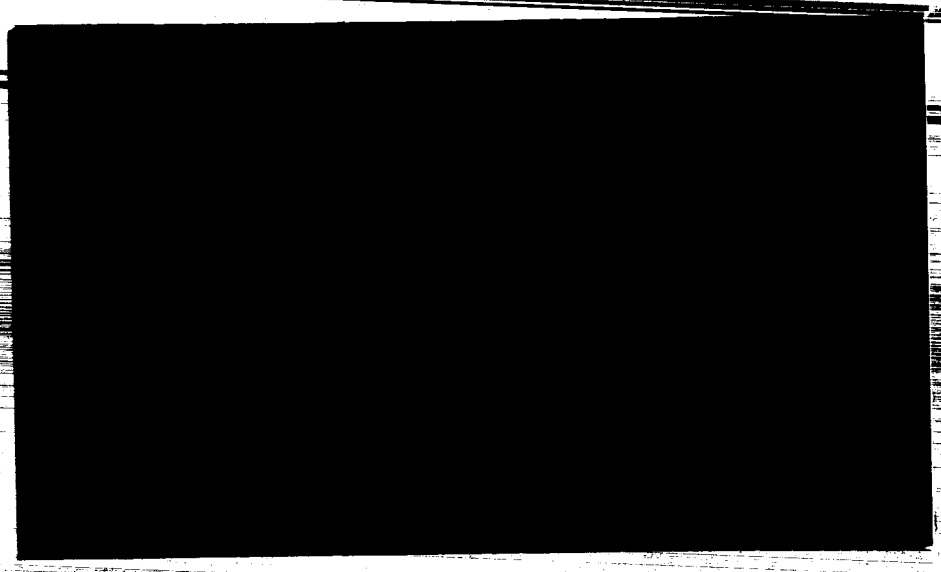

UNIVERSITY OF MARYLAND
COMPUTER SCIENCE CENTER

COLLEGE PARK, MARYLAND

Technical Report TR-65-20 August 1965
NsG-398

Non-linear Difference Equations and Gauss-Seidel Type Iterative Methods

by

James M. Ortega
Research Assistant Professor
and
Maxine L. Rockoff
Research Associate
Computer Science Center
University of Maryland
College Park, Maryland

The work reported here was in part supported by the National Aeronautics and Space Administration under Grant NsG-398 to the Computer Science Center of the University of Maryland.

CONTENTS

ABSTRACT

33281

An important class of methods for approximating solutions of non-linear systems of equations are the Gauss-Seidel or relaxation processes. The problem of obtaining asymptotic rates of convergence of these methods is treated here by linearization about a solution of the system. This leads to the study of the asymptotic behavior of solutions of perturbed linear difference equations and estimates for the decay rate of such solutions are obtained. These results are then applied to more general difference equations and, in particular, to the difference equations of typical Gauss-Seidel processes. This gives a precise determination of the asymptotic rate of convergence of these processes and is a generalization of known results for linear systems of equations. Application is made to a particular class of non-linear systems arising from mildly non-linear elliptic boundary value problems. In particular, estimates are given for optimum overrelaxation parameters and the results of numerical experiments are presented.

Author

Non-linear Difference Equations and Gauss-Seidel Type Iterative Methods

By James M. Ortega and Maxine L. Rockoff

1. Introduction

Several authors ([1]-[7]) have recently considered Gauss-Seidel type iterative processes (i.e. relaxation processes) for the approximation of solutions of a system of non-linear equations:

$$(1.1) \quad f(x) = 0 \qquad (f_i(x_1,\ldots,x_n) = 0, \quad i=1,\ldots,n).$$

No discussion of the asymptotic rate of convergence of these processes has yet been given, however, and the purpose of this paper is to set forth a generalization of the linear theory as described, for example, in [8] and [9].

If x^* is a solution of (1.1) and

$$(1.2) \qquad x^{(k+1)} = h(x^{(k)}), \quad k=0,1,\ldots$$

is an iterative process such that $x^* = h(x^*)$, then expansion of h about x^* leads to the error equation:

$$(1.3) \qquad e^{(k+1)} = He^{(k)} + r(e^{(k)}), \quad k=0,1,\ldots,$$

where $H = h'(x^*)$ is the Jacobian matrix of h at x^* and $e^{(k)} = x^{(k)} - x^*$. In Section 2 we study the perturbed linear difference equation (1.3) under various assumptions on H, r and $e^{(0)}$. In particular, we obtain estimates of the form

(1.4) $$\beta_0 k^K \lambda^k \leq \|e^{(k)}\| \leq \beta_1 k^K \lambda^k, \quad k=1,2,\ldots, \ 0 < \lambda < 1,$$

where K is a fixed integer, β_0 and β_1 are constants depending on $e^{(0)}$, and $\lambda \equiv \rho(H)$ is the spectral radius of H. These estimates are related to qualitative results of Panov [10] and may be of interest in themselves.

The iterative processes with which we are concerned are not naturally of the form (1.2); rather the iterates satisfy a more general difference equation

(1.5) $$g(x^{(k+1)}, x^{(k)}) = 0, \quad k=0,1,\ldots,$$

where g may be non-linear in $x^{(k+1)}$ as well as $x^{(k)}$. In Section 3, we obtain results on the asymptotic behavior of solutions of (1.5) by means of the implicit function theorem together with our previous results for (1.3). Then, in Section 4, we consider some typical relaxation processes as applied to (1.1) and, under suitable assumptions on f and hence g, we conclude that the asymptotic rate of convergence is given by $-\ln[\rho(H)]$. Here H is the matrix $[g_x]^{-1} g_y$ evaluated at (x^*,x^*) and g_x and g_y are the partial Frechet derivatives (Jacobian matrices) of g with respect to the first and second vector variables. This generalizes the corresponding result of the linear theory.

Finally, in Section 5, we treat a particular class of non-linear systems that arise, for example, as discrete analogues of certain mildly non-linear elliptic boundary value problems. Here

we are able to make useful a priori comparisons between different iterative processes and, moreover, obtain estimates for optimum over-relaxation parameters. Some numerical experiments supporting the theory are also included.

2. Pertrubed Linear Difference Equations

The following result is related to a theorem of Ostrowski [11, p. 119] on points of attraction.

Theorem 1: Suppose H is an nxn matrix with spectral radius $\rho(H) \equiv \lambda < 1$ and let $r: \Omega \subset R^n \to R^n$ denote a mapping of an open neighborhood Ω of the origin of R^n into R^n such that[1]

$$\frac{\|r(x)\|}{\|x\|} \to 0, \qquad \|x\| \to 0. \tag{2.1}$$

Then, given any constant δ with $\lambda < \lambda + \delta < 1$, there are an open neighborhood Ω' of the origin and a constant d such that for any initial vector $e^{(0)} \in \Omega'$, a solution $\{e^{(k)}\}$ of the difference equation (1.3) exists and satisfies

$$\|e^{(k)}\| \leqq d(\lambda+\delta)^k \|e^{(0)}\|, \quad k=0,1\ldots \tag{2.2}$$

Proof: Let (see, e.g., [12, p.46]) $\| \ \|'$ be a norm such that

$$\|H\|' \leqq \lambda + \delta/2.$$

Then there are positive constants c_1 and c_2 such that for all $x \in R^n$

$$c_1\|x\|' \leqq \|x\| \leqq c_2\|x\|', \tag{2.3}$$

and, by (2.1), there exists a $\sigma > 0$ so that

---

[1] Throughout the paper $\| \ \|$ denotes an arbitrary vector norm as well as the corresponding operator (lub) norm.

(2.4) $$\|r(x)\| \leq \frac{\delta}{2d}\|x\|, \quad \|x\| \leq \sigma, \quad d = c_2/c_1.$$

Now let $\Omega' = \{x \mid \|x\|' < \sigma/c_2\}$. Then $x\in\Omega'$ implies $\|x\|\leq\sigma$ and, by (2.3) and (2.4),

$$\|r(x)\|' \leq \frac{1}{c_1}\|r(x)\| \leq \frac{\delta}{2c_1d}\|x\| \leq \frac{\delta}{2}\|x\|', \quad x \in \Omega';$$

hence

$$\|e^{(1)}\|' \leq \|H\|'\,\|e^{(0)}\|' + \|r(e^{(0)})\|' \leq (\lambda+\delta)\,\|e^{(0)}\|', \quad e^{(0)} \in \Omega'.$$

Therefore $e^{(1)} \in \Omega'$ and, by induction, it follows that $e^{(k)} \in \Omega'$, $k=2,3,\ldots$, and

$$\|e^{(k)}\|' \leq (\lambda+\delta)^k\|e^{(0)}\|', \quad k=1,2,\ldots .$$

Then (2.2) follows using (2.3) and the proof is complete.

<u>Corollary:</u> Let H and r satisfy the conditions of Theorem 1 and let $\{e^{(k)}\}$ be any solution of (1.3) such that $e^{(k)}\to 0$ as $k\to\infty$. Then

(2.5) $$\limsup_{k\to\infty} \|e^{(k)}\|^{\frac{1}{k}} \leq \rho(H).$$

<u>Proof:</u> Since $e^{(k)}\to 0$, $\gamma = \limsup \|e^{(k)}\|^{\frac{1}{k}}$ exists and $\gamma\leq 1$. Suppose $\gamma>\lambda=\rho(H)$. Let $\delta=(\gamma-\lambda)/2$ and let Ω' be the neighborhood given by Theorem 1 for this δ. Then there exists an index k_o such that $e^{(k_o)} \in \Omega'$ and, by Theorem 1, the sequence $\{e^{(k)} \mid k = k_o, k_o+1, \ldots\}$ satisfies (2.2). Hence

$$\gamma = \limsup_{k\to\infty} \|e^{(k)}\|^{\frac{1}{k}} = \limsup_{k \geq k_o} \|\, e^{(k)}\|^{\frac{1}{k}}$$

$$= \limsup_{k \geq k_o} [d(\lambda+\delta)^k\,\|e^{(0)}\|]^{\frac{1}{k}} = \lambda + \delta < \gamma,$$

which is a contradiction.

Under additional assumptions on r we are able to sharpen the estimate (2.2).

Theorem 2: Suppose that $0<\lambda\equiv\rho(H)<1$ and r satisfies

$$\|r(x)\| \leq c\|x\|^{1+\epsilon}, \quad \epsilon > 0, \quad 0 < c < \infty, \quad x \in \Omega. \tag{2.6}$$

Let K+1 be the dimension of the largest Jordan block of H associated with an eigenvalue of modulus λ. Then for any solution $\{e^{(k)}\}$ of (1.3) such that $e^{(k)}\to 0$ as $k\to\infty$, there exists a constant β such that

$$\|e^{(k)}\| \leq \beta k^K \lambda^k, \quad k=1,2,\ldots . \tag{2.7}$$

Moreover, there exists a solution $\{e^{(k)}\}$ of (1.3) with $e^{(k)}\to 0$ as $k\to\infty$, and a constant $\beta_0>0$ such that

$$\|e^{(k)}\| \geq \beta_0 k^K \lambda^k, \quad k=0,1,\ldots . \tag{2.8}$$

Proof: Choose γ such that $\lambda^{\epsilon}<\gamma<1$, set $\delta=(\lambda\gamma)^{\frac{1}{1+\epsilon}}-\lambda>0$ and let Ω' be the neighborhood given by Theorem 1 for this δ. Then if $\{e^{(k)}\}$ is any solution of (1.3) such that $e^{(k)}\to 0$ as $k\to\infty$, there is a k_o for which $e^{(k)}\in\Omega'$, $k\geq k_o$. Consequently, we may assume, without loss of generality, that $e^{(k)}\in\Omega'$, $k=0,1,\ldots$.

Since $\{e^{(k)}\}$ is a solution of (1.3), we have

$$e^{(k)} = H^k e(0) + \sum_{j=0}^{k-1} H^j r(e^{(k-j-1)}) \equiv H^k e(0) + u^{(k)}, \quad k=1,2,\ldots . \tag{2.9}$$

Moreover, from (2.2), (2.6), and the definition of δ, it follows that

$$\|r(e^{(k)})\| \leq c[d(\lambda+\delta)^k \|e^{(0)}\|]^{1+\varepsilon} = d_o\lambda^{k}\gamma^{k}\|e^{(0)}\|^{1+\varepsilon}, \quad k=0,1,\ldots$$

where $d_o = cd^{1+\varepsilon}$. Therefore, since there exists a constant q, $1\leq q<\infty$, such that

$$\|H^j\| \leq qj^{k}\lambda^{j}, \quad j=1,2,\ldots ,$$

(see, e.g., [12, p. 183]), we obtain the estimate

$$\|u^{(k)}\| \leq \sum_{j=0}^{k-1} \|H^j\| \, \|r(e^{(k-j-1)})\|$$

$$\leq qd_o\|e^{(0)}\|^{1+\varepsilon}\sum\lambda^{j}j^{K}\lambda^{k-j-1}\gamma^{k-j-1}$$

$$\leq \mu k^{K}\lambda^{k}\|e^{(0)}\|^{1+\varepsilon}, \quad k=1,\ldots ,$$

where

$$\mu = qd_o(1-\gamma)^{-1}\lambda^{-1}. \tag{2.11}$$

Hence

$$\|e^{(k)}\| \leq \|H^k\| \, \|e^{(0)}\| + \|u^{(k)}\| \leq \lambda^{k}k^{K}[q \, \|e^{(0)}\| + \mu\|e^{(0)}\|^{1+\varepsilon}], \quad k=1,\ldots,$$

and this establishes (2.7).

Now let λ_1 be an eigenvalue of H of modulus λ with which there is associated a Jordan block of dimension $K+1$. Then there exists a principal vector v, with $\|v\| = 1$, and an eigenvector v_o such that

$$\lambda_1^{-k}k^{-K}H^{k}v\to v_o \neq 0, \quad k\to\infty. \tag{2.12}$$

Choose $e^{(0)}=\theta v \in \Omega'$ where θ satisfies $0 < \mu\theta^{\varepsilon} \leq \frac{1}{2}\|v_o\|$, with μ given

by (2.11). Then, by (2.10), we have

$$\|e^{(k)}\| = \lambda^{k}k^{K}\,\|\,\lambda_1^{-k}\,k^{-K}H^{k}e(0) + \lambda_1^{-k}\,k^{-K}u(k)\,\|$$

$$\geq \lambda^{k}k^{K}(\theta\|\lambda_1^{-k}\,k^{-K}H^{k}v\| - \mu\|\theta v\|^{1+\varepsilon})$$

$$\geq \lambda^{k}k^{K}\theta(\|\lambda_1^{-k}\,k^{-K}H^{k}v\| - \tfrac{1}{2}\|v_o\|), \quad k=0,1,\ldots$$

and it follows, using (2.12), that for some integer k_1,

$$\|e^{(k)}\| \geq \lambda^{k}k^{K}(\theta/4)\|v_o\|, \quad k\geq k_1. \tag{2.13}$$

But then there exists a β_0 with $0<\beta_0\leq(\theta/4)\|v_o\|$ such that (2.8) holds. For, otherwise, we would have $e^{(k_o)}=0$ for some $k_o<k_1$, and this would imply $e^{(k)}=0$ for all $k\geq k_o$.

Corollary: Let H and r satisfy the conditions of Theorem 2 with the exception that $\rho(H)=0$ is not excluded. Then

$$\sup\{\limsup_{k\to\infty}\,\|\,e^{(k)}\|^{\frac{1}{k}}\} = \rho(H), \tag{2.14}$$

where the supremum is taken over all solutions $\{e^{(k)}\}$ of (1.3) such that $e^{(k)}\to 0$ as $k\to\infty$.

Proof: By the corollary to Theorem 1, we have

$$\sup\{\limsup_{k\to\infty}\,\|\,e^{(k)}\|^{\frac{1}{k}}\} \leq \rho(H),$$

and if $\rho(H)=0$, equality holds in (2.14). If $\rho(H)\neq 0$, then the reverse inequality follows from (2.8). This completes the proof.

We conjecture that (2.14) holds if r satisfies only (2.1). However simple examples show that the estimates (2.7) and (2.8) do not hold even when n = 1.

3. Application to More General Difference Equations

We now apply the results of the previous section to difference equations of the form (1.2) and for notational convenience we define the following class of functions.

Definition 1: Let $g:D_g \subset R^n \times R^n \to R^n$ denote a mapping from a domain D_g in the product space $R^n \times R^n$ into R^n. Suppose $S \subset R^n$ is a non-empty set such that $S \times S \subset D_g$ and let x^* be a point in the closure of S. Then g is defined to belong to the class of functions $\mathfrak{F}(s;x^*)$ if for each initial vector $x^{(0)} \in S$, the difference equation (1.2) has a unique solution $\{x^{(k)}, k=0,1,\ldots\} \subset S$ which converges to x^*. Each solution $\{x^{(k)}\} \subset S$ of (1.3) will be called a g-sequence on S.

We note that if $g(x,y)=x-Hy-r(y)$, where H and r satisfy the conditions of Theorem 1, then $g \in \mathfrak{F}(\Omega';0)$ with Ω' the neighborhood given by the theorem.

If V is an open set in D_g, we write $g \in C^1(V)$ if all $2n^2$ partial derivatives of the components of g exist and are continuous on V. If, in addition, all $4n^3$ second partial derivatives exist and are continuous on V, we write $g \in C^2(V)$. Finally, we denote by g_x and g_y the nxn matrices:

$$g_x(x,y) = \left(\frac{\partial g_i}{\partial x_j}(x,y) \right), \quad g_y(x,y) = \left(\frac{\partial g_i}{\partial y_j}(x,y) \right)$$

Theorem 3: Let S' be an open neighborhood of a point $x^* \in R^n$.

Assume that $g \in C^1(S' \times S')$, g_x^{-1} is defined and continuous on $S' \times S'$ and $g(x^*,x^*) = 0$. Define

$$H \equiv -\left[g_x(x^*,x^*)\right]^{-1} g_y(x^*,x^*) \tag{3.1}$$

and suppose that $\rho(H) \equiv \lambda < 1$. Then there exists a neighborhood S of x^* such that $g \in \mathfrak{F}(S;x^*)$; moreover, each g-sequence on S satisfies

$$\limsup_{k\to\infty} \| x^{(k)} - x^* \|^{\frac{1}{k}} \leq \lambda. \tag{3.2}$$

If, in addition,

$$g \in C^2(S' \times S'), \tag{3.3}$$

and $\lambda \neq 0$, then for each g-sequence $\{x^{(k)}\}$ on S, there is a constant β such that

$$\|x^{(k)} - x^*\| \leq \beta \lambda^{k_k K}, \quad k=1,2,... \tag{3.4}$$

where K is defined as in Theorem 2. Moreover there exist a g-sequence $\{x^{(k)}\}$ on S and constant $\beta_0 > 0$ such that

$$\|x^{(k)} - x^*\| \geqq \beta_0 \lambda^{k_k K}, \quad k=1,2,... \quad . \tag{3.5}$$

Proof: By the implicit function theorem (see, e.g.,[13, p. 265]) there exist an open neighborhood T of x^* and a unique function h defined on T with the property that $(h(y),y) \subset S' \times S'$ and $g(h(y),y) = 0$ for all $y \in T$; i.e., the equation $g(x,y) = 0$ has a unique solution $x \in S'$ for all $y \in T$. Moreover $x^* = h(x^*)$, h' exists and is continuous on T and $h'(x^*) = H$.

Therefore, if $S \subset T$ is a neighborhood of x^*, then every g-sequence

$\{x^{(k)}\}$ on S satisfies $x^{(k+1)} = h(x^{(k)})$ and, consequently,

$$(3.6)\quad x^{(k+1)} - x^* = H(x^{(k)} - x^*) + r(x^{(k)} - x^*), \quad k=0,1,\ldots$$

where

$$r(x-x^*) = h(x)-h(x^*) - H(x-x^*).$$

Conversely, every sequence $x^{(k)} \to x^*$ satisfying (3.6) is a g-sequence on S. Then, since r satisfies (2.1) with $\Omega = \{x-x^* \mid x\in T\}$, all the conditions of Theorem 1 hold for (3.6). Hence $g \in \mathfrak{I}(S;x^*)$ where, if Ω' is the set given by Theorem 1, $S = \{x \mid x-x^* \in \Omega'\}$. Moreover, (3.2) follows immediately from the Corollary to Theorem 1.

Now if (3.3) is satisfied, it also follows from the implicit function theorem that h is twice continuously differentiable on T. Let T' be an open neighborhood of x^* such that $\bar{T}' \subset T$. Then the Taylor remainder theorem implies that the function r of (3.6) satisfies

$$\|r(x-x^*)\| \leqq c\|x-x^*\|^2, \quad x \in T', \ c<\infty.$$

Hence r satisfies (2.6) with $\Omega = \{x \mid x+x^* \in T'\}$ and therefore (3.4) and (3.5) are restatements of the conclusions (2.7) and (2.8) of Theorem 2.

4. Iterative Processes and Asymptotic Rates of Convergence

The results of the previous sections now permit the determination of asymptotic rates of convergence of certain iterative processes applied to the approximation of solutions of (1.1). The processes we consider are those whose iterates satisfy a difference equation of the form (1.2) and while this includes, by definition, all one step methods, our attention will be focused on relaxation type processes. The following, whose associated difference equations are defined by (4.1), (4.2) and (4.3), are typical.

I. The Jacobi-Newton-Process (J-N-P). (See [2] and [6].)

$$(4.1)\quad g_{I,i}(x^{(k+1)},x^{(k)}) \equiv \frac{\partial f_i}{\partial x_i}(x^{(k)})\left[x_i^{(k+1)}-x_i^{(k)}\right] + f_i(x^{(k)}) = 0, \quad \begin{matrix} i=1,\ldots,n, \\ k=0,1,\ldots . \end{matrix}$$

II. The Extrapolatated-Gauss-Seidel-Newton-Process (E-G-S-N-P). (See [2], [3, p. 136], [4] - [7].)

$$(4.2)\quad g_{II,\omega,i}(x^{(k+1)},x^{(k)}) \equiv \frac{\partial f_i}{\partial x_i}(x^{(k,i)})\left[x_i^{(k+1)}-x_i^{(k)}\right] + \omega f_i(x^{(k,i)} = 0, \quad \begin{matrix} i=1,\ldots,n, \\ k=0,1,\ldots \end{matrix}$$

where

$$x^{(k,i)} = (x_1^{(k+1)},\ldots,x_{i-1}^{(k+1)}, x_i^{(k)},\ldots, x_n^{(k)}).$$

III. The non-linear Gauss-Seidel (Liebmann) Process (G-S-P). (See [1], [3, p. 135] and [7].)

$$(4.3)\quad g_{III,i}(x^{(k+1)},x^{(k)}) \equiv f_i(x_1^{(k+1)},\ldots,x_i^{(k+1)},x_{i+1}^{(k)},\ldots,x_n^{(k)}) = 0, \quad \begin{matrix} i=1,\ldots,n, \\ k=0,1,\ldots . \end{matrix}$$

For a more general discussion of these and several related processes, including block forms of I, II and III, see [14].

Note that in order to carry out the G-S-P, a non-linear equation in a single variable must be solved for each i and k. Note also that if f is linear, i.e. $f(x) = Ax-b$, then I, II and III reduce to the usual cyclic Jacobi, cyclic Extrapolated Gauss-Seidel (SOR) and cyclic Gauss-Seidel processes respectively.

Now let $g \in \mathfrak{I}(S;x^*)$ (Definition 1) and define the quantity

$$\alpha(g;S) \equiv \sup \{\limsup_{k\to\infty} \|x^{(k)}-x^*\|^{\frac{1}{k}}\}, \tag{4.4}$$

where the supremum is over all g-sequences $\{x^{(k)}\}$ on S. Since $x^{(k)}-x^* \to 0$ as $k\to\infty$, the quantity in brackets in (4.4) always exists and in bounded by unity. Moreover, an argument using (2.3) easily shows that $\alpha(g;S)$ is independent of the particular norm. Therefore $\alpha(g;S)$ is well-defined and satisfies

$$0 \leqq \alpha(g;S) \leqq 1. \tag{4.5}$$

In general, of course, $\alpha(g;S)$ is dependent on S. However, the case of greatest interest is when S is a neighborhood of x^* and we have the following result.

<u>Lemma 1:</u> Let S and S' be open sets such that $x^* \in S \cap S'$ and assume that $g \in \mathfrak{I}(S;x^*) \cap \mathfrak{I}(S';x^*)$. Then $\alpha(g;S) = \alpha(g;S')$.

<u>Proof:</u> If $\{x^{(k)}\}$ is a g-sequence on S, then for some k_o, $x^{(k_o)} \in S \cap S' \subset S'$ and $\{x^{(k)} \mid k = k_o, k_o+1, \ldots\}$ is a g-sequence on S'. But since

$$\limsup_{k\to\infty} \|x^{(k)}-x^*\|^{\frac{1}{k}} = \limsup_{k\geqq k_o} \|x^{(k)}-x^*\|^{\frac{1}{k}}$$

it follows that $\alpha(g;S) \leqq \alpha(g;S')$. The reverse inequality follows by the symmetry of the argument.

It is clear that, without additional conditions on g, it is impossible to conclude more than (4.5) about the magnitude of $\alpha(g;S)$. Suppose, however, that $g(x,y) = x-Hy+c$, where c is a constant vector and $\rho(H)<1$. Then stated in our terms, a fundamental result in the theory of iterative processes for systems of linear equations is that there exists a unique $x^* \in R^n$ such that $g \in \mathfrak{F}(R^n;x^*)$, and

$$\alpha(g;R^n) = \rho(H). \tag{4.6}$$

The conclusions of Theorem 3 now provide a generalization of (4.6).

Corollary to Theorem 3: Let g satisfy the conditions of Theorem 3 with the exception of (3.3). Then there exists a neighborhood S of x^*, such that $g \in \mathfrak{F}(S;x^*)$ and

$$\alpha(g;S) \leqq \rho(H), \tag{4.7}$$

where H is defined by (3.1). If, in addition, g satisfies (3.3), then

$$\alpha(g;S) = \rho(H). \tag{4.8}$$

Proof: (4.7) follows immediately from (3.2). If $\rho(H)\neq 0$, (4.8) follows from (4.7) and (3.5). If $\rho(H) = 0$, then, by (4.7), $\alpha(g;S) = 0$ so that (4.8) holds in any case. This completes the proof.

We conjecture that (4.8) holds without the additional assumption (3.3); see the remarks following the Corollary to Theorem 2.

Clearly $\alpha(g;S)$ can be considered a measure of the slowest possible asymptotic convergence of any g-sequence to x^*. When applied to general one-step iterative processes, however, $\alpha(g;S)$ may give only a minimal amount of information. For example, suppose there exist constants γ and $p \geq 1$ such that for all g-sequences on S

$$\|x^{(k)}-x^*\| \leq \gamma \|x^{(k-1)}-x^*\|^p, \quad k=0,1,\ldots \quad . \tag{4.9}$$

If $p>1$ then $\alpha(g;S) = 0$; hence $\alpha(g;S)$ provides no basis for the comparison of different "higher order" methods. However, our interest here is in processes for which $p = 1$. In this case, $\alpha(g;S) \leq \gamma$ and the determination of $\alpha(g;S)$ may yield a sharper asymptotic convergence measure than the geometric estimate provided by (4.9).

In the sequel, it will be convenient, and consistent with the linear theory, to adopt the following terminology. If $g \in \mathfrak{F}(S;x^*)$, where S is a neighborhood of x^*, and (1.2) is the difference equation of an iterative process, then we shall define

$$R(g)= -\ln[\alpha(g;S)] \tag{4.10}$$

to be the <u>asymptotic rate of convergence</u> of the process (on S and hence, by Lemma 1, on any other neighborhood S' of x^* for which $g \in \mathfrak{F}(S';x^*)$). For example, if g_I is given by (4.1) and $g_I \in \mathfrak{F}(S;x^*)$ then we say that $R(g_I)$ is the asymptotic rate of convergence of the Jacobi-Newton-Process (or the a.r.c. of the J-N-P, for short.)

Moreover, if $g_I \in \mathfrak{F}(S;x^*)$ and $g_{III} \in \mathfrak{F}(S;x^*)$, we say that the J-N-P is <u>asymptotically faster</u> than the G-S-P if $R(g_I) > R(g_{III})$ or <u>asymptotically equivalent</u> if $R(g_I) = R(g_{III})$. Similar statements apply for comparison of the other processes.

We now consider the functions g_I, $g_{II,\omega}$ and g_{III} of (4.1), (4.2) and (4.3) in more detail. Assume that the function f of (1.1) satisfies, for some $x^* \in R^n$,

$$(4.11) \quad f \in C^3(S'), \quad S' = \{x \mid |x_i - x_i^*| < \delta \leqq +\infty, \quad i=1,\ldots n\},$$

and

$$(4.12) \quad f(x^*) = 0.$$

That is, f is defined and three times continuously differentiable on an open cube S' and the system $f(x) = 0$ has a solution $x^* \in S'$. Then the functions $g_I, g_{II,\omega}$ and g_{III} are defined and twice continuously differentiable on S' and S'; furthermore

$$(4.13) \quad g_I(x^*,x^*) = g_{II,\omega}(x^*,x^*) = g_{III}(x^*,x^*) = 0.$$

Let the Jacobian matrix f'(x) be written as

$$(4.14) \quad f'(x) = D(x) - E(x) - F(x), \quad x \in S',$$

where D, E and F are diagonal, strictly lower triangular and strictly upper triangular respectively, and assume that

$$(4.15) \quad \det[D(x)] \neq 0, \quad x \in S'.$$

Then g_x^{-1} exists and is continuous on S'xS' for each of g_I, $g_{II,\omega}$ and g_{III}. Moreover, if we denote by H_I, $H_{II,\omega}$ and H_{III}, the

matrices $-[g_x(x^*,x^*)]^{-1}\, g_y(x^*,x^*)$ for g_I, $g_{II,\omega}$ and g_{III} respectively, then

$$H_I = [D(x^*)]^{-1}[E(x^*) + F(x^*)], \tag{4.16}$$

$$H_{II,\omega} = [D(x^*) - \omega E(x^*)]^{-1}[(1-\omega)D(x^*) + \omega F(x^*)], \tag{4.17}$$

and

$$H_{III} = [D(x^*) - E(x^*)]^{-1}F(x^*) = H_{II,1}. \tag{4.18}$$

Now assume that

$$\rho(H_I) < 1, \quad \rho(H_{II,\omega}) < 1, \quad \rho(H_{III}) < 1. \tag{4.19}$$

Then all of the conditions of Theorem 3 are satisfied for each of g_I, $g_{II,\omega}$ and g_{III}, and the following theorem is simply a restatement, using (4.10), of the conclusions of Theorem 3 and its Corollary.

Theorem 4: Let f satisfy the conditions (4.11), (4.12) and (4.15) and assume (4.19) holds. Then there exist neighborhoods S_1, $S_{2,\omega}$ and S_3 of x* such that

$$g_I \in \mathfrak{F}(S_1;x^*),\ g_{II,\omega} \in \mathfrak{F}(S_{2,\omega};x^*),\ \text{and}\ g_{III} \in \mathfrak{F}(S_3;x^*).$$

Moreover,

$$R(g_I) = -\ln[\rho(H_I)], \quad R(g_{II,\omega}) = -\ln[\rho(H_{II,\omega})],$$

and

$$R(g_{III}) = -\ln[\rho(H_{III})] = R(g_{II,1}). \tag{4.20}$$

Therefore, under the assumptions of Theorem 4, the asymptotic rates of convergence of the processes I, II and III are determined

by the spectral radii of the matrices H_I, $H_{II,\omega}$ and H_{III} respectively; this generalizes the corresponding result in the linear theory. Of course, to compare these spectral radii and, moreover, to verify that the conditions (4.19) hold, we need to assume more about f; this we shall do in the following section. Note, however, that Theorem 4 already yields one interesting, although perhaps intuitively obvious, comparison, namely, that under the conditions of the theorem, the G-S-N-P and the G-S-P are asymptotically equivalent. (Of course, this says nothing about the <u>global</u> behavior of the processes; however, see the numerical experiments of the next section.)

5. Applications

The conditions imposed upon the function f in the previous section are admittedly stringent; in particular, the verification of (4.19) requires, in general, that the solution x^* of (1.1) be known. However, for certain functions arising in practice it may be quite simple to ascertain that all of these conditions are fulfilled and to make useful a priori comparisons of different iterative processes. In this section we continue the analysis of g_I, $g_{II,\omega}$ and g_{III} for a particular class of equations $f(x) = 0$ which arise, for example, as the discrete analogues of certain mildly non-linear elliptic boundary value problems of the type $\Delta u = \sigma(u)$ (see, e.g., [4]). The development is based upon the corresponding theory for linear problems as described, for example, in Forsythe and Wasow [8] and Varga [9] and we refer to these references for definitions of the terminology used here.

Consider the system of equations

$$f(x) \equiv Ax + \emptyset(x) = 0, \tag{5.1}$$

and assume that

(5.2) A is irreducibly diagonally dominant,

and

$$A = D - E - F, \quad D \geq 0, \quad E + F \geq 0, \tag{5.3}$$

where D, E, and F are diagonal, strictly lower triangular and strictly upper triangular respectively. About the non-linear

function ϕ, we assume that each component ϕ_i is a function of a single variable and that

$$\phi_i(x) \equiv \phi_i(x_i), \quad i=1,\ldots,n, \tag{5.4}$$

$$\phi_i \in C^3(-\infty,+\infty), \quad i=1,\ldots,n, \tag{5.5}$$

and

$$\phi_i'(t) \geq 0, \ -\infty<t<\infty, \quad i=1,\ldots,n. \tag{5.6}$$

Under these hypotheses it may be shown that (5.1) has a unique solution x*. (See, e.g., [4], or for symmetric A, [7]). Moreover, it is easily verified that (4.11) holds with $S' = R^n$. Furthermore, $f'(x) = D + \phi'(x) - E - F$ where, by (5.4) and (5.6), $\phi'(x)$ is a non-negative diagonal matrix. Hence using (5.2) and (5.3), (4.15) holds and, in order to apply Theorem 4, it remains to examine the spectral radii of the matrices H_I and $H_{II,\omega}$ of (4.16) and (4.17).

For notational convenience we define

$$H_I(Z) = Z^{-1}(E + F), \tag{5.7}$$

and

$$H_{II,\omega}(Z) = (Z-\omega E)^{-1}\,[(1-\omega)Z + \omega F], \tag{5.8}$$

for all non-singular diagonal matrices Z. Then $H_I = H_I(D + \phi'(x^*))$ and $H_{II,\omega} = H_{II,\omega}(D + \phi'(x^*))$ while $H_I(D)$ and $H_{II,\omega}(D)$ are the Jacobi and SOR matrices for the linear problem

$$Ax = 0. \tag{5.9}$$

Now for any diagonal matrix $D_1 \geqq D$ we observe directly from (5.7) and (5.8), using (5.3), that

$$0 \leqq H_I(D_1) \leqq H_I(D), \quad D_1 \geqq D$$

and

$$0 \leqq H_{II,\omega}(D_1) \leqq H_{II,\omega}(D), \quad D_1 \geqq D, \quad 0 < \omega \leqq 1.$$

Hence, by (5.2) and the Perron-Frobenius theory, we conclude that

$$\rho[H_I] \leqq \rho[H_I(D)] < 1, \tag{5.10}$$

and

$$\rho[H_{II,\omega}] \leqq \rho[H_{II,\omega}(D)] < 1, \quad 0 < \omega \leqq 1, \tag{5.11}$$

where equality holds in (5.10) and (5.11) if and only if $\emptyset'(x^*) = 0$. Moreover, by the Stein-Rosenberg Theorem (see [9, p. 68]),

$$0 < \rho(H_{II,1}) < \rho(H_I). \tag{5.12}$$

Finally it can be shown that (see [9, p. 92])

$$\rho(H_{II,\omega_2}) < \rho(H_{II,\omega_1}), \quad 0 < \omega_1 < w_2 \leqq 1. \tag{5.13}$$

Now, using (5.10) and (5.11), Theorem 4 may be applied and our conclusions thus far may be summarized as follows. Each function g_I, $g_{II,\omega}$, $0 < \omega \leqq 1$, and g_{III} is contained in $\mathfrak{F}(S';x^*)$ where S' is some neighborhood depending on the particular process. (Actually, under additional assumptions on $\emptyset$ or A, much more is known about the global convergence properties of the Gauss-Seidel and Gauss-Seidel-Newton processes. See [4] and [7]). By (5.10),

the asymptotic rate of convergence (a. r. c.) of the Jacobi-Newton process for (5.1) is not less than that of the Jacobi process for (5.9), and is greater if $\emptyset'(x^*) \neq 0$. Likewise, by (5.11), the a. r. c. of the G-S-N-P and (since, $H_{III} = H_{II,1}$) the a. r. c. of the G-S-P for (5.1) are equal and are not less than that of the G-S-P for (5.9). Moreover, by (5.12), the a. r. c. of the G-S-N-P for (5.1) is greater than that of the J-N-P for (5.1). Finally, by (5.13), the a. r. c. of the E-G-S-N-P is a monotonically increasing function of ω for $\omega \leqq 1$. Hence, under the conditions imposed on f, the only reason that could be advanced for under relaxing the E-G-S-N-P, is to improve the global convergence.

5.2 We now consider $g_{II,\omega}$ for $\omega > 1$ and in addition to the assumptions already made about (5.1) we add the following:

(5.14) A is 2-cyclic and consistently ordered,

(5.15) A is symmetric.

We can then apply all of the theory first developed by Young [15] for the point successive overrelaxation of linear problems.

We first note that if Z is any diagonal matrix, (5.14) implies that Z - E - F is again 2-cyclic and consistently ordered. Now let

(5.16) $Z \geqq D$.

Then $H_I(Z)$ exists and, by (5.15), is similar to a symmetric matrix;

hence $H_I(Z)$ has real eigenvalues. Furthermore the fundamental relationship

$$[\nu(Z) + \omega - 1]^2 = \nu(Z)\,\omega^2\,[\mu(Z)]^2 \tag{5.17}$$

holds between the eigenvalues $\mu(Z)$ of $H_I(Z)$ and the eigenvalues $\nu(Z)$ of $H_{II,\omega}(Z)$. Thus if $\omega_{opt}(Z)$ denotes that value of ω which minimizes $\rho[H_{II,\omega}(Z)]$, then it follows from (5.17) that

$$\omega_{opt}(Z) = \frac{2}{1 + \{1 - \rho^2[H_I(Z)]\}^{\frac{1}{2}}} \tag{5.18}$$

$$\rho[H_{II,\omega}(Z)] = \tfrac{1}{4}\{\omega\rho[H_I(Z)] + (\omega^2\rho^2[H_I(Z)] - 4(\omega - 1))^{\frac{1}{2}}\}^2, \quad 0 < \omega \leqq \omega_{opt}, \tag{5.19}$$

and

$$\rho[H_{II,\omega}(Z)] = \omega - 1, \quad \omega_{opt} \leqq \omega \leqq 2. \tag{5.20}$$

Now let D_1 be a diagonal matrix such that

$$D + \emptyset'(x^*) \leqq D_1. \tag{5.21}$$

Then

$$\rho[H_I(D_1)] \leqq \rho[H_I] \leqq \rho[H_I(D)], \tag{5.22}$$

and, using (5.18) through (5.22), the spectral radii of $H_{II,\omega}(D_1)$, $H_{II,\omega}$ and $H_{II,\omega}(D)$ are related as shown in Diagram A.

These results can be summarized as follows. Under the assumptions (5.1) through (5.6), (5.14) and (5.15), $g_{II,\omega}$ applied to (5.1) for any $0 < \omega < 2$ satisfies the conditions of Theorem 4; hence for each $0 < \omega < 2$, there is a neighborhood S_ω such that $g_{II,\omega} \in \mathfrak{F}(S_\omega; x^*)$. Moreover, there exists an optimum ω, such that

(5.23) $R(g_{II,\omega_{opt}}) > R(g_{II,\omega})$, $0<\omega<2$, $\omega\neq\omega_{opt} \equiv \omega_{opt}(D + \emptyset'(x^*))$,

and, for the matrix D_1 of (5.21), ω_{opt} satisfies

(5.24) $$\omega_{opt}(D_1) \leqq \omega_{opt} \leqq \omega_{opt}(D).$$

Finally,

$$R(g_{II,\omega}) \geqq -\ln\{\rho[H_{II,\omega}(D)]\}, \quad 0 < \omega < 2;$$

that is, the a. r. c. of the E-G-S-N-P is at least as great when the process is applied to (5.1) as when it is applied to (5.9).

5.3 Along with the E-G-S-N-P there is, of course, an extrapolated non-linear Gauss-Seidel-Process (E-G-S-P) whose iterates satisfy the difference equation:

(5.25) $$f_i(x_1^{(k+1)},\ldots,x_{i-1}^{(k+1)},x_i^{(k)}-\omega^{-1}[x_i^{(k+1)}-x_i^{(k)}],x_{i+1}^{(k)},\ldots,x_n^{(k)})=0, \begin{matrix} i=1,\ldots,n. \\ k=0,1,\ldots \end{matrix}$$

In the previous section we showed that the asymptotic rates of convergence of the E-G-S-P and the E-G-S-N-P were the same if $\omega = 1$. In an analogous way it may be shown that they are the same for all ω; hence the conclusions of this section for the E-G-S-N-P apply verbatim to the E-G-S-P.

5.4 In Table I we summarize the results of some numerical experiments in which the E-G-S-N-P and the E-G-S-P were applied to discrete analogues of the following boundary value problems (the domain of each problem is $\Omega = [0,1]\ x[0,1]$ and $\dot{\Omega}$ is the boundary of Ω):

a) $\Delta x = e^x$; $\quad x(s,t) = s + 2t, \quad (s,t) \in \dot{\Omega}$

b) $\Delta x = x^3$; $x(s,t) = s + 2t$, $(s,t) \in \dot{\Omega}$

c) $\Delta x = x^3$; $x(s,t) = 0$, $(s,t) \in \dot{\Omega}$

d) $\Delta x = 0$, $x(s,t) = 0$, $(s,t) \in \dot{\Omega}$

Experiments involving problem a) have been reported in [4].

In each case, the Laplacian operator was approximated by using the usual 5-point formula with h = .05. This gives for each problem a system of 361 equations of the form:

$$(5.26)\quad 4x_{i,j} - x_{i,j-1} - x_{i,j+1} - x_{i-1,j} - x_{i+1,j} + h^2\sigma(x_{i,j}) = 0,$$
$$i,j = 1,\ldots,N-1, N=h^{-1},$$

where $\sigma(x) = e^x, x^3, x^3$ and 0 for a), b), c), and d) respectively and the ordering of the grid points is left to right, bottom to top. It is easy to verify that each of the systems (5.26) satisfies the conditions (5.1)-(5.6), (5.14) and (5.15); hence the theory of this section applies.

Clearly c) and d) have the unique solutions $x \equiv 0$ and the same is true of their discrete analogues (5.26). Therefore, $\emptyset'(x^*) = 0$ and the optimum ω for the E-G-S-N-P and E-G-S-P applied to c) is the same as that for the linear problem d). Using (5.18) and the known eigenvalues of the corresponding Jacobi matrix, it may be computed exactly:

$$(5.27)\qquad \omega_{opt,\ c,d.} \doteq 1.73.$$

By (5.24), (5.27) also gives an upper bound for $\omega_{opt,a}$ and $\omega_{opt,b}$.

In fact, after approximate solutions of (5.26) were obtained, we computed

$$\omega_{opt,a} \doteq 1.705, \ \omega_{opt,b} \doteq 1.701$$

by approximating the spectral radii of the corresponding Jacobi matrices $H_I(D + \emptyset'(x^*))$.

We used starting approximations determined as follows:

1. The boundary conditions were linearly interpolated at each grid point.
2. A function value of 50 was taken at each grid point.

For problems c) and d) only 2. is applicable.

In Table I we tabulate the number of iterations and seconds of machine time required to satisfy a convergence test of $\|x^{(k+1)} - x^{(k)}\|_\infty \leqq 10^{-6}$. All calculations were done on an IBM-7094-II. The machine time includes a certain amount of printing and does not reflect an optimal program; however, it is the relative times that are of interest.

The E-G-S-P was carried out by first solving the individual equations of (5.26) by Newton's method (with a convergence criterion of 10^{-6} for the residual) and then extrapolating.

Table I shows that the use of the theoretical optimum ω produces the fastest convergence in most cases; this tends to substantiate the theory. However it is clear that the use of the optimum $\omega(\doteq 1.73)$ for the linear problem is almost as good. Perhaps the

most interesting fact is that the E-G-S-P and the E-G-S-N-P tend to take almost exactly the same number of iterations. Our theory predicts that asymptotically this is true but, of course, does not indicate that this should be true globally. Note, however, that for the problem a)2, the E-G-S-N-P took significantly more iterations than the E-G-S-P. This was due to the poor starting approximation and the resulting magnitude of e^{50}. In any case the E-G-S-N-P was always $1\frac{1}{2}$ to 2 times as fast as the E-G-S-P.

We also ran all of the test problems for $\omega = 1$; convergence was never achieved before the program was stopped after 300 iterations. Finally, we run numerous problems for $h = .1$ (81 equations) but there was no qualitative difference in the results.

ACKNOWLEDGEMENTS

We would like to express our thanks to Professor W. Rheinboldt of the University of Maryland for numerous enlightening discussions and to Messrs. R. Elkin and C. Henderson who performed the calculations of Section 5.

Problem ω	1.57	1.63	1.68	1.705	1.73	1.80
a)1 EGS	92/37	77/31	61/25	50/21	48/19	62/24
EGSN	92/22	77/19	61/15	50/13	48/12	62/15
a)2 EGS	133/65	111/56	88/46	73/40	70/38	88/47
EGSN	138/33	118/29	97/23	82/20	82/20	105/25
b)1 EGS	85/27	71/23	56/18	46/15	46/15	62/23
EGSN	85/17	71/14	56/11	46/9	46/9	62/12
b)2 EGS	88/30	88/30	74/26	63/23	67/24	89/32
EGSN	106/20	87/17	65/13	55/11	62/12	83/16
c)2 EGS	130/49	90/35	89/34	79/28	69/24	88/33
EGSN	138/27	114/22	91/18	80/14	62/12	82/16
d)2 EGS	157/22	133/18	107/15	94/13	72/10	87/12

Table 1. (iterations/seconds)

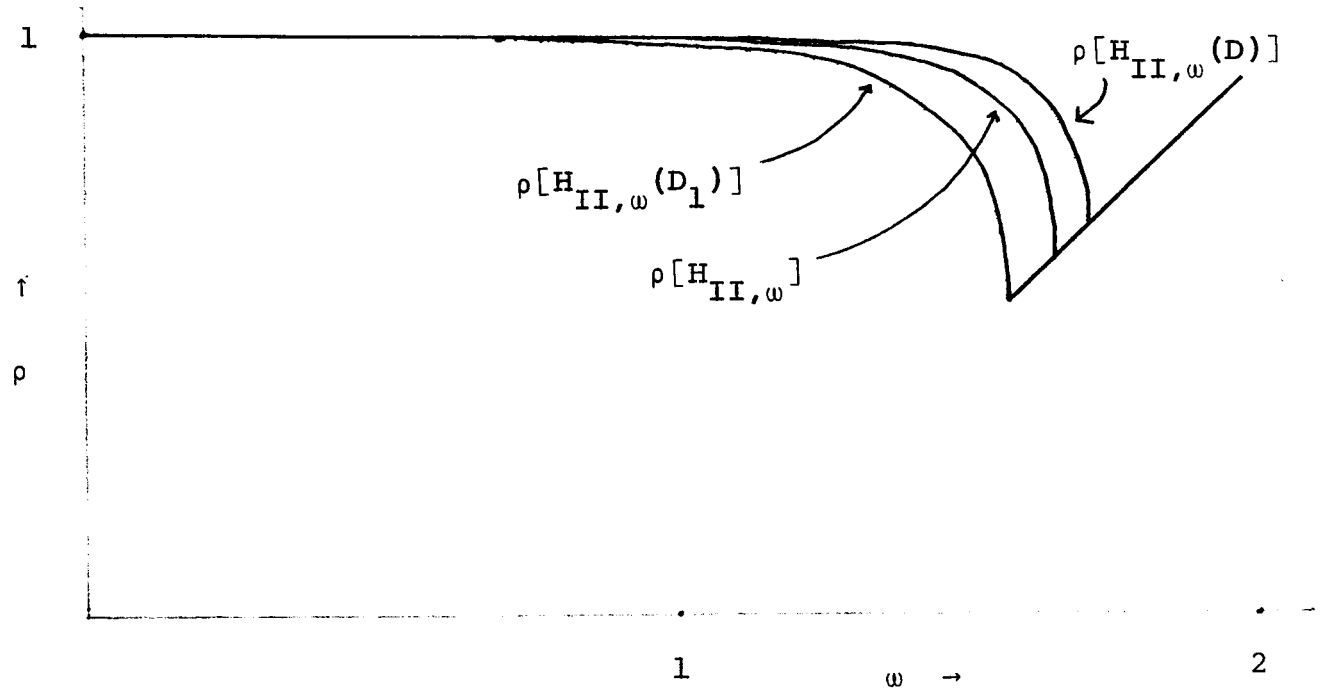

Diagram A

REFERENCES

[1] L. Bers, On mildly non-linear partial difference equations of elliptic type, J. Res. Nat. Bur. Standards, 51(1953), pp. 229-236.

[2] C. Bryan, An iterative method for solving non-linear systems of equations, Dissertation, University of Arizona, 1963.

[3] D. Greenspan, Introductory Numerical Analysis of Elliptic Boundary Value Problems, Harper and Row, New York, 1965.

[4] D. Greenspan and S. Parter, Mildly non-linear elliptic partial differential equations and their numerical solution, II, Numer. Math., 7 (1965), pp. 129-146.

[5] D. Greenspan and M. Yohe, On the approximate solution of $\Delta u = F(u)$, Comm. A.C.M., 6(1963), pp. 564-568.

[6] H. Lieberstein, Overrelaxation for non-linear elliptic partial differential equations, Math. Res. Cen. Tech. Rept. 80, University of Wisconsin, 1959.

[7] S. Schechter, Iteration methods for non-linear problems, Trans. Amer. Math. Soc., 104(1962), pp. 179-189.

[8] G. Forsythe and W. Wasow, Finite Difference Methods for Partial Differential Equations, Wiley, New York, 1960.

[9] R. Varga, Matrix Iterative Analysis, Prentice Hall, Englewood Cliffs, N. J., 1962.

[10] A. Panov, The behavior of the solutions of difference equations near a fixed point, Izv. Vyss. Ucebn. Zaved. Matematika, 12 (1959), pp. 174-183. (Russian)

[11] A. Ostrowski, Solution of Equations and Systems of Equations, Academic Press, New York, 1960.

[12] A. Householder, The Theory of Matrices in Numerical Analysis, Blaisdell, New York, 1964.

[13] J. Dieudonné, Foundations of Modern Analysis, Academic Press, New York, 1960.

[14] W. Rheinboldt and J. Ortega, Iterative Methods for Non-linear Operator Equations, to be published by Blaisdell Publishing Company, New York.

[15] D. Young, Iterative methods for solving partial difference equations of elliptic type, Trans. Amer. Math. Soc., 76 (1954), pp. 92-111.