SU326 P30-20

RICHARDSON'S NON-STATIONARY MATRIX ITERATIVE PROCEDURE

BY

R. S. ANDERSSEN AND G. H. GOLUB

STAN-CS-72-304

AUGUST 1972

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

SU326 P30-20

Richardson's Non-Stationary Matrix Iterative Procedure

by

R. S. Anderssen* and G. H. Golub**

* Computer Centre, Australian National University, Canberra, A.C.T. 2600, Australia.

** Computer Science Department, Stanford University, Stanford, California, 94305, U.S.A. The work of this author was in part supported by the AEC.

RICHARDSON'S NON-STATIONARY MATRIX ITERATIVE PROCEDURE

R. S. Anderssen and G. H. Golub

Computer Centre, Australian National University and
Computer Science Department, Stanford University

ABSTRACT

Because of its simplicity, Richardson's non-stationary iterative scheme is a potentially powerful method for the solution of (linear) operator equations. However, its general application' has more or less been blocked by

(a) the problem of constructing polynomials, which deviate least from zero on the spectrum of the given operator, and which are required for the determination of the iteration parameters of the non-stationary method, and

(b) the instability of this scheme with respect to rounding error effects.

Recently, these difficulties were examined in two Russian papers. In the first, Lebedev [15] constructed polynomials which deviate least from zero on a set of subintervals of the real axis which contains the spectrum of the given operator. In the second, Lebedev and Finogenov [11] gave an ordering for the iteration parameters of the non-stationary Richardson scheme which makes it a stable numerical process. Translation of these two papers appear as Appendices 1 and 2, respectively, in

this report. The body of the report represents an examination of the properties of Richardson'; non-stationary scheme and the pertinence of the two mentioned papers along with the results of numerical experimentation testing the actual implementation of the procedures given in them.

61. INTRODUCTION

Of the many methods proposed for the iterative solution of the linear system

$$\underset{\sim}{A}\underset{\sim}{u} = \underset{\sim}{f} \qquad (1.1)$$

where A is an $n \times n$ non-singular matrix, the simplest is the non-stationary method of Richardson [1], viz.

$$\underset{\sim}{u}^{(k+1)} = \underset{\sim}{u}^{(k)} - \alpha_k(\underset{\sim}{A}\underset{\sim}{u}^{(k)} - \underset{\sim}{f}) \quad (k = 1, 2, \ldots), \qquad (1.2)$$

where $\alpha_1, \alpha_2, \ldots$ are iteration parameters with $\alpha_k = \alpha_{k-N} (k > N)$. The given fixed integer N is called the period of the iteration (1.2). Though Richardson's original method was the stationary version of (1.2), viz.

$$\underset{\sim}{u}^{(k+1)} = \underset{\sim}{u}^{(k)} - \alpha(\underset{\sim}{A}\underset{\sim}{u}^{(k)} - \underset{\sim}{f}) \quad (\alpha = \text{const.}, k = 1, 2, \ldots), \qquad (1.3)$$

he observed that better convergence could be obtained if α_n varied with n. Along with other methods, Young [2] examined its use for the iterative solution of elliptic partial differential equations. In sub-sequent papers [3],[4], its numerical properties were examined in some detail. It was shown that

(i) in a certain sense, the choice

$$\alpha_k = 2[a+b - (b-a)\cos((2k-1)\pi/2N)]^{-1} \quad (k = 1, 2, \ldots, N), \qquad (1.4)$$

where $a \leq \lambda_j(A) \leq b$ $(j = 1,2, \ldots, n)$, gives optimal convergence properties to the $\underset{\sim}{u}^{(k)}$ defined by (1.2),

(ii) the method, at least when using the optimal choice (1.4) with the $\{\alpha_k\}$ in the order given there, is very sensitive to rounding error effects.

Independent studies of the method have been made by Birman [5] and Gavurin [6]. Since the publication of Young's paper [3], the method has been discussed in different contexts by Stiefel [7], Engeli et al [8; Chapter II], Golub and Varga [9] and Young [10]. An important result obtained during this period is that due to Young [10; §11.4], which shows that the non-stationary forms of (1.2) are related to semi-iterative forms of the stationary procedure (1.3):

Definition 1.1. Let $\{\gamma_{ij}\} = \{\gamma_{ij}\ (i = 1, 2, \ldots\ j = 1, 2, \ldots)\ i)\}$ denote a set of coefficients which satisfy

$$\sum_{j=1}^{i} \gamma_{ij} = 1 \qquad (i = 1, 2, \ldots) . \tag{1.5}$$

Given a sequence $\{\underset{\sim}{x}_j\} = \{\underset{\sim}{x}_j, j = 1, 2, \ldots)$ generated by (1.3), then

$$\underset{\sim}{y}_k = \sum_{j=1}^{k} \gamma_{kj}\, \underset{\sim}{x}_j \qquad (k = 1,2, \ldots) \tag{1.6}$$

defines a semi-iterative method with respect to the linear stationary procedure (1.3).

Theorem 1.1. Let A be a non-singular matrix. Given $\{\alpha_j\}$, there exists, for $\alpha \neq 0$, $\{\gamma_{ij}\}$ such that (1.5) is valid and such that the semi-iterative method based on (1.6) and the $\{\gamma_{ij}\}$ yields the same iterates for the starting vector $\underset{\sim}{v}^{(1)} = \underset{\sim}{u}^{(1)}$ as does Richardson's method

based on the $\{\alpha_k\}$ and the starting vector $u^{(1)}$. Conversely, given $\{\gamma_{ij}\}$ such that (1.5) is valid, then, for any $\alpha \neq 0$ and any i, there exists a set $\{\alpha_j\}$, such that the $u^{(i)}$, as determined by Richardson's method based on the $\{\alpha_k\}$ and starting with $u^{(1)}$, is the same as $v^{(n)}$ determined by the semi-iterative method based on (1.3) and the $\{\gamma_{ij}\}$ with $v^{(1)} = u^{(1)}$.

Compared with the semi-iterative method based on (1.3), which requires the triangular array $\{\gamma_{ij}\}$ defining (1.6) to be stored, Richardson's non-stationary scheme is the simpler to implement. However, its sensitivity to rounding error has more or less blocked its general use, especially since semi-iterative methods are less sensitive.

Closer examination of the Richardson scheme indicates that the mentioned sensitivity is a function of the order in which the $\{\alpha_k\}$ are taken. This was first observed by Young [3], [4] who examined a number of orderings and showed that some gave better results than others. However, the more fundamental question of the existence of orderings for which Richardson's method defines a stable numerical process was not examined.

We pause to mention further connections of Richardson's method (1.2) with other well known iterative techniques:

((i)) The optimal choice of the relaxation parameter for the SOR 'method is only known explicitly in special cases; e.g. the given positive definite matrix A has Property A {see Varga [19; Chapter 14]}. This is stronger than positive definiteness which is all that is required for the application of Richardson's non-stationary method.

((ii)) If the first order method (1.2) is replaced by the second order (Richardson) method.

$$\underset{\sim}{u}^{(k+1)} = \underset{\sim}{u}^{(k)} + \alpha\{A\underset{\sim}{u}^{(k)} + \underset{\sim}{g} - \underset{\sim}{u}^{(k-1)}\} + \beta\{\underset{\sim}{u}^{(k)} - \underset{\sim}{u}^{(k-1)}\} ,$$

then the problem of numerical instability is found to disappear {see Golub [18]}. The close connection between this second order method and the Chebyshev semi-iterative method, viz.

$$\underset{\sim}{u}^{(k+1)} = \omega_{k+1}\{A\underset{\sim}{u}^{(k)} + \underset{\sim}{g} - \underset{\sim}{u}^{(k-1)}\} + \underset{\sim}{u}^{(k-1)} \qquad (k > 1)$$

with

$$\omega_k = 1/(1 - \rho^2\omega_k/4),\ (k > 2),\ \omega_1 = 1,\ \omega_2 = 2/(2 - \rho^2),$$

where ρ denotes the spectral norm of A, has been examined in detail by Golub and Varga [19]. See also Varga [19] and Young [10; Chapter 16]. This semi-iterative method contains SOR as the special case $\omega_k = 2/(1+[1-\rho^2]^{\frac{1}{2}})$, $(k > 1)$.

((iii)) The advantage of either method mentioned in ((ii)) over a stable implementation of (1.2) is the choice of the iteration parameters. For the first order Richardson method, it is necessary to specify in advance the roots of the polynomial of degree N which deviate least from zero on o(A) for a given N. For the second order method, Golub and Varga [19] have shown that there exists, under a wide range of circumstances, an optimal choice of α and β, viz. $\alpha = 2/(1+[1-\mu^2]^{\frac{1}{2}})$ and $\beta = -1$. In the case of the Chebyshev semi-iterative method, the ω_k are generated sequentially as and when required.

Recently, Lebedev and Finogenov [11] {A translation is given in Appendix 2} examined this question and constructed an ordering of the α_k of (1.4) for which Richardson's method defines a stable numerical process. The construction of this ordering will be examined in §2, and the application of Richardson's method based on it for the solution of

different forms of Poisson's equation will be examined in §3. Lebedev and Finogenov did not examine whether their ordering is in any sense optimal or whether other orderings exist for which Richardson's method defines a stable numerical process.

The other difficulty associated with the efficient implementation of Richardson's method is the actual choice of the α_k. The choice based on (1.4) is more or less optimal if

(a) a and b are the exact lower and upper bounds for o(A), the spectrum of A, (assuming now that A is positive definite) and

(b) o(A) does not consist of widely separated disjoint subintervals in which the $\lambda_i(A)$ $\{i = 1, 2, \ldots n)$ lie.

In fact, the optimal choice of the α_k are the reciprocals of the roots of the polynomials of the form

$$P_N(t) = \prod_{k=1}^{N} (1 - \alpha_k t) \qquad (1.7)$$

with

$$P_N(0) = 1, \qquad (1.8)$$

which deviate least from zero on the set o(A). However, the actual construction of such polynomials is often blocked by

(α) the lack of knowledge about the structure of o(A) {the numerical determination of all the $\lambda_i(A)$ $(i = 1, 2, \ldots n)$ in general, involves more computation than the numerical evaluation of $A^{-1}f$}, and

(β) the fact that, for a given o(A), the construction of polynomials of the form (1.7) - (1.8) which deviate least from zero on o(A) can be a difficult, if not impossible, task.

For these reasons, the construction of the α_k for a given A has been based on the following approximate procedure:

1. Determine a region Ω

   ((i)) for which $\sigma(A) \in \Omega$, and

   ((ii)) such that the polynomials $P_N(t; \Omega)$ with $P_N(0; \Omega) = 1$ which deviate least from zero on Ω are known or can be constructed.

2. Set the α_k ($k = 1, \ldots, N$) as the reciprocals of the roots of the polynomial $P_N(t; \Omega)$.

The α_k of (1.4) correspond to the case when $\Omega = [a, b]$.

A number of authors, including Samokish [12], Achieser[13] and Markov [14], have examined cases when Ω is disjoint. The most recent analysis is that of Lebedev [15] {A translation is given in Appendix 1}, who examined the construction of polynomials $P_N(t; \Omega)$ when Ω consists of a number of disjoint subintervals of the real axis. This work is summarized for the two interval case in §4 and applied to a problem in §5.

§2. THE ORDERING OF LEBEDEV AND FINOGENW.

In this section, we describe the ordering of the iteration parameters of (1.4) which Lebedev and Finogenov [11] proved makes the numerical process defined by (1.2) stable.

For a given N, let $(\varphi_1, \varphi_2, \ldots, \varphi_N)$ denote the basic ordering of the iteration parameters defined by (1.4), viz.

$$\varphi_i = 2[b+a - (b-a)\cos((2i-1)\pi/2N)]^{-1} \quad (i = 1, 2, \ldots, N). \tag{2.1}$$

Then, different orderings of the $\{\varphi_k\}$ for the α_k in (1.2) can be defined as permutations on the set $(\varphi_1, \varphi_2, \ldots, \varphi_N)$. In particular, we define a one-to-one mapping between $(\varphi_1, \varphi_2, \ldots, \varphi_N)$ and $(\alpha_1, \alpha_2, \ldots, \alpha_N)$ with $\alpha_k = \varphi_{i_k}$ by the permutation

$$\nu_N = (i_1, i_2, \ldots, i_N). \tag{2.2}$$

Thus, any ordering of the iteration parameters (1.4) can be defined in terms of this permutation ν_N.

Let $N \in \{2^p, p = 0, 1, 2, \ldots\}$, $\nu_{2^0} = \nu_1 = (1)$, and

$$\nu_{2^{p-1}} = (j_1, j_2, \ldots, j_{2^{p-1}}) \tag{2.3}$$

Then, the permutation ν_N which defines the Lebedev-Finogenov ordering of (2.1) is constructed inductively with respect to (2.2) and $N \in \{2^p, p = 0, 1, 2, \ldots\}$ using

$$\kappa_{2^p} = (j_1, 2^{p+1}-j_1, j_2, 2^{p+1}-j_2, \bullet \ldots\ j_{2^{p-1}}, 2^{p+1}-j_{2^{p-1}}) . \quad (2.4)$$

In particular,

$$\kappa_2 = (1,2) ,$$

$$\kappa_4 = (1,4,2,3) ,$$

$$\kappa_8 = (1,8,4,5,2,7,3,6) ,$$

$$\kappa_{16} = (1,16,8,9,4,13,5,12,2,15,7,10,3,14,6,11) ,$$

$$\kappa_{32} = (1,32,16,17,8,25,9,24,4,29,13,20,5,28,12,21,2,31,15,18,7,26,$$
$$10,23,3,30,14,19,6,27,11,22).$$

An ALGOL procedure for generating κ_{2^p} for a given p is:

```
Procedure  Lebedev-Finogenov-Ordering (P, Kappa);
Value P;  Integer P;
Integer Array Kappa;
        Begin Comment For a given "P", this procedure generates the
        permutation "Kappa" of order 2↑P which defines the Lebedev-
        Finogenov ordering of period 2↑P. The array "Kappa" must be
        dimensioned externally with order 2↑P;
        Integer I, J, INT, INS;
        KAPPA [1]: = 1;
        INT: = 1;
        For I: = 1   step 1 until P do
             Begin INT: = 2 X INT;
             INS : = INT + 1;
             For J: = INT ÷ 2  step - 1 until 1 do
```

```
            Begin KAPPA [2 x J]: = INS - KAPPA [J];
            KAPPA [2 x J - 1]: = KAPPA [J]
            End
        End
    End Lebedev-Finogenov Ordering;
```

§3. APPLICATION OF THE LEBEDEV-FINOGENOV ORDERING.

In order to test the Lebedev-Finogenov ordering, we examine the type of matrix equation (1.1) for which iterative methods are best suited, viz. A is a sparse large rank matrix with $A\underset{\sim}{u}$ defined in a systematic manner which rules out the necessity to store A. In particular, we examine the following boundary value problem: using finite difference methods, construct a function $u(x,y)$, continuous on the unit square $S = \{(x,y);\ 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$ except possibly at the corner points, having first and second derivatives in the interior and satisfying Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -g(x,y)\ (u = u(x,y),\ (x,y) \in S), \tag{3.1}$$

and the boundary conditions

$$u(0,y) = \alpha(y),\ u(1,y) = \bar{\alpha}(y) \quad (0 \leq y \leq 1), \tag{3.2}$$

$$u(x,0) = p(x),\ u(x,1) = \bar{\beta}(x) \quad (0 < x < 1). \tag{3.3}$$

We introduce the grid

$$G = \{(ih,jh);\ i = 0, 1, \ldots, I,\ j = 0, 1, \ldots, I,\ Ih = 1\}$$

and the notation

$$u_{ij} = u(ih,jh) \quad ((ih,jh) \in G),$$

and we use the central difference approximations

$$\left[\frac{\partial^2 u(x,y)}{\partial x^2}\right]_{i,j} = (u_{i+1,j} - 2u_{i,j} + u_{i-1,j})/h^2 + O(h^2) \qquad (3.4)$$

$$\left[\frac{\partial^2 u(x,y)}{\partial y^2}\right]_{i,j} = (u_{i,j+1} - 2u_{i,j} + u_{i,j-1})/h^2 + O(h^2) \qquad (3.5)$$

which are valid if $\frac{\partial^4 u}{\partial x^4}$ and $\frac{\partial^4 u}{\partial y^4}$ exist and are bounded for (x,y) e Interior (S). Substitution of (3.4) and (3.5) in (3.1) yields, after neglecting truncation error, the following finite difference scheme for the approximate solution of (3.1)-(3.3):

$$4v_{ij} - v_{i-1,j} - v_{i,j+1} - v_{i,j-1} + h^2 g_{ij} = 0 \quad (i,j = 1, 2, \cdots, I-1), \qquad (3.6)$$

$$v_{0j} = \alpha_j, \quad v_{Ij} = \bar{\alpha}_j \qquad (j = 0, 1, \ldots, I), \qquad (3.7)$$

$$v_{i0} = \beta_i, \quad v_{iI} = \bar{\beta}_i \qquad (i = 1, 2, \ldots, I-1), \qquad (3.8)$$

where $\alpha_j = \alpha(jh)$, $\bar{\alpha}_j = \bar{\alpha}(jh)$, $\beta_i = \beta(ih)$ and $\bar{\beta}_i = \bar{\beta}(ih)$, and v_{ij} denotes the calculated value of u_{ij}.

Since (3.6)-(3.8) define a linear algebraic system of order $(I-2)^2$ for the determination of the finite difference solution (3.1)-(3.3) on the grid G, it can be solved by the non-stationary scheme (1.2). In fact, its implementation only involves the use of three matrix arrays

$$v_{ij}^{(1)}, \quad v_{ij}^{(0)} \quad \text{and} \quad g_{ij} \qquad (i, j = 0, 1, \ldots, I) \qquad (3.9)$$

in the following way:

(i) Set $v_{0j}^{(0)} = \alpha_j$, $v_{Ij}^{(0)} = \bar{\alpha}_j$, $v_{i0}^{(0)} = \beta_i$, $v_{iI}^{(0)} = \bar{\beta}_i$

$(j = 0, 1, 2, \ldots, I, \; i = 1, 2, \ldots, I-1)$,

$$v_{ij}^{(0)} = d_{ij} \qquad (i, j = 1, \ldots . I-1) \tag{3.10}$$

where the d_{ij} define the starting solution, and k = 1.

(ii) Compute

$$v_{ij}^{(1)} = v_{ij}^{(0)} - \alpha_k(r_{ij}) \text{ and } R = \max_{ij} |r_{ij}| \ (i,j=1,2,\ldots,I-1), \tag{3.11}$$

where

$$r_{ij} = 4v_{ij}^{(0)} - v_{i+1,j}^{(0)} - v_{i-1,j}^{(0)} - v_{i,j+1}^{(0)} - v_{i,j-1}^{(0)} + h^2 g_{ij}, \tag{3.12}$$

and then

$$v_{ij}^{(0)} = v_{ij}^{(1)} \qquad (i, j = 1, 2, \ldots . I - 1). \tag{3.13}$$

(iii) For a given positive value ϵ, if $R > \epsilon$, set $k = k + 1$ and return to (ii) then back to (iii); if $R < \epsilon$, stop as $u_{ij}^{(0)}$ $(i, j = 0, 1, \ldots . I)$ defines an approximate solution with the required accuracy. The final value of k gives the number of iterations required.

Note. Because (3.11) is only a three-level difference scheme, the storage requirement can be reduced to at most $(I+1)^2 + 3(I-1)$ -locations.

For the matrix defined by (3.6), the exact bounds for the eigenvalues of A are

$$a = \lambda_{min}(A) = 4(1-\cos \pi h), \ b = h,,(A) = 4(1+\cos \pi h). \tag{3.14}$$

Since the eigenvalues of A are dense in the interval [a,b], for sufficiently small h, it is not inapprdpriate to define Ω as the single interval [a,b]. The iteration parameters for (1.2) are therefore defined by (1.4). Using

§4. POLYNOMIALS WHICH DEVIATE LEAST FROM ZERO ON DISJOINT SUBINTERVALS OF THE HEAL LINE

For the general case, when

$$\Omega = \bigcup_{i=1}^{n} [a_{2i-1}, a_{2i}] \tag{4.1}$$

with $a_i < a_{i+1}$ $(i = 1, 2, \ldots, 2n-1)$ and $0 \notin [a_{2i-1}, a_{2i}]$ $(i=1, 2, \ldots n)$, Lebedev used a method of Markov [14] to construct the polynomial $P_N(t; \Omega)$, with $P_N(0;\Omega) = 1$, which deviates least from zero on Ω. The construction hinges on the validity of the following assumption which amounts to a restriction on the way in which the a_i $(i=1, 2, \ldots 2n)$ are chosen.

Assumption: There exists a polynomial $Q_n(t)$ of degree n, with leading coefficient one and $Q_n(0) = 0$, which maps all of the intervals $[a_{2i-1}, a_{2i}]$ $(i = 1, 2, \ldots n)$ onto one and the same interval $[m, M]$, with $mM > 0$, and which maps the ends of $[a_{2i-1}, a_{2i}]$ onto the ends of $[m, M]$. Further $Q_n(t)$ must be a monotone function as each of the intervals $[a_{2i-1}, \quad]$ which varies from m to M or from M to m.

Let $\eta_i(\tau) \in [a_{2i-1}, a_{2i}]$ $(i = 1, 2, \ldots n)$ denote the roots of the equation

$$Q_n(t) = \tau \tag{4.2}$$

for $\tau \in [m, M]$. Then, the required $P_N(t; \Omega)$ is defined in terms of (4.2) using the following sequence of steps:

(i) Let $N = jn$, where $j > 0$ is an integer, and set

$$P_N(t) = S_j(Q_n(t)) \tag{4.3}$$

where $S_j(z)$ is the polynomial of degree j which deviates least from zero on [m, M] and is normalized with respect to the condition $S_j(0) = 1$.

(ii) Observe that

$$S_j(z) = T_j((2z - M-m)/(M-m))/T_j(z_0) \tag{4.4}$$

where $T_j(\xi) = \cos(j \arccos \xi)$ is the Chebyshev polynomial of degree j, and

$$z_0 = -(M+m)\ (M-m)^{-1}, \quad |z_0| > 1. \tag{4.5}$$

(iii) Since

$$T_j(z) = \prod_{i=1}^{j} (z-z_i), \quad z_i = \cos((2i-1)\pi/2j), \tag{4.6}$$

it follows that

$$P_N(t) = \prod_{i=1}^{j} \left(\frac{Q_n(t) - \tau_i}{\tau_i} \right) \tag{4.7}$$

where

$$\tau_i = \tfrac{1}{2}(M+m + z_i(M-m)). \tag{4.8}$$

(iv) Using the definition of $\eta_i(s)$, it follows that

$$P_N(t) = \prod_{i=1}^{j} \prod_{s=1}^{n} \left(\frac{t - \eta_s(\tau_i)}{-\eta_s(\tau_i)} \right) = \prod_{i=1}^{j} \prod_{s=1}^{n} \left(1 - \frac{t}{\eta_s(\tau_i)} \right) \tag{4.9}$$

which is the required $P_N(t; \Omega)$ [see Lebedev [15; Lemma] for the proof].

Thus, the actual construction of $P_N(t;\Omega)$ and the required α_k ($k = 1, 2, \ldots, N$) involves the following steps:

1. On the basis of the restrictions contained within the Assumption, construct an Ω and a corresponding $Q_n(t)$.

2. Set $N = jn$, where $j > 0$ is an integer, and determine all the roots $\eta_s(\tau_i)$ ($s = 1, 2, \ldots, n$) of the polynomials

$$Q_n(t) = \tau_i \qquad (i = 1, 2, \ldots, j)$$

with the τ_i defined by (4.8).

3. Then $P_N(t;\Omega)$ is defined by (4.9) with the roots

$$\alpha_k^{-1} = [\eta_s(\tau_i)] \quad (k = (i-1)n+s),\ i = 1,2,\ldots,j,\ s=0,1,\ldots,n-1). \tag{4.10}$$

In order to apply the results of §2, it is first necessary to determine $(\varphi_1, \varphi_2, \ldots, \varphi_N)$ which corresponds to the α_k arranged in descending order of magnitude.

Since, for the case $n = 2$, $Q_2(t) = t(at+b)$ is symmetric with respect to the line $t = -b/2a$, it can only be used to generate the required transformation for $\Omega = [a_1, a_2] \cup [a_3, a_4]$ if

$$a_2 - a_1 = a_4 - a_3 = \text{const.}$$

Hence, given that $o(A) \subseteq D = [b_1, b_2] \cup [b_3, b_4]$, it is necessary to examine the optimum choice of the a_i ($i = 1, 2, 3, 4$) for this D. All the possibilities are examined in detail in Lebedev [15]. We only pause to examine the case which covers the problem to be considered in §5. For this problem, we have $a_2 - a_1 = a_4 - a_3$, $n = 2$ and N even. Thus, we can use the following explicit expressions for the α_k of

(4.10) which Lebedev [15] derived using the properties of $Q_2(t)$:

$$\alpha_{2k-1} = \{c+[\tau_k+c^2]^{\frac{1}{2}}\}^{-1},\ \alpha_{2k} = \{c-[\tau_k+c^2]^{\frac{1}{2}}\}^{-1} \quad (k = 1,2,\ldots,j) \tag{4.11}$$

with τ_k defined by (4.8), $M = -a_1a_4$, $m = -a_2a_3$, and $c = (a_2+a_3)/2$.

§5. AN APPLICATION WITH THE SPECTRUM ON TWO DISJOINT SUBREGIONS

In this section, we examine the use of the Lebedev-Finogenov ordering for the solution of a positive definite sparse matrix A with its spectrum contained on two equal-length disjoint subintervals. The φ_i ($i = 1, 2, \ldots, N$) of (2.1) will then be defined by the roots β_i of (4.11) arranged in descending order of magnitude.

We do this by examining the following problem: on the unit square S {we use the notation of §3}, use finite difference methods to construct a continuous function $u = u(x,y)$, $(x,y) \in S$, which satisfies

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + q(x)\, u = -g(x,y) \qquad (5.1)$$

and the boundary conditions (3.2) and (3.3). Following the procedure of §3, we introduce the grid G and the differencing (3.4) and (3.5). This yields the following finite difference scheme for the determination of $u = u(x,y)$ on the grid G:

$$\Delta_h v_{ij} + h^2 q_i\, v_{ij} + h^2\, g_{ij} = 0 \quad (i, j = 1, 2, \ldots, I-1) \qquad (5.2)$$

along with (3.7) and (3.8), where $\Delta_h v_{ij}$ is defined by

$$\Delta_h v_{ij} = 4v_{ij} - v_{i,j-1} - v_{i,j+1} - v_{i-1,j} - v_{i+1,j}. \qquad (5.3)$$

For the solution of this difference scheme, we use the following generalization of Richardson's non-stationary method (1.2):

$$B\underset{\sim}{u}^{(k+1)} = B\underset{\sim}{u}^{(k)} - \alpha_k(C\underset{\sim}{u}^{(k)} - \underset{\sim}{f}), \qquad (5.4)$$

where the α_k are now the reciprocals of the roots of the polynomials which deviate least from zero on the spectrum of $B^{-1}C$.

We denote by $\bar{\underset{\sim}{v}}$ {and $\bar{\underset{\sim}{g}}$} the vectors obtained by ordering the elements v_{ij} {and g_{ij}} (i, j = 1, 2, • **, I-l) in the following way:

$$v_k = v_{ij}, k = (j-1) + i(I-1) \quad (j=1, 2, \ldots . I-1, i=1, 2, \ldots . I-1). \qquad (5.5)$$

Using this ordering, we can write (5.2) as

$$L\bar{\underset{\sim}{v}} + h^2\bar{\underset{\sim}{g}} - \underset{\sim}{b} \equiv (A + h^2Q)\bar{\underset{\sim}{v}} + h^2\bar{\underset{\sim}{g}} - \underset{\sim}{b} = \underset{\sim}{0} \qquad (5.6)$$

where

$$Q = \text{diag}\,(q_1\tilde{I}, q_2\tilde{I}, \ldots . q_{I-1}\tilde{I})$$

with $\tilde{I}$ the unit matrix of order I-l,

$$A = \begin{bmatrix} A & -\tilde{I} & & & \\ -\tilde{I} & A & -\tilde{I} & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & -\tilde{I} & A \end{bmatrix} \quad \text{with } A = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & 4 & -1 & \\ & & \cdot & \cdot & \cdot \\ & & & -1 & 4 \end{bmatrix}, \qquad (5.7)$$

and

$$b = [\underset{\sim}{b}_1, \underset{\sim}{b}_2, \ldots \cdot, \underset{\sim}{b}_{I-1}]^T$$

where the vectors $\underset{\sim}{b}_k$ (k = 1, 2, ... I-l) are all of order I-l with

$$\underset{\sim}{b}_1 = [\alpha_1+\beta_1, \alpha_2, \ldots, \alpha_{I-2}, \alpha_{I-1} + \bar{\beta}_1]^T, \quad \underset{\sim}{b}_{I-1} = [\beta_{I-1}+\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\beta}_{I-1}+\bar{\alpha}_{I-1}]^T$$

and

$$\underset{\sim}{b}_k = [\beta_k, 0, \ldots, 0, \bar{\beta}_k]^T \qquad (k = 2, 3, \ldots, I-2).$$

Thus, the implementation of the generalized Richardson procedure for the solution of (5.2), (3.7) and (3.8) becomes:

((i)) Set k = 1, and using the ordering defined by (5.5) set

$$\underset{\sim}{v}^{(0)} = (d_{11}, d_{12}, d_{13}, \ldots d_{I-1,I-1}) \tag{5.8}$$

where the d_{ij} define the starting solution.

((ii)) Compute

$$\bar{r} = L\,\underset{\sim}{v}^{(0)} + h^2\bar{g} - \underset{\sim}{b}, \qquad R = \max_k |\bar{r}_k|, \tag{5.9}$$

and then

$$\underset{\sim}{v}^{(1)} = \underset{\sim}{v}^{(0)} - \alpha_k \Delta^{-1}\underset{\sim}{r}, \tag{5.10}$$

followed by

$$\underset{\sim}{v}^{(0)} = \underset{\sim}{v}^{(1)}.$$

((iii)) For a given positive value ε, if $R > \varepsilon$, set $k = k+1$ and return to ((ii)) and then back to ((iii)); if $R < \varepsilon$, stop as $\underset{\sim}{v}^{(0)}$ defines an approximate solution with the required accuracy. The final value of k gives the number of iterations required.

Note. In (5.10), the actual inversion of Δ is done using one of the recently developed direct methods which takes into full account the sparseness of Δ. See, for example, Buzbee et al [19].

For this implementation, the α_k must be the reciprocals of the roots of the polynomials which deviate least from zero on the spectrum of $\Delta^{-1}L$. Since the Δ of (5.7) coincides with the A of (3.14), we obtain that the spectrum of $\Delta^{-1}L$ must lie on the interval

$$[1+h^2\{\lambda_{min}(Q)/\lambda_{min}(A)\},\ 1+h^2\{\lambda_{max}(Q)/\lambda_{min}(A)\}]$$

if $\lambda_{min}(Q) \cdot \lambda_{max}(Q) \leq 0$, and on the interval

$$[1+h^2\{\lambda_{min}(Q)/\lambda_{max}(A)\},\ 1+h^2\{\lambda_{max}(Q)/\lambda_{min}(A)\}]$$

if $\lambda_{min}(Q)\lambda_{max}(Q) \geq 0$. If, in (5.1) and (5.2), we take

$$q(x) = q_k(x) = \begin{cases} -4k & (x \leq \frac{1}{2} - 1/4k) \\ (4k)^2(\frac{1}{2} - x) & (\frac{1}{2} - 1/4k \leq x \leq \frac{1}{2} + 1/4k) \\ +4k & (x \geq \frac{1}{2} + 1/4k) \end{cases}$$

with $2kh \geq 1$ and I odd, then the spectrum of $\Delta^{-1}L$ will be on the equal-length disjoint subintervals

$$\left[1 - \frac{k}{2}\left(\frac{\sin\frac{\pi h}{2}}{h}\right)^{-2},\ 1 - \frac{k}{2}\left(\frac{\cos\frac{\pi h}{2}}{h}\right)^{-2}\right],$$
$$\left[1 + \frac{k}{2}\left(\frac{\cos\frac{\pi h}{2}}{h}\right)^{-2},\ 1 + \frac{k}{2}\left(\frac{\sin\frac{\pi h}{2}}{h}\right)^{-2}\right].$$

Consequently, using the notation and results of §4, the φ_i of §2 become the roots β_i of (4.11) [arranged in descending order of magnitude) with

$$c = \tfrac{1}{2}, \quad M = -\left(1 - \frac{k^2}{4}\left(\frac{\sin\frac{\pi h}{2}}{h}\right)^4\right)$$

$$m = -\left(1 - \frac{k^2}{4}\left(\frac{\cos\frac{\pi h}{2}}{h}\right)^4\right)$$

Applying the Lebedev-Finogenov ordering to the φ_i (as detailed in §2), the following two problems based on (5.1), (3.7) and (3.8) were solved using the above implementation ((i)), ((ii)), and ((iii)):

Problem 5.1. The homogeneous problem

$$\alpha(y) = \bar{\alpha}(y) = \beta(x) = \bar{\beta}(x) = 0, \; g(x,y) = 0,$$

which has the exact solution $u(x,y) = 0$, $(x,y)\in S$. Along with $k = 4$ and $I = 32$, the starting solution was taken to be

$$d_{ij} = 1 \qquad (i,j = 1, 2, \ldots I\text{-}1).$$

Problem 5.2. The non-homogeneous problem with

$$a(y) = p(x) = 0, \; \bar{\alpha}(y) = \sin \pi y, \; \bar{\beta}(x) = \sin \pi x,$$

$$g(x,y) = -\{\pi^2(x^2+y^2) + q_k(x)\} \sin \pi xy$$

which has the exact solution

$$u(x,y) = \sin \pi xy \qquad ((x,y)\in S).$$

Along with k = 4 and I = 32, the starting solution was taken to be

$$d_{ij} = 0 \qquad (i,j = 1, 2, \ldots . I-1).$$

The actual numerical results are discussed in §6.

§6. NUMERICAL RESULTS AND CONCLUSIONS

Numerical experimentation with Problems 3.1-3.3 using the ordering of Young [2], [3] as well as that of Lebedev and Finogenov indicated that:

(i) Young's ordering allowed (1.2) to behave in'a stable manner for small N when using the floating point double precision arithmetic of the IBM 360/67 computer at the Computer Science Department at Stanford University. This is easily reconciled with' Young's finding since his computations were performed with the low precision fixed point arithmetic of the ORDVAC. However, with the α_k chosen in ascending order, even the use of floating point double precision arithmetic did not prevent the rapid onset of instability.

(ii) For N ~ 100, Richardson's non-stationary scheme (1.2) behaved in an unstable manner when using the ordering of Young. This is illustrated in Table 1, where we list the errors arising from the use of Young's ordering with N = 128 when Problem 1 was solved.

(iii) When using the Lebedev-Finogenov ordering, the non-stationary scheme (1.2) always behaved in a stable manner. This is illustrated in Table 2, where we list the errors arising from the use of the Lebedev-Finogenov ordering with N = 128 when Problem 1 was solved.

Further support for the validity of (iii) is contained in Tables 3 and 4. Here, we list the errors arising from the solution of Problems 3.2 and 3.3 using (1.2) with the Lebedev-Finogenov ordering.

While of interest in its own right, the stability of the non-stationary scheme (1.2) for the Lebedev-Finogenov ordering raises important practical questions. For example:

(i) Since this stability result applies to a wider class of matrices than covered by the Property A condition, do there exist classes of matrices for which Richardson's non-stationary scheme, with the Lebedev-Finogenov ordering, yields better results than SOR?

(ii) Do there exist other orderings for which the non-stationary scheme (1.2) is stable?

Though answers to such questions will be of interest, the practical importance of this result will depend on how good a method it proves to be for the type of problem and procedure discussed in §5 (see also Concus and Golub [20]). That it represents a reliable method for such problems is illustrated by the results of Tables 5 and 6. Here, we list the residuals arising from the solution of Problems 5.1 and 5.2 using the generalized Richardson procedure of §5 with the Lebedev-Finogenov ordering applied to the α_k of (4.11) arranged in descending order of magnitude to form the set $(\varphi_1, \varphi_2, \ldots \varphi_N)$.

TABLE 1.

The Solution of Problem 3.1 Using The Non-stationary Scheme (1.2) With Young's Ordering

Table of $v_{ij}^{(n)}$ - values

(N=128, n=128, I=20, h=1/20)

	j=0	j=4	j=8	j=12	j=16	j=20
i=0	0.0	0.0	0.0	0.0	0.0	0.0
i=4	0.0	3 ● QE+?	9.51E+7	-3.64E+7	-2.96E+7	0.0
i=8	0.0	1.7 9E+7	-4.72E+7	8.27E+7	-6.28E+7	0.0
i=12	0.0	-6.71E+7	-1.87E+8	1.68E+8	-1.29E+8	0.0
i=16	0.0	-4.77E+7	-4.17E+7	-1.13E+8	2.16E+7	0.0
i=20	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 2.

The Solution of Problem 3.1 Using The Non-stationary Scheme (1.2) With The Lebedev-Finogenov Ordering

Table of $v_{ij}^{(n)}$ - values

(N=128, n=128, I=20, h=1/20)

	j=0	j=4	j=8	j=12	j=16	j=20
i=0	0.0	0.0	0.0	0.0	0.0	0.0
i=4	0.0	8.56E-10	2.83E-9	2.83E-9	8.56E-10	0.0
i=8	0.0	2.83E-9	7.73E-9	7.73E-9	2.83E-9	0.0
i=12	0.0	2.83E-9	7.73E-9	7.73E-9	2.83E-9	0.0
i=16	0.0	8.56E-10	2.8E-9	2.83E-9	8.56E-10	0.0
i=20	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 3.

The Residual of Problem 3.2 Using the Non-stationary Scheme (1.2) With The Lebedev-Finogenov Ordering

Table of $r_{ij}^{(n)}$ - values

(N=128, n=128, I=20, h=1/20)

	j=0	j=4	j=8	j=12	j=16	j=20
i=0	0.0	0.0	0.0	0.0	0.0	0.0
i=4	0.0	1.18E-9	7.16E-10	7.16E-10	1.18E-9	0.0
i=8	0.0	-2.25E-10	-4.13E-10	-4.13E-10	-2.25E-10	0.0
i=12	0.0	-9.97E-10	-1.17E-9	-1.17E-9	-9.97E-10	0.0
i=16	0.0	-4.22E-10	-5.99E-10	-5.99E-10	-4.22E-10	0.0
i=20	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 4.

The Residual of Problem 3.3 Using the Non-stationary Scheme (1.2) With The Lebedev-Finogenov Ordering

Tables of $r_{ij}^{(n)}$ - values

(N=128 n=128 I=20 h=1/20)

	j=0	j=4	j=8	j=12	j=16	j=20
i=0	0.0	0.0	0.0	0.0	0.0	0.0
i=4	0.0	8.91E-11	-6.56E-10	-1.21-E-9	-5.85E-10	0.0
i=8	0.0	-5.63E-11	-1.41E-9	-2.33E-9	-1.21E-9	0.0
i=12	0.0	5.13E-10	-4.90E-10	-1.41E-9	-6.563-10	0.0
i=16	0.0	8.23E-10	5.14E-10	-5.63E-11	8.91E-11	0.0
i=20	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 5.

The Residual of Problem 5.1 Using the Non-stationary Scheme (1.2) With The Lebedev-Finogenov Ordering

Table of $r_{ij}^{(n)}$ - values

(N=8, n=32, I=32, $h=2^{-5}$)

	j=0	j=8	j=16	j=24	j=32
i=0	0.0	0.0	0.0	0.0	0.0
i=8	0.0	-1.90E-10	5.50E-12	-1.90E-10	0.0
i=16	0.0	2.69E-23	-6.03E-23	-3.79E-23	0.0
i=24	0.0	-1.42E-10	9.573-Q	-1.42E-10	0.0
i=32	0.0	0.0	0.0	0.0	0.0

TABLE 6.

The Residual of Problem 5.2 Using the Non-stationary Scheme (1.2) With The Lebedev-Finogenov Ordering

Table of $r_{ij}^{(n)}$ - values

(N=8, n=32, I=32, $h=2^{-5}$)

	j=0	j=8	j=16	j=24	j=32
i=0	0.0	0.0	0.0	0.0	0.0
i=8	0.0	6.22E-11	3.43E-11	8.68E-11	0.0
i=12	0.0	-9.12E-12	-2.70E-11	-5.723-11	0.0
i=16	0.0	3.79E-11	-1.87E-11	5.99E-11	0.0
i=32	0.0	0.0	0.0	0.0	0.0

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Richard Brent of the Australian National University for preparing the ordering program given in Section 2. We are very grateful to Mr. Robert Tomlinson for performing the calculations so carefully and conscientiously. Dr. J. Alan George was 'kind enough to provide his fast Poisson solver which was used to solve Problems 5.1 and 5.2.

REFERENCES

[1] L. F. Richardson, The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stress in a masonary dam, Phil. Trans. Roy. Soc. London Ser A 210 (1910), 307-357.

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

[3] D. M. Young, On Richardson's method for solving linear systems with positive definite matrices, J. Math. Phys. 32 (1954), 243-255.

[4] D. M. Young, On the solution of linear systems by iteration, Proc. Sixth Symp. in Appl. Math., Amer. Math. Soc. 6 (1956), 283-298.

[5] M. Sh. Birman, On a variant of the method of successive approximations, Vestn. LGU 9 (1952), 69-76.

[6] M. K. Gavurin, The use of polynomials of best approximation for the improvement of the convergence of iteration processes, Uspekhi Matem Nauk 5(3) (1950), 156-160.

[7] Eduard L. Stiefel, Kernel polynomials in linear algebra and their numerical application, Further Contributions to the Solution of Simultaneous Linear Equations and the Determination of Eigenvalues, Nat. Bureau of Standards, Appl. Math. Series No. 49, 1958.

[8] M. Engeli, Th. Ginsburg, H. Rutishauser and E. Stiefel, Refined Iterative Methods For Computation of the Solution and Eigenvalues of Selfadjoint Boundary Value Problems, Mitteilungen aus dem Institut für angewandte Mathematik, Nr. 8, Birkhäuser Verlag, Basel, 1959.

[9] G. H. Golub and R. S. Varga, Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second-order Richardson iterative methods, Numer. Math. 3 (1961), 147-168.

[10] David M. Young, Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.

[11] V. I. Lebedev and S. A. Finogenov, On the order of choice of the iteration parameters in the Chebyshev cyclic iteration method, Zhur. Vych. Mat. i Mat. Fiz. 11 (1971), 425-438.

[12] B. A. Samokish, On the rate of convergence of the method of steepest descent, Uspe'khi Matem Nauk 12(1) (1957), 238-240.

[13] N. I. Achieser, Ueber einige Funktionen die gegebenen Intervallen am wenigsten von Null abwichen, Izv. Kazanskogo Fiz-Matem. Ob-va, 1928.

[14] A. A. Markov, On functions with least deviation from zero, Izbr. Tr., M-L, OGEZ, 1948.

[15] V. I. Lebedev, Iterative methods for the solution of operator equations with their spectrum lying on several intervals, Zhur. Vych. Mat. i Mat. Fiz 9(6)(1969), 1247-1252.

[16] D. M. Young and C. H. Warlick, On the use of Richardson's method for the numerical solution of Laplace's equation on the ORDVAC, Ballistic Research Labs. Memorandum Report No. 707, Aberdeen Proving Ground, Maryland, 1953.

[17] R. S. Varga, Matrix Iterative Analysis, Prentice Hall, Englewood Cliffs, 1962.

[18] G. H. Golub, Bounds for the round-off errors in the Richardson second order method, BIT 2 (1962), 212-232.

[19] B. L. Buzbee, G. H. Golub, and C. W. Nielson, On direct methods for solving Poisson's equations, SIAM J. Numer. Anal. 7 (1970), 627-656.

[20] P. Concus and G. H. Golub, Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations, Computer Science Department Technical Report No. STAN-CS-72-278, April, 1972.

APPENDIX 1.

ITERATIVE METHODS FOR THE SOLUTION OF OPERATOR EQUATIONS WITH THEIR SPECTRUM LYING ON SEVERAL INTERVALS.*

V. I. Lebedev
(Moscow)

Let

$$Au = f \tag{1}$$

denote an equation in a Hilbert space H with A a bounded self-adjoint operator. Let $\sigma(A)$ denote the spectrum of A with $0 \notin \sigma(A)$. We examine for the solution of (1) the effectiveness of the use of cyclic iteration methods of period N [see [1]-[3]], viz.

$$u^{k+1} = u^k - \alpha_k(Au^k - f), \tag{2}$$

where the α_k are numerical parameters such that $\alpha_{k+N} = \alpha_k$. Performing N iterations with (2), we obtain

$$u^{k+N} = P_N(A)u^k + (I - P_N(A))A^{-1}f,$$

where the polynomials $P_N(t)$ have the form

$$P_N(t) = \prod_{k=1}^{N} (1 - \alpha_k t) \tag{3}$$

$$P_n(0) = 1. \tag{4}$$

*Translator's Note. First published in Zhurnal Vychislitel'noy Matematiki i Matematicheskoy Fiziki 9(6) (1969), 1247-1252.

Thus, the error $\epsilon^k = u - u^k$ satisfies the following recursive formula

$$\epsilon^{k+N} = P_N(A)\epsilon^k. \tag{5}$$

Hence, if the coefficients in (2) are. chosen so that the polynomials of the form (3) and (4) deviate least from zero (PDLZ $\sim$ polynomial which deviates least from zero) on the set o(A), then we obtain a sufficiently effective Chebyshev iteration method which gives for N iterations the maximum damping of all the errors $\epsilon^k \in M = \{\epsilon : ||\epsilon|| \leq c\}$.

The actual construction of such a polynomial, as a rule, does not appear to be feasible because either a known structure of o(A) is not available, or o(A) is such that it is difficult to construct a PDLZ. It is clear that the problem can be solved in the following simple minded way: Assume it is known that $\sigma(A) \in \Omega$ where Ω is such that a PDLZ can be constructed for it, then the α_k are the reciprocals of the roots of this polynomial.

Methods for the construction of iterative methods (2) with $\Omega = [m,M]$ and $mM > 0$ are well known [see [1] - [3]}. In [4], the choice of the α_k with N = 1 is based on a conformalmapping procedure when $\sigma(A)$ belongs to a set Ω in the complex plane and the complement of Ω is a connected region. The case, when Ω consists of the region [m, M] with $mM > 0$ and a point $\lambda > M$, was examined in [5]. Methods for the construction of PDLZ on two non-intersecting regions, with the coefficient of the highest term unity, were developed in [6]. We note that, when these two regions lie on opposite sides of the origin, the polynomials found in [6] can not be always used as a basis for the construction of PDLZ of the form (3) and (4).

In this article, we examine the construction of PDLZ of the form (3) and (4) when Ω consists of n intervals of the real axis. We make no assumption about whether A is positive or negative.

Case 1: $n \geq 1$. Let

$$\Omega = \bigcup_{i=1}^{n} [a_{2i-1}, a_{2i}],$$

where $a_i < a_{i+1}$ $(i = 1, 2, \ldots, 2n-1)$, $0 \notin [a_{2i-1}, a_{2i}]$ $(i=1, 2, \ldots, n)$. In addition, we require that the a_i satisfy the following algebraic condition: there exists an n-th degree polynomial $Q_n(t)$, with leading coefficient one and $Q_n(0) = 0$, which maps the intervals $[a_{2i-1}, a_{2i}]$ $(i = 1, 2, \ldots, n)$ as a whole onto one and the same interval $[m, M]$ with $mM > 0$ and maps the ends of $[a_{2i-1}, a_{2i}]$ onto the ends of $[m, M]$. We denote by $\eta_i(\tau) \in [a_{2i-1}, a_{2i}]$ $(i = 1, 2, \ldots, n)$ the roots of the equation

$$Q_n(t) = \tau$$

for $\tau \in [m, M]$. It is clear that $Q_n(t) - (M+m)/2$ is a PDLZ on Ω with leading coefficient one. It follows from this, that the modulus of $Q_n(t) - (M+m)/2$ takes its maximum value of $(M-m)/2$ on Ω and that its sign oscillates with respect to the following $n+1$ point of Ω : $a_1, a_2, a_4, \ldots, a_{2n}$ {see [7]}. Below, we require the fact that $Q_n(t)$ is a monotone function on each of the intervals $[a_{2i-1}, a_{2i}]$ which varies from m to M or M to m.

For the construction of the PDLZ, we make use of a well-known method [7], which is used in the actual construction of the polynomial

and in the proof that it is the required polynomial. Let $N = jn$, where $j > 0$ is an integer. We set

$$P_N(t) = S_j(Q_n(t)),$$

where $S_j(z)$ is the PDLZ on $[m, M]$ of degree j which is normalized with respect to the condition $S_j(0) = 1$. In fact, $S_j(z)$ satisfies the following explicit expression [7]

$$S_j(z) = T_j((2z - M - m)/(M - m))/T_j(z_0),$$

where $T_j(\xi) = \cos j \arccos \xi$ is the Chebyshev polynomial of degree j and $z_0 = -(M+m)\ (M-m)^{-1}$, $|z_0| > 1$. Since

$$T_j(z) = \prod_{i=1}^{j} (z - z_i),\quad z_i = \cos \frac{(2i - 1)\pi}{2j},$$

it follows that

$$P_N(t) = \prod_{i=1}^{j} \left(\frac{Q_n(t) - \tau_i}{\tau_i} \right) \tag{6}$$

where

$$\tau_i = 1/2(M + m + z_i(M - m)). \tag{7}$$

We transform (6) to obtain

$$P_N(t) = \prod_{i=1}^{j} \prod_{s=1}^{n} \left(\frac{t - h_s(\tau_i)}{-h_s(\tau_i)} \right) = \prod_{i=1}^{j} \prod_{s=1}^{n} \left(1 - \frac{t}{h_s(\tau_i)} \right). \tag{8}$$

Let

$$E_N = E_N(\Omega) = \max_{t \in \Omega} |w_N(t)|,$$

where $w_N(t)$ is the PDLZ on Ω of degree N which is normalized with respect to the condition $w_N(0) = 1$.

Lemma. Among all polynomials of degree N which are normalized with respect to the condition (4), $P_N(t)$ defined by (8) is the PDLZ on Ω.

For the proof of the Lemma, we note that $P_N(t)$ attains, with respect to its modulus, a maximum on Ω which equals $|T_j(z_0)|^{-1}$, and has an oscillating sign with respect to the following $jn+1$ points Ω_1 of Ω : a_{2i-1} $(i = 1, 2, \ldots, n)$, a_{2n} and $(j-1)$ internal points on each of the intervals $[a_{2i-1}, a_{2i}]$ $(i=1, 2, \ldots, n)$. In addition,

$$P_N(a_{2i+1}) = P_N(a_{2i}), \quad i = 1, 2, \ldots, n-1. \tag{9}$$

We assume that $\bar{P}_N(t)$ and not $P_N(t)$ is the PDLZ. We examine the behavior of the polynomial

$$\varphi_N(t) = P_N(t) - \bar{P}_N(t)$$

with degree less than or equal to N. It is clear that

$$\varphi_N(0) = 0. \tag{10}$$

On the other hand, $\varphi_N(t)$ changes sign at the N + 1 points of Ω_1, i.e., $\varphi_N(t)$ has no less than N roots on $[a_1, a_{2n}]$, and we conclude, as a consequence of (9), that not less than N roots of $\varphi_N(t)$ lie in Ω. Taking the degree of $\varphi_N(t)$ and condition (10) into account, we obtain that $\varphi_N(t) \equiv 0$. The resulting contradiction proves the Lemma.

Comparing (3) and (8), we see that the α_k of method (2) must take the values

$$\alpha_k = \eta_s^{-1}(\tau_i) \tag{11}$$

with $k = n(i-1) + s$ $(i = 1, 2, \ldots, j, s = 1, 2, \ldots, n)$. Hence,

$$E_N = |T_j(z_0)|^{-1} = 2((z_0 + \sqrt{(z_0^2 - 1)})^j + (z_0 - \sqrt{(z_0^2 - 1)})^j)^{-1} < 1,$$

and thus, $||\epsilon^{k+N}|| \leq E_N ||\epsilon^k||$. This leads to the problem concerning the existence of implementations of (2) for which the strong accumulation of rounding errors does not occur. We note that, for $n = 1$, we obtain Chebyshev's method for one interval.

Case 2: n=2. If the above assumptions regarding Ω hold, then for n=2 they imply that $a_2 - a_1 = a_4 - a_3 = h$. Let

$$c = (a_2 + a_3)/2, \qquad h_1 = (a_3 - a_2)/2,$$

then

$$Q_2(t) = t(t - 2c), \quad M = -a_1 a_4, \; m = -a_2 a_3, \tag{12}$$

$$z_0 = -(a_1 a_4 + a_2 a_3)/(a_1 a_4 - a_2 a_3),$$

$$a_{2i-1} = [c + \sqrt{(\tau_i + c^2)}]^{-1}, \; a_{2i} = [c - \sqrt{(\tau_i + c^2)}]^{-1}, (i=1, 2, \ldots, j.)$$

We evaluate $|T_j(z_0)|$. For this we put $t_0^2 = (1 + z_0)/2$. Calculating t_0, we obtain

$$t_0^2 = (1 - a_2 a_3 / a_1 a_4)^{-1}.$$

Therefore,

$$T_j(z_0) = T_j(T_2(t_0)) = T_{2j}(t_0),$$

and hence, for $t_0^2 > 0$, we have $t_0^2 > 1$ and

$$E_{2j} = 2\{[t_0 + \sqrt{(t_0^2 - 1)}]^j + [t_0 - \sqrt{(t_0^2 - 1)}]^j\}^{-1} \qquad (14)$$

and for $t_{02} = -\rho^2 \leq 0$, we have

$$E_{2j} = 2\{[\sqrt{(\rho^2 + 1)} + \rho]^j + [\sqrt{(\rho^2 + 1)} - \rho]^j\}^{-1} .$$

For $j > 1$ and $n = 2$, we examine the following problem. Assume it is known that $\sigma(A) \subseteq D = [b_1, b_2] \cup [b_3, b_4]$ where $b_1 \leq b_2 \leq b_3 \leq b_4$, $0 \notin D$, and let $\ell_1 = b_2 - b_1$, $\ell_2 = b_4 - b_3$ and $\Delta = \ell_1 - \ell_2$. How is the set Ω chosen so that the modulus of z_0 is a maximum?

Let $b_1 b_4 > 0$. It is sufficient to examine the case when $b_1 > 0$. If $\Delta > 0$, then we put $a_1 = b_1$ and $a_2 = b_2$. It is necessary to imbed $[b_3, b_4]$ in an interval $[a_3, a_4]$ of length ℓ_1. It follows from (12) that for Ω it is necessary to take the set

$$\Omega = [b_1, b_2] \cup [b_3 - \Delta, b_4]. \qquad (14)$$

If $\Delta < 0$, then we put $a_3 = b_3$, $a_4 = b_4$ and imbed $[b_1, b_2]$ in $[a_1, a_2]$. From (12) it follows that

$$\Omega = [b_1, b_2 - \Delta] \cup [b_3, b_4]. \qquad (15)$$

In this way, if Ω, defined by (14) or (15), is simply connected, then it is best to apply the Chebyshev method for $[b_1, b_4]$. If $a_2 < a_3$, it is easy to see that the examined method converges more quickly than the cyclic method (2) for the interval $[b_1, b_4]$ with period 2j.

Let $b_1 b_4 < 0$ and without loss of generality $\Delta \geq 0$. We put $a_1 = b_1$, $a_2 = b_2$ and imbed $[b_3, b_4]$ in $[a_3, a_4]$. From (12) it follows that

$$\Omega = [b_1, b_2] \cup [b_3, b_4 + \Delta].$$

Now, we construct a PDLZ on D of the form

$$P_2(t) = 1 + 2at + bt^2. \tag{16}$$

Then, for N = 2 and $\sigma(A) \subseteq D$, we have for (2) that

$$\alpha_{1,2} = -a \pm \sqrt{(a^2 - b)}.$$

As a preliminary, we introduce the notation

$$\varphi(x, y, z) = -(x + y)\,(x(y + z) + z(y - z))^{-1},$$

$$\psi(x, y, z) = 2(x(y + z) + z(y - z))^{-1},$$

$$\beta_{1,2}(x, y, z) = -2(x + y \pm \sqrt{[(x - z)^2 + (y - z)^2]})^{-1},$$

$$\gamma(x, y, z) = |(y - z)\,(z - x)\,(x(y + z) + z(y - z))^{-1}|.$$

1. Let $b_1 b_4 > 0$ and, without loss of generality, $b_1 > 0$. Then $P_2(t)$ will be a PDLZ, if, on D, it attains on three occasions the maximum modulus value with changing sign. We actually construct such a polynomial.

a. If $b_2 \geq (b_1 + b_4)/2$ or $b_3 \leq (b_1 + b_4)/2$, then, as is easily verified, the PDLZ on D is the same as the PDLZ on $[b_1, b_4]$, i.e.

$$\alpha_{1,2} = 2(b_4 + b_1 \pm (b_4 - b_1)/\sqrt{2})^{-1},$$

$$E_2 = (b_4 - b_1)^2((b_4 + b_1)^2 + 4b_4 b_1)^{-1}.$$

b. Let $b_2 < (b_1+b_4)/2$ and $b_3 > (b_1+b_4)/2$. If $\Delta \geq 0$, then the points of oscillation are b_1, b_2 and b_4.

We find from $P_2(b_1) = -P_2(b_2) = P_2(b_4)$ that

$$a = \varphi(b_1, b_4, b_2), \quad b = \psi(b_1, b_4, b_2), \quad \alpha_{1,2} = \beta_{1,2}(b_1, b_4, b_2),$$

$$E_2 = \gamma(b_1, b_4, b_2).$$

For $\Delta \leq 0$, the points of oscillation will be the points b_1, b_3 and b_4. We find from $P_2(b_1) = -P_2(b_3) = P_3(b_4)$ that

$$a = \varphi(b_1,b_4,b_3), \quad b = \psi(b_1,b_4,b_3), \quad \alpha_{1,2} = \beta_{1,2}(b_1,b_4,b_3)$$

$$E_2 = \gamma(b_1,b_4,b_3).$$

For the examined case 1, the results of [6] and [8] on the construction and evaluation of PDLZ of higher degree are applicable. In particular, in [6], the explicit form of the polynomial of third degree is presented.

2. The analysis of the case $b_1 b_4 < 0$ is somewhat more involved, since the system t^k (k = 1, 2, ... n) does not satisfy the Haar condition on $[b_1,b_4]$ {see [9]}. Initially, we note that $E_2 < 1$, i.e. the maximum of the modulus of the PDLZ is not attained inside D, -but is attained on the ends of the intervals $[b_1,b_2]$ and $[b_3,b_4]$.

Let $\Delta \geq 0$. We construct $P_2(t)$ under the assumption that $-P_2(b_1) = P_2(b_2) = P_2(b_3)$. Then $a = \varphi(b_2,b_3,b_1)$ and $b = \psi(b_2,b_3,b_1)$. The derivative of $P_2(t)$ becomes zero at the point $(b_2+b_3)/2 \notin D$, hence, $P_2(t)$ attains its extreme values on D at $t = b_i$ (i = 1, 2, 3, 4) with $P_2(b_2) = P_2(b_3) > 0$. As a consequence of the symmetry of $P_2(t)$

with respect to the straight line $t = (b_2+b_3)/2$ and $\Delta \geq 0$, we have $|P_2(b_4)| < |P_2(b_3)|$. We show that $P_,(t)$ is the PDLZ. In fact, if it is not $P_,(t)$, then let it be $\bar{P}_2(t)$. Then the second degree polynomial

$$\varphi_2(t) = P_2(t) - \bar{P}_2(t)$$

changes sign on $[b_1,b_2]$, and hence, there exists a zero of $\varphi_2(t)$ inside $[b_1,b_2]$. On $[b_2,b_3]$ the polynomial $\varphi_2(t)$ has two roots, since it has the same sign at the ends of $[b_2,b_3]$ and $\varphi_2(0) = 0$. Hence, $\varphi_2(t) \equiv 0$. The resulting contradiction shows that $P_2(t)$ is the PDLZ and that

$$\alpha_{1,2} = \beta_{1,2}(b_2,b_3,b_1), \quad E_2 = \gamma(b_2,b_3,b_1).$$

Analogously, if $\Delta < 0$, then the PDLZ on D is determined by

$$P_2(b_2) = P_2(b_3) = -P_2(b_4),$$

i.e.

$$a = \varphi(b_2,b_3,b_4),\ b = \psi(b_2,b_3,b_4), \quad \alpha_{1,2} = \beta_{1,2}(b_2,b_3,b_4)$$

$$E_2 = \gamma(b_2,b_3,b_4).$$

We compare the examined method with two known convergent iteration methods with $b_1b_4 < 0$. The method

$$u^{k+1} = u^k - \alpha^2 A(Au^k - f), \tag{17}$$

where α^2 is a constant, has each iteration defined by the operator $(1-\alpha^2A^2)$. We assume that A^2u^k is calculated for each iteration by multiplying

u^k twice with A. Analogously, the method [10]

$$u^{k+1} = u^k + (-1)^k \alpha(Au^k - f) \tag{18}$$

has every two iterations defined by the operator $(1 - 2\alpha^2 A)$. Comparing (17) and (18) with (5), we see that for the same number of operations the methods (17) and (18) guarantee a lower rate of convergence when either $A \neq 0$ or $A = 0$ and $|b_1| \neq |b_4|$.

Finally, we note that the examined iterative methods allow problems of the following kind to be solved: Let the selfadjoint operator B have eigenvalues λ_i such that either $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_k < \lambda_{k+1} \leq \cdots \leq \beta$, or $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \beta$. It is necessary to solve $Bu - \varkappa u = f$ for $\lambda_k < \varkappa < \lambda_{k+1}$ by the iteration method (2). For example, such problems arise when eigen value and functions are determined by a method with shift [2].

Received by the Editors: 24.05.1968.

REFERENCES

[1] L. F. Richardson, The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stress in a masonary dam, Phil. Trans. Roy. Soc. London A. 210(1910), 307-357.

[2] D. K. Faddeev and V. N. Faddeeva, Computational Methods of Linear Algebra, M, Fizmatgiz, 1960.

[3] M. Sh. Birman, Some estimates for the method of steepest descent, Uspekhi Matem Nauk 5(3)(1950), 152-155.

[4] V. N. Kublanovskaya, The application of analytic continuation to the change of variable method in numerical analysis, Tr. Matem. in - ta ANSSSR 53(1959), 145-185.

[5] B. A. Samokish, On the rate of convergence of the method of steepest descent, Uspekhi Matem Nauk 12(1)(1957), 238-240.

[6] N. I. Achieser, Ueber einige Funktionen die gegebenen Intervallen am wenigsten von Null abweichen, Izv. Kazanskogo Fiz - Matem. Ob-va, 198.

[7] A. A. Markov, On functions with least deviation from zero, Izbr. Tr., M-L, OGEZ, 1948.

[8] S. N. Bernstein, Extremal Properties of Polynomials and The Best Approximation of a Continuous Function of One Real Variable, M., GONTE, 1937.

[9] A. F. Timan, The Theory of Approximation of Functions of a Real Variable, M.,Fizmatigiz, 1960.

[10] H. Buckner, Ein unbeschränkt anwendbarer Iterationverfahren für Fredholmische Integralgleichungen, Math Nachr 2 (1949), 304-314.

APPENDIX 2.

ON THE ORDER OF CHOICE OF THE: ITERATION PARAMETERS IN THE CHEBYSHEV CYCLIC ITERATION METHOD.*

V. I. Lebedev and S. A. Finogenov
(Moscow)

A solution is given for the problem of ordering the iteration parameters in a cyclic iteration method** which can be used to solve Au = f. This solution guarantees a computationally stable form for the method.

In the 1950's, when it was proposed that the equation

$$Au = f \tag{1}$$

be solved by cyclic iteration methods [1-4]

$$u^{k+1} = u^k - \alpha_k(Au^k - f), \tag{2}$$

with the iteration parameters α_k ($\alpha_{k+N} = \alpha_k$) related to the roots of the Chebyshev polynomial:;, it was noted that, when solving (1) by such methods on a computer with a finite word length using fixed or floating point operations, one can estimate, for poorly conditioned A, the loss of significant figures in the intermediate and final results (in u^k), and the resulting significance of the intermediate calculations with respect to the initial accuracy of the data. Marked

*Translator's Note. First published in Zhurnal Vychislitel'noy Matematiki i Matematicheskoy Fiziki 11(2) (1971), 425-438.

**Translator's Note. In fact, Richardson's non-stationary method.

instability has to a significant degree blocked the wide application of these methods. It can be shown that this instability depends heavily on the order in which the iteration parameters α_k are used.

In [3] {see also [5] and [6]} some proposals regarding the use of the α_k have been given; but, as we shall explain, they do not eliminate but only reduce the instability within the method. No other investigations of this particular problem are known to the Authors. In this article, we are given an ordering for the iteration parameters α_k which guarantees a computationally stable implementation of the method.

We shall assume that (1) is defined on a Banach space $\underline{B}$, that $u, f \in \underline{B}$, and that A is a bounded operator which maps from $\underline{B}$ into B and has a complete system of normalized eigenvectors φ_n which correspond to its eigenvalues λ_n. Let $\sigma(A)$ {the spectrum of A} lie on the real axis, and m and M denote its exact lower and upper bounds with $0 < m \leq M$.

Let $\epsilon^k = u - u^k$ and $N > 0$ be some fixed integer. Then, in (2), the set $(\alpha_1, \alpha_2, \ldots, \alpha_N)$ is a permutation of the set $(\gamma_1, \gamma_2, \ldots, \gamma_N)$ where

$$(3) \qquad \gamma_i = 2\left(M + m - (M - m)\cos\frac{(2i - 1)\pi}{2N}\right)^{-1}, \quad i = 1, 2, \ldots, N.$$

The one-to-one mapping between $(\alpha_1, \alpha_2, \ldots, \alpha_N)$ and $(\gamma_1, \gamma_2, \ldots, \gamma_N)$ is defined by the permutation

$$\nu_N = (i_1, i_2, \ldots, i_N)$$

with $\alpha_k = \gamma_{i_k}$.

Let N belong to some increasing sequence of integers $\{N_i\}$ for which the permutations $\varkappa_{N_i}$ are defined. Our problem reduces to the determination of a permutation $\varkappa_N$ which guarantees a computationally stable implementation of (2) for small m/M and $N \in \{N_i\}$.

Performing N iterations with (2) using exact arithmetic, we obtain that

$$u^N = P_N(A)u^0 + (I - P_N(A))A^{-1}f,$$

where the polynomials $P_N(t)$ have the form

$$P_N(t) = \prod_{k=1}^{N} (1 - \alpha_k t) = \frac{T_N[(M + m - 2t)/(M - m)]}{T_N(\theta)},$$

with

$$T_N(t) = \cos(N \arccos t), \qquad \theta = \frac{M + m}{M - m} > 1.$$

Let

$$R_i^N(t) = \prod_{j=1}^{i} (1 - \alpha_j t), \qquad Q_i^N(t) = \prod_{j=i+1}^{N} (1 - \alpha_j t).$$

Then the errors ϵ^k $(k=1, 2, \ldots, N)$ satisfy

$$\epsilon^k = R_k^N(A)\epsilon^0, \tag{4}$$

and if

$$\epsilon^k = \sum_n \epsilon_n^k \varphi_n,$$

then we obtain that

$$\epsilon_n^k = R_k^N(\lambda_n)\,\epsilon_n^0. \tag{5}$$

Real computations on a computer, which has fixed or floating point operations with finite length words, are executed with rounding error as a result of which errors arise. This fact can be taken into account if it is assumed in (2) that $u^k + h^k$ is written instead of u^k, where h^k is the error added to u^k during the calculation of u^{k+1}. Independent of the actual rounding procedure on a given computer, the h^k can be interpreted as the result of errors which arise as a result of the rounding in the computation of u^k and in the final result. The h^k can be correlated among themselves.

Instead of (2), we obtain

$$u^{k+1} = u^k + h^k - \alpha_k(A(u^k + h^k) - f) \tag{6}$$

with

$$\epsilon^k = R_k^N(A)\epsilon^0 - \sum_{i=1}^{k-1} Q_i^N(A)h^i - h^k. \tag{7}$$

If

$$h^k = \sum_n \eta_n^k \varphi_n \quad ,$$

then

$$\epsilon_n^k = R_k^N(\lambda_n)\epsilon_n^0 - \sum_{i=1}^{k-1} Q_i^k(\lambda_n)\, h_n^i - h_n^k. \tag{8}$$

We now examine two reasons why the iteration process (2) loses computational stability and the accuracy of the approximate solution of (1) is reduced. Our optimal choice of $\varkappa_N$ is based on the removal from (2) of situations of this type.

Let

$$q_i^N = \max_{m \le t \le M} |Q_i^N(t)|, \quad r_i^N = \max_{m \le t \le M} |R_i^N(t)| , \tag{9}$$

$$t_i^N = q_i^N r_i^N , \quad q^N = \max_i q_i^N , \quad r^N = \max_i r_i^N . \tag{10}$$

The first reason: The loss of significant digits in the intermediate iterations (with $k < N$) which results in the loss of accuracy in the final solution and in the growth of $\|n^k\|$ or the accidental stopping of computations due to computer under or overflow.

The loss of significant digits will occur when $\|u^k\| >> \|u\|$ with $1 \le k \le N$ and $N \to \infty$. This is equivalent to the condition

$$\sup_k \|\varepsilon^k\| \gg \|u\| \quad \text{as } N \to \infty, \quad N \in \{N_i\} .$$

Taking (4), (5), (9) and (10) into account, we see that a substantial loss of accuracy due to the growth of $\|u^k\|$ does not occur for initial data which satisfies $\|\varepsilon^0\| < c_1$, if

$$r^N < c_2 , \tag{11}$$

where c_2 is a constant which depends on m/M but not on N.

The second reason: The growth of the quantity.

$$\sup_k \|Q_k^N(A) n^k\| / \|n^k\| \quad \text{as} \quad N \to \infty, \quad N \in \{N_i\},$$

which characterizes the instability in the computation.

Taking (7), (8), (9) and (10) into account, we see that the calculation will be stable with respect to permissible error, if

$$q^N < C_3, \tag{12}$$

where C_3 is a constant which depends on m/M but not on N.

Thus, conditions (11) and (12) insure that a substantial loss of accuracy in the final solution is avoided. In fact, they represent necessary conditions for the stability of any real implementation of (2) on a class of initial data which satisfy $\|\epsilon^0\| < C_1$. Sufficient conditions for the stability of such an algorithm depend on the method and order of the implementation of (2) and the type of rounding used as well as (11) and (12).

Initially, we examine the character of r^N and q^N for the simplest permutations; viz.

$$\pi_1^N = (1,2,\ldots,N) \text{ or } \pi_2^N = (N,N-1,\ldots,2,1).$$

As a preliminary, we introduce the notation which will be used in the proofs of the Lemmas:

$$\delta = \theta - 1, \quad \Delta = \delta/2, \quad \beta_k = (2k-1)\pi/2N,$$

$$z = (M + m - 2t) / (M-m).$$

Lemma 1. If $\varkappa_N$ corresponds to the permutation π_1^N, then

$$\left\{\left[1 + \frac{\Delta}{\cos^2(\pi/4N)}\right]\left[1 + \frac{\Delta}{\cos^2[\pi(2N - 2j - 1)/4N]}\right]\right\}^{(j-N)/2} \leq$$

$$\leq q_j^N \leq \left\{1 + \frac{\Delta}{\cos^2[\pi(N - j)/4N]}\right\}^{j-N}$$

Proof. It follows from the conditions of the Lemma and (3) that $\alpha_k = \gamma_k$. It is clear that

$$\max_{m \leq t \leq M} |1 - \alpha_k t| = (1 - \alpha_k m) < 1 \quad (k \geq N/2),$$

and hence, $q_j^N = Q_j^N(m)$ for $j \geq N/2$.

We now prove that for $j < N/2$

$$q_j^N = \max_{m \leq t \leq M} |Q_j^N(t)| = Q_j^N(m).$$

In order to do this, we examine the function

$$v(\varphi) = \left|\prod_{k=j+1}^{N-j}\left(\cos\varphi - \cos\frac{\pi}{2N}(2k - 1)\right)\right| =$$

$$= \left|\prod_{k=j+1}^{N/2}\left(\cos\varphi - \cos\frac{\pi}{2N}(2k - 1)\right)\left(\cos\varphi + \cos\frac{\pi}{2N}(2k - 1)\right)\right| =$$

$$= \left|\prod_{k=j+1}^{N-j}\sin\left(\varphi + \frac{\pi}{2N}(2k - 1)\right)\right|,$$

where $\cos\varphi = Z$.

Let $\varphi = \psi + \pi i/N$, where $0 \leq \psi < \pi/N$, $0 \leq i \leq N/2 - 1$, then

$$\nu\left(\psi + \frac{\pi}{N} i\right) = \left| \prod_{k=j+i+1}^{N-j+i} \sin\left(\psi + \frac{\pi}{2N}(2k - 1)\right) \right| =$$

$$= \left| \prod_{k=j+1+i}^{N-j} \sin\left(\psi + \frac{\pi}{2N}(2k - 1)\right) \right| \left| \prod_{k=N-j+1}^{N-j+i} \sin\left(\psi + \frac{\pi}{2N}(2k - 1)\right) \right| .$$

But

$$\left| \prod_{k=N-j+1}^{N-j+1} \sin\left(\psi + \frac{\pi}{2N}(2k - 1)\right) \right| \leq$$

$$\leq \left| \prod_{k=j+1}^{j+i} \sin\left(\psi + \frac{\pi}{2N}(2k - 1)\right) \right| ,$$

and consequently,

$$\nu\left(\psi + \frac{\pi}{N} i\right) \leq \nu(\psi).$$

Since $\nu(\psi)$ attains its maximum for $\psi = 0$ and differs from $Q_j^{N-j}(t)$ only by a constant factor, it follows that

$$q_j^N = \max_{m \leq t \leq M} |Q_j^N(t)| = Q_j^N(m).$$

But

$$Q_j^N(m) = \prod_{k=j+1}^{N} \frac{1 - \cos[\pi(2k-1)/2N]}{\theta - \cos[\pi(2k-1)/2N]} =$$

$$= \prod_{k=j+1}^{N} \left\{1 + \frac{A}{\sin^2[\pi(2k-1)/4N]}\right\}^{-1} =$$

$$= \prod_{k=1}^{\frac{N-j}{2}} \left\{1 + \frac{A}{\cos^2[\pi(2k-1)/4N]}\right\}^{-1} \times$$

$$\times \underbrace{\left\{1 + \frac{A}{\cos^2[\pi(2(N-j-k)+1)/4N]}\right\}^{-1}}_{I} \leq$$

$$\leq \prod_{k=1}^{N-j} \left\{1 + \frac{A}{\cos^2[\pi(N-j)/4N]}\right\}^{-1} = \left\{1 + \frac{\Delta}{\cos^2[\pi(N-j)/4N]}\right\}^{j-N}$$

On the other hand,

$$|Q_j^N(m)| > \prod_{k=1}^{(N-j)/2} \left[1 + \frac{A}{\cos^2(\pi/4N)}\right]^{-1} \times$$

$$\times \left\{1 + \frac{A}{\cos^2[\pi(2N-2j-1)/4N]}\right\}^{-1} =$$

$$= \left[1 + \frac{A}{\cos^2(\pi/4N)}\right]^{(j-N)/2} \left\{1 + \frac{A}{\cos^2[\pi(2N-2j-1)/4N]}\right\}^{(j-N)/2}.$$

Lemma 2. If $\varkappa_N = \pi_1^N$, $A = m/(M-m) < 1/2$ and $i \le N_0(N) = 1/2 + (2N/\pi)[(1-\Delta)/2]^{\frac{1}{2}}$, then

$$r_i^N \ge \left[\Delta + \frac{\pi^2}{16N^2}\right]^{\frac{1}{2}} \left[\frac{1 - \pi^2(2i+1)^2/16N^2}{A + \pi^2(2i+1)^2/16N^2}\right]^{i+1/2}.$$

Indeed, under the conditions of the Lemma $r_i^N = |R_i^N(M)|$, if

$$2i - 1 \le (2N/\pi) \arcsin[(1-\Delta^2)]^{\frac{1}{2}}$$

Noting that

$$\frac{1+\cos\beta_k}{6 + 2\sin^2(\beta_k/2)} > \frac{1 - (\beta_k/2)^2}{\Delta + (\beta_k/2)^2} > 1 \text{ when } \beta_k < \sqrt{[2(1-\Delta)]},$$

we obtain that

$$r_i^N = \prod_{k=1}^{i} \frac{1+\cos\beta_k}{\delta + 2\sin^2(\beta_k/2)} \ge \prod_{k=1}^{i} \frac{1 - x_k^2}{A + x_k^2} = w,$$

where

$$x_k = \frac{\beta_k}{2} = \frac{2k-1}{2N}\,\frac{\pi}{2}, \qquad i \le N_0(N).$$

But

$$\ln w = \frac{2N}{\pi} Z,$$

where

$$Z = \sum_{k=1}^{i} \ln \frac{1 - x_k^2}{\Delta + x_k^2} \frac{\pi}{2N} \geq v = \int_{\beta_1}^{\beta_{i+1}} \ln \frac{1 - x^2}{\Delta + x^2} dx =$$

$$= \left(x \ln \frac{1 - x^2}{\Delta + x^2} + 2 \arcsin x - 2\sqrt{\Delta} \operatorname{arctg} \frac{x}{\sqrt{\Delta}} \right) \Bigg|_{\beta_1}^{\beta_{i+1}} \geq$$

$$> x \ln \frac{1 - x^2}{\Delta + x^2} \Bigg|_{\beta_1}^{\beta_{i+1}} + \frac{(\beta_{i+1} - \beta_1)\, \beta_1^2\, \beta_{i+1}^2}{8}$$

At this stage, we make use of the inequality

$$\left(\arcsin x - \sqrt{\Delta} \operatorname{arctg} \frac{x}{\sqrt{A}} \right) \Bigg|_{\beta_1}^{\beta_{i+1}} \geq \frac{\beta_1^2\, \beta_{i+1}^2 (\beta_{i+1} - \beta_1)}{8}$$

and obtain

$$r_i^N \geq \exp\left(\frac{2N}{\pi} z\right) \geq \left[\frac{1 - (\pi(2i + 1)/4N)^2}{\Delta + (\pi(2i + 1)/4N)^2} \right]^{i+1/2} \times$$

$$\times \left[\frac{\Delta + (\pi/4N)^2}{1 - (\pi/4N)^2} \right]^{1/2} \qquad (i \leq N_0(N)) .$$

It follows from Lemma 2 that, for small Δ and large N, the values r_i^N grow strongly for $i \leq N_0(N)$. Since the quantities r_i^N and q_i^N for the permutation π_2^N correspond to the quantities q_{N-i}^N and r_{N-i}^N for the permutation π_1^N, it follows from Lemmas 1 and

2 that

Corollary 1. If $\kappa_N = \pi_2^N$, then

$$\left\{\left[1 + \frac{\Delta}{\cos^2(\pi/4N)}\right]\left[1 + \frac{\Delta}{\cos^2(\pi(2j-1)/4N)}\right]\right\}^{j/2} \leq$$

$$\leq r_j^N \leq \left[1 + \frac{\Delta}{\cos^2(\pi j/4N)}\right]^j \qquad (1 \leq j \leq N)$$

and

$$q_j^N \geq \left[\Delta + \frac{\pi^2}{16N^2}\right]^{1/2} \left[\frac{1 - (\pi/4N)^2(2(N-j)+J-f}{\Delta + (\pi/4N)^2(2(N-j)+1)^2}\right]^{N-j+1/2}$$

for $j > N - N_0(N)$.

Let $N = 2n$. We examine the permutation

$$\pi_3^N = (n,\ n+1,\ n-1,\ n+2,\ \ldots\ldots\ 1,\ 2n).$$

Noting that $T_{2n}(z) = T_n(T_2(z))$ and using the above results, we see that

$$(13) \qquad u^{k+2} = u^k - \frac{8}{(M-m)(2\theta^2 - 1 - \cos^2\beta_k)} \times$$

$$x((M+m)I - A)(Au^k - f), \qquad k = 0, 2, \ldots\ldots 2n.$$

Hence, the functions $R_{2i}^N(t)$ and $Q_{2i}^N(t)$ for π_3^N are equal to the functions $R_i^{N/2}(t)$ and $Q_i^{N/2}(t)$ for the permutations $\pi_2^{N/2}$ applied

to the cyclic method (13) for the solution of

$$((M + m)I - A)Au = ((M + m)I - A)f.$$

Therefore, for r_{2j}^N and q_{2j}^N corresponding to π_3^N, the inequalities of Corollary 1 are valid; viz.

$$\left\{\left[1 + \frac{\Delta_1}{\cos^2(\pi/2N)}\right]\left[1 + \frac{\Delta_1}{\cos^2(\pi(2j-1)/2N)}\right]\right\}^{j/2} \leq$$

$$\leq r_{2j}^N \leq \left[1 + \frac{\Delta_1}{\cos^2(\pi j/2N)}\right]^j \quad (1 \leq j < N),$$

$$q_{2j}^N \geq \left[\Delta_1 + \frac{\pi^2}{4N^2}\right]^{1/2}\left[\frac{1 - (\pi/2N)^2((N-j)+1)^2}{\Delta_1 + (\pi/2N)^2((N-j)+1)^2}\right]^{N-j+1/2}$$

for $2j > N - N_0(N)$ where $\Delta_1 = \theta^2 - 1$. The permutation which corresponds to $\pi_1^{N/2}$ in an analogous way is

$$\pi_4^N = (2n, 1, 2n-1, 2, \ldots, n+1, n).$$

In this way, we see that, when using the permutations π_2^N and π_3^N in the exact iteration process (2), the $||u^k||$ are suitably located on the real axis, and $||\epsilon^k||$ in (7) is constantly decreasing, but that the error due to rounding can grow strongly. These permutations were proposed in [3], [5] - [7]. For the permutations π_1^N and π_4^N the norm of the initial errors decreases for subsequent iterations, however $||u^k||$ can grow strongly and this can lead to the growth of $||n^k||$ and further to the accidental stopping of the computations.

Let $N \in \{2^p, p = 0, 1, 2, \ldots\}$. For this case, we construct a permutation for which

$$q^N < 1, \qquad r^N < c_1. \tag{14}$$

We define a recursive procedure for the construction of the permutation $\varkappa_N$ for $N = 2^n$ which insures that (14) is satisfied. For $N = 1$, the solution is obvious. For $N = 2^{p-1}$ with $p-1 < n$, let the required permutation be

$$\varkappa_{2^{p-1}} = (j_1, j_2, \ldots, j_{2^{p-1}}),$$

then the permutation of order 2^p is defined inductively as

$$\varkappa_{2^p} = (j_1, 2^p + 1 - j_1, j_2, 2^p + 1 - j_2, \ldots, j_{2^{p-1}}, 2^p + 1 - j_{2^{p-1}}). \tag{15}$$

putting $p = 0, 1, 2, \ldots$ we obtain the required set of permutations. For example, for $N = 16$,

$$\varkappa_{16} = (1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11).$$

Below, we shall shot. that for this method the ordering is such that the operators $(I - \alpha_k A)$ with large norm are uniformly distributed among the operator,. which decrease the norm of the error.

We explain in a different way the mentioned procedure for the construction of $\varkappa_{2^n}$. In order to do this, we put $P_N(t)$ in the form

$$P_N(t) = \frac{T_N(z)}{T_N(\theta)} = \frac{2T^2_{N/2}(z) - 1}{2T^2_{N/2}(\theta) - 1} = \frac{T_{N/2}(z) - \cos\sigma_1}{T_{N/2}(\theta) - \cos\sigma_1} \times$$

$$\times \frac{T_{N/2}(z) - \cos\sigma_2}{T_{N/2}(\theta) - \cos\sigma_2},$$

where

$$\sigma_1 = \pi/4, \qquad \sigma_2 = \pi - \sigma_1 .$$

We assume that the roots of the polynomial $r_1^{(1)}(z) = T_{N/2} - \cos\sigma_1$ precede the roots of $r_2^{(1)} = T_{N/2} - \cos\sigma_2$. In a similar way, we replace each of the polynomials $r_1^{(1)}$ and $r_2^{(1)}$ by the product of two, for example

$$r_1^{(1)}(z) = \frac{(T_{N/4}(z) - \cos(\sigma_1/2))(T_{N/4}(z) - \cos(\pi - \sigma_1/2))}{(T_{N/4}(\theta) - \cos(\sigma_1/2))(T_{N/4}(\theta) - \cos(\pi - \sigma_1/2))} = r_1^{(2)} r_2^{(2)},$$

and again we assume that the roots of $r_1^{(2)}$ precede $r_2^{(2)}$. We continue this process and determine by this method a sequence of roots for each $r_k^{(i)}$ $(i = 1, 2)$ up to the point where the degree of the polynomials $r_k^{(i)}$ become equal to one. As a result, we obtain a permutation $\varkappa_N$. Observing that the roots x_i $(i = 1, 2, \cdots, k)$ of the equation $T_k(x) = \beta$ satisfy

$$x_1 = \omega_0 \cos\frac{2(i-1)\pi}{k} \mp \sqrt{(1 - \omega_0^2)} \sin\frac{2(i-1)\pi}{k} \quad (i = 1, \ldots) \; kj,$$

where $\omega_0 = \cos(y/k)$, $\gamma = \arccos \beta$, it can be shown that this permutation $\varkappa_N$ coincides with the permutation (15).

Before estimating r^N and q^N for the permutation (15), we establish a series of subsidiary inequalities:

a. If $\theta = 1 + 6$, $\delta > 0$, then

$$T_i(\theta) \geq 1 + i^2\delta + \frac{i^2(i^2 - 1)}{6}\,\delta^2. \tag{16}$$

In fact,

$$2T_i(\theta) = (\theta + \sqrt{(\theta^2 - 1)})^i + (\theta - \sqrt{(\theta^2 - 1)})^i =$$

$$= [1 + (6 + \sqrt{(2\delta + \delta^2)})]^i + [1 + (6 - \sqrt{(\delta^2 + 2\delta)})]^i =$$

$$= 2 + 2i\delta + \frac{i(i - 1)}{2}(46 + 4\delta^2) + \ldots$$

$$\ldots > 2 + 2i^2\delta + \frac{i^2(i^2 - 1)\delta^2}{3}.$$

b. If $0 \leq \omega \leq \pi/2$, $n \geq 2$, then

$$\frac{\sin^2(\omega/2) + \mu_{nk}}{\cos^2(\omega/2)}\;\frac{\sin^2[(\pi - \omega)/2n] + \mu_k}{\cos^2[(\pi - \omega)/2n]} > \tag{17}$$

$$> \frac{\sin^2(\omega/2n) + \mu_k}{\cos^2(\omega/2n)}\;\frac{\cos^2(\omega/2) + \mu_{nk}}{\cos^2(\omega/2)},$$

where $\mu_i = T_i(\theta) - 1$.

The inequality (17) is established once we show that

$$(18)\quad \operatorname{tg}\frac{\omega}{2}\operatorname{tg}\frac{\pi-\omega}{2n}\geq \operatorname{tg}\frac{\omega}{2n},$$

$$(19)\quad \frac{\mu_{nk}}{\cos^2(\omega/2)}\,\frac{\sin^2[(\pi-\omega)/2n]}{\cos^2[(\pi-\omega)/2n]}+$$

$$+\frac{\mu_k}{\cos^2[(\pi-\omega)/2n]}\,\frac{\sin^2(\omega/2)}{\cos^2(\omega/2)}\geq\frac{\mu_k}{\cos^2(\omega/2n)}+$$

$$+\frac{\sin^2(\omega/2n)}{\cos^2(\omega/2n)}\,\frac{\mu_{nk}}{\cos^2(\omega/2)}.$$

For a proof of inequality (18), we replace ω by $\pi/2-\alpha$ in it and obtain

$$(20)\quad \left(\cos\frac{\alpha}{2}-\sin\frac{\alpha}{2}\right)\sin\frac{\pi/2+\alpha}{2n}\cos\frac{\pi/2-\alpha}{2n}\geq$$

$$\geq\left(\cos\frac{\alpha}{2}+\sin\frac{\alpha}{2}\right)\sin\frac{\pi/2-\alpha}{2n}\cos\frac{\pi/2+\alpha}{2n}.$$

Transforming (20), we obtain

$$\left(\cos\frac{\alpha}{2}-\sin\frac{\alpha}{2}\right)\left(\sin\frac{\pi}{2n}+\sin\frac{\alpha}{n}\right)\geq$$

$$\geq\left(\cos\frac{\alpha}{2}+\sin\frac{\alpha}{2}\right)\left(\sin\frac{\pi}{2n}-\sin\frac{\alpha}{n}\right),$$

which yields

$$\cos\frac{\alpha}{2}\sin\frac{\alpha}{n} \geq \sin\frac{\alpha}{2}\sin\frac{\pi}{2n},$$

and thus,

$$\frac{\sin(\alpha/n)}{\sin(\pi/2n)} \geq \operatorname{tg}\frac{\alpha}{2}. \tag{21}$$

We note that for $\alpha = 0$ and $\alpha = \pi/2$ the inequality (21) becomes an equality. However, since $\operatorname{tg}(\alpha/2)$ lies above and $\sin(\alpha/n)\sin^{-1}(\pi/2n)$ lies below the line $y = 2\alpha/\pi$ on the interval $(0, \pi/2)$, it follows that (21) must be satisfied at any interior point of this interval, and consequently, that (18) is satisfied.

Inequality (19) is equivalent to the following inequality

$$\frac{\mu_{nk}}{\mu_k}\left(\sin^2\frac{\pi-\omega}{2n}\cos^2\frac{\omega}{2n} - \sin^2\frac{\omega}{2n}\cos^2\frac{\pi-\omega}{2n}\right) \geq \tag{22}$$

$$\geq \cos^2\frac{\omega}{2}\cos^2\frac{\pi-\omega}{2n} - \sin^2\frac{\omega}{2}\cos^2\frac{\omega}{2n}.$$

We show that (22) holds if the following is valid:

$$\frac{\mu_{nk}}{\mu_k}\left(\sin^2\frac{\pi-\omega}{2n} - \sin^2\frac{\omega}{2n}\right) \geq \cos\omega.$$

Since $\mu_{nk}/\mu_k > n^2$, the last inequality follows from

$$n^2\sin\frac{\pi}{2n}\sin\frac{a}{n} \geq \sin\alpha, \tag{23}$$

if $\alpha = \pi/2 - \alpha$. This inequality is valid, since for $\alpha \geq 0$ the inequality (23) reduces to an equality, and since the derivative with respect to α of the left hand side is always greater than that of the right. In this way, we have established inequality (17).

Lemma. If κ_N is defined by (15), and

$$N - i = 2^{j_1} + 2^{j_2} + \cdots + 2^{j_t},$$

where $j_1 > j_2 > \cdots > j_t \geq 0$, $1 \leq t \leq \log_2 (N-i+1)$, then

$$q_i^N \leq \prod_{i=1}^{t} 2(1 + T_{2j_i}(\theta))^{-1} \leq \left(1 + \frac{(N-i)^2}{t}\Delta\right)^{-1}.$$

Proof. It can be shown that

$$(24) \quad a_i^N = \max_{m \leq t \leq M} \left| \frac{T_{2j_1}(z) - \cos\xi_1}{T_{2j_1}(\theta) - \cos\xi_1} \cdot \frac{T_{2j_2}(z) - \cos\xi_2}{T_{2j_2}(\theta) - \cos\xi_2} \cdots \frac{T_{2j_t}(z) - \cos\xi_t}{T_{2j_t}(\theta) - \cos\xi_t} \right|,$$

where all $\xi_i \geq \pi/2$, and hence, the ith component in (24) does not exceed $2(1 + T_{2j_i}(\theta))^{-1}$ or, if (6) holds, does not exceed $(1 + 2^{2j_i}\Delta)^{-1}$. But since

$$\left(\sum_{i=1}^{t} |a_i|\right)^2 \leq t \sum_{i=1}^{t} a_i^2,$$

it follows that

$$\prod_{i=1}^{t} (1 + 2^{2j_i}\,\Delta) \geq 1 + \Delta \sum_{i=1}^{t} 2^{2j_i} \geq 1 + \Delta \left(\sum_{i=1}^{t} 2^{j_i} \right)^2 t^{-1} \geq$$

$$\geq 1 + \frac{(N - i)^2}{t}\,\Delta.$$

This proves the Lemma.

Lemma 4. If $\varkappa_N$ is defined by (15), and

$$i = 2^{i_1} + 2^{i_2} + \ldots + 2^{i_s},$$

where $i_1 > i_2 \geq \ldots > i_s \geq 0$, then

$$r_i^N \leq \left(\frac{\pi^2}{16N^2} + \Delta \right)^{-1} \left(1 + \frac{(i-1)^2}{s}\,\Delta \right)^{-1}. \tag{25}$$

Proof. It can be shown that

$$|R_i^N(t)| = \left| \prod_{k=1}^{s} \frac{T_{2^{i_k}}(z) - \cos\sigma_k}{T_{2^{i_k}}(\theta) - \cos\sigma_k} \right|, \tag{26}$$

where all $\sigma_k < \pi/2$ and $\sigma_1 = \pi/2.2^{n-i_1}$, $\sigma_{k+1} = \pi - \sigma_k/2^{i_k - i_{k+1}}$. For $s \geq 1$, taking (16) and (17) into account, we obtain from (26) that

$$r_i^N \leq \max_{-1\leq z\leq 1} \left| \frac{T_{2^{i_s}}(z) - \cos\bar{\sigma}_s}{T_{2^{i_s}}(\theta) - \cos\bar{\sigma}_s} \right| \max_{-1\leq z\leq 1} \prod_{k=1}^{s-1} \left| \frac{T_{2^{i_k}}(z) + \cos\sigma_k}{T_{2^{i_k}}(\theta) - \cos\sigma_k} \right| \leq \tag{27}$$

$$\leq \left(\operatorname{tg}^2 \frac{\bar{\sigma}_s}{2} + 2^{2i_s}\,\Delta \right)^{-1} \prod_{k=1}^{s-1} 2\left(1 + T_{2^{i_k}}(\theta)\right)^{-1},$$

where $\sigma_s = \pi/2.2^{n-i_s}$. We make use of the method of estimation contained in the proof of Lemma 3. For $i_s = 0$ and $s > 1$, (25) follows at once from (26).

For $i_s > 0$, taking account of' the following inequality

$$tg^2 \frac{\pi}{2.2^{n-i_s}} + \frac{1}{2}(T_{2^{i_s}}(\theta) - 1) \geq \left(\frac{\pi^2}{16N^2} + \Delta\right)(1 + 2^{2i_s}\Delta),$$

which can be verified by means of inequality (16), we obtain from (27) that

$$r_i^N \leq \left(\frac{\pi^2}{16N^2} + \Delta\right)^{-1} \prod_{k=1}^{s} (1 + 2^{2i_k}\Delta)^{-1} <$$

$$\leq \left(\frac{\pi^2}{16N^2} + \Delta\right)^{-1} \left(1 + \frac{i^2}{s}\Delta\right)^{-1}.$$

In this way, the Lemma is proved.

Corollary 2. If

$$\varkappa_N = (N+1 - j_{N/2}, j_{N/2}, \ldots, N+1 - j_2, j_2, N+1 - j_1, j_1),$$

where j_k is defined by the recursive method (15), then

$$q_i^N < \left(\frac{\pi^2}{16N^2} + \Delta\right)^{-1} \left(1 + \frac{(N - i - 1)^2}{s}\Delta\right)^{-1},$$

$$r_i^N \leq \left(1 + \frac{i^2}{t}\Delta\right)^{-1}.$$

Corollary 3. If the order of the choice of' the α_k correspond:: to the permutation (15), then for $k > i$ we have

$$\max_{m \le t \le M} \left| \prod_{j=i}^{k} (1 - \alpha_j t) \right| \le r_i^N \quad .$$

Lemma 5. If the order of the choice of the α_k corresponds to the permutation (15), then

$$(28)\quad t_i^N \le \frac{16}{\pi^2} \frac{N^2}{T_N(\theta)} \le \frac{16}{\pi^2} \left(N^{-2} + \delta + \frac{\delta^2}{6} (8 - 1) \right)^{-1} .$$

In fact, since $R_i^N(t)$ and $Q_i^N(t)$ always contain a polynomial of the form

$$T_{2^{i+1}}(z) \pm \cos(\pi/2^{i+1}) \qquad (i = n-1, \ldots, m^*, n_0 > 0)$$

then

$$t_i^N \le 2N \prod_{i=1}^{n} \left(1 + \cos \frac{\pi}{2^{i+1}} \right) T_N^{-1}(\theta) = 2N^2 \prod_{i=1}^{n} \cos^2 \frac{\pi}{2^{i+2}} \cos^2 \frac{\pi}{2^{n+2}} .$$

But

$$\prod_{i=1}^{n} \cos \frac{\pi}{2^{i+2}} = \frac{\sin(\pi/4)}{\pi/4} \frac{\pi/2^{n+2}}{\sin(\pi/2^{i+2})} = \frac{1}{\sqrt{2} \cdot 2^n \sin(\pi/2^{n+2})} ,$$

and hence,

$$t_i^N \le \frac{1}{tg^2(\pi/2^{n+2}) T_N(\theta)} \le \frac{16}{\pi^2} \frac{N^2}{T_N(\theta)} .$$

ding inequality (16), we obtain (28).

Lemma 6. Let $\nu^N = \sum_{i=1}^{N} q_i^N$. Then

$$\nu^N \le \exp\left(\frac{1}{1+\delta} + \delta/\ln 4\right) \delta^{-1/\ln 4} \tag{29}$$

Proof. Along with ν^N we examine $\nu^{N/2}$. It is easy to see that

$$\nu^N \le \left(1 + \frac{T_{N/2}(\theta) - 1}{2}\right)^{-1} \nu^{N/2} + \nu^{N/2} = \nu^{N/2}\left(1 + \frac{2}{T_{2i}(\theta) + 1}\right) \le$$

$$\le \prod_{i=0}^{n-1}\left(1 + \frac{2}{T_{2i}(\theta) + 1}\right)$$

However,

$$\ln \prod_{i=0}^{n-1}\left(1 + \frac{2}{T_{2i}(\theta) + 1}\right) \le \sum_{i=0}^{n-1} \frac{2}{T_{2i}(\theta) + 1} \le \sum_{i=0}^{n-1} (1 + 2^{2i}\delta)^{-1}.$$

Together with this, we have

$$\sum_{i=0}^{n-1} (1 + 4^i\delta)^{-1} \le (1 + \delta)^{-1} + \int_0^{n-1} (1 + 4^x\delta)^{-1}\, dx \le$$

$$\le \frac{\delta - \ln\delta}{\ln 4} + (1 + \delta)^{-1},$$

which therefore establishes the validity of (29).

In the Table, values for r_1^{16} and q_i^{16} corresponding to $\varkappa_{16}$, defined by (15), are given with m/M = 0.01.

i	r_i^{16}	q_i^{16}	i	r_i^{16}	q_i^{16}
1	79.8	0.418	9	27.0	0.761
2	19.6	0.423	10	5.63	0.768
3	9.59	0.432	11	5.14	0.790
4	4.63	0.440	12	0.601	0.803
5	28.0	0.479	13	7.66	0.940
6	2.68	0.485	14	1.27	0.950
7	7.98	0.511	15	2.18	0.986
8	0.907	0.518	16	0.0812	---

For a more detailed study of the iteration process(2), it is necessary to introduce a <u>priori</u> assumptions about the nature of h^k. Concentrating on the situations which cause the iteration process to behave unfavorably {for example, when $_I|u^k|_I$ is larger or smaller than m/M) and taking into account the final ordering which insures that $||h^k||$ is proportional to $||u^k||$ for $||u^k|| \gg ||u||$, we make the assumption that the errors belong to the class

$$D = \{h^k : h^k = (k_1 + k_2 r_k^N) w^k, \; ||w^k|| < C\},$$

where C, k_1 and k2 {for example, $k_1 = O(||u||)$ and $k_2 = O(||\epsilon^0||)$} are positive constants which are independent of N but depend on m/M, the ordering in the computer and the implementation of (2).

Then, if $w^k = \Sigma w_n^k \varphi_n$, it follows that

$$\epsilon_n^N = R_i^N(\lambda_n)\epsilon_n^0 - \nu_n^N,$$

where

$$\nu_n^N = \sum_{i=1}^{N-1} Q_i^N(\lambda_n)(k_1 + k_2 r_i^N) w_n^i + (k_1 + k_2 r_N^N) w_n^N$$

Consequently,

$$||\nu^N|| \leq C \left(k_1 \sum_{i=1}^{N} q_i^N + k_2 \sum_{i=1}^{N} t_i^N \right) .$$

Taking the results of Lemmas 5 and 6 into account, we are led to the conclusion that if $n^k \in D$, then the calculation is stable.

We have also examined the ordering of the coefficients α_k when the spectrum $\sigma(A)$ lies on p intervals, the ends of which satisfy the conditions of [8] {see Appendix 1}. In this case, there exists a polynomial of degree p, $Q_p(t)$, with $Q_p(0) = 0$, which maps all the intervals onto one $[m, M]$ with $mM > 0$. Let $N = p2^n$, and $\varkappa_2^n$ be a permutation of the form (15). We denote by τ_i $(i=1, 2, \ldots, 2^n)$ the coefficients of the cyclic method of ordering for $[m, M]$ according to $\varkappa_2^n$. Now, we put $\alpha_k = \mu_s^{-1}(\tau_i)$, where

$$k = (i-1)p+s \qquad (i=1, 2, \bullet\,**, 2^n, s=1, 2, \bullet\,*=, p)$$

and the $\mu_s(\tau_i)$ are the roots of the equation

$$\tau_i Q_p(t) = 1$$

which are ordered with respect to increasing modulus.

It is clear that the above results can be extended to iteration methods of the following type {see [9]}

$$Bu^{k+1} = Bu^{k} - \alpha_k(Au^{k} - f).$$

Received by the Editors: 25.03.1970

REFEREPJCES

[1] M. Sh. Birman, On a variant of the method of successive approximations, Vestn. LGU 9(1952), 69-76.

[2] M. K. Gavurin, The use of polynomials of best approximation for the improvement of the convergence of iteration processes, Uspekhi Matem Nauk 5(3) (1950), 156-160.

[3] D. Young, On Richardson's method for solving linear systems with positive definite matrices, J. Math and Phys. 32(4)(1954), 243-255.

[4] D. K. Faddeev, On sequences of polynomials useful for the construction of iteration methods for the solution of systems of linear algebraic equations, Vestn. LGU 7(1958), 155-159.

[5] D. K. Faddeev and V. N. F'addeeva, Computational Methods of Linear Algebra, Fizmatgiz, 1963.

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

[7] D. Young and C. H. Warlick, On the use of Richardson's method for the numerical solution of Laplace's equation on the ORDVAC, Memorandum Rept. No. 707, Ballistics Res. Lab., Aberdeen Proving Ground, Md., 1953.

[8] V. I. Lebedev, Iterative methods for solution of operator equations with their spectrum lying on several intervals, Zh. Vychisl. Matem. i Matem. Fiz 9(6)(1969), 1247-1252.

[9] E. G. D'yakonov, On the construction of iterative methods based on the use of operators with equivalent spectra, Zh. Vychisl. Matem. i Matem. Fiz 6(1)(1966), 12-34.