable in a neighborhood of a solution x^* of $Fx = 0$ and if $F'(x)$ is nonsingular in a deleted neighborhood of x^*, then the process equation considered locally takes the form of a perturbed linear difference equation

$$(x(n+1)-x^*) = A(x(n)-x^*) + O(\|x(n)-x^*\|^2),\ n = 0,1,\dots ,$$

where $A = 0$ if $(F'(x^*))^{-1}$ exists. Local convergence is then equivalent to asymptotic stability of the trivial solution of

$$y(n+1) = Ay(n) + O(\|y(n)\|^2),\ n = 0,1,\dots .$$

Finally, a typical majorization approach (see, e.g., Ortega [1968]) for the same process involves bounding $\|x(n+1)-x(n)\|$ by real numbers a_{n+1}, $n = 0,1,\dots$. If F' exists and is Lipschitzian with constant γ, we obtain from Banach's lemma (with suitable $x(0)$) the difference equation

$$a_{n+1} = \frac{\frac{1}{2}\gamma\eta a_n^2}{1-\gamma\eta(\sum_{j=1}^{n} a_j)}, \quad n = 1,2,\ldots ,$$

where $\|F'(x(0))^{-1}\| \le \eta$. Semi-local convergence is now guaranteed when $a_1 = \|x(1)-x(0)\|$, η, and γ are such that $\{a_n\}$ is summable; for then $\{x(n)\}$ is a Cauchy sequence.

We begin in the first part with the localized process equation approach, which we call the direct approach. In Chapter 1, a general stability theory for the perturbed linear difference equation is developed, extending results of Perron [1929] and Ta Li [1934]. Using sum equations and a fixed point theorem on sequence spaces, we can give a new proof of the work of Coffman [1964] and extend it to the variable coefficient problem. These results are applied to iterative processes in Chapter 2 to prove point-of-attraction theorems in the sense of Ostrowski [1966] and Ortega and Rockoff [1966], including a new "singular" Newton theorem.

In Part II, the majorization approach is discussed. Chapter 3 consists of a collection of convergence theorems for isotone difference equations, the type which naturally occurs when norm estimates are employed for analyzing the process equation. In Chapter 4, these results are applied first to the case of majorizing sequences, as used by Kantorovich [1949] and Rheinboldt [1968]. Then a more general concept of majorizing systems of difference equations is introduced. This approach appears to contain all known convergence proofs for iterative processes which rely on norm estimates.

Notation, Definitions, and Background Theorems

Throughout this paper X, Y will denote p-dimensional linear spaces over the real or complex field ($\mathbb{R}$ or $\mathbb{C}$), with "$\|\cdot\|$" meaning a vector norm or its induced matrix norm, usually the ℓ_1-, ℓ_2-, or ℓ_∞-norm. Let $L(X,Y)$ denote the space of linear operators mapping X into Y, with $L(X) \equiv L(X,X)$. For $A \in L(X)$, $\rho(A)$, is the spectral radius of A (largest eigenvalue in modulus).

Since there seems to be no standard usage of the term "difference equation", we define it in the following general way.

Definition 1. Let S be a set of vector-valued sequences $\{x(n)\}$, $x(n) \in R^p$, $n = 0,1,\ldots$, containing the null sequence and Φ be a function mapping S into S. Then the relation

$$\Phi(\{x(n)\}) = \{0\} \tag{1}$$

defines a difference equation whose solution $\{x(n)\}$ is any member of S which satisfies (1).

Our interest will be in equations of a more special form than (1).

Definition 2. An m-th order difference equation in normal form is given by

$$x(k+1) = g(k,x(k),\ldots,x(k-m+1)),\ k = 0,1,\ldots ,$$

for an appropriately defined given function g.

Historically, the difference equation was defined in analogy with the differential equation, with the differences $\Delta_h^q(x(n))$ of a function x corresponding to the derivatives $x'(n)$. Here $\Delta_h^1(x(n)) = x(n+h)-x(n)$, $\Delta_h^2(x(n)) = \Delta_h^1(\Delta_h^1(x(n)))$, etc., and the equation took the form

$$F(k,\Delta_h^q x(k),\Delta_h^{q-1} x(k),\ldots,\Delta_h^1 x(k),x(k)) = 0,\ k = 0,1,\ldots .$$

We will avoid the term "recurrence equation", sometimes used as a synonym for "difference equation", for reasons of its connotations in the field of logic.

For the meaning of "iterative process" we have the following from Ortega and Rheinboldt [1970].

Definition 3. A family of operators $\{G_k\}$ with $G_k:D_k \subseteq X^{k+m} \to X$, $k = 0,1,\ldots$, defines an iterative process with m initial points and domain $D^* \subseteq D_0$ if for any $(x(0),\ldots,x(-m+1)) \in D^*$, the sequence $\{x(k)\}$ generated from

(2) $$x(k+1) = G_k(x(k),x(k-1),\ldots,x(-m+1)),\ k = 0,1,\ldots$$

exists. If $\lim_{k\to\infty} x(k) = x^*$, then x^* is a limit of the process. The process is ℓ-step if $m = \ell$ and (2) takes the form

(3) $$x(k+1) = G_k(x(k),\ldots,x(k-\ell+1)),\ k = 0,1,\ldots .$$

It is stationary if $G_k \equiv G$ for all k.

It is convenient at this time to define attraction in the sense of Ostrowski [1966].

<u>Definition 4.</u> For the iterative process $\{G_k\}$ with m initial points and domain D^* suppose there is a point $x^* \in X$ and a neighborhood U of $(x^*,\ldots,x^*)$ such that whenever $(x(0),\ldots,x(-m+1)) \in U$, the sequence $\{x(k)\}$ generated by (2) exists and converges to x^*. Then x^* is called a <u>point of attraction</u> for the process and U a domain of convergence.

Finally, we will be interested in the asymptotic speed at which sequences $\{x(n)\}$ converge to limits $x^* \in X$ and again take the following from Ortega and Rheinboldt [1970].

<u>Definition 5.</u> Let $\{x(n)\}$, $x(n) \in X$, be a sequence with limit x^* and for a given norm define the <u>quotient</u> and <u>root convergence factors</u> for $p \in [1,\infty)$

$$Q_p\{x(k)\} = \begin{cases} 0 & \text{if } x(k)=x^* \text{ for } k \geq k_0, \\ \limsup\limits_{k\to\infty} \dfrac{\|x(k+1)-x^*\|}{\|x(k)-x^*\|^p}, & \text{if } x(k)\neq x^* \text{ for all but finitely many } k, \\ \infty & \text{otherwise,} \end{cases}$$

$$R_p(\{x(k)\}) = \begin{cases} \limsup\limits_{k\to\infty} \|x(k)-x^*\|^{1/k} & \text{if } p = 1, \\ \limsup\limits_{k\to\infty} \|x(k)-x^*\|^{1/p^k}, & \text{if } p > 1. \end{cases}$$

If $C(\mathcal{J},x^*)$ denotes the set of all sequences generated by an iterative process $\mathcal{J}$ (or which solves a difference equation $\mathcal{J}$) and which converge to x^*, we can define, for $p \in [1,\infty)$,

$$Q_p(\mathcal{J},x^*) = \sup \{Q_p\{x(k)\} \mid \{x(k)\} \in C(\mathcal{J},x^*)\}$$

$$R_p(\mathcal{J},x^*) = \sup \{R_p\{x(k)\} \mid \{x(k)\} \in C(\mathcal{J},x^*)\}$$

The <u>Q- and R-order of $\mathcal{J}$ at x^*</u> are then given by

$$O_Q(\mathcal{J},x^*) = \begin{cases} \infty & \text{, if } Q_p(\mathcal{J},x^*)=0 \text{ for all } p, \\ \inf \{p \in [1,\infty) \mid Q_p(\mathcal{J},x^*) = \infty\}, & \text{otherwise,} \end{cases}$$

$$O_R(\mathcal{J},x^*) = \begin{cases} \infty & \text{, if } R_p(\mathcal{J},x^*)=0 \text{ for all } p, \\ \inf \{p \in [1,\infty) \mid R_p(\mathcal{J},x^*) = 1\}, & \text{otherwise.} \end{cases}$$

We say that $\mathcal{J}$ has convergence which is <u>R-superlinear</u>, <u>linear</u>, or <u>sublinear</u> as $R_1(\mathcal{J},x^*)$ is 0, between 0 and 1, or 1, respectively. Similar concepts relate to the Q_1 factors.

<u>Theorem 6</u>. $Q_p\{x(k)\}$ is norm dependent; $R_p\{x(k)\}$ is not. The Q- and R-orders of $\mathcal{J}$ at x^* are norm independent. Q_p is an isotone function of p which has values $0, \infty$, except possibly at one point p; R_p is an isotone function of p with values 0,1, except possibly at one point.

<u>Definition 7</u>. Let $\mathcal{J}_1$, $\mathcal{J}_2$ be different processes. Then $\mathcal{J}_1$ is <u>Q-faster</u> (<u>R-faster</u>) than $\mathcal{J}_2$ at x^* if there is some $p \in [1,\infty)$ such that

$$Q_p(\mathcal{J}_1,x^*) < Q_p(\mathcal{J}_2,x^*) \quad (R_p(\mathcal{J}_1,x^*) < R_p(\mathcal{J}_2,x^*))$$

Theorem 8. For two processes $\mathcal{J}_1$, $\mathcal{J}_2$ with limit x^*, if

$$O_Q(\mathcal{J}_1,x^*) > O_Q(\mathcal{J}_2,x^*) \quad (O_R(\mathcal{J}_1,x^*) > O_R(\mathcal{J}_2,x^*)) \; ,$$

then $\mathcal{J}_1$ is Q-faster (R-faster) than $\mathcal{J}_2$ at x^*. If $Q_p(\mathcal{J},x^*) < \infty$ for some p, then $O_Q(\mathcal{J},x^*) \geq p$. If $Q_q(\mathcal{J},x^*) > 0$ for some q, then $O_Q(\mathcal{J},x^*) \leq q$. Likewise for R, O_R with 1 replacing ∞.

To compare processes $\mathcal{J}_1$, $\mathcal{J}_2$ at x^*, we compare $O_Q(\mathcal{J}_1,x^*)$ with $O_Q(\mathcal{J}_2,x^*)$. If these are both equal to p, compare $Q_p(\mathcal{J}_1,x^*)$ with $Q_p(\mathcal{J}_2,x^*)$. Similarly for R-convergence.

Theorem 9. If $Q_p\{x(k)\} < \infty$ for some p, then for any $\varepsilon > 0$ there is a k_0 such that

$$\|x(k+1)-x^*\| \leq (Q_p+\varepsilon)\|x(k)-x^*\|^p \text{ for all } k \geq k_0$$

If $R_p\{x(k)\} < 1$, then for any $\varepsilon > 0$ there is a k_0 such that either

$$\|x(k)-x^*\| \leq (R_p+\varepsilon)^{p^k}, \text{ for } k \geq k_0, \; p > 1 \; ,$$

or

$$\|x(k)-x^*\| \leq (R_1+\varepsilon)^k, \text{ for } k \geq k_0, \; p = 1$$

Theorem 10. If $\lim\limits_{k\to\infty} x(k) = x^*$, then $R_1\{x(k)\} \leq Q_1\{x(k)\}$ in every norm. Hence $R_1(\mathcal{J},x^*) \leq Q_1(\mathcal{J},x^*)$ and $O_Q(\mathcal{J},x^*) \leq O_R(\mathcal{J},x^*)$

Theorem 11. For a process $\mathcal{J}$ and some $p \in (1,\infty)$ suppose tha each sequence of iterates $\{x(n)\}$ which converges to x^* satisfies $R_p\{x(n)\} < 1$. Then $O_R(\mathcal{J},x^*) \geq p$.

PART I

DIRECT APPROACH

As mentioned above, here we will generate difference equations by considering iterative methods locally. Specifically, suppose we have the single-step process

$$\text{(I-1)} \qquad x(n+1) = G(n,x(n)),\ n = 0,1,\ldots ,$$

for solving $\lim_{n\to\infty} G(n,x) = x$, where $G(n,\cdot):E \subseteq X \to Y$ for each n. Suppose that there is a point $x^* \in E$ such that $G(n,x^*) = x^*$ for all n and that the F-derivative $G'(n,x^*)$ exists. If E is open and convex, then for any $x(n) \in E$ we have

$$\begin{aligned} \text{(I-2)} \qquad x(n+1) - x^* &= G'(n,x^*)(x(n)-x^*) \\ &\quad + [G(n,x(n))-G(n,x^*)-G'(n,x^*)(x(n)-x^*)] \\ &\equiv G'(n,x^*)(x(n)-x^*) + \Phi(n,x(n)) , \end{aligned}$$

and

$$\lim_{x\to x^*} \frac{\|\Phi(n,x)\|}{\|x-x^*\|} = 0 \text{ for each } n .$$

This system is of the general form

$$\text{(I-3)} \qquad y(n+1) = A(n)y(n) + F(n,y(n)),\ n = 0,1,\ldots ,$$

with

$$F(n,y) = o(\|y\|) \text{ as } y \to 0 .$$

Chapter 1 is devoted to the study of the stability properties of the perturbed linear systems of difference equations (I-3). In Chapter 2, we return to iterative processes of the form (I-1), as well as of slightly more general forms, and apply the results of Chapter 1 to obtain point-of-attraction theorems by means of the error equation (I-2).

Chapter 1

Perturbed Linear Difference Equations

The purpose of this chapter is to present a general stability theory for the perturbed linear system of difference equations

$$(1.1) \qquad y(n+1) = A(n)y(n) + F(n,y(n)),\ n = 0,1,\ldots$$

where $y(n) \in \mathbb{C}^p$, $A(n) \in L(\mathbb{C}^p)$ for $n = 0,1,\ldots,$ and $F(n,\cdot)$ maps a domain in $\mathbb{C}^p$ into $\mathbb{C}^p$ for each n. In the form presented here, this theory appears to be new, although for certain special cases the results are essentially known.

From time to time we will refer to several types of stability for the trivial solution $\{0\}$ of a difference equation

$$(1.2) \qquad y(n+1) = H(n,y(n)),\ n = 0,1,\ldots;\ H(n,0) = 0 .$$

$\{0\}$ is <u>stable</u> if for any sufficiently small $\varepsilon > 0$ there is a $\delta \in (0,\varepsilon)$ such that $||y(n)|| \leq \varepsilon$, $n = 0,1,\ldots,$ for any solution $\{y(n)\}$ of (1.2) with $||y(0)|| < \delta$. It is <u>asymptotically stable</u> if whenever $||y(0)|| > \delta$ is sufficiently small, then the solution $\{y(n)\}$ of (1.2) has $\lim_{n\to\infty} y(n) = 0$. $\{0\}$ is <u>exponentially stable</u> if there are positive constants α,β and some $\varepsilon > 0$ such that for any $n_0 \geq 0$, $||y(n_0)|| < \varepsilon$ implies $||y(n)|| \leq \alpha||y(n_0)||e^{-\beta(n-n_0)}$. Finally, we call $\{0\}$ <u>conditionally stable</u> (or <u>conditionally asymptotically stable</u>) if for any sufficiently small $\varepsilon > 0$, there exists $y(0)$

with $0 < \|y(0)\| < \varepsilon$ such that the solution $\{y(n)\}$ of (1.2) satisfies $\|y(n)\| \leq \varepsilon$, $n = 0,1,\ldots$ (or $\lim_{n\to\infty} y(n) = 0$).

The chapter consists of three sections which we describe briefly below.

In Section 1 we consider the following "canonical" form of (1.1),

$$y(n+1) = D(n)y(n) + G(n,y(n)),\ n = 0,1,\ldots \tag{1.3}$$

Here $D(n)$ is a block diagonal $(p\times p)$-matrix with square blocks and $G(n,\cdot)$ maps a domain in $\mathbb{C}^p$ into $\mathbb{C}^p$. A general conditional asymptotic stability theory for such equations is developed under the assumption that the perturbation G either has a special form or is in some sense "small".

Section 2 is concerned with the special case of (1.1) for which $A(n)$ is a constant matrix. After transforming (1.1) into canonical form (1.3) with $D(n)$ constant, we apply the results of Section 1 to obtain a new proof of the results of Coffman [1964]. Finally, in Section 3, we turn to the general case of (1.1) in which $A(n)$ is actually variable. Again a change of variables reduces (1.1) to canonical form (1.3) and an application of the theory of Section 1 yields stability results, including an extension of the work of Ta Li [1934].

Section 1. Stability for the Canonical Equations

For equation (1.3) suppose that $D(n)$ has r blocks $D_j(n)$ of dimensions p_j, $j = 1,\ldots,r$, respectively. Since $\mathbb{C}^{p_1}\times\ldots\times\mathbb{C}^{p_r} = \mathbb{C}^p$, we can define the natural projections

$$P_j\colon \mathbb{C}^p \to \mathbb{C}^{p_j} \text{ by } P_j((x_1,x_2,\ldots,x_p)^T) \equiv (x_{p_1+\ldots+p_{j-1}+1},\ldots,x_{p_1+\ldots+p_j})^T$$

Then let $G_j(n,\cdot) \equiv P_jG(n,\cdot)$ and notice that $P_jD(n) = D_j(n)P_j$.

The standard product notation for matrices will have the order interpretation

$$\prod_{j=k}^{k+\ell} B(j) \equiv B(k+\ell)B(k+\ell-1)\ldots B(k+1)B(k)$$

with the usual conventions $\sum_{j=k}^{k-1} \cdot \equiv 0$, $\prod_{j=k}^{k-1} \cdot \equiv 1$.

For finding conditions on $D(n)$ and G which guarantee certain conditional stability properties of the trivial solution of (1.3) it is advantageous to use a sum equation which is equivalent to (1.3). Such forms were first considered by Perron [1929] and are analogous to the integral equation forms for ordinary differential equations.

Lemma 1.1. A sequence $\{z(n)\}$ with $z(n) \in \mathbb{C}^p$, $n = 0,1,\ldots,$ is a solution of (1.3) if and only if it satisfies

$$(1.4)\qquad P_iz(n+1) = \Big(\prod_{k=0}^{n} D_i(k)\Big)P_iz(0) + \sum_{j=0}^{n}\Big(\prod_{k=j+1}^{n} D_i(k)\Big)G_i(j,z(j)),$$

$$i = 1,2,\ldots,r;\ n = 0,1,\ldots$$

Moreover, if for some index i_0 $(\prod_{k=0}^{n} D_{i_0}(k))^{-1}$ exists for each n and is uniformly bounded by $\frac{1}{\alpha} > 0$, then any solution $\{z(n)\}$ of (1.3) with $\lim_{n\to\infty} z(n) = 0$ satisfies

$$(1.5) \qquad P_{i_0} z(n) = -\sum_{j=n}^{\infty} (\prod_{k=n}^{j} D_{i_0}(k))^{-1} G_{i_0}(j,z(j)), \; n = 0,1,\ldots \; .$$

On the other hand, if $\{z(n)\}$ is a sequence with $z(n) \in \mathbb{C}^p$ and satisfies for each i either (1.4) or (1.5), then $\{z(n)\}$ solves (1.3).

<u>Proof</u>: The first and last parts follow from a simple induction argument. For the second statement, rewrite (1.4) with $i = i_0$ as

$$P_{i_0} z(n+1) = (\prod_{k=0}^{n} D_{i_0}(k))[P_{i_0} z(0) + \sum_{j=0}^{n} (\prod_{k=0}^{j} D_{i_0}(k))^{-1} G_{i_0}(j,z(j))] \; .$$

But then

$$\|P_{i_0} z(n+1)\| \geq \alpha \|P_{i_0} z(0) + \sum_{j=0}^{n} (\prod_{k=0}^{j} D_{i_0}(k))^{-1} G_{i_0}(j,z(j))\|$$

and since $\lim_{n\to\infty} z(n) = 0$, we have

$$P_{i_0} z(0) = \lim_{n\to\infty} [-\sum_{j=0}^{n} (\prod_{k=0}^{j} D_{i_0}(k))^{-1} G_{i_0}(j,z(j))]$$

Hence

$$P_{i_0} z(n+1) = -(\prod_{k=0}^{n} D_{i_0}(k))[\sum_{j=n+1}^{\infty} (\prod_{k=0}^{j} D_{i_0}(k))^{-1} G_{i_0}(j,z(j))]$$

which is (1.5).

Equations (1.4) and (1.5) will be used next to define a certain function T which operates on sequence spaces. Let S denote the set of all sequences with range $\mathbb{C}^p$ and c_0 its subset consisting of those sequences which converge to 0. Clearly, under the ordinary operations of addition and scalar multiplication and with the induced norm $\|\{x(n)\}\| \equiv \sup\{\|x(n)\| \,|\, n=0,1,\ldots\}$ for any $\{x(n)\} \in C_0$ and any norm $\|\cdot\|$ on $\mathbb{C}^p$, C_0 is a Banach space (see Dunford and Schwartz [1958]).

We call the difference equation (1.3) admissible if there is an index $s \in [1,r]$ and numbers $\alpha \geq 1$, $0 < \lambda < 1$, $\beta \geq 0$ such that

$$(1.6) \qquad \Big\| \prod_{k=\ell}^{\ell+m} D_i(k) \Big\| \leq \alpha\lambda^{m+1} \text{ for all } \ell,\ m \geq 0,\ i = s,s+1,\ldots,r,$$

$$(1.7) \qquad \sum_{j=0}^{\infty} \Big\| \Big(\prod_{k=0}^{j} D_i(k)\Big)^{-1} \Big\| \leq \beta \text{ for } i = 1,2,\ldots,s-1$$

Notice that (1.7) is stronger than the assumption in the second part of Lemma 1.1. The reasons for these conditions will be discussed at the end of this chapter. For any admissible equation we can define the following operator T which maps a subset of S into S:

$$(1.8a) \qquad T(\{x(K)\})(0) \equiv x(0),$$

$$(1.8b) \qquad T(\{x(K)\})(n+1) \equiv \begin{pmatrix} -\displaystyle\sum_{j=n+1}^{\infty} \Big(\prod_{k=n+1}^{j} D_1(k)\Big)^{-1} G_1(j,x(j)) \\ \cdots \\ -\displaystyle\sum_{j=n+1}^{\infty} \Big(\prod_{k=n+1}^{j} D_{s-1}(k)\Big)^{-1} G_{s-1}(j,x(j)) \\ \Big(\displaystyle\prod_{k=0}^{n} D_s(k)\Big) P_s x(0) + \displaystyle\sum_{j=0}^{n} \Big(\prod_{k=j+1}^{n} D_s(k)\Big) G_s(j,x(j)) \\ \cdots \qquad\qquad \cdots \\ \displaystyle\prod_{k=0}^{n} D_r(k) P_r x(0) + \displaystyle\sum_{j=0}^{n} \Big(\prod_{k=j+1}^{n} D_r(k)\Big) G_r(j,x(j)) \end{pmatrix}$$

for $n = 0,1,\ldots$. By Lemma 1.1, if $\{y(n)\} \in S$ is in the domain of T and satisfies $T(\{y(n)\}) = \{y(n)\}$, then $\{y(n)\}$ solves (1.3). Hence, the approach to be taken here is to find fixed points of T in C_0. This is accomplished with the aid of Schauder's Theorem [1930], in the form: If T is a continuous operator which maps a compact, convex subset of a Banach space into itself, then T has a fixed point in that subset.

Again from Dunford and Schwartz [1958], a set $M \subseteq C_0$ is conditionally compact if and only if M is bounded and $\lim_{n\to\infty} x(n) = 0$ uniformly for all $\{x(n)\} \in M$. Thus, we consider the general compact, convex subsets of C_0 defined in the following way. Let $\{a_n\}$ be a nonnegative sequence with $\lim_{n\to\infty} a_n = 0$ and let $z(0) \in \mathbb{C}^p$. Then

$$(1.9) \qquad M = \{\{x(n)\} \in C_0 \mid P_j(x(0)-z(0)) = 0 \text{ for } j = s,s+1,\ldots,r, \text{ and } \|x(n)\| \leq a_n,\ n = 0,1,\ldots\}$$

is evidently compact and convex in C_0.

Sufficient conditions for the continuity of T are given next.

Lemma 1.2. For an admissible equation (1.3) suppose that the domain of the corresponding operator T of (1.8) contains a set $M \subseteq C_0$ of the form (1.9) with $T(M) \subseteq M$. If for each $n = 0,1,\ldots,$ $G(n,\cdot)$ is continuous on $\bar{S}(0,a_n)$ and $\|G(n,x)\| \leq \theta$ for $x \in \bar{S}(0,a_n)$, then T is continuous on M.

Proof: Let $\varepsilon > 0$ and $\{x(n)\} \in M$ with $\{v(n)\} \equiv T(\{x(n)\})$. For any $\{z(n)\} \in M$, we have $T(\{z(n)\}) \equiv \{w(n)\} \in M$ and $\|w(n)\| \leq a_n \leq \frac{\varepsilon}{2}$ for $n \geq n_0(\varepsilon) = n_0$. Thus $\|v(n)-w(n)\| \leq \varepsilon$ for $n \geq n_0$ and any $\{z(n)\} \in M$. Now, since $G(n,\cdot)$ is continuous, we can choose a $\delta_1 > 0$ such that whenever $\|\{z(n)\}-\{x(n)\}\| \leq \delta_1$, then

$$\|P_i(v(n)-z(n))\| \leq \sum_{j=0}^{n} \|\prod_{k=j+1}^{n} D_i(k)\| [\|G_i(j,x(j))-G_i(j,z(j))\|] \leq \varepsilon$$

for $i = s,\ldots,r$; $n = 0,1,\ldots,n_0$. Let $\eta,\theta,n_1 \geq n_0$ be such that $\|\prod_{k=0}^{j} D_i(k)\| \leq \eta$ for $j = 0,1,\ldots,n_0$; $i = 1,2,\ldots,s-1$ and

$$\sum_{j=n}^{\infty} \|(\prod_{k=0}^{j} D_i(k))^{-1}\| \leq \frac{\varepsilon}{4\eta\theta} \text{ for } n \geq n_1,\ i = 1,\ldots,s-1 .$$

Finally, choose $0 < \delta_2 \leq \delta_1$ such that whenever $\|\{x(n)\}-\{z(n)\}\| \leq \delta_2$ and $\{z(n)\} \in M$, then

$$\eta \sum_{j=1}^{n_1} \|(\prod_{k=0}^{j} D_i(k))^{-1}\| \|G(j,x(j))-G(j,z(j))\| < \frac{\varepsilon}{2} \text{ for } i = 1,2,\ldots,s-1$$

In that case, for $i = 1,\ldots,s-1$, $n = 0,1,\ldots,n_0$,

$$\|P_i(v(n)-w(n))\| \leq \|\sum_{j=n+1}^{\infty} (\prod_{k=n+1}^{j} D_i(k))^{-1}(G_i(j,z(j))-G_i(j,x(j)))\|$$

$$\leq \|\prod_{k=0}^{n} D_i(k)\| (\sum_{j=n+1}^{\infty} \|(\prod_{k=0}^{j} D_i(k))^{-1}\| \cdot \|G(j,z(j))-G(j,x(j))\|)$$

$$\leq \eta (\sum_{j=n+1}^{n_1} \|(\prod_{k=0}^{j} D_i(k))^{-1}\| \|G(j,z(j))- G(j,x(j))\|) + 2\eta\theta(\frac{\varepsilon}{4\eta\theta})$$

$$\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This completes the proof.

It remains then to find a set M of the form (1.9) with $T(M) \subseteq M \subseteq C_0$. Using norm estimates on the components of (1.8), we see that the existence of a sequence $\{a_n\}$ which satisfies the following inequalities is sufficient:

$$\sum_{j=n+1}^{\infty} \left\| \left(\prod_{k=n+1}^{j} D_i(k) \right)^{-1} \right\| \, \| G_i(j,x(j)) \| \leq a_{n+1}, \; i=1,\dots,s-1, \; n=0,1,\dots,$$

$$\left\| \prod_{k=0}^{n} D_i(k) \right\| \, \| x(0) \| + \sum_{j=0}^{n} \left\| \prod_{k=j+1}^{n} D_i(k) \right\| \, \| G_i(j,x(j)) \| \leq a_{n+1} \; ,$$

for $i=s,s+1,\dots,r$, $n=0,1,\dots,$ whenever $\{x(j)\} \in M$. If we assume that the equation (1.3) is admissible and that for some real sequence $\{a_n\}$ with $\lim_{n\to\infty} a_n = 0$

$$\sup_{\|x\| \leq a_j} \frac{\|G(j,x)\|}{\|x\|} \leq c + c_j \text{ for } j=0,1,\dots,$$

where $c \geq 0$, $c_j \geq 0$, then $\{a_n\}$ determines a set M with $T(M) \subseteq M$ provided that $\{a_n\}$ satisfies

(1.10a) $$\sum_{j=n+1}^{\infty} \left\| \left(\prod_{k=n+1}^{j} D_i(k) \right)^{-1} \right\| (c+c_j) a_j \leq a_{n+1} \text{ for } i=1,\dots,s-1; \; n=0,1,\dots$$

(1.10b) $$\alpha\lambda^{n+1} a_0 + \sum_{j=0}^{n} \alpha\lambda^{n-j} (c+c_j) a_j \leq a_{n+1} \text{ for } i=s,\dots,r; \; n=0,1,\dots$$

For solving (1.10) we have the following lemma.

Lemma 1.3. For the system (1.10) suppose $\{D_i(n)\}$, $i=1,\dots,s-1$, satisfy (1.7) and that $c_j \leq c^*$ for $j=0,1,\dots$. If

$$(1.11) \qquad (\alpha c+\alpha c^*+\lambda) < 1 \text{ and } \beta(c+c^*) < 1 ,$$

then $a_n = b_n\alpha\lambda^n$ solves (1.10) where $\{b_n\}$ is defined by

$$(1.12) \qquad b_1 = (1+\frac{c+c_0}{\lambda})b_0\alpha,\ b_{n+1} = (1+(\frac{\alpha}{\lambda})(c+c_n))b_n,\ n=1,2,\dots \ .$$

Furthermore, $\lim\limits_{n\to\infty} a_n = 0$ and $R_1\{a_n\} \leq (\lambda+\alpha c+\alpha c^*)$. If $\lim\limits_{n\to\infty} c_n = 0$, then $R_1\{a_n\} \leq \lambda+ c\alpha$.

Proof: An induction argument is used with the inequalities

$$\alpha\lambda a_0+\alpha(c+c_0)a_0 = \alpha\lambda\alpha b_0+\alpha(c+c_0)\alpha b_0 = b_0\alpha(1+\frac{c+c_0}{\lambda})\alpha\lambda = b_1\alpha\lambda = a_1 ,$$

$$\alpha\lambda^{n+1}a_0+\sum_{j=0}^{n}\alpha\lambda^{n-j}(c+c_j)a_j = \alpha\lambda^{n+1}[b_0\alpha(1+\frac{c+c_0}{\lambda})+\sum_{j=1}^{n}(\frac{\alpha}{\lambda})(c+c_j)b_j]$$

$$= \alpha\lambda^{n+1}[b_1+\sum_{j=1}^{n}(\frac{\alpha}{\lambda})(c+c_j)b_j] = b_{n+1}\alpha\lambda^{n+1} = a_{n+1} ,$$

and for $i=1,2,\dots,s-1$,

$$\sum_{j=n+1}^{\infty}\left\|\left(\prod_{k=n+1}^{j} D_i(k)\right)^{-1}\right\|(c+c_j)b_j\alpha\lambda^j$$

$$\leq b_{n+1}\alpha\lambda^{n+1}\left[\sum_{j=n+1}^{\infty}\left\|\left(\prod_{k=0}^{j} D_i(k)\right)^{-1}\right\|(c+c_j)\frac{b_j}{b_{n+1}}\frac{\lambda^j}{\lambda^{n+1}}\right]$$

$$\leq b_{n+1}\alpha\lambda^{n+1}[(\beta)(c+c^*)] \leq a_{n+1} .$$

Here we use

$$\frac{b_j}{b_{n+1}} \frac{\lambda^j}{\lambda^{n+1}} = \begin{cases} \prod_{k=n+1}^{j-1} (\alpha c+\alpha c_k+\lambda) \leq 1 & \text{if } j > n+1 \\ & \\ & \text{if } j = n+1 \end{cases}$$

For the last part, notice that

$$b_{n+1} = [\prod_{k=1}^{n} (1+ \frac{\alpha}{\lambda} (c+c_k))]b_1 \leq (1+ \frac{\alpha}{\lambda} (c+c^*))^n b_1$$

and so

$$\limsup_{n\to\infty} a_n^{1/n} \leq \limsup_{n\to\infty} [(1+(\frac{\alpha}{\lambda})(c+c^*))^{n-1} b_1 \alpha\lambda^n]^{1/n}$$

$$= (\lambda+\alpha c+\alpha c^*)$$

We are now prepared to give the main result of this section. Here $\|\cdot\|$ will refer to the maximum norm on $\mathbb{C}^p$.

<u>Theorem 1.4</u>. For an admissible difference equation of form (1.3) suppose that

(1.13a) (1.6) holds for at least one index, i.e., $s-1 \neq r$,

(1.13b) $G(n,x)$ is continuous in x for $\|x\| \leq \eta$ and $n=0,1,\ldots,$

(1.13c) $\begin{cases} \text{there is a constant } c \geq 0 \text{ and a nonnegative real-valued} \\ \text{function } c_n(x)\text{, defined for } \|x\| \leq \eta \text{ and } n=0,1,\ldots, \text{ with} \\ \lim\limits_{\substack{n\to\infty \\ x\to 0}} c_n(x) = 0,\ c_n(x) \leq c^*, \text{ and} \\ \|G(n,x)\| \leq (c+c_n(x))\|x\| \text{ for } \|x\| \leq \eta \end{cases}$

If b_0, c, and c^* are such that (1.11) holds and $\alpha b_0 \leq \eta$, then there is for each $y(0) \in \mathbb{C}^p$ with $\max_{s\leq i\leq r} \|P_i y(0)\| \leq \alpha b_0$ a solution $\{z(n)\} \in C_0$ of (1.3) with the properties:

(1.14a) $\quad P_i(z(0)-y(0)) = 0,\ i=s,s+1,\ldots,r,$

(1.14b) $\quad R_1\{z(n)\} \leq (\lambda+\alpha c)\ ,$

(1.14c) $\quad z(n)$ satisfies (1.4) for $i=s,\ldots,r$ and (1.5) for $i=1,2,\ldots,s-1$.

Proof: Let M be the set

$$\{\{z(n)\} \in C_0 \mid \|z(0)\| \leq \alpha b_0,\ P_i(z(0)-y(0)) = 0,\ i=s,s+1,\ldots,r,$$

$$\|z(n)\| \leq b_n \alpha\lambda^n,\ n \geq 0\}\ .$$

Here $\{b_n\}$ and $\{c_n\}$ are defined so that $\lim_{n\to\infty} c_n = 0$ and $\{b_n\}$ satisfies (1.12) with $\|G(n,x)\| \leq (c+c_n)\|x\|$ for $\|x\| \leq b_n\alpha\lambda^n$.

From the discussion above, M is compact and convex, and the operator T which corresponds to (1.3) maps M into M continuously. Hence T has a fixed point $\{z(n)\} \in M$ which by Lemma 1.1 solves (1.3). The estimate (1.14b) follows from Lemma 1.3 and the proof is complete.

Condition (1.6) was chosen so that it holds if $D_i(k)$ is constant with spectral radius less than one or if $\|D_i(k)\| \leq d_i(k)$ and $\limsup_{k\to\infty} d_i(k) < 1$. As for (1.7), notice that if $\alpha = 1$, then $a_n = b_n\lambda^n$ solves (1.10b) with equality. In that case conditions (1.7) and (1.11) guarantee that $\{a_n\}$ solves (1.10a) and is thus a minorant for the solutions of (1.10).

Although the sum equation form of a difference equation has been used often, the T-operator (1.8) formulation appears to be new. For the special case in which $D(n)$ is constant, Coffman [1964] has obtained the conclusion of Theorem 1.4 without using the T-operator. His proof uses the Brouwer theorem, and he points out the necessity of applying general fixed point results. Prior to his work, other authors including Perron [1929], Ta Li [1934], and Bellman [1947] have proved conditional stability results for special forms of $D(n)$ under more stringent conditions on the perturbation G. Their approach employs the method of successive approximations. Of interest in this regard is the paper of Panov [1959] in which a successive approximation argument allegedly proves a theorem equivalent to Theorem 1.4 for the case in which D is constant. The proof is clearly incorrect. A similar but correct proof for a weaker theorem was given by Bellman [1947]. Finally, an approach analogous to ours for differential equations has been used in special cases by several authors (see Coppel [1965], pp. 80 and 100).

We next prove a general result for the case in which G is more restricted. It contains all of the above cited theorems which can be proved by successive approximations. In that event, our operator T is contractive. The norm used here is the maximum norm on $\mathbb{C}^p$.

<u>Lemma 1.5</u>. For a constant $\eta > 0$ and $z(0) \in \mathbb{C}^P$, define the set $\widetilde{M} \subseteq S$ by

(1.15) $\quad \widetilde{M} = \{\{x(n)\} \in S \,|\, \|x(n)\| \leq \eta \text{ and } P_i(x(0)-z(0)) = 0 \text{ for } i=s,s+1,\dots,r\}.$

Suppose that (1.3) is admissible and that $T(\widetilde{M}) \subseteq \widetilde{M}$. If G satisfies

(1.16) $\quad \|G(n,x)-G(n,y)\| \leq \varepsilon\|x-y\| \text{ for } \|x\|,\|y\| \leq \eta \text{ and } n=0,1,\dots$

where $\gamma = \max\left(\beta\varepsilon, \frac{\alpha\varepsilon}{1-\lambda}\right) < 1$ and $\varepsilon > 0$, then T is contractive on $\widetilde{M}$ with constant γ, that is,

$$\|T\{x(n)\}-T\{y(n)\}\| \leq \gamma \|\{x(n)\}-\{y(n)\}\|$$

whenever $\{x(n)\},\{y(n)\} \in \widetilde{M}$.

<u>Proof</u>: For $\{x(n)\},\{y(n)\} \in \widetilde{M}$, let $\{v(n)\} = T(\{x(n)\})$ and $\{w(n)\} = T(\{y(n)\})$. Then for $i=1,2,\dots,s-1$,

$$\|P_i(w(n)-v(n))\| \leq \sum_{j=n}^{\infty} \|(\prod_{k=n}^{j} D_i(k))^{-1}\| \|G_i(j,x(j))-G_i(j,y(j))\|$$

$$\leq \beta \sup_{n\leq j<\infty} \|G(j,x(j))-G(j,y(j))\|$$

$$\leq \beta \sup_{n\leq j<\infty} \varepsilon\|x(j)-y(j)\| ,$$

and for $i=s,s+1,\dots,r$,

$$\|P_i(w(n)-v(n))\| \leq \sum_{j=0}^{n-1} \alpha\lambda^{n-j-1}\|G_i(j,x(j))-G_i(j,y(j))\|$$

$$\leq \left(\frac{\alpha}{1-\lambda}\right) \left(\max_{0\leq j\leq n-1} \varepsilon \|x(j)-y(j)\|\right) .$$

Finally, $\|w(n)-v(n)\| \leq (\max(\beta\varepsilon, \frac{\alpha\varepsilon}{1-\lambda}))(\|\{x(n)\}-\{y(n)\}\|)$.

The final result can be proved once a set $\tilde{M}$ with $T(\tilde{M}) \subseteq \tilde{M}$ is found.

<u>Theorem 1.6</u>. For an admissible equation (1.3) suppose that there are constants $\varepsilon, \eta, \theta$, $\delta \leq \eta$ such that (1.16) holds and $\|G(k,x)\| \leq \theta$ for all k and $\|x\| \leq \eta$. If $\gamma = \max(\beta\varepsilon, \frac{\alpha\varepsilon}{1-\lambda}) < 1$ and $\max(\beta\theta, \alpha\delta+\theta(\frac{\alpha\lambda}{1-\lambda})) \leq \eta$, then for any initial vector $z(0) \in \mathbb{C}^p$ with $\|z(0)\| < \delta$, there is exactly one solution $\{x(n)\}$ of (1.3) satisfying $P_i(z(0)-x(0)) = 0$ for $i=s,s+1,\ldots,r$ and $\|x(n)\| \leq \eta$, $n=0,1,\ldots$.

<u>Proof</u>: We first show that for the set $\tilde{M}$ of (1.15), $T(\tilde{M}) \subseteq \tilde{M}$. With $\{x(n)\} \in \tilde{M}$ and $\{v(n)\}$ as in Lemma 1.5, we have

$$\|P_i v(n)\| \leq \sum_{j=n}^{\infty} \|(\prod_{k=n}^{j} D_i(k))^{-1}\| \, \|G(j,x(j))\| \leq \beta\theta \leq \eta, \; i=1,\ldots,s-1$$

and

$$\|P_i v(n)\| \leq \alpha\lambda^n\delta + \sum_{j=0}^{n-1} \alpha\lambda^{n-j}\theta \leq \alpha\delta + \frac{\alpha\lambda}{1-\lambda}\theta \leq \eta, \; i=s,\ldots,r \, .$$

Thus, $T(\tilde{M}) \subseteq \tilde{M}$. Since $\tilde{M}$ is complete, the standard contraction mapping theorem yields the existence of a unique fixed point of T in $\tilde{M}$. In fact, the iterative process $\{x^{\mu+1}(n)\} = T(\{x^{\mu}(n)\})$, $\mu = 0,1,\ldots$ converges to this fixed point for any $\{x^0(n)\} \in \tilde{M}$.

Section 2. Equations with Constant Matrices

In this section and the next we will prove stability properties for the general equation (1.1) by first transforming it into canonical form (1.3) and then applying the results obtained above. We begin with a special form of (1.1) in which the matrices $A(n)$ are constant,

$$x(n+1) = Ax(n) + F(n,x(n)), \; n=0,1,\ldots \; . \tag{1.17}$$

Choose $S \in L(\mathbb{C}^p)$ such that $SAS^{-1} \equiv D = (d_{ij})$ is the Jordan form of A with $d_{ii} = \lambda_i$, $i=1,\ldots,p$; $d_{i+1,i}=\theta_i=0$ or 1, $i=1,\ldots,p-1$; $d_{ij} = 0$ otherwise. Suppose that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_p|$ and that for some $s \in \{1,\ldots,r\}$, $r \leq p$, the distinct moduli ρ_i of the λ_i satisfy

$$\rho_1 > \rho_2 > \cdots > \rho_{s-1} > 1 > \rho_s > \cdots > \rho_r \; . \tag{1.18}$$

Let D_i denote the square diagonal block of dimension p_i which has as diagonal entries all λ_j of modulus ρ_i and let P_i be the corresponding projection from $\mathbb{C}^p$ to $\mathbb{C}^{p_i}$. Then with the change of variables

$$y(n) = Sx(n), \; n = 0,1,\ldots \; , \tag{1.19}$$

we obtain an equation in canonical form

$$y(n+1) = Dy(n) + SF(n,S^{-1}y(n)) \equiv Dy(n) + G(n,y(n)), \; n=0,1,\ldots \; . \tag{1.20}$$

It is clearly admissible whenever (1.18) holds.

We are now prepared to prove a theorem for (1.20) and hence, by means of (1.19), for (1.17), and we begin by stating without proof a lemma of Coffman [1964]:

Lemma 1.7. Let $\{a_n\}$, $\{b_n\}$, $\{c_n\}$ be nonnegative sequences of real numbers satisfying the system of inequalities

$$(1.21a) \qquad a_{n+1} \leq \sigma a_n + c_n(a_n+b_n),\ n = 0,1,\ldots ,$$

$$(1.21b) \qquad b_{n+1} \geq \tau b_n - c_n(a_n+b_n),\ n = 0,1,\ldots ,$$

where $0 \leq \sigma < \tau$ and $\lim_{n\to\infty} c_n = 0$. If $\varepsilon > 0$ is given, then there exists a $c^* = c^*(\varepsilon) > 0$ such that whenever $a_0 < \varepsilon b_0$ and $0 \leq c_n \leq c^*$, $n=0,1,\ldots,$ we have $b_n > 0$ and $a_n < \varepsilon b_n$ for all n.

If $a_n + b_n \neq 0$ and $\{\phi_n\}$, $\{\psi_n\}$ are positive sequences for which

$$(1.22a) \qquad c_n \leq \phi_n \text{ and } \frac{\phi_{n+1}}{\phi_n} \geq \alpha > \frac{\sigma}{\tau} \text{ for } n \geq N,\ \lim_{n\to\infty} \phi_n = 0 ,$$

$$(1.22b) \qquad c_n \leq \psi_n \text{ and } \frac{\psi_{n+1}}{\psi_n} \leq 1,\ \lim_{n\to\infty} \psi_n = 0 ,$$

then either $\frac{a_n}{b_n}$ is defined for all large n and $\frac{a_n}{b_n} = O(\phi_n)$ or else $\frac{b_n}{a_n}$ is defined for all large n and $\frac{b_n}{a_n} = O(\psi_n)$.

An application of this and Theorem 1.4 yields a new proof of the following conditional stability result for (1.20) due to Coffman [1964]. Throughout, $\|\cdot\|$ will denote the appropriate ℓ_1-norm.

Theorem 1.8. For equation (1.20) suppose that D has at least one eigenvalue λ_p with $|\lambda_p| < 1$ and that $G(n,\cdot)$ is defined and continuous on $\bar{S} = \bar{S}(0,\eta)$, $\eta > 0$, with

(1.23a) $$\|G(n,y)\| \leq c^*\|y\| \text{ for } y \in \bar{S} \text{ and all } n ,$$

(1.23b) $$\lim_{\substack{n\to\infty \\ y\to 0}} \frac{\|G(n,y)\|}{\|y\|} = 0.$$

Then, if c^* is sufficiently small, for any $\gamma \in (0,1)$ there exists a $\delta = \delta(\gamma) \in (0,1)$ such that whenever $z \in \mathbb{R}^p$ satisfies

(1.24) $$\|P_{s+1}z\| + \ldots + \|P_r z\| \leq \gamma \|P_s z\| \text{ and } 0 < \|P_s z\| < \delta ,$$

then there exists a solution $\{y(n)\}$ of (1.20) with the properties

(1.25) $$P_i(y(0)-z) = 0 \text{ for } i = s,\ldots,r ,$$

(1.26) $$\lim_{n\to\infty} y(n) = 0 \text{ and } \lim_{n\to\infty} \|y(n)\|^{1/n} = \rho_s ,$$

(1.27) $\{y(n)\}$ satisfies (1.4)/(1.5).

Furthermore, if $y(n) \neq 0$ for n large and if ϕ_n,ψ_n are positive majorants for

$$c_n = \sup\left\{\left(\frac{\|G(n,y)\|}{\|y\|}\right) \;\middle|\; 0 < \|y\| \leq \|y(n)\|\right\} ,$$

which satisfy (1.22) with $\alpha > \max\left(\frac{\rho_s}{\rho_{s-1}}, \frac{\rho_{s+1}}{\rho_s}\right)$, then

(1.28a) $$\frac{\|P_i y(n)\|}{\|P_s y(n)\|} = O(\phi_n) \text{ for } i = s+1,\ldots,r ,$$

(1.28b) $$\frac{\|P_i y(n)\|}{\|P_s y(n)\|} = O(\psi_n) \text{ for } i = 1,2,\ldots,s-1 .$$

<u>Proof</u>: Let $P(a) \equiv \mathrm{diag}\,(a, a^2, \ldots, a^p) \in L(\mathbb{R}^p)$ for $a > 0$, and transform equation (1.20) by performing the change of variables

$$x(n) = P(a)y(n),\ n = 0,1,\ldots\ , \tag{1.29}$$

where $a > 0$ is for the moment unspecified. The new equation is

$$\begin{aligned} x(n+1) &= (P(a)DP(a)^{-1})x(n) + P(a)G(n,P(a)^{-1}x(n)) \\ &\equiv \bar{D}x(n) + \bar{G}(n,x(n)),\ n = 0,1,\ldots\ , \end{aligned} \tag{1.30}$$

where $\bar{G}(n,x) = P(a)G(n,P(a)^{-1}x)$ and $\bar{D} = (\bar{d}_{ij})$ has $\bar{d}_{ii} = \lambda_i$, $\bar{d}_{i+1,i} = a\theta_i$, and $\bar{d}_{ij} = 0$ otherwise. The purpose of this transformation is to reduce $\|D_i\|$ while preserving the properties of G. In fact, we have

(1.31a) $\|\bar{D}_i\| \leq \rho_i + a$, $\|\bar{D}_i^{-1}\| \leq (\rho_j - a)^{-1}$ for $a < \rho_j$, $\bar{D}_i = P_i\bar{D}$,

(1.31b) $\|P(a)\| \leq a$, $\|P(a)^{-1}\| \leq a^{-p}$ for $a \leq 1$,

(1.31c) $\|\bar{G}(n,x)\| \leq a\,\|G(n,P(a)^{-1}x)\| \leq ac^*\,\|P(a)^{-1}x\| \leq c^*a^{1-p}\,\|x\|$, $a \leq 1$,

for $\|P(a)^{-1}x\| \leq \eta$ or $\|x\| \leq a^p\eta$, $a \leq 1$,

(1.31d) $\displaystyle\lim_{\substack{n\to\infty \\ x\to 0}} \frac{\|\bar{G}(n,x)\|}{\|x\|} = 0$,

(1.31e) $\displaystyle\frac{1}{a}\|x(n)\| \leq \frac{\|x(n)\|}{\|P(a)\|} \leq \|y(n)\| = \|P(a)^{-1}x(n)\| \leq \frac{1}{a^p}\|x(n)\|$,

for $a \leq 1$, $n \geq 0$.

Accordingly, if we prove the theorem for equation (1.30) with a specific $a \in (0,1)$, then the same result follows for (1.20) from (1.29).

Choose $a \in (0,1)$ so that

$$\rho_s+a<1, \; \rho_{s-1}-a>1, \; \rho_s+a\leq\alpha(\rho_{s-1}-a), \tag{1.32}$$
$$\rho_{s+1}+a\leq\alpha(\rho_s-a), \quad \alpha \text{ from } (1.22)\ .$$

Then (1.30) is an admissible equation with $\|\prod_{k=\ell}^{\ell+m} \bar{D}_i\| \leq (\rho_s+a)^{m+1}$ for $i = s,\ldots,r$, and

$$\sum_{j=0}^{\infty} \|(\prod_{k=0}^{j} \bar{D}_i)^{-1}\| \leq \sum_{j=0}^{\infty} \left(\frac{1}{\rho_{s-1}-a}\right)^{j+1} < $$

for $i = 1,\ldots,s-1$. From this and (1.31) it is clear that Theorem 1.4 applies to (1.30). The conclusion is that for $\|P_i x(0)\|$, $i = s,\ldots,r$, and $a^{1-p}c^*$ sufficiently small, there is a solution $\{x(n)\}$ of (1.30) with $\lim_{n\to\infty} x(n) = 0$ and $\limsup_{n\to\infty} \|x(n)\|^{1/n} \leq \rho_s+a$. We have used here the equivalency of all $\mathbb{C}^p$-norms.

We next apply Lemma 1.7 to show that, under additional assumptions on $x(0)$, (1.26) and (1.28) hold. Let $\|P_I x\| = \sum_{i=1}^{s-1} \|P_i x\|_1$, $\|P_{II}x\| = \sum_{i=s+1}^{r} \|P_i x\|_1$, and $\bar{c}_n = \sup\{\frac{\|\bar{G}(n,x)\|}{\|x\|} \mid 0 < \|x\| \leq \|x(n)\|\}$ for $\{x(n)\}$ to be defined. A solution of (1.30) satisfies

$$\|P_{II}x(n+1)\| \leq (\rho_{s+1}+a)\,\|P_{II}x(n)\| + \bar{c}_n\|x(n)\| \ , \tag{1.33a}$$

$$\|(P_S+P_I)x(n+1)\| \geq (\rho_s-a)\,\|(P_S+P_I)x(n)\| - \bar{c}_n\|x(n)\| \tag{1.33b}$$

Furthermore, $\bar{c}_n \leq a^{1-p} c_n$, and from Lemma 1.7 we have that if

$$\|P_{II}x(0)\| < \varepsilon\|P_S x(0)\| \leq \varepsilon\|(P_S+P_I)x(0)\|$$

and $\|x(0)\|$, $\bar{c}_n$ are sufficiently small, then $\|(P_S+P_I)x(n)\| > 0$ and

$$\frac{\|P_{II}x(n)\|}{\|(P_S+P_I)x(n)\|} = O(a^{1-p}\phi_n) = O(\phi_n) \tag{1.34}$$

On the other hand, for this solution we also have

$$\|(P_{II}+P_S)x(n+1)\|_1 \leq (\rho_s+a)\ \|(P_{II}+P_S)x(n)\|_1 + \bar{c}_n\|x(n)\|_1 \tag{1.35a}$$

$$\|P_I x(n+1)\|_1 \geq (\rho_{s-1}-a)\ \|P_I x(n)\|_1 - \bar{c}_n\|x(n)\|_1 \tag{1.35b}$$

From Lemma 1.7 either

$$\frac{\|(P_S+P_{II})x(n)\|_1}{\|P_I x(n)\|_1} = O(\phi_n) \tag{1.36}$$

or

$$\frac{\|P_I x(n)\|_1}{\|(P_S+P_{II})x(n)\|_1} = O(\psi_n) \tag{1.37}$$

If (1.36) were true, then we would have

$$\frac{\|P_I x(n+1)\|_1}{\|P_I x(n)\|_1} \geq (\rho_{s-1}-a-\bar{c}_n) - \bar{c}_n \frac{\|(P_{II}+P_S)x(n)\|_1}{\|P_I x(n)\|_1} > 1$$

for n sufficiently large and $\lim_{n\to\infty} \|P_I x(n)\|_1 = \infty$. Hence (1.37) holds and combined with (1.32) yields (1.28). To prove (1.26) let $\gamma > 0$ be given and choose a norm $\|\cdot\|_\gamma$ such that $\|\bar{D}_s\|_\gamma \leq (\rho_s + \frac{\gamma}{2})$. Since $\lim_{n\to\infty} x(n) = 0$ and (1.31c), (1.28) hold, we can find $n_0 \geq 0$ such that $\|\bar{G}(n,x(n))\|_\gamma \leq \frac{\gamma}{2}\|P_s x(n)\|_\gamma$ for all $n \geq n_0$. Then we have

$$\|P_s x(n+1)\|_\gamma \leq \|\bar{D}_s\|_\gamma \|P_s x(n)\|_\gamma + \|\bar{G}_s(n,x(n))\|_\gamma$$

$$\leq (\rho_s + \gamma)\|P_s x(n)\|_\gamma \text{ for } n \geq n_0 .$$

Similarly we obtain $\|P_s x(n+1)\|_\gamma \geq (\rho_s - \gamma)\|P_s x(n)\|_\gamma$ for n large and $\rho_s \neq 0$, by means of the estimates

$$\|P_s x(n)\|_\gamma \leq \|D_s^{-1}\|_\gamma \|P_s x(n+1)\|_\gamma + \|D_s^{-1}\|_\gamma \|G_s(n,x(n))\|_\gamma$$

and $\|\bar{D}_s^{-1}\|_\gamma \leq (\frac{1}{\rho_s} + \gamma')$ for an appropriate $\gamma' > 0$. Since $\gamma > 0$ is arbitrary in both cases, we have

$$\lim_{n\to\infty} \|P_s x(n)\|_\gamma^{1/n} = \lim_{n\to\infty} \|P_s x(n)\|_1^{1/n} = \rho_s$$

and with (1.28) the proof is finished.

Notice that Lemma 1.7 was used only to obtain initial values such that $\lim_{n\to\infty} \|y(n)\|^{1/n} \geq \rho_s$ and (1.28) hold.

For completing the discussion of this conditional stability result, we use a simple transformation to prove the following corollary, also given by Coffman [1964]. Some credit is due to Panov [1964] for concluding essentially these results, but from those of Panov [1959], which, as mentioned, are based on erroneous proofs.

Corollary 1.9. The results of Theorem 1.8 hold when any index $\ell > s$ replaces s, i.e., for any $\rho_\ell < 1$. Also, the assumption that $\rho_{s-1} > 1$ may be weakened to $\rho_{s-1} \geq 1$ in (1.18).

For the proof choose ρ such that $\rho\rho_\ell < 1$, $\rho\rho_{\ell-1} > 1$, and $\rho > 1$, and make the substitution $z(n) = \rho^n y(n)$. Equation (1.20) becomes

$$z(n+1) = (\rho D)z(n) + \rho^{n+1}G(n, \frac{1}{\rho^n} z(n)), \; n=0,1,\ldots ,$$

and Theorem 1.8 applies. The largest $\rho_i < 1$ is $\rho\rho_\ell$ so that we have $\{z(n)\}$ with $\lim_{n\to\infty} \|z(n)\|^{1/n} = \rho\rho_\ell$. But $\|z(n)\|^{1/n} = \rho\|y(n)\|^{1/n}$. For the second part simply use the same substitution with $\rho > 1$, $\rho\rho_s < 1$.

Perron [1929] proved this result under the more stringent conditions (1.16) on G. However, the following partial converse is due to him. It was reproved by Coffman [1964] using Lemma 1.7 and an argument similar to that found at the end of the proof of Theorem 1.8.

Theorem 1.10. Suppose that D and G satisfy the conditions of Theorem 1.8. If $y(n)$ is any solution of (1.20) which satisfies $\lim_{n\to\infty} y(n) = 0$, $y(n) \neq 0$, then

$$\lim_{n\to\infty} \|y(n)\|^{1/n} = \rho_\ell \tag{1.38}$$

for some $\rho_\ell \leq 1$, and $\|P_j y(n)\|_1 = o(\|P_\ell y(n)\|_1)$ for $j \neq \ell$.

We finish this section with two special cases of (1.20) which are of particular interest. The first is for $\rho(D) = \rho(A) < 1$ and the second for diagonal D.

If $\rho(A) < 1$, then the $\{0\}$ solution of (1.17), under conditions (1.23) on the perturbation, is clearly asymptotically stable and stable. Notice however that no existence proof is required, so that F need not be continuous. This result was proved by Perron [1929, Theorems 5, 7, 11] in the following form.

Theorem 1.11. For equation (1.17) suppose (1.23) holds for and that $\rho(A) < 1$. Then $\{0\}$ is stable and asymptotically stable. In that case, whenever $\{y(n)\}$ solves (1.1) and $\lim_{n\to\infty} y(n) = 0$, $y(n) \neq 0$, then

$$\lim_{n\to\infty} \|y(n)\|^{1/n} = |\lambda_\ell| \leq \rho(A)$$

for some eigenvalue λ_ℓ of A. On the other hand, if $\rho(A) > 1$, then $\{0\}$ is unstable.

We remark here that Perron was apparently unaware of the Jordan form and used the Schur transformation to lower triangular form. Then by employing the similarity transformation with matrix $P(a)$, $a > 0$, as given above, he obtained a matrix whose diagonal elements were the eigenvalues of A and whose off-diagonal components were of order a, and hence could be made small. A similar approach was taken by Perron's student Ta Li for the variable matrix equation to be seen in Section 3.

The second special case occurs when $D = \operatorname{diag}(d_1,\ldots,d_p)$. It is of historical interest because it contains the well-known Poincaré-Perron difference equation of order p,

$$a_{n+p} + b_1^{(n)} a_{n+p-1} + \cdots + b_p^{(n)} a_n = 0, \quad n=0,1,\ldots \tag{1.39}$$

where a_i, $b_i^{(j)} \in \mathbb{C}$, $\lim_{n\to\infty} b_i^{(n)} = b_i \in \mathbb{C}$, and the roots of the "characteristic equation" $z^p + b_1 z^{p-1} + \ldots + b_{p-1} z + b_p = 0$ have distinct moduli. The first studies of perturbed linear equations began here.

For this case we have a statement about the Q-convergence rate. Phrased in the context of Theorem 1.8, the following general result arises.

<u>Corollary 1.12</u>. Assume that all of the hypotheses of Theorem 1.8 hold and that $\{y(n)\}$ is a solution of (1.20) satisfying (1.26), (1.27) and (1.28). Then, if $P_s y(n) \neq 0$, there is a constant β such that

$$\rho_s - a - \beta\phi(n) \leq \frac{\|P(a)y(n+1)\|_1}{\|P(a)y(n)\|_1} \leq \rho_s + a + \beta\phi(n) \tag{1.40}$$

for all n sufficiently large and $a > 0$ small enough. In particular, if D_s is a diagonal matrix, then

$$\frac{\|y(n+1)\|_1}{\|y(n)\|_1} = \rho_s + O(\phi_n) \text{ as } n \to \infty , \tag{1.41}$$

or

$$Q_1(\{y(n)\},0) = \rho_s .$$

<u>Proof</u>: In view of the inequalities

$$\begin{aligned}(\rho_s - a)\,\|P_s P(a)y(n)\|_1 - \phi(n)\,\|P_s P(a)y(n)\|_1 \\ \leq \|P_s P(a)y(n+1)\|_1 \leq (\rho_s + a)\|P_s P(a)y(n)\|_1 \\ + \phi(n)\|P_s P(a)y(n)\|_1 ,\end{aligned}$$

(1.40) follows directly from (1.28). If D_s is diagonal, the above estimates hold, with $P(a)$ and a omitted, yielding (1.41).

Section 3. Equations with Variable Matrices

In this section we consider the more difficult general equation (1.1) with variable $A(n)$. Of course, if $\lim_{n\to\infty} A(n) = A$ or $\|A(n)-A\|$ is sufficiently small, then the results of Section 2 can be applied to

$$y(n+1) = Ay(n) + [(A(n)-A)y(n)+F(n,y(n)]\,,\ n=0,1,\ldots\ .$$

However, we will give an analysis which does not require such assumptions.

Again we apply a change of variables to transform (1.1) into canonical form. The resulting equation will be

$$(1.42)\qquad y(n+1) = D(n)y(n) + [L(n)y(n)+H(n,y(n))]\,,\ n=0,1,\ldots$$

where $D(n) = \text{diag}\ (d_1(n),\ldots,d_p(n))$, $L(n)$ is strictly lower triangu and H has the properties of F. In applying the results of Section 1, we consider each $d_i(n)$ as a diagonal block $D_i(n)$, $i=1,\ldots,p$, and obtain the condition for admissibility (1.6)/(1.7) in terms of these scalars, i.e., there exist constants $\alpha \geq 1$, $0 \leq \lambda < 1$, $\beta \geq 0$ such that for some $s \in \{1,\ldots,p\}$

$$(1.43)\qquad \prod_{k=\ell}^{\ell+m} |d_i(k)| \leq \alpha\lambda^{m+1} \text{ for all } m,\ell \geq 0;\ i=s,s+1,\ldots,p\ ,$$

$$(1.44)\qquad \sum_{j=0}^{n} \Big(\prod_{k=0}^{j} \frac{1}{|d_i(k)|} \Big) \leq \beta \text{ for } n=0,1,\ldots;\ i=1,2,\ldots,s-1$$

The existence of a transformation of (1.1) into (1.42) which preserves the norm of $y(n)$ and the required properties of F is guaranteed by the following lemma due to Ta Li [1934].

Lemma 1.13. For any sequence of matrices $\{A(n)\}$ from $L(\mathbb{C}^P)$ there exists a sequence of unitary matrices $\{U(n)\}$ in $L(\mathbb{C}^P)$ such that

$$U(n+1)A(n)U(n)^{-1} \equiv D(n) + L(n), \quad n=0,1,\ldots ,$$

is lower triangular. If all $A(n)$ are real, then $U(n)$ can be chosen real.

The change of variables

$$(1.45) \qquad y(n) = U(n)x(n), \quad n=0,1,\ldots$$

applied to (1.1) yields

$$(1.46) \qquad y(n+1) = D(n)y(n) + [L(n)y(n)+U(n+1)F(n,U(n)^{-1}y(n))]$$

$$\equiv D(n)y(n) + L(n)y(n) + H(n,y(n)), \quad n=0,1,\ldots ,$$

and

$$\|y(n)\|_2 = \|x(n)\|_2, \; \|U(n)\|_2 = \|U(n)^{-1}\|_2 = 1$$

We first prove a conditional stability result for (1.42) which extends a theorem of Ta Li [1934] by weakening the assumptions on

$$(1.47) \qquad G(n,y) \equiv L(n)y + H(n,y)$$

from (1.16) to (1.48). Ta Li used the contractive property of the perturbation in a successive approximation argument as discussed in Theorem 1.6.

Theorem 1.14. Suppose that (1.42) is an admissible equation (i.e., (1.43)/(1.44) hold) with $s \leq p$. Assume further that there are constants $c, c^*, \eta > 0$ such that

(1.48a) $\|L(n)\| \leq c$, $n=0,1,\ldots$,

(1.48b) $H(n,y)$ is continuous in y for $\|y\| \leq \eta$ and $n=0,1,\ldots$

(1.48c) $\|H(n,y)\|_1 \leq c^*\|y\|_1$ for $\|y\|_1 \leq \eta$, $n=0,1,\ldots$, and

$$\lim_{\substack{n\to\infty \\ y\to 0}} \frac{\|H(n,y)\|}{\|y\|} = 0$$

Then, if $z(0) \in \mathbb{C}^p$ and $|P_i z(0)|$, $i=s,s+1,\ldots,p$, c^* are sufficiently small, there is a solution $\{y(n)\}$ of (1.42) with the properties

(1.49a) $P_i(y(0)-z(0)) = 0$, $i=s,\ldots,p$,

(1.49b) $\lim_{n\to\infty} y(n) = 0$,

(1.49c) $\limsup_{n\to\infty} \|y(n)\|^{1/n} \leq \lambda + O(\|y(0)\|)$ as $\|y(0)\| \to 0$

(1.49d) $\{y(n)\}$ satisfies (1.4), (1.5) .

Proof: Introduce the change of variable $x(n) = P(a)y(n)$, $n=0,1,\ldots$, $a > 0$, as in the proof of Theorem 1.8. Then (1.42) becomes

$$(1.50) \qquad x(n+1) = D(n)x(n) + (P(a)L(n)P(a)^{-1})x(n)$$

$$+ P(a)(H(n,P(a)^{-1}x(n))), \; n=0,1,\ldots ,$$

and now $(P(a)L(n)P(a)^{-1})_{ij}$ is $a^{i-1}(L(n))_{ij}$ for $i > j$ and 0 otherwise. Thus,

$$\|P(a)L(n)P(a)^{-1}x + P(a)H(n,P(a)^{-1}x)\|_1 \leq (ac+c^*a^{1-p}) \|x\|_1$$

for $\|P(a)^{-1}x\|_1 \leq \eta$, $a < 1$, and

$$\lim_{\substack{n\to\infty \\ x\to 0}} \frac{\|P(a)H(n,P(a)^{-1}x)\|}{\|x\|} = 0$$

Applying Theorem 1.4, we obtain the conclusion that for $|P_i x(0)|$, $i=s,\ldots,p$, ac, and $a^{1-p}c^*$ sufficiently small, $a \geq 0$ fixed, a solution $\{x(n)\}$ of (1.3) exists, converges to 0, and satisfies $\limsup_{n\to\infty} \|x(n)\|^{1/n} \leq (\lambda+ac)$, as well as (1.4), (1.5). From the inverse transformation $y(n) = P(a)^{-1}x(n)$, the theorem is proved.

We remark at this time that a more comprehensive result for this equation (1.42) which gives exact componentwise convergence behavior as in Theorem 1.8 and Corollary 1.9, is possible and can be proved using arguments analogous to those for the constant matrix case.

To apply this result for (1.1), we must find conditions on $A(n)$ which guarantee that $D(n) = (d_i(n))$ satisfies (1.43), (1.44). The following lemma, also due to Ta Li [1934], will enable us to give a general theorem. The idea here is to use an auxiliary non-homogeneous linear equation to determine the properties of $D(n)$ from those of $A(n)$.

Lemma 1.15. For a sequence of matrices $\{A(n)\}$ from $L(\mathbb{C}^p)$ consider the difference equation

$$y(n+1) = A(n)y(n) + \omega(n), \quad n = 0,1,\dots \tag{1.51}$$

If $A(n)$ is uniformly bounded and if (1.51) has at least one bounded solution for each bounded sequence $\{\omega(n)\} \in S$, then after a transformation by unitary operators $U(n)$ from Lemma 1.13, the matrix $(D(n)+L(n)) = U(n+1)A(n)U(n)^{-1}$ has diagonal entries $d_i(n)$ each of which satisfies exactly one of the following conditions:

$$\sum_{\ell=0}^{n} \Big(\prod_{k=\ell}^{n} |d_i(k)| \Big) \text{ is bounded in } n, \tag{1.52}$$

$$\sum_{\ell=n}^{\infty} \Big(\prod_{k=n}^{\ell} \frac{1}{|d_i(k)|} \Big) \text{ exists and is bounded for each } n. \tag{1.53}$$

In fact, the general bounded solution $\{y(n)\}$ of (1.51) has the property that $y(0)$ may be chosen from a q-parameter family where q is the number of indices i for which (1.52) holds.

As a final preparation for the main theorem we quote a theorem due to Hahn [1958] which relates condition (1.52) with (1.43).

Lemma 1.16. Let $\{a_n\}$ be a sequence of complex numbers such that either $a_j \neq 0$ for all j or the number of nonzero a_j's between any two consecutive a_j's which are zero is bounded. Then the sequence

$$b_n = \sum_{\ell=0}^{n} \left(\prod_{k=\ell}^{n} |a_k| \right)$$

is bounded if and only if there exist real numbers $\alpha \geq 1$, $0 < \lambda < 1$ such that

$$\prod_{k=\ell}^{\ell+m} |a_k| \leq \alpha\lambda^{m+1} \text{ for all } \ell, m \geq 0 .$$

Hahn proved this result for the case in which all a_n are nonzero. However, the above statement follows easily from his proof.

We can now prove the general theorem. It extends a result of Ta Li [1934] by weakening the conditions on both $A(n)$ and F.

Theorem 1.17. For equation (1.1) suppose $A(n)$ is uniformly bounded and F is defined and continuous in y for each n and $\|y\| \leq \eta$, $\eta > 0$. Assume that the auxiliary equation (1.51) has at least one bounded solution for each bounded sequence $\{\omega(n)\} \in S$ and that the linear equation

$$(1.54) \qquad y(n+1) = A(n)y(n), \; n = 0,1,\ldots$$

has exactly a q-dimensional manifold of vectors $y(0)$ which initiate bounded solutions. If F satisfies the conditions

(1.55a) $\quad \|F(n,y)\| \leq c^*\|y\|$ for $\|y\| \leq \eta$, $n = 0,1,\ldots$,

(1.55b) $$\lim_{\substack{n\to\infty \\ y\to 0}} \frac{\|F(n,y)\|}{\|y\|} = 0 \ ,$$

for c^* sufficiently small and if the number of nonsingular matrices $A(n)$ between any two consecutive singular $A(n)$'s is bounded, then there is a q-parameter family of vectors $y(0)$ which when sufficiently small initiates solutions $\{y(n)\}$ of (1.1) which converge to 0.

Proof: Apply the unitary change of variables (1.45) to each equation (1.1), (1.51), (1.54) to obtain

(1.1)' $\quad y(n+1) = (D(n)+L(n))y(n) + H(n,y(n))$

(1.51)' $\quad y(n+1) = (D(n)+L(n))y(n) + U(n+1)\omega(n)$

(1.54)' $\quad y(n+1) = (D(n)+L(n))y(n)$

From Lemma 1.15, the components $d_i(n)$ of $D(n)$ each satisfy one of (1.52), (1.53), and from Lemma 1.16, (1.52) is equivalent to (1.43). Next, the assumption about the solutions of (1.54) and hence those of (1.54)' implies that (1.43) holds for exactly q indices and (1.53) for exactly (p-q). Thus, the hypotheses of Theorem 1.14 are satisfied for equation (1.1)'. Finally, since $\|y(n)\|_2 = \|U(n)x(n)\|_2 = \|x(n)\|_2$, the conclusions of Theorem 1.14 hold for the original equation (1.1) and the proof is complete.

As at the end of Section 2, we will briefly discuss the case of total stability of the trivial solution of (1.1). For brevity, we use the term stable to include asymptotically stable.

For the basic theorem the assumptions are

(1.56) there exist constants $\alpha \geq 1$, $0 < \lambda < 1$ such that

$$\left\| \prod_{j=k}^{k+\ell} A(j) \right\| \leq \alpha\lambda^{\ell+1} \text{ for all } k, \ell \geq 0 ;$$

(1.57) $\frac{\|F(n,y)\|}{\|y\|}$ is bounded for all $n \geq 0$ and $\|y\| \leq \eta_1$, $\eta_1 > 0$;

(1.58) for $c^* = c^*(\alpha,\lambda) < \frac{1-\lambda}{\alpha}$, there exist $\eta_2 > 0$, $n_0 \geq 0$ such that

$$\|F(n,y)\| \leq c^*\|y\| \text{ for all } \|y\| \leq \eta_2,\ n \geq n_0 .$$

Notice that (1.57), (1.58) is a weaker restriction on F than that of Theorem 1.14 or 1.17. The reason for this is that we do not have to prove existence of specific starting vectors $y(0)$, and so F need not even be continuous. The condition (1.56) will be discussed in the Appendix.

The theorem can then be stated as follows:

Theorem 1.18. For equation (1.1) suppose $A(n)$, F satisfy conditions (1.56), (1.57), and (1.58). Then the trivial solution is stable.

Proof: First consider (1.1) for $n \geq n_0$. Norm estimates give

$$\|y(n+1)\| \leq \alpha\lambda^{n+1-n_0}\|y(n_0)\| + \sum_{k=n_0}^{n} \alpha\lambda^{n-k}c^*\|y(k)\|,$$

$$\text{for } n = n_0, n_0+1, \ldots; \ \|y(k)\| \leq \eta_2 .$$

Referring to Lemma 1.3, we see that $\|y(n)\|$ is bounded by a sequence $a_n = b_n\alpha\lambda^{n-n_0}$ where $\alpha b_{n_0} = \|y(n_0)\|$, $b_{n_0+1} = (1+\frac{c^*}{\lambda})\alpha b_{n_0}$, $b_{n+1} = (1+\frac{\alpha c^*}{\lambda})b_n$ for $n > n_0+1$. Since $\lambda+\alpha c^*<1$, $\lim_{n\to\infty} a_n = 0$ and in fact $R_1\{a_n\} \leq (\lambda+\alpha c^*)$. Hence there is a constant η_3 such that if $\|y(n_0)\| \leq \eta_3$, then $\|y(n)\| \leq \eta_2$ for $n \geq n_0$. From (1.57) we have a constant $c \geq 0$ such that for $\|y(n)\| \leq \eta_1$

$$\|y(n+1)\| \leq \|A(n)\|\|y(n)\| + \|F(n,y(n))\| \leq (\alpha+c)\|y(n)\|, \ n=0,1,\ldots .$$

Thus, if $\|y(0)\| \leq \max(\eta_3,\eta_1)/(\alpha+c)^{n_0+1}$, then $\|y(n_0)\| \leq \eta_3$.

This result appears to be new but is closely related to the following theorem of Hahn [1958]:

<u>Corollary 1.19</u>. If the trivial solution of the linear equation

$$y(n+1) = A(n)y(n), \ n = 0,1,\ldots \tag{1.59}$$

is exponentially stable, then so is the trivial solution of (1.1) provided that

$$\|F(n,y)\| \leq c^*\|y\|, \text{ for } \|y\| \leq \eta, \ n \geq n_0 , \tag{1.60}$$

with c^* sufficiently small, and $\frac{\|F(n,y)\|}{\|y\|}$ is bounded for $n \geq 0$, $\|y\| \leq \eta$.

<u>Proof</u>: From the condition on (1.59) we have the existence of

constants $\alpha,\beta,\varepsilon > 0$ such that for any $n_1 \geqslant 0$, $\|y(n_1)\| \leqslant \varepsilon$ implies $\|y(n)\| \leqslant \alpha\|y(n_1)\|e^{-\beta(n-n_1)}$ for $n \geqslant n_1$. But

$$\|y(n)\| = \|(\prod_{j=n_1}^{n-1} A(j))y(n_1)\|$$

and since

$$\|\prod_{j=n_1}^{n-1} A(j)\| = \frac{1}{\|x\|}\|(\prod_{j=n_1}^{n-1} A(j))x\|$$

for some $x \neq 0$, we have for $y(n_1) = \varepsilon\frac{x}{\|x\|}$,

$$\|y(n)\| = \|\prod_{j=n_1}^{n-1} A(j)\|\|\varepsilon\frac{x}{\|x\|}\| \leqslant \alpha\|y(n_1)\|e^{-\beta(n-n_1)}$$

or

$$\|\prod_{j=n_1}^{n} A(j)\| \leqslant \alpha(e^{-\beta})^{n-n_1}, \text{ for any } n \geqslant n_1 \geqslant 0 .$$

But this is exactly (1.56) with $e^{-\beta} \equiv \lambda$. From Theorem 1.18 and its proof we know that for $c^* < \frac{\lambda-\lambda^2}{\alpha}$ and $y(n_1)$ sufficiently small, the solution $\{y(n)\}$ of (1.1) satisfies $\|y(n)\| \leqslant \|y(n_1)\|\gamma\lambda_1^{n-n_1}$ where $\gamma > 0$, $0 < \lambda_1 < 1$ are constants independent of n_1, n. This completes the proof.

In the appendix we discuss sufficient conditions for (1.1) to be admissible, in particular (1.56).

Appendix. Conditions for Admissibility in the Total Stability Case

The purpose of this addition to Chapter 1 is to discuss and give sufficient conditions for (1.56), which yields admissibility for the total stability case. Recall that in Section 2 we assumed $\rho(A) < 1$, while for the variable coefficient problem of Section 3 we supposed that $D(n)$ satisfied (1.43) and $L(n)$ was bounded. Each implies (1.56). On the other hand, from the expansion of (1.1),

$$y(n+1) = \left(\prod_{j=0}^{n} A(j)\right)y(0) + \sum_{k=0}^{n}\left(\prod_{j=k}^{n} A(j)\right)F(k,y(k)),\ n=0,1,\dots ,$$

it is apparent that the condition that $\sum_{k=0}^{n}\left\|\prod_{j=k}^{n} A(j)\right\|$ is bounded is reasonable for total stability. But this condition and Lemma 1.16 yield (1.56) in the case that $A(n)$ is nonsingular for all n.

Because (1.56) seems to be so crucial, we should give some conditions for it which can be verified. For the two conditions mentioned above one is simple and the other uncheckable. The following lemma can be verified by a close examination of the proof of a result of Smith [1966, Theorem 4].

Lemma 1.20. Let $\{A(n)\}$ be a bounded sequence of matrices from $L(\mathbb{C}^p)$ with $\limsup_{n\to\infty} \rho(A(n)) < 1$. Then there exists an $\varepsilon > 0$ such that $\limsup_{n\to\infty} \|A(n+1)-A(n)\| \leq \varepsilon$ implies (1.56).

In fact, if there exists an $n_0 \geq 0$ and $\beta,\delta > 0$ such that

(1.61) $\quad \|A(n)\|_2 \leq \beta$ and $\rho(A(n)) \leq 1-\delta$ for all $n \geq n_0$ and $\beta,\delta > 0$, then ε can be chosen to be $(pW\ln[p(p-\delta)^{-1}W])^{-1}([\delta(p-\delta)]^{1/2}\ln(p(p-\delta)^{-1})$ where $W = (1+\beta)^{p-1}\ ([(p-1)\delta]^{-1}p)^{p-1/2}$.

It is interesting to note here that although Smith actually needed this result to prove a corollary equivalent to Corollary 1.19, he did not state it, and the corollary, based on a theorem of Hahn, was in fact not proved.

We state without proof a second lemma which follows from several estimates made by Smith and appears to be new.

Lemma 1.21. For the sequence of matrices $\{A(n)\}$, suppose there is another matrix $A \in L(\mathbb{C}^p)$ with $\rho(A) < 1$ and such that

$$\limsup_{n\to\infty} \|A(n)-A\|_2 < \left(\frac{1}{pe}\right) \frac{(1+|\lambda_1|)\prod_{i=1}^{p}(1-|\lambda_i|)^2}{(1+\|A\|_2)^{2p-2}}$$

where λ_i are the eigenvalues of A in order of decreasing modulus and e is the Euler number. Then (1.56) holds.

As a final remark, we point out that even the problem of finding sufficient conditions on $\{A(n)\}$ such that $\|\prod_{j=0}^{n} A(j)\|$ converges to 0 is nontrivial and that counterexamples can be given to show that (1.61) is not sufficient (see Smith [1966]).

Chapter 2

Convergence of Iterative Processes: Direct Local Approach

We return now to the iterative process (I-1) and consider first its stationary form:

$$(2.1) \qquad x(n+1) = G(x(n)), \ n = 0,1,\ldots \ .$$

The expansion (I-2) leads to

$$(2.2) \qquad y(n+1) = Ay(n) + F(y(n)), \ n = 0,1,\ldots \ ,$$

where

$$(2.3) \qquad y(n) = x(n) - x^* \in X, \ G'(x^*) = A \in L(X) \ ,$$

and, under the assumption that G is F-differentiable at x^*,

$$(2.4) \qquad F(y) = o(\|y\|) \text{ as } \|y\| \to 0$$

It was Ostrowski [1966] who first used the form (2.2) to prove local convergence results for (2.1).

<u>Theorem 2.1</u>. Suppose $G: E \subseteq X \to X$ has a fixed point $x^* \in \text{int}(E)$ at which G is F-differentiable. If $\rho \equiv \rho(G'(x^*)) < 1$, then x^* is a point of attraction of (2.1) and $R_1\{\mathcal{J}, x^*\} \leq \rho$. If $\rho > 1$, then x^* is a point of repulsion for (2.1), i.e., there is a ball $S(x^*)$ about x^* and a solid cone L with vertex x^* such that if $x(0) \in S(x^*) \cap L$, then some $x(n) = x^*$ or else some $x(n) \notin S(x^*)$.

Ostrowski's proof used (I-2) and the following matrix product theorem.

Theorem 2.2. Let $A \in L(X)$ and $\varepsilon > 0$. Then there exist positive constants $\eta = \eta(\varepsilon,A)$, $\sigma = \sigma(\varepsilon,A)$ such that when $U_i \in L(X)$ and $\|U_i\| \leq \eta$ for $i = 1,2,\ldots$, then

$$\left\| \prod_{i=1}^{n} (A+U_i) \right\| \leq \sigma(\rho(A)+\varepsilon)^n, \quad n = 1,2,\ldots \quad (2.5)$$

If $\rho(A) > 1 + \alpha$ and $\alpha > 0$, then there is a constant $\delta = \delta(A,\alpha) > 0$ such that, when $U_i \in L(X)$ and $\|U_i\| \leq \delta$, $i = 1,2,\ldots$, then the product $\prod_{i=1}^{n} (A+U_i)$ diverges.

Both of these theorems are direct results of Theorem 1.11 due to Perron [1929]. For the matrix products, notice that

$$\left\| \prod_{i=1}^{n} (A+U_i) \right\| = \sup_{\substack{y \neq 0 \\ y \in X}} \frac{1}{\|y\|} \left\| \prod_{i=1}^{n} (A+U_i) y \right\| = \sup_{\substack{y(1) \neq 0 \\ y(1) \in X}} \frac{\|y(n)\|}{\|y(1)\|}, \quad n = 1,2,\ldots ,$$

where $\{y(n)\}$ is the solution of the difference equation

$$y(n+1) = (A+U_n)y(n), \quad n = 1,2,\ldots \quad (2.6)$$

Under stronger assumptions on G, Ortega and Rockoff [1966] proved the following theorem which gives the exact asymptotic rate of convergence.

Theorem 2.3. Let $G: E \subseteq X \to X$ have a fixed point $x^* \in \operatorname{int}(E)$ at which G is F-differentiable and $\rho \equiv \rho(G'(x^*)) < 1$. Suppose further that there exist a neighborhood E_0 of x^* in E and constants $c, \varepsilon > 0$ such that

$$\|Gx - Gx^* - G'(x^*)(x-x^*)\| \leq c\|x-x^*\|^{1+\varepsilon} \quad \text{for all } x \in E_0 . \quad (2.7)$$

Then x^* is a point of attraction for (2.1). Furthermore, if $\rho > 0$, for any sequence of iterates $\{x(n)\}$ with $\lim_{n\to\infty} x(n) = x^*$, there is a constant η such that

$$\|x(n)-x^*\| \leq \eta n^{j_0}\rho^n, \quad n = 0,1,\dots , \tag{2.8a}$$

and for some sequence of iterates $\{x(n)\}$ there is a constant $\theta > 0$ with

$$\|x(n)-x^*\| \geq \theta n^{j_0}\rho^n, \quad n = 0,1,\dots . \tag{2.8b}$$

Here (j_0+1) is the dimension of the largest Jordan block in the Jordan form of $G'(x^*)$ associated with an eigenvalue of modulus ρ.

The authors of this theorem have noted that if (2.7) is weakened to the assumption that G is F-differentiable, then (2.8) need not hold, but that Coffman's theorem [1964] (see Theorem 1.8) can be used to show that

$$O_R(\mathcal{J}, x^*) = \rho \tag{2.9}$$

still holds. A direct proof based on Coffman's proof can be found in Ortega and Rheinboldt [1970a].

All of the above results are contained in the following general theorem. It is an application of the stability theory for difference equations with constant matrices, namely, Theorem 1.8, which is essentially due to Coffman [1964].

Theorem 2.4. Let $G: E \subseteq X \to X$ have a fixed point $x^* \in \text{int}(E)$ at which G is F-differentiable. If $\rho(G'(x^*)) < 1$, then x^* is a point of attraction of (2.1). In fact, if $G'(x^*)$ has Jordan form D with blocks D_i of dimension e_i associated with eigenvalues of modulus ρ_i, $i = 1,\ldots,r$, respectively, and if $1 > \rho_1 > \cdots > \rho_r$, then for each s with $1 \leq s \leq r$ there is an e_s-dimensional manifold of initial values $x(0)$ for which the solution $\{x(n)\}$ of (2.1) satisfies

$$\text{(2.10)} \qquad R_1\{x(n)\} = \rho_s, \ \lim_{n\to\infty} x(n) = x^* ,$$

$$\text{(2.11a)} \qquad P_s(x(n+1)-x^*) = D_s^{n+1}P_s(x(0)-x^*) + O(\|\sum_{j=0}^{n} D_s^{n-j}F_s(x(j))\|) ,$$

$$\text{(2.11b)} \qquad P_i(x(n+1)-x^*) = o(\|P_s(x(n+1)x^*)\|), \text{ as } n \to \infty, \text{ for } i \neq s ,$$

where $F(x) \equiv Gx - Gx^* - G'(x^*)(x-x^*)$ and $F_s = P_sF$. On the other hand, any sequence of iterates which converges to x^* satisfies

$$\text{(2.12)} \qquad R_1\{x(n)\} = \rho_j \text{ for some } j \text{ or } R_1\{x(n)\} = 0 ,$$

and

$$\text{(2.13)} \quad P_i(x(n+1)-x^*) = D_i^{n+1}P_i(x(0)-x^*) + O(\|\sum_{j=0}^{n} D_i^{n-j}F_i x(j)\|), \text{ for all } i.$$

To see how the estimates (2.8) are obtained under conditions (2.7) on G, consider the following result.

Corollary 2.5. In addition to the hypotheses of Theorem 2.4, assume that for some $\rho_s > 0$ there is an isotone function ϕ such that for sufficiently small $\|x-x^*\|$, the critical components of F, $F_s = P_sF$, satisfy

$$(2.14) \qquad \|F_sx\| \leq \phi(\|x-x^*\|)$$

Suppose further that for some constants $\alpha > 0$, $\beta > 0$ with $\rho_s-\alpha>0$, $\rho_s+\beta<1$, we have

$$(2.15) \qquad \sum_{j=0}^{\infty} \frac{1}{(\rho_s-\alpha)^{j+1}} \phi(\gamma(\rho_s+\beta)^j) < \infty \quad \text{for any } \gamma > 0 .$$

Then for any iterates $\{x(n)\}$ satisfying (2.10), (2.11) for this s we have

$$(2.16) \qquad P_s(x(n+1)-x^*) = D_s^{n+1}P_s(x(0)-x^*) + O(\|D_s^{n+1}\|) \text{ as } n \to \infty .$$

If $\rho_s = 0$, there is an integer $n_1 \geq 0$ such that

$$(2.17) \qquad P_s(x(n+1)-x^*) = O(\|\sum_{j=n-n_1}^{n} D_s^{n-j}F(x(j))\|) \text{ as } n \to \infty .$$

Proof: Since $\rho_s > 0$, for any small $\alpha > 0$, there is a norm $\|\cdot\|_\alpha$ for which $\|D_s^{-1}\|_\alpha \leq \frac{1}{\rho_s-\alpha}$. Since (2.10) holds, we can choose $n_0, \gamma_0 \geq 1$ such that $\|x(n)-x^*\|_\alpha \leq (\rho_s+\beta)^n$ for $n \geq n_0$ and $\|x(n)-x^*\| \leq \gamma_0(\rho_s+\beta)^n$ for $n < n_0$. Then from (2.15)

$$\|\sum_{j=0}^{n} D_s^{n-j}F_s(x(j))\|_\alpha \leq \|D_s^{n+1}\|_\alpha (\sum_{j=0}^{n} \|D_s^{-1}\|_\alpha^{j+1} \quad \varepsilon\phi(\delta\gamma_0(\rho_s+\beta)^j))$$

$$\leq \theta\|D_s^{n+1}\|_\alpha \quad \text{for } n = 0,1,\ldots$$

where θ is constant and ε,δ are equivalency constants for the norms

$\|\cdot\|$, $\|\cdot\|_\alpha$. (2.17) is clear.

In the case of Theorem 2.3, we have $\phi(\|x-x^*\|) = c\|x-x^*\|^{1+\varepsilon}$ from (2.7). But then condition (2.15) holds and according to (2.16) there is a sequence of iterates $\{x(n)\}$ satisfying (2.16) [or (2.17) if $\rho_s = 0$], which is equivalent to (2.8) for $\rho_s = \rho_r = \rho$.

As an illustration of the application of Theorem 2.4, we consider the well-known Newton process for solving $Fx = 0$. For the sake of completeness we recall first the standard attraction theorem for which $F'(x^*)$ is nonsingular. In Theorem 2.7 we then prove a new result for the case when $F'(x^*)$ is singular.

<u>Theorem 2.6</u>. Suppose $F\colon D \subseteq X \to Y$ is continuously F-differentiable in the open set D. If $Fx^* = 0$ for some $x^* \in D$ and $F'(x^*)$ is nonsingular, then x^* is a point of attraction of Newton's method

$$(2.18) \qquad x(n+1) = x(n) - F'(x(n))^{-1}Fx(n),\ n = 0,1,\ldots$$

and $O_R(\mathcal{J},x^*) \geq 1$.

<u>Proof</u>: Here $Gx = x - F'(x)^{-1}Fx$ for $x \in D$ whenever $F'(x)^{-1}$ exists. Since $F'(x)$ is continuous in D and $F'(x^*)^{-1}$ exists, there is a ball $S \subseteq D$ containing x^* in which $F'(x)^{-1}$ exists. Then

$$\|Gx-Gx^*\| = \|x-x^*-F'(x)^{-1}Fx\| \leq \|F'(x)^{-1}\|\|Fx-Fx^*-F'(x)(x-x^*)\|$$

$$= o(\|x-x^*\|) \text{ for } x \in S \text{ as } x \to x^*$$

Hence G is F-differentiable at x^* with $G'(x^*) = 0$. An application of Theorem 2.4 completes the proof.

For a case in which $F'(x^*)$ is singular we have the following theorem which corrects a result of Rall [1966]. The crucial point of the proof involves showing that $G'(x^*)$ exists and has eigenvalues 0 and $\frac{1}{2}$

<u>Theorem 2.7.</u> For the function $F: D \subseteq X \to Y$ with D open, suppose the following conditions hold:

(2.19a) F is twice differentiable on D and $F''(x)$ is continuous with modulus of continuity $\omega_1(r)$;

(2.19b) $Fx = 0$ has a solution $x^* \in D$;

(2.19c) $F'(x)$ is nonsingular in a deleted neighborhood S of x^* (i.e., $x^* \notin S$), and $S \subseteq D$;

(2.19d) if N_1 denotes the null manifold of $F'(x^*)$, then there is a constant K such that for any $m \in N_1$ and any $x \in X$, the inequality

$$\|F''(x^*)xm\| \geq K\|x\|\cdot\|m\|$$

holds.

Then x^* is a point of attraction for Newton's method (2.18). In fact, if X_1 is a complementary summand for N_1 in X $(X_1 \oplus N_1 = X)$ and P_{N_1}, P_{X_1} are the corresponding projections of X onto X_1, N_1 respectively, [i.e., if $m \in N_1$, $y \in X_1$, then $P_{N_1}(m+y) = m$, $P_{X_1}(m+y)=y$], then $P_{X_1}(x_n-x^*)$ converges to 0 superlinearly while $P_{N_1}(x_n-x^*)$ converges to 0 with $R_1 \leq \frac{1}{2}$. If $\omega_1(\delta) \leq \text{constant } (\delta)^\gamma$, $\gamma > 0$, then

$P_{X_1}(x_n - x^*)$ converges and $R_{(1+\gamma)} < 1$. Finally, for $Gx \equiv x - F'(x)^{-1}Fx$, we have

$$G'(x^*) = \frac{1}{2} P_{N_1} \tag{2.20}$$

Proof: Choose a nonsingular matrix T such that $TF'(x^*)(X_1) = X_1$. But the process (2.18) for $TF(x)$ is identical with that for $F(x)$:

$$x_{n+1} = x_n - F'(x_n)^{-1}Fx_n = x_n - (TF'(x_n))^{-1}TF(x_n) .$$

Also $TFx = 0$ if and only if $Fx = 0$. Hence, from the start we can simply assume that $F'(x^*)(X_1) = X_1$.

For $x^* \neq x \in S$, let $R_x = F'(x)(X_1)$ and $S_x = F'(x)(N_1)$, and let P_{R_x}, P_{S_x} be the corresponding projections of X onto R_x, S_x. Now $R_x \oplus S_x = X$, for if $z \in R_x$ and and $z \in S_x$, then $z = F'(x)y_1 = F'(x)m_1$ for $y_1 \in X_1$, $m_1 \in N_1$. Hence $y_1 = m_1 = 0$, and so $z = 0$.

Consider the expansions for $x \in S$:

$$Fx = Fx^* + F'(x^*)(x-x^*) + \frac{F''(x^*)}{2}(x-x^*)^2 + O(\omega_1(\|x-x^*\|)\|x-x^*\|^2) \tag{2.21}$$

$$Fx^* = Fx - F'(x)(x-x^*) + \frac{F''(x)}{2}(x-x^*)^2 + O(\omega_1(\|x-x^*\|)\cdot\|x-x^*\|^2) \tag{2.22}$$

$$F'(x) = F'(x^*) + F''(x^*)(x-x^*) + O(\omega_1(\|x-x^*\|)\|x-x^*\|) \tag{2.23}$$

$$F'(x^*) = F'(x) + F''(x)(x^*-x) + O(\omega_1(\|x-x^*\|)\|x-x^*\|) \tag{2.24}$$

Then, if $y \in X_1$ and $x \in S$, from (2.23)

$$F'(x)y = F'(x^*)y + F''(x^*)(x-x^*)y + O(\omega_1(\|x-x^*\|)\|x-x^*\|\|y\|)$$

2

This implies the existence of constants $K_1 > 0$, $\delta_1 > 0$ for which

$$(2.25) \qquad \|F'(x)y\| \geq K_1\|y\|$$

whenever $\|x-x^*\| \leq \delta_1$, $y \in X_1$. K_1 is independent of x,y.

Immediately from (2.25)

$$(2.26) \qquad \frac{\|F'(x)^{-1}r\|}{\|r\|} \leq \frac{1}{K_1} \text{ for all } r \in R_x \text{ and } \|x-x^*\| \leq \delta_1,\ r\neq 0.$$

For $m \in N_1$, $x \in S$ we have again from (2.23):

$$F'(x)m = F'(x^*)m + F''(x^*)(x-x^*)m + O(\omega_1(\|x-x^*\|)\|x-x^*\|\|m\|)$$

Since $F'(x^*)m = 0$, we have the existence of constants $K_2, \delta_2 > 0$, $\delta_2 < \delta_1$, for which

$$\|F'(x)m\| \geq K_2\|x-x^*\|\|m\| \text{ for } \|x-x^*\| \leq \delta_2,\ m \in N_1$$

Here we have used (2.19) and can write

$$(2.27) \qquad \frac{\|F'(x)^{-1}s\|}{\|s\|} \leq \frac{1}{K_2\|x-x^*\|} \text{ for } \|x-x^*\| \leq \delta_2,\ s \in S_x,\ s\neq 0$$

Notice that (2.26) and (2.27) more or less describe how "fast" $F'(x)$ becomes singular as x tends to x^*. With the special case in which F is scalar-valued in mind, we might say that F has a root of multiplicity two at x^*.

Next define $Gx \equiv x - F'(x)^{-1}Fx$ for $x \in S$ and Gx^* by $\lim_{\substack{x\to x^* \\ x\in S}} Gx$.

Now this limit exists and equals x^* as can be seen from the following:

$$F'(x)^{-1}Fx = F'(x)^{-1}[F'(x^*)(x-x^*)+\omega_2(x)]$$

$$= F'(x)^{-1}[F'(x^*)(P_{N_1}+P_{X_1})(x-x^*)+\omega_2(x)]$$

$$= F'(x)^{-1}F'(x^*)P_{X_1}(x-x^*) + F'(x)^{-1}\omega_2(x)$$

for $x \in S$ where $\omega_2(x) = O(\|x-x^*\|^2)$ as in (2.21). Referring to (2.26) and (2.27) we have $\|F'(x)^{-1}\| \leq \frac{1}{K_3\|x-x^*\|}$ for $\|x-x^*\| \leq \delta_2$; and so with (2.25),

$$\|F'(x)^{-1}Fx\| \leq \frac{1}{K_1}\|F'(x^*)P_{X_1}(x-x^*)\| + \frac{K_4\|x-x^*\|^2}{K_3\|x-x^*\|} \leq K_5\|x-x^*\|$$

for $\|x-x^*\| \leq \delta_2$, $x \in S$, K_5 constant. Thus, if we let $Gx^* \equiv x^*$, then G is continuous in $S \cup \{x^*\}$.

Consider an alternate estimate of $F'(x)^{-1}Fx$ for $x \in S$, $\|x-x^*\| \leq \delta$:

$$F'(x)^{-1}Fx = F'(x)^{-1}[-F'(x)(x^*-x) - \frac{F''(x)}{2}(x^*-x)^2+\omega_3(x)]$$

$$= x-x^* - F'(x)^{-1}\frac{F''(x)}{2}(x^*-x)^2 + \omega_4(x)$$

where $\omega_3(x) = O(\omega_1(\|x-x^*\|)\|x-x^*\|^2)$ and $\omega_4(x) = O(\omega_1(\|x-x^*\|)\|x-x^*\|)$ from (2.22) and $\|F'(x)^{-1}\| \leq \frac{1}{K_3\|x-x^*\|}$. But then

$$\text{(2.28)} \qquad \|Gx-Gx^*-\frac{1}{2}P_{N_1}(x-x^*)\| \leq \|x-x^*-F'(x)^{-1}Fx-\frac{1}{2}P_{N_1}(x-x^*)\|$$

$$\leq \|F'(x)^{-1}\frac{F''(x)}{2}(x^*-x) - \frac{1}{2}P_{N_1}\|\,\|x-x^*\|$$

$$+ O(\omega_1(\|x-x^*\|)\|x-x^*\|)$$

Since $F'(x)^{-1}F''(x)(x^*-x) = F'(x)^{-1}[F'(x^*)-F'(x)] + O(\omega_1(\|x-x^*\|))$ from (2.24), we have for $m \in N_1$, $x \in S$, $\|x-x^*\| \leq \delta_2$,

$$F'(x)^{-1}F''(x)(x^*-x)m = -m + O(\omega_1(\|x-x^*\|)\|m\|)$$

and for $y \in X_1$

$$F'(x)^{-1}F''(x)(x^*-x)y = -y + F'(x)^{-1}F'(x^*)y + O(\omega_1(\|x-x^*\|)\|y\|)$$

$$= O(\omega_1(\|x-x^*\|)\|y\|)$$

Hence, $\left\|F'(x)^{-1}\frac{F''(x)}{2}(x^*-x) - \frac{1}{2}P_{N_1}\right\| = O(\omega_1(\|x-x^*\|))$ as $x \to x^*$ and

(2.29) $$\left\|Gx-Gx^* - \frac{1}{2}P_{N_1}(x-x^*)\right\| = O(\omega_1(\|x-x^*\|)\|x-x^*\|) \text{ as } x \to x^*$$

or $G'(x^*) = \frac{1}{2}P_{N_1}$. The estimate

$$Gx = Gx^* + \frac{1}{2}P_{N_1}(x-x^*) + O(\omega_1(\|x-x^*\|)\|x-x^*\|)$$

yields

$$x(n+1) - x^* = \frac{1}{2}P_{N_1}(x(n)-x^*) + O(\omega_1(\|x(n)-x^*\|)\|x(n)-x^*\|)$$

and the proof is complete.

It is clear that superlinear convergence can be restored by altering the process equation. Use $H(x) = G(x) - \frac{1}{2}P_{N_1}x$ so that

$$H'(x^*) = G'(x^*) - \frac{1}{2}P_{N_1} = 0 ,$$

and the corrected process is

$$x(n+1) = H(x(n)),\ n = 0,1,\ldots .$$

A theorem similar to the one above for roots of multiplicity greater than two can be formulated, but with additional conditions of the type (2.19d) on the derivatives of F.

Finally, we present the following examples which show that (2.19d) cannot be substantially weakened. Consider

$$(2.30a) \qquad F((x_1,x_2)^T) = (x_1^{2p}-x_2^2, x_1x_2)^T$$

for $p \geq 2$ fixed, and

$$(2.30b) \qquad F((x_1,x_2)^T) = \begin{cases} (e^{-1/x_1}-x_2^2,\ x_1 \cdot x_2)^T & \text{for } x_1 \neq 0 \\ (-x_2^2, 0)^T & \text{for } x_1 = 0 \end{cases}$$

In both cases, 0 is the only root of F in a neighborhood of 0, $F'((x_1,x_2)^T)$ is nonsingular for $x \neq 0$, $\|x\|$ small, and $F'(0) = 0$. Also

$$F''(0)\binom{\alpha}{\beta} = \begin{pmatrix} 0 & -2\beta \\ \beta & \alpha \end{pmatrix} \neq 0 \text{ for } \binom{\alpha}{\beta} \neq 0 .$$

However, with $x_2(0) = 0$ and $x_1(0) \in (0,1)$, we have

$$x_2(n) \equiv 0 \text{ and } \begin{cases} x_1(n+1) = \left(\frac{2p-1}{2p}\right) x_1(n) & \text{for (2.30a)} \\ x_1(n+1) = x_1(n)[1-x_1(n)] & \text{for (2.30b)} \end{cases}$$

Thus, the rates of convergence are given by $R_1 = \frac{2p-1}{2p}$, $R_1 = 1$, respectively.

Let us turn now to an application of the variable matrix theory of Section 3 in Chapter 1, in particular, of Theorem 1.18.

Theorem 2.8. For the iterative process

$$x(n+1) = G(n,x(n)),\ n = 0,1,\ldots \tag{2.31}$$

suppose that $G(n,\cdot)$ maps the open, convex domain $D \subseteq X$ into X, that there is an $x^* \in D$ with $G(n,x^*) = x^*$, and that $G(n,x)$ is uniformly differentiable in x at x^* for $n = 0,1,\ldots$ (i.e., for any $\varepsilon > 0$ there is a $\delta > 0$ such that when $x \in D$, $\|x-x^*\| < \delta$, we have $\|G(n,x) - G(n,x^*) - G'(n,x^*)(x-x^*)\| \leq \varepsilon\|x-x^*\|$ for all n). If, for some $a > 1$, $0 \leq \lambda < 1$ and for all $k,\ell \geq 0$ the following holds

$$\left\| \prod_{j=k}^{k+\ell} G'(j,x^*) \right\| \leq a\lambda^{\ell+1} \tag{2.32}$$

then x^* is a point of attraction for (2.15) and $R_1 \leq \lambda$.

Proof: For any $x(n) \in D$, we have

$$x(n+1) - x^* = G(n,x(n)) - G(n,x^*) = G'(n,x^*)(x(n)-x^*) + \Phi(n,x(n)) .$$

But

$$\frac{\|\Phi(n,x)\|}{\|x-x^*\|} \leq \frac{\|G(n,x)-G(n,x^*)-G'(n,x^*)(x-x^*)\|}{\|x-x^*\|}$$

so that

$$\lim_{x-x^*\to 0} \frac{\|\Phi(n,x)\|}{\|x-x^*\|} = 0$$

and convergence is uniform in n. Theorem 1.18 then yields the result.

As an application we have the following result of Ortega and Rheinboldt [1970].

Corollary 2.9. Let $G: D \times D_h \subseteq \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^p$ and assume that there is an $x^* \in \text{int}(D)$ with $x^* = G(x^*,h)$ for all $h \in D_h$. Assume that the functions $G(\cdot,h): D \subseteq \mathbb{R}^p \to \mathbb{R}^p$ are uniformly differentiable at x^* and that $G'(x^*,h) = H^{q(h)}$ where $H \in L(\mathbb{R}^p)$, $\rho(H) < 1$, $q(h)$ is a positive integer. Then there is a real number $\delta > 0$ such that for any $x(0) \in S = S(x^*,\delta)$ and any sequence $\{h^k\} \subseteq D_h$, the vectors

$$(2.33) \qquad x(k+1) = G(x(k),h^k),\ k = 0,1,\ldots$$

are well-defined, $\lim_{k\to\infty} x(k) = x^*$, and $R_1\{x(k)\} \leqslant \rho(H^m)$ where $m = \liminf_{k\to\infty} q(h^k)$.

Proof: Let $\varepsilon > 0$ be given. It is well-known that we may choose a norm $\|\cdot\|$ on $\mathbb{R}^p$ such that $\|H\| \leqslant (\rho(H)+\varepsilon) < 1$ and then

$$\left\|\prod_{j=k}^{k+\ell} G'(x^*,h^j)\right\| = \|H^{q(k)+q(k+1)+\ldots+q(k+\ell)}\|$$

$$\leqslant \|H\|^{q(k)+\ldots+q(k+\ell)}$$

$$\leqslant \left(\frac{(\rho(H)+\varepsilon)^{q(k)+\ldots+q(k+\ell)}}{[(\rho(H)+\varepsilon)^m]^{\ell+1}}\right) [(\rho(H)+\varepsilon)^m]^{\ell+1}$$

By the definition of m, the first factor above is bounded uniformly for all $k,\ell \geqslant 0$. Hence (2.32) of Theorem 2.8 holds and we see that x^* is a point of attraction for (2.33) and that $R_1\{x(k)\} \leqslant (\rho(H)+\varepsilon)^m$. Since $\varepsilon > 0$ was arbitrary, the theorem is proved.

The authors of Corollary 2.9 have applied it to the process

$$x(k+1) = x(k) - [I+H(x(k))+\ldots+H^{mk}(x(k))]B^{-1}(x(k))Fx(k), \quad k=0,1,\ldots$$

where $F'(x) = B(x) - C(x)$, $H(x) = B^{-1}(x)C(x)$.

Finally, we should remark that this analysis is not restricted to single-step processes. Consider the m-step method

$$(2.33) \qquad x(n+1) = G(n,x(n),\ldots,x(n-m+1)), \quad n = 0,1,\ldots$$

where $G(n,\ldots): D_n \subseteq X^m \to X$ and

$$(2.34) \qquad x^* = G(n,x^*,\ldots,x^*) \in D_n \text{ for each } n.$$

According to Voigt [1969], if $G(n,\ldots)$ is F-differentiable in each variable at x^*, then for the associated process

$$\begin{aligned} &(x(n+1),x(n),\ldots,x(n-m+2)) \\ (2.35) \qquad &= (G(n,x(n),\ldots,x(n-m+1)),x(n),\ldots,x(n-m+2)) \\ &\equiv H(n,x(n),\ldots,x(n-m+1)), \quad n = 0,1,\ldots, \end{aligned}$$

or $z(n+1) = H(n,z(n))$, $n = 0,1,\ldots$, we have

$$(2.36) \qquad H'(n,x^*,\ldots,x^*) = \begin{pmatrix} H_1(n) & H_2(n) & & & H_m(n) \\ I & O & \cdots & & O \\ O & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ O & \cdots & O & I & O \end{pmatrix}$$

where $H_i(n) = \partial_i G(n,x^*,\ldots,x^*)$, $i = 1,\ldots,m$. The multistep process (2.33) can then be considered a single-step method and the results given above applied. Of course, some difficulty is carried with the analysis of the block companion matrix (2.36). (See Voigt [1969] for further discussion of this problem.) Furthermore, with multistep processes we usually expect superlinear convergence; this will be considered in Part II.

PART II

MAJORIZATION APPROACH

Whereas in Part I the iterative process equation was analyzed directly and locally, here it is studied in terms of majorizing sequences or systems which arise from norm estimates.

We begin in Chapter 3 with several results for the type of difference equations which commonly occurs in this approach, namely, the isotone (that is, in each variable monotonically nondecreasing) equations:

(II-1) $\quad a_{n+1} = \phi(a_n, a_{n-1}, \ldots, a_{n-k}),\ n = 0,1,\ldots\ ,$

with ϕ isotone. The theorems obtained deal primarily with determining domains of initial vectors $(a_0, a_{-1}, \ldots, a_{-k})$ for which the solution $\{a_n\}$ of (II-1) is convergent or summable. Rate of convergence estimates are also of interest.

In Chapter 4, principles by which difference equations are generated from iterative processes through norm estimates are discussed and illustrated. An application of the results of Chapter 3 then yields local and semi-local convergence theorems for the associated processes.

Chapter 3

Convergence Domains for Isotone Difference Equations

In this chapter we consider the scalar difference equation

$$(3.1) \qquad a_{n+1} = \phi(n,a_n,a_{n-1},\ldots,a_0,a_{-1},\ldots,a_{-m+1}),\ n=0,1,\ldots,$$

where $\phi(n,\ldots):D_n \subseteq [0,\infty)^{n+m} \to [0,\infty)$ is isotone and the sets D_n are products of $(n+m)$ intervals of the form $[0,a)$, $[0,a]$, or $[0,\infty)$.

As usual, we say that a solution $\{a_n\}$ of (3.1) exists if $(a_n,\ldots,a_0,\ldots,a_{-m+1})^T \in D_n$ for all n and (3.1) is satisfied. Let $\mathcal{D}_E^0$ denote the set of all initial vectors $\underset{\sim}{a}^0 = (a_0,a_{-1},\ldots,a_{-m+1})^T \in D_0$ for which the solution $\{a_n\}$ of (3.1) exists, $\mathcal{D}_C^0$ the set of those $\underset{\sim}{a}^0$ for which $\{a_n\}$ converges to 0, and $\mathcal{D}_S^0$ the set for which $\{a_n\}$ is summable. Clearly, $\mathcal{D}_S^0 \subseteq \mathcal{D}_C^0 \subseteq \mathcal{D}_E^0$.

The purpose of this chapter is to find sufficient conditions for $\underset{\sim}{a}^0$ to be contained in (1) $\mathcal{D}_E^0$, (2) $\mathcal{D}_C^0$, or (3) $\mathcal{D}_S^0$. With (2) we will include the estimation of convergence rates which in turn leads to results for (3). Anticipating the applications of Chapter 4, we will at times use special forms of (3.1).

The first simple lemma treats (1) and gives information about the structure of $\mathcal{D}_C^0$, $\mathcal{D}_S^0$. Here and throughout the chapter $\underset{\sim}{a} \leq \underset{\sim}{b}$ for $\underset{\sim}{a},\underset{\sim}{b} \in \mathbb{R}^p$ refers to the natural (componentwise) partial ordering of $\mathbb{R}^p$.

Lemma 3.1. If $\underset{\sim}{a}^0 \in \mathcal{D}_E^0$ (or $\mathcal{D}_C^0$, or $\mathcal{D}_S^0$) and $\underset{\sim}{0} \leq \underset{\sim}{b}^0 \leq \underset{\sim}{a}^0$, then $\underset{\sim}{b} \in \mathcal{D}_E^0$ (or $\mathcal{D}_C^0$, or $\mathcal{D}_S^0$). In fact, if $\{b_n\}$, $\{a_n\}$ are the respective solutions with initial data $\underset{\sim}{b}^0 \leq \underset{\sim}{a}^0$, then $0 \leq b_n \leq a_n$ for all n.

The proof follows immediately from an induction argument with the inequalities

$$0 \leq b_1 = \phi(0,b_0,\ldots,b_{-m+1}) \leq \phi(0,a_0,\ldots,a_{-m+1}) = a_1 ,$$

$$0 \leq b_{n+1} = \phi(n,b_n,\ldots,b_{n-m+1}) \leq \phi(n,a_n,\ldots,a_{n-m+1}) = a_{n+1}$$

We define a set $\mathcal{D}$ in $\mathbb{R}^p$ to be order-convex (O-convex) if whenever $\underset{\sim}{x},\underset{\sim}{y} \in \mathcal{D}$ and $\underset{\sim}{x} \leq \underset{\sim}{y}$, then the order interval $\{\underset{\sim}{z} \in \mathbb{R}^p | \underset{\sim}{x} \leq \underset{\sim}{z} \leq \underset{\sim}{y}\}$ is contained in $\mathcal{D}$. The previous lemma states that if D_0 is O-convex, then each of the sets $\mathcal{D}_E^0$, $\mathcal{D}_C^0$, $\mathcal{D}_S^0$ is either empty or it is O-convex and contains $\underset{\sim}{0}$.

An important property of a nonempty, O-convex set $\mathcal{D} \subseteq \mathbb{R}^m$ with $\underset{\sim}{0} \in \mathcal{D}$ is that its intersection with any one of the 2^m coordinate subspaces ($\underset{\sim}{0}$,span $(\underset{\sim}{e}_1),\ldots,$span $(\underset{\sim}{e}_m)$, span $(\underset{\sim}{e}_1,\underset{\sim}{e}_2),\ldots,\mathbb{R}^m$), called a face of $\mathcal{D}$, is again O-convex and contains $\underset{\sim}{0}$. A point $\underset{\sim}{a} \in \mathcal{D}$ will be called an inner point of $\mathcal{D}$ if it is an interior point of any one of the faces of $\mathcal{D}$ in the relative topology of the intersection of $C^+ = \{\underset{\sim}{z} \in \mathbb{R}^m | \underset{\sim}{z} \geq 0\}$ with the subspace which contains the face. Thus, when $\mathcal{D} \neq \emptyset$, $\underset{\sim}{0}$ is an inner point. It is clear then that $\underset{\sim}{a} = (a_1,\ldots,a_m)^T$ is an inner point of $\mathcal{D}$ if and only if there is a point $\underset{\sim}{b} = (b_1,\dots,b_m)^T \in \mathcal{D}$ such that $a_i < b_i$ if $a_i \neq 0$ and $b_j = 0$ if $a_j = 0$. Of particular interest is the following result which will be used several times in this chapter.

<u>Lemma 3.2</u>. Let $\mathcal{D}$ be an O-convex set in $C^+ \subseteq \mathbb{R}^m$ with $\underset{\sim}{0} \in \mathcal{D}$ and $\{a_n\}$ a sequence of nonnegative numbers which has limit 0. If $\underset{\sim}{a}^n = (a_n,\dots,a_{n-m+1})^T \in \mathcal{D}$ for all n, then there is an integer $N \geq 0$ for which $\underset{\sim}{a}^N$ is an inner point of $\mathcal{D}$.

<u>Proof</u>: To each $\underset{\sim}{a}^n$ corresponds the face of $\mathcal{D}$ of smallest dimension to which it belongs. Since $\underset{\sim}{a}^n$ converges to $\underset{\sim}{0}$ and since there are only finitely many faces, either $\underset{\sim}{a}^N = \underset{\sim}{0}$ for some N or else there are infinitely many of the $\underset{\sim}{a}^n$, say $\underset{\sim}{a}^{n_\nu}$, $\nu = 1,2,\dots,$ in one face which are nonzero and which converge to $\underset{\sim}{0}$. In the first case we are finished since $\underset{\sim}{0}$ is an inner point. In the second, each of the $\underset{\sim}{a}^{n_\nu}$ has the same zero components and no others. Hence $\underset{\sim}{a}^{n_\nu}$ eventually must be in the interior of the face.

For the next several results, we restrict ourselves to the m-th order special form of (3.1)

$$a_{n+1} = \phi(n,a_n,\dots,a_{n-m+1}),\ n=0,1,\dots, \tag{3.2a}$$

and its stationary case

$$a_{n+1} = \phi(a_n,\dots,a_{n-m+1}),\ n=0,1,\dots \tag{3.2b}$$

The following lemma and corollary deal with the question of when the initial vector $\underset{\sim}{a}^0 = (a_0,\dots,a_{-m+1})^T$ is in $\mathcal{D}^0_E$ or $\mathcal{D}^0_C$. The combination of the two is a natural and minor extension of Kantorovich's lemma (see Rheinboldt [1968]) which appears to be new in this form.

Lemma 3.3. For $\rho > 0$, suppose $\phi(n,\ldots):[0,\rho]^m \to [0,\infty)$ is isotone and that $\phi(\cdot,a,b,\ldots,c)$ is antitone in n, n=0,1,..., for each $(a,b,\ldots,c)^T \in [0,\rho]^m$. If the initial values $\underset{\sim}{a}^0 = (a_0,\ldots,a_{-m+1})$ satisfy $\rho \geq a_{-m+1} \geq \ldots \geq a_{-1} \geq a_0 \geq 0$, as well as $0 \leq a_1 \leq a_0$, then $\underset{\sim}{a}^0 \in \mathcal{D}_E^0$ and the solution $\{a_n\}$ is nonincreasing and converges to some $b^* \geq 0$. Thus, if $b^* = 0$, then $\underset{\sim}{a}^0 \in \mathcal{D}_C^0$.

Proof: Again we use an induction argument with the inequalities

$$0 \leq a_2 = \phi(1,a_1,a_0,\ldots,a_{-m+2}) \leq \phi(0,a_0,a_{-1},\ldots,a_{-m+1}) = a_1 \leq \rho ,$$

$$0 \leq a_{n+1} = \phi(n,a_n,\ldots,a_{n-m+1}) \leq \phi(n-1,a_{n-1},\ldots,a_{n-m}) = a_n \leq \rho .$$

Corollary 3.4. If $\underset{\sim}{a}^0$ and ϕ are as in Lemma 3.3 with ϕ continuous and of the form (3.2b), then the solution $\{a_n\}$ of (3.2) converges to the largest number $a^* \in [0,a_1]$ for which $\phi(a^*,\ldots,a^*) = a^*$. If $a^* = 0$, then $\underset{\sim}{a}^0 \in \mathcal{D}_C^0$.

For the proof, the induction now becomes

$$a^* = \phi(a^*,\ldots,a^*) \leq \phi(a_n,\ldots,a_{n-m+1}) = a_{n+1}$$

so that $a^* \leq b^* = \lim_{n\to\infty} a_n \leq a_1$. Since $b^* = \phi(b^*,\ldots,b^*)$ by continuity, we have $b^* \leq a^* \leq a_1$ which yields $b^* = a^*$.

Two examples of the type to be encountered in Chapter 4 are

$$a_{n+1} = \frac{1}{2} a_n^2 + \frac{3}{4} a_{n-1}, n=0,1,\ldots , \tag{3.3}$$

$$(3.4) \qquad a_{n+1} = \frac{1}{2}\left(\frac{a_n}{(1-a_n)^2} \right), \; n=0,1,\ldots \quad .$$

For (3.3), if $0 \leq a_i \leq \frac{1}{2}$, $i = -1,0$, then $a_1 \leq a_0$. An application of Corollary 3.4 and Lemma 3.2 gives $\lim_{n\to\infty} a_n = 0$, i.e., $\underset{\sim}{a}^0 = (a_0,a_{-1})^T \in \mathcal{D}_C^0$. The function ϕ of (3.4) is increasing on $(0, 1-\frac{\sqrt{2}}{2})$ with the endpoints as fixed points. Since $a_1 = \phi(a_0) < a_0$ for $a_0 \in (0,1-\frac{\sqrt{2}}{2})$, a_n converges to 0 monotonically.

We remark at this time that a result similar to Corollary 3.4 holds for $\underset{\sim}{a}^0$ satisfying $0 \leq a_{-m+1} \leq \cdots \leq a_0 \leq a_1 \leq \rho$ provided that $\phi(\rho,\ldots,\rho) \leq \rho$. Then $\{a_n\}$ is nondecreasing and converges to the smallest $a^* \in [a_1,\rho]$ for which $\phi(a^*,\ldots,a^*) = a^*$.

We now turn to a special form of (3.2), namely,

$$(3.5) \qquad a_{n+1} = \phi(n,a_n,\ldots,a_{n-m+1}) = \sum_{i=1}^{d} \phi_i(n,a_n,\ldots,a_{n-m+1}), \; n=0,1,\ldots \; .$$

Here d,m are positive integers and $\phi_i(n,\ldots):D_n \subseteq [0,\infty)^m \to [0,\infty)$ is isotone for $i=1,2,\ldots,d$, $n=0,1,\ldots$. The sets D_n are products of m intervals of form $[0,a]$, $[0,a)$, $[0,\infty)$ and are assumed to be 0-convex. In order to distinguish between domains $\mathcal{D}_E^0$, $\mathcal{D}_C^0$, $\mathcal{D}_S^0$ belonging to ϕ or ϕ_i, we use the notation $\mathcal{D}_E^0(\phi_i)$, etc. In the following four results, rate of convergence estimates will be obtained for solutions $\{a_n\}$ of (3.5) with $\underset{\sim}{a}^0$ already known to be in $\mathcal{D}_C^0(\phi)$. The idea here is to use information about the related equations

(3.6a) $$b_{n+1} = \phi_i(n, b_n, \ldots, b_{n-m+1}),\ n=0,1,\ldots,$$

(3.6b) $$\underset{\sim}{b}^0 = (b_0, \ldots, b_{-m+1})^T = \underset{\sim}{a}^0$$

The first result gives an upper bound for the speed of convergence of $\{a_n\}$.

Theorem 3.4. Suppose $\underset{\sim}{a}^0 \in \mathcal{D}_E^0(\phi)$. Then $\underset{\sim}{b}^0 \in \mathcal{D}_E^0(\phi_i)$ and $0 \leq b_n \leq a_n$. If $\underset{\sim}{a}^0 \in \mathcal{D}_C^0(\phi)$, then $R_p\{a_n\} \geq R_p\{b_n\}$ for $p \in [1,\infty)$.

Proof: A simple induction argument is used with

$$b_1 = \phi_i(0, b_0, \ldots, b_{-m+1}) \leq \sum_{i=1}^{d} \phi_i(0, a_0, \ldots, a_{-m+1}) = a_1 ,$$

$$b_{n+1} = \phi_i(n, b_n, \ldots, b_{n-m+1}) \leq \sum_{i=1}^{d} \phi_i(n, a_n, \ldots, a_{n-m+1}) = a_{n+1}$$

We next determine lower bounds on the speed of convergence of the solution $\{a_n\}$ of (3.5). Again the component functions ϕ_i of (3.6) are the means for this. In Theorem 3.5 a condition on each ϕ_i, which is sufficient for R-linear convergence of each solution $\{b_n\}$ of (3.6) with $\underset{\sim}{b}^0 \in \mathcal{D}_C^0$, provides for linear convergence of the solution $\{a_n\}$ of (3.5) with $\underset{\sim}{a}^0 \in \mathcal{D}_C^0$. A similar approach is used in Theorem 3.6 and its corollary for the more difficult R-superlinear case.

Theorem 3.5. In (3.5) suppose that to each ϕ_i there corresponds some $\lambda_i \in (0,1)$ such that whenever $0 \leq a \leq \rho$, for some $\rho > 0$, we have

$$(3.7) \qquad \phi_i(n, a\lambda_i^{m-1}, a\lambda_i^{m-2}, \ldots, a\lambda_i^0) \leqslant a\lambda_i^m, \; i=1,\ldots,d; \; n=0,1,\ldots$$

where $\sum_{i=1}^{d} \lambda_i \equiv \lambda < 1$. Then for any $\underset{\sim}{a}^0 \in \mathcal{B}_C^0(\phi)$, we have

$$(3.8) \qquad R_1\{a_n\} \leqslant \lambda < 1 ,$$

and then

$$(3.9) \qquad R_1(\phi,0) \leqslant \lambda < 1 .$$

Proof: Let $\hat{\lambda} = \min_{1\leqslant i\leqslant d} \lambda_i$. Since $\lim_{n\to\infty} a_n = 0$, we can choose an integer $n_0 \geqslant 0$ so that $a_n \leqslant \rho\hat{\lambda}^{m-1}$ for all $n \geqslant n_0$. Then with $b = \rho(\frac{\hat{\lambda}}{\lambda})^{m-1}$ and for $j = 0,1,\ldots,m-1$, we have

$$a_{n_0+j} \leqslant \rho\hat{\lambda}^{m-1} \leqslant \rho \left(\frac{\hat{\lambda}}{\lambda}\right)^{m-1}\lambda^{m-1} \leqslant b\lambda^j .$$

For the induction argument assume that $a_{n_0+k} \leqslant b\lambda^k$ for $k = 0,1,\ldots,\ell$ where $\ell \geqslant m-1$. Then

$$a_{n_0+\ell+1} = \sum_{i=1}^{d} \phi_i(n_0+\ell, a_{n_0+\ell}, \ldots, a_{n_0+\ell-m+1})$$

$$\leqslant \sum_{i=1}^{d} \phi_i\left(n_0+\ell, b\frac{\lambda^\ell}{\lambda_i^{m-1}}\lambda_i^{m-1}, \ldots, b\frac{\lambda^{\ell-m+2}}{\lambda_i}\lambda_i, b\frac{\lambda^{\ell-m+1}}{\lambda_i^0}\lambda_i^0\right)$$

Now

$$\frac{\lambda^{\ell-k}}{\lambda_i^{m-1-k}} = \frac{\lambda^\ell}{\lambda_i^{m-1}} \left(\frac{\lambda_i}{\lambda}\right)^k \leqslant \frac{\lambda^\ell}{\lambda_i^{m-1}} \text{ for } k=0,1,\ldots,m-1,$$

and $b\frac{\lambda^\ell}{\lambda_i^{m-1}} \leqslant \rho$ for $\ell \geqslant m-1$. We then obtain from (3.7) with $a = b\left(\frac{\lambda^\ell}{\lambda_i^{m-1}}\right)$

$$a_{n_0+\ell+1} \leq \sum_{i=1}^{d} b \frac{\lambda^\ell}{\lambda_i^{m-1}} \lambda_i^m = b\lambda^\ell \sum_{i=1}^{d} \lambda_i = b\lambda^{\ell+1}$$

Hence, $a_{n_0+n} \leq b\lambda^n$ for all $n \geq 0$ and so

$$\limsup_{n\to\infty} a_n^{1/n} = \limsup_{n\to\infty} (a_{n_0+n})^{\frac{1}{n_0+n}} \leq \limsup_{n\to\infty} (b\lambda^n)^{\frac{1}{n_0+n}} = \lambda$$

For the superlinear case there is a difficulty which must be overcome. As an illustration, consider the equation

$$a_{n+1} = \frac{(n+1)^2}{(n+2)} a_n^2 \ .$$

One would expect that for $a_0 \in \mathcal{S}_C^0$, $R_2\{a_n\} < 1$. However, if $a_0 = 1$, then $a_n = \frac{1}{n+1}$ and the convergence is only sublinear. The problem is that solutions which do not begin at an inner point of $\mathcal{S}_C^0$ may behave quite differently from those which do. This should explain the hypotheses of the following theorem, which we will discuss further later on.

Theorem 3.6. For the functions ϕ_i of (3.5) suppose there is a number $p > 1$ such that for any $\lambda \in (0,1)$ and any $(s_{m-1},\ldots,s_0)^T \in D_n$ we have

$$\phi_i(n,s_{m-1}\lambda^{p^{m-1}},\ldots,s_0\lambda^{p^0}) \leq \phi_i(n,s_{m-1},\ldots,s_0)\lambda^{p^m} \tag{3.10}$$

$$i=1,\ldots,d;\ n=0,1,\ldots$$

If $\underset{\sim}{a}^0$ is an inner point of $\mathcal{S}_C^0$, then the corresponding solution $\{a_n\}$ of (3.5) satisfies

$$(3.11) \qquad R_p\{a_n\} < 1$$

If all solutions of $\{a_n\}$ are uniformly bounded, then for $\underset{\sim}{a}^0$ to be in $\mathcal{D}_C^0$ and for (3.11) to hold it is sufficient that $\underset{\sim}{a}^0$ be an inner point of $\mathcal{D}_E^0$.

Proof: In either case there is an initial vector $\hat{\underset{\sim}{a}}^0 = (\hat{a}_0,\ldots,\hat{a}_{-m+1})^T$ such that the corresponding solution $\{\hat{a}_n\}$ is bounded and satisfies $\limsup_{n\to\infty} \hat{a}_n^{1/p^n} \leq 1$ and such that $a_j < \hat{a}_j$ whenever $a_j \neq 0$, $j=-m+1,\ldots,0$. Hence we can choose $\lambda \in (0,1)$ so that

$$a_j \leq \hat{a}_j \lambda^{p^{j+m-1}}$$

for $j = -m+1,\ldots,0$. To use induction, assume that the above inequality holds for $j = -m+1,\ldots,n$, and compute

$$\begin{aligned} a_{n+1} &\leq \sum_{i=1}^{d} \phi_i(n,\hat{a}_n\lambda^{p^{n+m-1}},\ldots,\hat{a}_{n-m+1}\lambda^{p^n}) \\ &= \sum_{i=1}^{d} \phi_i(n,\hat{a}_n(\lambda^{p^n})^{p^{m-1}},\ldots,\hat{a}_{n-m+1}(\lambda^{p^n})^{p^0}) \\ &\leq \sum_{i=1}^{d} \phi_i(n,\hat{a}_n,\ldots,\hat{a}_{n-m+1})(\lambda^{p^n})^{p^m} = \hat{a}_{n+1}\lambda^{p^{n+m}} \end{aligned}$$

Thus, for all integers $n \geq 0$, $a_n \leq \hat{a}_n\lambda^{p^{n+m-1}}$, and

$$\begin{aligned} \limsup_{n\to\infty} a_n^{1/p^n} &\leq \left(\limsup_{n\to\infty} (\hat{a}_n)^{1/p^n}\right)\left(\limsup_{n\to\infty} \lambda^{p^{n+m-1-n}}\right) \\ &\leq 1\cdot \limsup_{n\to\infty} \lambda^{p^{m-1}} \leq \lambda < 1 . \end{aligned}$$

For the mentioned example $a_{n+1} = \frac{(n+1)^2}{n+2} a_n^2$, $n = 0,1,\ldots$, we see that the conditions of the theorem are satisfied for $p = 2$ but that $a_0 = 1$ is not an inner point of $\mathcal{S}_C^0$. Notice also that here and in the theorem we are unable to conclude anything about $O_R(\phi,0)$ because we do not know the behavior of all solutions with $\underset{\sim}{a}^0 \in \mathcal{S}_C^0$.

Nevertheless, the following corollary shows that for the stationary form of (3.5) the conclusions of the theorem hold for any $\underset{\sim}{a}^0 \in \mathcal{S}_C^0$.

<u>Corollary 3.7</u>. If ϕ of (3.5) satisfies the conditions of Theorem 3.6 and if ϕ is stationary, then

$$O_R(\phi,0) \geq p\ . \tag{3.12}$$

<u>Proof</u>: Lemma 3.2 implies that if $\underset{\sim}{a}^0 \in \mathcal{S}_C^0$, then eventually $\underset{\sim}{a}^n = (a_n,\ldots,a_{n-m+1})^T$ is an inner point of $\mathcal{S}_C^0$, say for $n = N \geq 0$. Since ϕ is stationary, a reindexing in which a_n is replaced by a_{n-N} does not affect the proof of Theorem 3.6. The conclusion is that $R_p\{a_n\} < 1$ for any solution $\{a_n\}$ which converges to 0. Theorem 11 yields (3.12).

This reasoning is not valid in the nonstationary case since there it is not sufficient that $\underset{\sim}{a}^n$ be an inner point of $\mathcal{S}_C^0$ for some n. Rather, for some n, $\underset{\sim}{a}^n$ must be an inner point of $\mathcal{S}_C^n$, the set of all initial vectors $\underset{\sim}{b}^n$ for which the solution $\{b_k\}$ of

$$b_{k+1} = \phi(k,b_k,\ldots,b_{k-m+1}),\ k = n,n+1,\ldots \tag{3.13}$$

converges to 0. In the stationary case, $\mathcal{S}_C^n \equiv \mathcal{S}_C^0$ for all n, but otherwise $\mathcal{D}_C^n$ may vary and could, for example, shrink to $\{\underset{\sim}{0}\}$ as n increases. Thus the mere assumption that $\underset{\sim}{a}^n$ converges to $\underset{\sim}{0}$ does not guarantee that $\underset{\sim}{a}^N$ is an inner point of $\mathcal{D}_C^N$ for any N. In the example above with $a_0 = \underset{\sim}{a}^0 = 1$, for no N is $\underset{\sim}{a}^N = \frac{1}{N+1}$ an inner point of $\mathcal{D}_C^N$.

There are of course sufficient conditions on the domains $\mathcal{D}_C^n$ under which $\underset{\sim}{a}^n$ is an inner point of $\mathcal{D}_C^n$ for some n whenever $\{a_n\}$ is convergent to 0, and we mention two of them. First, if for some $N_0 \geq 0$

$$(3.14a) \qquad D_C^n \subseteq D_C^{n+1}, \text{ for all } n \geq N_0$$

then by Lemma 3.2, $\underset{\sim}{a}^{N_1}$ is an inner point of $\mathcal{D}_C^{N_0} \subseteq \mathcal{D}_C^{N_1}$ for some $N_1 \geq N_0$. A second sufficient condition is that for some $N_0 \geq 0$, there is a vector $\underset{\sim}{a} \in (0,\infty)^m$ such that

$$(3.14b) \qquad a \in \bigcap_{k=N_0}^{\infty} \mathcal{D}_C^k$$

In this case, when $\lim_{n\to\infty} a_n = 0$, there must be some $N_1 \geq N_0$ such that $\underset{\sim}{a}^{N_1} \leq \underset{\sim}{a}$ with no components equal, and $\underset{\sim}{a}^{N_1}$ is an inner point of $\mathcal{D}_C^{N_1}$.

We should mention here that Theorems 3.4, 3.5, 3.6 and the corollary can be significantly generalized with only minor changes in the proofs. For example, in place of $\sum\phi_i$ we could use $\mathcal{F}(\phi_1,\ldots,\phi_d)$ where $\mathcal{F}$ is an isotone, q-homogeneous functional on $[0,\infty)^d$; or instead of λ^k, λ^{p^k} we might consider λ^{γ_k} where $\{\gamma_k\}$ is a nonnegative, increasing, unbounded sequence.

Theorems 3.4 through Corollary 3.7 appear to be new results. For the special case

$$\phi_i(n,x,y,\ldots,z) = L_i\, x^{\alpha_i} y^{\beta_i} \ldots z^{\gamma_i},\ i=1,2,\ldots,d;\ n=0,1,\ldots,$$

where $\alpha_i, \beta_i, \ldots, \gamma_i, L_i$ are nonnegative, $\delta_i = \alpha_i + \beta_i + \ldots + \gamma_i \geq 1$, Schmidt [1968] has proved Theorem 3.5 and stated without proof Corollary 3.7. In the former, the assumptions are that $\delta_i \geq 1$ and $\sum_{i=1}^{d} L_i = L < 1$ with $L = \lambda$. In the latter, δ_i must be greater than one, and p is the smallest of the dominant positive roots of the polynomials $x^m - \alpha_i x^{m-1} - \ldots - \gamma_i = 0$, $i=1,\ldots,d$. Contained here are the conclusions of Traub [1964] for equations of the form

$$a_{n+1} = a_n \sum_{j=0}^{m} c_j a_{n-j},\ n=0,1,\ldots,\quad c_i \text{ constant}.$$

An example of an application of Theorem 3.6 which is more general than these will be encountered in Chapter 4 and has the form

$$\phi_i(n,x,y,\ldots,z) = 2^n L_i\, x^{\alpha_i} y^{\beta_i} \ldots z^{\gamma_i},\ i=1,\ldots,d;\ n=0,1,\ldots,$$

with $\delta_i > 1$ for all i. In that case, p is the same as above.

We now turn to the problem of finding sufficient conditions for $\underset{\sim}{a}^0$ to be contained in $\mathcal{D}_S^0$. Although Theorems 3.5 and 3.6 provide some, we will often be interested in estimating such quantities as $\sum_{j=0}^{n} a_j$ and $\sum_{j=n}^{\infty} a_j$. In certain cases, the difference equation possesses a structure which allows a more explicit statement about $\mathcal{D}_S^0$.

Let us consider equations of the form

$$a_{n+1} =$$

(3.15a)
$$[\psi(a_n, \sum_{j=0}^{n} a_j, a_{n-1}, \sum_{j=0}^{n-1} a_j, \ldots, a_{n-m+1}, \sum_{j=0}^{n-m+1} a_j)]a_{n-d}, \; n=0,1,\ldots$$

(3.15b)
$$\underset{\sim}{a}^0 = (a_0, a_{-1}, \ldots, a_{-m+1}) \geq \underset{\sim}{0}, \; \sum_{j=0}^{-s} a_j \equiv a_0 \text{ for } s=0,1,\ldots,m-1 ,$$

where $\psi:[0,\rho]^{2m} \to [0,\infty)$ is isotone, $\rho > 0$, and d is a fixed integer in $[0,m-1]$. However, we can better appreciate the approach by first discussing the special case of equation (3.15) with $d = 0$, $m = 1$,

(3.15)'
$$a_{n+1} = [\psi(a_n, \sum_{j=0}^{n} a_j)] \, a_n, \; n = 0,1,\ldots \; .$$

We want to find the largest number $A > 0$ such that for $a_0 \in [0,A]$ or $[0,A)$ we have $\sum_{j=0}^{\infty} a_j < \infty$. Let $\psi(a_n, \sum_{j=0}^{n} a_j) \equiv \beta_n$ for $n = 0,1,\ldots,$ so that

(3.16)
$$a_{n+1} = \beta_n a_n = (\prod_{j=0}^{n} \beta_j) a_0 \text{ and } \sum_{k=0}^{n} a_k = \sum_{k=0}^{n} (\prod_{j=0}^{k} \beta_j) a_0 \; .$$

Under the not unreasonable assumption that all $\beta_j \neq 0$, we know from the result of Hahn [see Lemma 1.16] that $\sum_{k=0}^{n} (\prod_{j=0}^{k} \beta_j)$ is bounded if and only if there are real numbers $a \geq 1$, $0 < \lambda < 1$ such that

(3.17)
$$\prod_{j=k}^{k+\ell} \beta_j \leq a\lambda^{\ell+1} \text{ for all } k,\ell \geq 0 \quad .$$

We conclude that $\sum_{j=0}^{n} a_j$ is bounded if and only if there exists a,λ such that

$$\prod_{j=k}^{k+\ell} \psi(a_j, \sum_{i=0}^{j} a_i) \leq a\lambda^{\ell+1} \text{ for all } k,\ell \geq 0 \tag{3.18}$$

In such an event, we have $a_{n+1} = (\prod_{j=0}^{n} \beta_j)a_0 \leq a\lambda^{n+1}a_0$ and

$$\sum_{k=0}^{n} a_k =$$

$$a_0 + \sum_{k=1}^{n} (\prod_{j=0}^{k-1} \beta_j)a_0 \leq a_0(1+a\sum_{k=1}^{n} \lambda^k) \leq a_0(1+a(\frac{\lambda}{1-\lambda})(1-\lambda^n)),$$

for $n = 0,1,\ldots$. This yields a sufficient condition for a_0 to be in $\mathcal{D}_S^0$, namely that there exist $a \geq 1$, $0 < \lambda < 1$ with

$$\prod_{j=k}^{k+\ell} \psi(a\lambda^j a_0, a_0(1+\frac{a\lambda}{1-\lambda}(1-\lambda^j))) \leq a\lambda^{\ell+1} \text{ for all } k,\ell \geq 0 \ . \tag{3.19}$$

Since (3.19) is difficult to verify, we give the alternate, simpler condition

$$\psi(a_0, \frac{a_0}{1-\lambda}) \leq \lambda < 1 \tag{3.20}$$

which implies (3.19) with $a = 1$, i.e.,

$$\prod_{j=k}^{k+\ell} \psi(\lambda^j a_0, a_0(1+\frac{\lambda}{1-\lambda}(1-\lambda^j))) \leq \prod_{j=k}^{k+\ell} \psi(a_0, \frac{a_0}{1-\lambda}) \leq \lambda^{\ell+1}$$

The main idea here is that since ψ is isotone, we can find a_0 so small that $a_{n+1} \leq \lambda a_n$ for all n and some $\lambda \in (0,1)$. Thus, there is a certain contractive relation between a_{n+1} and a_n which in turn guarantees the summability of $\{a_n\}$. In fact, $\sum a_j$ is compared with the geometric series $\sum \lambda^j$. For the more general equation (3.15), we extend this approach, using a condition corresponding to (3.20).

<u>Theorem 3.8</u>. For ψ of (3.15) suppose there is a $\lambda \in (0,1)$ and an initial vector $\underset{\sim}{a}^0$ such that

$$0 \leq a_j \leq a_0 \leq \rho(1-\lambda) < \rho \quad \text{for } j = -m+1,\ldots,-1 \tag{3.21a}$$

$$\psi\left(a_0, \frac{a_0}{1-\lambda}, a_0, \frac{a_0}{1-\lambda}, \ldots, a_0, \frac{a_0}{1-\lambda}\right) \leq \lambda^{d+1} \tag{3.21b}$$

Then $\underset{\sim}{a}^0 \in \mathcal{D}_S^0$ and, in fact, for $n = 0,1,\ldots,$ we have

$$0 \leq a_n \leq a_0\lambda^n \leq \rho \ , \tag{3.22a}$$

$$\sum_{j=0}^{n} a_j \leq \frac{a_0}{1-\lambda} \ , \tag{3.22b}$$

$$\sum_{j=n}^{\infty} a_j \leq \frac{\lambda^n}{1-\lambda} \ a_0 \tag{3.22c}$$

<u>Proof</u>: The induction proposition is

$$a_j \leq \lambda^{d+1} a_{j-(d+1)} \ , \quad j = 1,2,\ldots \ . \tag{3.23}$$

For $j = 1$, $a_1 = \psi(a_0,a_0,a_{-1},a_0,\ldots,a_{-m+1},a_0)a_{-d} \leq \lambda^{d+1}a_{-d}$. Assume that (3.23) holds for $j = 1,2,\ldots,n$. Then

$$a_j \leq \lambda^{d+1}a_{j-(d+1)} \leq (\lambda^{d+1})^2 a_{j-2(d+1)} \leq \cdots \leq (\lambda^{d+1})^s a_{j-s(d+1)},$$

$$j=1,2,\ldots,n,$$

where $-m+1 \leq j-s(d+1) \leq 0$ or $\frac{j}{d+1} \leq s \leq \frac{j+m-1}{d+1}$. Hence, we have

(3.24a) $$a_j \leq \lambda^j a_0 \quad \text{for } j = 1,2,\ldots,n \ ,$$

(3.24b) $$\sum_{j=0}^{k} a_j \leq a_0 \sum_{j=0}^{k} \lambda^j \quad \text{for } k = 0,1,\ldots,n \ .$$

Finally,

$$a_{n+1} = \psi(a_n, \sum_{j=0}^{n} a_j, \ldots, a_{n-m+1}, \sum_{j=0}^{n-m+1} a_j) a_{n-d}$$

$$\leq \psi(a_0, \frac{a_0}{1-\lambda}, \ldots, a_0, \frac{a_0}{1-\lambda}) a_{n-d} \leq \lambda^{d+1} a_{n-d} \ ,$$

and so (3.23) holds for $j = 1,2,\ldots$ and (3.22) follows from (3.24).

For an example of the application of this theorem, consider

(3.25) $$a_{n+1} = \frac{Aa_n^2 + Ba_n a_{n-1}}{1-B(\sum_{j=0}^{n} a_j + \sum_{j=0}^{n-1} a_j)} \ , \quad n = 0,1,\ldots \ ,$$

where $A,B,C > 0$ and we require $B(\sum_{j=0}^{n} a_j + \sum_{j=0}^{n-1} a_j) < 1$. We wish to find a_0,λ,ρ such that

(3.26a) $$\frac{a_0}{1-\lambda} \leq \rho < \frac{1}{2B}$$

(3.26b) $$\frac{(A+B)a_0}{1-2B(\frac{a_0}{1-\lambda})} \leq \lambda \ .$$

Since we are interested in finding the largest $a_0 \in \mathcal{D}_S^0$, we rewrite (3.26) as

$$a_0 \leq \rho(1-\lambda) < \frac{1-\lambda}{2B} \tag{3.27a}$$

$$a_0 \leq \frac{\lambda(1-\lambda)}{(A+B)+\lambda(B-A)} \tag{3.27b}$$

By maximizing the right-hand sides of (3.27) for $\lambda \in (0,1)$, we obtain the best $a_0 \in \mathcal{S}_S^0$ available from the theorem. In any event

$$a_{n+1} \leq \frac{Aa_n^2 + Ba_n a_{n-1}}{1-2\rho B}, \quad n = 1,2,\ldots$$

and Corollary 3.7 gives $O_R(\phi,0) \geq p = \frac{1+\sqrt{5}}{2}$

Several remarks are in order here. First of all, the proof of Theorem 3.8 can be adapted for use on equations of the form (3.15) with functions ψ depending on additional quantities such as $\sum_{j=2}^{n} a_j$, etc. Secondly, although the theorem gives sufficient conditions for $\underset{\sim}{a}^0$ to be contained in $\mathcal{S}_S^0$, a much more valuable piece of information is to know the "best" initial points in $\mathcal{S}_S^0$. These points are actually the maximal elements of chains or totally ordered subsets of $\mathcal{S}_S^0$, which exist by Zorn's Lemma when $\mathcal{S}_S^0$ is bounded. The problem of actually finding such points in the general case is unsolved. However, for $m = 1$, $d = 0$, Rheinboldt [1968] has found a device whereby such a point is made available in certain cases. Let us begin with an example and then discuss this approach in general.

The following difference equation arises in the proof of the Newton-Kantorovich theorem (see Rheinboldt [1968]):

$$a_{n+1} = \frac{\frac{1}{2} K a_n^2}{1-K(\sum_{j=0}^{n} a_j)} \quad , \; n = 0,1,\dots ; \; a_0 \geq 0 \; . \tag{3.28}$$

If we apply Theorem 3.8, we require $\frac{a_0}{1-\lambda} < \frac{1}{K}$ and $\frac{\frac{1}{2} K a_0}{1-K(\frac{a_0}{1-\lambda})} < \lambda$ for

some $\lambda \in (0,1)$. The maximum a_0 is obtained for $\lambda = -1+\sqrt{2}$ and is $a_0 = \frac{2(3-2\sqrt{2})}{K} \approx .\frac{36}{K}$. Notice, however, that if we replace a_n by $t_n - t_{n-1}$, a_0 by t_0, and set $t_{-1} = 0$ for $n = 0,1,\dots$, (3.28) becomes

$$t_{n+1} - t_n = \frac{\frac{1}{2} K(t_n - t_{n-1})^2}{1-Kt_n} \quad , \; n=0,1,\dots ; \; t_n < \frac{1}{K} ; \; t_0 > 0 \tag{3.29}$$

Moreover, the sequence $\{t_n\}$ satisfies (3.29) if and only if it also satisfies

$$t_{n+1} = t_n + \frac{\frac{1}{2} Kt_n^2 - t_n + t_0}{1-Kt_n} \equiv \chi(t_n, t_0), \; n=0,1,\dots ; \; t_n < \frac{1}{K} . \tag{3.30}$$

The function χ is called a <u>first integral</u> for (3.29). Since $t_n = \sum_{j=0}^{n} a_j$, we are interested in finding the largest $t_0 (=a_0)$ for which the solution $\{t_n\}$ of (3.30) is bounded. But $\chi(t,t_0)$ is isotone and continuous in t and $\chi(t_{-1},t_0) = \chi(0,t_0) = t_0 > 0$ so that from the remark following Corollary 3.4, t_n converges to the smallest solution $t^*(t_0)$ of $\chi(t^*(t_0),t_0) = t^*(t_0)$ in $[t_0,\infty)$. For $t_0 < \frac{1}{2K}$, $t^*(t_0) = \frac{1-\sqrt{1-2a_0K}}{K}$ and so if $0 \leq a_0 < \frac{1}{2K}$, then $\sum_{j=0}^{\infty} a_j \leq (\frac{1}{K})(1-\sqrt{1-2a_0K})$. The conclusion is that $[0, \frac{1}{2K}) \subseteq \mathcal{D}_S^0$, a significant improvement over the interval $[0, \frac{0.36}{K})$ found by Theorem 3.8. Notice also that if $t_0 = \frac{1}{2K}$, then $t^*(t_0) = \frac{1}{K}$ and so $[0, \frac{1}{2K})$ is the maximal set in $\mathcal{S}_S^{\cup}$ for which $\sum_{n=0} a_j < \frac{1}{K}$, i.e., for which $a_n \geq 0$ for all $n = 0,1,\ldots$.

We summarize the approach of this example for the general equation

(3.31a) $$a_{n+1} = \phi(a_n, \sum_{j=0}^{n} a_j),\ n = 0,1,\ldots ,$$

(3.31b) $$a_0 \geq 0 ,$$

where $\phi:[0,\rho_1] \times [0,\rho_2] \to [0,\infty)$ for some $\rho_2 > \rho_1 > 0$ is isotone. Note that (3.31) is slightly more general than (3.15) with m=1, d=0 and that there is really no contraction principle used here.

Suppose that for each $a_0 \in \mathcal{S}_E^0$, $\{a_n(a_0)\}$ denotes the solution of (3.31) with initial values a_0. Define the new variables $t_n(a_0)$ by

(3.32) $$t_{n+1}(a_0) - t_n(a_0) = a_{n+1}(a_0) \text{ for } n = 0,1,\ldots,$$

$$t_{-1}(a_0) = 0 \quad , \quad t_0(a_0) = a_0$$

Then $\sum_{j=k}^{\ell} a_j(a_0) = t_\ell(a_0) - t_{k+1}(a_0)$ and $\{t_n(a_0)\}$ is an increasing sequence for each a_0. The condition $\lim_{n\to\infty} t_n(a_0) = t^*(a_0) \leq \rho_2$ implies that $a_0 \in \mathcal{S}_S^0$ and in fact $\sum_{j=0}^{\infty} a_j(a_0) = t^*(a_0)$. Now, suppose that there is a function $\mathfrak{X}(t,a_0)$ for each $a_0 \in \mathcal{S}_E^0$ such that $\{t_n(a_0)\}$ solves

(3.33a) $$t_{n+1}(a_0) = \mathfrak{X}(t_n(a_0),a_0),\ n = 0,1,\ldots$$

(3.33b) $$t_{-1}(a_0) = 0$$

if and only if $\{t_n(a_0)\}$ solves the transformed version of (3.31)

$$t_{n+1} - t_n = \phi(t_n - t_{n-1}, t_n), \ n = 0,1,\ldots , \tag{3.34a}$$

$$t_0 = a_0, \ t_{-1} = 0 . \tag{3.34b}$$

In this case, $\mathcal{X}$ is called a <u>first integral</u>. Notice that $\mathcal{X}$ is isotone since ϕ is. If $\mathcal{X}$ is continuous in t for each fixed a_0, and if $t^*(a_0) \neq 0$ is the smallest solution of $\mathcal{X}(t,a_0) = t \leq \rho_2$ (if it exists), then since $\mathcal{X}(0,a_0) = \mathcal{X}(t_{-1}(a_0),a_0) = t_0(a_0)$ $a_0 > 0$, we obtain $\lim_{n\to\infty} t_n(a_0) = t^*(a_0) \leq \rho_2$ from the remark following Corollary 3.4. It follows then that

$$\mathcal{D}_S^0 = \{a_0 \in \mathcal{D}_E^0 \mid t^*(a_0) \leq \rho_2\},$$

and we seek

$$\sup \{a_0 \in \mathcal{D}_E^0 \mid t^*(a_0) \leq \rho_2\}$$

If $\mathcal{X}(t,a)$ is defined only for a in a proper subset $\hat{\mathcal{D}}_E^0$ of $\mathcal{D}_E^0$, then this approach yields the "best" a_0 in $\mathcal{D}_S^0 \cap \hat{\mathcal{D}}_E^0$.

Of course, this idea can be used for equations more general and of higher order than (3.31). However, to find the first integral is often a more difficult problem than that of determining $\mathcal{D}_S^0$.

Chapter 4

Convergence of Iterative Processes: Majorization Approach

This chapter is divided into three sections, the first two of which discuss a majorizing sequence approach to proving local and semi-local convergence results for iterative methods. In the third, the more general majorizing system concept is introduced and applied. The importance of the results of Chapter 3 will become apparent as the solution of a difference equation emerges in the central role of majorant. We stress at this time that the emphasis in this chapter will be on techniques for proving convergence rather than on any particular iterative method or class of methods.

Section 1. Local Convergence, Majorizing Sequences

We will be concerned here with local results for iterative processes. To prove that $x^* \in \mathbb{R}^p$ is a point of attraction for an m-step method $\mathcal{J}$, we must show the existence of m neighborhoods $U_j \subseteq \mathbb{R}^p$ of x^* with the property that whenever the initial vectors $x(i)$ satisfy $x(j) \in U_j$ for each $j = -m+1,\ldots,0$, then the iterates $\{x(n)\}$ generated by $\mathcal{J}$ exist and converge to x^*. The following observation describes the majorization approach of this section which will be used to prove local convergence.

Local Majorant Principle. Let $x^* \in \mathbb{R}^p$ and E_j, $j = -m+1,\ldots,0,1,\ldots,$ be a sequence of neighborhoods of x^* in $\mathbb{R}^p$. Suppose that there is a sequence of isotone functions $\phi(n,\ldots):D_n \subseteq [0,\infty)^m \to [0,\infty)$, $n=0,1,\ldots,$

such that whenever the iterates $x(j)$, $j = -m+1,\ldots,0,\ldots,n$, obtained from an m-step process $\mathcal{J}$ exist and satisfy $x(j) \in E_j$, then the next iterate $x(n+1)$ exists in E_{n+1} and the estimate

$$(4.1) \qquad \|x(n+1)-x^*\| \leq \phi(n,\|x(n)-x^*\|,\ldots,\|x(n-m+1)-x^*\|)$$

holds.

A simple induction argument shows that if $x(j) \in E_j$ for $j = -m+1,\ldots,0$ and if $(a_0,\ldots,a_{-m+1}) \equiv \underset{\sim}{a}^0 \in \mathcal{D}_C^0(\phi)$ for a_j satisfying $\|x(j)-x^*\| \leq a_j$, $j = -m+1,\ldots,0$, then the entire sequence $\{x(n)\}$ of $\mathcal{J}$-iterates exists with $x(j) \in E_j$ and satisfies

$$(4.2) \qquad \|x(j)-x^*\| \leq a_j,\ j = -m+1,\ldots,0,1,\ldots\ ,$$

where

$$(4.3) \qquad a_{n+1} = \phi(n,a_n,\ldots,a_{n-m+1}),\ n = 0,1,\ldots\ .$$

In this case, $\lim_{n\to\infty} \|x(n)-x^*\| \leq \lim_{n\to\infty} a_n = 0$ and thus we may use

$$E_j \cap S(x^*,a_j) \equiv U_j,\ \text{for } j = -m+1,\ldots,0$$

Apparently, $R_p\{x(n)\} \leq R_p\{a_n\}$ for all $p \in [1,\infty)$. To make a statement about $O_R(\mathcal{J},x^*)$, we must have estimates of $R_p\{x(n)\}$ for <u>every</u> sequence of iterates $\{x(n)\}$ which converges to x^*. Generally, there is no reason to expect that every convergent sequence eventually enters the set for which (4.1) is satisfied. However, if $\mathcal{J}$ is stationary, then (4.1) will hold and in that case if $R_p\{x(n)\} < 1$ for some $p \geq 1$ and all convergent sequences, then by Theorem 11,

$$O_R(\mathcal{J},x^*) \geq p$$

We call (4.3) a <u>majorizing difference equation</u> for $\mathcal{J}$ and its solution $\{a_n\}$ a <u>majorizing sequence</u>.

In practice, (4.1) is not easily verified. However, if $\mathcal{J}$ is given by

(4.5) $\quad x(n+1) = G(n,x(n),\ldots,x(n-m+1)),\ n = 0,1,\ldots$

where $G(n,\ldots): \prod_{j=n-m+1}^{n} E_j \to \mathbb{R}^p$, then a sufficient condition is

(4.6) $\quad \|G(n,y_1,\ldots,y_m)-x^*\| \leq \phi(n,\|y_1-x^*\|,\ldots,\|y_m-x^*\|)$

for all $y_i \in E_{n-i}$. A more general condition involves a functional relationship between G and ϕ which we write for the simple case $m = 1$:

(4.7) $\quad \|G(n,G(n-1,x))-x^*\| \leq \phi(n,\|G(n-1,x)-x^*\|),\ n = 0,1,\ldots ,$

for $G(n-1,x) \in E_n$ and $G(-1,x) \equiv x$.

We will now discuss several examples which illustrate the local majorant principle.

<u>Example 4.1</u>. Consider first the general m-step stationary process

(4.5a) $\quad x(n+1) = G(x(n),\ldots,x(n-m+1)),\ n = 0,1,\ldots .$

Assume that for some $x^* \in X$, $\rho > 0$, $G:(\bar{S}(x^*,\rho))^m \subseteq X^m \to X$, and

$$(4.8) \qquad G(x^*,\ldots,x^*) = x^*$$

$$(4.9) \qquad \|G(x,y,\ldots,z)-x^*\| \leq \phi(\|x-x^*\|,\|y-x^*\|,\ldots,\|z-x^*\|)$$

for $x,y\ldots,z$ in $\bar{S}(x^*,\rho)$, where

$$(4.10) \qquad \phi(a,b,\ldots,c) = \sum_{i=1}^{d} L_i a^{\alpha_i} b^{\beta_i} \ldots c^{\gamma_i} ,$$

and $\alpha_i,\beta_i,\ldots,\gamma_i$, $L_i \geq 0$, $\delta_i = \alpha_i + \beta_i + \ldots + \gamma_i \geq 1$.

Evidently, the majorizing equation is

$$(4.11) \qquad a_{n+1} = \phi(a_n,a_{n-1},\ldots,a_{n-m+1}),\ n = 0,1,\ldots$$

which has been discussed in Chapter 3 as an application of Theorem 3.5 and Corollary 3.7. The conclusion is that if either

$$(4.12a) \qquad \delta_i > 1 \text{ for } i = 1,\ldots,d$$

or if

$$(4.12b) \qquad \sum_{i\in I} L_i \leq \Lambda < 1 \text{ where } I \text{ is the set of indices } i \text{ for which } \delta_i = 1 ,$$

then x^* is a point of attraction of (4.5). If (4.12a) holds, then $O_R(\mathcal{J},x^*) \geq p$ where p is the smallest of the dominant positive roots of the polynomials $x^m-\alpha_i x^{m-1}-\ldots-\gamma_i = 0$, $i = 1,2,\ldots,d$. In case (4.12b), $R_1(\mathcal{J},x^*) \leq \Lambda$.

The convergence result has been given by Schmidt [1968]. However, the role of the difference equation was not discussed. The problem, of course, is to determine the function ϕ of (4.9), and this will be done in the next example.

Example 4.2 (Newton and Secant Type Processes). Consider m-step processes of form

(4.13) $$x(n+1) = x(n) - J(n, x(n), \ldots, x(n-m+1))^{-1} Fx(n), \quad n = 0,1,\ldots ,$$

where $F\colon E \subseteq X \to X$ and $J(n,\ldots)\colon E^m \subseteq X^m \to L(X)$, $n = 0,1,\ldots$.
Assume that $x^* \in E$ solves $F(x) = 0$. For the many known results about this method, the usual estimate takes the form

(4.14) $$\begin{aligned} \|x(n+1)-x^*\| \leq {} & \|J(n,x(n),\ldots,x(n-m+1))^{-1}\| \\ & \cdot \|Fx(n)-Fx^*-J(n,x(n),\ldots,x(n-m+1))(x(n)-x^*)\| \end{aligned}$$

and gives rise to the following kinds of assumptions.

(4.15a) F has a Hölder continuous F-derivative on the open and convex set E with constants $L \geq 0$, $0 < \alpha \leq 1$.

(4.15b) $F'(x^*)^{-1}$ exists and is bounded by β.

(4.15c) Whenever $x(j)$, $j = n-m+1,\ldots,n$, exists in E, then

$$\begin{aligned} & \|J(n,x(n),\ldots,x(n-m+1))-F'(x(n))\| \\ & \qquad \leq \gamma(n,\|x(n)-x^*\|,\ldots,\|x(n-m+1)-x^* \end{aligned}$$

where $\gamma(n,\ldots)$ is isotone.

Then, if $x(j)$, $j = n-m+1,\ldots,n$, exists in E and if

$$(4.16)\qquad \beta[\gamma(n,\|x(n)-x^*\|,\ldots,\|x(n-m+1)-x^*\|) + L\|x(n)-x^*\|^{\alpha}] < 1$$

we have, from Banach's Lemma,

$$(4.17)\qquad \|x(n+1)-x^*\| \leq \frac{\beta[\frac{L}{1+\alpha}\|x(n)-x^*\|^{\alpha}+\gamma(n,\|x(n)-x^*\|,\ldots)]}{1-\beta[\gamma(n,\|x(n)-x^*\|,\ldots)+L\|x(n)-x^*\|^{\alpha}]}$$

Hence, the majorizing equation is

$$(4.18)\qquad a_{n+1} = \frac{\beta[\frac{L}{1+\alpha}a_n^{\alpha}+\gamma(n,a_n,\ldots,a_{n-m+1})]a_n}{1-\beta[La_n^{\alpha}+\gamma(n,a_n,\ldots,a_{n-m+1})]}$$

This equation is of the form dealt with in Theorem 3.8 and the conclusion is that if there exists some $a>0$ such that

$$(4.19)\qquad \frac{\beta[\frac{L}{1+\alpha}a^{\alpha}+\gamma(n,a,a,\ldots,a)]}{1-\beta[La^{\alpha}+\gamma(n,a,\ldots,a)]} \leq \lambda < 1,\ n = 0,1,\ldots,$$

then any initial vector $\underset{\sim}{a}^0 = (a_0,\ldots,a_{-m+1}) \leq (a,\ldots,a)$ is in $\mathcal{D}_C^0$ for (4.18). For the process (4.13) under conditions (4.15), (4.19) it follows from the majorant principle that x^* is a point of attraction. The convergence rate is then estimated from that of $\{a_n\}$, applying Theorem 3.5 and Corollary 3.7.

Of course, under alternate conditions different difference equations occur. For instance, if (4.15) is replaced by

(4.15a') F has an F-derivative on the open, convex set E which is continuous with modulus of continuity ω_1, (4.15b') J is independent of n and $J(x^*,\ldots,x^*)^{-1}$ exists and is bounded by β,

(4.15c') For any $y(j) \in E$, $j = 1,2,\ldots,m$,

$$\|J(y(1),\ldots,y(m))-F'(y(1))\|$$

$$\leq \gamma(\|y(1)-x^*\|,\ldots,\|y(m)-x^*\|)$$

where $\gamma\colon [0,\rho]^m \to [0,\infty)$ is isotone, $\rho > 0$,

then in place of (4.18) we obtain

$$(4.18') \qquad a_{n+1} = \frac{\beta[\int_0^1 \omega_1((1-t)a_n)dt+\gamma(a_n,\ldots,a_{n-m+1})]a_n}{1-\beta[\gamma(0,0,0,\ldots,0)+\gamma(a_n,\ldots,a_{n-m+1})+\omega_1(a_n)]}$$

Let us show now how several explicit processes fit into this framework.

For Newton's method, $\gamma \equiv 0$ and under (4.15), (4.18) is

$$a_{n+1} = \frac{\beta[\frac{L}{1+\alpha}\, a_n^{1+\alpha}]}{1-\beta L a_n^{\alpha}}$$

The process is stationary so that $O_R(\mathcal{J},x^*) \geq 1+\alpha$.

With (4.15) one approximate Newton process has $\|J(x)-F'(x)\| \leq \delta_0+\delta_1\|x-x^*\|$ and then

$$a_{n+1} = \frac{\beta[\frac{L}{1+\alpha}\, a_n^{\alpha}+\delta_0+\delta_1 a_n]a_n}{1-\beta[La_n^{\alpha}+\delta_0+\delta_1 a_n]}$$

We require $\beta\delta_0 < \frac{1}{2}$ for convergence and in that case $O_R(\mathcal{J},x^*) \geq 1+\alpha$

if $\delta_0 = 0$ and $R_1(\mathcal{J},x^*) \leq \frac{\beta\delta_0}{1-\beta\delta_0}$ if $\delta_0 \neq 0$.

Another approximate Newton process assumes (4.15') with (4.15a) and again $\gamma(a) = \delta_0 + \delta_1(a)$. Then we obtain the majorizing equation

$$a_{n+1} = \frac{\beta[\frac{L}{1+\alpha} a_n^{\alpha}+\delta_0+\delta_1 a_n]a_n}{1-\beta[2\delta_0+\delta_1 a_n+La_n^{\alpha}]}$$

and for convergence we need $\beta\delta_0 < \frac{1}{3}$. $O_R(\mathcal{J},x^*) \geq 1+\alpha$ if $\delta_0 = 0$ and $R_1(\mathcal{J},x^*) \leq \frac{\beta\delta_0}{1-2\beta\delta_0} < 1$ if $\delta_0 \neq 0$.

For the two-point secant process, we assume (4.15) with $\gamma(a,b) \leq K(a+b)$. Then we have

$$a_{n+1} = \frac{\beta[\frac{L}{1+\alpha} a_n^{\alpha}+K(a_n+a_{n-1})]a_n}{1-\beta[La_n^{\alpha}+K(a_n+a_{n-1})]} ,$$

and $O_R(\mathcal{J},x^*) \geq \min\,(1+\alpha, \frac{1+\sqrt{5}}{2})$.

In Steffensen's method, assume (4.15) with $\gamma(a) \leq Ka$ to obtain

$$a_{n+1} = \frac{\beta[\frac{L}{1+\alpha} a_n^{\alpha}+Ka_n]a_n}{1-\beta[La_n^{\alpha}+Ka_n]}$$

and $O_R(\mathcal{J},x^*) \geq \min\,(1+\alpha,2)$.

Finally, we can discuss the so-called combined methods as considered for example by Schmidt and Schwetlick [1968]. In these an integer n_0 is given and the operator J is kept fixed on each index set

$$\ell(n_0+1) \leq n \leq (\ell+1)(n_0+1)-1 \text{ for } \ell = 0,1,\ldots .$$

If we use the assumptions for the two-point secant process and the notation

$$x_{n,\nu} \equiv x((n_0+1)n+\nu) \text{ for } \nu = 0,1,\ldots,n_0 \ ,$$

we can write the method as

(4.20a) $\quad x_{n,\nu+1} = x_{n,\nu} - J(y(n),y(n-1))^{-1}F(x_{n,\nu}) , \ \nu = 0,1,\ldots,n_0,$

$$n = 0,1,\ldots \ ,$$

(4.20b) $\quad x_{n,0} = y(n) \ ,$

(4.20c) $\quad y(n+1) = x_{n,n_0+1}; \ y(0), \ y(-1)$ given

In this case

$$\|J(y(n),y(n-1)) - F'(x_{n,\nu})\| \leq K(\|y(n)-x^*\|+\|y(n-1)-x^*\|)$$

$$+ L\|y(n)-x_{n,\nu}\|^{\alpha}$$

for $\nu = 0,1,\ldots,n_0$. With $a_{n,\nu}$ corresponding to $\|x_{n,\nu}-x_{n,\nu-1}\|$, the majorizing equation is

$$a_{n,\nu+1} = \frac{\beta[\frac{L}{1+\alpha} a_{n,0}^{\alpha}+K(a_{n,0}+a_{n-1,0})+L(a_{n,0}+a_{n,\nu})^{\alpha}](a_{n,\nu})}{1-\beta[La_{n,0}^{\alpha}+K(a_{n,0}+a_{n-1,0})+L(a_{n,0}+a_{n,\nu})^{\alpha}]}$$

where $a_{n,n_0+1} = a_{n+1,0}$. Theorem 3.8 again applies to give the result that x^* is a point of attraction of the process, meaning that for

$y(-1)$, $y(0)$ sufficiently close to x^*, $\lim_{n\to\infty} y(n) = \lim_{n\to\infty} x_{n,\nu} = x^*$ for each $\nu = 0,1,\ldots,n_0$. For the convergence rate, notice that if $\lim_{n\to\infty} a_{n,\nu} = 0$, then eventually $a_{n,\nu+1} \leq a_{n,\nu}$ and $a_{n+1,0} \leq a_{n,n_0}$ for $n = 0,1,\ldots;\ \nu = 0,\ldots,n_0$. Hence, we have for a constant $\eta > 0$

$$a_{n,\nu+1} \leq (\eta a_{n-1,0}^{\alpha})a_{n,\nu} \text{ for } \nu = 0,1,\ldots,n_0 \text{ and } n \text{ sufficiently large.}$$

Then $a_{n+1,0} = a_{n,n_0+1} \leq \eta^{n_0+1}(a_{n-1,0}^{\alpha})^{n_0+1} a_{n,0}$ for n sufficiently large. Hence, the sequence $\{a_{\nu,0}\}$ satisfies the difference equation

$$b_{n+1} = \delta(b_{n-1})^{\alpha(n_0+1)} \cdot b_n,\ n = 0,1,\ldots ,$$

where δ is constant, and the convergence rate according to Corollary 3.7 is $\frac{1}{2} + \sqrt{1+4(\alpha(n_0+1))}/2$. Hence, if $\mathcal{J}$ refers to the process (4.20) with iterates $\{y(n)\}$, then $O_R(\mathcal{J},x^*) \geq \frac{1}{2} + \sqrt{1+4(\alpha(n_0+1))}/2$. The actual iterates $x_{n,\nu}$ have

$$O_R \geq \left(\frac{1+\sqrt{1+4(\alpha(n_0+1))}}{2}\right)^{1/(n_0+1)}$$

These examples were used for illustrative purposes and are either known or extensions of known results (see Schmidt [1968], Ortega and Rheinboldt [1970], Rheinboldt [1968]).

Section 2. Semi-Local Convergence, Majorizing Sequences

For the proof of semi-local convergence results we will employ a technique similar to that of Section 1. For an m-step process, $\mathcal{J}$, we seek explicit conditions for the initial vectors $x(j)$, $j = -m+1,\ldots,0$, under which the iterates $x(n)$ converge to some point x^*, but in this case the existence of x^* is not a priori assumed.

The approach used here involves majorizing the numbers $\{\|x(n)-x(n-1)\|\}$ by a summable sequence $\{a_n\}$. It generalizes the theory of Rheinboldt [1968] in which the sequence $\{a_n\}$ is bypassed and replaced by $\{t_n-t_{n-1}\}$, as discussed in Chapter 3. His idea arose from the majorizing functions of Kantorovich [1949] and has also been used by Ortega [1968].

Our technique can be described generally in the following way.

Basic Semi-Local Majorant Principle. Suppose that there are sets $E_i \subseteq X$, $i = -m+1,\ldots,0,1,\ldots$, and isotone functions $\phi(j,\ldots):D_j \subseteq [0,\infty)^{j+m-1} \to [0,\infty)$, $j = 0,1,\ldots$, such that whenever $x(i) \in E_i$ for $i = -m+1,\ldots,n$, $n \geqslant 0$, and

$$(\|x(n)-x(n-1)\|,\ldots,\|x(1)-x(0)\|,\ldots,\|x(-m+2)-x(-m+1)\|)^T \in D_n$$

then $x(n+1)$ exists, lies in E_{n+1}, and satisfies

$$\|x(n+1)-x(n)\| \leqslant \phi(n,\|x(n)-x(n-1)\|,\ldots,\|x(-m+2)-x(-m+1)\|) \tag{4.30}$$

Assume now that $\underset{\sim}{a}^0 = (a_0,a_{-1},\ldots,a_{-m+2}) \in \mathcal{D}_S^0(\phi)$ and let $\{a_n\}$ be the corresponding solution of

$$(4.31) \qquad a_{n+1} = \phi(n,a_n,\ldots,a_{-m+2}),\ n = 0,1,\ldots .$$

A simple induction argument shows that if $x(j) \in E_j$ for $j = -m+1,\ldots,0$ and if $\|x(j)-x(j-1)\| \leq a_j$ for $j = -m+2,\ldots,0$, then the entire sequence of iterates $\{x(n)\}$ from $\mathcal{J}$ exists and satisfies

$$(4.32) \qquad \|x(n)-x(n-1)\| \leq a_n,\ n = -m+2,\ldots,0,1,\ldots .$$

In that case, the summability of $\{a_n\}$ implies that $\{x(n)\}$ is a Cauchy sequence in X and thus converges to some $x^* \in X$. Moreover, the following estimates hold.

$$(4.33a) \qquad \|x(k+\ell)-x(k)\| \leq \sum_{j=k+1}^{k+\ell} a_j \text{ for } k \geq -m+1,\ \ell \geq 1 ,$$

$$(4.33b) \qquad \|x^*-x(k)\| \leq \sum_{j=k+1}^{\infty} a_j \text{ for } k \geq -m+1 .$$

As before, we call (4.31) a (semi-local) <u>majorizing difference equation</u> for $\mathcal{J}$ and $\{a_n\}$ a <u>majorizing sequence</u>.

In practice, condition (4.30) is not easily verified. However, when $\mathcal{J}$ has the form

$$(4.34) \qquad x(n+1) = G(n,x(n),\ldots,x(n-m+1)),\ n = 0,1,\ldots ,$$

then a sufficient condition for (4.30) is

$$\|G(n,y_n,\ldots,y_{n-m+1}) - G(n-1,y_{n-1},\ldots,y_{n-m})\|$$

$$\leq \hat{\phi}(n,\|y_n-y_{n-1}\|,\ldots,\|y_{n-m+1}-y_{n-m}\|,\|y_n-x(0)\|,\ldots,\|y_{n-m}-x(0)\|)$$

where $\hat{\phi}(n,\ldots)$ is isotone and has an appropriate domain of definition.

In that event,

$$\phi(n,\|x(n)-x(n-1)\|,\ldots,\|x(-m+2)-x(-m+1)\|)$$

$$= \hat{\phi}(n,\|x(n)-x(n-1)\|,\ldots,\|x(n-m+1)-x(n-m)\|,$$

$$\sum_{j=1}^{n}\|x(j)-x(j-1)\|,\ldots,\sum_{j=1}^{n-m}\|x(j)-x(j-1)\|)$$

In analogy with (4.7), a more general condition involving a functional relationship among the G's can be used. For example, in the case $m = 1$, we could assume

$$\|G(n+1,G(n,x))-G(n,x)\| \tag{4.35}$$
$$\leq \hat{\phi}(n+1),\|G(n,x)-x\|,\|G(n,x)-x(0)\|,\|x-x(0)\|)$$

Here $\phi(n,a_n,\ldots,a_1) = \hat{\phi}(n,a_n,\sum_{j=1}^{n} a_j,\sum_{j=1}^{n-1} a_j)$.

The requirement that whenever $x(j) \in E_j$, $j = -m+1,\ldots,n$, exist, then $x(n+1)$ is in E_{n+1} can be guaranteed by the assumption that

$$\bar{S}(x(\ell),\sum_{j=\ell+1}^{n+1} a_j) \subseteq E_{n+1}$$

for at least one $\ell \in \{-m+1,\ldots,0,1,\ldots,n+1\}$.

Under certain conditions on the function G of (4.34), the limit x^* of the iterates is a "fixed point" of G.

Lemma 4.3. Suppose that $\{x(n)\}$ is a sequence of iterates generated by (4.34) with $\lim_{n\to\infty} x(n) = x^*$, that $\lim_{n\to\infty} G(n,x^*,\ldots,x^*) \equiv G(\infty,x^*,\ldots,x^*)$ exists, and that $\{G(n,\ldots)\}$ is equicontinuous in a neighborhood of $(x^*,\ldots,x^*)$. Then

$$(4.36) \qquad x^* = G(\infty, x^*, \ldots, x^*) \ .$$

The proof follows from the estimate

$$\|x^*-G(\infty,x^*,\ldots,x^*)\| \leq \|x^*-x(n+1)\|$$

$$+ \|G(n,x^*,\ldots,x^*)-G(n,x(n),\ldots,x(n-m+1))\|$$

$$+ \|G(n,x^*,\ldots,x^*) - G(\infty,x^*,\ldots,x^*)\| \ .$$

Under the assumptions of Lemma 4.3, it is of interest to know domains in which x^* is the unique solution of (4.36). Of course, such domains depend on the structure of G and have nothing intrinsically to do with (4.34). However, the existence of the convergent sequence $\{x(n)\}$ and the majorizing function ϕ can in certain cases give enough information for the determination of uniqueness regions. The following lemma contains the uniqueness results of Rheinboldt [1968].

<u>Lemma 4.4</u>. Suppose that $\{x(n)\}$ is formed by (4.34) with $x(n) \in \bar{S} \subseteq \bigcap_{j=-m+1}^{\infty} E_j$, and $\lim_{n\to\infty} x(n) = x^*$. Assume further that the following four conditions hold: (a) $\{G(n,\cdot)\}$ is an equicontinuous family on $\bar{S}^m$, (b) $\lim_{n\to\infty} G(n,z,\ldots,z) \equiv G(\infty,z,\ldots,z)$ exists for all $z \in \bar{S}$, (c) $\|G(n,z,\ldots,z)-G(\infty,z,\ldots,z)\| \leq \lambda_n$, and (d)

$$\|G(n,x(n),\ldots,x(n-m+1))-G(n,z,\ldots,z)\|$$

$$\leq \psi(n,\|x(n)-z\|,\ldots,\|x(n-m+1)-z\|,$$

$$\|x(n)-x(0)\|,\ldots,\|x(n-m+1)-x(0)\|,\|z-x(0)\|)$$

for $n = 0,1,\ldots$ where $\psi(n,\ldots)$ is isotone on an appropriate domain. If the convergence set $\mathcal{D}_C^0(\psi)$ of

$$(4.37) \qquad b_{n+1} = \psi(n,b_n,\ldots,b_{n-m+1},\|x(n)-x(0)\|,\ldots,\|x(n-m+1)-x(0)\|,b_0)+\lambda_n$$

contains $(\hat{b}_0,\ldots,\hat{b}_0)$ for some $\hat{b}_0 > 0$, then x^* is the unique solution of (4.36) in $\bigcap_{j=-m+1}^{0} \bar{S}(x(j),\hat{b}_0)$.

Proof: Suppose that $G(\infty,y^*,\ldots,y^*) = y^* \in \bigcap_{j=-m+1}^{0} \bar{S}(x(j),\hat{b}_0)$, and let $\beta_n \equiv \|x(n)-y^*\|$ for $n = -m+1,\ldots,0,1,\ldots$. Then

$$\begin{aligned}\beta_{n+1} = \|x(n+1)-y^*\| &\leq \|G(n,x(n),\ldots,x(n-m+1))-G(n,y^*,\ldots,y^*)\| \\ &\quad + \|G(n,y^*,\ldots,y^*)-G(\infty,y^*,\ldots,y^*)\| \\ &\leq \psi(n,\|x(n)-y^*\|,\ldots,\|x(n-m+1)-y^*\|,\|x(n)-x(0)\|,\ldots, \\ &\qquad \|x(n-m+1)-x(0)\|,\|y^*-x(0)\|) \\ &\quad + \lambda_n .\end{aligned}$$

Define $\{b_n\}$ by (4.37) with $b_j = \hat{b}_0$ for $j = -m+1,\ldots,0$. But then a simple induction argument shows that $\beta_n \leq b_n$ for all n. Since

$$\lim_{n\to\infty} \|x(n)-y^*\| = \lim_{n\to\infty} \beta_n = 0 ,$$

we have $x^* = y^*$.

Let us now consider several examples of the semi-local majorant principle.

Example 4.5. We discuss first the simple one-step process

$$x(n+1) = G(n,x(n)), \quad n = 0,1,.. \tag{4.38}$$

when $G(n,\cdot):E \subseteq X \to X$, $n = 0,1,\ldots$. Assume that G satisfies (4.35) with $\hat{\phi}(n,a,b,c) = \hat{\phi}(n,a) = \alpha(n)a$ for some non-negative sequence $\{\alpha(j)\}$. This has been called an iterated contraction condition (see Ortega and Rheinboldt [1970]). The majorizing equation is

$$a_{n+1} = \hat{\phi}(n,a_n) = \alpha(n)a_n, \quad n = 1,2,\ldots . \tag{4.39}$$

For any $a_1 > 0$, the solution of (4.39) is summable if and only if

$$\sum_{j=1}^{n} \left(\prod_{k=1}^{j} \alpha(k) \right) \tag{4.40}$$

is bounded for all n. If this is true for some bound $M > 0$, then we have the conclusion that for any $x(0) \in E$ with $\bar{S}(x(0),Ma_1) \subseteq E$ and $a_1 = \|G(0,x(0))-x(0)\|$, the sequence $\{x(n)\}$ generated by (4.38) exists in E with $\lim_{n\to\infty} x(n) = x^*$ and $\|x(n)-x^*\| \leq \sum_{j=n}^{\infty} \left(\prod_{k=1}^{j} \alpha(k)\right)a_1$. If $\{G(n,\cdot)\}$ is equicontinuous on E and if $\lim_{n\to\infty} G(n,x^*) = x^*$, then $x^* = G(\infty,x^*)$.

As a point of interest, in the case that $\{\alpha(j)\}$ is a positive sequence, the boundedness of (4.40) is equivalent to the condition that there exist constants $a \geq 1$, $0 \leq \lambda < 1$ such that

$$\prod_{j=k}^{k+\ell} \alpha(j) \leq a\lambda^{\ell+1} \text{ for } k,\ell \geq 0 .$$

This is exactly the theorem of Hahn (see Lemma 1.16).

If uniqueness results are desired, we assume that under the above conditions

$$\|G(n,x(n))-G(n,z)\| \leq \psi(n,\|x(n)-z\|,\|x(n)-x(0)\|,\|z-x(0)\|)$$

for all $z \in \bar{\bar{S}}(x(0),b_0)$, $b_0 > 0$, $\psi(n,\ldots)$ isotone. If $b_0 \in \mathcal{S}_C^0(\psi)$ for

$$b_{n+1} = \psi(n,b_n,\|x(n)-x(0)\|,b_0),\ n = 0,1,\ldots ,$$

then x^* is the unique solution of $x^* = G(\infty,x^*)$ in $\bar{\bar{S}}(x(0),b_0)$.

Example 4.6. As a second example we have the general approximate Newton iteration as considered by Rheinboldt [1968] and seen in Section 1. The process is

$$(4.41)\qquad x(n+1) = x(n) - A(x(n))^{-1}F(x(n)),\ n = 0,1,\ldots$$

where $F\colon E \subseteq X \to Y$, $A(x) \in L(X,Y)$. The most common estimate is

$$\|x(n+1)-x(n)\| \leq \|A(x(n))^{-1}\|\|Fx(n)-Fx(n-1)-A(x(n-1))(x(n)-x(n-1))\|;$$

it predicates assumptions of the following type.

(4.42a) F has a Hölder-continuous F-derivative on the open, convex set E with constants $L \geq 0$, $0 < \alpha \leq 1$.

(4.42b) $A\colon E \subseteq X \to L(X,Y)$ and for some $x(0) \in E$, $\omega_1 > 0$, $\eta \geq 0$

$$\|A(x)-A(x(0))\| \leq \eta\|x-x(0)\|^{\omega_1} \text{ whenever } x \in E ,$$

(4.42c) $\|F'(x)-A(x)\| \leq \delta_0+\delta_1\|x-x(0)\|^{\omega_2}$, $\omega_2 > 0$, $\delta_i \geq 0$,

whenever $x \in E$,

(4.42d) $A(x(0))^{-1}$ exists and is bounded by β

If $x(n-1)$, $x(n) \in E$ exist and $\beta\eta\|x(n)-x(0)\|^{\omega_1} < 1$, then $x(n+1)$ exists and

$$\|x(n+1)-x(n)\|$$

$$\leq \frac{\beta\left(\frac{L}{1+\alpha}\|x(n)-x(n-1)\|^{\alpha}+(\delta_0+\delta_1\|x(n-1)-x(0)\|^{\omega_2})\right)\|x(n)-x(n-1)\|}{1-\beta\eta\|x(n)-x(0)\|^{\omega_1}}$$

Accordingly, the majorizing equation is

$$a_{n+1} = \left[\frac{\beta}{1-\beta\eta\left(\sum_{j=1}^{n} a_j\right)^{\omega_1}}\right]\left[\frac{L}{1+\alpha}\, a_n^{\alpha}+\delta_0+\delta_1\left(\sum_{j=1}^{n-1} a_j\right)^{\omega_2}\right] a_n \,. \tag{4.43}$$

This is exactly the type of difference equation treated in Theorem 3.8 and the conclusion is that a choice of $a_1 > 0$ as a solution of

$$0 \leq \left[\frac{\beta}{1-\beta\eta\left(\frac{a_1}{1-\lambda}\right)^{\omega_1}}\right]\cdot\left[\frac{L}{1+\alpha}\, a_1^{\alpha}+\delta_0+\delta_1\left(\frac{a_1}{1-\lambda}\right)^{\omega_2}\right] \leq \lambda$$

for some $\lambda \in (0,1)$, gives $a_1 \in \mathcal{D}^0_S$. To guarantee the existence of a_1 and λ we require $\beta\delta_0 < 1$. If a_1 and λ are found, $\|x(1)-x(0)\| \leq a_1$ and $\bar{S}(x(0), \frac{a_1}{1-\lambda}) \subseteq E$, then according to the majorant principle all $x(n)$ of the iterative process (4.41) exist and satisfy

$$\lim_{n\to\infty} x(n) = x^*,\ Fx^* = 0\ ,$$

$$\|x(n+1)-x(n)\| \leqslant \lambda^n a_1\ ,$$

$$\|x^*-x(n)\| \leqslant a_1 \sum_{j=n}^{\infty} \lambda^j = \frac{a_1\lambda^n}{1-\lambda}$$

$$\|x(n)-x(0)\| \leqslant a_1 \sum_{j=0}^{n-1} \lambda^j \leqslant \frac{a_1}{1-\lambda}$$

Theorem 3.5 and Corollary 3.7 provide the following R-convergence rate results.

$$R_1\{a_n\} \leqslant \left[\frac{\beta}{1-\beta\eta\left(\frac{a_1}{1-\lambda}\right)^{\omega_1}}\right]\left[\delta_0+\delta_1\left(\frac{a_1}{1-\lambda}\right)^{\omega_2}\right] \text{ if } \delta_0+\delta_1 \neq 0$$

$$R_p\{a_n\} < 1 \text{ where } p = (1+\alpha) \text{ if } \delta_0 = \delta_1 = 0$$

Apparently, since $\|x(n)-x^*\| \leqslant \sum_{j=n+1}^{\infty} a_j$, the same estimates hold for $\{x(n)-x^*\}$. Since this is not a local theorem, we cannot expect to get O_R-estimates.

In certain cases, there is a first integral for (4.43) and better initial values can be found. For example, if $\delta_0 = \delta_1 = 0$ and $\alpha = 1 = \omega_1$, then (3.28) replaces (4.43). Rheinboldt [1968] has found first integrals for more general equations than (3.28), but not for (4.43).

An additional example of the semi-local majorant principle is found in the next section.

Section 3. Majorizing Systems

The final approach considered here is a natural generalization of the majorant sequence technique of the previous two sections. Although it also applies to local results, we will be content to investigate its use in proving semi-local theorems. It arose from a futile attempt to apply the majorizing sequence approach of Rheinboldt [1968] to processes other than approximate Newton methods. As we have seen in Section 2, for iterates $\{x(n)\}$ an estimation of the crucial quantity $\|x(n+1)-x(n)\|$ in terms of $\|x(n)-x(n-1)\|, \ldots,$ $\|x(n-m+1)-x(0)\|$, etc. can yield satisfactory results in many cases. However, in some instances either this approach is not feasible or else it is better strategy to use estimates which take advantage of the structure of the original system. Thus we are led to estimate other quantities besides $\|x(n+1)-x(n)\|$ and as a result to consider systems of difference equations. The examples given below illustrate the value of this approach, and it appears that all known convergence proofs which rely on norm estimates actually use the principle which we propose here. In particular, this idea has been used implicitly by authors including Kantorovich [1948], Schmidt [1968], Bittner [1963], Chen [1964], and Kivistek [1961]. Its explicit formulation as given here appears to be new.

Since the choice of the quantities to be majorized depends on the particular process at hand, no formal presentation of this majorizing procedure in terms of several general theorems can be given.

However, the procedure can be described generically in the following way.

<u>Extended Semi-Local Majorant Principle</u>. Let $\mathscr{I}$ be an m-step iterative process with the following properties.

(i) There are sets $E_j \subseteq X$, $j = -m+1,\ldots,0,1,\ldots,$ such that whenever the iterates $x(j) \in E_j$ exist for $j = -m+1,\ldots,n$, then the next iterate $x(n+1)$ exists and lies in E_{n+1}.

(ii) Let $Q_i(j,x(j),\ldots,x(-m+1))$, $i = 1,2,\ldots,i_0$, be certain selected quantities at each step j of $\mathscr{I}$, where $Q_i(j,\ldots):E_j x \ldots x E_{-m+1} \subseteq X^{m+j} \to X$. If the iterates $x(j) \in E_j$, $j = -m+1,\ldots,n$, exist and if the vectors $\underset{\sim}{a}(j)$ in $[0,\infty)^{i_0}$ satisfy

$$(4.45) \qquad \|Q_i(j,x(j),\ldots,x(-m+1))\| \leq a_i(j),\ i=1,\ldots,i_0;\ j=0,1,\ldots\ ,$$

then

$$\|Q_i(n+1,x(n+1),\ldots,x(-m+1))\| \leq a_i(n+1),\ i = 1,\ldots,i_0$$

where

$$(4.46) \qquad a_i(n+1) = \phi_i(n,\underset{\sim}{a}(n),\ldots,\underset{\sim}{a}(0)),\ i = 1,\ldots,i_0\ ,$$

and $\phi_i(n,\ldots):D_n \subseteq [0,\infty)^{(n+m)i_0} \to [0,\infty)$ is isotone for each i,n fixed.

The system (4.46) is called a <u>majorizing system</u> of difference equations.

(iii) The quantity $\|x(n)-x(n-1)\|$ is bounded by a function of $\underset{\sim}{a}(n)$, say

$$\|x(n)-x(n-1)\| \leq \psi(n,\underset{\sim}{a}(n)) \equiv d(n) \tag{4.47}$$

Then, as in Section 2, the summability of $\{d_n\}$ establishes the convergence of $x(n)$ to some limit x^*. Hence, if there is an initial vector $\underset{\sim}{a}(0) \geq \underset{\sim}{0}$ such that the solution $\{\underset{\sim}{a}(n)\}$ of the system of difference equations (4.46) exists and such that $\{d_n\}$ is summable, then a simple induction argument shows that whenever $x(j) \in E_j$, $j = -m+1,\ldots,0$, and (4.45) is satisfied for $j = 0$, we have the existence of a limit x^* of $x(n)$ and the estimates

$$\|x(n+\ell)-x(n)\| \leq \sum_{j=n+1}^{n+\ell} d_j, \; n \geq 0, \; \ell \geq 0 \, , \tag{4.48a}$$

$$\|x^*-x(n)\| \leq \sum_{j=n+1}^{\infty} d_j, \; n \geq 0 \, . \tag{4.48b}$$

In practice, the quantities $Q_i(j,x(j),\ldots,x(-m+1))$ are usually any of $x(j)-x(j-1)$, $x(j)-x(0)$, $P(x(j))$ for some function P, etc. If only the first one of these is needed, we obtain the majorizing sequence approach of Section 2 and $\psi(n,\underset{\sim}{a}(n)) \equiv a(n) = d(n)$. The condition $x(n) \in E_n$ can again be guaranteed by

$$\bar{S}(x(\ell), \sum_{j=\ell+1}^{n} d_j) \subseteq E_n \tag{4.49}$$

for some $\ell \in \{0,1,\ldots,n\}$.

As in Sections 1 and 2, the convergence proof for the iterative process now rests with an analysis of the corresponding majorizing system of difference equations. Let us now illustrate the use of this principle with some examples.

<u>Example 4.7</u> (Newton and Secant Type Processes). We return to the following general method discussed in Section 1,

$$(4.50) \qquad x(n+1) = x(n) - J(n,x(n),\ldots,x(n-m+1))^{-1}F(x(n)),\ n=0,1,\ldots$$

where $F\colon E \subseteq X \to Y$ and $J(n,\ldots)\colon E^m \subseteq X^m \to L(X,Y)$. There is a wide variety of approaches to proving semi-local results for this process, and so it provides a good illustration of the extended majorant principle.

Consider first the estimate

$$(4.51) \qquad \|x(n+1)-x(n)\| \leq \|J(n,x(n),\ldots,x(n-m+1))^{-1}\| \cdot \|Fx(n)-Fx(n-1)-J(n-1,x(n-1),\ldots,x(n-m))(x(n)-x(n-1))\|$$

which predicates certain assumptions on F and J. These might be of the following type as used, for example, in the "consistent approximation" formulation of Ortega (see Ortega and Rheinboldt [1970]).

(4.52a) F has a Hölder continuous F-derivative on the open, convex set E with constants $L \geq 0$, $0 < \alpha \leq 1$.

(4.52b) Whenever $x(i) \in E$, $i = n-m+1,\ldots,n$, are generated by (4.50), then

$$\|J(n,x(n),\ldots,x(n-m+1))-F'(x(n))\| \leq \eta(n,x(n),\ldots,x(n-m+1)) \equiv \eta_n$$

where $\eta(n,\ldots)\colon E^m \subseteq X^m \to [0,\infty)$.

(4.52c) $F'(x(0))^{-1}$ exists and is bounded by β

The functions Q_i are in this case $F(x(n))$, $F'(x(n))^{-1}$, $J(n,x(n),\ldots,x(n-m+1))^{-1}$ and $x(n)-x(n-1)$. If the numbers $a_i(j)$ satisfy

(4.53a) $\quad \|x(j)-x(j-1)\| \leq a_1(j),\ j = 1,2,\ldots,n$

(4.53b) $\quad \|Fx(j)\| \leq a_2(j),\ j = 0,1,\ldots,n-1$

(4.53c) $\quad \Gamma_j \equiv \|J(j,x(j),\ldots,x(j-m+1))^{-1}\| \leq a_3(j),\ j = 0,1,\ldots,n-1$

whenever the $x(j)$ exist for $j = -m+1,\ldots,n$, and

(4.53d) $\quad \|F'(x(j))^{-1}\| \leq a_4(j),\ j = 0,1,\ldots,n-1,\ a_4(0) \equiv \beta$

then we have

(4.54) $\quad \|x(n+1)-x(n)\| \leq a_1(n+1) \equiv a_2(n)a_3(n)$,

(4.55) $\quad \|Fx(n)\| \leq a_2(n) \equiv \frac{L}{1+\alpha}(a_1(n))^{1+\alpha}+\eta_{n-1}a_1(n)$

(4.56a) $$\Gamma_n \leq a_3(n) \equiv \frac{a_3(n-1)}{1-a_3(n-1)[\eta_n+\eta_{n-1}+La_1^{\alpha}(n)]}$$

and $a_3(0) \equiv \frac{\beta}{1-\beta\eta_0}$,

or

(4.56b) $$\Gamma_n \leq a_3(n) \equiv \frac{\beta}{1-\beta[\eta_n+L(\sum_{j=1}^{n} a_1(j))^{\alpha}]}$$

or

$$(4.56c) \qquad \Gamma_n \leq a_3(n) \equiv \frac{a_4(n)}{1-a_4(n)\eta_n}$$

$$a_4(n) \equiv \frac{a_4(n-1)}{1-a_4(n-1)L_1(a_1(n))^{\alpha}}$$

where the denominators of (4.56) must be positive, since Banach's lemma has been used. The estimates follow from

$$\Gamma_n \leq (\Gamma_{n-1})[1-\Gamma_{n-1}\|J(n,x(n),\ldots,x(n-m+1))-J(n-1,x(n-1),\ldots,x(n-m))\|]^{-1}$$

$$\leq (\Gamma_{n-1})[1-\Gamma_{n-1}(\|J(n,x(n),\ldots,x(n-m+1))-F'(x(n))\|$$

$$+ \|F'(x(n))-F'(x(n-1))\|$$

$$+ \|J(n-1,x(n-1),\ldots,x(n-m))-F'(x(n-1))\|)]^{-1} ,$$

or

$$\Gamma_n \leq \Gamma_0[1-\Gamma_0\|J(n,x(n),\ldots,x(n-m+1))-F'(x(0))\|]^{-1}$$

$$\leq \Gamma_0[1-\Gamma_0(\|J(n,x(n),\ldots,x(n-m+1))-F'(x(n))\|+\|F'(x(n))-F'(x(0))\|)]^{-1}$$

or

$$\|F'(x(n))^{-1}\| \leq \|F'(x(n-1))^{-1}\| [1-\|F'(x(n-1))^{-1}\|\|F'(x(n))-F'(x(n-1))\|]^{-1}$$

A majorizing system is thus defined by (4.54), (4.55), and one of (4.56). Of course, the quantity η_j is still unspecified until the method (that is, J) is defined. For instance, we have

$$(4.57a) \qquad \eta_n = \eta \cdot a_1(n), \quad \eta \text{ constant}$$

in the case of the secant process, and

(4.57b) $\qquad \eta_n = \theta a_2(n), \quad \theta$ constant

for Steffensen's method.

For the same estimate (4.51), (4.52) could be replaced by the alternate conditions below which come from Schmidt's [1961] "Steigung formulation.

(4.58a) Whenever $x(j)$, $j = n-m,\ldots,n$, exist in E, then

$$Fx(n) - Fx(n-1) = J(n,x(n),\ldots,x(n-m+1))(x(n)-x(n-1))$$

(4.58b) and $\|J(n,x(n),\ldots,x(n-m+1))-J(n-1,x(n-1),\ldots,x(n-m))\|$

$$\leq \theta(n,x(n),\ldots,x(n-m)) \equiv \theta_n ,$$

where $\theta(n,\ldots):E^{m+1} \subseteq X^{m+1} \to [0,\infty)$

(4.58c) $J(0,x(0),\ldots,x(-m+1))^{-1}$ exists and is bounded in norm by γ

In this case, the discussion given above leads to a system of the form

(4.59) $\qquad a_1(n+1) = a_2(n)a_3(n)$

(4.60) $\qquad a_2(n) = \theta_n a_1(n)$

(4.61) $\qquad a_3(n) = \dfrac{a_3(n-1)}{1-a_3(n-1)\theta_n} \quad$ or $a_3(n) = \dfrac{a_3(0)}{1-a_3(0)(\sum_{j=1}^{n} \theta_j)}, \; a_3(0) \equiv \gamma$

Again, θ_n is left unspecified, but could be one of

(4.62a) $\theta_n \equiv \eta(a_1(n)+a_1(n-1))$, η constant (Secant method)

(4.62b) $\theta_n \equiv \theta(a_2(n-1))$, θ constant (Steffensen's method)

Other norm estimates besides (4.51) can also be used. For instance, Bittner [1963] has chosen

(4.63)
$$\begin{aligned}\|x(n+1)-x(n)\| &\leq \|(I+J(n,x(n),\ldots,x(n-m+1))-F'(x(n)))^{-1}\| \\ &\cdot\|F'(x(n))^{-1}F'(x(n-1))\|\|F'(x(n-1))^{-1}Fx(n)\|,\ n=0,1,\ldots\end{aligned}$$

In this case the basic assumptions for F and J could be

(4.64a) F has an F-derivative on the open, convex set E and for some fixed $x(0) \in E$

$$\|F'(x(-1))^{-1}[F'(y)-F'(x)]\| \leq L\|y-x\|^{\alpha}$$

$$\text{for all } x,y \in E,\ L \geq 0,\ 0 < \alpha \leq 1 .$$

(4.64b) When the quantity $F'(x(n))^{-1}(J(n,x(n),\ldots,x(n-m+1)-F'(x(n)))$ exists, then it is bounded in norm by ζ_n

We then have

(4.65a) $\|x(n+1)-x(n)\| \leq a_1(n+1) \equiv a_2(n)a_3(n)a_4(n)$,

$n=0,1,\ldots,\ a_1(0)=a_1(1)$,

(4.65b) $$\|[I+J(n,x(n),\ldots,x(n-m+1)-F'(x(n))]^{-1}\|$$

$$\leq a_2(n) \equiv (1-\zeta_n)^{-1},\ n = 0,1,\ldots\ ,$$

(4.65c) $$\|F'(x(n))^{-1}F'(x(n-1))\| \leq [1-\|I-F'(x(n-1))^{-1}F'(x(n))\|]^{-1}$$

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

$$\leq \left(1-[\prod_{j=0}^{n-1}\|F'(x(j))^{-1}F'(x(j-1))\|]\cdot\|F'(x(0))^{-1}(F'(x(n-1))-F'(x(n)))\|\right)^{-1}$$

$$\leq a_3(n) \equiv [1-(\prod_{j=0}^{n-1} a_3(j))L(a_1(n))^{\alpha}]^{-1},\ n = 1,2,\ldots\ ,$$

where $a_3(0)$ is given;

(4.65d) $$\|F'(x(n-1))^{-1}Fx(n)\|$$

$$\leq \|F'(x(n-1))^{-1}[Fx(n)-Fx(n-1)-F'(x(n-1))(x(n)-x(n-1))]\|$$

$$+ \|F'(x(n-1))^{-1}[F'(x(n-1))-J(n-1,x(n-1),\ldots,x(n-m))]\|\cdot\|x(n)-x(n-1)\|$$

$$\leq \|\int_0^1 F'(x(n-1))^{-1}[F'(x_\theta)-F'(x(n-1))](x(n)-x(n-1))d\theta\|$$

$$+ \zeta_{n-1}a_1(n)$$

$$\leq a_4(n) \equiv (\prod_{j=0}^{n-1} a_3(j))\frac{L(a_1(n))^{1+\alpha}}{1+\alpha} + \zeta_{n-1}a_1(n),\ n=1,2,\ldots\ ,$$

where $a_4(0)$ is given, and the denominators of the quantitites in (4.65b,c) must be positive and $\{\zeta_n\}$ is unspecified.

Let us now analyze a typical system of each type (4.54)-(4.56), (4.59)-(4.61), and (4.65), and then obtain a convergence statement for the associated process.

<u>Sample Process</u> (4.54)-(4.56). We consider the method of Steffensen

$$x(n+1) = x(n) - J(x(n))^{-1}Fx(n),\ n = 0,1,... \tag{4.66}$$

for solving $Fx = 0$ under assumptions (4.52) with $\alpha = 1$ and (4.57b). Then we have the majorizing system

$$a_1(n+1) = a_2(n)a_3(n),\ n = 0,1,... , \tag{4.54'}$$

$$a_2(n) = \frac{L}{2}(a_1(n))^2 + \theta a_2(n-1)a_1(n),\ n = 1,2,... , \tag{4.55'}$$

$$a_3(n) = a_3(n-1)\{1-a_3(n-1)[\theta a_2(n)+\theta a_2(n-1)+L(a_1(n))^2]\}^{-1} \tag{4.56a'}$$

$$a_3(0) = \frac{\beta}{1-\beta\theta a_2(0)},\ n = 1,2,...$$

The exact solution is unknown. However, by means of the following crude estimates the system can be simplified. Let

$$a_4(n) = (La_3(n)+2\theta)a_1(n+1)$$

Then

$$a_2(n) \leq \frac{1}{2}\ (La_3(n-1)+2\theta)a_2(n-1)a_1(n) \equiv \frac{1}{2}\,a_4(n-1)a_2(n-1)$$

$$a_3(n) \leq \frac{a_3(n-1)}{1-a_4(n-1)} \leq 2a_3(n-1) \text{ when } a_4(n-1) \leq \frac{1}{2}$$

$$a_4(n) \leq L(2a_3(n-1)+2\theta)(2a_3(n-1)a_2(n))$$

$$\leq L(2a_3(n-1)+2\theta)a_3(n-1)a_4(n-1)a_2(n-1) \leq 2(a_4(n-1))^2 \text{ where } a_4(n-1) \leq \frac{1}{2}$$

Thus, we have the new system

$$\begin{aligned} &b_1(n+1) = b_2(n)b_3(n),\ n = 0,1,\ldots \\ &b_2(n) = \frac{1}{2}\,b_4(n-1)b_2(n-1),\ n = 1,2,\ldots \\ &b_3(n) = 2b_3(n-1),\ n = 1,2,\ldots \\ &b_4(n) = 2(b_4(n-1))^2,\ n = 1,2,\ldots\ , \end{aligned} \tag{4.67}$$

with $a_i(n) \leq b_i(n)$ where $a_i(0) = b_i(0)$ for $i = 2,3,4$. By choosing $b_4(0) \leq \frac{1}{2}$, we obtain $b_4(n) \leq \frac{(2b_4(0))^{2^n}}{2} \leq \frac{1}{2}$ so that the condition $a_4(n) \leq \frac{1}{2}$ holds. But then, using an induction argument, we have

$$a_1(n+1) \leq b_1(n+1) = b_2(n)b_3(n)$$

$$= 2^n b_3(0)\left(\prod_{j=0}^{n-1} \frac{1}{2}\,b_4(j)\right)b_2(0) \leq \frac{(2b_4(0))^{2^n-1}}{2^n}\,b_1(1)\ ,$$

and

3

$$\sum_{j=1}^{n} a_1(j) \leq 2b_1(1) = 2a_1(1) ,$$

$$b_2(n) + \sum_{j=1}^{n} b_1(j) \leq \left(\frac{1}{4}\right)^{n+1} b_2(0)+2b_1(0) \leq \frac{1}{4} a_2(0)+2a_1(1)$$

$$\sum_{j=n+1}^{\infty} a_1(j) \leq \frac{a_1(1)}{b_4(0)} \frac{(2b_4(0))^{2^n}}{2^n} = \frac{a_1(1)}{a_4(0)} \frac{(2a_4(0))^{2^n}}{2^n}$$

For the iterative process (4.66), the conclusion derived from the majorant principle is that if

(4.68a) $\|J(x(0))^{-1}\| \leq a_3(0)$,

(4.68b) $\|F(x(0))\| \leq a_2(0)$,

(4.68c) $a_4(0) = (La_3(0)+2\theta)a_3(0)a_2(0) \leq \frac{1}{2}$,

(4.68d) $\frac{1}{4} a_2(0) + 2a_2(0)a_3(0) \leq r$,

(4.68e) the domain E contains the set $\bar{S}(x(0),r)$

then the iterates $\{x(n)\}$ generated by (4.66) exist, $x(n) \in E$, $F(x(n))-x(n) \in E$, and $\lim_{n\to\infty} x(n) = x^*$. Moreover,

$$\|x(n)-x^*\| \leq \frac{a_2(0)a_3(0)}{a_4(0)} \frac{(2a_4(0))^{2^n}}{2^n} , \; Fx^* = 0 .$$

A similar result has been obtained by Johnson and Scholz [1968] under the conditions (4.58). These authors did not recognize the significance of (4.54')-(4.56a') and derived only (4.67). Of interest is the fact that (4.67) is the same system obtained by Kantorovich [1948] in his proof of convergence of Newton's method. Apparently better initial conditions could be found if it were possible to solve the system (4.54')-(4.56a').

<u>Sample Process</u> (4.59)-(4.61). We consider next the secant method

$$x(n+1) = x(n) - J(x(n),x(n-1))^{-1}Fx(n), \; n = 0,1,... \tag{4.69}$$

under the conditions (4.58) and (4.62a). The analysis given below improves the result of Schmidt [1961] who has obtained the same majorizing system:

$$a_1(n+1) = a_2(n)a_3(n), \; n = 0,1,... ,$$

$$a_2(n) = \eta(a_1(n)+a_1(n-1))a_1(n), \; n = 1,2,... ,$$

$$a_3(n) = a_3(n-1)\{1-a_3(n-1)(\eta)(a_1(n)+a_1(n-1))\}^{-1}, \; n=1,2,...$$

For convenience, let $a_4(n) = a_3(n-1)\eta(a_1(n)+a_1(n-1))$. Then we have the equivalent system

$$a_1(n+1) = \frac{a_4(n)a_1(n)}{1-a_4(n)} \;, \; n = 0,1,...$$

$$a_2(n) = \eta(a_1(n)+a_1(n-1))a_1(n), \; n = 1,2,... \tag{4.70}$$

$$a_3(n) = \frac{a_3(n-1)}{1-a_4(n)} \;, \; n = 1,2,... .$$

If $a_4(1) < (1-\frac{\sqrt{2}}{2}) < \frac{1}{2}$ and $a_1(0) = a_1(1)$, then $a_1(2) = \frac{a_4(1)}{1-a_4(1)} a_1(1)$.

and

$$a_4(2) \leq \frac{1}{2} \frac{a_4(1)}{(1-a_4(1))^2} < (1-\frac{\sqrt{2}}{2})$$

Continue by induction to obtain

$$a_4(n+1) \leq \frac{1}{2} \frac{a_4(n)}{(1-a_4(n))^2} , \quad a_4(n) < (1-\frac{\sqrt{2}}{2}) < \frac{1}{2} ,$$

$$a_4(n+1) \leq a_4(n), \quad n = 1,2,\ldots$$

By Kantorovich's Lemma (see Corollary 3.4) $a_4(n)$ decreases to 0.
Hence $a_1(n)$ and $\frac{a_4(n)}{1-a_4(n)}$ are decreasing with

$$\frac{a_4(n)}{1-a_4(n)} < \frac{1-\frac{\sqrt{2}}{2}}{\frac{\sqrt{2}}{2}} = \sqrt{2} - 1 < \frac{1}{2}$$

and finally $a_1(n) \leq (\sqrt{2}-1)a_1(n-1)$ for $n = 2,3,\ldots$. In fact,

$$\sum_{j=1}^{n} a_1(j) = \sum_{j=1}^{n} (\prod_{k=1}^{j-1} \frac{a_4(k)}{1-a_4(k)}) a_1(1) \leq (\frac{1-a_4(1)}{1-2a_4(1)}) a_1(1) .$$

The conclusion obtained from this analysis of (4.70) and from the majorant principle is as follows. If

(4.71a) $\quad J(x(0),x(1))^{-1}$ exists and is bounded by $a_3(0)$

(4.71b) $\quad \max(\|x(1)-x(0)\|,\|x(0)-x(-1)\|) \leq a_1(1) = a_1(0)$

(4.71c) $\quad a_4(1) = a_3(0)\eta(a_1(0)+a_1(1)) < 1 - \frac{\sqrt{2}}{2}$ and $a_1(1) \leq (\frac{1-2a_4(1)}{1-a_4(1)})r$

(4.71d) the set $E \supseteq \bar{S}(x(0),r) \equiv \bar{S}$

then the entire sequence of iterates $\{x(n)\}$ for (4.69) under (4.58) and (4.62a) exists and converges to $x^* \in \bar{S}$. Furthermore, we have the estimate

$$\|x^*-x(n+1)\| \leq \left(\frac{1-a_1(1)}{1-2a_1(1)}\right) a_1(1) \cdot \prod_{j=1}^{n} \left(\frac{a_4(j)}{1-a_4(j)}\right)$$

where $a_4(n+1) = \frac{1}{2} \dfrac{a_4(n)}{(1-a_4(n))^2}$. Finally, $Fx^* = 0$ from (4.60), and

$$R_p(\{x(n)\}) < 1 \text{ with } p = \frac{1+\sqrt{5}}{2}$$

from (4.60) and Theorem 3.6.

Sample Process (4.65). Bittner has used for (4.64b)

$$\zeta_n = \frac{a_5(n)L(a_1(n))^\alpha}{1+\alpha}, \quad n = 0,1,\ldots,$$

$$a_5(n) = \prod_{j=0}^{n-1} a_3(j), \quad n = 1,2,\ldots; \ a_5(0) \text{ given}$$

But then

$$(4.72) \qquad a_5(n) = a_3(n-1)a_5(n-1) = \frac{a_5(n-1)}{1-a_5(n-1)L(a_1(n-1))^\alpha}, \quad n = 1,2,\ldots$$

and where $a_1(0) = a_1(1)$ is given.

$$(4.73) \qquad a_1(n+1) = 1\Big/\left(\frac{(1-a_5(n)L(a_1(n))^{\alpha})}{(1+\alpha)}\right)\left(1/(1-a_5(n)L(a_1(n))^{\alpha})\right)$$

$$\cdot(a_5(n)L(a_1(n))^{\alpha}+a_5(n-1)L(a_1(n-1))^{\alpha})\left(\frac{a_1(n)}{1+\alpha}\right), \ n=1,2,\ldots \ .$$

The equation can be expanded to yield

$$a_5(n) = \frac{a_5(0)}{1-a_5(0)L\sum_{j=0}^{n-1}(a_1(j))^{\alpha}}, \ n = 1,2,\ldots \ ;$$

which reduces (4.72), (4.73) to a single equation of the form of Section 2. The initial data $\hat{a}_1(0),\hat{a}_1(1),\hat{a}_5(0),\hat{a}_5(1)$ can be determined from Theorem 3.8 in such a way that $\{a_1(n)\}$ is summable. In that case, we have from Corollary 3.8,

$$O_R \geqslant \frac{1+\sqrt{1+4\alpha}}{2}$$

For the iterative process the conclusion is that if $x(0),x(-1)$ satisfy

$$(4.74a) \qquad \|x(j)-x(j-1)\| \leqslant \hat{a}_1(j), \ j = 0,1,$$

$$(4.74b) \qquad \|F'(x(0))^{-1}F'(x(-1))\| \leqslant \hat{a}_5(1) \ ,$$

$$(4.74c) \qquad \zeta_j \leqslant \frac{\hat{a}_5(j)L(a_1(j))^{\alpha}}{1+\alpha}, \ j = 0,1,$$

$$(4.74d) \qquad E \supseteq \bar{S}(x(-1), \sum_{j=0}^{\infty} a(j))$$

then the iterates $\{x(n)\}$ exist, remain in E, and converge to some x^*. The usual error estimates determined from (4.48) with $d_j = a_1(j)$ hold.

We give two final examples which serve as illustrations of further characteristics of the extended majorant principle. The first is a semi-local theorem for the combined method of Schmidt and Schwetlick [1968] considered in Section 1. This example, which appears to be new, shows that a rather complicated process can be attacked with relative ease when a majorizing system of difference equations is used. The second example concerns a process for which convergence has been proved by Kivistek [1961].

<u>Example 4.8</u>. For the iterative process (4.20) suppose that J,F satisfy the conditions (4.52) with $\alpha = 1$. If we assume that $\|F'(x)-J(x,y)\| \preccurlyeq M\|x-y\|$ for all $x,y \in \bar{S}(y(0),r)$, $r > 0$, then we obtain the majorizing system

$$\text{(4.75a)} \qquad a_1(n,\nu+1) = a_2(n,\nu)a_3(n)$$

$$\text{(4.75b)} \qquad a_1(n,0) = a_1(n-1,k+1) \text{ for } n = 1,2,\ldots, \text{ and } a_1(0,0),\ a_1(0,1) \text{ given,}$$

$$\text{(4.75c)} \qquad a_2(n,\nu) = \frac{L}{2}(a_1(n,\nu))^2 + La_1(n,\nu)\left(\sum_{j=1}^{\nu-1} a_1(n,j)\right) + M\left(\sum_{j=1}^{k+1} a_1(n-1,j)\right)a_1(n,\nu) \text{ for } n=0,1,\ldots;\ \nu = 1,2,\ldots,k$$

$$\text{(4.75d)} \qquad a_2(n,0) = \frac{3L}{2}(a_1(n-1,k+1))^2 + M\left(\sum_{j=1}^{k+1} a_1(n-1,j)\right)a_1(n-1,k+1) \quad \text{for } n = 1,2,\ldots ,$$

$$(4.75e) \qquad a_3(n) = \frac{a_3(0)}{1-a_3(0)(M(\sum_{j=1}^{k+1} a_1(n-1,j)+a_1(0,0))+L(\sum_{i=0}^{n-1} \sum_{j=1}^{k+1} a_1(i,j))}$$

for $n = 1,2,\ldots;\ a_3(0)$ given

As above, Banach's lemma has been used for (4.75e) and the estimates are

$$\|x_{n,\nu+1}-x_{n,\nu}\| \leq \|J(y(n),y(n-1))^{-1}\|\|Fx_{n,\nu}\| \ ,$$

$$\|Fx_{n,\nu}\| \leq \|Fx_{n,\nu}-Fx_{n,\nu-1}-F'(x_{n,\nu-1})(x_{n,\nu}-x_{n,\nu-1})\|$$

$$+ \|F'(x_{n,\nu-1})-J(y(n),y(n-1))\|\|x_{n,\nu}-x_{n,\nu-1}\|$$

$$\leq \frac{L}{2}\|x_{n,\nu}-x_{n,\nu-1}\|^2 + L\|x_{n,\nu}-x_{n,\nu-1}\|\|y(n)-x_{n,\nu-1}\|$$

$$+ M\|y(n)-y(n-1)\|\|x_{n,\nu}-x_{n,\nu-1}\| \ ,$$

$$\|J(y(n),y(n-1))^{-1}\| \leq \frac{\|J(y(0),y(-1))^{-1}\|}{1-\|J(y(0),y(-1))^{-1}\|\|J(y(n),y(n-1))-J(y(0),y(-1))\|}$$

$$\|J(y(n),y(n-1))-J(y(0),y(-1))\| \leq M(\|y(n)-y(n-1)\| + \|y(0)-y(-1)\|)$$

$$+ L\|y(n)-y(0)\|$$

The relations then are

$$\|y(0)-y(-1)\| \leq a_1(0,0),\|x_{0,1}-y(0)\| \leq a_1(0,1),$$

$$\|x_{n,\nu+1}-x_{n,\nu}\| \leq a_1(n,\nu+1)$$

$$\|Fx_{n,\nu}\| \leq a_2(n,\nu)$$

$$\|J(y(n),y(n-1))\| \leq a_3(n)$$

An induction argument, based on the idea of Theorem 3.8, gives the following conclusion for the system (4.75). Let $\lambda \in (0,1)$ and choose $a_1(0,0), a_1(0,1)$, $a_3(0)$ such that

$$c_0 = a_3(0)\left(Ma_1(0,0)+ka_1(0,1) + \frac{La_1(0,1)}{1-\lambda}\right) < 1$$

$$d_0 = \frac{a_3(0)}{1-c_0}\ ,\ d_0\left(\frac{L}{2} + Lk + Mk\right)\cdot a_1(0,1) \leq \lambda < 1,\ \frac{a_1(0,1)}{1-\lambda} \leq r\ .$$

Then $a_3(n) \leq d_0$ and $a_1(n,\nu+1) \leq \lambda a_1(n,\nu)$ for $n = 0,1,\ldots;\ \nu = 0,1,\ldots,$ $(n,\nu) \neq (0,0)$. Moreover

$$\sum_{i=0}^{n} \sum_{j=1}^{k+1} a_1(i,j) \leq \frac{a_1(0,1)}{1-\lambda} \leq r\ .$$

Applied to the iterative process, this yields convergence of $y(n) \in \bar{S}(0,r)$ to some $y^* \in \bar{S}$ when $y(0)$, $y(-1)$ are so chosen that

$$\|J(y(0),y(-1))^{-1}\| \leq a_3(0)\ ,$$

$$\|y(0)-y(-1)\| \leq a_1(0,0),\ \|x_{0,1}-y(0)\| \leq a_1(0,1).$$

Moreover, from (4.75d), $Fy^* = 0$ and $O_R(\mathcal{J},y^*) \geq \frac{1+\sqrt{4k+5}}{2}$ from Section 1, as well as

$$\|y(n)-y^*\| \leq \sum_{i=n}^{\infty} \sum_{j=1}^{k+1} a_1(i,j)$$

Example 4.9. A semi-local convergence theorem for the following iterative process for solving $F(x) = 0$, $F:X \to X$, has been given by Kivistek [1961]:

$$(4.76) \qquad x(n+1) = x(n) - \frac{(Fx(n), F'(x(n))Fx(n))}{\|F'(x(n))Fx(n)\|^2} F(x(n)), \quad n=0,1,\ldots .$$

Under the assumptions that F is twice F-differentiable and

$$(4.77a) \qquad \|Fx(0)\| \leq a_2(0) ,$$

$$(4.77b) \qquad \|(F'(x(0))h,h)\| \geq \frac{1}{a_3(0)} \|h\|^2 \text{ for all } h \in X ,$$

$$(4.77c) \qquad \|F'(x)\| \leq A, \|F''(x)\| \leq B \text{ for all } x \in \bar{S}(x(0),r) ,$$

the estimates

$$\|x(n+1)-x(n)\| \leq \frac{\|Fx(n)\|^2}{\|F'(x(n))Fx(n)\|} \leq a_3(n)\|Fx(n)\|$$

$$(4.78) \qquad \|Fx(n)\| \leq \|Fx(n-1)+F'(x(n-1))(x(n)-x(n-1))\| + \frac{1}{2} B\|x(n)-x(n-1)\|^2$$

$$\leq \left[\left(1 - \frac{1}{a_3^2(n)A^2}\right)^{1/2} + \frac{1}{2} B\, a_3^2(n)\|Fx(n-1)\|\right]\|Fx(n-1)\|$$

$$a_3(n+1) \leq \frac{a_3(n)}{1-a_3^2(n)B\|Fx(n)\|}$$

yield the majorizing system

$$a_1(n+1) = a_3(n)a_2(n) ,$$

$$a_2(n) = \left[\left(1- \frac{1}{a_3^2(n)A^2}\right)^{1/2} + \frac{1}{2} Ba_3^2(n)a_2(n-1)\right] a_2(n-1)$$

$$a_3(n) = \frac{a_3(n-1)}{1-a_3^2(n-1)Ba_2(n-1)}$$

Notice that the quantity $a_1(n)$ which bounds $\|x(n)-x(n-1)\|$ does not appear in the last two equations. Kivistek has shown that convergence occurs for $A^2Ba_2(0)a_3^4(0) < \frac{1}{9}$.

The fact that this process (4.76) is not an approximate Newton method as well as the estimates (4.78) seem to indicate that no improvement can be made in estimating $\|x(n+1)-x(n)\|$ in terms of $\|x(n)-x(n-1)\|$, etc. as is the basis of Section 2. Hence, the claim may be made that the methods of Section 2 cannot yield results as general as those of this section.

REFERENCES

Baluev, A., On the Method of Chaplygin, Dokl. Akad. Nauk SSSR 83 (1952), 781-784.

Bellman, R., On the Boundedness of Solutions of Nonlinear Differential and Difference Equations, Trans. AMS 62 (1947), 357-386.

Bittner, L., Mehrpunktverfahren zur Auflösung von Gleichungssystemen, Z. Angew. Math. Mech. 43 (1963), 111-126.

Chen, K., Generalization of Steffensen's Method for Operator Equations in Banach Space, Comment. Math. Univ. Carolinae 5 (1964), 47-77.

Coffman, C., Asymptotic Behavior of Solutions of Ordinary Difference Equations, Trans. AMS 110 (1964), 22-51.

Coppel, W., Stability and Asymptotic Behavior of Differential Equations, Heath, (1965).

Dunford, N., and Schwartz, J., Linear Operators I, Interscience, (1958).

Hahn, W., Über die Anwendung der Methode von Ljapunov auf Differenzengleichungen, Math. Annal. 136 (1958), 430-441.

Johnson, L., and Scholz, D., On Steffensen's Method, SIAM J. Num. Anal. 5 (1968), 296-302.

Kantorovich, L., Functional Analysis and Applied Mathematics (Russian), Uspehi Mat. Nauk 3 (1948), 89-185; translated by C. Benster as Nat. Bur. Standards Report 1509, (1952).

Kantorovich, L., On Newton's Method, Trudy Mat. Inst. Steklov 28 (1949), 104.

Kivistek, L., A Modification of the Minimum-Residual Iteration Method for the Solution of Equations Involving Nonlinear Operators, Dokl. Akad. Nauk 136 (1961), 22.

Ortega, J., The Newton-Kantorovich Theorem, Amer. Math. Monthly 75 (1968), 658-660.

Ortega, J., and Rheinboldt, W., Local and Global Convergence of Generalized Linear Iterations, in Numerical Solution of Nonlinear Problems, J. Ortega and W. Rheinboldt, Editors, SIAM, (1970a).

Ortega, J., and Rheinboldt, W., Iterative Solution of Nonlinear Equations in Several Variables, in press, Academic Press, (1970).

Ortega, J., and Rockoff, M., Nonlinear Difference Equations and Gauss-Seidel Type Iterative Methods, SIAM J. Numer. Anal. 3 (1966), 497-513.

Ostrowski, A., Solution of Equations and Systems of Equations, Academic Press, (1960), Second Edition (1966).

Panov, A., The Behavior of the Solutions of Difference Equations Near a Fixed Point (Russian), Izv. Vyss. Ucebn. Zaved. Matematika 12 (1959), 174-183.

Panov, A., Qualitative Behavior of the Trajectories of a System of Difference Equations in the Neighborhood of a Singularity (Russian), Izv. Vyss. Ucebn. Zaved. Matematika 40 (1964), 111-115.

Perron, O., Über Stabilität und Asymptotisches Verhalten der Lösungen eines Systems endlicher Differenzengleichungen, J. Reine Angew. Math. 161 (1929), 41-64.

Rall, L., Convergence of the Newton Process to Multiple Solutions, Num. Math. 9 (1966), 23-37.

Rheinboldt, W., A Unified Convergence Theory for a Class of Iterative Processes, SIAM J. Num. Anal. 5 (1968), 42-63.

Schauder, J., Der Fixpunktsatz in Funktionalräumen, Studia Math. 2 (1930), 171-180.

Schmidt, J., Eine Übertragung der Regula Falsi auf Gleichungen in Banach-Räumen, Z. Angew. Math. Mech. 41 (1961), T61-T63 and 43 (1963), 1-8, 97-110.

Schmidt, J., Ein Konvergenzsatz für Iterationsverfahren, Math. Nachr. 37 (1968), 67-81.

Schmidt, J., and Schwetlick, H., Ableitungsfreie Verfahren mit höherer Konvergenzgeschwindigkeit, Computing 3 (1968), 215-226.

Smith, R., Sufficient Conditions for Stability of a Solution of Difference Equations, Duke Math. J. 33 (1966), 725-734.

Ta Li, Die Stabilitätsfrage bei Differenzengleichungen, Acta Math. 63 (1934), 99-141.

Traub, J., Iterative Methods for the Solution of Equations, Prentice-Hall, (1964).

Voigt, R., Rates of Convergence for Iterative Methods for Nonlinear Systems of Equations, Ph.D. Thesis, University of Maryland, (1969).