UNIVERSITY OF NIŠ
FACULTY OF ELECTRONIC ENGINEERING

NUMERICAL METHODS AND APPROXIMATION THEORY

Niš, September 26-28, 1984

Edited by G. V. Milovanović

Niš, 1984

UNIVERSITY OF NIŠ

FACULTY OF ELECTRONIC ENGINEERING

NUMERICAL METHODS AND APPROXIMATION THEORY

Niš, September 26-28, 1984

Edited by G. V. Milovanović

Niš, 1984

ORGANIZING COMMITTEE

Chairman: G.V. Milovanović (Niš)

Secretary: M.A. Kovačević (Niš)
Members: P.M. Vasić (Beograd)
B.M. Damnjanović (Kragujevac)
R.Ž. Djordjević (Niš)
P.B. Madić (Beograd)
I.Ž. Milovanović (Niš)
M.S. Petković (Niš)
Lj.R. Stanković (Niš)
D.Dj. Tošić (Beograd)
D. Herceg (Novi Sad)

This publication was in part supported by the Regional Association for Science - Niš and University of Niš.

Published by: Faculty of Electronic Engineering, Beogradska 14,
P.O. Box 73, 18000 Niš, Yugoslavia

Printed by: Prosveta, Niš

Number of copies: 300

PREFACE

The conference "Numerical Methods and Approximation Theory" was held at the Faculty of Electronic Engineering, University of Niš, September 26-28, 1984. It was attended by 46 mathematicians from several universities.

These proceedings contain most of the papers presented during the conference in the form in which they were submitted by the authors. Typing, grammatical and other errors were not, except in some isolated cases, edited out of the received material.

The topic treated cover different problems on numerical analysis and approximation theory.

September 1984 G.V. Milovanović

CONTENTS

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

GAUSSIAN ELIMINATION FOR DIAGONALLY DOMINANT MATRICES

Zvonimir Bohte, Marko Petkovšek

ABSTRACT:
Wilkinson [1] *proved that the property of columnwise diagonal dominancy is preserved during the Gaussian elimination. This is true only for exact arithmetic. In this paper a corresponding theorem for floating point arithmetic is proved.*

GAUSOVA ELIMINACIJA ZA DIJAGONALNO DOMINANTNE MATRICE.
Wilkinson [1] *je dokazao da se osobina dijagonalne dominantnosti po kolonama u toku Gausove eliminacije ne narušava. To je tačno samo za egzaktnu aritmetiku. U ovom radu je dokazana odgovarajuća teorema za aritmetiku u pomičnom zarezu.*

1. INTRODUCTION

Let A be a real square matrix of order n. The Gaussian elimination for the solution of the system of linear equations

$$Ax = b$$

yields a set of equivalent systems

$$A^{(r)}x = b^{(r)}\ , \quad r = 1,\dots,n$$

where $A^{(1)} = A$, $b^{(1)} = b$ and $A^{(n)}$ is an upper triangular matrix. The matrix $A^{(r)}$ has the following block structure

$$(1) \qquad A^{(r)} = \begin{bmatrix} U_r & X_r \\ \emptyset & A_r \end{bmatrix}$$

where U_r is an upper triangular matrix of order r-1 and A_r a square matrix of order n-r+1.

Wilkinson [1] proved: If the original matrix A is columnwise diagonally dominant, i.e. if

$$|a_{kk}| > \sum_{\substack{i=1 \\ i\neq k}}^{n} |a_{ik}|\ , \quad k = 1,\dots,n$$

then the matrix A_r is also columnwise diagonally dominant, i.e.

$$|a_{kk}^{(r)}| > \sum_{\substack{i=r\\ i\neq k}}^{n} |a_{ik}^{(r)}| \ , \quad k = r,\ldots,n$$

for all r = 2,...,n-1. He also proved that

$$\max_{i,k,r} |a_{ik}^{(r)}| \leq 2.\max_{i,k} |a_{ik}| .$$

Unfortunately, the presence of rounding errors may destroy the original diagonal dominancy. Therefore, to ensure the nonfailure of the method it is necessary to require more than just a mere diagonal dominancy.

In the analysis of rounding errors we shall use the equation

$$(2) \qquad fl(x\circ y) = (x\circ y)(1 + e) \ , \quad |e| \leq u$$

where x and y are any standard floating point numbers and $fl(x\circ y)$ denotes the computed result of any of the four arithmetic operations. We shall suppose that the relative error of an arithmetic operation is bounded by unit rounding error which is normally

$$u = b^{1-t}/2 \quad \text{(for rounding)}$$
$$= b^{1-t} \quad \text{(for chopping)}$$

where t is the length of the mantissa in the base b (usually 2 or 10). It is of course assumed also that during the computation no overflow or underflow occurs.

In the following we shall leave out all work with the right-hand sides.

2. THE ALGORITHM AND ERROR ANALYSIS

We denote the current calculated matrix at the r-th step by $B^{(r)}$. It has the same block structure as the matrix (1)

$$B^{(r)} = \begin{bmatrix} V_r & Y_r \\ \emptyset & B_r \end{bmatrix}$$

t is assumed that the matrix $A = B^{(1)}$ is the matrix stored n the computer.

The algorithm for the calculation of the upper triangular matrix $B^{(n)}$ is as follows:

```
        $r = 1,\ldots,n-1$:
            $i = r+1,\ldots,n$:
(3)             $m_{ir} = fl(b_{ir}^{(r)}/b_{rr}^{(r)})$
                $k = r+1,\ldots,n$:
(4)                 $b_{ik}^{(r+1)} = fl(b_{ik}^{(r)} - fl(m_{ir}b_{rk}^{(r)}))$
```

Let us denote

$$h_r = \max|b_{ik}^{(r)}| \ , \quad i,k = r,\ldots,n \tag{5}$$

and

$$h = \max h_r \ , \quad r = 1,\ldots,n \tag{6}$$

Using (2) in (3) and (4) we have

$$m_{ir} = q_{ir}(1 + x_{ir}) \ , \quad i = r+1,\ldots,n$$

and

$$b_{ik}^{(r+1)} = (b_{ik}^{(r)} - m_{ir}b_{rk}^{(r)}(1 + y_{ik}^{(r)}))(1 + z_{ik}^{(r)}) \ , \quad i,k = r+1,\ldots,n \tag{7}$$

where

$$q_{ir} = b_{ir}^{(r)}/b_{rr}^{(r)} \tag{8}$$

and

$$|x_{ir}|, \ |y_{ik}^{(r)}|, \ |z_{ik}^{(r)}| \leqq u \tag{9}$$

Let us suppose that

$$|q_{ir}| \leqq 1 \tag{10}$$

We can write equation (7) in the form

$$b_{ik}^{(r+1)} = b_{ik}^{(r)} - q_{ir}b_{rk}^{(r)} + d_{ik}^{(r)} \ , \quad i,k = r+1,\ldots,n \tag{11}$$

where

$$d_{ik}^{(r)} = -q_{ir}b_{rk}^{(r)}(x_{ir} + y_{ik}^{(r)} + x_{ir}y_{ik}^{(r)})(1 + z_{ik}^{(r)}) + \\ +(b_{ik}^{(r)} - q_{ir}b_{rk}^{(r)})z_{ik}^{(r)}$$

Then we can obtain the bound for $d_{ik}^{(r)}$ using (5), (9) and (10)

$$|d_{ik}^{(r)}| \leqq h_r(2u+u^2)(1+u) + 2uh_r = (4 + 3u + u^2)uh_r \tag{12}$$

Now, we can formulate the theorem.

3. THE THEOREM

Let A be a columnwise diagonally dominant matrix of order n and furthermore, let

$$(13) \qquad |a_{kk}| > \sum_{\substack{i=1 \\ i\neq k}}^{n} |a_{ik}| + cun(n-1)|a_{kk}| \ , \quad k = 1,\ldots,n$$

where $c = 4 + 3u + u^2$, and u is the unit rounding error.

Then the following is true for $r = 1,\ldots,n$:

(i) the matrix B_r is columnwise diagonally dominant and furthermore,

$$|b_{kk}^{(r)}| > \sum_{\substack{i=r \\ i\neq k}}^{n} |b_{ik}^{(r)}| + cu(n-r+1)(n-r)|a_{kk}| \ , \quad k = r,\ldots,n$$

$$\text{(ii)} \quad \sum_{i=r}^{n} |b_{ik}^{(r)}| \leqq \sum_{i=1}^{n} |a_{ik}| + cu(2n-r)(r-1)|a_{kk}| \ , \quad k = r,\ldots,n$$

$$\text{(iii)} \quad |b_{ik}^{(r)}| \leqq (2 - cu(n-r+1)(n-r))|a_{kk}| \ , \quad i,k = r,\ldots,n$$

PROOF. We shall prove the theorem by the mathematical induction with respect to r. Let $r = 1$. Then, since $B^{(1)} = B_1 = A$, proposition (i) coincides with (13). Obviously, (13) implies that $cun(n-1) < 1$. Therefore, (ii) and (iii) hold trivially for $r = 1$.

Let propositions (i) - (iii) hold for some r, $1 \leqq r \leqq n-1$, and let $r+1 \leqq k \leqq n$. From (11) and (8) we obtain

$$(14) \qquad \sum_{\substack{i=r+1 \\ i\neq k}}^{n} |b_{ik}^{(r+1)}| \leqq |b_{rk}^{(r)}|/|b_{rr}^{(r)}| \sum_{\substack{i=r+1 \\ i\neq k}}^{n} |b_{ir}^{(r)}| + \sum_{\substack{i=r+1 \\ i\neq k}}^{n} |b_{ik}^{(r)}| + \sum_{\substack{i=r+1 \\ i\neq k}}^{n} |d_{ik}^{(r)}|$$

From (i) and (8) it follows that the inequality (10) holds. Therefore, we can use the bound (12) in (14). From (i) it follows

$$\sum_{\substack{i=r+1 \\ i\neq k}}^{n} |b_{ir}^{(r)}| < |b_{rr}^{(r)}| - |b_{kr}^{(r)}|$$

rom (14) we have

$$\begin{aligned}\sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r+1)}| &\leqq |b_{rk}^{(r)}|(|b_{rr}^{(r)}| - |b_{kr}^{(r)}|)/|b_{rr}^{(r)}| + \\ &\quad + \sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r)}| + cuh_r(n-r-1) = \\ &= \sum_{\substack{i=r\\ i\neq k}}^{n} |b_{ik}^{(r)}| - |q_{kr}||b_{rk}^{(r)}| + cuh_r(n-r-1)\end{aligned}$$

Finally, from (i), (iii), (11) and (12) it follows

$$\begin{aligned}\sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r+1)}| &< |b_{kk}^{(r)}| - cu(n-r+1)(n-r)|a_{kk}| - \\ &\quad - |q_{kr}||b_{rk}^{(r)}| + 2cu(n-r-1)|a_{kk}| \leqq \\ &\leqq |b_{kk}^{(r+1)} - d_{kk}^{(r)}| - cu((n-r)(n-r-1) + 2)|a_{kk}| \leqq \\ &\leqq |b_{kk}^{(r+1)}| + 2cu|a_{kk}| - cu((n-r)(n-r-1) + 2)|a_{kk}| \leqq \\ &\leqq |b_{kk}^{(r+1)}| - cu(n-r)(n-r-1)|a_{kk}|\end{aligned}$$

which proves (i).

To prove (ii), note that

$$\sum_{i=r+1}^{n} |q_{ir}| < 1 \tag{15}$$

because B_r is columnwise strictly diagonally dominant. Therefore, (11), (12), (15) and (iii) imply that

$$\begin{aligned}\sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| &\leqq \sum_{i=r+1}^{n} |b_{ik}^{(r)}| + |b_{rk}^{(r)}| \sum_{i=r+1}^{n} |q_{ir}| + \sum_{i=r+1}^{n} |d_{ik}^{(r)}| \leqq \\ &\leqq \sum_{i=r}^{n} |b_{ik}^{(r)}| + 2cu(n-r)|a_{kk}|\end{aligned}$$

Then, using (ii) it follows

$$\begin{aligned}\sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| &\leqq \sum_{i=1}^{n} |a_{ik}| + cu(2n-r)(r-1)|a_{kk}| + \\ &\quad + 2cu(n-r)|a_{kk}| = \\ &= \sum_{i=1}^{n} |a_{ik}| + cu(2n-r-1)r|a_{kk}|\end{aligned} \tag{16}$$

and we have obtained the same inequality (ii) in which r is replaced by r+1.

If we proceed and use the inequality (13) in (16) we get

$$\sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| \leq 2|a_{kk}| - cun(n-1)|a_{kk}| + cu(2n-r-1)r|a_{kk}| =$$

$$= 2|a_{kk}| - cu(n-r)(n-r-1)|a_{kk}|$$

Therefore, for each pair $i,k = r+1,\dots,n$

$$|b_{ik}^{(r+1)}| \leq (2 - cu(n-r)(n-r-1))|a_{kk}|$$

which proves (iii).

4. CONCLUSIONS

The assumptions of the Theorem are sufficient to ensure that the Gaussian elimination in floating point cannot break down. All the quotients m_{ir} are bounded in modulus by 1 and the pivotal growth of the computed elements is bounded by 2. Therefore, in view of Wilkinson's error analysis [1] the Gaussian elimination for matrices which satisfy (13) is numerically stable.

The Theorem also enables us to determine the minimal length of the mantissa which ensures that the breakdown of the Gaussian elimination cannot occur. Let the matrix A be such that

$$|a_{kk}| \geq d \sum_{\substack{i=1 \\ i \neq k}}^{n} |a_{ik}| , \quad k = 1,\dots,n$$

The following table shows the minimal length of the mantissa in dependance on d and n with rounding in base 10.

	minimal length of the mantissa		
d	n = 5	n = 10	n = 100
1·001	6	7	9
1·01	5	6	8
1·1	4	5	7
1·5	4	4	6
2	3	4	6

REFERENCES:

1. WILKINSON J.H.: *Error analysis of direct methods of matrix inversion.* J. ACM 8 (1961), 281 - 330.

erical Methods and
Approximation Theory
Niš, September 26-28, 1984

N SOME NUMERICAL PROPERTIES OF INFINITE-DIMENSIONAL SIMPLEX

Miloš M. Laban

BSTRACT:

tarting from an analytic model of infinite-dimensional simlex in Banach space, the possibility of it`s good approximaion by one of it`s finite-dimensional subsimplexes is obseved. The class of simplexes, where such a approximation is ossible to eighter make or not are established by a seque of theorems. Herefrom, the members of the class of limited nfinite-dimensional simplexes with vertices making the ortogonal system, could not be approximated on such a way.

NEKIM NUMERIČKIM OSOBINAMA BESKONAČNO-DIMENZIONOG SIMPLEKA. Polazeći od analitičkog modela beskonačno-dimenzionog simpleksa u Banahovom prostoru, ispituje se mogućnost njegove dobre aproksimacije jednim njegovim konačno-dimenzionim podsimpleksom. Nizom teorema utvrđuju se klase simpleksa kod kojih je takva aproksimacija moguća i one kod kojih nije moguća. Tako se dobija da oni iz klase ograničenih beskonačno-dimenzionih simpleksa čija temena čine ortogonalan sistem, ne mogu biti aproksimirani na pomenuti način.

1. BASIC IDEAS

A finite-dimensional simplex in mathematics and applications is widely threated notion. There exists a great number of articles on analytic-geometrical properties of a n-dimensional simplex and, consequently, numerical applications. The notion of infinite-dimensional simplex is introduced by Bastiani in [1], and is developed in topological sense by Maserick in [4], Phelps in [5], Lau in [3] and Höllein in [2]. For a difference of such a direction, we shall deal with the analytic-geometrical approach to this notion, keeping on mind that infinite-dimensional simplex would be natural generalization of a finite-dimensional case as much as possible. At the same time, we shall insist on the results which are suitable for the numerical practice.

At first, we shall show that it is possible to make such a construction in at least infinite-dimensional Banach space.

Theorem 1: Let X be Banach space and let $x_0, x_1, \dots, x_n, \dots$ be such a vectors in X that $\{x_1 - x_0, \dots, x_n - x_0, \dots\}$ is the infinite unconditional set of linearly independent vectors. Let us denote

$$S = \left\{ \sum_{n=0}^{+\infty} \theta_n x_n \;\middle|\; \sum_{n=0}^{+\infty} \theta_n = 1 \;;\; \theta_0, \theta_1, \theta_2, \dots \geqslant 0 \;;\; \sum_{n=0}^{+\infty} \theta_n x_n \text{ converges} \right\}$$

$$T = \left\{ \sum_{n=0}^{k} \theta_{i_k} x_{i_k} \;\middle|\; \sum_{n=0}^{k} \theta_{i_k} = 1 \;;\; \theta_{i_0}, \dots, \theta_{i_k} \geqslant 0 \;; \{i_0, \dots, i_k\} \subset \subset \{0, 1, 2, \dots\} \right\}$$

Then $\overline{S} = \overline{T}$, where A denotes the closure of set A.

Proof: 1° Let y be an arbitrary vector from $\overline{S}$. Then there exists sequence $(y_j)(j=1,2,\dots)$ of vectors from S such that

$$(1) \qquad \lim_{j \to +\infty} y_j = y \quad \left(y_j = \sum_{n=0}^{+\infty} {}^j\theta_n x_n \;;\; \sum_{n=0}^{+\infty} {}^j\theta_n = 1 \;;\; {}^j\theta_n \geqslant 0 \right)$$

states. Let us denote

$$y_j' = \sum_{n=0}^{j-1} {}^j\theta_n x_n + \left(1 - \sum_{n=0}^{j-1} {}^j\theta_n\right) x_j \quad (j = 2, 3, \dots)$$

It is easy to verify that

$$(2) \qquad y_j' \in T \quad (j = 2, 3, \dots) \quad .$$

Further on, we have

$$(3) \quad \|y_j' - y_j\| = \left\| \left(1 - \sum_{n=0}^{j} {}^j\theta_n\right) x_j - \sum_{n=j+1}^{+\infty} {}^j\theta_n x_n \right\| \leqslant \left(1 - \sum_{n=0}^{j} {}^j\theta_n\right) \|x_j\| + + \left\| \sum_{n=j+1}^{+\infty} {}^j\theta_n x_n \right\| \quad .$$

Since $\sum_{n=0}^{+\infty} {}^j\theta_n = 1$, it follows

$$(4) \qquad \lim_{j \to +\infty} \left(1 - \sum_{n=0}^{j} {}^j\theta_n\right) = 0 \quad ,$$

and by $\sum_{n=0}^{+\infty} {}^j\theta_n x_n = y_j$ we obtain

$$(5) \qquad \lim_{j \to +\infty} \left\| \sum_{n=j+1}^{+\infty} {}^j\theta_n x_n \right\| = 0 \quad .$$

If we, now, let $j \to +\infty$ in (3), then accordingly to (4) and (5) we have

$$\lim_{j\to+\infty} \| y'_j - y_j \| = 0 \quad ,$$

wherefrom and (1) it follows

$$\lim_{j\to+\infty} y'_j = y \quad .$$

Herefrom, with the regard to (2), we obtain $y \in \overline{T}$. Consequently $\overline{S} \subseteq \overline{T}$.

2^o Let z be an arbitrary vector from $\overline{T}$. Then there exists sequence $(z_j)(j=1,2,\dots)$ of vectors from T such that

$$\lim_{j\to+\infty} z_j = z \; \Big(\; z_j = \sum_{n=0}^{k_j} {}^j\eta_n {}^j x_n \; ; \; \sum_{n=0}^{k_j} {}^j\eta_n = 1 \; ; \; {}^j\eta_0, \dots, {}^j\eta_{k_j} \geq 0 ; \; ; \{ {}^j x_0, \dots, {}^j x_{k_j} \} \subset \{ x_0, x_1, \dots \} \Big)$$

states. Let us denote

$$z'_j = \sum_{n=0}^{+\infty} {}^j\eta'_n \cdot x_n \quad ,$$

where

$${}^j\eta'_n = \begin{cases} {}^j\eta_i & , \; x_n = {}^j x_i \\ 0 & , \; x_n \notin \{ {}^j x_0, \dots, {}^j x_{k_j} \} \end{cases} \quad .$$

It is obvious that $z'_j \in S$ $(j=1,2,\dots)$. Since $z'_j = z_j$, it follows $\lim_{j\to+\infty} z'_j = z$, hence $z \in \overline{S}$. Consequently $\overline{T} \subseteq \overline{S}$ and the proof is completed.

This theorem allows us to use the following notion of a infinite-dimensional simplex:

Definition 1: Set $\overline{S}$ we shall call the infinite-dimensional simplex (IDS in the further text) with vertices $x_0, x_1, x_2, \dots$ and denote $S(x_0, x_1, x_2, \dots)$. At the same time set

$$p(x_{j_1}, x_{j_2}, \dots, x_{j_k}, \dots) \stackrel{def}{=} \overline{x_{j_1} + L(x_{j_2} - x_{j_1}, \dots, x_{j_k} - x_{j_1}, \dots)}$$

(L denotes lineal) we shall call the face of $S(x_0, x_1, x_2, \dots)$, if $\{ x_{j_1}, x_{j_2}, \dots, x_{j_k}, \dots \}$ is the finite (or infinite) set of different vectors which is subset of $\{ x_0, x_1, x_2, \dots \}$.

The following theorem(obviously true) points out that such a notion of IDS keeps a number of very important and for application ruther useful properties of it's finite-dimensional generator.

Theorem 2: Let $\{x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots\}$ be a finite (or infinite) set of different vectors which is subset of $\{x_o, x_1, \ldots\}$. Then:

$$1^\circ \quad S(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) \subseteq S(x_o, x_1, x_2, \ldots) \ ;$$

$$2^\circ \quad S(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) \subset p(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) \ ;$$

$$3^\circ \quad p(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) = p(x_{j_2}, x_{j_1}, \ldots, x_{j_k}, \ldots) \ ;$$

$$4^\circ \quad p(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) = x_{j_1} + \overline{L(x_{j_2} - x_{j_1}, \ldots, x_{j_k} - x_{j_1} \ \ldots)}.$$

2. APPROXIMATION

Let $S(x_o, x_1, \ldots)$ be an IDS. Naturally, the possibility of replacing such a simplex with a finite-dimensional one (FDS in the further text) is of the great importance.

At first, if $\sup\|x_n\| = +\infty$, then $S(x_o, x_1, \ldots)$ is unlimited set and, consequently, it is not possible to replace it with an FDS which is necessary a limited set. If $\sup\|x_n\| < +\infty$, then we have the following results:

Theorem 3: Let $\{x_o, x_1, \ldots\}$ be a orthogonal set and $\inf\|x_n\| = =\alpha > 0$. Then for each vector y from an arbitrary finite-dimensional subsimplex there exists a set $Y(y)$ such that the following conditions are fulfilled:

1° $Y(y) \subseteq S(x_o, x_1, \ldots)$;

2° $Y(y)$ is itself an IDS ;

3° $(\forall x)(x \in Y(y))(\|x - y\| \geq \frac{\alpha}{2})$.

Proof: Without loosing the generality in proof, we can observe FDS $S(x_o, x_1, \ldots, x_k)$ and such $y \in S(x_o, x_1, \ldots, x_k)$ that

$$y = \sum_{n=o}^{k} \varphi_n x_n \quad \Big(\sum_{n=o}^{k} \varphi_n = 1 \ ; \ \varphi_n \geq 0 \ (n = 0, 1, \ldots, k) \Big)$$

where $\varphi = \varphi_k = \max\{\varphi_n \mid n = 0, 1, \ldots, k\}$.

Case 1: $\varphi \geq \frac{1}{2}$. Then $Y(y) = S(x_{k+1}, x_{k+2}, \ldots)$.

Really, let $x \in S(x_{k+1}, x_{k+2}, \ldots)$. Since

$$x-y\|^2 = \|x\|^2 + \|y\|^2 \geq \|y\|^2 \geq \varphi_k^2\cdot\|x_k\|^2 \geq \frac{1}{4}\alpha^2 \quad ,$$

; follows 3^o.

Case 2: $\varphi < \frac{1}{2}$ .Then $Y(y)=S(\{(1-\varphi)x_{k+1}+\varphi x_{k+2},(1-\varphi)x_{k+1}+\varphi x_{k+3},\dots\})$ .Let us denote

$$x' = \sum_{n=k+1}^{+\infty} \varphi'_n((1-\varphi)x_{k+1}+\varphi x_{n+1}) \ ,$$

here $\sum\limits_{n=k+1}^{+\infty} \varphi'_n = 1$, $\varphi'_n \geq 0$ $(n=k+1,k+2,\dots)$.Since

$$x' = (\sum_{n=k+1}^{+\infty}\varphi'_n)(1-\varphi)x_{k+1} + \sum_{n=k+1}^{+\infty}\varphi'_n\cdot\varphi x_{n+1} = (1-\varphi)x_k + \sum_{n=k+1}^{+\infty}\varphi'_n\cdot\varphi\cdot x_{n+1}$$

t follows $x' \in S(x_o,x_1,\dots)$,becouse

$$(1-\varphi)+\sum_{n=k+1}^{+\infty}\varphi'_n\varphi = 1-\varphi+\varphi = 1 \quad .$$

n the base of definition 1 we can now conclude that the co-dition 2^o is fulfilled.Further on,we have

$$\|x'-y\|^2 = \|x'\|^2 + \|y\|^2 \geq \|x'\|^2 \geq (1-\varphi)^2\|x_{k+1}\|^2 > \frac{\alpha^2}{4}$$

i.e. $\|x'-y\| > \frac{\alpha}{2}$ .Let now x be an arbitrary vector from $Y(y)$.According to definition 1,there exists sequence

$$x'_j = \sum_{n=k+1}^{+\infty}\varphi'_n(j)((1-\varphi)x_{k+1}+\varphi x_{n+1})(j=1,2,\dots) \text{ such that } x=\lim_{j\to+\infty} x'_j.$$

Since $\|x'_j-y\| > \frac{\alpha}{2}$ $(j=1,2,\dots)$,there exists such a natural number j_o that

$$\big|\|x'_{j_o}-y\| - \|x'_{j_o}-x\|\big| \geq \frac{\alpha}{2}$$

states,hence 3^o is satisfied and the proof is completed.

Remark 1: The last theorem in the other words means that the good approximation of an IDS by one of it's FDS is not possible in that case,in spite of the fact that such a IDS is limited set.Therefore,it makes a sence to develope the theory on such a simplex,which is done in [6] already.

The next theorem shows that somewhere on IDS the desirable approximation is possible in local view.

Theorem 4: Let $\sup\|x_n\| < +\infty$ and let $\varepsilon > 0$ be a arbitrary real number.Let,further, $y=\sum\limits_{n=o}^{+\infty}\theta_n x_n$ $(\sum\limits_{n=o}^{+\infty}\theta_n = 1 \ ; \ \theta_n \geq 0 \ (n=0,1,\dots))$

be such a vector that $\sum\limits_{n=o}^{k}\theta_n > 1 - \frac{\varepsilon}{4\sup\|x_n\|}$

states.Then for each $x \in S(x_o,x_1,\dots)\cap K(y,\frac{\varepsilon}{2})$,there exists $y' \in S(x_o,x_1,\dots,x_k)$ such that $\|x-y'\| < \varepsilon$ is valid,where

$K(y,\frac{\varepsilon}{2})$,as usual,denotes $\{x \mid \|x-y\|<\frac{\varepsilon}{2}\}$.

Proof: We shall demonstrate that $y`=(1-\sum_{n=o}^{k-1}\theta_n)x_k+\sum_{n=o}^{k-1}\theta_n x_n$ satisfies the proposition.Really,

$$\|x-y`\|\leqslant\|x-y\|+\|y-y`\|<\frac{\varepsilon}{2}+(1-\sum_{n=o}^{k}\theta_n)\cdot\|x_k\|+\sum_{n=k+1}^{+\infty}\theta_n\|x_n\|<$$

$$<\frac{\varepsilon}{2}+\frac{\varepsilon\cdot\|x_k\|}{4\sup\|x_n\|}+\sup\|x_n\|\cdot(1-\sum_{n=o}^{k}\theta_n)<\frac{\varepsilon}{2}+\frac{\varepsilon}{4}+\sup\|x_n\|\frac{\varepsilon}{4\sup\|x_n\|}=\varepsilon$$

and the proof is completed.

The sufficient conditions when the absolute error made in replacing the IDS by it`s FDS is lower then given $\varepsilon>0$, followed in the next two theorems:

Theorem 5: If $\|x_n\|<\frac{\varepsilon}{3}$ $(n>k)$,then for each $y\in S(x_o,x_1,\ldots.)$ there exists $y`\in S(x_o,x_1,\ldots.,x_k)$ such that $\|y-y`\|<\varepsilon$.

Proof: Let $y=\lim_{j\to+\infty}y_j$,where $y_j=\sum_{n=o}^{+\infty}{}^j\theta_n x_n$, $\sum_{n=o}^{+\infty}{}^j\theta_n=1$ and $\theta_n\geqslant 0$ $(n=0,1,\ldots.)$ .Let us,further,denote $y`=\sum_{n=o}^{k-1}{}^r\theta_n x_n+(1-\sum_{n=o}^{k-1}{}^r\theta_n)x_k$,where $\|y-y_r\|<\frac{\varepsilon}{3}$ .Now we have

$$\|y-y`\|\leqslant\|y-y_r\|+\|y_r-y`\|<\frac{\varepsilon}{3}+(1-\sum_{n=o}^{k}{}^r\theta_n)\|x_k\|+\sum_{n=k+1}^{+\infty}{}^r\theta_n\|x_n\|<$$

$$<\frac{\varepsilon}{3}+(1-\sum_{n=o}^{k}{}^r\theta_n)\frac{\varepsilon}{3}+(1-\sum_{n=o}^{k}{}^r\theta_n)\frac{\varepsilon}{3}<\varepsilon$$,which proves the theorem.

As a direct consequence of this theorem we obtain

Theorem 6: If $\lim_{n\to+\infty}x_n=a$ (a is vector) ,then for each $\varepsilon>0$, there exists an FDS which is ε-approximation of IDS $S(x_o-a,x_1-a,\ldots)$.

REFERENCES:

1. BASTIANI A.: Cónes convexes et pyramides convexes.Ann.Ins Fourier 9(1959),249 -292.
2. HÖLLEIN H.: Polytope in lokalkonvexex Räumen.Math.Ann. 22 (1977),65 - 85.
3. LAU K.S.: Infinite dimensional polytopes.Math.Scand. 32 (1972),193 - 213.
4. MASERICK P.H.: Convex polytopes in linear spaces.Illinois J.Math. 9(1965),623 - 635.
5. PHELPS R.R.: Infinite dimensional compact convex polytope Math.Scand. 24(1969),5 - 26.
6. LABAN M.: Neki geometrijski problemi Hilbertovih prostora Ph.D. Thesis,Beograd 1980.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

SOME SUFFICIENT CONDITIONS FOR CONVERGENCE OF AOR-METHOD

Ljiljana Cvetković, Dragoslav Herceg

ABSTRACT

'e consider AOR (Accelerated Overrelaxation method for a system of n linear equations with n unknowns Ax = b, where the matrix A has nonvanishing diagonal elements. If A is strictly diagonally dominant we improve the convergence intervals, given in |5|, for σ and ω. We also consider the convergence intervals for some matrices, which are not strictly diagonally dominant.

NEKI DOVOLJNI USLOVI ZA KONVERGENCIJU AOR-POSTUPKA. *Posmatramo AOR (Accelerated Overrelaxation) postupak za rešavanje sistema n linearnih jednačina sa n nepoznatih Ax = b, gde matrica A ima nenula dijagonalne elemente. Ako je A strogo dijagonalno dominantna, poboljšavamo intervale konvergencije, date u |5|, za σ i ω. Takodje, posmatramo intervale konvergencije za neke matrice, koje nisu strogo dijagonalno dominantne.*

1. INTRODUCTION

We consider a system of n linear equations with n unknowns, written in the matrix form

$$Ax = b,$$

where the matrix $A = |a_{ij}|$ has nonvanishing diagonal elements, and AOR (Accelerated overrelaxation) method for the numerical solution of this linear system. This iterative method was presented by Hadjidimos in |1|, 1978. By splitting A into the sum D-S-T, where $D = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ and S and T are the strictly lower and upper triangular parts of A multiplied by -1, the corresponding AOR scheme has the following form:

(1) $(E-\sigma L)x^{k+1} = ((1-\omega)E+(\omega-\sigma)L+\omega U)x^{k} + \omega c, \ k=0,1,\dots,$

where $L = D^{-1}S$, $U = D^{-1}T$, $c = D^{-1}b$, E is the unit matrix of order n, σ is the acceleration parameter, $\omega \neq 0$ is the overrelaxation parameter and $x^{0} \in C^{n}$ is arbitrary. The iterative matrix of scheme (1) is given by

$$M_{\sigma,\omega} = (E-\sigma L)^{-1}((1-\omega)E+(\omega-\sigma)L+\omega U).$$

We get bounds for the spectral radius $\rho(M_{\sigma,\omega})$ of the matrix $M_{\sigma,\omega}$ in form $\rho(M_{\sigma,\omega}) \leq G$ and then from $G<1$ we get sufficient conditions for the convergence of AOR method.

For $A = |a_{ij}| \in C^{n,n}$ (= set of complex nxn matrices) we define for $i=1,2,\dots,n$

$$P_i(A) = \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad Q_i(A) = \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|,$$

$$e_i = P_i(L), \quad f_i = P_i(U), \quad \tilde{e}_i = Q_i(L), \quad \tilde{f}_i = Q_i(U),$$

$$e_{\alpha,i} = \alpha e_i + (1-\alpha)\tilde{e}_i, \quad f_{\alpha,i} = \alpha f_i + (1-\alpha)\tilde{f}_i .$$

2. CONVERGENCE OF THE AOR METHOD

Theorem 1. *Let* $A = [a_{ij}] \in C^{n,n}$, $a_{ii} \neq 0$, $i=1,2,\dots,n$ *and* $\alpha \in [0,1]$. *Then for* $\omega,\sigma \in \mathbb{R}$, $\omega\neq 0$, $|\sigma| e_{\alpha,i} < 1$, $i=1,2,\dots$ $\dots,n$, $\rho(M_{\sigma,\omega})$ *satisfies the following:*

$$\min_{1\leq i\leq n} \frac{|1-\omega| - |\omega-\sigma| e_{\alpha,i} - |\omega| f_{\alpha,i}}{1+|\sigma| e_{\alpha,i}} \leq \rho(M_{\sigma,\omega}) \leq$$

$$\leq \max_{1\leq i\leq n} \frac{|1-\omega| + |\omega-\sigma| e_{\alpha,i} + |\omega| f_{\alpha,i}}{1-|\sigma| e_{\alpha,i}} .$$

Proof. We prove the upper bound for $\rho(M_{\sigma,\omega})$. Let λ be any eigenvalue of $M_{\sigma,\omega}$ and suppose that

$$|\lambda| > \frac{|1-\omega|+|\omega-\sigma|e_{\alpha,i}+|\omega|f_{\alpha,i}}{1-|\sigma|e_{\alpha,i}}, \quad i=1,2,\dots,n .$$

After some manipulations we have

$$|\lambda+\omega-1| > |\omega+\sigma(\lambda-1)|e_{\alpha,i}+|\omega|f_{\alpha,i}, \quad i=1,2,\dots,n ,$$

$$|b_{ii}| > \alpha P_i(B)+(1-\alpha)Q_i(B), \quad i=1,2,\dots,n,$$

where $B=[b_{ij}]\in C^{n,n}$, $B=(\lambda+\omega-1)E-(\omega+\sigma(\lambda-1))L-\omega U$. Then theorem 2.5.2 from [2] shows that $\det B\neq 0$. Since $(E-\sigma L)\cdot$ $\cdot(\lambda E-M_{\sigma,\omega})=B$ and $\det(E-\sigma L)=1$ it follows $\det(\lambda E-M_{\sigma,\omega})\neq 0$. This contradicts the singularity of $\lambda E-M_{\sigma,\omega}$.

The lower bound for $\rho(M_{\sigma,\omega})$ one proves similarly.

Theorem 2. Let $A=[a_{ij}]\in C^{n,n}$, $a_{ii}\neq 0$, $i=1,2,\dots$ $\dots,n$. *Then for* $\omega,\sigma\in\mathbb{R}$, $\omega\neq 0$, $|\sigma|(e_i+e_j)<2$, $i\neq j$, $i,j=$ $=1,2,\dots,n$, $\rho(M_{\sigma,\omega})$ *satisfies the following:*

$$\min_{i\neq j}\frac{2|1-\omega|-|\omega-\sigma|(e_i+e_j)-|\omega|(f_i+f_j)}{2+|\sigma|(e_i+e_j)} \le \rho(M_{\sigma,\omega}) \le$$

$$\le \max_{i\neq j}\frac{2|1-\omega|+|\omega-\sigma|(e_i+e_j)+|\omega|(f_i+f_j)}{2-|\sigma|(e_i+e_j)}$$

Proof. We prove only upper bound for $\rho(M_{\sigma,\omega})$. The lower bound we obtain similarly. Suppose that $M_{\sigma,\omega}$ has an eigenvalue λ with

$$|\lambda| > \frac{2|1-\omega|+|\omega-\sigma|(e_i+e_j)+|\omega|(f_i+f_j)}{2-|\sigma|(e_i+e_j)}, \quad i\neq j ,$$

$$i,j=1,2,\dots,n .$$

From this inequality follows that

$$|\lambda+\omega-1| > |\omega+\sigma(\lambda-1)| \frac{e_i+e_j}{2} + |\omega| \frac{f_i+f_j}{2} , \quad i\neq j, i,j=1,2,\dots,n,$$

$$|\lambda+\omega-1| > \frac{1}{2} (P_i(B)+P_j(B)), \quad i\neq j, \quad i,j=1,2,\dots,n,$$

where B is defined as in the proof of theorem 1. Since $b_{ii} = =\lambda+\omega-1$, $i=1,2,\dots,n$ and

$$\frac{1}{2} (P_i(B) + P_j(B)) \geq \sqrt{P_i(B)P_j(E)} \quad ,$$

we have now

$$|b_{ii}|\,|b_{jj}| > P_i(B)P_j(B), \quad i\neq j, \quad i,j=1,2,\dots,n.$$

But then, theorem 2.4.1 from [2] shows that $\det B \neq 0$. This contradicts the singularity of $\lambda E - M_{\sigma,\omega}$.

Theorem 1 contains as a special case ($\alpha=1$) theorem 1 of [3], where the matrix A must be strictly diagonally dominant. In our case it is sufficient that A has nonvanishing diagonal elements.

Under assumptions of theorem 1 of [3] our theorem 2 holds, but the converse is not true.

Theorem 3. Let $A = [a_{ij}] \in C^{n,n}$, $a_{ii} \neq 0$, $i=1,2,\dots$ $\dots,n$ *and* $\alpha \in [0,1]$.

Then the AOR method converges for

(a) $$\max_i (e_{\alpha,i}+f_{\alpha,i}) < 1, \quad 0 < \omega < \min_i \frac{2}{1+e_{\alpha,i}+f_{\alpha,i}} ,$$

$$\max_i \frac{-\omega(1-e_{\alpha,i}-f_{\alpha,i})+2\max(0,\omega-1)}{2\, e_{\alpha,i}} < \sigma < \min_i \frac{\omega(1+e_{\alpha,i}-f_{\alpha,i})+2\min(0,1-\omega)}{2\, e_{\alpha,i}}$$

or

(b) $$\max_{i\neq j} (e_i+e_j+f_i+f_j) < 2, \quad 0 < \omega < \frac{4}{2+e_i+e_j+f_i+f_j} ,$$

$$\max_{i\neq j} \frac{-\omega(2-e_i-e_j-f_i-f_j)+4\max(0,\omega-1)}{2(e_i+e_j)} < \sigma < \min_{i\neq j} \frac{\omega(2+e_i+e_j-f_i-f_j)+4\min(0,1-\omega}{2(e_i+e_j)}$$

Proof. We consider (a) and theorem 1. Similarly one can show the convergence of AOR method in case (b) using theorem 2.

shall prove that for all $i=1,2,\dots,n$ holds

2)
$$\left.\begin{array}{c} e_{\alpha,i}+f_{\alpha,i}<1,\quad 0<\omega<\dfrac{2}{1+e_{\alpha,i}+f_{\alpha,i}} \\[2ex] \dfrac{-\omega(1-e_{\alpha,i}-f_{\alpha,i})+2\max(0,\omega-1)}{2e_{\alpha,i}}<\sigma<\dfrac{\omega(1+e_{\alpha,i}-f_{\alpha,i})+2\min(0,1-\omega)}{2e_{\alpha,i}} \end{array}\right\}\Longrightarrow$$

3)
$$\frac{|1-\omega|+|\omega-\sigma|e_{\alpha,i}+|\omega|f_{\alpha,i}}{1-|\sigma|e_{\alpha,i}}<1.$$

ince for σ and ω from (a) we have $|\sigma|e_{\alpha,i}<1$, theorem 1 nd (3) show that $\rho(M_{\sigma,\omega})<1$.

o prove implication (2) $\Longrightarrow$ (3) we consider the next cases.

ase I: $0<\omega\leq 1$, $\dfrac{-\omega(1-e_{\alpha,i}-f_{\alpha,i})}{2e_{\alpha,i}}<\sigma\leq 0$.

hen $1-\omega+\omega e_{\alpha,i}-\sigma e_{\alpha,i}+\omega f_{\alpha,i}<1+\sigma e_{\alpha,i}$, which is equivalent to (3).

Case II: $0<\omega\leq 1$, $0<\sigma\leq\omega$.

Then $1-\omega+\omega e_{\alpha,i}-\sigma e_{\alpha,i}+\omega f_{\alpha,i}<1-\sigma e_{\alpha,i}$, since $e_{\alpha,i}+f_{\alpha,i}<1$.

Case III: $0<\omega\leq 1$, $\omega<\sigma<\dfrac{\omega(1+e_{\alpha,i}-f_{\alpha,i})}{2e_{\alpha,i}}$.

Then $1-\omega+\sigma e_{\alpha,i}-\omega e_{\alpha,i}+\omega f_{\alpha,i}<1-\sigma e_{\alpha,i}$.

Case IV: $1<\omega<\dfrac{2}{1+e_{\alpha,i}+f_{\alpha,i}}$, $\dfrac{\omega+\omega e_{\alpha,i}+\omega f_{\alpha,i}-2}{2e_{\alpha,i}}<\sigma\leq 0$.

Then $\omega-1+\omega e_{\alpha,i}-\sigma e_{\alpha,i}+\omega f_{\alpha,i}<1+\sigma e_{\alpha,i}$.

Case V: $1<\omega<\dfrac{2}{1+e_{\alpha,i}+f_{\alpha,i}}$, $0<\sigma\leq\omega$.

Then $\omega-1+\omega e_{\alpha,i}-\sigma e_{\alpha,i}+\omega f_{\alpha,i}<1-\sigma e_{\alpha,i}$.

Case VI: $1<\omega<\dfrac{2}{1+e_{\alpha,i}+f_{\alpha,i}}$, $\omega<\sigma<\dfrac{-\omega+\omega e_{\alpha,i}-\omega f_{\alpha,i}+2}{2e_{\alpha,i}}$

Then $\omega-1+\sigma e_{\alpha,i}-\omega e_{\alpha,i}+\omega f_{\alpha,i}<1-\sigma e_{\alpha,i}$.

Remark. If in case (a) of theorem 3 we assume $\alpha = 1$, then for strictly diagonally dominant matrices AOR method converges if

$$0 < \omega < \min_i \frac{2}{1+e_i+f_i},$$

$$\max_i \frac{-\omega(1-e_i-f_i)+2\max(0,\omega-1)}{2e_i} < \sigma < \min_i \frac{\omega(1+e_i-f_i)+2\min(0,1-\omega)}{2e_i}$$

This convergence intervals for ω and σ are larger than the corresponding intervals from theorem 3 of [5].

REFERENCES

1. HADJIDIMOS A.: *Accelerated Overrelaxation Method.* Math. Comp., v. 32 (1978), 149-157.
2. MARKUS M.,MINK H.: Obzor po teorii matric i matričnyh neravenstv, Moskva 1972.
3. MARTINS M.: *On an Accelerated Overrelaxation Iterative Method for Linear Systems With Strictly Diagonally Dominant Matrix.* Math. Comp., v. 35 (1980), 1269-1273.
4. MARTINS M.: *Note on Ireducible Diagonally Dominant Matrices and the Convergence of the AOR Iterative Method.* Math. Comp., v. 37 (1981), 101-103.
5. MATRINS M.: *An improvent for the Area of Convergence of the Accelerated Overrelaxation Iterative Method.* Anal. Numér. Théor.Approx., T 12, No 1 (1983), 65-76.
6. VARGA R.S.: Matrix Iterative Analysis. Englewood Cliffs, New York 1962.

Numerical Methods and
Approximation Theory
Niš, September 26 - 28, 1984

SOME MODIFIED SQUARE ROOT ITERATIONS FOR THE SIMULTANEOUS DETERMINATION OF MULTIPLE COMPLEX ZEROS OF A POLYNOMIAL

Miodrag S. Petković, Lidija V. Stefanović

ABSTRACT:

Applying Newton's and Halley's correction, some modifications of square root method, suitable for simultaneous finding multiple complex zeros of a polynomial with the known order of multiplicity, are obtained in the paper. The convergence order of the proposed (total-step) methods is five and six respectively. Further improvements of these methods are performed by approximating to all zeros in a serial fashion using new approximations immediately they become available (the so-called Gauss-Seidel approach). Faster convergence is attained without additional calculations. The lower bounds of the R-order of convergence for the serial (single-step) methods are given. The considered iterative processes are illustrated numericaly in the example of an algebraic equation.

NEKE MODIFIKOVANE KVADRATNO KORENSKE ITERACIJE ZA SIMULTANO ODREDJIVANJE VIŠESTRUKIH KOMPLEKSNIH NULA POLINOMA. *Primenjujući Newtonovu i Halleyevu korekciju u radu su dobijene neke modifikacije metoda kvadratnog korena, pogodne za simultano nalaženje višestrukih kompleksnih nula polinoma poznatog reda višestrukosti. Red konvergencije predloženih (total-step) metoda je pet i šest respektivno. Dalja poboljšanja ovih metoda su postignuta aproksimirajući sve nule u serijskom postupku korišćenjem novih aproksimacija odmah kada postanu dostupne (tzv. Gauss-Seidelov pristup). Brža konvergencija je dobijena bez dodatnih izračunavanja. Za serijske (single-step) metode date su donje granice R-reda konvergencije. Razmatrani iterativni procesi ilustrovani su numerički na primeru algebarske jednačine.*

1. INTRODUCTION

The iterative methods for the simultaneous determination of multiple zeros of a polynomial have been developed during the last decade as extensions of the known methods for simple zeros. M.R. Farmer and G. Loizou [4] have derived a class of iterative methods with arbitrary order of convergence. The basic imperfection of methods from this class with high convergence order (greater than three) is a demand for great number of numerical operations, which decrease their effectiveness. Several modifications of the basic Maehly's method [10], which enable very fast convergence by reasonably small numerical operations, have been proposed in [12]. In recent years a lot of attention has been given to the study of this topics in interval arithmetics (see [6], [7], [15], [16]).

In this paper we give some modifications of square root method (also known as Ostrowski's method [14]) which provide: (i) simultaneous determination of multiple polynomial zeros whose the multiplicities are known; (ii) acceleration of convergence with small number of additional calculations in relation to the basic method.

2. SOME MODIFICATIONS OF SQUARE ROOT ITERATIONS

Consider a monic polynomial P of degree $n \geqq 3$

$$P(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0 = \prod_{j=1}^{k} (z - r_j)^{m_j} \quad (a_i \in C)$$

with real or complex zeros $r_1, \ldots, r_k$ having the order of multiplicity $m_1, \ldots, m_k$ recpectively, where $m_1 + \cdots + m_k = n$. Let $z_1, \ldots, z_k$ be distinct reasonably good approximations to these zeros and $\hat{z}_i$ be the next approximation to r_i using some iterative sheme.

Let m be the multiplicity of the zero r of P. By the functions

$$u(z) = \frac{P'(z)}{P(z)}, \qquad v(z) = \frac{P''(z)}{P'(z)}$$

we define

(1) $$G(z) = u(z)\left[u(z) - v(z)\right] \quad \text{(Ostrowski's function)},$$

(2) $$N(z) = -\frac{m}{u(z)} \quad \text{(Newton's correction)},$$

(3) $$H(z) = 2\left[v(z) - (1 + \frac{1}{m})\, u(z)\right]^{-1} \quad \text{(Halley's correction)}.$$

We recall that the correction terms (2) and (3) appear in the iterative formulas

(4) $\hat{z}_i = z_i + N(z_i)$ (Shröder's modification of Newton's method for multiple zeros, see [17]),

(5) $\hat{z}_i = z_i + H(z_i)$ (modification of Halley's method, introduced by Hansen and Patrick [8] for multiple zeros),

the convergence order two and three recpectively. We note that the order of multiplicity in the iterative formulas (4) and (5) take the values $m = m_i$ $(i = 1, \dots, k)$.

Using the logarithmic derivative of P we obtain

$$-\frac{d^2}{dz^2}\ln P(z) = \frac{P'(z)^2 - P(z)P''(z)}{P(z)^2} = G(z) = \sum_{j=1}^{k} m_j(z-r_j)^{-2}.$$

The value of Ostrowski's function at the point $z = z_i$ is

$$G(z_i) = \sum_{j=1}^{k} m_j(z_i - r_j)^{-2},$$

wherefrom

$$(6) \qquad r_i = z_i - \sqrt{m_i}\left[G(z_i) - \sum_{j\neq i} m_j(z_i - r_j)^{-2}\right]_*^{-1/2} \qquad (i=1,\dots,k).$$

The symbol * denotes that one of two values of square root is chosen. One criterion for the choice of the appropriate value of square root has been established by Gargantini [7]. If all zeros of P are real, then this criterion reduces to the choice of sign which coincides to the sign of (real value) PP'.

Setting $r_i \cong z_i$ in (6) and taking some approximations of r_j on the right-hand side of the identity (6), some modified iterative processes of square root type for simultaneous finding of multiple complex zeros of a polynomial can be obtained from (6). The convergence analyses of these methods is essentially the same to that of the iterative methods considered in [1], [2, Ch. 8], [11], and so, it will be omitted. For the serial (single-step) methods, where new approximations are used in the same iteration, we shall use the concept of the R-order of convergence (see [13]). The R-order of convergence of an iterative process IP with the limit point given by the vector $\mathbf{r} = [r_1 \cdots r_k]^T$ (where $r_1, \dots, r_k$ are polynomial zeros) will be denoted by $O_R((IP), \mathbf{r})$.

1° For $r_j := z_j$ $(j \neq i)$ we get from (6) the parallel (total-step) square root iteration (shortly TS):

$$(7) \qquad \hat{z}_i = z_i - \sqrt{m_i}\left[G(z_i) - \sum_{j\neq i} m_j(z_i - z_j)^{-2}\right]_*^{-1/2} \qquad (i = 1,\dots,k).$$

This method has been considered in [15] as a special case of the generalised root iteration. It has been proved that the convergence order of TS-method (7) is **four**. Note that the iterative method of the form (7) in terms of circular regions has been analysed by Gargantini [7].

2° Taking $r_j := \hat{z}_j$ $(j < i)$ and $r_j := z_j$ $(j > i)$ in (6), we obtain the serial (single-step) square root iteration (SS):

$$(8)\quad \hat{z}_i = z_i - \sqrt{m_i}\left[G(z_i) - \sum_{j<i} m_j(z_i-\hat{z}_j)^{-2} - \sum_{j>i} m_j(z_i-z_j)^{-2}\right]_*^{-1/2}$$

$$(i = 1,\ldots,k).$$

It has been proved in [16] that the R-order of convergence of SS-method is at least $3 + \mu_k \in (4, \frac{27}{5})$, where $\mu_k \in (1, \frac{12}{5})$ is the unique positive zero of the equation $\mu^k - \mu - 3 = 0$ $(k \geq 2)$.

3° Putting $r_j := z_j + N(z_j)$ $(i \neq j)$ in (6), where $N(z_j)$ is Newton's correction given by (2), we obtain the parallel (total-step) square root method with Newton's correction (TSN):

$$(9)\quad \hat{z}_i = z_i - \sqrt{m_i}\left[G(z_i) - \sum_{j\neq i} m_j(z_i - z_j - N(z_j))^{-2}\right]_*^{-1/2}$$

$$(i = 1,\ldots,k).$$

Using similar procedure as in [11], it can be proved that the convergence order of the modified method (9) is **five**.

4° The iterative process (9) can be accelerated by approximating all zeros in a serial fashion, i.e. using new approximations immediately they become available (the so-called Gauss-Seidel approach). In this way, substituting $r_j := \hat{z}_j$ $(j < i)$, $r_j := z_j + N(z_j)$ $(j > i)$ in (6), we derive single-step method with Newton's correction (SSN):

$$(10)\quad \hat{z}_i = z_i - \sqrt{m_i}\left[G(z_i) - \sum_{j<i} m_j(z_i-\hat{z}_j)^{-2} - \sum_{j>i} m_j(z_i-z_j-N(z_j))^{-2}\right]_*^{-1/2}$$

$$(i = 1,\ldots,k).$$

For the iterative process (10) we can prove the following statement concerning the convergence order:

THEOREM 1: *The lower bound of the R-order of convergence of the iterative method (10) is given by*

$$O_R((10),\mathbf{r}) \geq 3 + \tau_k \in (5,7)$$

where $\tau_k \in (2,4)$ is the unique positive root of the equation

$$\tau^k - 2^{k-1}(\tau + 3) = 0 \quad (k \geq 2).$$

Similar as for TSN-method, we can apply Halley's corrction (3) for multiple zeros. Taking $r_j := z_j + H(z_j)$ $(j \neq i)$ in 6), we obtain total-step method with Halley's correction (TSH):

$$(11)\qquad \hat{z}_i = z_i - \sqrt{m_i}\left[G(z_i) - \sum_{j \neq i} m_j (z_i - z_j - H(z_j))^{-2}\right]_*^{-1/2} \qquad (i = 1,\dots,k).$$

The iterative method constructed on the basis of formula (11) has the convergence order equal to **six**.

6^o Finally, setting $r_j := \hat{z}_j$ $(j < i)$, $r_j := z_j + H(z_j)$ $(j > i)$ in (6), we obtain single-step method with Halley's correction (SSH):

$$(12)\qquad \hat{z}_i = z_i - \sqrt{m_i}\left[G(z_i) - \sum_{j < i} m_j (z_i - \hat{z}_j)^{-2} - \sum_{j > i} m_j (z_i - z_j - H(z_j))^{-2}\right]_*^{-1/2} \qquad (i = 1,\dots,k).$$

The following assertion for the method (12) is valid:

THEOREM 2: *The lower bound of the R-order of convergence of the iterative method (12) is given by*

$$O_R((12), r) \geqq 3(1 + \sigma_k) \in (6, 8),$$

where $\sigma_k \in (1, \frac{5}{3})$ *is the unique positive root of equation* $\sigma^k - \sigma - 1 = 0$ $(k \geqq 2)$.

The increase of convergence of single-step methods (8), (10) and (12) (in a serial fashion), compared to the corresponding total-step methods (7), (9) and (11) (in a parallel fashion), is larger if the number of different zeros is smaller. The acceleration of convergence is attained without additional calculations; moreover, single-step methods occupy less storage space in digital computer (because the calculated approximations immediately take positions of the former ones).

In practical realization of the iterative methods (9)-(12) with Newton's and Halley's corrections, before determination of new approximations it is desirable to evaluate u(z) and v(z) and then, by (1), (2) and (3) calculate G(z) and the wanting corrections N(z) or H(z). In such a way, the methods with correction terms claim slightly more of numerical operations compared to the basic fourth order method (7). This point at the effectiveness of the proposed modifications of square root methods.

3. NUMERICAL RESULTS

In practice, it is convenient to apply a three-stage globally convergent composite algorithm (see [4]):

(a) Find an inclusion region of the complex plane containing all the zeros of a polynomial.

(b) Apply a slowly convergent search algorithm to obtain initial approximations to the zeros and calculate their respective multiplicities. The multiplicities of these approximations can be estimated, for example, using ([9])

$$m_i = \lim_{z_i \to r_i} u'(z).$$

Other limiting formulas are described in [3], [18] etc.

(c) Improve starting approximations with a rapidly convergent iterative processes (for example, applying any of the algorithms (7)-(12)) to any required accuracy.

In this section we shall apply the considered iterative methods (7) - (12) of square root type for the stage (c). In order to test these methods the routine on FORTRAN was realised on HONEYWELL 66 system in double precision arithmetic (about 18 significant digits). Before calculating new approximations the values $u(z^{(\lambda)})$ and $v(z^{(\lambda)})$ ($\lambda = 1,2,\ldots$ is the iteration index, $i = 1,\ldots,k$), necessary for evaluation of Ostrowski's function (1), where calculated. The same values were used for calculation of Newton's and Halley's corrections in the formulas (9) - (12).

The proposed modifications were illustrated numerically in the example of the polynomial

$$P(z) = z^9 - 7z^8 + 20z^7 - 28z^6 - 18z^5 + 110z^4 - 92z^3 + 44z^2 + 345z+225$$

whose zeros are $r_1 = 1 + 2i$, $r_2 = 1 - 2i$, $r_3 = -1$, $r_4 = 3$ with the multiplicities $m_1 = 2$, $m_2 = 2$, $m_3 = 3$, $m_4 = 2$. As the initial approximations to these zeros the following complex numbers were taken:

$z_1^{(0)} = 1.8 + 2.7i$, $z_2^{(0)} = 1.8 - 2,7i$, $z_3^{(0)} = -0.3 - 0.8i$,
$z_4^{(0)} = 2.3 - 0.7i$.

spite of crude initial approximations, the presented iterative methods demonstrate very fast convergence. Numerical results, obtained in the second iteration, are displayed in Table 1.

	i	Re $\{z_i^{(2)}\}$	Im $\{z_i^{(2)}\}$
TS (7)	1	0.999999853800923892	2.000000112716998844
	2	0.999999826741999847	-2.000000351383949125
	3	-0.999999859207295616	-8.18×10^{-7}
	4	3.000000527270300803	-3.48×10^{-8}
SS (8)	1	0.999999939617346251	1.999999964305993363
	2	1.000000861310650873	-2.000000509862992614
	3	0.999999999709498985	1.35×10^{-9}
	4	3.000000000000030662	7.16×10^{-14}
TSN (9)	1	0.999999455077856744	2.000000212961094747
	2	1.000000018147137107	-2.000000068835695135
	3	0.999999974528732211	3.43×10^{-8}
	4	3.000000722708680682	-9.58×10^{-8}
SSN (10)	1	0.999999894885117145	2,000000042747320793
	2	0.999999994177457521	-2.000000000709903145
	3	-1.000000000007845003	3.82×10^{-11}
	4	2.999999999999997525	-6.58×10^{-15}
TSH (11)	1	1.000000000098386276	1.999999999890580897
	2	1.000000000450329186	-2.000000000521585396
	3	-0.999999999986166747	-2.93×10^{-12}
	4	3.000000000368704406	-6.92×10^{-10}
SSH (12)	1	1.000000000032764666	2.000000000002146278
	2	1.000000000000674921	-1.999999999997025086
	3	-1.000000000000001322	2.74×10^{-15}
	4	3.000000000000000383	$-2,01 \times 10^{-16}$

Table 1

We note that the lower bounds of the suggested serial methods in our example $(k=4)$ are $O_R((8),r) \geq 4.453$, $O_R((10),r) \geq 5.586$ and $O_R((12),r) \geq 6.662$.

REFERENCES

1. ALEFELD G., HERZBERGER J.: On the convergence speed of some algorithms for the simultaneous approximation of polynomial roots. SIAM J. Numer. Anal. 2(1974),237-243.
2. ALEFELD G., HERZBERGER J.: Einführung in die Intervallrechnung. B.I. Wissenschaftsverlag, Zürich 1974.
3. DEKKER T. J.: Newton-Laguerre iteration. Colloq. Internat. CNRS no. 165. Programmation en mathématiques numériques 1968, 189-200.
4. FARMER M.R., LOIZOU G.: An algorithm for the total or partial factorization of a polynomial. Math. Proc. Camb. Phil. Soc. 82 (1977), 427-437.
5. GARGANTINI I.: Parallel Laguerre iterations: Complex case. Numer. Math. 26 (1976), 317-323.
6. GARGANTINI I.: Further applications of circular arithmetic: Schroeder-like algorithms with error bounds for finding zeros of polynomials. SIAM J. Numer. Anal. 3 (1978), 497-510.
7. GARGANTINI I.: Parallel square-root iterations for multiple roots. Comput. Math. with Appl. 6 (1980), 279-288.
8. HANSEN E., PATRICK M.: A family of root finding methods. Numer. Math. 27 (1977), 257-269.
9. LAGOUANELLE J. L.: Sur une méthode de calcul de l'ordre de multiplicité des zéros d'un polynome. C.R. Acad. Sci. Paris A 262 (1966), 626-627.
10. MAEHLY H. J.: Zur iterativen Auflösung algebraischer Gleichungen. Z. Angew. Math. Phys. 5 (1954), 260-263.
11. MILOVANOVIĆ G. V., PETKOVIĆ M. S.: On the convergence order of a modified method for simultaneous finding polynomial zeros. Computing 30 (1983), 171-178.
12. MILOVANOVIĆ G.V., PETKOVIĆ M.S.: Metodi visokog reda za simultano odredjivanje višestrukih nula polinoma. Zbornik V znanstvenog skupa PPPR, Stubičke Toplice 1983, 95-99.
13. ORTEGA J.M., RHEINBOLDT W.C.: Iterative solution of nonlinear equations in several variables. Academic Press, New York 1970.
14. OSTROWSKI A.M.: Solution of equations and systems of equations. Academic Press, New York 1966.
15. PETKOVIĆ M.S.: Generalised root iterations for the simultaneous determination of multiple complex zeros. ZAMM 62 (1982), 627-630.
16. PETKOVIĆ M.S., STEFANOVIĆ L.V.: On the convergence order of accelerated root iterations. Numer. Math. (to appear).
17. SCHRÖDER E.: Über unendlich viele Algorithmen zur Auflösung der Gleichungen. Math. Ann. 2 (1870), 317-365.
18. TRAUB J.F.: Iterative methods for the soilution of equations. Englewood Cliffs, New Jersey, Prentice Hall 1964.

Numerical Methods and
Approximation Theory
Niš, September 26 - 28, 1984

THE GENERALIZATION OF TEN RATIONAL APPROXIMATIONS OF ITERATION FUNCTIONS

Dušan V. Slavić

ABSTRACT:

I.Newton (1676), E. Halley (1694), P. L. Čebyšev (1838), E. T. Whittaker (1918), E.Durand (1960) and J.F.Traub (1961) gave the one-point iteration functions for solving the equation $f(x) = 0$. The general result is given here which contains the mentioned functions as particular cases or gives corrections of some coefficients in these functions in order to increase the convergency order of the methods. In addition, the questions of autorship priorities are considered.

GENERALIZACIJA DESET RACIONALNIH APROKSIMACIJA ITERACIONIH FUNKCIJA. I. Newton (1676), E. Halley (1694), P. L. Čebyšev (1838), E. T. Whittaker (1918), E. Durand (1960) i J. F. Traub (1961) dali su jednotačkaste iteracione funkcije za rešavanje jednačine $f(x) = 0$. Ovde je dat opšti rezultat koji sadrži pomenute iteracione funkcije kao posebne slučajeve ili daje korekcije nekih koeficijenata u tim funkcijama u cilju povećanja reda konvergencije metoda. Pored toga, razmatrana su i pitanja autorskog prioriteta.

Let u, A, B, C be defined by

$$u = f/f', \quad A = f''/(2f'), \quad B = f'''/(6f'), \quad C = f^{IV}/(24f'),$$

let r be the order of convergency of the method and let $x_{n+1} = y_r(x_n)$. The classical results then become:

(1) $y_2 = x - u$ Newton

(2) $y_3 = x - u/(1 - Au)$ Halley

(3) $y_3 = x - u - Au^2$ Čebyšev

(4) $y_4 = x - u - Au^2 - (2A^2 - B)u^3$ Čebyšev

(5) $y_4 = x - u(1 - Au)/(1 - 2Au + Bu^2)$ Whittaker

(6) $y_2 = x - u/(1 - 2Au)$ Durand

(7) $y_2 = x - u(1 - 2Au)/(1 - 3Au + 3Bu^2)$ Durand

(8) $y_4 = x - u(1-3Au+3Bu^2)/(1-4Au+(2A^2+4B)u^2-4Cu^3)$ Durand

(9) $y_4 = x - u\,(A-(A^2-B)\,u)\,/\,(A-(2\,A^2-B)\,u)$ Durand

(10) $y_4 = x - u\,/(1 - A\,u - (A^2-B)\,u^2)$ Traub

The literature is full of disagreements conserning the authors of these formulas. It is claimed that already Heron (two millenia ago) had known the iteration procedure $x_1 > 0$, $x_{n+1} = (x_n + z/x_n)/2$ tending to $z^{1/2}$, which is a particular case of formula (1) for $f = x^2 - z$ $(z > 0)$.

The method of tangents (1) is related to the names: Ch'in Chiushao (1247), F.Viète (1600), T.Harriot (1611), A. Girard (1629), W.Oughtred (1647), I.Newton (1664, 1666, 1669, 1674, 1676, ...), J.Wallis (1685), J.Raphson (1690), ...

The method of tangent hyperbolas (2) is related to the names: E.Halley (1694), J.H.Lambert (1770), P.Barlow (1814), Hutton, E.Kobald (1891), E.T.Whittaker (1918), J.V.Uspensky (1927), V.A.Bailey (1941), J.S.Frame (1944), H.S.Wall (1948), H.J.Hamilton (1950), G.S.Salehov (1951), ...

The method of osculatory inverse polynomials (3) and (4) is related to the names: L.Euler (1748), H.Bürmann (1799), P.S.Čebyšev (1838), E.Schröder (1870), E.Bodewig (1935), ...

The method (5) is related to the names: H.Wronski(1811), A.de Morgan (1868), E.T.Whittaker (1918), H.J.Hamilton (1946), I.Kiss (1954), R.W.Snyder (1955), E.Durand (1960), V.L.Zaguskin (1960), A.P.Domorjad - D.K.Lika (1965), ...

The uniform and simple manner of writing the iteration functions enables one to see more easily the iterations betveen then. Each formula from (2) to (10), neglecting the higher degrees of u, becomes formula (1). Neglecting the term with u^2 formula (10) becomes (2). Neglecting the term with u^3 formula (4) becomes (3).

Let a, b, c be arbitrary parameters. Formula (10) is equivalent to

$$y_4 = x - u\,\frac{(1 + aA\,u + (bA^2 + cB)\,u^2)}{(1 - Au - (A^2-B)\,u^2)\,(1 + aA\,u + (bA^2 + cB)\,u^2)}\,,$$

wherefrom, upon neglecting the terms with u^3 and u^4, we get

$$11)\quad y_4 = x - u\,\frac{1 + aA\,u + (bA^2 + cB)\,u^2}{1 + (a-1)A\,u + ((b-a-1)A^2 + (c+1)B)\,u^2}\ .$$

Neglecting terms with u^2 from (11) it stems:

$$(12)\quad y_3 = x - u\,(1 + aA\,u)\,/\,(1 + (a-1)A\,u)\ .$$

Neglecting terms containing u, from formula (12) it stems (1).

From (12) for $a = 0$ it follows (2), while for $a = 1$ it follows (3). For $a = -1$ we get the correction of formula (6)

$$(13)\quad y_3 = x - u\,(1 - A\,u)\,/\,(1 - 2A\,u)\ .$$

Formula (13) stems also from (5) by neglecting the terms with u^2.

From (11) for $a = 1 \wedge b = 2 \wedge c = 1$ it stems (4), for $a = -1 \wedge b = c = 0$ it stems (5), for $a = -b = c \rightarrow +\infty$ it stems (9), for $a = b = c = 0$ it stems (10).

For $a = -2 \wedge b = c = 0$ or $a = -2 \wedge b = -1 \wedge c = 0$ from formula (11) it stems the correction of formula (7)

$$(14)\quad y_4 = x - u\,(1 - 2A\,u)\,/\,(1 - 3A\,u + (A^2 + 3B)\,u^2)\ ,$$

$$(15)\quad y_4 = x - u\,(1 - 2A\,u - (A^2 - 2B)\,u^2)\,/\,(1 - 3A\,u + 3B\,u^2)\ .$$

For $a = -3 \wedge b = 0 \wedge c = 3$ from formula (11) it stems the simplified formula (8)

$$(16)\quad y_4 = x - u\,(1 - 3A\,u + 3B\,u^2)\,/\,(1 - 4A\,u + (2A^2 + 4B)\,u^2)\ .$$

Formulas (1), (12), (11) are general rational approximations of one-point iteration functions for solving the equations in a sufficiently close neighborhood of the equation root. About the stages of solving the equation, see Slavić (1982).

*

A.Dordević, N.Z.Klem, G.V.Milovanović, D.S.Mitrinović, N.Obradović, D.B.Popović, D.D.Tošić, P.M.Vasić have read this paper in manuscript and have made some valuable remarks and suggestions.

REFERENCES

1. ČEBYŠEV P.L.: Vyčislenie korney uravnenija. Moskva 1848.

2. DURAND E.: Solutions numériques des équations algébriques, I. Paris 1960.

3. HALLEY E.: A new, exact and easy method of finding the roots of equations generally, and that without any previous reduction. Phil.Trans.Roy.Soc.London 18(1694)136-145.

4. NEWTON I.: Letter to G.W.Leibniz, 13.6.1676.

5. SLAVIĆ D.V.: Solution of equations by function modification. Univ.Beograd. Publ.Elektrotehn.Fak. Ser.Mat.Fiz. N^{o}735-762 (1982) 127-129.

6. TRAUB J.F.: On a class of iteration formulas and some historical notes. Comm.ACM 4(6) (1961) 276-278.

7. WHITTAKER E.T.: A formula for the solution of algebraic or transcendental equations. Proc.Edinburgh Math.Soc. 36(1918) 103-106.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

ONE-POINT ITERATION FUNCTIONS OF ARBITRARY CONVERGENCE ORDER

Dušan V. Slavić

ABSTRACT:

The approximation of iteration functions for solving the equation $f(x) = 0$ of an arbitrary convergence order, containing the values of the function f and its derivatives only at one point, are dealt with in the present paper. Though the methods are with the arbitrary convergence order r, the coefficients of methods up to $r = 5$ were calculated effectively here. All the methods dealt with here contain the Newton tangent method as their basic approximation for $r = 2$. Two general one-point iteration functions are introduced.

JEDNOTAČKASTE ITERACIONE FUNKCIJE PROIZVOLJNOG REDA KONVERGENCIJE. Ovde su razmatrane aproksimacije iteracionih funkcija za rešavanje jednačina $f(x) = 0$ proizvoljnog reda konvergencije koje sadrže vrednosti funkcije f i njenih izvoda samo u jednoj tački. Iako su metodi sa proizvoljnim redom konvergencije r, ovde su koeficijenti metoda efektivno izračunati do $r = 5$. Sve metode ovde razmatrane sadrže Newtonov metod tangenata kao svoju osnovnu aproksimaciju za $r = 2$. Uvedene su dve opšte jednotačkaste iteracione funkcije.

Let r be the convergence order of the method, y_r the iteration function $x_{n+1} = y_r(x_n)$ and let

$$u = \frac{f}{f'}, \quad A = \frac{f''}{2f'}, \quad B = \frac{f'''}{6f'}, \quad C = \frac{f^{IV}}{24f'}, \quad D = \frac{f^{V}}{120f'}.$$

P.L.Čebyšev and others (see [11]) gave the results which, in the notations given here, can be presented as

$$(1) \qquad y_r = x - u \sum_{k=0}^{r-2} p_k u^k ,$$

where

$$p_0 = 1, \quad p_1 = A, \quad p_2 = 2A^2 - B, \quad p_3 = 5A^3 - 5AB + C, \quad \dots$$

The expansion (1) is equivalent to the power series of the inverse function.

E.T.Whittaker gave the formula

$$(2)\qquad y_r = x - u - \frac{A\ u^2}{1\begin{vmatrix}1 & A\\ u & 1\end{vmatrix}} - \frac{\begin{vmatrix}A & B\\ 1 & A\end{vmatrix}\ u^3}{\begin{vmatrix}1 & A\\ u & 1\end{vmatrix}\begin{vmatrix}1 & A & B\\ u & 1 & A\\ 0 & u & 1\end{vmatrix}} - \frac{\begin{vmatrix}A & B & C\\ 1 & A & B\\ u & 1 & A\end{vmatrix}\ u^4}{\begin{vmatrix}1 & A & B\\ u & 1 & A\\ 0 & u & 1\end{vmatrix}\begin{vmatrix}1 & A & B & C\\ u & 1 & A & B\\ 0 & u & 1 & A\\ 0 & 0 & u & 1\end{vmatrix}} - \dots$$

where on the right hand side r terms are to be taken. Formula (2) contains the Halley formula.

$$y_3 = x - u/(1 - A u) .$$

E.T.Hamilton provided the method

$$(3)\qquad y_r = x - u\ R_{r-1} / R_r ,$$

where

$$R_1 = R_2 = 1, \quad R_3 = 1 - A u, \quad R_4 = 1 - 2 A u + B u^2,$$
$$R_5 = 1 - 3 A u + (A^2 + 2 B) u^2 - C u^3, \quad \dots$$

Method (3) is equivalent to method (2).

E.Durand gave an analogous result:

$$(4)\qquad y_r = x - u\ T_{r-1} / T_r ,$$

where

$$T_1 = 1, \quad T_2 = 1 - 2 A u, \quad T_3 = 1 - 3 A u + 3 B u^2,$$
$$T_4 = 1 - 4 A u + (2 A^2 + 4 B) u^2 - 4 C u^3,$$
$$T_5 = 1 - 5A u + (6A^2 + 5 B) u^2 - (5AB + 5C) u^3 + 5 D u^4, \quad \dots$$

Starting from (1), by means of the formula

$$(5)\qquad q_0 = 1, \qquad q_k = - \sum_{i=1}^{k} p_i\ q_{k-i} \quad (k > 0) ,$$

we get the formula

$$(6)\qquad y_r = x - u \Bigg/ \left(\sum_{k=0}^{r-2} q_k u^k \right),$$

with the coefficients:

$$q_0 = 1, \quad q_1 = -A, \quad q_2 = -(A^2 - B), \quad q_3 = -(2A^3 - 3AB + C), \quad \dots$$

Equation (6) contains the Traub formula

$$y_4 = x - u / (1 - Au - (A^2 - B) u^2) .$$

Starting from (6), by means of equation (5), we get the ntinued fraction

$$7) \quad y_r = x - \frac{u}{s_2 -}\,\frac{u}{s_3 -}\,\cdots\,\frac{u}{s_{r-1} -}\,\frac{u}{s_r}$$

here

$$s_2 = 1, \quad s_3 = \frac{1}{A}, \quad s_4 = \frac{A^2}{A^2-B}, \quad s_5 = \frac{(A^2-B)^2}{A^4-A^2B-B^2+AC}, \quad \ldots$$

ormula (7) contains the Halley formula

$$y_3 = x - u/(1-A\,u),$$

s well as the Durand formula

$$y_4 = x - u\,(A-(A^2-B)\,u)\,/\,(A-(2A^2-B)\,u)\,.$$

If the numerator and the denominator of the fraction in 6) are multiplied by the expansion

$$8) \quad 1 + aA\,u + (bA^2+cB)\,u^2 + (dA^3+eAB+gC)\,u^3 + \ldots,$$

where a, b, c, d, e, g, ... are arbitrary coefficients, then by the method of undefined coefficients the following expansion is obtained:

$$(9) \quad y_r = x - u\left(\sum_{k=0}^{r-2} v_k\,u^k\right)\Bigg/\left(\sum_{k=0}^{r-2} w_k\,u^k\right),$$

where

$$\begin{aligned} v_0 &= 1, & w_0 &= 1,\\ v_1 &= a\,A, & w_1 &= (a-1)\,A,\\ v_2 &= b\,A^2 + c\,B, & w_2 &= (b-a-1)\,A^2 + (c+1)\,B,\\ v_3 &= dA^3+eAB+gC, & w_3 &= (d-b-a-2)A^3+(e-c+a+3)AB+(g-1)C, \ \ldots \end{aligned}$$

Formula (9) contains Slavić's formulas

$$y_3 = x - u\,(1+aA\,u)\,/\,(1+(a-1)A\,u)\,.$$

$$y_4 = x - u\,\frac{1+aA\,u+(bA^2+cB)\,u^2}{1+(a-1)A\,u+((b-a-1)A^2+(c+1)B)\,u^2}\,.$$

If the numerator and the denominator of the fraction in (6) are multiplied by an arbitrary parameter t and if +1 -1 are added to the denominator, we get

$$y_r = x - t\,u \Big/ \left(t - 1 + \left(1 + \sum_{k=1}^{r-2} t\, q_k\, u^k\right)\right) .$$

Upon squaring the expression in brackets we get

$$(10) \quad y_r = x - t\,u \Big/ \left(t - 1 + \left(\sum_{k=0}^{r-2} h_k u^k\right)^{1/2}\right),$$

where

$$h_0 = 1, \quad h_1 = -2\,t\,A, \quad h_2 = t(t-2)\,A^2 + 2t\,B,$$
$$h_3 = 2t(t-2)\,A^3 - 2t(t-3)\,AB - 2t\,C, \quad \ldots$$

Formula (10) contains: for $t = 2$ the Euler formula

$$y_3 = x - 2u/(1-(1-4\,A\,u)^{1/2}),$$

for $t = 1$ the Ostrowski formula or the Durand formula

$$y_3 = x - u/(1-2\,A\,u)^{1/2} ,$$

for $t = n/(n-1)$ the Laguerre formula

$$y_3 = x - n\,u \,/\, (1 + ((n-1)^2 - 2n(n-1)\,A\,u)^{1/2})$$

(n is the degree of the polynomial whose zero is sought), the general Hansen-Patrick formula

$$y_3 = x - t\,u \,/\, (t - 1 + (1 - 2tA\,u)^{1/2}),$$

and for $t = 2$ the Traub formula

$$y_4 = x - 2\,u \,/\, (1 + (1 - 4A\,u + 4B\,u^2)^{1/2}) .$$

Equations (9) and (10) are generalization of more above mentioned one-point iteration functions.

*

A. Đorđević, G. V. Milovanović, D. S. Mitrinović, N. Obradović D. B. Popović, D. Đ. Tošić, P. M. Vasić have read this paper in manuscript and have made some valuable remarks and sugestions.

REFERENCES

1. ČEBYŠEV P. L.: Vyčeslenie korney uravnenija. Moskva 1848.

2. DURAND E.: Solutions numériques des équations algébriques, I. Paris 1960.

3. EULER L.: Opera Omnia, Ser. I, Vol. X (1755) 422-455.

. HALLEY E.: A new, exact and easy method of finding the roots of equations generally, and that without any previous reduction. Phil.Trans.Roy.Soc.London 18(1694) 136-145.

5. HAMILTON H.J.: Roots of equations by functional iteration. Duke Math.J. 13(1946) 113-121.

6. HANSEN E., PATRICK M.: A family of root finding methods. Numer.Math. 27(1977) 257-269.

7. LAGUERRE E.N.: Sur une méthode pour obtener par approximation les recines d'une équation algébrique qui a toutes ses racines réelles. Nouvelles Ann.de Math. 2^{e} séries 19(1880) 88-103.

8. NEWTON I.: Letter to G.W.Leibniz, 13.6.1676.

9. OSTROWSKI A.: Solution of equations and systems of equations. New York 1960.

10. SLAVIĆ D.V.: Solution of equations by function modification. Univ.Beograd Publ.Elektrotehn.Fak.Ser.Mat.Fiz. N^{o}735-762 (1982) 127-129.

11. SLAVIĆ D.V.: The generalization of ten rational approximations of iteration functions. These publications.

12. TRAUB J.F.: Iterative methods for the solution of equations. Englewood Cliffs 1964.

13. WHITTAKER E.T.: A formula for the solution of algebraic or transcendental equations. Proc.Edinburgh Math.Soc. 36(1918) 103-106.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON THE CHOICE OF THE INITIAL APPROXIMATION IN SOLVING OF THE OPERATOR EQUATIONS BY THE NEWTON-KANTOROVIČ METHOD

Milenko Ćojbašić

BSTRACT:

he iterative procedure (see [2]) for the choice of the initial approximaion is generalized for the case of solving the equation P(x)=0, where P s a Frechet differentiable operator in a Banach space X. Separately, we onsider the case when P is a integral operator. A numerical example is iven.

IZBORU POČETNE APROKSIMACIJE PRI REŠAVANJU OPERATORSKIH JEDNAČINA NEWON-KANTOROVIČEVOM METODOM. *Iterativni postupak (v. [2]) za izbor početne roksimacije, generalisan je na slučaj rešavanja jednačine P(x)=0, gde e P Frechet diferencijabilan operator u Banach-ovom prostoru X. Razmara se primena na integralne jednačine. Dat je numerički primer.*

1. INTRODUCTION

Let P denote a Frechet differentiable operator in a Banach space X. To find a solution x=x* of the equation

(1) $P(x) = 0,$

one often applies Newton-Kantorovič's method, which consists of the construction of the sequence $\{x_n\}$ defined by

(2) $x_{n+1} = x_n - [P'(x_n)]^{-1} . P(x_n) \qquad n=0,1,2,\ldots,$

starting from some suitable chosen $x_0 \in X$. The sufficient conditions for the success of this procedura are given by the famous theorem of L.V.Kantorovič [1]:

THEOREM 1. If the conditions are satisfied

1) For the initial approximation x_0, the operator $P'(x_0)(\in B(X,Y))$ has inverse, and $\|\Gamma_0\| \le B_0$

2) $\|P(x_0)\| \le \eta_0$

3) Second derivative $P''(x)$ is bounded in the region defined by (4); i.e. $\|P''(x)\| \le K$;

4) The constants B_0, η_0, K satisfy the inequality

(3) $h = B_0^2 \eta_0 K \le \frac{1}{2}$.

Then the equation (1) has the solution x*, which can be find in the ball defined by

(4) $$\|x-x_0\| \leq N(h_0).\eta_0 = \frac{1-\sqrt{1-2h_0}}{h_0}.\eta_0 \;;$$

and the successive approximants x_n of the iterative procedure (2) converge to x^*. For the rapidity of convergence is valid

$$\|x_n-x^*\| \leq \frac{1}{2^{n-1}}.(2h_0)^{2^n-1}.\eta_0 \;.$$

Now let the operator P be integral operator defined by

(5) $$y(s) = x(s) - \int_0^1 K(s,t,x(t))dt;$$

and the sequence $x_n(s)$ is formed ot the next way: the initial approximation $x_0(s)$ is given. The next approximation $x_1(s)$ is defined from the linear integral equation

$$x_1(s)-x_0(s)-\int_0^1 K'_x(s,t,x_0(t))(x_1(t)-x_0(t))dt = \varepsilon_0(s),$$

where

$$\varepsilon_0(s) = \int_0^1 K(s,t,x_0(t))dt-x_0(s).$$

The inequality (3) in this case becomes

(6) $$h = (B+1)^2.\eta.K \leq \frac{1}{2},$$

where, for the initial approximation $x_0(s)$, the kernel $K'_x(s,t,x_0(t)) = K(s,t)$ has the resolvent $G(s,t)$ and

$$\int_0^1 |G(s,t)|dt \leq B; \quad 0 \leq s \leq 1,$$

where η, K have the same meaning as in the theorem 1.

2. THE CHOICE OF THE INITIAL APPROXIMATION

One of the most difficult problems in solving the equation (1) by the Newton-Kantorovič method is the choice of the initial approximation x_0. In the paper [2] is given an iterative procedure for defining the initial approximation in solving the nonlinear system of equation by the Newton-Kantorovič method, which after finite number of steps automatically becomes the Newton-Kantorovič method. We will generalize the method on the case in solving the operator equation (1).

The iterative procedure (2) is replaced by

(7) $$x_{n+1} = x_n - [P'(x_n)]^{-1}[P(x_n)-\alpha_n P(x_0)] \qquad (n=0,1,\ldots),$$

where

(8) $$\alpha_n = \max\left[0, 1-\frac{1}{2K\|P(x_0)\|}\left(\frac{1}{\|[P'(x_n)]^{-1}\|^2} + \frac{3}{4}\sum_{i<n}\frac{1}{\|[P'(x_i)]^{-1}\|^2}\right)\right],$$

The equation (7) can be taken in as the realization of the Newton-Kantorovič method for the equation

(9) $$P(x)-\alpha_n P(x_0)=0 \;, \qquad \alpha_n \in [0,1].$$

LEMMA 1. If the operator $[P'(x_0)]^{-1}$ exists then:

(a) The condition (3) is satisfied for each x_n, which is obtained y the Newton-Kantorovič method for the equation (9); i.e.exists $[P'(x_n)]^{-1}$ nd

$$2K\| [P'(x_n)]^{-1} \|^2 . \| P(x_n)-\alpha_n P(x_0) \| \leq 1 \tag{10}$$

(b) α_n is non increasing seqence; i.e. $\alpha_{n+1} \leq \alpha_n$.

Proof. We prove the lemma by induction (see[2]and [5]). For n=0 the statement is trivial. We suppose that the inequality (10) is valid. Then we get for (n+1)-st step

$$\| x_{n+1}-x_n \| \leq \| [P'(x_n)]^{-1} \| . \| P(x_n)-\alpha_n P(x_0) \| \leq \frac{1}{2K\| [P'(x_n)]^{-1} \|} . \tag{11}$$

Now let us prove that $[P'(x_{n+1})]^{-1}$ exists. Using (11) we get

$$|I-[P'(x_n)]^{-1}.P'(x_{n+1}) \| \leq \| [P'(x_n)]^{-1} \| . \| P'(x_{n+1})-P'(x_n) \| \leq$$
$$\leq K\| [P'(x_n)]^{-1} \| . \| x_{n+1}-x_n \| \leq \frac{1}{2} <1.$$

Using the Banach theorem we conclude that the operator

$$H = (I-(I-[P'(x_n)]^{-1}.P'(x_{n+1}))) \tag{12}$$

has inverse and that is $\| H^{-1} \| \leq 2$. From (12) we simply get

$$\| H^{-1} \| = \| [P'(x_{n+1})]^{-1}.P'(x_n) \| \leq 2,$$

and it follows that exists $[P'(x_{n+1})]^{-1}$ ant that is

$$\| [P'(x_{n+1})]^{-1} \| \leq 2. \| [P'(x_n)]^{-1} \| . \tag{13}$$

Using (13) we get

$$\alpha_n-\alpha_{n+1} = \frac{1}{2K\| P(x_0) \|} . \left(\frac{1}{\| [P'(x_{n+1})]^{-1} \|^2} - \frac{1}{4\| [P'(x_n)]^{-1} \|^2}\right)>0. \tag{14}$$

Now, using the analogous Taylor's formulae (see [1]) for differentiable operators we find

$$\| P(x_{n+1})-\alpha_{n+1} P(x_0) \| \leq \| P(x_{n+1})-P(x_n)-P'(x_n)(x_{n+1}-x_n) \| + \| P(x_0) \| . (\alpha_n-\alpha_{n+1}) \leq$$
$$\leq \frac{\| P''(x_n) \| . \| x_{n+1}-x_n \|^2}{2} + \| P(x_0) \| . (\alpha_n-\alpha_{n+1}).$$

Finally using (11) and (14) we prove that the inequality (1o) is valid, which together with (14) proves the lemma.

Let us consider a convex region G which includes the solution x* of the equation (1). Suppose in G, for the operator $P \in C^2(G)$, exists $|P'(x)|^{-1}$ for each $x \in G$, and $P(x_1) \neq P(x_2)$ for $x_1 \neq x_2$; $x_1, x_2 \in G$. Then x* is the unique solution of the equation (1) in G.

THEOREM 2. For each $x_0 \in G$ the iterative procedure (7) for finite number of steps n_0 leads to the point x_0, for which the condition (3) of the newton-Kantorović method is satisfied, and $\alpha_n=0$, for $n\geq n_0$.

Proof. We first prove that the sequence $\| [P'(x_n)]^{-1} \|$ Is bounded. We suppose the opposite; i.e. that $\| [P'(x_n)]^{-1} \| \to \infty, n\to\infty$. By the lemma $\alpha_n \to \alpha \in [o,1]$, then by (10) $P(x_n)\to \alpha P(x_0)$. From the definition of the region G and characteristics of mapping P, we conclude that P(G) is a convex region, $P(x_0) \in P(G)$ and $P(x^*)=0 \in P(G)$. Therefore is $\alpha P(x_0) \in P(G)$. But then

$$x = P^{-1} (\alpha P(x_0)) \in G,$$

and $x_n \to x$. So

$$\| [P'(x_n)]^{-1} \| \to \| [P'(x)]^{-1} \| \to \infty ,$$

which is in contradiction with assumption. Thus $\| [P'(x_n)]^{-1} \| \leq L < \infty$. Using (8) and the lemma we get that for

$$n \geq n_0 \geq (2KL^2 \| P(x_0) \| -1),$$

$\alpha_n = 0$ and the condition for applying the Newton-Kantorovič method is satisfied.

NOTE 1. In the paper [2] is considered the case when P is the system of nonlinear equations.

We suppose that for the integral equation (5) the condition (6) is not satisfied. Using the lemma 1 for defining the initial approximation we get the iterative procedure

(15) $$\Delta x_n(s) - \int_0^1 K'_x(s,t,x_n(t))\Delta x_n(t)dt = \varepsilon_n(s) - \alpha_n \varepsilon_0(s),$$

where

(16) $$\varepsilon_n(s) = \int_0^1 K(s,t,x_n(s))dt - x_n(s).$$

Then α_n is expressed by (8), where $[P'(x_n)]^{-1}$ is the operator defined with

(17) $$\Delta x_n(s) = \varepsilon_n(s) - \alpha_n \varepsilon_0(s) - \int_0^1 G_n(s,t)(\varepsilon_n(s) - \varepsilon_0(s))dt,$$

and $G_n(s,t)$ is the resolvent of the integral equation with the kernel $K'_x(s,t,x_n(t))$. Using theorem 2 the successive approximative which are get by solving the linear integral equation (15) lead to x_{n_0} for which is the condition (6) for applying the Newton-Kantorovič method is satisfied.

3. NUMERICAL EXAMPLE

The integral equation is given

(18) $$x(s) = 1 - 0.4854.s + s^2 + \int_0^1 st \text{ arc tg} x(t)\, dt,$$

whose exact solution is $x^*(s) = 1 + s^2$. Let us try to use the Newton-Kantorovič method for solving the equation (18), with the initial approximation $x_0(t)=1$. As the kernel

$$k(s,t) = K'_x(s,t,x_0(t)) = \frac{st}{3} .$$

is degenerated, according [3] a resolvent can be find from the integral equation for a resolvent, and we get

$$G(s,t) = \frac{3}{5}.st.$$

Using (16) and the estimation for K(see [4]), we can find B,η,K,h_0

$$B = \max_s \int_0^1 |G(s,t)|dt = \frac{3}{10},\ \eta = \max_s|\varepsilon_0(s)| = 0,9073,$$

$$K = \max_{s,t} |K''_{u^2}(s,t,u)| = 0,6495,\ h_0=(B+1)^2\eta K=0,9959>\frac{1}{2}.$$

So, we can not use the Newton-Kantorovič method. Let us apply the iterative procedure (15) for defining the initial approximation. We easily get $\alpha_0=0,4979$. By solving the integral equation (15) for $n = 0$ we get

$$\Delta x_0(s)=0,5021\ s^2+0,0195s;\ x_1(s)=x_0(s)+\Delta x_0(s)= = 1+0,0195s + 0,5021\ s^2.$$

Likely for $x_1(s)$ we define the constants η_1 and B_1:$\eta_1=0,4416$, $B_1=0,2222$ (see[5]). Now it is

$$h_1 = (B_1+1)^2\eta_1.K = 0,4284 < \frac{1}{2}.$$

So, the condition for using the Newton-Kantorovič method with the initial approximation $x_1(s)=1+0,0195.s + 0,5021.s^2$, is satisfied. For the next iteration we get

$$\Delta x_1(s) = 0,4979\ s^2-0,0135.s;\ x_2(s) = x_1(s)+\Delta x_1(s)=s^2+1+0,0060.s.$$

Since the exact solution is $x^*(s)=1+s^2$, that is the maximal error

$$\max_s|x^*(s)-x_2(s)| = \max_s|0,006.s| = 0,06<10^{-2}.$$

NOTE 2. In the paper [4] for $x_0(s) = \frac{3}{2}$ one obtains $h_0=0,451<0,5$ so it is possible to use the Newton-Kantorovič method immediatelly. Here $x_1(s)=s^2+0,0067\ s+1$.

REFERENCES

1. KANTOROVIČ L.V.: Funkcional'nyj analiz i prikladnaja matematika. Uspehi Mat.Nauk. 3(1948), 89-185.

2. KUL'ČICKIJ O.JU., ŠIMELEVIČ L.I.: O nahoždenii načal'nogo približenija dlja metoda Newtona. Ž.Vyčisl. Mat. i Mat. Fiz. 14(1974),1016-1018.

3. KRASNOV M.L.: Integral'nye uravnenija, Moskva,1975.

4. ZAGADSKIJ D.M.: Približennoe rešenie nelinejnyh integral'nyh uravnenij., Ph.thesis. Pedag.inst.im. A.I.Gercena, Leningrad, 1946.

5. ĆOJBAŠIĆ M.M.: Neki aspekti primene Newton-Kantorovičeve metode, Mr. thesis, PMF Beograd, 1982.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

NUMERICAL SOLUTION OF THE FREDHOLM INTEGRAL EQUATION OF THE FIRST KIND WITH LOGARITHMIC SINGULARITY IN THE KERNEL

Tomaž Slivnik, Gabrijel Tomšič

ABSTRACT:

The paper describes a numerical method for the solution of the Fredholm integral equation of the first kind with logarithmic singularity in the kernel. The method is based on proper substitution for the singularity and on the use of generalized quadrature formulas which allow a faster convegence.

NUMERIČNA REŠITEV FREDHOLMOVE INTEGRALSKE ENAČBE PRVE VRSTE Z LOGARITMIČNO SINGULARNOSTJO V JEDRU. *V članku je opisana numerična metoda za rešitev Fredholmove integralske enačbe z logaritmično singularnostjo v jedru. V metodi je uporabljena posebna substitucija in posplošene kvadraturne formule, ki omogočajo hitro konvergenco.*

1. INTRODUCTION

The solutions of electrostatic problems can be often formulated by the Fredholm integral equations of the first kind. For instance the charge distribution $\sigma(x)$ on the surface of the microstrip transmission line is given in the following form

$$(1) \qquad 1 = \int_{-1}^{1} \sigma(y)G(x,y)\,dy \qquad -1<x<1$$

where

$$G(x,y) = A \sum_{n=1}^{\infty} K^{n-1} \ln\left|\frac{4n^2 + (\frac{x-y}{d})^2}{4(n-1)^2+(\frac{x-y}{d})^2}\right|$$

where A, K<1, d are given constants.

It is well-known that the numerical solution of Fredholm integral equations is not numerically stable process, namely the condition numbers of matrices become with the order of matrices larger and larger. To obtain stable solutions some kind of regularization must be used. Nevertheless in many cases of Fredholm equations of the first kind are solved and very usable results are obtained by using standard numerical processes (with no regularization),[3]. In all such cases the kernel has logarithmic singularity. In this paper the numerical method for the solution of equation (1) is described. For the improvment of the convergence the Richardson extrapolation technique can be used.

2. STATEMENT OF THE PROBLEM

We are trying to find

$$Q = \int_{-1}^{1} \sigma(y)\,dy$$

where $\sigma(y)$ is the solution of the equation (1). The kernel G(x,y) has a logarthmic singularity

$$(2) \qquad G(x,y) = C \ln|x-y| + K(x,y)$$

where K(x,y) is a continous fun ction. It is known that the solution has singularities at the both ends of the interval $[-1,1]$ and $\sigma(y)$ can be represented as ([3])

$$\sigma(y) = \frac{f(y)}{\sqrt{1 - y^2}}$$

where f(y) is the continous function.

3. METHOD FOR THE SOLUTION

For numerical treatment of the equation (1) we apply the generalized quadrature formulas introduced by K.Atkinson, [1]. By introducing new variables

$$x = \cos\alpha$$

$$y = \cos\beta$$

we get

$$(3) \qquad 1 = \int_0^{\pi} S(\alpha)\left[\ln|\cos\alpha - \cos\beta| + H(\alpha,\beta)\right] d\alpha$$

$$\beta \in [0, \pi]$$

where $S(\alpha) = \sigma(\cos\alpha)\sin\alpha$ and $H(\alpha,\beta)$ are continous functions. The kernel can be rewritten

$$\ln|\cos\alpha - \cos\beta| = \ln\left|\frac{\sin\frac{\alpha-\beta}{2}}{\frac{\alpha-\beta}{2}}\right| + \ln\left|\frac{\sin\frac{\alpha+\beta}{2}}{(2\pi-\alpha-\beta)(\alpha+\beta)}\right| +$$

$$+ \ln|\alpha-\beta| + \ln|\alpha+\beta| + \ln|2\pi - \alpha - \beta|,$$

where the first two terms are continous, the last three terms are singular. Continous parts can be approximated by using standard quadrature formulas, the singular parts are approximated by introducing the generalized quadrature formulas of the Newton-Cotes type.

For instance by using the "midpoint rule" we obtain

$$\int_0^{\pi} S(\alpha)\ \ln|\cos\beta_i - \cos\alpha|\,d\alpha = \sum_{j=1}^{n} \alpha_{ij} S(\beta_j)$$

where

$$\beta_i = (i - \tfrac{1}{2})h, \quad h = \frac{\pi}{n}$$

and

$$\alpha_{ij} = \ln\left|\frac{\sin\frac{\beta_i-\beta_j}{2}}{(2\pi-\beta_i-\beta_j)(\beta_i+\beta_j)}\right| + h\ln\left|\frac{\sin\frac{\beta_i+\beta_j}{2}}{\frac{\beta_i - \beta_j}{2}}\right| +$$

$$+\ 3\,h\ln h + h\left|\phi_0(i-j) + \phi_0(1-i-j) + \phi_0(2n-i-j+1)\right|$$

$$\phi_0(\ell) = \int_0^1 \ln|1+\frac{1}{2} - u|\,du = (1+\frac{1}{2})\ln|1+\frac{1}{2}| - (1-\frac{1}{2})\ln|1-\frac{1}{2}| - 1$$

Now the well-known method gives a system of linear equations, which can be solved by standard methods. Observing that

$$Q = \int_{-1}^{1} \sigma(y)dy = \int_0^{\pi} S(\alpha)\ d\alpha$$

the quantity Q may be computed.

4. THE RICHARDSON EXTRAPOLATION

Suppose that Q can be written in the form

$$(4) \qquad Q = Q(h) + A\,h^{\alpha} + B\,h^{\alpha+1} + \ldots$$

where coefficients A,B,... are independent of h. If we consider only the first term of the series (4), we get

$$Q \approx Q(h) + A\,h^{\alpha}$$
$$Q \approx Q(\tfrac{h}{2}) + A(\tfrac{h}{2})^{\alpha}$$

hence

$$\frac{Q - Q(h)}{Q - Q(\frac{h}{2})} \approx 2^{\alpha} = r$$

Suppose that (4) is valid, but we do not know the value of α.

Nevertheless we can compute α experimently for some cases, for whic the exact solution is known. The equation

$$\int_{-1}^{1} \ln|x - y| f(y)dy = 1$$

has the exact solution ([4])

$$f(y) = -\frac{1}{\pi \ln 2\sqrt{1 - y^2}}$$

and

$$\int_{-1}^{1} f(y)dy = -\frac{1}{\ln 2} = -1{\cdot}442695$$

By the numerical way (midpoint rule) we get the following results

n	Q	r
1	-1·56525	4·6
2	-1·47283	4·01
4	-1·45021	4·008
8	-1·44457	4·03
16	-1·44316	

and we can assume that the convergence of the method is quadratic.

The well-known Richardson´s elimination gives the table

-1·56525	-1·44202
-1·47283	-1·44267
-1·45021	-1·44269
-1·44457	-1·44269
-1·44316	

and it is evident that the second column converges very fast to the solution.

We compute extensive tables of Q for the different values of constants K and d. When using, for instance, generalized Simpson´s rule, the four point approximation completely agrees with the results which can be find in the literature, [2].

REFERENCES

[1] ATKINSON K.E.: Extension of the Nyström method for the numerical solution of linear integral equations of the second kind, MRC Report, 1966

[2] SILVESTER P.: TEM wave properties of microstrip transmission lines, Proc.IEEE London, vol.115,No.1, 1968

[3] SILVESTER P., BENEDEK P.: Electrostatic of the Microstrip Revisited, IEEE Trans, on MTT, Nov. 1972

[4] STAKGOLD I.: Boundary Value Problems of Mathematical Physics, MacMillan Co., New York, 1968

Numerical Methods and
Approximation Theory
Niš, September 26 - 28, 1984

ON A CLASS OF COMPLEX POLYNOMIALS HAVING ALL ZEROS IN A HALF DISC

WALTER GAUTSCHI AND GRADIMIR V. MILOVANOVIĆ

ABSTRACT:

e study the location of the zeros of the polynomial $p_n(z) = \pi_n(z) - i\,\theta_{n-1}\pi_{n-1}(z)$*, where* $\{\pi_k\}$ *is a system of monic polynomials orthogonal with respect to an even weight function on* $(-a,a)$*,* $0 < a < \infty$*, and* θ_{n-1} *s a real constant. We show that all zeros of* p_n *lie in the upper half isc* $|z| < a \wedge \mathrm{Im}\, z > 0$*, if* $0 < \theta_{n-1} < \pi_n(a)/\pi_{n-1}(a)$*, and in the lower alf disc* $|z| < a \wedge \mathrm{Im}\, z < 0$*, if* $-\pi_n(a)/\pi_{n-1}(a) < \theta_{n-1} < 0$*. The ultrasphrical weight function is considered as an example.*

O KLASI KOMPLEKSNIH POLINOMA KOJI IMAJU SVE NULE U POLUKRUGU. *U radu se razmatra problem lokalizacije nula polinoma* $p_n(z) = \pi_n(z) - i\theta_{n-1}\pi_{n-1}(z)$*, gde je* $\{\pi_k\}$ *sistem moničnih polinoma ortogonalnih u odnosu na parnu težinsku funkciju na* $(-a,a)$*,* $0 < a < \infty$*, a* θ_{n-1} *realna konstanta. Dokazujemo da sve nule polinoma* p_n *leže u gornjem polukrugu* $|z| < a \wedge \mathrm{Im}\, z > 0$ *, ako je* $0 < \theta_{n-1} < \pi_n(a)/\pi_{n-1}(a)$*, a u donjem polukrugu* $|z| < a \wedge \mathrm{Im}\, z < 0$ *, ako je* $-\pi_n(a)/\pi_{n-1}(a) < \theta_{n-1} < 0$ *. Kao primer razmatrana je ultrasferna težinska funkcija.*

1. INTRODUCTION

In a series of papers, Specht [2] studied the location of the zeros of polynomials expressed as linear combinations of orthogonal polynomials. He obtained various bounds for the modulus of the imaginary part of an arbitrary zero in terms of the expansion coefficients and certain quantities depending only on the respective orthogonal polynomials. Giroux [1] sharpened some of these results by providing bounds for the sum of the moduli of the imaginary parts of all zeros. In the process of doing so, he also stated as a corollary the following result.

Theorem A. *Let*

$$f(x) = (x-x_1)(x-x_2)\dots(x-x_n),$$
$$g(x) = (x-y_1)(x-y_2)\dots(x-y_n),$$

with $x_1 < y_1 < x_2 < \dots < y_{n-1} < x_n$. *Then, for any real number* c, *the zeros of the polynomial* $h(x) = f(x) + icg(x)$ *are all in the half strip* $\operatorname{Im} z \geq 0$, $x_1 \leq \operatorname{Re} z \leq x_n$, *or all are in the conjugate half strip*.

Here we consider special linear combinations of the form

$$p_n(z) = \pi_n(z) - i\theta_{n-1}\pi_{n-1}(z), \tag{1.1}$$

where $\{\pi_k\}$ is a system of monic polynomials orthogonal with respect to an even weight function on $(-a,a)$, $0 < a < \infty$, and θ_{n-1} is a real constant. We combine Theorem A with Rouché's theorem to show, in this case, that all zeros of p_n, under appropriate restrictions on θ_{n-1}, are contained in a half disc of radius a. The result is illustrated in the case of Gegenbauer polynomials.

2. LOCATION OF THE ZEROS OF $p_n(z)$

Let $\omega(x)$ be an even weight function on $(-a,a)$, $0<a<\infty$. Then the monic polynomials orthogonal with respect to $\omega(x)$ satisfy a three-term recurrence relation of the form

$$\begin{cases} \pi_{k+1}(z) = z\pi_k(z) - \beta_k\pi_{k-1}(z), & k=0,1,\dots, \\ \pi_{-1}(z) = 0, \quad \pi_0(z) = 1, \end{cases} \tag{2.1}$$

where $\beta_k > 0$. Since $\pi_k(-z) = (-1)^k\pi_k(z)$, $k=0,1,\dots,$ the polynomial (1.1) can be expanded in the form

$$p_n(z) = z^n - i\theta_{n-1}z^{n-1} + \dots,$$

so that

$$\sum_{k=1}^{n} \zeta_k = i\theta_{n-1},$$

hence

$$\sum_{k=1}^{n} \operatorname{Im}\zeta_k = \theta_{n-1},$$

here $\zeta_1, \zeta_2, \ldots, \zeta_n$ are the zeros of the polynomial (1.1).

By Theorem A and (2.2) all zeros of the polynomial (1.1) lie in the half strip

$$\text{(2.3)} \qquad \operatorname{Im} z > 0, \quad -a < \operatorname{Re} z < a \quad \text{if } \theta_{n-1} > 0,$$

or

$$\text{(2.3')} \qquad \operatorname{Im} z < 0, \quad -a < \operatorname{Re} z < a \quad \text{if } \theta_{n-1} < 0,$$

strict inequality holding in the imaginary part, since $p_n(z)$ for $\theta_{n-1} \neq 0$ cannot have real zeros. Of course, if $\theta_{n-1} = 0$, all zeros lie in $(-a,a)$.

Let D_a be the disc $D_a = \{z : |z| < a\}$ and ∂D_a its boundary. We first prove the following auxiliary result.

Lemma. <u>For each</u> $z \in \partial D_a$ <u>one has</u>

$$\text{(2.4)} \qquad \left| \frac{\pi_k(z)}{\pi_{k-1}(z)} \right| \geqq \frac{\pi_k(a)}{\pi_{k-1}(a)}, \qquad k = 1,2,\ldots .$$

Proof. Let $r_k(z) = \pi_k(z)/\pi_{k-1}(z)$ and $z \in \partial D_a$. We seek lower bounds r_k (not depending on z) of $|r_k(z)|$ for $z \in \partial D_a$,

$$|r_k(z)| \geqq r_k, \qquad z \in \partial D_a.$$

From the recurrence relation (2.1) there follows

$$\text{(2.5)} \qquad r_k(z) = z - \frac{\beta_{k-1}}{r_{k-1}(z)}, \qquad k = 2,3,\ldots,$$

where $r_1(z) = z$. We can take, therefore,

$$\text{(2.6)} \qquad r_1 = a, \quad r_k = a - \frac{\beta_{k-1}}{r_{k-1}}, \qquad k = 2,3,\ldots .$$

Using the usual notation of continued fraction, we obtain from (2.6)

$$r_k = a - \frac{\beta_{k-1}}{a -} \; \frac{\beta_{k-2}}{a -} \cdots \frac{\beta_1}{a} .$$

It is easily seen that $r_k = \pi_k(a)/\pi_{k-1}(a)$, $k \geqq 1$. Indeed, using (2.1),

$$r_k = a - \frac{\beta_{k-1}}{r_{k-1}} = a - \beta_{k-1}\frac{\pi_{k-2}(a)}{\pi_{k-1}(a)} = \frac{\pi_k(a)}{\pi_{k-1}(a)} . \square$$

By a similar argument one could show that

$$2a - \frac{\pi_k(a)}{\pi_{k-1}(a)} \geqq \left|\frac{\pi_k(z)}{\pi_{k-1}(z)}\right| , \quad z \in \partial D_a ,$$

but this will not be needed in the following.

Theorem. *If the constant* θ_{n-1} *satisfies* $0 < \theta_{n-1} < \pi_n(a)/\pi_{n-1}(a)$, *then all zeros of the polynomial* (1.1) *lie in the upper half disc*

$$|z| < a \quad \wedge \quad \operatorname{Im} z > 0 .$$

If $-\pi_n(a)/\pi_{n-1}(a) < \theta_{n-1} < 0$, *then all zeros of* (1.1) *are in the lower half disc*

$$|z| < a \quad \wedge \quad \operatorname{Im} z < 0 .$$

Proof. By (2.4) we have

$$\left|\frac{\pi_n(z)}{\pi_{n-1}(z)}\right| \geqq \frac{\pi_n(a)}{\pi_{n-1}(a)} , \quad z \in \partial D_a , \tag{2.7}$$

hence, if $\pi_n(a)/\pi_{n-1}(a) > |\theta_{n-1}|$,

$$|\pi_n(z)| > |\theta_{n-1}\pi_{n-1}(z)| , \quad z \in \partial D_a .$$

Applying Rouché's theorem to (1.1), we conclude that all zeros of the polynomial p_n lie in the open disc D_a. Combining this with (2.3) or (2.3´), we obtain the assertions of the theorem. □

3. EXAMPLE: GEGENBAUER POLYNOMIALS

We now consider the ultraspherical weight function $\omega(x) = (1-x^2)^{\lambda-1/2}$ $(\lambda > -1/2)$ on $(-1,1)$. In this case, a=1, and $\pi_k(z) = \frac{k!}{2^k(\lambda)_k} C_k^\lambda(x)$, where $C_k^\lambda(x)$ is the Gegenbauer polynomial and $(\lambda)_k$ Pochhammer's symbol, $(\lambda)_k = \lambda(\lambda+1)\ldots(\lambda+k-1)$.

Since

$$\frac{\pi_k(1)}{\pi_{k-1}(1)} = \frac{k}{2(\lambda+k-1)} \cdot \frac{C_k^\lambda(1)}{C_{k-1}^\lambda(1)} = \frac{2\lambda+k-1}{2(\lambda+k-1)}$$

nd

$$D^m C_k^\lambda(x) = 2^m(\lambda)_m C_{k-m}^{\lambda+m}(x), \qquad m \leqq k,$$

here D is the differentiation operator, our theorem implies he following

Corollary. *Let* $\pi_k(z)$ *denote the monic Gegenbauer polynomial of degree* k *with parameter* λ. *If the constant* θ_{n-1} *satisfies* $0 < \theta_{n-1} < \dfrac{2\lambda+n-1}{2(\lambda+n-1)}$, *then all zeros of the polynomial* $_n(z) = \pi_n(z) - i\theta_{n-1}\pi_{n-1}(z)$ *and of its derivatives lie in the pper half disc* $|z|<1 \wedge \mathrm{Im}\, z>0$. *If* $-\dfrac{2\lambda+n-1}{2(\lambda+n-1)} < \theta_{n-1} < 0$, *then they are all in the lower half disc* $|z|<1 \wedge \mathrm{Im}\, z<0$.

The upper bound $(2\lambda+n-1)/(2(\lambda+n-1))$ for $|\theta_{n-1}|$ becomes $n/(2n-1)$ in the case of Legendre polynomials ($\lambda=1/2$), and 1/2 in the case of Chebyshev polynomials ($\lambda=0$).

REFERENCES

1. GIROUX A.: *Estimates for the imaginary parts of the zeros of a polynomial.* Proc. Amer. Math. Soc. 44(1974), 61 - 67.

2. SPECHT W.: *Die Lage der Nullstellen eines Polynomes, I - IV.* Math. Nachr. 15(1956), 353 - 374; 16(1957), 257 - 263; 16 (1957), 369 - 389: 21(1960), 201 - 222.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON THE OPTIMAL CIRCULAR CENTERED FORM

Ljiljana D. Petković

BSTRACT:

ome including circular approximations of the closed set $\{f(z) : z \in Z\}$ f is a closed complex function and $Z = \{c\,;\,r\}$ is a disk) in the cenered form $\{f(c)\,;\,R\}$ are considered in this paper. The optimal centeed form $\{f(c)\,;\,R_0\}$, where $R_0 = \max |f(c) - f(z)|$ $(z \in Z)$, is compared o the centered forms which use Taylor's series. The optimal radius R_0 s determined for some standard (library) functions.

I OPTIMALNOJ KRUŽNOJ CENTRALNOJ FORMI. *U ovom radu su razmatrae neke uključujuće kružne aproksimacije zatvorenog skupa $\{f(z) : z \in Z\}$ f je zatvorena kompleksna funkcija i $Z = \{c\,;\,r\}$ je disk) u centralnoj ormi $\{f(c)\,;\,R\}$. Optimalna centralna forma $\{f(c)\,;\,R_0\}$, gde je $R_0 =$ ax $|f(c) - f(z)|$ $(z \in Z)$, uporedjena je sa centralnim formama koje koriste Taylorov red. Za neke standardne (bibliotečke) funkcije odredjen je optimalni radijus R_0.*

Let $Z = \{w : |w-z| \leq r\} = \{z\,;\,r\}$ $(z \in C,\ r \geq 0)$ be a disk with the center z and the radius r. The set of all disks will be denoted with K(C). Let f be a complex valued function of a complex variable, analitical on the union of all disks which belong to the set $U \subseteq K(C)$, such that the set

$$f^*(Z) = \{f(z) : z \in Z\} = \bigcup_{z \in Z} \{f(z)\}$$

is closed for each $Z \in U_1 \subseteq U$. $f^*(Z)$ will be called *the closed united extension* of the function f over Z. Since the closed region $f^*(Z)$ is not a disk in general, it is of interest for evaluation in circular arithmetic to introduce a disk W which includes the closed set $f^*(Z)$, that is $W \supseteq f^*(Z)$.

The circular region F(Z) such that the inclusion $F(Z) \supseteq f^*(Z)$ holds for each $Z \in U_1$ is *the inclusive disk* for the range $f^*(Z)$. Covering of the exact range $f^*(Z)$ by the inclusive disk will be sometimes called including approximation, or, shorter *I-approximation*. Obviously, I-approximation is better if the quotient area$\{f^*(Z)\}$/area$\{F(Z)\}$ is closer to 1.

In this paper we shall consider inclusive disks of the form $F(Z)= \{f(c) ; R\}$ $(Z = \{c ; r\})$. This form is centered and it will be shortly called *the C-form.*

Among all inclusive disks with the C-form, the best I-approximation is obtained by the disk with the radius

(1) $$R = R_0 = \max_{z \in Z} |f(z) - f(c)| \, .$$

The disk $F_0(Z) = \{f(c) \; ; R_0\}$ is called *the inclusive disk with the optimal C-form.*

Using computer programs, the special attention is dedicated to the computation of standard (subroutine library) functions(e^z, $\ln z$, $\operatorname{arc\,tg} z$, $\sin z$, $\cos z$, z^n, $z^{1/n}$). For some of these functions it is possible to find the optimal radius R_0 according to (1). In a such procedure the following simple assertions, which we give without the proof, will be used.

LEMMA 1: *Let L be a closed region in complex plane. If there exists $w \in L$ such that $|f(\zeta)| \leqq |f(w)|$ for each $\zeta \in L$ then*

$$\max_{\zeta \in L} |f(\zeta)| = |f(w)|.$$

LEMMA 2: *([1]), p. 70). The inequality*

$$|e^z - 1| \leqq e^{|z|} - 1$$

is valid for arbitrary $z \in C$.

LEMMA 3: *Let u and v be real functions of a real variable $t \in [a,b]$, and let $f(t) = u(t) + iv(t)$. If f is R-integrable function, then*

$$\left|\int_a^b f(t)dt\right| \leqq \int_a^b |f(t)|dt.$$

LEMMA 4: *([4], p. 370). If the condition $a_k \geqq 0$ holds for all $k = 0,1,\ldots$ in the disk $|z| < A$, then*

$$\left|\sum_{k=0}^{+\infty} a_k z^k\right| \leqq \sum_{k=0}^{+\infty} a_k |z|^k \quad for \ |z| < A.$$

Let Γ be the boundary of the disk $Z = \{c ; r\}$, $c = |c|e^{i\gamma}$ and let $\Omega = [0 , 2\pi)$, $p = \frac{r}{|c|}$. Then, an arbitrary point $z \in \Gamma$ can be expressed by

$$z = c + re^{i\theta} = c(1 + pe^{i\omega}) \quad (\theta, \omega \in \Omega).$$

We shall now determine the optimal radius R_0 for some standard functions.

$f(z) = e^z$.

Using Lemma 1 and Lemma 2 we get

$$R_0 = \max_{z\in\Gamma} |e^z - e^c| = |e^c| \max_{\theta\in\Omega} |e^{re^{i\theta}} - 1| = |e^c|(e^r - 1).$$

$f(z) = \ln z$.

Let $0 \notin Z$, i.e. $p < 1$ is valid, and let $w = pe^{i\omega}$. Since $|w| < 1$, with regard to Lemma 3 it follows

$$|\ln(1+w)| = \left|\int_0^p \frac{e^{i\omega}}{1+te^{i\omega}}\,dt\right| = \left|\int_0^p \frac{dt}{1+te^{i\omega}}\right| \leqq \int_0^p \left|\frac{1}{1+te^{i\omega}}\right| dt$$

$$= \int_0^p \frac{dt}{(1+t^2+2t\cos\omega)^{1/2}} \leqq \int_0^p \frac{dt}{1-t} = -\ln(1-p).$$

Since $z = c(1+w)$, on the basis of Lemma 1 we obtain

$$R_0 = \max_{z\in\Gamma} |\ln z - \ln c| = \max_{z\in\Gamma} |\ln(1+w)| = -\ln(1-p) .$$

$f(z) = z^n$.

Let $g(\omega) = (1+pe^{i\omega})^n - 1$, $\omega \in \Omega$. Since

$$R_0 = \max_{z\in\Gamma} |z^n - c^n| = |c|^n \max_{\omega\in\Omega} |g(\omega)| = |c|^n \max_{\omega\in\Omega} \left|\sum_{k=1}^{n} \binom{n}{k}(pe^{i\omega})^k\right|,$$

in view of Lemma 4 (taking $a_k = \binom{n}{k}$ and $\zeta = pe^{i\omega}$) we obtain

$$R_0 = |c|^n \max_{\omega\in\Omega} |g(\omega)| = |c|^n g(0) = (|c|+r)^n - |c|^n .$$

$f(z) = z^{1/n}$.

We shall consider only the case $0 \notin Z = \{c\,;\,r\}$ $(p < 1)$. Let $h(\omega) = 1 - (1+pe^{i\omega})^{1/n}$, $\omega \in \Omega$. Using the development in binomial series, we find

$$R_0 = \max_{z\in\Gamma} |z^{1/n} - c^{1/n}| = |c|^{1/n} \max_{\omega\in\Omega} |h(\omega)|$$

$$= |c|^{1/n} \max_{\omega\in\Omega} \left|\sum_{k=1}^{+\infty} \left|\binom{1/n}{k}\right| (-pe^{i\omega})^k\right|.$$

Applying Lemma 4 for $a_k = \left|\binom{1/n}{k}\right|$ and $\zeta = -pe^{i\omega}$, we obtain

$$\left|\sum_{k=1}^{+\infty}\left|\binom{1/n}{k}\right|\left|(-pe^{i\omega})^k\right| \leq \sum_{k=1}^{+\infty}\left|\binom{1/n}{k}\right|p^k = 1-(1-p)^{1/n} = h(\pi),$$

so that

$$R_0 = |c|^{1/n} \max_{\omega\in\Omega} h(\omega) = |c|^{1/n} h(\pi) = |c|^{1/n} - (|c|-r)^{1/n} .$$

Another type of inclusive circular extension with the C-form, based on the development of analytical function by Taylor series, was considered in [5]:

Let f be an analytical function, defined on the union of all disks that belong to the set $U \subseteq K(C)$, such that $f^*(Z) = \{f(z) : z \in Z\}$ is a closed region for each $Z = \{c\,;\,r\} \in U_1 \subseteq U$. Then, for the closed united extension $f^*(Z)$ we have

$$(2) \qquad f^*(Z) \subseteq F_T(Z) = \{f(c)\,;\,R_T\}, \qquad R_T = \sum_{k=1}^{+\infty}\frac{|f^{(k)}(c)|r^k}{k!} .$$

The disk $F_T(Z)$ is called *Taylor's inclusive disk* and R_T *Taylor's radius*.

We shall now compare Taylor's form with the optimal C-form. Let $z = c + re^{it}$ $(t \in \Omega)$ be a point on the boundary Γ of the disk $Z = \{c\,;\,r\}$ and let $u^* \in \Gamma$ be the point which maximizes $|f(z) - f(c)|$. Then $R_0 = |f(u^*)-f(c)|$. Using Taylor series we get

$$R_0 = |f(u^*) - f(c)| = \left|\sum_{k=1}^{+\infty}\frac{f^{(k)}(c)(u^*-c)^k}{k!}\right|$$

$$\leq \sum_{k=1}^{+\infty}\frac{|f^{(k)}(c)|r^k}{k!} = R_T,$$

which means $F_0(Z) \subseteq F_T(Z)$ for each $Z \in U_1 \subseteq K(C)$, i.e. the I-approximation by the disk with optimal C-form is better than by Taylor's inclusive disk. On the other hand, we use clearly defined procedure to evaluate R_T, while the evaluation of R_0 is more complicated and often leads to hard extremal problems. For this reason, it is of interest to apply the disk $F_T(Z)$ instead of $F_0(Z)$, specially in the cases where R_T is close to R_0. It can be shown that for the above considered standard functions the equality $R_0 = R_T$ is valid (see [6]). In the remaining cases the inequality $R_0 < R_T$ holds and, consequently, $F_0(Z) \subset F_T(Z)$ (see Börsken [2]).

Taylor's centered form (2) uses Taylor's development of an analytical function f around the center c of domain $Z = \{c\,;\, r\}$. By expanding f(z) as Taylor's series around the origin (Maclaurin's series), we obtain

$$f(z) = \sum_{k=0}^{+\infty} \frac{f^{(k)}(0)}{k!} z^k . \tag{3}$$

In the sequel we shall use the following formula for the power of a disk $Z = \{c\,;\, r\}$

$$Z^k = \{c\,;\, r\}^k = \{c^k\,;\, (|c|+r)^k - |c|^k\}, \tag{4}$$

which can be obtained from the definition for multiplication of two disks, introduced by Gargantini and Henrici [3].

Natural circular extension of (3) using (4) results in a circular interval

$$F_P(Z) = \sum_{k=0}^{+\infty} \frac{f^{(k)}(0)}{k!} Z^k = \left\{ \sum_{k=0}^{+\infty} \frac{f^{(k)}(0)}{k!} c^k ; \sum_{k=1}^{+\infty} \frac{|f^{(k)}(0)|}{k!}((|c|+r)^k - |c|^k) \right\}$$

or

$$F_P(Z) = \{f(c)\,;\, R_P\} \supseteq \{f(z)\,;\, z \in Z\} = f^*(Z), \tag{5}$$

where

$$R_P = \sum_{k=1}^{+\infty} \frac{|f^{(k)}(0)|}{k!}((|c|+r)^k - |c|^k). \tag{6}$$

The inclusive disk (5) will be called *the power centered form.*

According to the developing procedure it is normally to expect that the power centered form is worse I-approximation for the exact region $f^*(Z)$ than the Taylor's centered form. Thus, we conjecture

$$F_T(Z) \subseteq F_P(Z). \tag{7}$$

The inclusion (7) leads to an equivalent condition in the form of inequality $R_T \leq R_P$, i.e.

$$\sum_{k=1}^{+\infty} \frac{|f^{(k)}(c)|}{k!} r^k \leq \sum_{k=1}^{+\infty} \frac{|f^{(k)}(0)|}{k!}((|c|+r)^k - |c|^k). \tag{8}$$

EXAMPLE 1. Let

$$q(z) = \sum_{k=0}^{n} a_k z^k \quad (a_k \in C)$$

be a polynomial of degree n. On the basis of (2), (5) and (6) we find

$$Q_T(Z) = \left\{ q(c)\,;\, \sum_{k=1}^{n} \frac{|q^{(k)}(c)|\, r^k}{k!} \right\}$$

and

$$Q_P(Z) = \{q(c) \; ; \; \sum_{k=1}^{n} |a_k|((|c| + r)^k - |c|^k)\} \; .$$

The inequality

$$\sum_{k=1}^{n} \frac{|q^{(k)}(c)|r^k}{k!} \leq \sum_{k=1}^{n} |a_k|((|c| + r)^k - |c|^k)$$

has been proved in [7], which gives $Q_T(Z) \subseteq Q_P(Z)$.

EXAMPLE 2. Let $f(z) = e^z$ and $Z = \{c \; ; r\}$ be arbitrary disk. Then

$$F_T(Z) = \{e^c \; ; \; |e^c|(e^r - 1)\},$$

$$F_P(Z) = \{e^c \; ; \; e^{|c|}(e^r - 1)\}.$$

Since $|e^c| \leq e^{|c|}$ it follows $F_T(Z) \subseteq F_P(Z)$ in the case of exponential function.

The above examples confirm the conjecture (7). It is interesting that other considered examples also verify the inclusion (7). But, we are not able to prove the inequality (8) in general case (for arbitrary f) so that the conjectrure (7) remains as an *open problem.*

REFERENCES

1. ABRAMOWICZ M., STEGUN I.A.: Handbook of Mathematical Functions. New York 1965.

2. BÖRSKEN N.C.: *Komplexe Kreisstandardfunktionen.* Freiburger Intervall-Berichte 2(1978).

3. GARGANTINI I., HENRICI P.: *Circular arithmetic and determination of polynomial zeros.* Numer. Math. 18 (1972), 305 - 320.

4. GARNIR H.G.: Fonctions de variables reelles I, Louvain-Paris 1963.

5. PETKOVIĆ L., PETKOVIĆ M.: *The representation of complex circular functions using Taylor series.* ZAMM 61 (1981), 661 - 662.

6. PETKOVIĆ L.: *O najboljim aproksimacijama zatvorenih oblasti pomoću diskova.* In: Proceedings of V Conference PPPR, Stubičke Toplice 1983, 83 - 88.

7. ROKNE J., WU T.: *The circular complex centered form.* Computing 28 (1982), 17 - 30.

Numerical Methods and
Approximation Theory
Niš, September 26 - 28, 1984

TWO METHODS FOR THE CURVE DRAWING IN THE PLANE

Dobrilo Đ. Tošić, Dejan V. Tošić

ABSTRACT:

Two methods for the drawing of the curve, given by the equation $F(x,y) = 0$, are presented. The first method is based on a differential equation $y' = -F_x/F_y$, which enables the the prediction of the next point of a given curve. The position of a predicted point is corrected. The second method has a random choice of the point with the same correction as in the first method. The corresponding program package TPLTS is realised in BASIC.

DVA METODA ZA CRTANJE KRIVIH U RAVNI. Data su dva metoda za crtanje krive date jednačinom $F(x,y) = 0$. Prvi metod je zasnovan na formiranju diferencijalne jednačine $y' = -F_x/F_y$, koja omogućava predikciju sledeće tačke krive. Položaj ove tačke je korigovan. Drugi metod ima slučajni izbor tačke sa istom korekcijom kao o prvom metodu. Odgovarajući programski paket TPLTS realizovan je u BASICu.

1. INTRODUCTION

The curve tracing and curve drawing belongs to the classic exercise. Many papers and books are devoted to the qualitative curve representation (by searching of the characteristic points- particularly singular points), to the investigation of the behaviour of the curve in the neighbourhood those points, the position of asymptots, number of branches, etc. Thus we can draw a skech of the curve which gives some approximation to the truth.

If the equation of the curve is given in an explicit or parametric form, then we have trivial case. The special case appears when the equation is presented in implicit form, i.e. $F(x,y) = 0$, where F is an differentiable function.

In the present paper we expose two methods for the curve drawing. This implies that the curve is to be drawn, with the utmost possible degree of accuracy, has to be considered for this purpose at large number of points on the curve.

The first method will be called "implicit" and the second one "random". Obtained results for the implicit method can be applied in solving of the differential equations, which are given in the implicit form.

2. IMPLICIT METHOD

The method is based on the following principle. Let $F(x,y)=0$ be the equation of the curve, where F is differentiable function. We calculate partial derivatives $\frac{\partial F}{\partial x}=$ $=F_x(x,y)$ and $\frac{\partial F}{\partial y}=F_y(x,y)$. Since $dF=0$, we obtain the differential equation

$$y' = -\frac{F_x(x,y)}{F_y(x,y)} .$$

Let $M_o(x_o,y_o)$ be given initial point belonging to the curve $F(x,y)=0$. We introduce the step h and parameters S_x and S_y in the set $\{-1, 1\}$. Those parameters give the code of initial direction. For example, if $S_x=1$ and $S_y=1$, we take $\Delta x=h$ and $\Delta y=h$; if $S_x=-1$ and $S_y=1$, then $\Delta x=$ $=-h$ and $\Delta y=h$.

First of all we will calculate $F_x(x_o,y_o)$ and $F_y(x_o,y_o)$. If $|F_x(x_o,y_o)| \leq |F_y(x_o,y_o)|$, i.e. $|y'| \leq 1$, by applying the simple Euler method, the prediction of the next point $M_1(x_1,y_1)$ can be obtained, where

$$x_1 = x_o + S_x h, \quad y_1 = y_o + S_x h \frac{F_x(x_o,y_o)}{F_y(x_o,y_o)} .$$

If $|F_x(x_o,y_o)| > |F_y(x_o,y_o)|$, i.e. $|y'| > 1$, then

$$x_1 = x_o + S_y h \frac{F_y(x_o,y_o)}{F_x(x_o,y_o)} , \quad y_1 = y_o + S_y h .$$

The position of the predicted point M_1 can be corrected by the Newton-Raphson method. If

$$|F_x(x_1,y_1)| > |F_y(x_1,y_1)| ,$$

we correct only x_1 by the formula

$$x_1^* = x_1 - \frac{F(x_1,y_1)}{F_x(x_1,y_1)} .$$

In the another case, when $|F_x(x_1,y_1)| \leq |F_y(x_1,y_1)|$, the ordinate y_1 is to be corrected by

$$y_1^* = y_1 - \frac{F(x_1,y_1)}{F_y(x_1,y_1)} .$$

There is also the second method for correction. Let us observe the surface $z = F(x,y)$ and the tangent-plane at the point $(x_1,y_1,F(x_1,y_1))$. The orthogonal projection of the point $M_1(x_1,y_1)$ to the straight line, which is the intersection of tangent-plane with xy-plane, has coordinates

$$x_1^* = x_1 - \frac{F(x_1,y_1)\ F_x(x_1,y_1)}{F_x(x_1,y_1)^2 + F_y(x_1,y_1)^2} ,$$

$$y_1^* = y_1 - \frac{F(x_1,y_1)\ F_y(x_1,y_1)}{F_x(x_1,y_1)^2 + F_y(x_1,y_1)^2} .$$

After the process of a correction we can calculate new coefficients S_x and S_y. Namely, if $\operatorname{sgn}(x_1 - x_o) > 0$, then $S_x = 1$, and if $\operatorname{sgn}(x_1 - x_o) \leq 0$, we have $S_x = -1$. The parameter S_y can be obtained by the analogous procedure.

New point $M_2(x_2,y_2)$ we will obtain by the same method, etc.

The application of above described implicit method can be inhibited in the vicinity of the singular points, because partial derivatives F_x and F_y are in the nearest of zero. Besides, some branch of the curve can be lost, particularly in the case of complex curves. Those difficulties can be avoided by the following method.

3. RANDOM METHOD

The application of random method is oriented to the curve drawing in a given domain in the xy-plane. A domain is usually chosen to be the rectangle R, bounded by lines $x=a$, $x=b$, $y=c$, $y=d$. The equation of the curve is again $F(x,y)=0$, where F is differentiable function.

First ao all, we choose coordinates of the point belonging to the rectangle R by the random generator with the uniform distribution. Afterwards, applying the above described methods, we try to correct predicted point, obtained in such manner. If corrected point does not belong to the curve (with given degree of accuracy), we choose new point, etc.

The random method is completely oriented for the use on computers. It is very efficient for the drawing of curves possessing singular points and several branches. The random method can be succesfully coupled with the implicit method, where parameters S_x and S_y are also at random choosen.

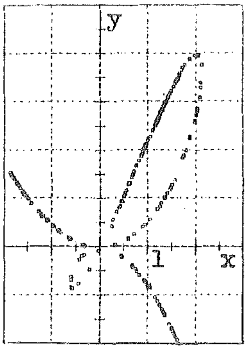
4. PROGRAM REALIZATION

Both methods are realized by the program TPLTS (Tošić PLoT Software) in the BASIC. The concept of the program is realised to be interactive. The modul UNICS (UNIversal Coordinate System) for the drawing of the frame, coordinate net, axis, etc., is particularly developed.

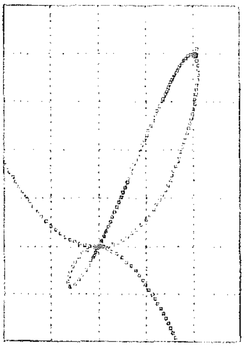
The input activity includes the forming of labels for F, F_x, F_y, the enter of number of curve points, maximum number of corrections of one point, the correction code (a choice of the method for correction), the code of the method for drawing (random, implicit or coupled), the tolerance of the function value (usually 10^{-6}) which is a criterion for the break of the correction, the step h for implicit method, the code of a initial direction (S_x and S_y), coordinates of initial point. After the execution of the program there is a possibility for the restart of some parts of the program, in the aim to obtain new points.

As an illustration of the implicit and random method, wo curves are plotted in figures. Those examples are taken rom the extraordinary book: Persival Frost: An elementary treatise of curve tracing. Fourth edition, London 1926. Macmillan and Co.

1° $x^5 - x^3 y - x y^2 + y^3 = 0$.

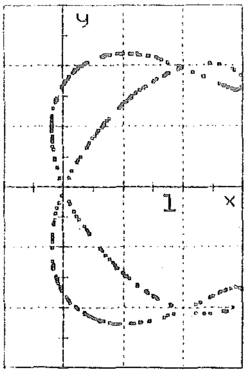


Random

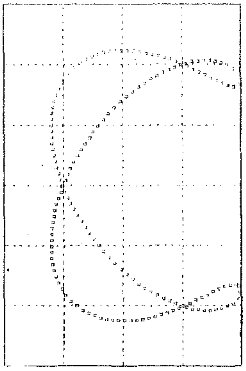


Implicit

2° $(y^2 - x^2)(x - 1)(x - \frac{3}{2}) = 2 (y^2 + x^2 - 2x)^2$.



Random



Implicit

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

OPERATING WITH A SPARSE NORMAL EQUATIONS MATRIX

Brankica Cigrovski, Miljenko Lapaine, Svetozar Petrović

BSTRACT:
n natural and applied sciences, especially in geodesy, symetric normal equations matrices $Q=A^tA$ containing relatively ew nonzero elements occur quite often. Then, one frequently earranges the matrix Q (interchanging columns and rows simulaneously) to bring it to a form more convenient for further reatment. The authors of the present paper have come to see hat in such cases it would make sense to rearrange first the riginal matrix A, and only then to form $Q=A^tA$. So they have elaborated one such procedure and tested it.

OPERIRANJE S MATRICOM NORMALNIH JEDNADŽBI KOJA SADRŽI RELATIVNO MALO ELEMENATA RAZLIČITIH OD NULE. *U prirodnim i tehničkim naukama, a posebno u geodeziji nerijetko se javljaju simetrične matrice normalnih jednadžbi $Q=A^tA$ koje sadrže relativno malo elemenata različitih od nule. Tada se često matrica Q preuredjuje (medjusobnim zamjenama stupaca i istovremeno redaka) da bi bila pogodnija za daljnju obradu. U ovom radu se uočava da bi u takvim slučajevima imalo smisla najprije preurediti polaznu matricu A, pa tek onda formirati $Q=A^tA$. Autori su razradili jedan takav postupak i testirali ga.*

In [2] it was necessary to estimate the accuracy of the 2nd order triangulation net which consisted of 88 trigonometric points interconnected by 666 observed directions. In fact, it was to evaluate the net as a whole, as well as certain parts of it.

On the basis of r.m.s. errors from station adjustment one can obtain only inner accuracy, i.e. the precision of observations. The outer accuracy can be determined only from true errors, the so called misclosures f of triangles. Since in the considered net directions were observed and not angles, misclosures were mutually dependent quantities, therefore one should calculate the r.m.s. error of an observed direction using the formula

$$M_p = \pm \sqrt{\frac{f^tQ^{-1}f}{n}} \tag{1}$$

(e.g. [7] p. 132, [1] p. 257, [11] p. 3). The symmetric nxn-matrix $Q=A^tA$ is the so-called correlation matrix (the normal equations matrix), the mxn-matrix A being a condition equations matrix, and f is a nx1-matrix of misclosures. The number of triangles is n and that of directions m. In our examples n varied from 19 to 180 (see Table 1), m being at most 666.

In all more precise geodetic operations (e.g. 1st and 2nd order triangulation) it is required to give an accuracy estimate prior to adjustment. Therefore, it had been common in geodetic practices to determine the r.m.s. error of an observed direction, using, instead of the strict formula (1), the approximate, much simpler Ferrero's formula (e.g. [7] p.132, [1] p.257, [11] p.11) where the computation had been done on the basis of misclosures f alone, without matrix Q.

Of course, to solve the proposed problem using the formula (1), one doesn't really have to compute the matrix Q^{-1}. It is possible to calculate $f^tQ^{-1}f$ at once, by transforming into triangular form the matrix, which is obtained from Q by adding f^t as the last row, f as the last column, and a zero at the (n+1,n+1)-position. As a result of reduction, the value of $f^tQ^{-1}f$ appears at the position (n+1,n+1). The reduction itself can be carried out by some of the known methods. Our choice was the Cholesky method, adapted for the later described storage scheme for the elements of

The decision was made to solve the problem by using a small desktop computer HP 9845A, the only computer existing at the Geodetic faculty in Zagreb. Namely, the matrix A has been composed by a human being, not by some machine. Thus it was to assume (which was confirmed later) that it would be necessary to correct data repeatedly, of course, using the computer (together with the knowledge of the matrix A special properties) also for the diagnosis of errors. Hence, it seemed more rational (and more interesting) to deal with the problem of handling a greater quantity of data by means of a little computer, which was at hand at every moment, than with frequent visits to some mightier computing system situated in some other institution. Besides, one can imagine that in some future investigations even bigger matrices may appear and the appropriate big computer need not be always at hand, even need not exist at all. Hence, we believe generally that it makes sense to try to exploit every particular computer as efficiently as possible.

Operating with the original matrix A was made easier by its very ;pecial form. Its dimension was up to 666x180, but each column contained)nly 6 nonzero elements, each of them being either +1 or -1.

```
  -2    -10    135     -6    -26      .........
   4     13    162      8     28      .........
   9     91   -148    128   -172      .........
 -13    -97   -165     29    174      .........
 107   -107   -662   -170   -190      .........
-118    110    666    172    193      .........
```

Fig. 1 Storing the matrix A

Thus it was possible to accomodate that matrix into an array not greater than 6x180. Fig. 1 presentsa part of that array for the case n=180. E.g. from the third column of that printout one can read that the third column of matrix A contains:

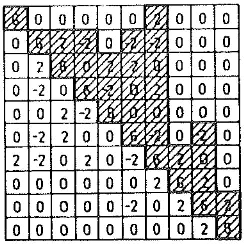
+1 in 135th, 162th and 666th rows,

-1 in 148th, 165th and 662th rows,

0 in all remaining rows.

The computation of individual elements of the matrix $Q=A^tA$ from the elements of the mentioned array presents no considerable problem. The problem is how to store them (suitably for further treatment), because the structure of Q is much more irregular than the structure of A. The number of nonzero elements is not constant but varies from column to column. Also, it cannot be predicted in advance.

We tried out the well-known column by column storage scheme for symmetric matrices - only from the first nonzero element in the column to the diagonal (see [6], as well as Fig.2). When doing so, the profile of the matrix (the quantity of elements to be stored, and to be operated on subsequently) reduces essentially, but still insufficiently for our computer.



6	0	0	0	0	0	2	0	0	0
0	6	2	-2	0	-2	-2	0	0	0
0	2	6	0	2	2	0	0	0	0
0	-2	0	6	-2	0	2	0	0	0
0	0	2	-2	6	0	0	0	0	0
0	-2	2	0	0	6	-2	0	-2	0
2	-2	0	2	0	-2	6	2	0	0
0	0	0	0	0	0	2	6	2	0
0	0	0	0	0	-2	0	2	6	2
0	0	0	0	0	0	0	0	2	6

Fig.2 Profile of matrix Q
(hatched)

The known methods for acting in such cases, e.g. column interchanges (with simultaneous interchanges of corresponding rows) for the symmetric matrix Q in order to reduce the profile

)

:here are a lot of papers dealing with that, let us mention at least [3], [4],[5],[8],[9],[10]) were out of question. Namely, one would have to use a considerable part of central memory to accomodate the program. Thus, there would be a lot of I/O operations in the course of rearranging the matrix Q (which was too large for the central memory). The computer in question uses a tape cartridge as mass storage medium, hence the procedure would progress rather slowly.

Having all that in mind, we concluded that it would be more favorable first to rearrange the matrix A appropriately, and only then to form $Q=A^tA$. The algorithm should be as simple as possible and the computer program short, so that almost everything could happen inside the central unit.

As normal equations (with the matrix of the form $Q=A^tA$) appear frequently in technology and in applied sciences, frequently exactly such whose matrix contains a lot of zeroes (e.g. in geodesy it is the consequence of the properties of geodetic nets, compare e.g. [8]), we consider that the approach proposed in the preceeding paragraph makes much sense. Regarding the realisation of that approach, one could probably also find other solutions, perhaps more elegant or better then ours, which is in its turn very simple and turned out well on examples.

The idea was to rearrange the matrix Q to look as "diagonal" as possible. Namely, as easily seen from Fig.3c. and 3d., the elements q_{li} of the matrix Q i-th column are zero for $l<j$ and for $l>k$, i.e. with A becoming "more diagonal", Q also becomes such and its profile reduces.

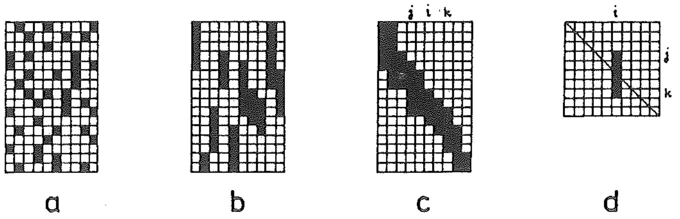


Fig.3 a. the original matrix A
b. A after completition of the first step (row interchanges)
c. A after second step (column interchanges)
d. i-th column of the matrix Q formed from rearranged A (c.)
(Blackened areas contain nonzero elements)

The "pushing" of the matrix A nonzero elements towards diagonal was realized in two steps. The first step represents the "compression" of each individual column. It started by putting six rows which contain nonzeros in the first column to first six rows. Then, the columns which have nonzero element in the first row were found one by one, and the rows containing remaining nonzeros of the column in question arranged in sequence following the rows already arranged up to then (the row which once exchanged place with some preceeding row remains there till the end of procedure, a once processed column is considered never again). After that followed the search for not yet considered columns having nonzeros in the second column etc. The procedure was completed when there remained no untreated column. In this way the matrix A was transformed from the form in Fig.3a. to the one in Fig.3b.

The second step consisted in column interchanges in order to change the situation in Fig.3b. to the one in Fig.3c. It was carried out by arranging the columns in a sequence using as a criterion the last nonzero element in each row (the first nonzero would do equally well, one could also use both of them): columns having nonzeros in the last row became last columns and so on.

	number of elements of the upper triangle		share of	upper trian. diagonals with nonzer.		profile		ratio of
n	total	nonzero	nonzeros	orig.	rearr.	orig.	rearr.	profiles
19	190	56	29%	18	9	123	110	1.12
152	11628	561	4.8%	145	24	5170	2200	2.35
162	13203	755	5.7%	160	35	6269	2725	2.30
173	15051	773	5.1%	171	34	6863	2745	2.50
180	16290	631	3.9%	173	24	6852	2324	2.95

Table 1

From the matrix A rearranged in this manner, the matrix Q was formed having a smaller profile than when formed from the original matrix A. As easily seen from the Table 1, the considered examples showed a significant decrease not only of the profile but also of the bandwidth. Hence, it was also possible to carry out the reduction to triangle in another way - to use some algorithm for banded matrices.

For the time being one can hold that the efficiency of the procedure increases when the format of the matrix Q and the share of zeros in it increase.

Finally, the question emerges: Is rearranging the matrix A before forming Q worth the trouble only in such very special cases when matrix A has the structure described above? We are not of that opinion. Namely, if A had approximately the same share of nonzero elements as in our examples, but the elements were disposed in an irregular pattern and assuming

.....e heterogenous values, A would fit into an array of something more than twofold magnitude, which would be still bearable. In that case e.g. for n=180, in place of an 6x180-array, one would need approximately 13x180 of storage space. Table 1 shows that it would be still less even than the profile of the matrix Q formed from the rearranged A.

REFERENCES

1. BJERHAMMAR A.: *Theory of errors and generalized matrix inverses*, Elsevier Scientific Publishing Company, Amsterdam 1973
2. CIGROVSKI B.: *Accuracy analysis of the observations for the part of the second order triangulation net in the boundary region of SR Croatia and SR BandH (Croatoserbian)*, master´s thesis (in preparation for publication)
3. DUFF I.S., REID J.K.: *A comparison of sparsity orderings for obtaining a pivotal sequence in Gaussian Elimination*, J. Inst. Maths Applics 14(1974), 281-291
4. FUCHS G. von, ROY J.R., SCHREM E.: *Hypermatrix solution of large sets of symmetric positive-definite linear equations*, Computer Methods in Applied Mechanics and Engineering 1(1972), 197-216
5. GIBBS N.E., POOLE W.G.Jr., STOCKMEYER P.K.: *An algorithm for reducing the bandwidth and profile of a sparse matrix*, SIAM J. Numer. Anal. 2(1976), 236-250
6. JENNINGS A.: *A compact storage scheme for the solution of symmetric linear simultaneous equations*, The Computer Journal 9(1966), 281-285
7. MIHAILOVIĆ K.: *Geodezija II (I part), (Serbocroatian)*, Gradjevinska knjiga, Beograd 1974
8. STEIDLER F.: *Zur Lösung grosser schwach besetzter Normalgleichungssysteme in der geodätischen Netzausgleichung*, Zeitschrift für Vermessungswesen 3(1982), 97-108
9. TEWARSON R.P.: *Sparse matrices*, Academic Press, New York and London 1973
10. WAI-HUNG LIU, SHERMAN A.H.: *Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices*, SIAM J. Numer Anal. 2(1976), 198-213.
11. WOLF H.: *Ausgleichungsrechnung II*, Ferd. Dümmlers Verlag, Bonn 1979

...erical Methods and
Approximation Theory
Niš, September 26 - 28, 1984

APPROXIMATION OF 2π - PERIODIC FUNCTIONS BY FUCTIONS OF SHORTER PERIOD

Slobodan D. Miloradović

ABSTRACT:

In this paper we give the value of the exact approximation, for a fixed, 2π - periodic function by functions of period 2π/k where k is a natural number larger than 1. We determine also the exact approximation of a fixed 2π - periodic fuction by funktions of period 2πm/k, where m and k are two mutually prime integers. Them the equality of both approximations is proved. An example which illustrates these results is given at the end.

APROKSIMACIJE 2π - PERIODIČNIH FUNKCIJA FUNKCIJAMA KRAĆE PERIODE. U radu se daje vrednost tačne aproksimacije, fiksirane, 2π - periodične funkcije funkcijama perioda 2π/$_k$ gde je k prirodan broj veći od 1. Takodje se utvrljuje tačna aproksimacija, fiksirane, 2π - periodičke funkcije funkcijama perioda 2πm/k, gde su m i k prirodni, uzajamno prosti, brojevi. Zatim se dokazuje jednakost jedne i druge aproksimacije. Na kraju se daje primer koji lustruje navedene rezultate.

Let C[a,b], as usaul be the space of real continuous functions f defined on the interval [a,b] with norm

$$||f||_{C[a,b]} = \max_{x\in[a,b]} |f(x)|,$$

and let C denote the space of periodic, real and continuous functions f defined on the real line (-∞, +∞), whose period is 2π multiplied by a rational number, with norm

$$||f||_C = \max_x |f(x)|,$$

The set of all periods of a functions f is denoted by Ω_f. For example, if f is a 2π - periodic function, we shall write $2\pi\in\Omega_f$.

In this paper weshall find the value of the exact approximation of a fixed 2π - periodical function f, by a functions ψ of period 2πm/$_k$, where m and k are two mutually prime integers, that is, we shall find the value of

$$\inf_{\psi,\, 2\pi m/_k\in\Omega_\psi} ||f-\psi||_C.$$

Let us mention that, if two functions have different periods, in order to find their distance in the metric C it is enough to find their distance on the smallest inteval in which the periods of the functions are contained a whole number of times. Thus for example, if $2\pi\in\Omega_f$ and $2\pi m/_k \in \Omega_{\bar{\psi}}$ then we have

$$||f - \psi||_C = ||f - \psi||_{C[0,2\pi m]}.$$

Lemma 1. Let $f\in C$, $2\pi\in\Omega_f$ and

$$\overline{f}_k(x) = \max_{0\le s\le k-1} f(x + \frac{2\pi s}{k}),\ \underline{f}_k(x) = \min_{0\le s\le k-1} f(x + \frac{2\pi s}{k}),\ k, s\in N,\ k\ge 2.$$

Then $\overline{f}_k$ and $\underline{f}_k$ are continuous and $\frac{2\pi}{k}\in\Omega_{\overline{f}_k}$ and $2\pi/_k\in\Omega_{\underline{f}_k}$.

Proof. The fact that f is continuous is equivalent to the fact that for every $\varepsilon>0$ there is an $\delta(\varepsilon)>0$ such that

$$(1)\qquad |x'-x''| < \delta(\varepsilon)\Rightarrow|f(x') - f(x'')| < \varepsilon$$

According to equality

$$(2)\qquad |(x' + \frac{2\pi s}{k}) - (x'' + \frac{2\pi s}{k})| = |x' - x''|$$

from (1) and (2) it follws that $|x' - x''| < \delta(\varepsilon)$ implies

$$|f(x' + \frac{2\pi s}{k}) - f(x'' + \frac{2\pi s}{k})| < \varepsilon,\ (0\le s\le k-1),$$

that is

$$-\varepsilon + f(x'' + \frac{2\pi s}{k}) < f(x' + \frac{2\pi s}{k}) < f(x'' + \frac{2\pi s}{k}) + \varepsilon\ (0\le s\le k-1),$$

hence

$$-\varepsilon + \max_{0\le s\le k-1} f(x'' + \frac{2\pi s}{k}) < \max_{0\le s\le k-1} f(x' + \frac{2\pi s}{k}) < \max_{0\le s\le k-1} f(x'' + \frac{2\pi s}{k}) + \varepsilon,$$

that is

$$|\overline{f}_k(x') - \overline{f}_k(x'')| = |\max_{0\le s\le k-1} f(x' + \frac{2\pi s}{k}) - \max_{0\le s\le k-1} f(x'' + \frac{2\pi s}{k})| < \varepsilon,$$

which means that $\overline{f}_k$ is continuous. Accordingly it is proved in the same way that $\underline{f}_k$ is continuous. Since $2\pi\in\Omega_f$, it follows that

$$\overline{f}_k(x + \frac{2\pi}{k}) = \max_{0\le s\le k-1} f(x + \frac{2\pi}{k} + \frac{2\pi s}{k}) = \max_{0\le s\le k-1} f(x + \frac{2\pi(s+1)}{k}) = \max_{0\le s\le k-1} f(x + \frac{2\pi s}{k})$$

that is $2\pi/_k\in\Omega_{\overline{f}_k}$. Similarly, $2\pi/_k\in\Omega_{\underline{f}k}$.

Theorem 2. If $f\in C$, $2\pi\in\Omega_f$, $d_k = \frac{\overline{f}_k - \underline{f}_k}{2}$, then

$$(3)\qquad \inf_{\psi, 2\pi/_k\in\Omega_\psi} ||f - \psi||_{C[0,2\pi]} = ||d_k||_{C[0,2\pi/_k]}.$$

Proof. According to lemma 1 the functions

$$d_k = \frac{\overline{f}_k - \underline{f}_k}{2} \quad \text{and} \quad \psi_k = \frac{\overline{f}_k + \underline{f}_k}{2}$$

ore from C and $2\pi/_k\in\Omega_{dk}$, $2\pi/_k\in\Omega_{\psi_k}$. Since for $x\in[0,2\pi]$

$$f(x) - \psi_k(x) \leqslant \overline{f}_k(x) - \psi_k(x) = d_k(x),$$

$$f(x) - \psi_k(x) \geqslant \underline{f}_k(x) - \psi_k(x) = -d_k(x),$$

then

$$|f(x) - \psi_k(x)| \leqslant d_k(x)$$

and

$$(4) \qquad ||f - \psi_k||_{C[0,2\pi]} \leqslant ||d_k||_{C[0,2\pi/_k]}.$$

Since d_k is continuous on the closed interval, then there is $x_0\in[0,\frac{2\pi}{k}]$ such that $d_k(x_0) = ||d_k||$. For any function $\psi\in C$, $2\pi/_k\in\Omega_\psi$, we have

$$||f - \psi||_{C[0,2\pi]} \geqslant \max_{0\leqslant s\leqslant k-1} |f(x_0 + \tfrac{2\pi s}{k}) - \psi(x_0 + 2\pi s/_k)| =$$

$$= \max_{0\leqslant s\leqslant k-1} |f(x_0 + \tfrac{2\pi s}{k}) - \psi(x_0)| =$$

$$= \max_{0\leqslant s\leqslant k-1} (\max((f(x_0 + 2\pi s/_k) - \psi(x_0), \psi(x_0) - f(x_0 + \tfrac{2\pi s}{k}))) =$$

$$= \max(\max_{0\leqslant s\leqslant k-1} f(x_0 + \tfrac{2\pi s}{k}) - \psi(x_0), \psi(x_0) - \min_{0\leqslant s\leqslant k-1} f(x_0 + \tfrac{2\pi s}{k}))=$$

$$= \max(\overline{f}_k(x_0) - \psi(x_0), \psi(x_0) - \underline{f}_k(x_0)) \geqslant d_k(x_0) = ||d_k||_{C[0,\frac{2\pi}{k}]}$$

that is

$$(5) \qquad ||f - \psi||_{C|0,2\pi|} \geqslant ||d_k||_{C[0,\frac{2\pi}{k}]}.$$

From (4) and (5) we get (3). Theorem 2, in the form of a lemma was proved and used in [1] for determining the exact upper limit of Fourier coefficients on class $H[\delta]_1$ in the space L of integrable functions.

Corollary 3. If $f\in C$, $2\pi\in\Omega_f$, $M = \max_x f(x)$, $m = \min_x f(x)$, then

$$(6) \qquad \inf_{\psi,\frac{2\pi}{k}\in\Omega_\psi} ||f - \psi|| \leqslant \frac{M-m}{2}.$$

We have equality in (6) for such function f for which there is a point x_0 in which $M = \overline{f}_k(x_0)$ and $m = \underline{f}_k(x_0)$.

Corollary 4. If $f \in C$, $2\pi \in \Omega_f$, then

$$\lim_{k\to\infty} \inf_{\psi, \frac{2\pi}{k} \in \Omega_\psi} ||f - \psi||_{C[0,2\pi]} = \frac{M-m}{2} .$$

The sequence $\left(\inf_{\psi, \frac{2\pi}{k} \in \Omega_\psi} ||f - \psi||_{C[0,2\pi]} \right)_{k=2}^{\infty}$ is in general not monotone, which is evident from an example given at the end of this paper. It would be of interest to describe the class of functions $f \in C$, $2\pi \in \Omega_f$, for which the upper set would be monotone.

If a fixed function $f \in C$, $2\pi \in \Omega_f$, is approximated with functions $\psi \in C$, $2\pi m/_k \in \Omega_\psi$, where m and k are mutually prime numbers,then by putting

$$\mathring{f}_k(x) = \max_{0 \leqslant s \leqslant k-1} f(x + \frac{2\pi ms}{k}), \; \underset{\circ}{f}_k(x) = \min_{0 \leqslant s \leqslant k-1} f(x + \frac{2\pi ms}{k}),$$

$$d_k^* = \frac{\mathring{f}_k - \underset{\circ}{f}_k}{2} ;$$

and using lemma 1, we prove, in a similar way like theorem 2.

Theorem 5. If $f \in C$, $2\pi \in \Omega_f$, them

$$(7) \qquad \inf_{\psi, \frac{2\pi m}{k} \in \Omega_\psi} ||f - \psi||_{C[0,2\pi m]} = ||d_k^*||_{C[0, \frac{2\pi m}{k}]} .$$

Since from $\frac{2\pi}{k} \in \Omega_\psi$ it follows that $\frac{2\pi m}{k} \in \Omega_\psi$ then

$$\{\psi: \frac{2\pi}{k} \in \Omega_\psi\} \subset \{ \psi : \frac{2\pi m}{k} \in \Omega_\psi\},$$

and for $f \in C$, $2\pi \in \Omega_f$,

$$(8) \qquad \inf_{\psi, \frac{2\pi}{k} \in \Omega_\psi} ||f - \psi||_{C[0,2\pi]} \geqslant \inf_{\psi, \frac{2\pi m}{k} \in \Omega_\psi} ||f - \psi||_{C[0,2\pi m]}$$

It could be expected that strict that strict inequality is not excluded in (8). We sholl prove, howerer, that both sides of inequality (8) are always equal we shall need.

Lemma 6. Let m and k be mutually prime numbers. By dividing all terms of sequence

(9) $\quad m, 2m, \ldots , (k-1)m$

by k, we obtain the sequence of remainders

(10) $\quad r_1, r_2, \ldots r_{k-1}$

Then (10) is a permutation of the sequence

(11) $\qquad 1, 2, \dots, k-1.$

Proof. It is clear that by dividing by k any term from set (9) we get a remainder which is smaller than k, as well as that the number of terms of (10) is k-1. Accordingly, to prove that the sequence (10) is a permutation of the sequence (11) is enough to prove that there are not equal terms between the terms of (10). Let us suppose the opposite,that is that among terms of set (10) there are two terms lm and nm, $l<n\leq k-1$, which when divided by k give the same remainder:

$$lm = kp + r_n,$$
$$nm = kg + rn,\ p,\ g \in N \cup \{0\},\ r_n < k,\ r_n \in N.$$

By substracting the first equality from the second equality we get

$$(n-l)m = k(g-p).$$

Hence, considering the fact that numbers m and k are mutually it follows that the number n-l is divisible by k, and that does not agree with inequality $0 < n-l < k$.

Theorem 7. If $f \in C$, $2\pi \in \Omega_f$, and m and k are mutually prime numbers, then it follws that

(12) $$\underset{\psi,\ \frac{2\pi}{k} \in \Omega_\psi}{\text{in f}} \ ||f - \psi||_{C[0,2\pi]} = \underset{\psi,\frac{2\pi m}{k} \in \Omega_\psi}{\text{in f}} \ ||f - \psi||_{C[0,2\pi m]}.$$

Proof. Due to lemma 6 $m\cdot s = g_s \cdot k + r_s$, $g_s \in N \cup \{0\}$, where $(r_s)_{s=1}^{k-1}$ is a permutation of sequence (11), and having in view that $2\pi \in \Omega_f$ we get

$$\dot{f}_k(x) = \max_{0\leq s\leq k-1} f(x + \frac{2\pi ms}{k}) = \max_{0\leq s\leq k-1} f(x + 2\pi g_s + \frac{2\pi r_s}{k}) = \max_{0\leq s\leq k-1} f(x + \frac{2\pi s}{k}) =$$
$$= \overline{f}_k(x).$$

Also $\dot{f}_k(x) = \underline{f}_k(x)$ and $\frac{2\pi}{k} \in \Omega_{\dot{f}_k}$, $\frac{2\pi}{k} \in \Omega_{\dot{f}_k}$. According to that we have $2\pi/_k \in \Omega_{d_k^*}$ and for every $x \in [0, \frac{2\pi}{k}]$ $d_k(x) = d_k^*(x)$, that is

(13) $$||d_k||_{C[0,\frac{2\pi}{k}]} = ||d_k^*||_{C[0,\ \frac{2\pi m}{k}]}.$$

From (3), (7) and (13) we get (12).

If m>k then we get $2\pi m_{/k} > 2\pi$, so that Theorem 5 gives a result about approximation of a function $f \in C$, $2\pi \in \Omega_f$, by functions whose period is larger than 2π, also. But theorem 7 confirms that approximation of 2π - periodic function whose periods are larger than 2π are equal with approximations of functions whose periods are smaller than 2π which is in accord with the title of thos paper.

Example 8. Let $f(x) = \cos x$. Then we have

$$(14)\qquad \inf_{\psi,\ \frac{2}{k}\in\Omega_\psi} \|f-\psi\|_{C[0,2\pi]} = \begin{cases} 1, & n = 2s,\ s\in N \\ \cos\frac{\pi}{2k}, & k = 2s+1,\ s\in N. \end{cases}$$

Proof. Let $k = 2s$, $s\in N$. Since $f(x) = \cos x$ it follows that

$$\overline{f}_k(x) = \begin{cases} \cos x, & x\in\left[0, \frac{\pi}{2s}\right] \\ \cos\left(x + \frac{2s-1}{s}\pi\right), & x\in\left[\frac{\pi}{2s}, \frac{\pi}{s}\right], \end{cases}$$

$$\underline{f}_k(x) = \begin{cases} \cos(x+\pi), & x\in\left[0, \frac{\pi}{2s}\right] \\ \cos\left(x + \frac{s-1}{s}\pi\right), & x\in\left[\frac{\pi}{2s}, \frac{\pi}{s}\right], \end{cases}$$

$$d_k(x) = \begin{cases} \cos x, & x\in\left[0, \frac{\pi}{2s}\right], \\ \cos\left(x - \frac{\pi}{s}\right), & x\in\left[\frac{\pi}{2s}, \frac{\pi}{s}\right], \end{cases}$$

and $\|d_k\|_{C[0,\frac{2\pi}{k}]} = 1$. According to theorem 2 we get the first part of equality (14). If $k = 2s+1$. Then we get

$$\overline{f}_k(x) = \begin{cases} \cos x, & x\in\left[0, \frac{\pi}{2s+1}\right] \\ \cos\left(x + \frac{4\pi s}{2s+1}\right), & x\in\left[\frac{\pi}{2s+1}, \frac{2\pi}{2s+1}\right] \end{cases}$$

$$\underline{f}_k(x) = \cos\left(x + \frac{2\pi s}{2s+1}\right), \quad x\in\left[0, \frac{2\pi}{2s+1}\right],$$

$$d_k(x) = \begin{cases} \sin\left(x + \frac{\pi s}{2s+1}\right)\cos\frac{\pi}{2(2s+1)}, & x\in\left[0, \frac{\pi}{2s+1}\right], \\ \sin\left(x + \frac{\pi(s-1)}{2s+1}\right)\cos\frac{\pi}{2(2s+1)}, & x\in\left[\frac{\pi}{2s+1}, \frac{2\pi}{2s+1}\right] \end{cases}$$

and $\|d_k\|_{C[0,\frac{2\pi}{2s+1}]} = \cos\frac{\pi}{2(2s+1)}$, which proves the other part of equality (14).

References

[1] МИЛОРАДОВИЧ С. О коэффициентах Фурье класса $W^rH[\delta_0]_1$. - Publications de l´institut mathématique, Beograd, tome

Numerical Methods and
Approximation Theory
Niš, September 26 - 28, 1984

ON THE STRONG SUMMABILITY (C,α) OF TRANSFORMATIONS OF SIMPLE AND MULTIPLE TRIGONOMETRIC FOURIER SERIES

Vladimir N. Savić

ABSTRACT:

While working on the above subject I found inequalities followed by a number of results on the strong summability (C,α) of transformations of simple and multiple trigonometric Fourier series.

O JAKOJ SUMABILNOSTI (C,α) TRANSFORMACIJA PROSTIH I VIŠESTRUKIH TRIGONOMETRIJSKIH FURIJEOVIH REDOVA. U radu su dokazane nejednakosti iz kojih sleduje niz rezultata o jakoj sumabilnosti (C,α) transformacija prostih i višestrukih trigonometrijskih Furijeovih redova.

Definition 1. Let $f\in L[-\pi,\pi]$,

$$(1)\qquad \frac{a_0}{2}+\sum_{k=1}^{\infty}a_k\cos kx+b_k\sin kx$$

Fourier series of f, $s_n(x;f)$ the sequence of its partial summs, $\alpha>0$, n natural number or zero,

$$A_n^\alpha=\binom{n+\alpha}{n},$$

$$(2)\qquad \sigma_n^\alpha(x,f)=\frac{1}{A_n^\alpha}\sum_{\nu=0}^{n}A_{n-\nu}^{\alpha-1}s_\nu(x,f)=\frac{1}{\pi}\int_0^\pi[f(x+t)+f(x-t)]K_\nu^\alpha(t)\,dt.$$

(C,α)transformation of the series(1), $T=(a_{n,k})$regular matrix and $p>0$. If

$$\lim_{n\to\infty}\sum_{k=0}^{\infty}a_{n,k}\left|\sigma_k^\alpha(x,f)-f(x)\right|^p=0$$

we say that series (1) (H,p,T,α) (or strong)is summabil in point x towards $f(x)$. In addition to the abovestated, if $f\in C[-\pi,\pi]$ and

$$\lim_{n\to\infty}\left\|\sum_{k=0}^{\infty}a_{n,k}\left|\sigma_k^\alpha(x,f)-f(x)\right|^p\right\|_C=0$$

we say that series (1)(H,p,T,α)(or strong) is summabil uniformly at $[-\pi,\pi]$ towards the f function.

<u>Theorem 1.</u> Let $f\in C[-\pi,\pi]$ and $\|f\|_C\le M$, then for each $p>0,\ k\ge 1$ and $\alpha>0$

$$\frac{1}{k+1}\sum_{\nu=k-1}^{2k-1}|\sigma_\nu^\alpha(x,f)|^p\le C(p,\alpha)M^p$$

where $C(p,\alpha)$ is a positive constant depending only on p and α.

<u>Proof.</u> It is sufficient to examine the case $0<\alpha<1$. From (2), putting it as

$$\sigma_n^\alpha(x,f)=\frac{1}{\pi}\left(\int_0^{\frac{1}{k}}+\int_{\frac{1}{k}}^{\pi}\right),$$

and from

$$|K_\nu^\alpha(t)|<\nu+1\quad;\quad|K_\nu^\alpha(t)|<A(\alpha)\nu^{-\alpha}t^{-(\alpha+1)}$$

for $\nu=1,2,3,\dots$; $0<t<\pi$ there follows

$$(3)\qquad\left|\frac{1}{\pi}\int_0^{\frac{1}{k}}\right|\le 2M,\ \left|\frac{1}{\pi}\int_{\frac{1}{k}}^{\pi}\right|<A_1(\alpha)M$$

where $A(\alpha)$, $A_1(\alpha)$ are the positive constants depending only on α

From inequality

$$|a+b|^p\le 2^p(|a|^p+|b|^p)\qquad(\forall a,b\in\mathbb{R})(\forall p>0)$$

and (3) there follows

$$\frac{1}{k+1}\sum_{\nu=k-1}^{2k-1}|\sigma_\nu^\alpha(x,f)|^p\le 2^p(2^p+A_1(\alpha))M.$$

<u>Theorem 2.</u> Let $f\in C[-\pi,\pi]$ then for each $p>0, k\ge 1,\ \alpha>0$

$$\sum_{\nu=2k-2}^{\infty}|\sigma_\nu^\alpha(x,f)-f(x)|^p\alpha_\nu\le C_1(p,\alpha)\sum_{\nu=k-1}^{\infty}[E_\nu(f)]^p\alpha_\nu$$

where α_ν is any non-negative, not-increasing numbers, $E_\nu(f)$ the best approximation of the f function to the trigonometric polynomials of the degree $\le\nu$ in distanceof the space $C[-\pi,\pi]$, and $C_1(p,\alpha)$ a positive constant depending only on p and α.

<u>Proof.</u> According to Theorem 1 there follows

$$\sum_{\nu=2k-2}^{\infty}|\sigma_\nu^\alpha(x,f)-f(x)|^p\alpha_\nu=\sum_{n=1}^{\infty}\sum_{\nu=2^nk-2}^{2^{n+1}k-3}|\sigma_\nu^\alpha(x,f)-f(x)|^p\le$$

$$\le 2^{p+1}C(p,\alpha)\sum_{n=1}^{\infty}2^nk[E_{2^nk-2}(f)]^p\alpha_{2^nk-2}\le$$

$$\le 2^{p+2}C(p,\alpha)\sum_{n=1}^{\infty}\sum_{\nu=2^{n-1}k-1}^{2^nk-2}[E_\nu(f)]^p\alpha_\nu=C_1(p,\alpha)\sum_{\nu=k-1}^{\infty}[E_\nu(f)]^p\alpha_\nu.$$

Definition 2. Let $F_n \downarrow 0$ any sequence and C_F collection of the function $f \in \mathbb{C}[-\pi,\pi]$ for which $E_n(f) \leq F_n$.

Theorem 3. Let $f \in C_F$ and T=(a_{nk}) a non-negative regular matrix the elements of which satisfy the condition

$$a_{n,k+1} \leq a_{n,k} \qquad (\forall n = 0,1,2,\dots)$$

Then, there exists the sequence $p(n)\uparrow+\infty$ such that

$$\lim_{n\to\infty} \left\| \left\{ \sum_{\nu=0}^{\infty} a_{n\nu} \left| \sigma_\nu^\alpha(x,f) - f(x) \right|^{p(n)} \right\}^{\frac{1}{p(n)}} \right\|_{\mathbb{C}} = 0 .$$

Theorem is proved as in (2).

Theorem 4. If $f \in C_F$ and α_ν is any sequence of non-negative, not increasing numbers, then for each $p>0, k\geq 1$ and $\alpha>0$ the following inequality is correct:

$$(4) \quad \sum_{\nu=2k-2}^{\infty} F_\nu^p \alpha_\nu \leq \sup_{f\in C_F} \left\| \sum_{\nu=2k-2}^{\infty} \left| \sigma_\nu^\alpha(x,f) - f(x) \right|^p \alpha_\nu \right\|_{\mathbb{C}} \leq$$

$$\leq C_1(p,\alpha) \sum_{\nu=k-1}^{\infty} F_\nu^p \alpha_\nu$$

Proof. The right hand inequality results from Theorem 2.

Let

$$f_0(x) = \sum_{n=1}^{\infty} (F_{n-1} - F_n) \cos nx .$$

Then

$$f_0(0) - \sigma_\nu^\alpha(0, f_0) = F_\nu , \quad f_0 \in C_F$$

and

$$\left\| \sum_{\nu=2k-2}^{\infty} \left| \sigma_\nu^\alpha(x,f_0) - f_0(x) \right|^p \alpha_\nu \right\|_{\mathbb{C}} \geq \sum_{\nu=2k-2}^{\infty} F_\nu^p \alpha_\nu$$

by which Theorem 4 is proved.

From inequality (4), it results that the approximation rate to the strong means, (C,α) $(\alpha>0)$transformations of trigonometric Fourier series of the function $f\in C_F$(for the whole CF collection) cannot be improved.

Similarly, we testify and prove the theorems relating to (C,α) $(\alpha>0)$ transformations of multiple trigonometric Fourier series.

REFERENCES

1. Bari N.K.:Trigonometrijski redovi,Fizmatgiz,Moskva 1961.

2. Gogoladze L.D.:O jakom sumiranju prostih i višestrukih trigonometrijskih Furijeovih redova,Zbornik radova:"Neka pitanja teorije funkcija",tom 2.izdanje Tbilisijskog univerziteta,Tbilisi 1981.str.5.-30.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

N ESTIMATION FOR REMAINDER OF ANALYTICAL FUNCTION IN TAYLOR'S SERIES

etar M. Vasić, Igor Ž. Milovanović and Josip E. Pečarić

BSTRACT:
n this paper some estimations of difference modul of analyical functions and referred Taylor's polynomials are given. he obtained results are illustrated on a certain concrete f cases, i.e. on exponential, sinhyperbolic and cosinhyperolic function.

OCENI OSTATKA ANALITIČKE FUNKCIJE U TAYLOROVOM RAZVOJU. U ovom radu date su neke ocene modula razlika analitičkih funkcija i odgovarajućih Taylorovih polinoma. Dobijene ocene su ilustrovane na konkretnim slučajevima, tj. na eksponencijalnoj, sinushiperboličnoj i cosinushiperboličnoj funkciji.

We shall prove first a more general result for analytical functions:

THEOREM 1. Let $z\mapsto f(z)$ be an analytical function in the circle $|z|< R$. Let functions $z\mapsto f^{(k)}(z)$, $k\in N_0$, map real axis in real axis. If a natural number r exist, $1\le r\le n$, so that $0\le f^{(n+k)}(0)\le f^{(r+k)}(0)$, $k\in N$, the inequality

$$(1)\quad \frac{|z|^{r+1}}{(r+1)!}\left|f(z)-\sum_{k=0}^{n}\frac{f^{(k)}(0)}{k!}z^k\right| \le \left(\frac{z^{n+1}}{(n+1)!}f(z)-\sum_{k=0}^{r}\frac{f^{(k)}(0)}{k!}|z|^k\right)$$

holds.

Proof. Assume that all conditions given in theorem 1 are fullfiled. Then

$$|f(z)-\sum_{k=0}^{n}\frac{f^{(k)}(0)}{k!}z^k| = |\sum_{k=n+1}^{+\infty}\frac{f^{(k)}(0)}{k!}z^k| \le \sum_{k=n+1}^{+\infty}\frac{f^{(k)}(0)}{k!}|z|^k$$

$$\le \frac{|z|^{n-r}}{(n+1)\dots(r+2)}\left(\frac{|z|^{r+1}}{(r+1)!}f^{(n+1)}(0)+\frac{|z|^{r+2}}{(r+2)!}f^{(n+2)}(0)+\dots\right)$$

$$\le \frac{|z|^{n-r}}{(n+1)\dots(r+2)}\left(\frac{|z|^{r+1}}{(r+1)!}f^{(r+1)}(0)+\frac{|z|^{r+2}}{(r+2)!}f^{(r+2)}(0)+\dots\right)$$

$$= \frac{|z|^{n-r}}{(n+1)\dots(r+2)}\left(f(|z|)-\sum_{k=0}^{r}\frac{f^{(k)}(0)}{k!}|z|^{k}\right),$$

wherefrom the inequality (1) is obtained.

In the similar way the following theorem can be proved.

THEOREM 2. Let $z\mapsto f(z)$ and $z\to g(z)$ be analytical functions in the circle $|z|<R$. Let functions $z\mapsto f^{(k)}(z)$ and $z\mapsto g^{(k)}(z)$, $k\in N_0$, map real axis in real axis. If a natural number r exist, $1\le r\le n$, so that $0\le f^{(n+k)}(0)\le g^{(r+k)}(0)$, $k\in N$, then the inequality

$$(2)\quad \frac{|z|^{r+1}}{(r+1)!}\left|f(z)-\sum_{k=0}^{n}\frac{f^{(k)}(0)}{k!}z^{k}\right|\le\frac{|z|^{n+1}}{(n+1)!}\Big(g(|z|)-\sum_{k=0}^{r}\frac{g^{(k)}(0)}{k!}|z$$

holds.

On the basis of inequalities (1) and (2) we shall give some approximations for concrete analytical functions.

If we put $f(z)=e^{z}$, then $f^{(k)}(0)=1$, $k\in N_0$, on the basis of the inequality (1) we obtain

$$(3)\quad \frac{|z|^{r+1}}{(r+1)!}\left|e^{z}-\sum_{k=0}^{n}\frac{z^{k}}{k!}\right|\le\frac{|z|^{n+1}}{(n+1)!}\left(e^{|z|}-\sum_{k=0}^{r}\frac{|z|^{k}}{k!}\right).$$

The inequality (3) (see [1]) is a generalization of Garnir's inequality (see for example [2,p.323])

$$\left|e^{z}-\left(1+\frac{z}{1!}+\dots+\frac{z^{n}}{n!}\right)\right|\le\frac{|z|^{n+1}}{(n+1)!}e^{|z|}.$$

If we put that

$$T_{2n-1}(z) = z + \frac{z^3}{3!} + \dots + \frac{z^{2n-1}}{(2n-1)!}, \quad T_{-1}(z) = 0,$$

and

$$T_{2n-2}(z) = 1 + \frac{z^2}{2!} + \dots + \frac{z^{2n-2}}{(2n-2)!}, \quad T_{-2}(z) = 0,$$

then for $f(z) = \mathrm{sh} z$ and $f(z) = \mathrm{ch}\, z$ we obtain inequalities

$$|\mathrm{sh}\, z - T_{2n-1}(z)| \leq \frac{(2n-2m+1)!}{(2n+1)!} |z|^{2m} (\mathrm{sh}|z| - T_{2n-2m-1}(|z|))$$

and

$$|\mathrm{ch}\, z - T_{2n-2}(z)| \leq \frac{(2n-2m)!}{(2n)!} |z|^{2m} (\mathrm{ch}|z| - T_{2n-2m-2}(|z|))$$

for $m=1,\dots,n$, respectively.

In the same way on the basis of tne inequality (2) we obtain inequalities

$$|\mathrm{sh}\, z - T_{2n-1}(z)| \leq \frac{(2n-2m)!}{(2n+1)!} |z|^{2m+1} (\mathrm{ch}|z| - T_{2n-2m-2}(|z|))$$

and

$$|\mathrm{ch}\, z - T_{2n-2}(z)| \leq \frac{(2n-2m+1)!}{(2n)!} |z|^{2m-1} (\mathrm{sh}|z| - T_{2n-2m-1}(|z|))$$

for $m = 0,1,\dots,n$.

On the basis of P.R.Beesack's remark (see [2]) we shall prove the following result:

THEOREM 3. If z is the complex number, $|z| \leq \sqrt{(2n+2)(2n+3)}$ then the inequality

$$|\mathrm{sh}\, z - T_{2n-1}(z)| \leq \frac{(2n+2)(2n+3)}{(2n+1)!} \; \frac{|z|^{2n+1}}{(2n+2)(2n+3) - |z|^2}$$

holds.

Proof. As

$$|\mathrm{sh}\, z - T_{2n-1}(z)| \leq \frac{|z|^{2n+1}}{(2n+1)!}\left(1 + \frac{|z|^2}{(2n+2)(2n+3)} + \dots\right)$$

$$\leq \frac{|z|^{2n+1}}{(2n+1)!}\left(1 + \frac{|z|^2}{(2n+2)(2n+3)} + \frac{|z|^4}{(2n+2)^2(2n+3)^2} + \dots\right)$$

$$= \frac{|z|^{2n+1}}{(2n+1)!} \frac{(2n+2)(2n+3)}{(2n+2)(2n+3) - |z|^2} ,$$

the wanted inequality is obtained.

In the similar way the following result is obtained:

THEOREM 4. If z is the complex number, $|z| < \sqrt{(2n+1)(2n+2)}$ then the inequality

$$|\,\mathrm{ch}\, z - T_{2n-2}(z)\,| \leq \frac{(2n+1)(2n+2)}{(2n)!} \frac{|z|^{2n}}{(2n+1)(2n+2) - z^2}$$

holds.

REFERENCES

1 . MARTIĆ B. : Some inequalities connected with exponential function. Mat.Vesnik 12(17) (1975),163-166.

2 . MITRINOVIĆ D.S.(In cooperation with P.M.Vasić): Analytic Inequalities. Berlin-Heidelberg-New York, 1970.

...merical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON A METHOD OF NUMERICAL DIFFERENTATION

Bogdan M. Damnjanović

ABSTRACT:

A method for numerical differentation of a function assigned tabelarly is described in the paper. The orthogonal system

$$\pi^{-1}\cos x,\quad \pi^{-1}\sin x,\quad \pi^{-1}\cos 2x,\quad \pi^{-1}\sin 2x,\ldots$$

is used.

O METODU ZA NUMERIČKO DIFERENCIRANJE. U radu je opisan metod numeričkog diferenciranja funkcije zadate tabelarno. Korišćen je ortogonalan sistem

$$\pi^{-1}\cos x,\quad \pi^{-1}\sin x,\quad \pi^{-1}\cos 2x,\quad \pi^{-1}\sin 2x,\ldots\ .$$

Let $f(x_1), f(x_2), \ldots, f(x_n)$ be the function values, found by a measuring for a series of real arguments $x_1, x_2, \ldots, x_n$. Denote the most convenient empirical formula concerning f by f_β. For the function f we assume that:

1° it is defined over $[-\pi, \pi]$,

2° it has continuous derivatives f' and f'' on $[-\pi,\pi]$,

3° the following inequalities are true

$$(1)\qquad \int_{-\pi}^{\pi}(f''(x))^2\,dx < \infty\ ,$$

$$(2)\qquad \|f - f_\beta\| = \left(\int_{-\pi}^{\pi}[f(x) - f_\beta(x)]^2\,dx\right)^{1/2} \leqq \beta,$$

where $\beta > 0$ is assigned.

On the basis of the known function f_β, the construction of a polynomial $P_{n,\beta}(x)$ which approximates evenly the function $f'(x)$ over $(-\pi,\pi)$, is presented in this paper. This procedure is as follows:

The known function $f_ß$ is approximated by a polynomial $q_{n+1}(x)$ of degree n+1, such one that the inequality

(3) $$\| f_ß - q_{n+1} \| \leqq ß$$

holds. The polynomial q_{n+1} is obtained developing $f_ß$ by Fourier' series using the orthogonal system

(*) $$\pi^{-1} \cos x, \quad \pi^{-1} \sin x, \quad \pi^{-1} \cos 2x, \quad \pi^{-1} \sin 2x \ , \ \ldots,$$

namely,

(4) $$q_{n+1}(x) = \frac{1}{\pi} \sum_{k=1}^{n+1} (a_k \cos kx + b_k \sin kx),$$

where

(5) $$a_k = \int_{-\pi}^{\pi} f_ß(x) \cos kx \, dx, \qquad b_k = \int_{-\pi}^{\pi} f_ß(x) \sin kx \, dx,$$

and n is chosen so that (3) is valid.

According to (2) and (3) it follows

(6) $$\|f - q_{n+1}\| \leqq \|f - f_ß\| + \|f_ß - q_{n+1}\| \leqq 2ß.$$

If the function f' is marked by $f'(x) = u_o(x)$, then under the condition (1) $u_o(x)$ is the unique solution of the integral equation

(7) $$\int_{-\pi}^{x} u(t) \, dt = f(x).$$

Let (7) be written in the operator form

(8) $$Au = f.$$

Introducing the functionalls

(9) $$I[u, q_{n+1}] = \|Au - q_{n+1}\|^2$$

and

(10) $$K[u] = \| u' \|^2 ,$$

the approximate solution of the equation (7) (i.e. (8)) can be found by minimizing the functional $K[u]$ under the condition

(11) $$I[u, q_{n+1}] \leqq 4ß^2$$

(according to (6)). Denote a such minimum by $u_ß(x)$. Then, it is easy to prove that

(12) $$I[u_ß, q_{n+1}] = 4ß^2.$$

In the following, we will show that $u_{ß}(x)$ is the required minimum if the condition

13) $$ß < \frac{1}{2}||q_{n+1}||$$

is satisfied.

Let

$$Au = v$$

be Fourier's development of v using the orthogonal system (*), namely

$$v(x) \sim \frac{1}{\pi}\sum_{k=1}^{\infty}(c_k \cos kx + d_k \sin kx).$$

Since $u(x) = v'(x)$, $u'(x) = v''(x)$, it follows

(14) $$u(x) \sim \frac{1}{\pi}\sum_{k=1}^{\infty}[c_k(-k \sin kx) + d_k k \cos kx],$$

and

$$u'(x) \sim -\frac{1}{\pi}\sum_{k=1}^{\infty}k^2(c_k \cos kx + d_k \sin kx),$$

where

$$c_k = -k^{-2}\int_{-\pi}^{\pi}u'(x) \cos kx\, dx,$$

$$d_k = -k^{-2}\int_{-\pi}^{\pi}u'(x) \sin kx\, dx.$$

Now, we have

$$K[u] = \int_{-\pi}^{\pi}[u'(x)]^2 dx = \int_{-\pi}^{\pi}[-\frac{1}{\pi}\sum_{k=1}^{\infty}k^2(c_k \cos kx + d_k \sin kx)]^2 dx$$

$$= \frac{1}{\pi}\sum_{k=1}^{\infty}k^4(c_k{}^2 + d_k{}^2)$$

and

$$I[u,q_{n+1}] = \int_{-\pi}^{\pi}\left[\int_{-\pi}^{\pi}u(x)dx - q_{n+1}(x)\right]^2 dx = \int_{-\pi}^{\pi}[v(x) - q_{n+1}(x)]^2 dx$$

$$= \int_{-\pi}^{\pi}\left[\frac{1}{\pi}\sum_{k=1}^{\infty}(c_k \cos kx + d_k \sin kx) - \frac{1}{\pi}\sum_{k=1}^{n+1}(a_k \cos kx + b_k \sin kx)\right]^2 dx$$

$$= \frac{1}{\pi}\sum_{k=1}^{n+1}\left[(c_k - a_k)^2 + (d_k - b_k)^2\right] + \sum_{k=n+2}^{\infty}(c_k{}^2 + d_k{}^2).$$

The coefficients c_k and d_k are defined by looking for the minimum of the functional $K[u]$ under the condition $I[u,q_{n+1}] = 4ß^2$ using the method of regulation:

)

$$\Phi[u,q_{n+1}] = K[u] + \eta I[u,q_{n+1}], \quad I[u,q_{n+1}] = 4\beta^2.$$

If we put $\lambda = \frac{1}{\eta} > 0$, then $\Phi[u,q_{n+1}]$ becomes

$$\Phi[u,q_{n+1}] = \frac{1}{\pi}\sum_{k=1}^{n+1} [(c_k - a_k)^2 + (d_k - b_k)^2]$$

$$+ \sum_{k=n+2}^{\infty} (c_k^2 + b_k^2) + \frac{\lambda}{\pi}\sum_{k=1}^{\infty} k^4(c_k^2 + b_k^2)$$

$$= \frac{1}{\pi}\sum_{k=1}^{n+1} [(c_k^2 - a_k^2) + (d_k - b_k)^2 + \frac{\lambda}{\pi} k^4(c_k^2 + b_k^2)]$$

$$+ \sum_{k=n+2}^{\infty} (\frac{1}{\pi} + \frac{\lambda}{\pi} k^4)(c_k^2 + d_k^2).$$

Since the second sum is nonnegative, for the minimum is supposec that $c_k \to 0$ and $d_k \to 0$ for $k \geqq n+2$. Then, according to

$$\frac{\partial\Phi}{\partial c_k} = 0 \quad \text{and} \quad \frac{\partial\Phi}{\partial d_k} = 0 ,$$

we find

(15) $$c_k = \frac{\pi a_k}{1 + \lambda k^4}, \quad d_k = \frac{\pi b_k}{1 + \lambda k^4}.$$

On the basis of (14) we obtain

(16) $$u_\beta(x) = \frac{1}{\pi}\sum_{k=1}^{n+1} \frac{a_k \cos kx + b_k \sin kx}{1 + \lambda k^4} ,$$

where λ is determined from

(17) $$\| u_\beta - q_{n+1} \|^2 = 4\beta^2,$$

that is

(18) $$\frac{1}{\pi}\sum_{k=1}^{n+1} \lambda^2 k^8 (a_k^2 + b_k^2)(1 + \lambda k^4)^{-2} = 4\beta^2.$$

The left-hand side of (18) is monotonically increasing function of λ which tends to

$$\frac{1}{\pi}\sum_{k=1}^{n+1} (a_k^2 + b_k^2)$$

when $\lambda \to \infty$. Besides, $\beta \to 0$ if $\lambda \to 0$ so that

(19) $$\frac{1}{\pi}\sum_{k=1}^{n+1} (a_k^2 + b_k^2) > 4\beta^2,$$

[illegible] from we immediately (13). Therefore, $u_{\beta}(x)$ will be mi-imum of $\Phi[u,q_{n+1}]$ if (13) is valid.

It is easy to prove that $u_{\beta}(x)$ is minimum of $\Phi[u,q_{n+1}]$ for each $x \in (-\pi,\pi)$ and $\lim\limits_{\beta \to 0} u_{\beta}(x) = u_o(x) = f'(x)$ evenly on $(-\pi,\pi)$.

According to the above, the polynomial $P_{n,\beta}(x)$ has the form

$$P_{n,\beta}(x) = \frac{1}{\pi}\sum_{k=1}^{n+1} \frac{a_k \cos kx + b_k \sin kx}{1 + \lambda k^4},$$

where a_k, b_k and λ are defined by (5) and (18).

REFERENCES

1. BAHVALOV N.: Numerical methods. Nauka, Moskva 1973.

2. DOLGOPOLOVA T., IVANOV V.: O čislennom differencirovanii. Ž. vičesl. matem. i matem. fiz. 3 (1966),570-576.

3. MOROZOV A.: O zadače diferencirovania i nektorih algoritmah približenia eksperimentalnoj informacii. Vičisl. metodii i programirovanie, MGU 1970, 46-62.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON AN APPLICATION OF HERMITE'S INTERPOLATION POLYNOMIAL AND SOME RELATED RESULTS

GRADIMIR V. MILOVANOVIĆ AND JOSIP E. PEČARIĆ

ABSTRACT:
In this paper we gave generalizations and improvements of integral inequalities from [1] *and* [2]. *In the proof we used the well-known result for the error of Hermite's interpolation polynomial. Some similar results are also given.*

O JEDNOJ PRIMENI HERMITEOVOG INTERPOLACIONOG POLINOMA I NEKIM SRODNIM REZULTATIMA. *U radu su date generalizacije i poboljšanja integralnih nejednakosti iz* [1] *i* [2]. *U dokazu je korišćen poznati rezultat za grešku Hermiteovog interpolacionog polinoma. Neki slični rezultati su takodje dati.*

1. INTRODUCTION

In the journal Amer. Math. Monthly the following two problems ([1],[2]) are posed:

1^o Suppose $f(x)$ has a continuous $(2m)$-th derivative on $a \leqq x \leqq b$, that $|f^{(2m)}(x)| \leqq M$, and that $f^{(r)}(a) = f^{(r)}(b) = 0$ for $r=0,1,\dots,m-1$. Show that

$$(1) \qquad \left|\int_a^b f(x)dx\right| \leqq \frac{(m!)^2 M}{(2m)!(2m+1)!}(b-a)^{2m+1}.$$

2^o Let $f:[a,b] \to R$ be a continuous function which is twice differentiable in (a,b) and satisfies $f(a) = f(b) = 0$. Prove that

$$(2) \qquad \int_a^b |f(x)|dx \leqq \frac{1}{12} M(b-a)^3,$$

where $M = \sup|f''(x)|$ for $x \in (a,b)$.

The solution of first problem is given in [3].

The inequalities (1) and (2) are related to IYENGAR's inequality [4, pp. 297-298][1].

In this paper we shall prove some inequalities which generalize (1) and (2) in many senses.

Let us define the two-parameter clas of polynomials $P_n^{(m,k)}$ $(0 \leq m \leq k < n;\ m,k,n \in N)$ by means of

$$P_n^{(m,k)}(x) \equiv P_n^{(m,k)}(x;a,b) =$$

$$= C_n^{(m,k)}(a,b)\ (x-a)^m \int_b^x (x-a)^{k-m}(x-b)^{n-k-1}dx$$

where a and b are real parametars and

$$C_n^{(m,k)}(a,b) = \frac{(-1)^{n-k}(n-m)!}{m!(k-m)!(n-k-1)!}(b-a)^{m-n}.$$

For this polynomials the following relations hold:

$$\frac{d^i}{dx^i}P_n^{(m,k)}(x)\Big|_{x=a} = \delta_{im} \quad (i=0,1,\ldots,k;\ \delta_{im} \text{ is the CRONECKER symbol}),$$

$$\frac{d^i}{dx^i}P_n^{(m,k)}(x)\Big|_{x=b} = 0 \quad (i=0,1,\ldots,n-k-1),$$

$$P_n^{(m,k)}(x) = C_n^{(m,k)}(a,b) \sum_{i=0}^{k-m} \frac{(b-a)^i}{n-m-i}\binom{k-m}{i}(x-a)^m(x-b)^{n-m-i}$$

$$\int_a^b P_n^{(m,k)}(x)dx = \frac{(n-m)!}{(n+1)!}\binom{k+1}{m+1}(b-a)^{m+1}.$$

If the values of derivatives of function f in $x=a$ and $x=b$ are known, using polynomials $P_n^{(m,k)}$, HERMITE's interpolation polynomial can be represented in the following form:

$$S_{n,k}(x) = \sum_{m=0}^{k-1} P_{n-1}^{(m,k-1)}(x;a,b)f^{(m)}(a) + \\ + \sum_{m=0}^{n-k-1} P_{n-1}^{(m,n-k-1)}(x;b,a)f^{(m)}(b).$$

[1] On some generalizations IYENGAR's inequality see [5-7].

2. MAIN RESULT

We use the following notation

$$M^{[r]}(f;p) = \left(\frac{\int_a^b p(x)\,|f(x)|^r dx}{\int_a^b p(x)dx} \right)^{1/r}, \quad g(x) = f(x) - S_{n,k}(x).$$

THEOREM 1. Let $x \mapsto f(x)$ ne a n-times differentiable function such that $|f^{(n)}(x)| \leqq M \quad (\forall x \in (a,b))$. If $x \mapsto p(x)$ is an integrable function on (a,b) such that

$$0 < c \leqq p(x) \leqq \lambda c \quad (\lambda \geqq 1,\ x \in [a,b]),$$

the following inequality

$$(3) \quad M^{[r]}(g;p) \leqq \frac{MC(b-a)^n}{n!} \left(\frac{\lambda B(rk+1, r(n-k)+1)}{C^r + (\lambda-1)B(rk+1, r(n-k)+1)} \right)^{1/r} \quad (r>0)$$

holds, where B is beta function and $C = k^k(n-k)^{n-k}/n^n$.

Proof. Since $|f^{(n)}(x)| \leqq M$, the inequality

$$\left| f(x) - S_{n,k}(x) \right| \leqq \frac{M}{n!} \left|(x-a)^k (x-b)^{n-k}\right|$$

is valid, wherefrom (for $r > 0$)

$$(4) \qquad M^{[r]}(g;p) = \frac{M}{n!} \left(\frac{\int_a^b p(x)(x-a)^{rk}(b-x)^{r(n-k)}dx}{\int_a^b p(x)dx} \right)^{1/r}.$$

According to J. KARAMATA's inequality [8] (see also [5]) we have

$$\frac{\int_a^b p(x)(x-a)^{rk}(b-x)^{r(n-k)}dx}{\int_a^b p(x)dx} \leqq \frac{\lambda N \mu}{N + (\lambda-1)\mu},$$

where

$$N = C^r(b-a)^{nr} \text{ and } \mu = (b-a)^{nr} B(rk+1, r(n-k)+1),$$

which combined with (4) gives (3).

From Theorem 1, we directly get the following result:

COROLLARY 1. Let $x \mapsto f(x)$ be a n-times differentiable function such that $|f^{(n)}(x)| \leqq M$ $(\forall x \in (a,b))$ and let $f^{(i)}(a) = 0$ $(i = 0,1,\ldots,k-1)$ and $f^{(i)}(b) = 0$ $(i = 0,1,\ldots,n-k-1)$. Then

$$\left(\frac{1}{b-a}\int_a^b |f(x)|^r dx\right)^{1/r} \leqq \frac{M(b-a)^n}{n!} B(rk+1, r(n-k)+1)^{1/r} \quad (r>0). \tag{5}$$

For $n = 2m$, $k = m$, $r = 1$, inequality (5) reduces to

$$\int_a^b |f(x)| dx \leqq \frac{M(b-a)^{2m+1}(m!)^2}{(2m)!(2m+1)!} \tag{6}$$

which generalize (2), and which is evidently stronger than the inequality (1).

COROLLARY 2. Let function $x \mapsto f(x)$ satisfy the conditions as in Corollary 1. If $x \mapsto p(x)$ is arbitrary nonnegative function, then

$$M^{[r]}(f;p) \leqq \frac{Mk^k(n-k)^{n-k}}{n!n^n}(b-a)^n \quad (r > 0). \tag{7}$$

REMARK 1. Corollary 2 can be formally obtained from Theorem 1 putting $\lambda \to +\infty$. Using N.ÅSLUND's result ([9]), the inequality (7) can be substituted by a somewhat simpler but weaker inequality

$$M^{[r]}(f;p) \leqq \frac{M}{n!}\binom{n}{k}^{-1}(b-a)^n \quad (r > 0).$$

3. SOME SIMILAR RESULTS

According to the results from the previous section and the inequality $\left|\int_a^b h(x)dx\right| \leqq \int_a^b |h(x)| dx$, we obtain the following inequality

$$\left|\int_a^b f(x)dx - \sum_{k=1}^{m} \frac{(2m-k)!}{(2m)!}\binom{m}{k}(b-a)^k\left(f^{(k-1)}(a) - (-1)^k f^{(k-1)}(b)\right)\right| \leqq \frac{M(m!)^2(b-a)^{2m+1}}{(2m)!(2m+1)!}. \tag{8}$$

REMARK 2. If $f^{(k-1)}(a) = (-1)^k f^{(k-1)}(b)$ $(k=1,\ldots,m)$, inequality (8) reduces to (1).

THEOREM 2. Let $I_n = \{0,1,\ldots,n\}$ and let $\{P_k\}_{k\in I_n}$ be a harmonic sequence of polynomials on $[0,1]$ $(P_n'(x) = P_{n-1}(x))$. If $x\mapsto f(x)$ is n-times differentiable function such that $|f^{(n)}(x)| \leqq M$ $(\forall x\in(a,b))$, then

$$\left|P_0\int_a^b f(x)dx - \sum_{k=1}^{n}(-1)^k(b-a)^k(P_k(0)f^{(k-1)}(a)-P_k(1)f^{(k-1)}(b))\right| \tag{9}$$
$$\leqq M(b-a)^{n+1}\int_0^1 |P_n(t)|dt.$$

Proof. If $h(t) = f(a+t(b-a))$ we have $\int_a^b f(x)dx = (b-a)\int_0^1 h(t)dt$, wherefrom, applying integration by parts on the last integral, we obtain

$$\int_0^1 h(t)dt = h(1) - \int_0^1 th'(t)dt. \tag{10}$$

Since $P_1(t) = P_0t + P_1(0)$ $(P_0(t) = P_0)$, equality (10) may be represented in the form

$$P_0\int_0^1 h(t)dt = P_1(1)h(1)-P_1(0)h(0) - \int_0^1 P_2'(t)h'(t)dt.$$

By succesive integration by parts of $\int_0^1 P_2'(t)h'(t)dt$ (n-1)-times, we obtain

$$P_0\int_0^1 h(t)dt = \sum_{k=1}^{n}(-1)^k(P_k(0)h^{(k-1)}(0)-P_k(1)h^{(k-1)}(1)$$
$$+(-1)^n\int_0^1 P_n(t)h^{(n)}(t)dt,$$

from where (9) follows.

COROLLARY 3. Let function $x\mapsto f(x)$ satisfy the conditions as in Theorem 2 and let $f^{(k)}(b) = (-1)^{k-1}f^{(k)}(a)$ $(k=0,\ldots,n-1)$. Then

$$\left|\int_a^b f(x)dx\right| \leq \frac{M(b-a)^{n+1}}{2^n(n+1)!}. \tag{11}$$

To prove this, take $P_n(t) = \frac{1}{n!}(t-1/2)^n$, in Theorem 2.

REMARK 3. The inequality (11) is obtained in [6] with somewhat stricter conditions for f.

REFERENCES

1. ANON: Problem E 2155. Amer. Math. Monthly 76 (1969), 188.
2. ZAIDMAN S.: Problem E 2622. Ibid. 83 (1976), 740-741.
3. TORCHINSKY A.: Solution of the Problem E 2155. Ibid. 76 (1969), 1142-1143.
4. MITRINOVIĆ D.S. (In cooperation with P.M.VASIĆ): Analytic Inequalities. Berlin-Heidelberg-New York, 1970.
5. VASIĆ P.M. and MILOVANOVIĆ G.V.: On an inequality of Iyengar. Univ.Beograd.Publ.Elektroteh.Fak.Ser.Mat.Fiz. No 544-576 (1976), 18-24.
6. MILOVANOVIĆ G.V. and PEČARIĆ J.E.: Some considerations on Iyengar's inequality and some related applications. Ibid. No 544-576 (1976), 166-170.
7. MILOVANOVIĆ G.V.: O nekim funkcionalnim nejednakostima. Ibid. No 599 (1977), 1-59.
8. KARAMATA J.: O prvom stavu srednjih vrednosti odredjenih integrala. Glas srpske kraljevske akademije CLIV Beograd (1933), 119-144.
9. ÅSLUND N.: The fundamental theorems of information theory II. Nordisk. Mat. Tidskr. 9 (1961), 105.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

CLASSIFICATION OF FORMULAS FOR N-DIMENSIONAL POLYNOMIAL INTERPOLATION

Dušan V. Slavić, Milorad J. Stanojević

BSTRACT:

e increased significance of interpolation in complicated computer calcution of the function values for either one or several variables imposes the ed for evaluation of the possibilities and the efficiency of some interpolaon formulas. The paper gives a classification of the formulas for polynoial interpolation in the n-dimensional space (n = 1, 2, 3, ...).

LASIFIKACIJA FORMULA ZA N-DIMENZIONALNU POLINOMSKU INTERPOLACIJU. Povećani značaj interpolacije u komplikovanim kompjuterskim izračunavanjima vrednosti funkcije jedne i više promenljivih nameće potrebu za vrednovanjem mogućnosti i efikasnosti pojedinih interpolacionih formula. U radu je data klasifikacija formula za polinomsku interpolaciju u n-dimenzionalnom prostoru (n = 1, 2, 3, ...).

CLASSIFICATION

The classification is hierarchical with the following order of priorities: according to the number of space dimensions as the most important criterion, according to the related number of nodes for the same number of space dimensions, according to the positions of the nodes and in relation to the algebraic accuracy when the above three conditions are satisfied.

The algebraic accuracy of each formula is expressed by a polynomial, having coefficients with arbitrary values, so that the formula is exact for that polynomial (not only approximate).

Known formulas are transformed in this paper in order to enable the classification.

In applications, the cases with symmetric node positions are of special interest. In the notation used in this paper the subscripts with the function denote node positions.

ONE-DIMENSIONAL INTERPOLATION

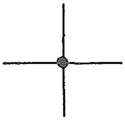
In the case of one-dimensional interpolation the general formula is

$$f(x) \cong \sum_{j=1}^{n} \left(\prod_{\substack{k=1\\k\neq j}}^{n} \frac{x - x_k}{x_j - x_k} \right) f(x_j), \tag{1}$$

usually referred to as Lagrange formula containing the following four known specific formulas, for example in [1].

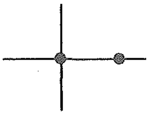
<u>$f(x) = a$</u>

$$f(x) = f_0$$



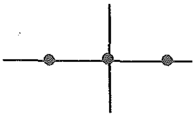
<u>$f(x) = a + bx$</u>

$$f(x) = (1 - x)f_0 + xf_1$$



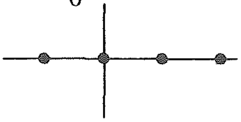
<u>$f(x) = a + bx + cx^2$</u>

$$f(x) = \frac{1}{2}x(x - 1)f_{-1} + (1 - x^2)f_0 + \frac{1}{2}x(x + 1)f_1$$



<u>$f(x) = a + bx + cx^2 + dx^3$</u>

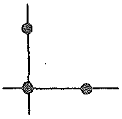
$$f(x) = -\frac{1}{6}x(x - 1)(x - 2)f_{-1} + \frac{1}{2}(x^2 - 1)(x - 2)f_0 - \frac{1}{2}x(x + 1)(x - 2)f_1 + \frac{1}{6}x(x^2 - 1)f_2$$

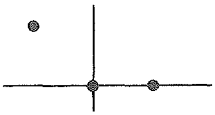


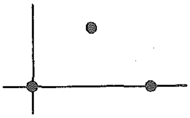
TWO-DIMENSIONAL INTERPOLATION

In the case of the two-dimensional interpolation three-point formulas are presented first. These formulas are function approximation by means o: the plane $z = A + Bx + Cy$.

<u>$f(x,y) = A + Bx + Cy$</u>





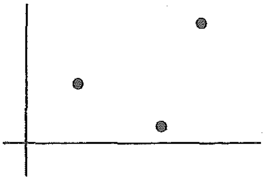


$$f(x,y) = (1 - x - y)f_{0,0} + xf_{1,0} + yf_{0,1}$$

$$f(x,y) = (1 - x - 2y)f_{0,0} + (x + y)f_{1,0} + yf_{-1,1}$$

$$f(x,y) = \frac{1}{2}(1 - x - y)f_{-1,0} + yf_{0,1} + \frac{1}{2}(1 + x - y)f_{1,0}$$

ıhe general three-point formula has been considered, for example, Young and Gregory and it reads:



$$f(x_i,y_i) = f_i$$

2) $$f(x,y) = (\alpha_1 f_1 + \alpha_2 f_2 + \alpha_3 f_3)/d$$

$$\alpha_1 = (x - x_2)(y - y_3) - (x - x_3)(y - y_2)$$

$$\alpha_2 = (x - x_3)(y - y_1) - (x - x_1)(y - y_3)$$

$$\alpha_3 = (x - x_1)(y - y_2) - (x - x_2)(y - y_1)$$

$$d = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

The formula (2) is equivalent to the following formula given by erezin - Žitkov (r and R are vectors)

$$r_k = (x - x_k)\vec{i} + (y - y_k)\vec{j}$$

$$r_{k\ell} = (x_k - x_\ell)\vec{i} + (y_k - y_\ell)\vec{j}$$

$$R_{k\ell} = (y_k - y_\ell)\vec{i} - (x_k - x_\ell)\vec{j}$$

$$f(x,y) = \frac{(r_2R_{23})}{(r_{12}R_{23})} f(x_1,y_1) + \frac{(r_3R_{31})}{(r_{23}R_{31})} f(x_2,y_2) + \frac{(r_1R_{12})}{(r_{31}R_{12})} f(x_3,y_3).$$

This formula may be written in a shorter way, as follows

(3) $$f(x,y) = \sum_3 \frac{(r_2R_{23})}{(r_{12}R_{23})} f(x_1,y_1),$$

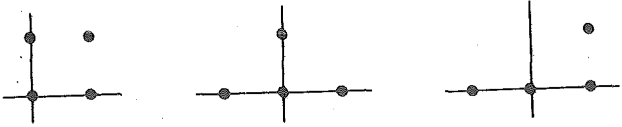
where 3 represents the number of cyclic permutation.

Four-point formulas which include the term with xy are presented here

$\underline{f(x,y) = A + Bx + Cy + Exy}$

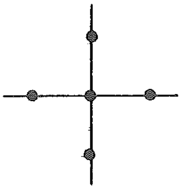
$$f(x,y) = (1 - x)(1 - y)f_{0,0} + x(1 - y)f_{1,0} + (1 - x)yf_{0,1} + xyf_{1,1}$$

$$f(x,y) = \frac{1}{2}x(x - 1)f_{-1,0} + (1 - x^2 - y)f_{0,0} + \frac{1}{2}x(x + 1)f_{1,0} + yf_{0,1}$$



$$f(x,y) = \frac{1}{2}x(x-1)f_{-1,0} + (1-x^2)f_{0,0} + (\frac{1}{2}x(x+1) - y)f_{1,0} + yf_{1,1}.$$

The following five-point formula instead of the term with xy contains the terms with x^2 and y^2

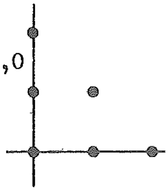


$\underline{f(x,y) = A + Bx + Cy + Dx^2 + Fy^2}$

$$f(x,y) = (1 - x^2 - y^2)f_{0,0} + \frac{1}{2}x(x+1)f_{1,0} + \frac{1}{2}y(y+1)f_{0,1} + \frac{1}{2}x(x-1)f_{-1,0} + \frac{1}{2}y(y-1)f_{0,-1}.$$

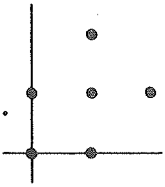
The following six-point formulas contain all the terms mentioned earlier

$\underline{f(x,y) = A + Bx + Cy + Dx^2 + Exy + Fy^2}$



$$f(x,y) = \frac{1}{2}(1-x-y)(2-x-y)f_{0,0} + x(2-x-y)f_{1,0} + y(2-x-y)f_{0,1} + \frac{1}{2}x(x-1)f_{2,0} + xyf_{1,1} + \frac{1}{2}y(y-1)f_{0,2}$$

$$f(x,y) = (1-x)(1-y)f_{0,0} + (x - \frac{1}{2}y)(1-y)f_{1,0} + (y - \frac{1}{2}x)(1-x)f_{0,1} + \frac{1}{2}x(x-1)f_{2,1} + (x(1-x) + y(1-y) + xy)f_{1,1} + \frac{1}{2}y(y-1)f_{1,2}.$$



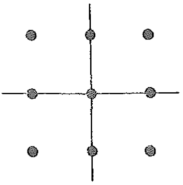
The most general six-point formula has been considered by Berezin and Žitkov, who present a complicated formula [2].

This formula can be presented in a simpler way by the notation of the formula (3), as follows

$$f(x,y) = \sum_{6} \frac{\begin{vmatrix} (r_2R_{23})(r_4R_{45}) & (r_{62}R_{23})(r_{64}R_{45}) \\ (r_2R_{24})(r_3R_{35}) & (r_{62}R_{24})(r_{63}R_{35}) \end{vmatrix}}{\begin{vmatrix} (r_{12}R_{23})(r_{14}R_{45}) & (r_{62}R_{23})(r_{64}R_{45}) \\ (r_{12}R_{24})(r_{13}R_{35}) & (r_{62}R_{24})(r_{63}R_{35}) \end{vmatrix}} f(x_1,y_1).$$

The following nine-point formula has the same order of accuracy.

$$\begin{aligned} f(x,y) &= (1 - x^2)(1 - y^2)f_{0,0} + \frac{1}{2}x(1 + x - xy^2)f_{1,0} \\ &+ \frac{1}{2}y(1 + y - x^2y)f_{0,1} - \frac{1}{2}x(1 - x + xy^2)f_{-1,0} \\ &- \frac{1}{2}y(1 - y + x^2y)f_{0,-1} + \frac{1}{4}xy(1 + xy)f_{1,1} \\ &- \frac{1}{4}xy(1 - xy)f_{1,-1} + \frac{1}{4}xy(1 - xy)f_{-1,-1} + \frac{1}{4}xy(1+xy)f_{-1,1}. \end{aligned}$$



J. H. Lambert [4] gave the formula which can be concisely written as

$$(4) \qquad f(x,y) = \sum_{m=0}^{n} \sum_{k=0}^{m} \binom{x}{m-k}\binom{y}{k} \Delta_{m-k,k}f_{0,0},$$

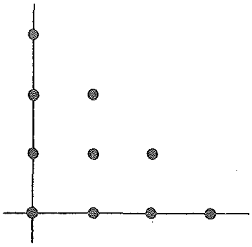
where:

$$\begin{aligned} \Delta_{0,0}f_{p,q} &= f_{p,q} \\ \Delta_{1,0}f_{p,q} &= f_{p+1,q} - f_{p,q} \\ \Delta_{0,1}f_{p,q} &= f_{p,q+1} - f_{p,q} \\ \Delta_{n+1,m}f_{p,q} &= \Delta_{n,m}\Delta_{1,0}f_{p,q} \\ \Delta_{n,m+1}f_{p,q} &= \Delta_{n,m}\Delta_{0,1}f_{p,q} . \end{aligned}$$

A specific case of the ten-point formula (4) is

$$\underline{f(x,y) = A + Bx + Cy + Dx^2 + Exy + Fy^2 + Gx^3 + Hx^2y + Ixy^2 + Jy^3}$$

$$\begin{aligned} f(x,y) &= \frac{1}{6}(1 - x - y)(2 - x - y)(3 - x - y)f_{0,0} \\ &+ \frac{1}{2}x(2 - x - y)(3 - x - y)f_{1,0} \\ &+ \frac{1}{2}y(2 - x - y)(3 - x - y)f_{0,1} \\ &+ \frac{1}{2}y(y - 1)(3 - x - y)f_{2,0} \\ &+ xy(3 - x - y)f_{1,1} + \frac{1}{2}x(x - 1)(3 - x - y)f_{0,2} \\ &+ \frac{1}{6}x(x - 1)(x - 2)f_{3,0} + \frac{1}{2}xy(x - 1)f_{2,1} \\ &+ \frac{1}{2}xy(y - 1)f_{1,2} + \frac{1}{6}y(y - 1)(y - 2)f_{0,3} . \end{aligned}$$

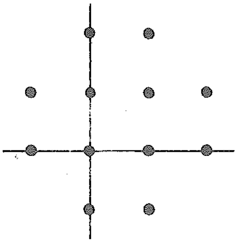


F. B. Hildebrand gives the following formula

$$\underline{f(x,y) = A + Bx + Cy + Dx^2 + Exy + Fy^2 + Gx^3}$$

$$\underline{+ Hx^2y + Ixy^2 + Jy^3 + Lx^3y + Nxy^3}$$

$$\begin{aligned} f(x,y) = {} & \frac{1}{2}(1 - x)(1 - y)\left[2 + x(1 - x) + y(1 - y)\right] f_{0,0} \\ & + \frac{1}{2}x(x - y)\left[2 + x(1 - x) + y(1 - y)\right] f_{1,0} \\ & + \frac{1}{2}xy\left[2 + x(1 - x) + y(1 - y)\right] f_{1,1} \\ & + \frac{1}{2}(1 - x)y\left[2 + x(1 - x) + y(1 - y)\right] f_{0,1} \\ & + \frac{1}{6}x(x - 1)(x - 2)(y - 1)f_{-1,0} \\ & + \frac{1}{6}(x - 1)y(y - 1)(y - 2)f_{0,-1} \\ & + \frac{1}{6}xy(y - 1)(2 - y)f_{1,-1} \\ & + \frac{1}{6}x(1 - x^2)(y - 1)f_{2,0} \\ & + \frac{1}{6}x(x^2 - 1)yf_{2,1} + \frac{1}{6}xy(y^2 - 1)f_{1,2} \\ & + \frac{1}{6}(x - 1)y(1 - y^2)f_{0,2} + \frac{1}{6}x(x - 1)(2 - x)yf_{-1,1}. \end{aligned}$$

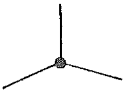


THREE-DIMENSIONAL INTERPOLATION

For interpolation in the three-dimensional space the following formulas are given

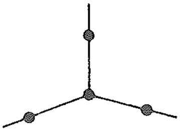
$$\underline{f(x,y,z) = \alpha}$$

$$f(x,y,z) = f_{0,0,0}$$



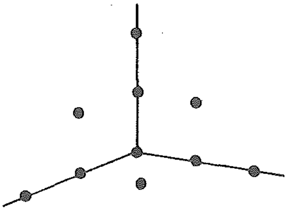
$$\underline{f(x,y,z) = \alpha + \beta x + \gamma y + \delta z}$$

$$f(x,y,z) = (1 - x - y - z)f_{0,0,0} + xf_{1,0,0} + yf_{0,1,0} + zf_{0,0,1}$$



$$\underline{f(x,y,z)=\alpha+\beta x+\gamma y+\delta z+\varepsilon x^{2}+\zeta xy+\eta y^{2}+\theta yz+\iota z^{2}+\kappa xz}$$

$$\begin{aligned}f(x,y,z)=&\frac{1}{2}(1-x-y-z)(2-x-y-z)f_{0,0,0}\\&+x(2-x-y-z)f_{1,0,0}\\&+y(2-x-y-z)f_{0,1,0}\\&+z(2-x-y-z)f_{0,0,1}\\&+\frac{1}{2}x(x-1)f_{2,0,0}+xyf_{1,1,0}\\&+\frac{1}{2}y(y-1)f_{0,2,0}+yzf_{0,1,1}\\&+\frac{1}{2}z(z-1)f_{0,0,2}+xzf_{1,0,1}.\end{aligned}$$



N-DIMENSIONAL INTERPOLATION

Suppose that the four-dimensional-space points are provided

$$f(a_i,\ b_j,\ c_k,\ d_\ell),$$

where:

$$i=1(1)n_a,\ \ j=1(1)n_b,\ \ k=1(1)n_c,\ \ \ell=1(1)n_d.$$

The aim of interpolation is the calculation of the function value $f(a,b,c,d)$.

By intersecting the hyperplanes the following formula is obtained

$$f(a_i,b_j,c_k,d)\simeq\sum_{\ell=1}^{n_d}\left(\prod_{\substack{m=1\\m\neq\ell}}^{n_d}\frac{d-d_m}{d_\ell-d_m}\right)f(a_i,b_j,c_k,d_\ell),$$

$$i=1(1)n_a,\ \ j=1(1)n_b,\ \ k=1(1)n_c.$$

By intersecting the planes the following formula is obtained

$$f(a_i,b_j,c,d)\simeq\sum_{k=1}^{n_c}\left(\prod_{\substack{m=1\\m\neq k}}^{n_c}\frac{c-c_m}{c_k-c_m}\right)f(a_i,b_j,c_k,d)$$

$$i=1(1)n_a,\ \ j=1(1)n_b.$$

By intersecting straight lines the following formula is obtained

$$f(a_i,b,c,d) \simeq \sum_{j=1}^{n_b} \left(\prod_{\substack{m=1\\ m\neq j}}^{n_b} \frac{b - b_m}{b_j - b_m} \right) f(a_i,b_j,c,d), \qquad i=1(1)n_a.$$

By applying the mentioned Lagrange formula (1) it can be obtained

$$f(a,b,c,d) \simeq \sum_{i=1}^{n_a} \left(\prod_{\substack{m=1\\ m\neq i}}^{n_a} \frac{a - a_m}{a_i - a_m} \right) f(a_i,b,c,d).$$

For the dimension number greater than four, the beginning of the procedure is analogous, and the end is the same as given in the above algorithm.

When dealing with the interpolation having a larger number of points in the n-dimensional space the simplicity of notation is important.

A more extensive work on formulas with a larger number of points is expected in future. The aim of this paper is not to deal with the formulas based on a larger number of points, because the application of these formulas is smaller due to larger computation procedures. Due to the limited length of the paper some more complex, but useful formulas, are not included

For this reason the paper contains only the most important formulas for computer calculation.

G. Alikalfić, A. Djordjević, A. Fišer-Popović, D. T. Jovanović, D. S. Mitrinović have read this paper in manuscript and have made some valuable remarks and suggestions.

REFERENCES

1. ABRAMOWITZ M., STEGUN I. A.: Handbook of Mathematical Function, National Bureau of Standars, New York 1964.
2. BEREZIN I. S., ŽITKOV N. P.: Numerička analiza (Numeričke metode), Naučna knjiga, Beograd 1963.
3. BUCKINGHAM R. A.: Numerical Methods, Pitman and Sons, London 1962.
4. DAVIS H. T.: Tables of the Higher Mathematical Function, The Principia Press, Bloomington, Indiana 1933.
5. HILDEBRAND F. B.: Interpolation to Numerical Analysis, McGraw-Hill, New York 1974.
6. STEFFENSEN J. F.: Interpolation, Chelsea Pub. Company, New York 1950.
7. YOUNG D. M., GREGORY R. T.: A Survey of Numerical Mathematics, Vol. I, Addison - Wesley Series in Mathematics, Reading 1972.

merical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON APPROXIMATIONS OF SOLUTIONS OF SECOND ORDER LINEAR DIFFERENTIAL EQUATIONS

Božo Vrdoljak

ABSTRACT:

The paper deals with second order linear differential equation with functional coefficients. For the corresponding sufficient conditions we obtain results on approximation of certain classes of Cauchy's solutions and on behaviour and stability of all solutions. The results were obtained by transforming the second order linear equation to a respective linear system of equations and by studying solutions of that system with respect to the "circular" neighbourhood of an integral curve. The obtained results are generalized to a quasi-linear equation as well.

O APROKSIMACIJAMA RJEŠENJA LINEARNE DIFERENCIJALNE JEDNADŽBE DRUGOG REDA. U radu se proučava linearna diferencijalna jednadžba drugog reda s funkcionalnim koeficijentima. Uz odgovarajuće dovoljne uvjete dobivaju se rezultati o aproksimaciji određenih klasa Cauchyevih rješenja, kao i o ponašanju i stabilnosti svih rješenja. Do rezultata se dolazi transformiranjem linearne jednadžbe drugog reda na odgovarajući linearan sistem jednadžbi i promatranjem rješenja tog sistema u odnosu na odgovarajuću "kružnu" okolinu neke integralne krivulje. Dobiveni rezultati se poopćuju i na kvazilinearnu jednadžbu.

Let us consider the equation

(1) $$y''+p(t)y'+q(t)y=f(t),$$

where $p,q,f\in C(I)$, $I=\langle \underline{t},\infty\rangle$. Let $y=\Psi(t)$, $\Psi\in C^1(I)$ be an arbitrary solution of equation (1). We shall use functions $\beta,\rho\in C^1(I)$, $\rho(t)>0$ on I and notations $p_o=p(t_o)$, $\beta_o=\beta(t_o)$, $\rho_o=\rho(t_o)$, $\Psi_o=\Psi(t_o)$, $y_o=y(t_o)$, $y_o'=y'(t_o)$.

THEOREM 1. Let us take functions β and ρ such that

(2) $$(\beta'+\beta^2+\beta p+q-1)^2<4(-\beta-p-\rho'/\rho)(\beta-\rho'/\rho) \text{ on } I.$$

(a) If

(3) $$\beta-\rho'/\rho<0 \text{ on } I,$$

then all solutions $y=y(t)$ of equation (1) satisfying initial condition

(4) $$(y_o-\Psi_o)^2+(y_o'-\beta_o y_o)^2\leqslant\rho_o^2\ ,\quad t_o\in I$$

satisfy also condition

(5) $$|y(t)-\Psi(t)|<\rho(t) \quad \text{for every} \quad t>t_o .$$

(b) If

(6) $$\beta-\rho'/\rho>0 \quad \text{on } I,$$

then problem (1)-(4) has at least one solution satisfying condition (5).

Proof. For equation (1) let us introduce the substitute

(7) $$y'=x+\beta(t)\,y ,$$

where $x=x(t)$ is a new unknown function. Equation (1) is transformed to the system of equations

(8) $$\begin{aligned} x' &= -(\beta+p)\,x-(\beta'+\beta^2+\beta p+q)y+f \\ y' &= x+\beta y . \end{aligned}$$

Let $K=\{(x,y,t):\ x=\varphi(t),\ y=\Psi(t),\ t\in I\}$, where $\varphi\in C^1(I)$, $\varphi(t_o)=0$ is an integral curve of system (8). Let $\Omega=R^2\times I$ and

$$\omega=\left\{(x,y,t)\in\Omega \ :\ \rho^{-2}(t)\left[(x-\varphi(t))^2+(y-\Psi(t))^2\right]<1\right\}$$

be open sets. Let $\tau(t)$ be a tangential vector or the integral curve $(x(t), y(t), t)$ of system (8) in points of surface $\partial\omega$ $(\partial\omega=C\ell\,\omega\setminus\omega)$, and let $\nu(t)$ be vector of external normal on surface $\partial\omega$, i.e.

$$\begin{aligned} \tau(t)=(x'(t),y'(t),1),\ \ \nu(t)=((x-\varphi)\rho^{-2},(y-\Psi)\rho^{-2}, \\ -\left[(x-\varphi)^2\rho'+(x-\varphi)\rho\varphi'+(y-\Psi)^2\rho'+(y-\Psi)\rho\Psi'\right]\rho^{-3}). \end{aligned}$$

Let us consider now the scalar product $P(t)=(\tau|\nu)$ in the points of surface $\partial\omega$. We have

(9) $$\begin{aligned} P(t)=(-\beta-p-\rho'/\rho)(x-\varphi)^2\rho^{-2}+(\beta-\rho'/\rho)(y-\Psi)^2\rho^{-2}+ \\ +(1-\beta'-\beta^2-\beta p-q)(x-\varphi)\rho^{-1}(y-\Psi)\rho^{-1} . \end{aligned}$$

Let us note that $P(t)$ is a quadratic symmetric form

$$P(t)=a_{11}X^2+2a_{12}XY+a_{22}Y^2 ,$$

where

$$a_{11}=-\beta-p-\rho'/\rho, \quad a_{12}=a_{21}=(1-\beta'-\beta^2-\beta p-q)/2 , \quad a_{22}=\beta-\rho'/\rho,$$

(10) $$X=(x-\varphi)\rho^{-1} \ , \quad Y=(y-\Psi)\rho^{-1} .$$

Moreover, it is sufficient to note the following.

(a) Conditions (2) and (3) grant conditions $-a_{11}>0$, $a_{11}a_{22}-a_{12}^2>0$ on I, and according to Sylvester's criterion it follows that $P(t)<0$ on I. Relation $P(t)<0$ means that set $\partial\omega$ is a set of points of strict entrance for integral curves of system (8) with respect to sets ω and Ω . Consequently, all solutions of system (8) satisfying initial condition

$$x_o^2 + (y_o - \Psi_o)^2 \leqslant \rho_o^2 , \quad t_o \in I$$

$(x_o = x(t_o))$ satisfy also condition

$$(x(t) - \varphi(t))^2 + (y(t) - \Psi(t))^2 < \rho^2(t) \quad \text{for every} \quad t > t_o .$$

Since, in view of (7), $x_o = y_o' - \beta_o y_o$ all solutions of equation (1) satisfying initial condition (4) satisfy also condition (5).

(b) Conditions (2) and (6) grant conditions $a_{11} > 0$, $a_{11}a_{22} - a_{12}^2 > 0$ on I, and it follows that $P(t) > 0$ on I. Thus $\partial \omega$ is a set of points of strict exit of integral curves of system (8) with respect to sets ω and Ω . Hence, according to retraction method ([4]), there exists at least one integral curve of system (8) which belongs to set ω for every $t \in I$. Consequently, problem (1)-(4) has at least one solution satisfying condition (5).

Let us note that conditions of Theorem 1 are simplified if functions β and ρ are taken in a special form.

COROLLARY 1 $(\rho(t) \equiv r)$. (a) $(\beta(t) \equiv -1)$ If

$$p > 1, \quad 0 < p - 2\sqrt{p-1} < q < p + 2\sqrt{p-1} \quad \text{on } I,$$

then all solutions of equation (1) satisfying initial condition

$$(y_o - \Psi_o)^2 + (y_o' + y_o)^2 \leqslant r^2 ,$$

where r is a positive constant, satisfy also condition

$$|y(t) - \Psi(t)| < r \quad \text{for every} \quad t > t_o \tag{11}$$

(b) $(\beta(t) \equiv 1)$ If

$$p < -1, \quad 0 < -p - 2\sqrt{-p-1} < q < -p + 2\sqrt{-p-1} \quad \text{on } I ,$$

then at least one of solutions of equation (1) which satisfy initial condition

$$(y_o - \Psi_o)^2 + (y_o' - y_o)^2 \leqslant r^2$$

satisfies also condition (11).

COROLLARY 2 $(\beta(t) \equiv 0)$. Let

$$(q-1)^2 < 4(p + \rho'/\rho)\rho'/\rho \text{ on } I.$$

(a) If $\rho' > 0$ on I, then all solutions of equation (1) which satisfy initial condition

$$(y_o - \Psi_o)^2 + y_o'^2 < \rho_o^2 , \quad t_o \in I \tag{12}$$

satisfy also condition (5).

(b) If $\rho' < 0$ on I, then problem (1)-(12) has at least one solution satisfying condition (5).

COROLLARY 3 $(\beta(t) \equiv -p(t))$. Let

$$(q - p' - 1)^2 < 4(p + \rho'/\rho)\rho'/\rho \quad \text{on } I.$$

(a) If $\rho' > 0$ on I, then all solutions of equation (1) satisfying initial condition

$$(y_0 - \Psi_0)^2 + (y_0' + p_0 y_0)^2 \leqslant \rho_0^2 \tag{13}$$

satisfy also condition (5).

(b) If $\rho' > 0$ on I, then problem (1)-(13) has at least one solution satisfying condition (5).

THEOREM 2. (a) If there exist functions β and ρ such that

$$2\beta + p \leqslant 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2\cdot(\beta + p + \rho'/\rho) \tag{14}$$

or

$$2\beta + p \geqslant 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2\cdot(-\beta + \rho'/\rho) \tag{15}$$

on I, then statement (a) of Theorem 1 holds true.

(b) If there exist functions β and ρ such that

$$2\beta + p \geqslant 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2(-\beta - p - \rho'/\rho)$$

or

$$2\beta + p \leqslant 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2(\beta - \rho'/\rho)$$

on I, then statement (b) of Theorem 1 holds true.

Proof. Let us use here the first part of the proof of Theorem 1 until the formation of the scalar product $P(t)$ according to formula (9). We shall also use notations (10). It is sufficient to note that the following estimates for $P(t)$ hold true.

(a) Since $ab \leqslant (a^2 + b^2)/2$ for every $a, b \in R$, on $\partial\omega$ it is valid

$$P(t) \leqslant (-\beta - p - \rho'/\rho)X^2 + (\beta - \rho'/\rho)Y^2 + |1 - \beta' - \beta^2 - \beta p - q| \cdot (X^2 + Y^2)/2 \equiv \bar{P}(t).$$

In view of (14) on $\partial\omega$ it is valid

$$\bar{P}(t) = (-\beta - p - \rho'/\rho + |1 - \beta' - \beta^2 - \beta p - q|/2)(X^2 + Y^2) + (2\beta + p)Y^2 =$$
$$= (-\beta - p - \rho'/\rho + |1 - \beta' - \beta^2 - \beta p - q|/2) + (2\beta + p)Y^2 < 0.$$

Moreover, in view of (15) on $\partial\omega$ it is valid

$$\bar{P}(t) = (\beta - \rho'/\rho + |1 - \beta' - \beta^2 - \beta p - q|/2) - (2\beta + p)X^2 < 0.$$

(b) Here it should be noted that on $\partial\omega$

$$P(t) \geqslant (-\beta - p - \rho'/\rho)X^2 + (\beta - \rho'/\rho)Y^2 - |1 - \beta' - \beta^2 - \beta p - q|(X^2 + Y^2)/2 \equiv \underline{P}(t)$$

nd that $\underline{P}(t) > 0$.

Conditions of Theorem 2 are simplified if functions β and ρ are taken in a special form. For example $\beta(t) \equiv 0, -1, 1$; $\rho(t) \equiv \rho_0 e^{-s(t-t_0)}$, $s \in R$.

Using the obtained results and the known properties valid for the linear differential equation, we can draw the following conclusions related to the questions of stability and approximation of solutions of equation (1).

1. If function ρ is bounded on I, then in cases (a) of all given statements we have stability of all solutions of equation (1) with the function of stability ρ ([5]). If $\rho(t) \to 0$, $t \to \infty$, we have asymptotic stability of all solutions with the function of stability ρ.

2. If $\rho(t) \to \infty$, $t \to \infty$, then in cases (b) of all given statements we have instability of all solutions of equation (1) with the function of instability .

3. Considering the approximation of certain classes of solutions the results given in cases (a) with the bounded function ρ are very significant. Approximation is particularly good when $\rho(t) \to 0$, $t \to \infty$. In that case we have precise asymptotic behaviour of certain Cauchy's solutions. Consequently it should be noted that conditions of Theorems 1 and 2 do not change if instead of function $\rho(t)$ we take function $C\rho(t)$, where $C>0$ is an arbitrary constant.

In the case of a homogeneous equation ($f(t) \equiv 0$) it is interesting to consider the behaviour of solutions in the neighbourhood of a trivial solution $y(t) \equiv 0$ (case $\Psi(t) \equiv 0$).

Remark. Statements of Theorems 1 and 2 are completely valid also for a quasi-linear equation

$$y'' + p(y,t)y' + q(y,t)y = f(y,t),$$

where functions p, q and f satisfy the conditions necessary for the existence and uniqueness of solutions on $R \times I$, only if the respective conditions of Theorems 1 and 2 ($p = p(y,t)$, $q = q(y,t)$) hold true on $\partial\omega$.

Example 1. All the solutions of the Bessel's equation

$$t^2 y'' + ty' + (t^2 - \lambda^2)y = 0, \quad \lambda \in R$$

which satisfy the initial condition

$$(y_0 - \Psi_0)^2 + (y_0' + y_0/2t_0)^2 \leqslant \rho_0^2, \quad t_0 \in I,$$

where $\underline{t} > |\lambda^2 - 1/4|/(1-2s)$, $s \in R$, $0<s<1/2$, then also satisfy condition

$$|y(t) - \Psi(t)| < \rho_o(t_o/t)^s \quad \text{for every} \quad t > t_o.$$

For example, for $\lambda = 1/2$, $\Psi(t) \equiv 0$ all solutions of the Bessel`s equation satisfying initial conditions $y_o = 0$, $|y_o'| \leqslant \rho_o$, $t_o > 0$ satisfy also condition $|y(t)| < \rho_o(t_o/t)^s$ for every $t > t_o$. Let us note that for $\lambda = 1/2$ Bessel`s equation has a general solution $y = (C_1 \cos t + C_2 \sin t)/\sqrt{t}$.

The proof of this result follows from Theorem 1 when $\beta(t) \equiv -1/2t$, $\rho(t) \equiv \rho_o(t_o/t)^s$.

Example 2. Let us take equation

$$y'' + p(t)(y' + y) + q(t)y = f(t),$$

where $p(t) \geqslant 2$, $|q(t)| < 2(1-s)$, $s \in R$, $0 \leqslant s < 1$, on I, $\underline{t} > 0$. All solutions of this equation which sastify the initial condition

$$(y_o - \Psi_o)^2 + (y_o' + y_o)^2 \leqslant \rho_o^2, \quad t_o \in I$$

also satisfy condition

$$|y(t) - \Psi(t)| < \rho_o e^{-s(t-t_o)} \quad \text{for every} \quad t > t_o.$$

It is interesting to consider the case $f(t) \equiv 0$, $\Psi(t) \equiv 0$.

This result follows from Theorem 2 when $\beta(t) \equiv -1$, $\rho(t) \equiv \rho_o e^{-s(t-t_o)}$.

REFERENCES

1. HATVANI L. : On the asymptotic behaviour of the solution of $(p(t)x')' + q(t)f(x) = 0$. Publicationes Math. Debrecen, 19 (1972), 225-237.

2. NEJMAN F. : Teorija global`nyh svojstv obyknovennyh linejnyh differencial`nyh uravnenij n-go porjadka. Differenc. uravnenija, 19 (1983), 799-808.

3. READ T. T. : Second-order differential equations with small solutions, J. Math. Anal. Appl. 77 (1980), 165-174.

4. VRDOLJAK B. : Curvilinear "tubes" in the retraction method and the behaviour of solutions for the system of differential equations. Mat. vesnik 4 (17) (32) (1980), 381-392.

5. VRDOLJAK B. : On behaviour of solutions of system of linear differential equations (in preparation).

Faculty of Civil Engineering, University of Split,
V.Masleše bb, P.O. Box 389, 58000 Split, Yugoslavia

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON HYPOTHESIS TESTING IN SPLINE REGRESSION

K. Surla, E. Nikolić, Z. Lozanov

ABSTRACT:

Algorithm given in [2] *for determing the number and the position of knots of the spline function is modified according to statistical tests in* [8] *for fitting the cubic spline regression. Several theorems conected with testing continuity of the third derivative are proved.*

O TESTIRANJU HIPOTEZA U SPLAJN REGRESIJI: *Algoritam dat u* [2] *za odredjivanje broja i pozicija čvorova splajna je modifikovan u skladu sa statističkim testovima datim u* [8] *za formiranje kubne splajn regresije. Dokazane su neke teoreme vezane za testiranje hipoteza o neprekidnosti trećeg izvoda.*

1. In modeling, curve fitting or recovering functions contaminated by noise, the problem of determining the minimal number and optimal positions of knots is still open, although some attempts had been made, see [5], [10], [11]. In [2], [3] an algorithm for automatic determination of the number of knots and their positions for fitting a least square spilne of k-th degree is given. The objectives are (i) the given values of dependent variable should be fitted closely enough, (ii) the approximating spline should be smooth enough, in the sence that the discontinuities in its k-th derivative are as small as possible. Also, it is presumed that the data are not contaminated by noise. In fitting spline regression curve to discrete, noisy obsrevations, besides the problem of choice of knots, occurs the problem of their statistical testing. These problems are investigated in [1], [4], [7], [8], [10], [11]. Hypothesis testing in B-spline regression is investigated in [8] and [11]. This paper is an attempt to applicate the results obtained in [2] to B-spline regression, and to mcdify the algorithm in [2] according to the statistical tests in [8].

2. Given the measured function values y_q, at the points x_q, $q=1,\dots,m$, $x_q<x_{q+1}$, coisider the model

$$y = XC + \varepsilon \tag{1}$$

$$X = \begin{bmatrix} B_{-3}(x_1)\dots B_g(x_1) \\ \vdots \qquad\qquad \vdots \\ B_{-3}(x_m)\dots B_g(x_m) \end{bmatrix}, \quad C = \begin{bmatrix} c_{-3} \\ \vdots \\ c_g \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{bmatrix}$$

$\varepsilon \sim N(0,\sigma^2 I)$, $B_i(x)$ are B-splines functions of the third degree on the grid $a = \lambda_0 < \lambda_1 < \ldots < \lambda_g < \lambda_{g+1} = b$, with additonal knots $\lambda_{-3} = \lambda_{-2} = \lambda_{-1} = a$, $b \neq \lambda_{g+2} = \lambda_{g+3} = \lambda_{g+4}$. We suppose that there must be at least one subset of g+4 strictly increasing values x_{q_i}, $(i=-3,\ldots,g)$ such that

(2) $\quad x_{q_i-4} < \lambda_i < x_{q_i}$ (condition Schoenberg and Whitney).

The approximation criterion is as in [2]

Minimize $\sum_{q=1}^{g} \left(\sum_{i=-3}^{g} C_i a_{i,q} \right)^2$, subject to the constraint

(3) $$\sum_{q=1}^{m} \left(y_q - \sum_{i=-3}^{g} C_i B_i(x_q) \right)^2 < S$$

where S is given, nonnegative constant (smoothing factor).

(4) $\quad a_{i,q} = B_i(\lambda_q+0) - B_i(\lambda_q-0)$, $(q=1,\ldots,g)$, $(i=-3,\ldots g)$.

We remark that our spline function $S(x) = \sum_{i=-3}^{g} C_i B_i(x)$ becomes a single polynomial on $[x_q, x_r]$ if $\sum_{i=-3}^{g} C_i a_{i,j} = 0$, $(j=q+1,\ldots,r-1)$, $a \leq x_q, x_r \leq b$.

In [2] it is shown that problem (3) has a solution and that the algorithm given there leads to the number of knots wich

$$F(p) = \sum_{q=1}^{m} \left(y_q - \sum_{i=-3}^{g_k} C_i(p) B_i(x_q) \right)^2 \leq S,$$

where p^{-1} is Lagrange's multiplier of problem (3). Also, the relation between the parameter p and the number of knots is given. So, for $p=\infty$ we get the least square spline, and for p=0 we get the least square polynomial.

3. For testing the continuity of the third derivative at the knot λ_j we test the following hypothesis (see [8]):

$$H_0 \; : \; \sum_{i=-3}^{g} C_i \left(B_i(\lambda_j+0) - B_i(\lambda_j-0) \right) = 0$$

We use the statistics

$$F = \frac{(1'C)^2 (1''(XX')^{-1} 1)^{-1}}{y(I-X(XX')^{-1}X')y} (m-g-4) \quad \text{or} \quad F = \frac{(1'C)^2}{\hat{V}(1'C)}$$

where $\hat{V}(1'C)$ is estimated variance of linear combination $1'C$, $1' = (a_{-3,j}, a_{-2,j}, \ldots, a_{q,j})$. Under null hipothests F has Fisher distribution $F_{1,m-g-4}$. Using the fact that the linear combination which is tested is contrast, (see [8]), it is enough to evaluate g+3 components of vector 1. If hypothesis H_0 is accepted, we shall say that knot λ_j is statisticaly not significant. Statisticaly not significant knots we shall denote by $\bar{\lambda}_j$, $(j=1,\ldots,\bar{g})$.

4. Denote the spline function with knots $\lambda_j (j=0,\ldots,g+1)$ by $S_g(x)$. The idea of our algorihtm is: Determine the least square spline $S_0(x)$ (single polynomial). If the sum of squares of residuals for $S_0(x)$ is less than S, $S_0(x)$ is the solution to our problem. If not, we determine

successive least square splines $S_{g_j}(x)$, $j=1,2,\ldots$ untill we find

$$F_{g_j}=\sum_{q=1}^{m}(y_q-S_{g_j}(x_q))^2 \leqslant S$$

isfied. The additional number of knots Δg_j and their positions in each eration is determined according to algorithm I (see [2]):

$$\Delta g_j=\begin{cases}1 & j=0\\ \min\{\Delta 1,\Delta 2,\max\{1,\Delta 3,\Delta 4\}\}, & j=1,2,\ldots\end{cases} \quad \text{with}$$

$=2\Delta g_{j-1}$, $\Delta 2=m-4-g_j$, $\Delta 3=[\Delta g_{j-1}/2]$, $\Delta 4=[(F_{g_j}-S)\Delta g_{j-1}/|F_{g_{j-1}}-F_{g_j}|]$

e additional knots are then located inside the intervals $[\lambda_i,\lambda_{i+1}]$ with rgest partial sum of squares of residuals. For details see [2].

Let knots λ_j $(j=0,\ldots,g_{j+1})$ be determinied. As the next iteration we:

) Determine spline $S_{g_j}(x)$.

i) Test all knots λ_i $(i=1,\ldots,g_j)$ using the given statistical test in 3.

ii) If $F_{g_j}\leqslant S$, $S_{g_j}(x)$ is the solution to our problem. If not we go to (*iv*)

) The additional knots are determinated according to the algorithm I .

) New set of knots λ_j $(i=0,\ldots,g_{j+1}+1)$ is formed takeing statisticaly significant knots from (*ii*) and additional knots from (*iv*).

(*vi*) Put $g_j=g_{j+1}$ and go to (*i*)

From $\overline{g}_j<g_{j+1}$ and $\overline{g}_j<g_j$ it follows $F_{\overline{g}_j}\geqslant F_{g_{j+1}}$ and $F_{\overline{g}_j}\geqslant F_{g_j}$. Supposing that the sum of squares of residuals will not be significantly changed by substituteing knots λ_i $(i=1,\ldots,g_j)$ by $\overline{\lambda}_i$ $(i=1,\ldots,\overline{g}_j)$ we can conclude that the relation (5) will be satisfied after finite number of iterations. Namely, for maximal number of knots $g=m-4$ we get a interpolating spline, i.e. $F_{m-4}=0$. As the values F_{g_j} and $F_{g_{j+1}}$ are evaluated the relation $F_{g_{j+1}}\not\approx F_{g_j}$ can be checked. If significant deviation occurs, it is possible to take g_j instead $\overline{g}_j$. This algorithm, compared with algorithm in [2] changes the position of knots, and we get a curve with statisticaly significant knots.

5. Before proving several statements, we shall introduce some definitions and notations. Denote by $P_i(x)=a_ix^3+b_ix^2+c_ix+d_i$ the restriction of function $S(x)$ on interval $[\lambda_i,\lambda_{i+1}]$ and put $P=\{P_i(x),\ i=0,\ldots,g\}$. The relations between coeficients of $P_i(x)$ and coeficients of $S(x)$ are given in [9].

DEFINITION 1. Polynomials $P_i(x)$ and $P_{i-1}(x)$ are not statisticaly different at a prescribed level of significance α, if their corresponding coeficients are not statisticaly different at the level α.

DEFINITION 2. Function $\overline{S}(x)$ is statistical equivalent of spline $S(x)$ at the prescribed level of significance α if (*i*) $\overline{S}(x) = Q_i(x) = \overline{a}_i x^3 + \overline{b}_i x^2 + \overline{c}_i x + \overline{d}_i$, $x \in [\mu_i, \mu_{i+1}]$, $M \subseteq \Lambda$, where $M = \{\mu_i,\ i=0,\dots,r+1\}$, $\Lambda = \{\lambda_i,\ i=0,\dots,g+1\}$, and $Q \subseteq P$, where $Q = \{Q_i(x),\ i=0,1,\dots,r\}$.

(*ii*) for fixed values x, $S(x)$ and $\overline{S}(x)$ are evaluated using the polynomials whicw are not statisticaly different at level α.

(*iii*) Discontinuities of function $\overline{S}(x)$, its first and second derivatives at points μ_i are not statisticaly significant at the level α.

Lema 1. If hypothesis H_0 is accepted at the level α, then polynomials $P_{j-1}(x)$ and $P_j(x)$ are not statisticaly different.

Proof. Hypothesis H_0 is equivalent to the hypothesis

$$H_0' : 6(a_j - a_{j-1}) = 0 \quad \text{and} \quad H_0' : \quad A(a_j - a_{j-1}) = 0$$

where A is constant. Spline $S(x)$ and its first and second derivative is continuous at point λ_j, so

$$-2(b_j - b_{j-1}) = 6\lambda_j(a_j - a_{j-1}), \quad \lambda_j(b_j - b_{j-1}) = -(c_j - c_{j-1}), \quad -\lambda_j^3(a_j - a_{j-1}) = d_j - d_{j-1}.$$

It follows that hypothesis of equality of correstonding coeficients of polynomials $P_{j-1}(x)$ and $P_j(x)$ are accepted at level α.

Theorem 1. Let there exists at most one statisticaly not significant knot λ_j between two statisticaly significant knots of spline $S(x)$. Function $\overline{S}(x)$ where $\overline{S}(x) = P_{j-1}(x)$ for $x \in [\overline{\lambda}_{j-1}, \overline{\lambda}_{j+1}]$ and $S(x) = \overline{S}(x)$ otherwise is statistical equivalent of spline $S(x)$ at level α.

Proof. We know that

$$P_{j-1}(\lambda_{j+1}) - P_{j+1}(\lambda_{j+1}) = (\lambda_{j+1}^3 - 3\lambda_{j+1}^2\lambda_j + 3\lambda_j^2\lambda_{j+1} - \lambda_j^3)(a_j - a_{j-1}),$$

$$P_{j-1}(\lambda_{j+1}) - P_{j+1}(\lambda_{j+1}) = (3\lambda_{j+1}^2 - 6\lambda_j\lambda_{j+1} + 3\lambda_j^2)(a_j - a_{j-1}),$$

$$P_{j-1}(\lambda_{j+1}) - P_{j+1}(\lambda_{j+1}) = (6\lambda_{j+1} - 6\lambda_j)(a_j - a_{j-1}).$$

So statement follows from Lema 1.

Theorem 2. Let $\Lambda_j = \{\lambda_i,\ i=0,\dots,g_j\}$ is set of knots of spline $S_{g_j}(x)$ $j=1,2$. If $g_1 > g_2$ and $\Lambda_1 \supseteq \Lambda_2$ then:

$$(6) \qquad F = \frac{F_{g_2} - F_{g_1}}{F_{g_1}} \; \frac{m-g_1-4}{g_1-g_2} \sim F_{m-g_1-4,\, g_1-g_2}.$$

Proof. From $F_{g_1}/\sigma^2 \sim \chi^2_{m-g_1-4}$, $F_{g_2}/\sigma^2 \sim \chi^2_{m-g_2-4}$, and $F_{g_2}/\sigma^2 > F_{g_1}/\sigma^2$ it follows that $F_{g_2}/\sigma^2 - F_{g_1}/\sigma^2 \sim \chi^2_{g_1-g_2}$ [6]. According to Fisher-Cohran theorem it follows that $F_{g_1}/\sigma^2 - F_{g_2}/\sigma^2$ and F_{g_1}/σ^2 are indipendent and that (6) is true.

Theorem 2. can be used to estimate the upper limit of increasing of the sum of squares of residuals of the spline finction with smaler number of knots, when positions of knots is not changed. For prescribed evel of significance α value F_α can be found such that

$$P\{0 < F < F_\alpha\} = 1-\alpha.$$

hen

$$0 < F_{g_2} - F_{g_1} < F_\alpha \, F_{g_1} \, \frac{g_1-g_2}{m-g_1-4} \quad \text{with probability } \alpha.$$

REFERENCES

[1]. Buse, A., and Lim, L., Cubic Spline as a Special Case of Restricted Least Squares, J. Amer. Stat. Assoc. 72, (1977), 64-68

[2]. Dierckx, P., A Fast Algorithm for Smoothing Data on a Rectangular Grid while Using Spline Functions, SIAM J. Numer. Anal., Vol 19, No 6, (1982), 1286-1303

[3]. Dierckx, P., Algorithm for Smoothing Data with Periodic and Parametric Splines, Comp. Graph. & Image Proc., 20,(1982), 171-174

[4]. Gallant, A.R. and Fuller, W.A., Fitting Segmented Polynomial Regression Models Whose Join Points Have to Be Estimated, J. Amer. Stat. Assoc. Vol 68, No 341, (1973), 144-147

[5]. Reinch, C.H., Smoothing by Spline Functions II, Numer. Math. 16, (1971), 451-454

[6]. Seber, Linejnyj Regresionnyj Analuz, Mir, Moskva, 1980.

[7]. Smith, P., Splines as a Useful and Convenient Statistical Tool, The Amer. Stat. 33, (1979), 57-62

[8]. Smith, P., Hypothesis testing in B-spline Regression, Commun. Stat. Simul. Compt. 11(2), (1982), 143-157

[9]. Surla, K., Jerinuć, LJ., Lozanov, Z., and Kovačević, R., An Application of Spline Functions in Data Analysis, V znan. skup "Proj. prorač. pomoću računala, Stubičke Toplice, Zbor. rad., (1983), 77-81

[10]. Wahba, G., and Wold, S., A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation, Comm. Stat. 4(1), (1975, 1-17

[11].Wold, S., Spline Functions in Data Analysis, Technometrics 16, (1974), 1-11

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

APPROXIMATION IN DISCRETE CONVEXITY CONES

Ivan B. Lacković , Ljubiša M. Kocić

ABSTRACT:

The necessary and sufficient conditions for positivity of linear continuous operators on a cone of convex sequences are given. The main theorem is based on the representation of every sequence as a limith (in d_s metric, given by (2)) of sequences $u^{(n)}$ given by (6). This is a discrete analogue of the result given in [2], and a generalization of result from [1].

APROKSIMACIJE U DISKRETNIM KONUSIMA KONVEKSNOSTI. Dobijeni su potrebni i dovoljni uslovi za pozitivnost linearnih neprekidnih operatora na konusu konveksnih nizova. Teorema se bazira na reprezentaciji svakog niza kao granice (u metrici d_s , koja je data sa (2)) nizova $u^{(n)}$ datih sa (6). Dobijeni rezultati predstavljaju diskretnu analogiju rezultata iz [2], i generališu rezultate iz [1].

1. ALGEBRA

In this paper, the following denotations will be used: $N = \{1, 2, 3, \ldots \}$, $N_o = N \cup \{0\}$, $x = (x_o, x_1, \ldots) = (x_k)(k \in N_o)$, S - the set of all sequences x . Further, the sequences $e_n \in S$ $(n \in N_o)$ is defined by

(1) $$e_n = (\delta_{nk})(k \in N_o),$$

where δ_{nk} is Kronecker's delta, $\delta_{nk} = \begin{cases} 0, & k \neq n , \\ 1, & k = n. \end{cases}$. For two sequences $x,y \in S$ we write $x = y$ if $x_k = y_k$ for every $k \in N_o$, and $x + y = (x_k + y_k)(k \in N_o)$. If $\lambda \in R$ then λx means $(\lambda x_k)(k \in N_o)$. Thus, S together with defined operations consist a linear space over the field R, with the sequences $e_n (n \in N_o)$ as a base.

The sequence $x = (x_k)(k \in N_o)$ is convex if $\Delta^2 x_n = x_{n+2} - 2x_{n+1} + x_n \geqslant 0$ $(n \in N_o)$. The set of all convex sequences will be denoted by K. It is known, that K is a cone in S.

Further, let $D \subset R$ be a nonempty set. With F(D) we will denote the set of all functions $f: D \Rightarrow R$. For the operator $A: S \Rightarrow F(D)$ we say that it is linear if for every $x,y \in S$ and

$\lambda, \mu \in R$ the equality $A(\lambda x + \mu y) = \lambda Ax + \mu Ay$, holds.

If for $x \varepsilon S$ and $f_x \varepsilon F(D)$ we have $f_x = Ax$, then we write $Ax \geqslant 0$ if $f_x(t) \geqslant 0$, for every $t \varepsilon D$. Similarly, $Ax = 0$ if $f_x(t) = 0$ for every $t \varepsilon D$.

2. TOPOLOGY

Let $x = (x_k)$ and $y = (y_k)$ be two sequences from S, and $d_s(x,y)$ be a distance between x and y, introduced with

$$(2) \qquad d_s(x,y) = \sum_{k=o}^{+\infty} 2^{-k} \frac{|x_k - y_k|}{1 + |x_k - y_k|} \; .$$

Now, (S, d_s) is a metric space with finite metric: $d_s(x,y) < +\infty$, for every $x,y \varepsilon S$. The sequence $x^{(n)} \varepsilon S$ converges to $x \varepsilon S$ in metric d_s if $d(x^{(n)}, x) \to 0$ when $n \to \infty$. Then we write $x^{(n)} \xrightarrow{d_s} x$, or $\lim\limits_{n} x^{(n)} = x$ in metric d_S, and say that x i a d_s- limit for $x^{(n)}$.

Now we have:

Lemma 1. Every sequence $u = (u_k)(k \varepsilon N_o)$ from S is a d_s-limi of the sequences $u^{(n)}$ $(n \varepsilon N_o)$ having the form

$$(3) \qquad u^{(n)} = \sum_{k=o}^{n} u_k e_k .$$

Proof. We have

$$d_s(u^{(n)}, u) = \sum_{i=n+1}^{+\infty} 2^{-i} \frac{|u_i|}{1 + |u_i|} \leqslant \sum_{i=n+1}^{+\infty} 2^{-i} = 2^{-n} \text{, wherefrom}$$

$d_s(u^{(n)}, u) \to 0$ when $n \to \infty$. Consequently,

$$u = \lim_n u^{(n)} = \lim_n \left(\sum_{k=o}^{n} u_k e_k \right) = \sum_{k=o}^{+\infty} u_k e_k \quad (\text{in } d_s\text{- metric})$$

for every $u \varepsilon S$.

3. REPRESENTATIONS

From the lemma 1 we have that the representation

$$(4) \qquad u = \sum^{+\infty} u_\nu e_\nu \; ,$$

holds for every $u=(u_k)\varepsilon S$. However, we need the following statement concerning representations.

Let the sequences E_o, E_1 and W_k be defined with

$$(5) \qquad E_o = \sum_{k=o}^{+\infty} e_k \; , \quad E_1 = \sum_{k=o}^{+\infty} ke_k \; , \quad W_k = \sum_{i=k+2}^{+\infty} (i-k-1)e_i \quad (k\varepsilon N_o) \; .$$

Now, we have

Theorem 1. (a) Every sequence of the form

$$(6) \qquad u^{(n)} = \lambda^{(n)}E_o + \mu^{(n)}E_1 + \sum_{k=o}^{n} c_k^{(n)}W_k \qquad (n\varepsilon N_o) \; ,$$

where $\lambda^{(n)}$, $\mu^{(n)} \varepsilon R$, $c_k^{(n)} \geqslant 0$ $(k\varepsilon N_o)$ for fixed n, is convex, i.e. depands to the cone K.

(b) Every sequence $u\varepsilon K$ is a limit (in d_s metric) of sequences $u^{(n)}$ given by (6).

Proof. (a) It is obvious that $E_o=(1,1,1,\ldots)$ and $E_1=(0,1,2,3,\ldots)$, i.e. $\Delta^2E_{ok} = \Delta^2E_{1k} = 0$ $(k\varepsilon N_o)$. From (5) we also have $W_{ki}=\begin{cases} 0 \; , & 0 \leqslant i < k+2 \\ \binom{i+1}{2} \; , & i \geqslant k+2 \end{cases}$ which gives $\Delta^2W_{ki}=\begin{cases} 0, & 0 \leqslant i < k-2 \\ 1, & i \geqslant k-2 \end{cases}$, i.e. $\Delta^2W_{ki} \geqslant 0$ for every $i\varepsilon N_o$ and $k\varepsilon N_o$. In virtue of nonnegativity of $c_k^{(n)}$ we have from (6) $\Delta^2u^{(n)} \geqslant 0$. Accordingly, $u^{(n)}$ is convex for every $n\varepsilon N_o$.

(b) Substituting the obvious identity

$$u = u_o + k\Delta u_o + \sum_{i=o}^{k-2} (k-i-1)\, \Delta^2 u_i$$

into (4) we have

$$(7) \quad u = u_oe_o + u_1e_1 + \sum_{k=2}^{+\infty} (u_o + k\Delta u_o + \sum_{i=o}^{k-2} (k-i-1)\Delta^2 u_i)\, e_k \; .$$

After some transformations (7) get the form

$$u = u_o(\sum_{k=o}^{+\infty} e_k) + (\Delta u_o)(\sum_{k=o}^{+\infty} ke_k)$$

$$+ \sum_{k=o}^{+\infty} (\Delta^2 u_k)(\sum_{i=k+2}^{+\infty} (i-k-1)e_i) \; ,$$

which is a d_s-limit of the sequence

$u^{(n)} = u_o \sum_{k=o}^{+\infty} e_k + \Delta u_o \sum_{k=o}^{+\infty} k e_k + \sum_{k=o}^{n} (\Delta^2 u_k)(\sum_{i=k+2}^{+\infty} (i-k-1)e_i)$, or by notation introduced by (5) $u^{(n)} = u_o E_o + (\Delta u_o)E_1 + \sum_{k=o}^{n} (\Delta^2 u_k) W_k$. Thus, we have $d_s(u^{(n)}, u) = \sum_{k=n+2}^{+\infty} 2^{-k} \frac{|u_k - v_k|}{1 + |u_k - v_k|} \quad 2^{-n}$, for every fixed n, where v_k $(k \geqslant n+2)$ is k-th therm of $u^{(n)}$. So, $d_s(u^{(n)}, u) \to 0$ $n \to +\infty$ i.e. u is a d_s-limit of the sequences $u^{(n)}$ given by (6

4. APPLICATIONS

Using the theorem 1 we can obtain the following theorem

Theorem 2. Let the operator $A: S \to F(D)$ be linear and continu over the sequences in S. Then, for every $u \varepsilon S$, the implication

(8) $\qquad u \varepsilon K \Longrightarrow A u \geqslant 0$

holds if and only if

(9) $\qquad A E_o = A E_1 = 0,$

(10) $\qquad A W_k \geqslant 0 \quad (k \varepsilon N_o).$

Proof. i) Suppose that (8) holds for every $u \varepsilon S$. Then, i we choose $u = E_o (u = E_1)$ we have that $u \varepsilon K$ which imply $Au \geqslant 0$, i. $AE_o \geqslant 0$ $(AE_1 \geqslant 0)$. But $-u \varepsilon K$ too. Thus, $A(-E_o) \geqslant 0$ or $A(-E_1) \geqslant 0$ wherefrom $AE_o = AE_1 = 0$. By theorem 1-a $W_k \varepsilon K$ $(k \varepsilon N_o)$, so, in vi tue of (8) we have $AW_k \geqslant 0$ for every $k \varepsilon N_o$.

/ ii) Suppose now that (9) and (10) holds. Then, on the basis of theorem 1-b, every sequence $u \varepsilon K$ is d_s-limit of the sequence $(u^{(n)})(n \varepsilon N_o)$ given by (6). This means that $u = \lim_n u^{(n)}$, wherefrom, accordingly with continuity of A over t sequences in S, we have

$$A u = A (\lim_n u^{(n)}) = \lim_n (A u^{(n)}),$$

and, in virtue of linearity of A over the sequences in S, we get

$$A u = \lim_n \left(\lambda^{(n)} (AE_o) + \mu^{(n)} (AE_1) + \sum_{k=o}^{n} c_k^{(n)} (AW_k) \right)$$

and, if (9) and (10) holds, and keeping in the mind that $c_k^{(n)} \geqslant$

ve ınaly obtain $Au \geqslant 0$.

Remark 1. It is easy to see that our theorem 2 generalizes the :esult of theorem 4 in [1]. In this case an operator A have the form of triangular matrix.

Remark 2. The representation (6) is a discrete analogue of the relation (6) in [2]. The sequences W_k we can call a discrete splines.

REFERENCES

1. MITRINOVIĆ D.S., LACKOVIĆ I.B., STANKOVIĆ M.S.: Addenda to the monograph "Analytic Inequalities" II. On some convex sequences connected with N.Ozeki's result. Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No 634 - No 677 (1969), 3-24
2. VASIĆ P.M., LACKOVIĆ I.B.: Notes on convex functions II: On continuous linear operators defined on a cone of convex functions. ibid. No 602 - No 633 (1978), 53-59.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

APPROXIMATION OF CONVEX FUNCTIONS BY FIRST DEGREE SPLINES

Ljubiša M. Kocić

ABSTRACT:
A method for approximation of functions convex on a finite interval by picewise affine function is developed. For the segments of approximating function we used the support affine function f in prescribed points, and its graph lies not up then the graph of f .

APROKSIMACIJA KONVEKSNIH FUNKCIJA SPLAJNOVIMA PRVOG STEPENA. U radu je razvijen metod aproksimacije konveksne funkcije, deo po deo afinom funkcijom. Segmenti aproksimirajuće funkcije su potporne afine funkcije aproksimirane funkcije f u zadatim tackama, a njen grafik leži ne iznad grafika funkcije f .

1. INTRODUCTION

The problem of approximation of one variable function by first degree splines (picewise affine functions, polygonal lines) is minutely studied from many autors,and for classes $C[a,b]$ and $C^2[a,b]$, as well as the interpolated classes H_ω and WH_ω , see,for example, [1] and [3] . A lot of results are oriented t(applications on computers [7]. In all appearance, this kind of approximation is especially important for convex functions. Erly results was obtained by K. TODA [6] and T. POPOVICIU [4]. They were shown the following theorem:

Theorem 1. The first degree spline function

$$(1) \qquad (S_nf)(x) = px + q + \sum_{k=o}^{n} c_k(x - x_k)_+ \ , \quad x \in [a,b] \ , \quad n \in N \ ,$$

where $p, q \varepsilon R$, $c_k \geqslant 0$, $x_k \varepsilon [a,b]$ $(k = 0, 1, \ldots, n)$ is convex on $[a,b]$. Furthermore, every convex function, defined on $[a,b]$ is the uniform limit of the sequence $S_n f$ of the form (1), where $x_k \varepsilon [a,b]$, $p = [x_o, x_1; f]$, $q = f(x_o) - x_o [x_o, x_1; f]$ and $c_k = \frac{2}{n}(b-a)[x_k, x_{k+1}, x_{k+2}; f]$ $(k = 0, 1, \ldots, n-2)$.

Of course, the spline $S_n f$ interpolates f in the knoots x_k $[a,b]$, and if we introduce for $x \varepsilon [a,b]$

(2) $$\text{epi}\, f = \left\{ (x, y) \varepsilon R^2 \mid y \geqslant f(x) \right\},$$

(3) $$\text{hyp}\, f = \left\{ (x,y) \varepsilon R^2 \mid y \leqslant f(x) \right\},$$

it is easy to see that $(S_n f)(x) \varepsilon$ epi f, for $n \varepsilon N$, and $x \varepsilon$ $[a,b]$.

On the basis of theorem 1, P. M. VASIĆ and I. B. LACKOVIĆ vere proved an important theorem on the positivity of linear operators [8, p. 55]. But, the attempt to formulate an analogue theorem for functions of two (or more) variables, based on TODA - POPOVICIU type theorem shall not be successful The reason is that the coefficients, corresponding to c_k will not be nonnegative for convex function $(x,y) \to f(x,y)$. By the other words, a polygonal surface, inscribed in the graph of $f(x,y)$, must not have a nonnegative coefficients. In this sense a polyhedral surface, circumscribed around $f(x,y)$ will be much convinient. Thus, we shall develope this kind of approximation for one variable function. This circumscribed spline wil be denoted by $s_n f$, because it is a kind of lower bound for f as $S_n f$ is a kind of its upper bound. Also, $(s_n f)(x) \varepsilon$ hyp f.

2. PRELIMINARY LEMAS

Function $f: [a,b] \to R$ is convex on $[a,b]$ if the inequality $f(\lambda u + (1-\lambda)v) \leqslant \lambda f(u) + (1-\lambda) f(v)$ holds for every u, v

a, b] and $\lambda \varepsilon (0, 1)$. The function f is called strictly con-ex if above inequality is strict. Let $K[a, b]$ and $K^+[a, b]$ enote the cones of convex and strictly convex functions on a, b] , continuous from the left (right) at the edge point a b). The function $x \rightarrow p(t,x)$ is called support function of con-ex function f if the following conditions are fulfiled :

$$f(x) \geqslant p(t,x), \quad x \varepsilon [a,b] \ , \ x \neq t \quad \text{and} \quad f(t) = p(t,t).$$

or strictly convex functions the following lemma takes place see for ex. [5]):

emma 1. Let $f:[a,b] \rightarrow R$ be strictly convex on $[a,b]$, and let $'_+(x)$ be the right derivative of f in x. Then

a) f'_+ is increasing function on $[a,b]$,

b) for every $a \leqslant u < v \leqslant b$ holds

$$f'_+(u) < \frac{f(v) - f(u)}{v - u} < f'_+(v).$$

Let $f \varepsilon K^+[a,b]$, and $x_1 < x_2 < \dots < x_n$ be a set of knoots from (a,b). The first order spline, circumscribed around the graph of f is given by

$$(4) \qquad (s_n f)(x) = \sup_{1 \leqslant i \leqslant n} \{ p(x_i,x) \} , \quad x \varepsilon [a,b] ,$$

where $p(x_i,x)$ is affine support function of f in the point x_i:

$$(5) \qquad p(x_i,x) = f(x_i) + f'_+(x_i)(x - x_i) . \ \square$$

So we have

Lemma 2. a) $s_n f$ interpolates f in the knoots $x_1, \dots, x_n$,

b) $s_n f$ is convex on $[a,b]$.

Proof. a) Let $1 \leqslant k \leqslant n$ be a fixed number. Then, $\sup_{1 \leqslant i \leqslant n} \{p(x_i,x_k)\} = p(x_k,x_k) = f(x_k)$ which is an interpolate property.

b) As every function $x \mapsto p(x_i,x)$ $(i = 1, 2, \dots, n)$ is convex, so $\{ p(x_i,x) \mid 1\le i \le n \}$ is a family of convex functions. But, it is known ([5]) that $\sup\limits_{1\le i\le n} p(x_i,x)$ is also convex.

Now, let x_k and x_{k+1} be two adjacent knoots. Correspodent support lines are $p(x_k,x)$, $p(x_{k+1},x)$. Note that the equation $p(x_k,x) = p(x_{k+1},x)$ have a solution

$$(6) \qquad t_k = \frac{f'_+(x_{k+1})x_{k+1} - f'_+(x_k)x_k + f(x_k) - f(x_{k+1})}{f'_+(x_{k+1}) - f'_+(x_k)}$$

if $f'_+(x_{k+1}) \neq f'_+(x_k)$ i.e. if $f \in K^+[a,b]$. The points t_k (k= 1, 2, ..., n-1) are the abscisses of vertex of the polygonal line which consists the graph of the spline $s_n(x)$.

Lemma 3. If $f \in K^+[a,b]$ then the inequalities

$$(7) \qquad x_k < t_k < x_{k+1}$$

holds for every $k = 1, 2, \dots, n-1$.

Proof. From lemma 1 - b) , we have that $f'_+(x_k) < \dfrac{f(x_{k+1}) - f(}{x_{k+1} - x_k}$

i.e. from $x_{k+1} - x_k > 0$ follows $(x_{k+1}-x_k)f'_+(x_k)<f(x_{k+1}) - f($ wherefrom we get $x_{k+1}f'_+(x_{k+1}) - x_kf'_+(x_k) + f(x_k) - f(x_{k+1}) < x_k \cdot (f'_+(x_{k+1}) - f'_+(x_k))$ or $t_k < x_{k+1}$. In the similar way, from $f(x_{k+1}) - f(x_k) < f'_+(x_{k+1})(x_{k+1} - x_k)$ we get $f'_+(x_{k+1})x_{k+1} - f'_+(x_k)x_k + f(x_k) - f(x_{k+1}) > x_k(f'_+(x_{k+1}) - f'_+(x_k))$, i.e. $t_k >$

In the seqel, we introduce a v-shaped function v_k w

$$v_k(x) = \sup_{x_k \le x \le x_{k+1}} \{ p(x_k,x) \ , \ p(x_{k+1},x) \} \ ,$$

which approximates f on $[x_k, x_{k+1}]$. Let E_k be defined wi $E_k = f - v_k$, and $\|\cdot\|$ be the sup norm. Then we have

Lemma 4. $\|E_k\| = f(t_k) - v_k(t_k)$, for every $f \in K^+[a,b]$.

Proof. On the basis of definition of E we have

$$(8) \qquad E(x) = \begin{cases} f(x) - p(x_k,x) \ , & x \in [x_k, t_k), \\ f(x) - p(x_{k+1},x). & x \in [t_k, x_{k+1}]. \end{cases}$$

We shall prove that $E(x)$ monotonely increasing on (x_k, t_k). Let $x_k < x < y < t_k$. Then, we can find $\lambda \varepsilon (0, 1)$ so that $x = \lambda x_k + (1-\lambda)y$, which, with strict convexity of the function f gives

$$(9) \qquad f(x) < \lambda f(x_k) + (1-\lambda) f(y) .$$

From (8), (5) and (9) we have $E(x) = f(x) - f(x_k) - f'_+(x_k)(x-x_k)$ $\lambda f(x_k) + (1-\lambda)f(y) - f(x_k) - f'_+(x_k)(x-x_k) = (1-\lambda)[f(y)-f(x_k)]$ $- f'_+(x_k)(\lambda x_k + (1-\lambda)y - x_k) = (1-\lambda)[f(y) - f(x_k) - f'_+(x_k)(y-x_k)]$ $= (1-\lambda)E(y)$ for $x \varepsilon [x_k, t_k)$, which means that $E(x) < E(y)$, for $x_k < x < y < t_k$, in virtue of inequality $0 < \lambda < 1$. Thus, E is increasing on $[x_k, t_k)$.

In the quite similar way one can prove that E is decreasing on $[t_k, x_{k+1}]$. Being a continuous function, $E(x)$ attains its maximal value in t_k, i.e. $E_k = \sup_{[x_k, x_{k+1}]} \{E(x)\} = E(t_k) = f(t_k) - v_k(t_k)$. □

3. APPROXIMATION

On the basis of previous lemmas ve can state

Theorem 2. a) The spline s_n have explicit form

$$(10) \qquad (s_n f)(x) = A x + B + \sum_{k=1}^{n-1} d_k (x - t_k)_+ , \quad x \varepsilon [a, b] ,$$

where $A = (T_1 - T_o)/(t_1 - t_o)$, $B = (t_1 T_o - t_o T_1)/(t_1 - t_o)$, $T_k = p(x_k, t_k)$, $d_k = f'_+(x_{k+1}) - f'_+(x_k)$, t_k is given by (6) and $t_o = a$.

b) For every $f \varepsilon K^+[a, b]$, $(s_n f)(x)$ approximates f uniformly on $[a, b]$ when $n \to \infty$ and $\max_{1 \leq k \leq n-1} (x_{k+1} - x_k) \to 0$.

Proof. If we put $T_k = p(x_k, t_k)$, then the vertex of the polygonal line (4) have coordinates (t_k, T_k). This line, being a graph of the first degree spline have the form (10), where d_k must be a difference between the slope of the right support line, $f'_+(x_{k+1})$ and the left one, $f'_+(x_k)$. Of course, $d_k \geqslant 0$ $k = 1, \ldots, n-1$. The proof of b) follows from lemma 4 . Namely, for $x \varepsilon [x_k, x_{k+1}]$, we have $|f(x) - v_k(x)| \leqslant E(t_k) = f(t_k) - f(x_k) - f'_+(x_k)(t_k - x_k)$, and if we put $h = x_{k+1} - x_k$, then $|f(t_k) - f(x_k)| \leqslant \omega(f,h)$, and also $t_k - x_k < h$, wherefrom $|f(x) - v_k(x)| < \omega(f,h) + h \to 0$, and accordingly, $f(x) - s_n(x) \to 0$, when $n \to \infty$ and $\max_k (x_{k+1} - x_k) \to 0$.

Sofar we deal only with strictly convex functions. What we have pointed out it is that no difficulties when we pass to convex functions. Namely, these subintervals of [a,b] on which f is affine, must be excluded, and remainded graph will be a strictly convex function. Now, we underline that the form of the spl ne does not new. There is no difference, in formal sense, betwee S_nf and s_nf. However, we have $s_nf \leqslant f \leqslant S_nf$ on [a,b], and n ε From this reason, we call S_nf an upper spline and s_nf a low spline of the convex function f . There is, also, a middle spl ne, have been studying by M. GAVRILOVIĆ in [2] and provid the mini-max approximation. The spline S_nf exists for every tinuous function. But, the middle and the lower spline do exis only for convex functions.

REFERENCES

1. ALBERG J.H., NILSON E.N., WALSH J.L.: The theory of splines and their applications (in russian, by complements of S.B. Stečkin and Yu.N. Subotin), Moscow 1972.
2. GAVRILOVIĆ M.M.: Optimal approximation of convex curves by functions which are picewise linear. J. Math. Anal. Appl. 5 (1975), 260 - 282.
3. MALOZEMOV V.N.: Ob otklonenii lomanyh. Vestnik Leningrad. U niv. 21 (1966), 150-153.
4. POPOVICIU, T.: Sur certaines inegalites qui caracterisent l fonctions convexes. A. Sti. Univ. "Al. I. Cusa" Iași Sect. Math. (N.S.) 11 B (1965), 155-146.
5. ROBERTS A.W., VARBERG D.E.: Convex functions. New York - Lo don 1973.
6. TODA K.: A method of approximation of convex functions. Tôh ku Math. J. 42 (1936), 311-317.
7. PAVLIDIS T.: Polygonal approximation by Newton's method. IEEE Trans. Computers. C-26 (1977), 800-80).
8. VASIĆ P.M., LACKOVIĆ I.B.: Notes on convex functions II: On continuous linear operators defined on a cone of convex fun ctions. Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fi No 602 - No 633 (1978), 53 - 59.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON THE SPLINE SOLUTIONS OF BOUNDARY VALUE PROBLEMS OF THE SECOND ORDER

Katarina Surla

ABSTRACT:
A tridiagonal difference scheme is developed for the boundary value problem (1). This scheme was derived by cubic spline according to Il'in [2]. The fitting factor of the form $\sigma_i = \frac{2sh^2(h_i\sqrt{q_i}/2)}{h_i^2 q_i}$ *was used in order to eliminate the condition* $h_i^2 q_i \leq 6$. *Error estimations for the solution and its derivatives are also given. In some cases these estimations appeared to be optimal* $(\beta_a = 0, h_o = Mh^2)$.

O SPLAJN REŠENJIMA KONTURNIH PROBLEMA DRUGOG REDA. *Posmatrana je tridiagonalna diferencna šema za konturni problem (1). Šema je izvedena primenom kubnog splajna prema [3]. U nameri da se eliminiše uslov* $h_i q_i^2 \leq 6$ *uveden je fiting faktor oblika* $\sigma_i = \frac{2sh^2(h_i\sqrt{q_1}/2)}{h_i^2 q_i}$. *Ocene greške za rešenje i izvode su date. U nekim slučajevima procene su optimalne* $(\beta_a = 0, h_o = Mh^2)$.

Consider the problem

$$
(1)\quad \begin{cases} -y''+q(x)y=f(x), \quad g(x)\geq 0,\ x\in[a,b],\ (a,b\in \mathbb{R}), \\ \alpha_a y(a)+\beta_a y'(a)=\gamma_a \qquad , \ |\alpha_a|+|\beta_a|\neq 0, \\ \alpha_b y(b)+\beta_b y'(b)=\gamma_b \qquad , \ |\alpha_b|+|\beta_b|\neq 0. \end{cases}
$$

The approximate solution of the problem (1) we wont to obtain in the form of the cubic spline $v(x)\in C^2[a,b]$ on the grid

$$a=x_o<x_1<\ldots<x_{n+1}=b$$

The restriction $v(x)$ on $[x_i,x_{i+1}]$ is $v_i(x)$, $v_i(x)=v_i^{(o)}+$ $+v_i^{(1)}(x-x_i)+\frac{1}{2}v_i^{(2)}(x-x_i)+\frac{1}{2}v_i^{(3)}(x-x_i)^3$ $(i=0,1,\ldots,n)$ where $v_i^{(k)}$ are constants which approximating $y_i^{(k)}=y^{(k)}(x_i)$. Using the equations

$$-\sigma_i v_i^{(2)}+q_i v_i^{(o)}=f_i \quad (i=0,1,\ldots,n)$$

$$-\sigma_{n+1}(v_n^{(2)}+h_n v_n^{(3)})+q_{n+1}(v_n^{(o)}+v_n^{(1)}h_n+\frac{h_n^2}{2}v_n^{(2)}+$$
$$+\frac{h_n^3 v_n^{(3)}}{6})=f_{n+1}$$

and the suppositions on the continuity we obtain

$$(2)\quad L^h v_i^{(o)}=k_i v_{i-1}^{(o)}+\ell_i v_i^{(o)}-m_i v_{i+1}^{(o)}=R_i, \quad (i=0,1,\ldots,n)$$

where $h_i=x_{i+1}-x_i$, $h=\max\limits_i h_i$, $h_{n+1}=h_n$

$$k_i=(1-\frac{h_{i-1}^2 q_{i-1}}{6\ \sigma_{i-1}})\frac{1}{h_{i-1}}, \quad m_i=(1-\frac{h_i^2 q_{i+1}}{6\ \sigma_{i+1}})\cdot\frac{1}{h_i}$$

$$\ell_i=(1+\frac{h_i^2 q_i}{3\ \sigma_i})\cdot\frac{1}{h_i}+(1+\frac{h_{i-1}^2 g_i}{3\ \sigma_i})\cdot\frac{1}{h_{i-1}}$$

$$R_i=\frac{h_{i-1}f_{i-1}}{6\sigma_{i-1}}+\frac{f_i}{3\sigma_i}(h_{i-1}+h_i)+\frac{f_{i+1}h_i}{6\sigma_{i+1}}, \quad (i=1,\ldots,n-1)$$

$$k_o=0,\ \ell_o=\frac{3\sigma_o+h_o^2 q_o}{3\sigma_o h_o}\beta_a-\frac{\alpha_a}{3\sigma_o}, \quad m_o=\frac{(6\sigma_1-h_o^2 q_1)\beta_a}{6\ \sigma_1\ h_o}$$

$$R_o=-\gamma_a-\beta_a\frac{h_o}{6}(\frac{2f_o}{\sigma_o}+\frac{f_1}{\sigma_1})$$

$$\varsigma_n = (1 - \frac{h_{n-1}^2 q_{n-1}}{6\ \sigma_{n-1}}) \cdot \frac{1}{h_{n-1}}, \quad m_n = 0$$

$$\ell_n = \frac{1}{h_{n-1}}(1 + \frac{h_{n-1}^2 q_n}{3\sigma_n}) + VF^{-1}, \quad V = q_{n+1} - A(\frac{q_n C}{\sigma_n} + \frac{6\alpha_b}{h_n B}) + \frac{q_n}{\sigma_n} E$$

$$F = h_n q_{n+1} - A \cdot D, \quad A = -\sigma_{n+1} + \frac{q_{n+1} h_n^2}{6}$$

$$D = (6\alpha_b h_n + 6\beta_b)/(h_n \cdot B), \quad B = \alpha_b h_n + 3\beta_b ,$$

$$C = \frac{3\alpha_b h_n + 6\beta_b}{\alpha_b h_n + 3\beta_b}, \quad E = -\sigma_{n+1} + \frac{q_{n+1} h_n^2}{2}$$

$$R_n = h_{n-1}(\frac{f_{n-1}}{6\sigma_{n-1}} + \frac{f_n}{3\sigma_n}) + \frac{1}{F}\ [f_{n+1} + \frac{f_n E}{\sigma_n} - A(\frac{f_n C}{\sigma_n} + \frac{6\gamma_b}{h_n B}\ ]$$

$$\sigma_i = \frac{2sh^2(h_i\sqrt{q_i}/2}{h_i^2\ q_i}$$

The constants $v_i^{(k)}$, k=1,2,3 we obtain from the relations

$$\alpha_a v_o^{(o)} + \beta_a v_o^{(1)} = \gamma_a$$

$$v_{i-1}^{(1)} = (\alpha_i v_i^{(o)} - \beta_i v_{i-1}^{(o)} - s_i) \cdot h_{i-1}^{-1}, \quad (i=1,n)$$

$$\alpha_i = 1 - \frac{h_{i-1}^2 q_i}{6\ \sigma_i}, \quad \beta_i = 1 + \frac{h_{i-1}^2 q_{i-1}}{3\ \sigma_{i-1}}, \quad s_i = -\frac{h_{i-1}^2}{2}\ (\frac{2f_{i-1}}{\sigma_{i-1}} + \frac{f_i}{\sigma_i})$$

$$v_n^{(1)} = v_{n-1}^{(1)} + r_n - \frac{h_{n-1}}{2}(\frac{q_n v_n^{(o)}}{\sigma_i} + \frac{q_{n-1} v_{n-1}^{(o)}}{\sigma_{n-1}}),$$

$$r_n = -\frac{h_{n-1}}{2}\ (\frac{f_{n-1}}{\sigma_{n-1}} + \frac{f_n}{\sigma_n})$$

$$-\sigma_i v_i^{(2)} + q_i v_i^{(o)} = f_i, \quad v_i^{(2)} = v_{i-1}^{(2)} + h_{n-1} v_{i-1}^{(3)}, (i=1,2,\ldots,n)$$

$$-\sigma_{n+1}(v_n^{(2)} + h_n v_n^{(3)}) + q_{n+1}\ (\ v_n^{(o)} + v_n^{(1)} h_n + h_n^2 v_n^{(2)}/2 +$$

$$+ h_n^3 v_n^{(3)}\ /6) = f_{n+1}$$

Similar to [2] it can be shown that $z_i^{(k)} = y_i^{(k)} - v_i^{(k)}$ satisfy the equations

$$L^h z_i^{(o)} = -k_i z_{i-1}^{(o)} + \ell_i z_i^{(o)} - m_i z_{i+1}^{(o)} = -\psi_i, \quad (i=0,\dots,n) \tag{3}$$

$$\psi_i = \phi_{i+1}^{(2)} h_i^{-1} - \phi_i^{(2)} h_{i-1}^{-1} + \phi_i^{(1)}, \quad (i=1,\dots,n-1)$$

$$\psi_o = \phi_1^{(2)} \beta_a h_o^{-1}$$

$$\psi_n = -\phi_n^{(2)} h_n^{-1} + \phi_n^{(1)} - \frac{1}{F}\left[\eta_{n-1} - \psi - \frac{6\psi_b A}{h_n B} - \frac{\eta_n}{\sigma_n}(E - A\cdot C)\right]$$

$$\eta_i = y_i^{(2)}(\sigma_i - 1), \quad \phi_i^{(2)} = \psi_i^{(o)} + h_{i-1}^2\left(\frac{\eta_{i-1}}{6\sigma_{i-1}} + \frac{\eta_i}{\sigma_i} - \frac{\psi_i^{(2)}}{6}\right)$$

$$\phi_i^{(1)} = \psi_i^{(1)} - \frac{h_{i-1}}{2}\psi_i^{(2)} + \frac{h_{i-1}}{2}\left(\frac{\eta_{i-1}}{\sigma_{i-1}} + \frac{\eta_i}{\sigma_i}\right)$$

$$\psi = -\sigma_{n+1}\psi_{n+1}^{(2)} + q_{n+1}\psi_{n+1}^{(o)}, \quad \psi_b = -\alpha_b \psi_{n+1}^{(o)} - \beta_b \psi_{n+1}^{(1)}$$

$$\psi_i^{(k)} = \frac{y^{IV}(\theta_i) h_{i-1}^{4-k}}{(4-k)!} \quad (k=0,1,2), \quad y_i \le \theta_i \le x_{i+1}$$

$$\alpha_a z_o^{(o)} + \beta_a z_o^{(1)} = 0 \tag{4}$$

$$z_i^{(1)} = z_{i-1}^{(1)} + \frac{h_{i-1}}{2}\left(\frac{q_{i-1} z_{i-1}^{(o)}}{\sigma_{i-1}} + \frac{q_i z_i^{(o)}}{\sigma_i}\right) + \phi_i^{(1)}, \quad (i=1,\dots,n) \tag{5}$$

$$\alpha_i z_i^{(o)} = \beta_i z_{i-1}^{(o)} + h_{i-1} z_{i-1}^{(1)} + \phi_i^{(2)}, \quad i=1,\dots,n) \tag{6}$$

$$z_i^{(2)} = (\eta_i + q_i z_i^{(o)})\sigma_i^{-1} \tag{7}$$

$$z_{i-1}^{(3)} = (z_i^{(2)} - z_{i-1}^{(2)} - \psi_i^{(2)}) h_{i-1}^{-1} \tag{8}$$

$$-\sigma_{n+1}(z_n^{(2)} + h_n z_n^{(3)}) + q_{n+1}\left(z_n^{(o)} + h_n z_n^{(1)} + \frac{h_n^2 z_n^{(2)}}{2} + \frac{h_n^3 z_n^{(3)}}{6}\right) = -\eta_{n-1} - \psi \tag{9}$$

THEOREM 1. Let $6\sigma_i - h_{i-1}^2 q_i \ge 0$, $\alpha_a\beta_a \le 0$, $\alpha_b\beta_b \ge 0$ and at leats one of q_k $(k=i-1,i,i+1)(q_{-1}=0)(i=0,1,\dots,n)$ is different from zero. Then the matrix of the system (2) is nverse monotone.

COROLLARY 1. The condition $6\sigma_i - h_{i-1}^2 q_i \ge 0$ in theorem 1 we can be replaced by $q_i(h_i^2 - h_{i-1}^2) + 6 \ge 0$.

THEOREM 2. Let the boundary value problem (1) has a unique solution $y(x) \in C^4[a,b]$. Let conditions of theorem 1 are fulfilled.

Then for $\beta_a \neq 0$ the following holds

$$|z_i^{(k)}| \leq Mh^2 \quad (k=0,1,2), \quad |z_i^{(3)}| \leq Mh$$

and for $\beta_a = 0$

$$|z_i^{(k)}| \leq Mh \quad (k=0,1,2), \quad |z_i^{(3)}| \leq M$$, where M denotes different constants, independent on h.

Proof.

$$(10) \quad \Delta_i = -k_i + \ell_i - m_i = \frac{h_{i-1}\ \sigma_{i+1}(2g_i\sigma_{i-1}+q_{i-1}\sigma_i)+h_{\ i-1}(2q_i\sigma_{\ i+1}+q_{i+1}\sigma_i)}{6\ \sigma_{i-1}\sigma_i\sigma_{i+1}} > Mh$$

$(i=1,\ldots,n)$

$$(11) \quad \Delta_o = \frac{\beta_o(\beta_1-\alpha_1)}{h_o} - \alpha_a > Mh$$

$$(12) \quad \Delta_n = \frac{1}{h_n}(\beta_n-\alpha_n) + \frac{h_{n-1}}{2}\left(\frac{q_n}{\sigma_n} + \frac{q_{n-1}}{n-1}\right) + VF^{-1} \geq Mh, \ M > 0$$

From (10),(11) and (12) we obtain that $||A^{-1}|| \leq \frac{1}{\max\limits_i \Delta_i} \leq Mh^{-1}$. A is matrix of the system (2).

Since $|\sigma_i - 1| \leq Mh^2$ we have $|\psi_i| = O(h^3)$ and then

$$|z_i^{(o)}| \leq ||A^{-1}||\ |\psi_i| \leq Mh^2$$

The estimates for $|z_i^{(k)}|$ $(k=1,2,3)$ we obtain from relations (4)-(9).

THEOREM 3. Let $h_i=h=\text{const}$, $\beta_a=\beta_b=0$ and the conditions of the theorem 1 are fulfilled. Then

$$|z_i^{(o)}| \leq Mh^3 \ ; \quad |z_i^{(k)}| \leq Mh^2; \quad |z_i'''| \leq Mh \ , \quad (k=1,2)$$

Proof. A simple calculation shows that $\bar{A} \geq h^{-1}B$, where $\bar{A}$ is matrix determined by (3) for $i=1,\ldots,n$, and $z_o^{(o)}=0$, $B=\{b_{ij}\}$ $(i,j=1,\ldots,n)$ is tridiagonal matrix with $b_{ii}=2$, $b_{i-1,i}=-1$, $(i=2,\ldots,n)$, $b_{i+1,i}=-1$ $(i=1,\ldots,n-1)$.

The solution of the system

$Bu = \omega$, $\omega = \max_i |\psi_i| h$ has the form $u_i = \frac{i(n+1-i)}{2}\omega$. Since $|z_i^{(o)}| \leq u_i$ we have $|z_1^{(o)}| \leq u_1 = O(h^3)$, $|z_n^{(o)}| \leq u_n = O(h^3)$. Then form (3) for i=2 we obtain $z_2^{(o)} = O(h^3)$ and then by induction we can conclude that $|z_i^{(o)}| = O(h^3)$ $(i=2,\ldots,n)$. The estimates for derivatives we get from (5)-(9).

THEOREM 3. Let $\beta_a = 0$ and $h_o = Mh^2$. Then $|z_i^{(k)}| \leq Mh^{4-k}$ $(k=0,1,2,3)$.

Proof. See [4].

REFERENCES

1. Doolan, E.P., Miller J.J.H., Sehilders W.H.A., *Uniform numerical methods for problems with Initial and Boundary Layers. BOOLE Press, Dubline, 1980.*
2. Il´in V.P., *O splaynovyh rešenijah obyknovenyh differencial´nyh uravnenij. Žurnal. vyčisl. mat. i mat. fiz., No 3, (1978), 621-627.*
3. Surla K., Kulpinski M., *O splajn rešenjima običnih diferencijalnih jednačina, Zbornik radova, V Znanstveni skup Proračunavanje i projektovanje pomoću računala, Stubičke Toplice, 1983, 137-141.*
4. Surla K., *Accuracy increase for some spline solutions of two-point boundary value problems, Rev. of Research Fac. of Sci. Univ. of Novi Sad, Volume 13 (1983) (in press).*
5. Jain M.K., Aziz T., *Numerical solution of stiff and convection-diffusion equations using adaptive spline function approximation. Appl. Math. Modelling, Vol. 7, February (1983), 57-62.*
6. Zavjalov, Ju.S., Kvasov B.I., Mirošničenko Z.L., *Metody splajn funkcii, Moskva 1980.*

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

MESH CONSTRUCTION FOR NUMERICAL SOLUTION OF A TYPE OF SINGULAR PERTURBATION PROBLEMS

Relja Vulanović

BSTRACT:

he singular perturbation problem (1) is considered. It is olved numerically by classical difference schemes on a nonniform mesh. The discretization mesh is construced in a pecial way, which gives linear convergence uniform in small erturbation parameter.

KONSTRUKCIJA MREŽE ZA NUMERIČKO REŠAVANJE JEDNOG TIPA SINGULARNIH PERTURBACIONIH PROBLEMA

Posmatra se singularni perturbacioni problem (1) koji se rešava numerički pomoću klasičnih diferencnih šema na neekvidistantnoj mreži. Mreža diskretizacije se konstruiše na specijalan način, tako da se dobija linearna konvergencija uniformna po malom perturbacionom parametru.

1. INTRODUCTION

We consider the two point boundary value problem

(1a) $L_\varepsilon u := \varepsilon^2 u'' + xb(x)u' - c(x)u = f(x)$, $x \in I = [0,1]$

(1b) $u(0) = U_0$, $u(1) = U_1$,

with basic asumptions

$$b,c,f \in C^2(I)$$

$$b(x) > 2\beta > 0, \quad c(x) \geq \gamma > 0, \quad 2b(0) < c(0) \,,$$
$$0 < \varepsilon \leq \varepsilon_0 \,.$$

This problem was solved in [2] by a special method which gives linear convergence uniform in small perturbation parameter ε. Our method seems to be somewhat simpler. It is based on the idea of Bahvalov, [1], that was generalized in [3] and uses a special mesh construction. We also achieve linear convergence uniform in ε but with less constraints - in [2] it was assumed: $b,c,f \in C^3(I)$, $3b(0)<c(0)$.

Now we shall give some estimates for the derivatives of the solution $u_\varepsilon \in C^4(I)$ to the problem (1). We use the result from Theorem 2. from [2]. Each positive constant independent of ε and of discretization mesh will be denoted by M.

THEOREM 1. For the solution u_ε to the problem (1) the following estimates hold:

$$(2) \quad |u_\varepsilon^{(i)}(x)| \le M(\varepsilon^{-i'} + \varepsilon^{-i}\exp(-\beta(\frac{x}{\varepsilon})^2)) ,$$

$$i=0,1,2,3,$$

where $\qquad i' = \max(0,i-2)$.

Proof. Using the same proof as in [2], we obtain

$$(3) \quad |u_\varepsilon^{(i)}(x)| \le M(1 + \varepsilon^{-i}\exp(-h(x)/\varepsilon^2)),\ i=0,1,2,$$

where $h(x) = \int_0^x tb(t)dt$. We use this inequality to get (2) for $i=0,1,2$. This part of the proof needs the asumption $u_\varepsilon \in C^4(I)$ and $2b(0) < c(0)$.

Let us now obtain the estimate (2) for i=3. Differentiating both sides of (1a) twice, we have

$$\varepsilon^2 u^{IV} + xb(x)u''' = g(x),$$

where, according to (3):

$$|g(x)| \le M(1 + \varepsilon^{-2}\exp(-h(x)/\varepsilon^2)).$$

Now it follows

$$u'''(x) = \exp(-h(x)/\varepsilon^2)\left[u'''(0) + \varepsilon^{-2}\int_0^x g(t)\exp(h(t)/\varepsilon^2)dt\right].$$

Since from [2] we have $|u'''(0)| \le M\varepsilon^{-3}$, we get:

$$|u'''(x)| \le M(A+B+C),$$

with

$$A = \varepsilon^{-3}\exp(-h(x)/\varepsilon^2) ,$$

$$B = \varepsilon^{-2}\int_0^x \exp((h(t)-h(x))/\varepsilon^2)dt ,$$

$$C = \varepsilon^{-4}x \exp(-h(x)/\varepsilon^2) .$$

Now we have

$$A,C \le M\varepsilon^{-3}\exp(-\beta(x/\varepsilon)^2)$$

and, see [2],

$$B \le \frac{1}{\varepsilon\beta y} (1-\exp(-\beta y^2)), \quad y = x/\varepsilon .$$

Hence,

$$B \le M\varepsilon^{-1}$$

and the theorem is proved.

2. THE MESH CONSTRUCTION

Let us denote by q a fixed number, $q \in (0,1)$, independent of ε and take $0 < a < q/\varepsilon_o$. Let for $t \in [0,q)$:

$$\phi(t) = t/(q-t), \quad \psi(t) = a\varepsilon\phi(t).$$

We have $\phi^{(k)}(t) > 0$, $k=1,2$. The mesh points are given by

$$x_i = \lambda(t_i), \quad t_i = i/n, \quad i=0,1,\dots,n,$$

where $n \in \mathbb{N}$, $n > 4/q$, and

$$\lambda(t) = \begin{cases} \psi(t) , & t \in [0,\alpha] \\ \psi(\alpha) + \psi'(\alpha)(t-\alpha), & t \in [\alpha,1] \end{cases} .$$

Here $(\alpha,\psi(\alpha))$ denotes the contact point of the tangent line taking the value 1 at 1, to the curve $\psi(t)$. Since we have $a\varepsilon < q$ it follows $\psi'(0) < 1$ and $\alpha \in (0,q)$ uniquely exists. For α we can get:

$$\alpha = (q-(a\varepsilon q(1-q+a\varepsilon))^{1/2})/(1+a\varepsilon) .$$

Note: The function $\lambda(t)$ which we give here is one of the class of functions that was constructed in [3] for a different type of problem, namely - the problem (1) with $b(x)\equiv 0$.

On this mesh we form the discretization of the problem (1):

$$u_o = U_o,$$

$$(4) \qquad L_h u_i := \varepsilon^2 D'' u_i + x_i b(x_i) D' u_i - c(x_i) u_i = f(x_i), \quad i=1,2,\dots,n-1,$$

$$u_n = U_1,$$

where

$$D'' u_i = 2(h_{i+1}u_{i-1} - (h_i+h_{i+1})u_i + h_i u_{i+1})/(h_i h_{i+1}(h_i+h_{i+1})),$$

$$D' u_i = (u_{i+1}-u_i)/h_{i+1},$$

$$h_i = x_i - x_{i-1}, \quad i=1,2,\dots,n.$$

3. CONVERGENCE UNIFORM IN ε

Because of $c(x) \geq \gamma > 0$ we can easyly get that the scheme (4) is stable uniformly in ε, see [3], for instance.

Now we shall state our main result.

THEOREM 2. For the solution u_ε to the problem (1) and for the solution u_i to (4) we have

$$|u\ (x_i) - u_i| \leq M \frac{1}{n}, \quad i=0,1,\dots,n.$$

Proof. We only have to prove consistency uniform in ε, i.e.

$$(5) \qquad |r_i| \leq M \frac{1}{n}, \quad i=1,2,\dots,n-1,$$

where

$$r_i = L_h u_\varepsilon(x_i) - (L_\varepsilon u_\varepsilon)(x_i).$$

Let $v_i = \exp(-\beta(x_i/\varepsilon)^2)$. We have

(6a) $|r_i| \leq M \frac{1}{n} (1+P_i+Q_i)$,

with

(6b) $P_i = \lambda'(t_{i+1})\frac{1}{\varepsilon} v_{i-1}$,

(6c) $Q_i = \lambda(t_i)\lambda'(t_{i+1})\varepsilon^{-2}v_i$.

Another estimate for r_i is

(7) $|r_i| \leq M(\varepsilon^2 + v_{i-1} + x_i + \frac{x_i}{\varepsilon} v_i)$.

The proof now follows the same way as in [1] (see [3] as well).

1° We first consider the case $t_{i-1} \geq \alpha$. Then we have $x_{i-1} \geq \lambda(\alpha) = a\varepsilon\phi(\alpha)$ and

$$v_i \leq v_{i-1} \leq \exp(-\beta a^2\phi^2(\alpha)) \ .$$

Because of $\lambda(t)$, $\lambda'(t) \leq M$, $t \in I$, we conclude

$$P_i, Q_i \leq M ,$$

and (5) is proved in this case.

2° Now let $t_{i-1} < \alpha$ and $t_{i-1} \leq q - \frac{4}{n}$. Then $t_{i+1} < q$ and $q - t_{i+1} \geq \frac{1}{2} (q - t_{i-1})$. From (6b) we can get

$$P_i \leq a \ \phi'(t_{i+1})v_{i-1} \leq$$

$$\leq M(q-t_{i-1})^{-2}\exp(-\beta a^2 (\frac{t_{i-1}}{q-t_{i-1}})^2) \leq M \quad .$$

Similarly, from (6c) we have

$$Q_i \leq a^2\phi(t_i)\phi'(t_{i+1})v_i \leq M$$

and (6a) give us (5).

3° The last case is

$$q - \frac{4}{n} < t_{i-1} < \alpha.$$

From this inequality it follows

$$q - \alpha < \frac{4}{n}$$

and

(8) $$\sqrt{\varepsilon} < M\,\frac{1}{n}\,,$$

because

$$q - \alpha > \frac{\sqrt{(1-q)aq\,\varepsilon}}{1+q}$$

Now $x_{i-1} > a\varepsilon\phi(q-\frac{4}{n})$, (notice $q-\frac{4}{n} > 0$), and we get

(9) $$v_{i-1} \le M\,\frac{1}{n}\ .$$

Similarly:

(10) $$\frac{x_i}{\varepsilon}\,v_i \le M\sqrt{v_i} \le M\,\frac{1}{n}\ .$$

For x_i we have

$$x_i = \lambda(t_i) \le \lambda(t_{i-1}) + M\,\frac{1}{n}\,,$$

and

$$\lambda(t_{i-1}) < \lambda(\alpha) < M\sqrt{\varepsilon}\,,$$

hence, using (8) we get

(11) $$x_i < M\,\frac{1}{n}$$

Now from (7-11) it follows (5) and the theorem is proved.

REFERENCES

1. Bahvalov,N.S.: *K optimizacii metodov rešenija kraevyh zadač pri naličii pograničnogo sloja, Ž. vyčisl. mat. i mat fiz., T9, No 4 (1969), 841-859.*
2. Lisejkin,V.D.: *O čislennom rešenii obyknovennogo differencial'nogo uravnenia vtorogo porjadka s malym parametrom pri staršej proizvodnoj, Čislennye metody meh.splošnoj sredy, T13, No 3 (1982), 71-80.*
3. Vulanović,R.: *Numeričko rešavanje konturnog problema drugog reda sa malim parametrom, magistarski rad, Novi Sad, 1983.*

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

AN ITERATIVE SOLUTION OF SOME DISCRETE ANALOGUES OF A MILDLY NONLINEAR BOUNDARY VALUE PROBLEM

Dragoslav Herceg, Ljiljana Cvetković

BSTRACT:

n this paper we consider numerical solution of the system of nonlinear equations $A(x)x=BFx$ by the iteration $x^o \in \mathbb{R}^n$, $x^{k+1}=A(x^k)^{-1}BFx^k$, $k=0,1,\ldots$. We apply our main result on some discrete analogues of a mildly nonlinear boundary problem, which are given in [1]. The results of [2] and [3] are the special cases of ours.

ITERATIVNO REŠAVANJE NEKIH DISKRETNIH ANALOGONA BLAGO NELINEARNIH KONTURNIH PROBLEMA. *U radu se posmatra numeričko rešavanje nelinearnog sistema jednačina $A(x)x=BFx$ iterativnim postupkom $x^o \in \mathbb{R}^n$, $x^{k+1}=A(x^k)^{-1}BFx^k$, $k=0,1,2,\ldots$. Naš glavni rezultat primenjujemo na neke diskretne analogone blago nelinearnog konturnog problema koji su dati u [1]. Rezultati iz [2] i [3] sadržani su u našem kao posebni slučajevi.*

1. INTRODUCTION

We shall consider a system of nonlinear equations

$$A(x)x = BFx, \tag{1}$$

where $A(x)$, $B \in \mathbb{R}^{n,n}$ (= set of all nxn real matrices) and where F is the nonlinear mapping of $\mathbb{R}^n$ into itself.

The i-th equation of (1) reads

$$\sum_{j=1}^{n} (A(x))_{ij}x_j = \sum_{j=1}^{n} B_{ij}(Fx)_j \ .$$

We abreviate this as

$$((A(x))_{i1},\ldots,\underline{(A(x))_{ii}},\ldots,(A(x))_{in})=$$
$$=(B_{i1},\ldots,\underline{B_{ii}},\ldots,B_{in}) \ ,$$

where we shall leave out zero entries and where we shall write common factors of the entries of the respective matrices in front of the prentheses. The diagonal elements are underlined.

The iteration which we shall consider for the solution of (1) is

$$(2) \qquad x^o \in \mathbb{R}^n \ , \quad A(x^k)x^{k+1} = BFx^k, \ k=0,1,\ldots \ .$$

If $A(x)$ is regular matrix for all $x \in \mathbb{R}^n$, the iteration (2) can be writen in the from

$$(3) \qquad x^o \in \mathbb{R}^n \ , \quad x^{k+1} = Tx^k, \quad k=0,1,\ldots \ ,$$

where $Tx = (A(x))^{-1}BFx$.

In next section we shall prove under certain assumptions on $A(x)$, B and F that T is contractive. Then the convergence of (3) follows from a well-known contraction-mapping theorem.

We apply our theorem to some discrete analogues of a milldly nonlinear boundary value problem of the form (4). These schemes occur frequently in the literature, see [1]. The special case of our theorem for the schem (5) was considered in [2] and [3]. The assumption in [3] was stronger than the one in [2].

For any step width $h=(n-1)^{-1}$, $n>2$, $n \in \mathbb{N}$, we define the grid $I_h = \{t_i = (i-1)h\colon \ i=1,2,\ldots,n\}$. For the numerical solution of problem

$$(4) \qquad \begin{aligned} -u'' + q(u)u &= f(t,u), \quad t \in [0,1] \\ u(0) = u(1) &= 0, \end{aligned}$$

we form the next discrete analogues of form (1). Let F is the nonlinear mapping of $\mathbb{R}^n$ into itself which assigns to $x \in \mathbb{R}^n$ the element $Fx \in \mathbb{R}^n$ whose i-th component is given via

$$(Fx)_i = f(t_i,x_i), \ i=1,2,\ldots,n.$$

The matrices $A(x)$ and B are defined by

(5) $h^{-2}(-1,\underline{2+h^2q(x_i)},-1) = (\underline{1})$ for $i=2,3,\dots,n-1$, (second order approximation),

(6) $\frac{h^{-2}}{12}(1,-16,\underline{30+12h^2q(x_i)},-16,1)=(\underline{1})$, for $i=3,4,\dots,n-2$, (fourth order approximation), and second order approximation for $i=2,n-1$ as in (5),

(7) $\frac{h^{-2}}{180}(-2,27,-270,\underline{490+180h^2q(x_i)},-270,27,-2)=(\underline{1})$ for $i=4,5,\dots,n-3$, (sixth order approximation), and fourth order approximation for $i=3,n-2$ as in (6),
and a fourth order unsymmetric approximation

$$\frac{h^{-2}}{12}(-10,\underline{15+12h^2q(x_i)},4,-14,6,-1)=(\underline{1}) \quad \text{for } i=2,$$

$$\frac{h^{-2}}{12}(-1,6,-14,4,\underline{15+12h^2q(x_i)},-10)=(\underline{1}) \quad \text{for } i=n-1.$$

In (5),(6) and (7) we have $(\underline{1})=(\underline{0})$ for $i=1,n$.

The solution $x=[x_1,x_2,\dots,x_n]^T \in \mathbb{R}^n$ of (1) is the numerical solution of the boundary value problem (4), i.e. $x_i \approx u(t_i)$, $i=1,2,\dots,n$.

THE CONVERGENCE ANALYSIS

Theorem. *Let* $A(x)=[a_{ij}(x)] \in \mathbb{R}^{n,n}$ *is inverse-monotone matrix for all* $x \in \mathbb{R}^n$ *and let* BF *is Frechèt-differentiable in* $\mathbb{R}^n$. *Suppose that*

$$\max_{1\le i\le n} \sum_{k=1}^{n} \sum_{j=1}^{n} \left|\frac{\partial a_{ik}}{\partial x_j}(x)\right| \le M, \quad \|BFx\|_\infty \le M_o, \quad \|(BF)'(x)\|_\infty \le M_1,$$

$$\|A^{-1}(x)\|_\infty \le \alpha, \quad x \in \mathbb{R}^n, \quad \alpha^2 M_o M + \alpha M_1 < 1.$$

Then the equation (1) has a unique solution and the sequence $x^o, x^1, x^2, \dots$, *generated by (3) converges to this solution.*

Proof. We shall prove that $\|T'(x)\|_\infty < 1$, where $T'(x)$ is Frechet derivative of $Tx = A(x)^{-1}BFx$. Let $C(x) = [c_{ij}(x)] = A(x)^{-1}$ and let $y \in \mathbb{R}^n$. From [3] we have

$$T'(y) = (C(x)BFy)'(y) + C(y)(BF)'(y).$$

Since $\| C(y) \|_\infty \le \alpha$, $\| (BF)'(y) \|_\infty \le M_1$ it follows

$$\| C(y)(BF)'(y) \|_\infty \le \alpha M_1 .$$

Let $G = [g_{ij}] = (C(x)BFy)'(y)$, $H_p = [\frac{\partial a_{ij}}{\partial x_p}] \in \mathbb{R}^{n,n}$, $p=1,2,\dots,n$.

Then

$$g_{ij} = \sum_{k=1}^{n} \frac{\partial c_{ik}(y)}{\partial x_j}(BFy)_k = \sum_{k=1}^{n} (C H_j C)_{ik}(BFy)_k = (C H_j CBFy)_i .$$

Since $C(y) \ge 0$, it follows

$$\max_{1\le i\le n} \sum_{j=1}^{n} |g_{ij}| \le \max_{1\le i\le n} \sum_{j=1}^{n} (C|H_j|C|BFy|)_i =$$

$$\max_{1\le i\le n} (C \sum_{j=1}^{n} |H_j|C|BFy|)_i = \| C(\sum_{j=1}^{n} |H_j|)C|BFy| \|_\infty \le$$

$$\|C\|_\infty \| \sum_{j=1}^{n} |H_j| \|_\infty \|C\|_\infty \| BFy\|_\infty \le \alpha^2 MM_o .$$

Now we have

$$\| T'(y) \|_\infty \le \alpha^2 MM_o + \alpha M_1 < 1 .$$

Theorem is proved.

APPLICATION TO THE PROBLEM (4)

We apply our theorem cn discrete analogues for (4), which are defined by (5),(6) and (7). First we summarize some properties of the matrices A(x) and $B = \mathrm{diag}(0,1,\dots,1,0)$ as defined by the schemes (5)-(7). The functions q(t) and f(t,u) are assumed to satisfy the conditions

$$\begin{gathered} q \in C^1(\mathbb{R}) , \quad f \in C^1(I \times \mathbb{R}) \\ |q'(t)| \le M, \ t \in \mathbb{R} , \quad |f(t,u)| \le M_o, \quad |\frac{\partial f}{\partial x}(t,u)| \le M_1, \ (t,u) \in I \times \mathbb{R} \\ -\lambda < \mu \le q(t) \le h^{-2} q_+ , \quad t \in \mathbb{R} , \end{gathered} \tag{8}$$

for some real μ, where λ and q_+ depend upon the scheme as follows. Let A_o is the matrix A(x) for $q(t) \equiv 0$, $t \in \mathbb{R}$. Then $A(x) = A_o + Q(x)$, where $A_o \in \mathbb{R}^{n,n}$ is independent of x and $Q(x) = \mathrm{diag}(0, q(x_2), \dots, q(x_{n-1}), 0)$. The matrix A_o is inverse-monotone, [1], and there exists the smallest positive eigenvalue λ to the eigenvalue problem $A_o x = \lambda B x$. From [1] we have that

$_{)}+ D$ is inverse-monotone for any diagonal matrix D whose .agonal elements are all in $(-\lambda, h^{-2}q_+]$. Next table schows a $_{1+}$ of this type where $q_+ = \infty$ means that $(-\lambda, u^{-2}q_+] = (-\lambda, \infty)$.

Scheme	(5)	(6)	(7)
q_+	∞	3	1/18

ow, from (8) follows that A(x) is inverse-monotone and

$$0 \leq A(x)^{-1} \leq (A_o + \mu B)^{-1} \quad \text{for} \quad x \in \mathbb{R}^n ,$$

$$\| A(x)^{-1} \|_\infty \leq \alpha ,$$

here

$$\alpha = \| (A_o + \mu B)^{-1} \|_\infty$$

epends on the scheme. Since

$$\max_{1 \leq i \leq n} \sum_{k=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial a_{ik}}{\partial x_j} (x) \right| = \max_{1 \leq i \leq n} \left| \frac{\partial a_{ii}}{\partial x_i} (x) \right| =$$

$$= \max_{1 \leq i \leq n} |q'(x_i)| \leq M ,$$

the assumptions of our theorem are satisfied. Then for any of the schemes (5)-(7) there exists the unique solution and the sequence $x^o, x^1, x^2, \ldots$ generated by (3) converges to this solution.

For any of the schemes (5)-(7) we have $\lambda(h) \leq \lambda$, [1] , where

$$\lambda(0) = \Pi^2, \quad \lambda(h) = 2h^{-2}(1 - \cos \Pi h), \quad h > 0.$$

This implies that $-\lambda < \mu$ is satisfied if $\mu > -\lambda(h)$. Now we can give easily computed estimates for $h > 0$ such that the condition $\mu > -\lambda(h)$ holds true for the schemes (5)-(7) if and only if $\mu > -\Pi^2$. We note that $\lambda(h)$ is monotone decreasing as a function of h and that $\lambda(h) \geq 8$ for $h \in |0,0.5]$. So, if $\mu > -8$, we have $\mu > -\lambda(h)$ for all $h \in |0,0.5|$. The restriction on h are described by

$$-\lambda(h) < \mu \leq h^{-2} q_+ .$$

The computable bounds of $\|(A_o + \mu B)^{-1}\|_\infty$ are given in [1]. So we have

$$\| (A_o + \mu B)^{-1} \| \leq \begin{cases} 1/8 & \text{for} \quad \mu = 0 \\ d - \mu^{-1} & \text{for} -\Pi^2 < \mu < 0, \end{cases}$$

where

$$d = (\mu\cos(0.5\beta))^{-1}, \quad 0 < \beta < \Pi, \quad \mu h^2 = 2(1-\cos\beta h) .$$

Now we can easyly see that the condition $\alpha^2 MM_o + \alpha M_1 < 1$ reduces to the case when $\alpha = 1/8$ for $\mu = 0$ and $\alpha = d - \mu^{-1}$ for $-\Pi^2 < \mu < 0$.

REFERENCES

1. BOHL E., LORENZ J.: *Inverse monotonicity and difference schemes of higher order*. A summary for two-point boundary value problems, Aeg.Math.19, (1979),1-36.
2. HERCEG D.: *Iterativno rešavanje diskretnog analogona jednog nelinearnog konturnog problema*. V Znanstveni skup Proračunavanje i projektovanje pomoću računala, Stubičke Toplice, 1983, 131-135.
3. POPOVICI P.: *An iterative solution of nonlinear finite-difference equations*. An.Univ. Timisoara, seria st.matematice, Vol. XX (1982), 49-57.

erical Methods and
Approximation Theory
Niš, September 26-28, 1984

ONE WAY OF DISCRETIZATION OF CHAPLYGIN'S METHOD

Dušan D. Tošić

BSTRACT:

ıaplygin's method (described in [1] and [3]) is an anaytic and iterative method for two-sided approximation to ıe solution of ordinary differential equations. This methi is difficult for practical applications in analytic form. ı this work one way of discretization of Chaplygin's meıod is proposed. Chapllygin's approximations are calculaэd by using the interpolation and the numerical integratiı. Some examples with the cubic spline interpolation and impson's rule are presented.

EDAN NACIN DISKRETIZACIJE ČAPLIGINOVE METODE. Čapliginova ..ɜtoda (opisana u [1] i [3]) je analitička i iterativna metoda za dvo-stranu aproksimaciju rešenja običnih diferencijalnih jednačina. Ova metoda je teška za praktične primene u analitičkoj formi. U ovom radu predložen je jedan način diskretizacije Čapliginove metode. Čapliginove aproksimacije se izračunavaju korišćenjem interpolacije i numeričke integracije. Navedeni su primeri sa interpolacijom pomoću kubnih splajnova i integracijom pomoću Simpsonovog pravila.

1. INTRODUCTTION

Let us consider the initial value problem:

(1) $$y' = f(x,y), \quad y(a) = y_0 .$$

We seek the solution $y(x)$ of (1) on the discrete point set $G_h = \{ x_i \mid x_i = a + ih, \; i=o,\dots,n, \; b-a = nh \}$. Suppose that the solution of (1) exists and that the conditions for the application of Chaplygin's method are satisfied (see [1]). If we denote by $u_k(x)$ and $v_k(x)$ upper and lower bounding Chaplygin's approximations order k, it holds ([1]):

(2) $$\max_{x\in[a,b]} | u_k(x) - v_k(x) | < \frac{C}{2^{2^k}} , \quad (C \in R^+).$$

In [4] and [6] some shortcomings and problems related to Chapligin's method are pointed out. However, if the discretization of Chaplygin's method is made successfully, there

..re some of cases where this method may be useful ([5]). For example, the error estimating in some of numerical methods for (1), may be based on Chaplygin's method.

2. DISCRETIZATION

If we introduce the following notation:

$$(3)\qquad p_k(x) = -\frac{\partial f(x,u_k(x))}{\partial y}$$

$$(4)\qquad q_k(x) = -\frac{f(x,v_k(x))-f(x,u_k(x))}{v_k(x)-u_k(x)}$$

and

$$(5)\quad I(a(x),b(x)) = \exp(-\int_{x_0}^{x} a(t)dt)(\ y_0 + \int_{x_0}^{x} (f(t,b(t)) + a(t)b(t))\exp(\ \int_{x_0}^{t} a(z)dz)dt),$$

then we have:

$$(6)\qquad u_{k+1}(x) = I(p_k(x),u_k(x))$$

$$(7)\qquad v_{k+1}(x) = I(q_k(x),v_k(x))\ .$$

The arising problem is to discretize the expression (5). Suppose that the values $u_k(x_i)$ and $v_k(x_i)$, ($x_i \in G_h$) are known. Performing the interpolation of functions $u_k(x)$ and $v_k(x)$ on the interval $[a,b]$ we get the polynomials $P_{u_k}(x)$ and $P_{v_k}(x)$. By using $P_{u_k}(x)$ and $P_{v_k}(x)$ we can calculate $u_k(x)$ and $v_k(x)$ with some accuracy, for each $x \in [a,b]$. This possibility allows using the numerous formulas for the numerical integration in (5). We want to know the truncation error made when the expression (5) is calculated. The following theorem is related to this problem.

THEOREM . If it holds:

(a) the values of the functions $a(x)$ and $b(x)$ in (5) are calculated with accuracy not lesser than $O(h^e)$, as $h \to 0$,

(b) each integral in (5) is calculated with the truncation error not greater than $O(h^s)$, as $h \to 0$,

then the expression $I(a(x),b(x))$ may be calculated on the set G_h with the accuracy $O(h^r)$, where $r=\min(e,s)$.

PROOF For $x_i \in G_h$ ($i = 1,\ldots,n$) we have:

$$(8)\qquad \int_{x_0}^{x_i} a(t)dt = S_1^h + R_1 + R_2$$

where R_1 is the roundoff error and R_2 is the truncation error of numerical integration. According to (b) we have:

$$(9)\qquad \int_{x_0}^{x_i} a(x)dx = S_1^h + O(h^r)$$

where $r = \min(e,s)$. From (9) we get:

$$(10)\qquad \exp(- \int_{x_0}^{x_i} a(x)dx) = C^h + O(h^r), \quad \text{as} \quad h \to 0.$$

Let be:

$$(11)\qquad g(x) = (f(x,b(x))+a(x)b(x))\exp(\int_{x_0}^{x} a(t)dt).$$

For the expression (5) we may write:

$$(12)\qquad I(a(x),b(x)) = \exp(- \int_{x_0}^{x} a(t)dt)(y_0 + \int_{x_0}^{x} g(t)dt).$$

By using (10) and (11) (for $x_i \in G_h$) we get:

$$g(x_j) = (f(x_j,b(x_j)) + a(x_j)b(x_j))(C_j^h + O(h^r))$$

and according to (a):

$$g(x_j) = g_j^h + O(h^r).$$

Now we obtain (like as in (9)):

$$(13)\qquad \int_{x_0}^{x_i} g(x)dx = S_2^h + R_1 + R_2$$

$$= S_2^h + O(h^r).$$

Finally from (12), (10) and (13) it follows that:

$$(14)\qquad I(a(x_i),b(x_i)) = C^h y_0 + C^h S_2^h + O(h^r)$$

$$= I^h + O(h^r)$$

for $x_i \in G_h$, as $h \to 0$ and the theorem is proved.

Denoting by u_{ki}^h and v_{ki}^h discrete Chaplygin's approximations in the point $x_i \in G_h$, from (6), (7) and (14) we have:

$$(15)\qquad u_k(x_i) = u_{ki}^h + O(h^r)$$

$$(16)\qquad v_k(x_i) = v_{ki}^h + O(h^r)$$

for k-th iteration. By using (2), (15) and (16) we make the following estimation:

$$|u_k(x_i)-v_k(x_i)| < |u_{ki}^h-v_{ki}^h| + O(h^r) + \frac{C}{2^{2^k}}$$

The error estimation for the solution $y(x)$ of (1) is based on the inequalities:

(17) $$u_k(x) < y(x) < v_k(x)$$

where $x \in [a,b]$ and $k=0,1,\dots$

3. NUMERICAL EXAMPLES

In the following examples we apply the results of the previous section. As polinomial $P_{u_k}(x)$ (i.e. $P_{v_k}(x)$) a cubic spline is used. Thus, the functions from (5) are calculated with the accuracy $O(h^4)$ (see [2]). We use Simpson's rule for numerical integration. Therefore, $R_1 = R_2 = O(h^4)$ in (8) and (13).

EXAMPLE 1. It is intended to solve the initial value problem

(18) $$y' = y^2 - y\sin x + \cos x, \qquad y(0) = 0,$$

by the previous method, using a steplength $h=0.1$ on the interval $[0,1]$. As the initial approximations we choose

$$\left.\begin{aligned} u_o(x_i) &= \sin x_i - 0.01i \\ v_o(x_i) &= \sin x_i + 0.01i. \end{aligned}\right. \qquad (i = 0,1,\dots,10)$$

(Similar results are obtained for $u_o(x_i) = \sin x_i - 0.1i$ and $v_o(x_i) = \sin x_i + 0.1i$, $i = 0,1,\dots,10$.) The results are presented in the table 1.

Table 1

x_i	$\bar{u}^h_{1i}$	$\bar{v}^h_{1i}$	$\bar{u}^h_{2i}$	$\bar{v}^h_{2i}$
0.1	0.09983016	0.09983686	0.09983352	0.09983291
0.2	0.19864265	0.19869641	0.19866954	0.19866895
0.3	0.29542913	0.29561211	0.29552049	0.29551989
0.4	0.38919984	0.38963886	0.38941868	0.38941806
0.5	0.47899265	0.47986351	0.47942589	0.47942527
0.6	0.56388187	0.56541504	0.56464278	0.56464220
0.7	0.64298689	0.64547439	0.64421786	0.64421744
0.8	0.71547964	0.71928375	0.71735594	0.71735601
0.9	0.78059514	0.78615534	0.78332613	0.78332729
1.0	0.83762731	0.84547896	0.84146881	0.84147252

The theoretical solution of (18) is $y(x)=\sin x$. The numerical results in table 1 are according to the theoretical consideration in the section 2.

EXAMPLE 2. Cosider the initial value problem:

(19) $y' = x^2 + y^2 - 32.1$, $y(0) = 0.1$,

on the interval $[0,1]$. Let be:

$$G_h = \left\{ x_i \mid x_i = 0.2i,\ i=0,\ldots,5 \right\}.$$

As the initial approximations we take:

$$u_o(x_i) = -2 \qquad (i=1,\ldots,5)$$
$$v_o(x_i) = -6.$$

In the table 2 the numerical results obtained in 5 iterations are presented. We give, also, results obtained by the method Runge-Kutta with the truncation error $O(h^5)$.

Table 2

x_i	$\bar{u}^h_{5i}$	$\bar{v}^h_{5i}$	Runge-Kutta
0.2	-4.6109031	-4.6105720	-4.39460...
0.4	-5.5755214	-5.5754871	-5.11619...
0.6	-5.6738443	-5.6738269	-5.39515...
0.8	-5.6625736	-5.6625541	-5.50713...
1.0	-5.6339439	-5.6339192	-5.53980...

The numerical results presented in this paper (and a lot of others numerical results) are obtained on the microcomputer COMODORE 64.

REFERENCES

1. Березин И.С, Жидков Н.П.: Методы вычислений 2, ФИЗМАТГИЗ, Москва, 1954
2. BURDEN R.L.,FAIRES J.,REYNOLDS A.C.: Numerical analysis, Prindle, Weber & Schmidt, Boston, 1981.
3. Чаплыгин С. А.: Изабранные труды по механике и математике, ГИЗ, Москва, 1954
4. TOŠIĆ D.: Some problems related to discretization of Chaplygin's method, Mat. vesnik, 35 (1983), 305 - 318.
5. VIDOSSICH G.: Chaplygin's Method is Newton's Method, Jour. of math. anlysis and appl. 66 (1978), 188 - 206.
6. MOORE R.E.: Two-sided approximation to solutions of nonlinear operator equations - a comparison of methods from classical analysis, functional analysis and interval analysis, Inter. Symposium, Karlsruhe, Springer-Verlag, Berlin,..., NewYork 1975, 31-47.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

ON A CONVERGENCE OF THE DIFFERENCE SCHEMES FOR THE EQUATION OF VIBRATING STRING

Boško S. Jovanović , Lav D. Ivanović

ABSTRACT:
In this note we inspect the convergence of the difference schemes for the equation of vibrating string, for the case when a generalized solution of the homogeneous boundary value problem belongs to a Sobolev-Slobodetsky space. The results for the elliptic and parabolic case are presented in [2, 3] .

O KONVERGENCIJI DIFERENCIJSKIH SHEMA ZA JEDNAČINU ŽICE KOJA TREPERI. U radu se ispituje konvergencija diferencijskih shema za jednačinu žice koja treperi, u slučaju kad generalisano rešenje konturnog problema pripada prostoru Soboljeva-Slobodeckog. Analogni rezultati za eliptički i parabolički slučaj dobijeni su u [2, 3] .

We will consider the first mixed homogeneous boundary value problem for the equation of vibrating string:

$$(1)\qquad \begin{aligned} &\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x,t) \quad , \qquad (x,t) \in Q = (0,1) \times (0,T] \, , \\ &u(0,t) = u(1,t) = 0 \quad , \qquad t \in (0,T] \, , \\ &u(x,0) = \frac{\partial u(x,0)}{\partial t} = 0 \quad , \qquad x \in [0,1] \, . \end{aligned}$$

Throught the note we will assume that the generalized solution of (1) belongs to a Sobolev-Slobodetsky space $W_2^s(Q)$, $1 \leq s \leq 4$, [4]. For such solutions one can construct a linear extension for $t < 0$ remaining in the same class [4]. By $\|\cdot\|_{s,Q}$ we will denote the norm and by $|\cdot|_{s,Q}$ the senior seminorm in $W_2^s(Q)$.

Pick a nonnegative integer n and let $h = 1/n$. We define a uniform grid ω_h with the step h over $(0,1)$. In the same way we define a uniform grid ω_τ with the step

$\tau = T/(m+0.5)$ over $(-0.5\tau, T]$ and put $Q_{h\tau} = \omega_h \times \omega_\tau$. We will assume that $c_1 h \leqslant \tau \leqslant c_2 h$, $c_1 , c_2 = \text{const} > 0$. If v is a function defined over $Q_{h\tau}$ by v^j we will denote its restriction for $t = (j-0.5)\tau$.

We introduce the difference operators v_x, $v_{\bar{x}}$, v_t and $v_{\bar{t}}$ in the standard way [6]. By $\|\cdot\|_h$ and $(\cdot,\cdot)_h$ we will denote the difference analogs of the norm and scalar product over $L_2(0,1)$. In the space of discrete functions, which are defined over ω_h and which are equal to zero over the boundary knots the operator:

$$\Lambda v = - v_{x\bar{x}}$$

is selfadjoint and positive definite. Therefore the norm

$$\|v\|_{\Lambda^{-1}} = (\Lambda^{-1} v , v)_h^{1/2}$$

can be defined. Also the norms over $Q_{h\tau}$ will be:

$$\|v\|^{(1)}_{2,\infty,h} = \max_k \left(\|v_t^k\|_h + 0.5 \, \|v_x^k + v_x^{k+1}\|_h\right) \quad ,$$

$$\|v\|^{(o)}_{2,\infty,h} = \max_k \left(\|v_t^k\|_{\Lambda^{-1}} + 0.5 \, \|v^k + v^{k+1}\|_h\right) \quad .$$

Due to the fact that $f(x,t)$ need not to be continuous, it seems natural to approximate $f(x,t)$ by some mean values over $Q_{h\tau}$. Let T be Steklov's mollifier defined by:

$$T\, g(x) = \int_{-0.5}^{0.5} g(x + h\, x')\, dx'$$

and $T^o g(x) = g(x)$, $T^k g(x) = T(T^{k-1} g(x))$, $k = 1,2,\ldots$ By $T^{k,r}$ we will denote the product of T^k over x, and T^r over t .

Then we will approximate (1) by a weighted difference scheme ($a = \text{const} > o$) :

$$\text{(2)} \qquad \begin{aligned} & v^j_{t\bar{t}} = a\, v^{j+1}_{x\bar{x}} + (1-2a)\, v^j_{x\bar{x}} + a\, v^{j-1}_{x\bar{x}} + T^{2,2} f^j , \\ & v^j = 0 \quad \text{for} \quad x = 0 \quad \text{and} \quad x = 1 , \\ & v^o = v^1 = 0 \quad . \end{aligned}$$

When $u \in W_2^s(Q)$, $2 \leqslant s \leqslant 4$, we will denote $z = = u - v$. Function z is well defined over $Q_{h\tau}$ satisfying:

$$z^j_{t\bar{t}} = a\, z^{j+1}_{x\bar{x}} + (1-2a)\, z^j_{x\bar{x}} + a\, z^{j-1}_{x\bar{x}} + \varphi^j_{x\bar{x}} + \psi^j_{x\bar{x}} \;,$$

$$z^j = 0 \quad \text{for} \quad x = 0 \quad \text{and} \quad x = 1 \;,$$

$$z^0 = u(x, -0.5\,\tau),$$

$$z^1 = u(x, 0.5\,\tau),$$

where

$$\varphi^j = T^{0,2}\, u^j - \; a\, u^{j+1} + (1-2a)\, u^j + a\, u^{j-1} \;,$$

$$\psi^j = u^j - T^{2,0}\, u^j \quad .$$

When $u \in W^s_2(Q)$, $1 \leqslant s \leqslant 3$, we will denote $\hat{z} = T^{2,0} u - v$. Furthermore, we will assume that the solution $u(x,t)$ can be extended outside Q so that the extension is odd over x, and remaining in the same class. The function $\hat{z}$ is defined over the grid $Q_{h\tau}$ satisfying:

$$\hat{z}^j_{t\bar{t}} = a\, \hat{z}^{j+1}_{x\bar{x}} + (1-2a)\, \hat{z}^j_{x\bar{x}} + a\, \hat{z}^{j-1}_{x\bar{x}} + \hat{\varphi}^j_{x\bar{x}} \;,$$

$$\hat{z}^j = 0 \quad \text{for} \quad x = 0 \quad \text{and} \quad x = 1 \;,$$

$$\hat{z}^0 = T^{2,0}\, u(x, -0.5\,\tau),$$

$$\hat{z}^1 = T^{2,0}\, u(x, 0.5\,\tau),$$

where

$$\hat{\varphi}^j = T^{0,2}\, u^j - T^{2,0}\left[\, a\, u^{j+1} + (1-2a)\, u^j + a\, u^{j-1} \right] .$$

Using the method of energy inequalities [6] one can prove the following a priori estimates:

$$(3)\quad \|z\|^{(1)}_{2,\infty,h} \leqslant C \left[\|z^0_t\|_h + \|z^0_x + z^1_x\|_h + \tau \sum_{j=1}^{m} \left(\|\varphi^j_{x\bar{x}}\|_h + \|\psi^j_{t\bar{t}}\|_h \right) \right]$$

$$(4)\quad \|\hat{z}\|^{(0)}_{2,\infty,h} \leqslant C \left(\|\hat{z}^0_t\|_{\Lambda^{-1}} + \|\hat{z}^0 + \hat{z}^1\|_h + \tau \sum_{j=1}^{m} \|\hat{\varphi}^j_x\|_h \right) \quad .$$

The convergence rate estimates in this note are based on the following generalization of the Branble-Hilbert lemma.

LEMMA: Let $[s]^-$ be a nonnegative integer, $[s]^- < s \leqslant [s]^- + 1$ and let $P_{[s]^-}$ be the set of polynomials of degree $\leqslant [s]^-$. If $\eta = \eta(u)$ is a bounded, linear functional over $W^s_2(Q)$ such that $P_{[s]^-} \subset \text{Kernel}(\eta)$, then for every $u \in W_2^s(Q)$ the following inequality is valid:

$$|\eta(u)| \leq C \; |u|_{s,Q} \qquad , \qquad C = C(Q,s) = \text{const} \; .$$

The proof of lemma follows from the Dupont-Scott theoreme [1].

Functionals $\varphi_{x\bar{x}}$, $\psi_{t\bar{t}}$, z_t^0 and $(z_x^0 + z_x^1)$ are bounded and linear over $W_2^s(Q)$ for $s > 2$ while $P_3 \subset \subset \text{Kernel}(\varphi_{x\bar{x}})$, $P_3 \subset \text{Kernel}(\psi_{t\bar{t}})$, $P_2 \subset \text{Kernel}(z_t^0)$, $P_2 \subset \text{Kernel}(z_x^0 + z_x^1)$. Using the lemma one obtains the following estimates:

$$(5) \qquad \tau \sum_{j=1}^{m} \left(\| \varphi^j_{x\bar{x}} \|_h + \| \psi^j_{t\bar{t}} \|_h \right) \leq C \; h^{s-2} \, |u|_{s,Q} \; , \quad 2 < s \leq 4$$

$$(6) \qquad \| z_t^0 \|_h + \| z_x^0 + z_x^1 \|_h \leq C \; h^{s-1.5} \, |u|_{s,Q_\tau} \; , \quad 2 < s \leq 3 ,$$

where $Q_\tau = (0,1) \times (-0.5\,\tau, 0.5\,\tau)$. Using (3),(5),(6) and

$$(7) \qquad |u|_{k,Q_\tau} \leq C \; F(h,a) \, \| u \|_{k+a,Q} \quad , \quad k = 0,1,2, \ldots ,$$

where

$$F(h,a) = \begin{cases} h^a & , \quad 0 \leq a < 0.5 , \\ h^{0.5} \, |\ln h| & , \quad a = 0.5 , \\ h^{0.5} & , \quad 0.5 < a \leq 1 \end{cases}$$

(see [5]) one obtains the following convergence rate estimate for the difference scheme (2) :

$$\| z \|^{(1)}_{2,\infty,h} \leq C \; h^{s-2} \, \| u \|_{s,Q} \qquad , \qquad 2 < s \leq 4 \quad .$$

Similarly from the lemma and from (7) one obtains:

$$(8) \quad \| \hat{z}^0 + \hat{z}^1 \|_h + \tau \sum_{j=1}^{m} \| \hat{\varphi}^j_x \|_h \leq C \; h^{s-1} \, \| u \|_{s,Q} \quad , \quad 1 < s \leq 3 ,$$

$$(9) \quad \| \hat{z}^0_t \|_{\Lambda^{-1}} \leq C \; h^{s-1.5} \, |w|_{s,Q_\tau} \quad , \quad 2 \leq s \leq 3 ,$$

where

$$w(x,t) = \int_0^x u(x',t) \, dx' - \int_0^1 \int_0^{x''} u(x',t) \, dx' \, dx'' .$$

From (4),(8) and (9) expressing w by u , one obtains the following convergence rate estimate for difference scheme

(2) :

$$\|\hat{z}\|^{(o)}_{2,\infty,h} \leqslant C\, h^{s-1} \left(\|u\|_{s,Q} + \|f\|_{p(s),s-1,Q}\right) \quad ,$$

$1 < s \leqslant 3$, where $p(s) = \max\{0, s-2\}$ and $\|f\|_{p,q,Q}$ the norm of the anisotropic Sobolev-Slobodetsky space $W_2^{p,q}(Q) = L_2(0,T; W_2^p(0,1)) \cap W_2^q(0,T; L_2(0,1))$ (see [4]).

REFERENCES

1. DUPONT T., SCOTT R.: Polynomial approximation of functions in Sobolev spaces. Math. Comput. 34 (1980), 441 - 463.
2. IVANOVIĆ L. D., JOVANOVIĆ B. S., SÜLI E. E. : On the convergence of difference schemes for the Poisson's equation. IV seminar primijenjene matematike, Split 1984.
3. JOVANOVIĆ B. S., IVANOVIĆ L. D., SÜLI E. E. : On the rate of convergence of difference schemes for heat transfer equation. IV seminar primijenjene matematike, Split 1984.
4. LIONS J. L., MAGENES E. : Problèmes aux limites non homogènes et applications I,II. Dunod, Paris 1968.
5. OGANESYAN L. A., RUHOVEC L. A. : Variacionno-raznostnye metody rešeniya elliptičeskih uravnenij. AN ArmSSR, Erevan 1979.
6. SAMARSKIJ A. A. : Teoriya raznostnyh shem. Nauka, Moskva 1983.

Department of Mathematics
Faculty of Sciences
11 000 Belgrade, Studentski trg 16
Yugoslavia

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

APPROXIMATION AND REGULARIZATION OF CONTROL PROBLEM GOVERNED BY PARABOLIC EQUATION

Lav D. Ivanović , Boško S. Jovanović

ABSTRACT:

A finite dimensional approximation of distributed control problem governed by the heat transfer equation is considered. We prove that a sequence of finite dimensional problem solution converge to the original solution. Also, we construct a minimizing sequence which converge to the optimal control.

APROKSIMACIJA I REGULARIZACIJA PROBLEMA OPTIMALNOG UPRAVLJANJA SISTEMIMA PARABOLIČKOG TIPA. U ovom radu razmatra se konačno dimanziona aproksimacija zadatka optimalnog upravljanja sistemom opisanim jednačinom provodjenja toplote. Dokazuje se da niz konačno dimenzionih rešenja konvergira ka originalnom rešenju i vrši se regularizacija tj. konstruiše se minimizirajući niz koji konvergira ka optimalnom upravljanju.

We shall consider the following optimal control problem:

(1) $$J(v)=\int_{Q_T} f(u(x,t),u_x(x,t))dxdt \longrightarrow \inf_U$$

(2) $$u_t=\Delta u+v_o \quad , \quad (x,t)\in Q_T=(0,1)^2 \times (0,T]$$

(3) $$u(x,t)=0 \quad , \quad (x,t)\in \Omega \times (0,T] \quad , \quad \Omega=(0,1)^2$$

(4) $$u(x,0)=v_1(x) \quad , \quad x\in\Omega \quad .$$

It is well known [5] that if $v_o\in H^{2r,r}(Q_T)$, $v_1\in H^{2r+1}(\Omega)$ than exists a unique solution of (2),(3),(4) $u\in H^{2(r+1),r+1}(Q_T)$ for $r\geqslant 0$ and the following estimate valid:

$$(5)\qquad \|u\|_{H^{2r+2,r+1}(Q_T)} \leqslant C(\|v_o\|_{H^{2r,r}(Q_T)} + \|v_1\|_{H^{2r+1}(\Omega)}).$$

We shall denote $v=(v_o,v_1)$ and $U=\{ v\in X_r=H^{2r,r}(Q_T)\times H^{2r+1}(\Omega): \|v\|_{X_r}\leqslant R$. Throughout the note we shall assume that f is a convex function and

$$(6)\qquad |f(a_o,a_1,a_2)-f(b_o,b_1,b_2)| \leqslant g(a_o,b_o)\sum_{i=o}^{2}|a_i-b_i|$$

where $g(a_o,b_o)$ is a positive bounded function over bounded sets.

Lemma. The solution of (1)-(4) exists.

Proof. it is easy to show that $J(u)$ is lower weakly semicontinuous function over U and U is a weakly compact set in X_r. By [7 p.47] follows the lemma.

To solve the problem (1)-(4) we shall construct a sequence of finite dimensional problems of nonlinear programming [3], [7] .

Let Ω_h be a uniform grid with the step $h=1/n$ over Ω and let ω_τ be a uniform grid over $(0,T]$ with the step $\tau = T/m$. In this note we shall assume that constants $c_1,c_2\geqslant 0$ exists such that $c_1h^2\leqslant\tau\leqslant c_2h^2$.

In the set of discrete functions over $Q_{h\tau}=\Omega_h\times\omega_\tau$ we shall introduce the following norms:

$$\|y\|_{o,o,h}=(\tau\sum_{j=o}^{m-1} o.25\,\|y^{j+1}+y^j\|_h^2)^{1/2}$$

$$\|y\|_{1,o,h}=(\tau\sum_{j=o}^{m-1}\sum_{i+1}^{2} o.25\,\|(y^{j+1}+y^j)_{x_i}\|_h^2)^{1/2}$$

$$\|y\|_{2,1,h}=(\tau\sum_{j=o}^{m-1}\|y_t\|_h^2 + o.25\sum_{i=1}^{2}\|(y^{j+1}+y^j)_{x_i\bar{x}_i}\|_h^2 +$$

$$o.25\,\|(y^{j+1}+y^j)_{x_1x_2}\|_h^2)^{1/2}$$

here $y^j = y/_{t=j\tau}$. We used the standard notation [6].

Let T be a Steklov mollifier [4], [6]. By T^{m_1,m_2} we shall denote the product of T^{m_1} over x_1 and T^{m_2} over x_2. By T^{m_1,m_2,m_3} we shall denote the product of T^{m_1} over x_1, T^{m_2} over x_2 and T^{m_3} over t [4].

The problem (2)-(4) will be approximated by the difference scheme of alternating directions [4], [6]:

$$(7) \qquad L_1y=(y^{j+0.5}-y^j)/0.5\tau - y^{j+0.5}_{x_1\bar{x}_1} - y^j_{x_2\bar{x}_2} = \hat{v}_0^{j+0.5}$$

$$(8) \qquad L_2y\pm(y^{j+1}+y^{j+0.5})/0.5\tau - y^{j+0.5}_{x_1\bar{x}_1} - y^{j+1}_{x_2\bar{x}_2} = \hat{v}_0^{j+0.5}$$

$$(9) \qquad y^0 = \hat{v}_1$$

where $y^{j+0.5}$ denotes the value of y on the auxiliary time slice $t=(j+0.5)\tau$, $\hat{v}_0^{j+0.5}=(T^{2,2,1}v_0)^{j+0.5}$ and $\hat{v}_1 = T^{2,2}v_1$.

Useing (5) and the discrete solution estimates [4] from Dupont-Scott theorem [2] follows :

$$(10) \qquad \|z\|_{2,1,h} \leqslant C\, h^{2r} \|v\|_{X_r} \qquad 0<r\leqslant 1$$

$$(11) \qquad \|z\|_{1,0,h} \leqslant C\, h^{2r+1} \|v\|_{X_r} \qquad 0<r\leqslant 0.5$$

$$(12) \qquad \|z\|_{0,0,h} \leqslant C\, h^2 \|v\|_{X_0} \quad .$$

The cost function J(u) will be approximated by the following equation :

$$I_n(y) = \sum_{Q_{h\tau}} \gamma_\kappa h^2 \tau f(y, y_{x_1}, y_{x_2}) \quad .$$

Now we can formulate the sequence of finite dimensional problems :

(13) $$I_n \longrightarrow \inf_{W_n^{p,q,s}}$$

(14) $$L_1 y = w_o \quad , L_2 y = w_o \quad , \quad y^o = w_1$$

where the operators L_1, L_2 are defined by (7),(8) and the set $W_n^{p,q,s} = \left\{ w=(w_o, w_1) \in Y_n = H_n^{p,q} X \, H_n^s \; : \; \|w\|_{Y_n} \leqslant R \right\}$.

We denoted by $H_n^{p,q}, H_n^s$ the discrete analogues of the Sobolev spaces $H^{p,q}(Q_T), H^s(\Omega)$ $p,q,s \in N \cup \{0\}$.

Since (13),(14) is the mathematical programming problem one can prove that a solution y_n^* of (13),(14) exists.

Furthermore we shall construct operators $Q_n : X_r \rightarrow Y_n$ and $P_n : Y_n \rightarrow X_r$ by

(15) $$Q_n(v) = (Q_n^o(v_o) \; , Q_n^1(v_1)) = (T^{2,2,1} v_o , T^{2,2} \; v_1 \;)$$

(16) $$P_n \; (w) = (\; \bar{w}_o \; , \; \bar{w}_1)$$

where $\bar{w}_o \in H^{2,1}(Q_T)$ and $\bar{w}_1 \in H^3(\Omega)$ are the interpolants defined in [1] .

Theorem 1. If the above assumptions are valid than

$$\lim_{n \to \infty} I_n^* = J_* = \inf_U J(u) \text{ and for } o < r \leqslant o.5$$

(17) $$| I_n^* - J_* | \leqslant C \; h^{2r+1} \quad .$$

Proof. Useing the technique developed in [7] from (11),(15),(16) one can prove that the conditions of theorem 3. [7 p.311] are satisfied so the theorem 1. follows.

Now, we shall introduce Tichonov functional [7] as:

$$T_n(w) = I_n(w) + \alpha_n \|w\|^2_{Y_n} \; .$$

Let w_n^* be a sequence of discrete controls such hat :

$$T_n^* = \inf_{W_n} T_n(w) \leq T_n(w_n^*) \leq T_n^* + \mu_n .$$

Sequences α_n and μ_n are positive and $\lim_{n\to\infty} \alpha_n = = \lim_{n\to\infty} \mu_n = 0$.

Theorem 2. If the theorem 1. is valid and if

$$(18) \qquad \lim_{n\to\infty} (h_n^{2r+1} + \mu_n)/\alpha_n = 0 \quad \text{for } 0 < r \leq 0.5$$

than

$$\lim_{n\text{---}} J(P_n(w_n^*)) = J_* \quad \text{and} \quad \lim_{n\to\infty} \| P_n(w_n^*) - v^* \|_{X_r} = 0$$

where $J(v^*) = J_*$.

Proof. Useing the same technique as in [7] from the estimate (11) follows the theorem.

Remark. If the cost functional is of the form

$$J(u) = \int_{Q_T} f(u, u_x, u_{xx}, u_t) dxdt$$

than in (17),(18) $2r+1$ must be replaced by $2r$. If the cost functional is

$$J(u) = \int_{Q_T} f(u) dx \, dt$$

and if $r=o$ than in (17),(18) $2r+1$ must be replaced by 2 .

REFERENCES

1. CIARLET P.: The finite element method for elliptic problems. North-Holland, Amsterdam 1978.
2. DUPONT T., SCOTT R.: Polynomial approximation of functions in Sobolev spaces. Math. Comput. 34 (198o), 441-463.
3. EVTUSHENKO Yu.: Metody resheniya ekstremaljnyh zadach i ih primenjenye v sistemah optimizacii. Nauka ,Moskva 1982.
4. JOVANOVIĆ B.S., IVANOVIĆ L.D., SÜLI E.E.: On the rate of convergence of difference schemes for heat transfer equation . IV seminar primijenjene matematike, Split 1984.
5. LIONS J.L., MAGENES E.: Non-Homogeneous Boundary Value Problems and Applications. Springer-Verlag, Berlin 1972.
6. SAMARSKIJ A.A.: Teoriya raznostnyh shem. Nauka , Moskva 1983.
7. VASILJEV F.P.: Metody resheniya ekstremaljnyh zadach. Nauka, Moskva 1981.

Department of Mathematics
Faculty of Sciences
11 ooo Belgrade, Studentski trg 16
Yugoslavia

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

SOLUTIONS OF THE GRID LAPLACE EQUATION DEFINED IN CORNERS

Desanka P. Radunović

ABSTRACT:

Solutions of difference schemes, defined on the rectangular grid for Dirichlet and mixed boundary problems for the Laplace equation in corners $3\pi/2$ and 2π are obtained. From their asymptotic expansions it can be seen that the orders of errors are $O(h^{2/\nu}/|z|^{1/\nu})$, $\nu=3/2,2,3,4$, and that in some cases the accuracy can be improved by the appropriate choise of the grid parameters.

REŠENJA MREŽNE JEDNAČINE LAPLACEA DEFINISANE U UGLOVIMA. Odredjuju se rešenja diferencijskih shema, definisanih na pravougaonoj mreži, kojima se aproksimiraju Dirichletov i mešoviti granični zadatak za jednačinu Laplacea u uglovima $3\pi/2$ i 2π. Iz asimptotskih razvoja dobijenih rešenja sledi da je red greške aproksimacije $O(h^{2/\nu}/|z|^{1/\nu})$, $\nu=3/2,2,3,4$, i da se u izvesnim slučajevima odgovarajućim izborom parametara mreže može povećati tačnost aproksimacije.

In this paper, in a context of studies of an accuracy of classic difference approximations of nonsmooth solutions of boundary problems for differential equations, we obtain solutions and their asymptotic expansions of difference problems that approximate on the rectangular grid

$$\Omega = \{ (x,y) \mid x=mh,\ y=nh',\ m,n \in Z,\ h'=\theta h,\ h,\theta>0 \}$$

following problems: to find the continuous function v, not identically equal to zero, harmonic in the corner $0<\varphi<\nu\pi$, $0<r<\infty$ ((r,φ) polar coordinates), $\nu=3/2,2,3,4$, equal to zero on the positive part of the x-axis and on the line $\nu\pi$, and which does not grow too rapidly at the infinity, i.e.

$$\lim_{r\to\infty} \frac{v}{r^{2/\nu}} = 0.$$

Solutions of these problems, for corresponding ν, are (from [2])

$$v = C\,\mathrm{Im}\,z^{1/\nu} = C\,r^{1/\nu}\sin\frac{\varphi}{\nu}, \qquad C = \mathrm{const},\quad z = re^{i\varphi},$$

so, their first derivatives have integrable singularities at the origin. For $\nu=3$ the initial problem can be replaced by the equivalent one, defined in the corner $0 \leqslant \varphi \leqslant 3\pi/2$, $0 \leqslant r < \infty$ with the boundary condition $\frac{\partial v}{\partial x} = 0$ on the line $\varphi = 3\pi/2$. Similarly, for $\nu = 4$ the initial problem can be replaced by the equivalent one defined in the plane with a crack, where the crack is on the positive part of the x-axis and at the lower edge of the crack the boundary condition $\frac{\partial v}{\partial y} = 0$ is given.

Let us define the one parameter difference operator family in order to approximate the Laplace operator (see [3])

$$\Lambda_\alpha u \equiv u_{\bar{x}x} + u_{\bar{y}y} - \alpha h^2 \frac{1+\theta^2}{2} u_{\bar{x}x\bar{y}y}, \qquad \alpha > -1/2,$$

and difference operator families in order to approximate the boundary conditions of the second type

$$\lambda_{3/2} u \equiv u_{\bar{x}} - \frac{h}{2} u_{\bar{y}y} - \alpha h^2 \frac{1+\theta^2}{2} u_{\bar{x}\bar{y}y} \qquad \text{on the line } 3\pi/2,$$

and

$$\lambda_2 u \equiv u_{\bar{y}} - \frac{\theta h}{2} u_{\bar{x}x} - \alpha h^2 \frac{1+\theta^2}{2} u_{\bar{x}x\bar{y}} \qquad \text{on the line } 2\pi$$

(for $\alpha > -1/2$ Λ_α is the elliptic operator, [4]). Let us denote

$$\Omega_\nu = \{(x,y) \mid (x,y) \in \Omega,\ 0 < \varphi < \nu\pi,\ 0 < r < \infty\},$$

$$\Gamma_\nu = \{(x,y) \mid (x,y) \in \Omega,\ \varphi = \nu\pi,\ 0 < r < \infty\},$$

$$\Gamma_0 = \{(x,y) \mid (x,y) \in \Omega,\ x>0,\ y = +0\}, \qquad \bar{\Gamma}_0 = \Gamma_0 \cup \{(0,0)\},$$

$$\Gamma_2 = \{(x,y) \mid (x,y) \in \Omega,\ x>0,\ y = -0\}.$$

Solutions of the following problems will be determined:

PROBLEM 1. ($\nu = 3/2$) $\quad \Lambda_\alpha u = 0$, $(x,y) \in \Omega_{3/2}$, $\lim\limits_{r\to\infty} u/r^{4/3} = 0$,

$u = 0$, $(x,y) \in \bar{\Gamma}_0$, $\quad u = 0$, $(x,y) \in \Gamma_{3/2}$, $\quad u(-h,0) = A$.

PROBLEM 2. ($\nu = 2$) $\quad \Lambda_\alpha u = 0$, $(x,y) \in \Omega_2$, $\lim\limits_{r\to\infty} u/r = 0$,

$u = 0$, $(x,y) \in \bar{\Gamma}_0$, $\quad u = 0$, $(x,y) \in \Gamma_2$, $\quad u(-h,0) = A$.

PROBLEM 3. ($\nu = 3$) $\quad \Lambda_\alpha u = 0$, $(x,y) \in \Omega_{3/2}$, $\lim\limits_{r\to\infty} u/r^{2/3} = 0$,

$u = 0$, $(x,y) \in \bar{\Gamma}_0$, $\quad \lambda_{3/2} u = 0$, $(x,y) \in \Gamma_{3/2}$, $\quad u(-h,0) = A$.

PROBLEM 4. ($\nu = 4$) $\quad \Lambda_\alpha u = 0$, $(x,y) \in \Omega_2$, $\lim\limits_{r\to\infty} u/r^{1/2} = 0$,

$u = 0$, $(x,y) \in \bar{\Gamma}_0$, $\quad \lambda_2 u = 0$, $(x,y) \in \Gamma_2$, $\quad u(-h,0) = A$.

THEOREM 1. The solutions of the problems 1-4, for corresponding ν, are

$$u(x_m, y_n) = \begin{cases} \dfrac{A}{2\pi F(-z_1)} \displaystyle\int_0^{2\pi} \frac{e^{i\xi}[q^n(\xi)\, e^{im\xi} - 1]}{(1-e^{i\xi})^{1+1/\nu}(1-z_1 e^{i\xi})^{1-1/\nu}} F\Big(\frac{e^{i\xi}-z_1}{1-z_1 e^{i\xi}}\Big)\, d\xi, & \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n>0, & \\ u(x_m, y_{|n|}) + 2\cos\frac{\pi}{\nu} u(x_{-m}, y_{|n|}), \quad n<0, & \end{cases}$$

ıere $m = 0, \pm1, \pm2, \ldots,$

?) $F(z) = {}_2F_1\left(\frac{1}{2} - \frac{1}{\nu},\ 1 - \frac{1}{\nu};\ \frac{3}{2};\ z\right)$

3 a hypergeometric function,

3) $z_1 = \{\sqrt{[(1+\theta^2)(1+2\alpha)]} - 1\}/\{\sqrt{[(1+\theta^2)(1+2\alpha)]} + 1\},$

ıd

4) $$q(\xi) = \frac{\sqrt{[\cos^2\frac{\xi}{2} + (1+\theta^2)(1+2\alpha)\sin^2\frac{\xi}{2}]} - \theta\sin\frac{\xi}{2}}{\sqrt{[\cos^2\frac{\xi}{2} + (1+\theta^2)(1+2\alpha)\sin^2\frac{\xi}{2}]} + \theta\sin\frac{\xi}{2}}\ .$$

The proof is similar to the proof of the corresponding theorem for square grid given in [1]. First we define problems 1. and 3. for $3\pi/2 < \varphi < 2\pi$ by appropriate transformations. Then, making use of the discrete Fourier transformation over the argument x, and solving the difference equation for the argument y, we obtain the Fourier image of the solution. If we now apply the inverse Fourier transformation, we obtain the solution in the form

$$(5) \quad u(x_m, y_n) = \frac{1}{2\pi}\begin{cases} \displaystyle\int_0^{2\pi} \tilde{u}_0(\xi)[q^n e^{im\xi} - 1]\, d\xi, & m=0,\pm1,\pm2,\ldots,\ n=0,1,2,\ldots, \\ \displaystyle\int_0^{2\pi} \tilde{v}_0(\xi)[q^{|n|} e^{im\xi} - 1]\, d\xi, & m=0,\pm1,\pm2,\ldots,\ n=-0,-1,-2,\ldots \end{cases}$$

(for $n=+0$ and $n=-0$ we have different forms of the solution, as there is a crack on the positive part of the x-axis). $\tilde{u}_0$ and $\tilde{v}_0$ are the Fourier images of the traces of the solution on the lines $y=+0$ and $y=-0$, and $q(\xi)$ is the function given by (4). If we demand, for every defined problem, that the function (5) satisfies boundary conditions, we obtain the problem of coupling analitic functions on the unit circle in the complex plane. This problem can be reduced to the singular integral equation

$$\int_0^1 \frac{\psi(t)}{t-x}\, dt + \cos\frac{\pi}{\nu} \int_0^1 \frac{t\,\psi(t)}{1-xt}\, dt = 0,$$

and its solution is

$$\psi(x) = C\, x^{-1\;2}(1-x)^{-1/\nu}\, {}_2F_1\left(\frac{1}{2} - \frac{1}{\nu},\ 1 - \frac{1}{\nu};\ \frac{3}{2};\ x\right), \qquad C = \text{const.}$$

Returning back to initial variables and using the given condition at the point (-h,0) to determine the constant C, we obtain the solution (1).

THEOREM 2. The asymptotic expansions of the solutions of the problems 1-4, for corresponding ν, are

$$u(x_m,y_n) = -\frac{A\nu(1+z_1)^{1/\nu}}{4\pi(1-z_1)F(-z_1)h^{1/\nu}} \operatorname{Im}\{2^{2/\nu}[(1+\theta^2)(1+2\alpha)]^{-1/(2\nu)}\Gamma(-\frac{1}{\nu})z^{1/\nu}$$

$$(6) \qquad -2^{-2/\nu}[(1+\theta^2)(1+2\alpha)]^{1/(2\nu)}\Gamma(\frac{1}{\nu})\frac{h^{2/\nu}}{z^{1/\nu}} + 2^{-2(2-1/\nu)}\frac{2-\frac{1}{\nu}}{3}\Gamma(2-\frac{1}{\nu})[(1+\theta^2)$$

$$(1+2\alpha)]^{-1/(2\nu)}[(1+\theta^2)(1+6\alpha)\frac{\bar{\beta}}{\beta} + 2(\theta^2-1)\frac{h^2}{z^{2-1/\nu}}\} + O(\frac{h^{2+2/\nu}}{|z|^{2+1/\nu}}),$$

where $F(z)$ and z_1 are given by (2) and (3), and $z = x_m + iy_n = h\beta$.

For the proof of this theorem, we obtain more convenient expression of the solution (1) with substitutions $\zeta = e^{i\xi}$ and $z = (\zeta - z_1)/(1 - z_1\zeta)$

$$(7) \qquad u(x_m,y_n) = -\frac{A\nu(1+z_1)^{1/\nu}}{\pi(1-z_1)F(-z_1)} \operatorname{Im} \int\limits_{\substack{|z|=1 \\ \operatorname{Im} z>0}} (1-z)^{-1/\nu} d[F(z)\ddot{q}^n(z)(\frac{z+z_1}{1+z_1z})^m],$$

where $\ddot{q}(z) = \hat{q}(\zeta) = q(\xi)$ and $F(z) = F[(e^{i\xi} - z_1)/(1 - z_1e^{i\xi})]$. The proof is realised in two steps. In the first step we prove that the essential contribution to the integral (7) is given by the integral in the close neighbourhood of the point $z=1$. The contour of integration is destorted and by estimating the integral for separate parts of the contour, we obtain

$$(8) \qquad u(x_m,y_n) = -\frac{A\nu(1+z)^{1/\nu}}{\pi(1-z_1)F(-z_1)} \operatorname{Im} \int\limits_C (1-z)^{-1/\nu} d[F(z)e^{-w(z)\beta}] + O(|\beta|^{-N}),$$

for any $N>0$. $v(z) = w(z)\beta$ is defined by the expresion

$$e^{-v(z)} = \ddot{q}^n(z)(\frac{z+z_1}{1+z_1z})^m.$$

The curve C is $C = \{z \mid \operatorname{Im} v(z) = 0\}$ for $z \in \{z \mid |z-1| < \delta, \operatorname{Im} z > 0\}$, i.e. C is contained in the close neighbourhood of the point $z=1$.

In the second step of the proof, the function in the integral is approximated by the partial sum of the asymptotic series written for $z=1$, and obtained integral is calculated analitically. With regard to the features of the function w, its inverse function $z = z(w)$ for $|w| < \delta_1$ exists. So, in the integral (8) we can use the new argument of tion w. We suppose some series representations as functions of w r $(1-z)^{-1/\nu}$ and F(z). and after some estimates we have

$$u(x_m,y_n) \sim \frac{A\nu(1+z_1)^{1/\nu}}{\pi(1-z_1)F(-z_1)} \operatorname{Im}\{\sum_{k=0}^{\infty}[p_k(\nu)\Gamma(k-\frac{1}{\nu})\beta^{1/\nu}-p_k(-\nu)\Gamma(k+\frac{1}{\nu})\beta^{-1/\nu}]\beta^{-k}\}, \tag{9}$$

here the coefficients $p_k(\nu)$ are determined by the series expansion of he function

$$P(w,\nu) = w^{1+1/\nu}\frac{1}{4(1-z)\sqrt{z}}\left(\frac{1+\sqrt{z}}{1-\sqrt{z}}\right)^{1/\nu}\frac{dz}{dw} \equiv \sum_{k=0}^{\infty} p_k w^k .$$

e express z by w and obtain

$$p_0(\nu) = -2^{-2(1-1/\nu)}\left(\frac{1-z_1}{1+z_1}\right)^{1/\nu},$$

$$p_2(\nu) = -2^{-2(3-1/\nu)}\left(\frac{1-z_1}{1+z_1}\right)^{1/\nu}\left(2-\frac{1}{\nu}\right)\frac{1}{3}[(1+\theta^2)(1+6\alpha)\frac{\bar{\beta}}{\beta}+2(\theta^2-1)],$$

....

$$p_{2k+1}(\nu) = 0, \quad k=0,1,2,\dots .$$

If we put these obtained coefficients in (9), as $z=h\beta$, we have (6).

The difference problem 2. we can define on the displaced rectangular grid

$$\Omega_\varepsilon = \{(x,y)\,|\,x=(m+\varepsilon)h,\ y=nh',\quad m,n\in Z,\ h'=\theta h,\ h,\theta>0\}.$$

Its solution is given also by the expression (1) and its asymptotic expansion is determined by the following theorem:

THEOREM 3. The asymptotic expansion of the solution of the problem 2., defined on the grid Ω_ε, is

$$u(x_m,y_n) = \frac{2A}{\sqrt{[\pi(1-z_1)h]}}\operatorname{Im}\{z^{1/2}-\frac{1}{2}(\varepsilon-\frac{\sqrt{[(1+\theta^2)(1+2\alpha)]}}{4})\frac{h}{z^{1/2}} - \tag{10}$$
$$-\frac{1}{2^6}[8\varepsilon^2-4\varepsilon\sqrt{[(1+\theta^2)(1+2\alpha)]}+(1+\theta^2)(1+6\alpha)\frac{\bar{\beta}}{2\beta}+\theta^2-1]\frac{h^2}{z^{3/2}}\}+O(\frac{h^3}{|z|^{5/2}}),$$

where $z=h(m+\varepsilon+in\theta)=h\beta_\varepsilon$.

For the proof of the theorem we put $\nu=2$ and replace β with β_ε in the expression (6), where $\beta=\beta_\varepsilon(1-\varepsilon/\beta_\varepsilon)$.

From the asymptotic expansion (6) we can conclude that the accuracy of the approximation is lower when the boundary corner is greater (it depends on ν), and that difference schemes of the higher order accuracy do not provide better approximations. With the choise of α, θ and ε such as $\varepsilon = \sqrt{[(1+\theta^2)(1+2\alpha)]/4}$, the problem 2. can be approximated in such a way that the order of an error is $O(h^2)$ for $|z| = O(1)$. For the scheme with $\alpha = -1/6$, the choise $\theta = \sqrt{2}$ and $\varepsilon = 1/(2\sqrt{2})$ provides the order of the accuracy $O(h^3)$.

REFERENCES:

1. АНДРЕЕВ В.Б.: Сеточные аппроксимации негладких решений дифференциальных уравнений, Дис.докт.физ.-мат.наук, Москва, 1983.

2. FOX L.,SANKAR R.: Boundary singularities in linear elliptic differential equations, J.Inst.Maths.Applics., 5(1969), 340-350.

3. САМАРСКИ А.А.,АНДРЕЕВ В.Б.: Разностные методы для эллиптических уравнений, Наука, Москва, 1976.

4. THOMEE V.,WESTERGREN B.: Elliptic difference equations and interior regularity, Numer.Math., 11(1968), 196-210.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

CONNECTION BETWEEN ONE PROBLEM IN ELASTICITY THEORY AND THE METHOD OF APPROXIMATE SOLVING OF CARLEMANN'S BOUNDARY VALUE PROBLEM

Miloš S. Čanak

ABSTRACT:

In this paper we consider one problem in elasticity theory which appears to be Carlemann's boundary value problem for analitic functions. For approximate solution of Carlemann's boundary value problem we take the exact solution of corresponding approximate problem and, after that, we estimate the error.

VEZA IZMEDJU JEDNOG PROBLEMA TEORIJE ELASTIČNOSTI I PRIBLIŽNOG REŠAVANJA GRANIČNOG PROBLEMA CARLEMANN-A: Razmatra se jedan problem teorije elastičnosti koji se svodi na granični problem Carlemann-a za analitičke funkcije. Za približno rešenje problema Carlemann-a uzima se tačno rešenje odgovarajućeg približnog problema, a zatim se daje ocena greške.

In this paper we consider the following problem: Find such solution of biharmonic equation

(1) $\Delta^2 u = 0 \quad , \quad y > 0$

which satisfies the following conditions

(2) $u(x,0) = 0 \quad , \quad -\infty < x < \infty$

(3) $u_y(x,0) - \sigma(x) u_{yy}(x,0) = h(x) \quad , \quad -\infty < x < \infty$

and which is bounded when $y \to \infty$,if $h(x)$ is a given continuous function and

$$\sigma(x) = \frac{ae^{-x}+b}{e^{-x}+c} \quad , \quad /a,b,c - \text{const.}/ \quad .$$

In this case the function $u(x,y)$ represents the displacement from the equilibrium position of elastic plate which covers a halfplane and which is fixed along the line $y=0$ by elastic hinge with a variable stiffness.

If we apply on biharmonic equation

$$(4) \qquad u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0$$

the Fourier transformation

$$(5) \qquad \mathcal{F}\left\{\frac{\partial^{p+q}u}{\partial x^p \partial y^q}\right\} = (-ix)^p \frac{d^q}{dy^q} U(x,y)$$

we get the ordinary differential equation

$$(6) \qquad U_{yyyy} - 2x^2 U_{yy} + x^4 U = 0$$

with general solution

$$(7) \qquad U(x,y) = c_1(x)e^{xy} + c_2(x)ye^{xy} + c_3(x)e^{-xy} + c_4(x)ye^{-xy} \quad .$$

But,if we looking for the particular solution which satisfies the condition $U(x,0)=0$ and which is bounded when $y \to \infty$,we have that

$$(8) \qquad U(x,y) = yC(x)e^{-|x|y} \quad ,$$

$$C(x) = \begin{cases} c_4(x) \ , & x > 0 \\ c_2(x) \ , & x < 0 \end{cases} \quad , \qquad c_2(0) + c_4(0) = 0 \quad .$$

If we substitute the value of $\sigma(x)$ in the condition (3) we get that

$$e^{-x}u_y(x,0) + cu_y(x,0) - ae^{-x}u_{yy}(x,0) - bu_{yy}(x,0) = e^{-x}h(x) + ch(x)$$

or,if we introduce the notation

9) $u_y(x,0)-au_{yy}(x,0)-h(x) = \varphi(x)$

n the condensed form

(10) $cu_y(x,0)-bu_{yy}(x,0)-ch(x) = -e^{-x}\varphi(x)$.

When we apply the Fourier transformation (5) on the equation (10) it transformes into

(11) $cU_y(x,0)-bU_{yy}(x,0)-cH(x) = -\Phi(x+i)$.

Since $U_y(x,0)=C(x)$ and $U_{yy}(x,0)=-2|x|C(x)$, substituting these values into (11) we hawe

(12) $C(x)\cdot(c+2b|x|) = cH(x)-\Phi(x+i)$.

Application of Fourier transformation (5) on the relation (9) gives

(13) $C(x)\cdot(1+2a|x|) = H(x)+\Phi(x)$.

Elimination of $C(x)$ from (12) and (13) gives

(14) $\Phi(x) = -\frac{1+2a|x|}{c+2b|x|}\Phi(x+i)+cH(x)\cdot\frac{1+2a|x|}{c+2b|x|} - H(x)$.

Relation (14) represents, so called, Carlemann's boundary value problem for determining analitic function $\Phi(z)$.

The most of the theory of Carlemann's boundary value problem is developed by soviet authors and analitic solving methods are given in details in [1] .

Nevertheless, in many cases it is more convenient to apply approximate solving methods. In monography [2] pp 156-158, the following theorem, which enables approximate solution of problem (14), is formulated and proved.

Theorem T: Given the Carlemann's boundary value problem

(15) $K\Phi \equiv \Phi(x)+[1+D(x)]\Phi(x+i) = G(x)$, $-\infty < x < \infty$

and corresponding approximate problem

(16) $\tilde{K}\tilde{\Phi} \equiv \tilde{\Phi}(x)+[1+\tilde{D}(x)]\tilde{\Phi}(x+i) = \tilde{G}(x)$, $-\infty < x < \infty$.

Suppose that for coefficient by $\Phi\ (x+i)$ the following conditions are fulfiled

$$(17)\qquad \begin{cases} 1+\widetilde{D}(x)\neq 0 \quad , \quad \widetilde{D}(x)\text{-continuous} \quad , \quad \widetilde{D}(\pm\infty)=0 \quad , \\ \operatorname{Ind}\left[1+\widetilde{D}(x)\right]=0 \end{cases}$$

and that it may be factorized in the following way

$$(18)\qquad 1+\widetilde{D}(x)=\frac{\widetilde{N}(x)}{\widetilde{N}(x+i)}$$

where function $\widetilde{N}(z)$ is continuous, bounded and analitic in the belt $0\leqslant \operatorname{Im} z\leqslant 1$ and nonzero in that domain.

Let us introduce the notation

$$M=\max\left\{\max_x |\widetilde{N}(x)|\ ,\max_x |\widetilde{N}(x+i)|\right\} \quad .$$

Let, further, in Carlemann's boundary value problem (15) function $D(x)$ be bounded and such that

$$(19)\qquad M\max_x \left|\frac{D(x)-\widetilde{D}(x)}{\widetilde{N}(x)}\right|<1 \quad .$$

Then for any right-hand side $G(x)$ from $L_2(-\infty,\infty)$, problem (15) has the unique solution in belt $0\leqslant \operatorname{Im} z\leqslant 1$. That solution is determined by formula

$$(20)\qquad \Phi=\widetilde{\Phi}+\left[I+\widetilde{K}^{-1}(K-\widetilde{K})\right]^{-1}\widetilde{K}^{-1}(G(x)-K\widetilde{\Phi})$$

where $\widetilde{\Phi}=\widetilde{\Phi}(x)$ is the solution of approximate Carlemann's problem (16). The difference between solutions is estimated by the inequality

$$(21)\qquad \|\Phi-\widetilde{\Phi}\|_{L_2}\leqslant\frac{\|\widetilde{K}^{-1}(G(x)-K\widetilde{\Phi})\|_{L_2}}{1-\|\widetilde{K}^{-1}(K-\widetilde{K})\|_{L_2}}$$

where the inverse operator $\widetilde{K}^{-1}$ is determined by formula

$$(22)\qquad \widetilde{K}^{-1}\widetilde{G}\equiv\widetilde{N}(x)\mathcal{F}\left[\frac{1}{1+e^{-t}}\mathcal{F}^{-1}\left(\frac{\widetilde{G}}{\widetilde{N}}\right)\right] \quad .$$

Let us now apply the mentioned theorem on the approxi- te solving of Carlemann's problem (14) .Instead of the whole per halfplane we'll obtain only the belt $0 < \mathrm{Im}\ z < 1$ and hoose, for sake of easier computation, that $a = 1/2$. $b = 1/2$ nd $c = 2$. Then Carlemann's boundary value problem (14) appears o be

$$\Phi(x) + \frac{1+|x|}{2+|x|}\Phi(x+i) = 2H(x)\cdot\frac{1+|x|}{2+|x|} - H(x) \qquad , \tag{23}$$

here function $\Phi(x)$ has to be analitic in the belt $0 < \mathrm{Im}\ z < 1$ nd for every $y \in [0,1]$ satisfies the inequality

$$\int_{-\infty}^{\infty} |\Phi(x+iy)|^2\, dx \leq C \qquad . \tag{24}$$

The free term $G(x) = 2H(x)\frac{1+|x|}{2+|x|} - H(x)$ is given in $_2(-\infty , \infty)$ .Let us choose in our case that $1+\widetilde{D}(x) = \frac{x^2+25}{x^2+36}$.

/Coefficient $1+\widetilde{D}(x)$ of the approximate problem is choosen in the form of rational function in order to avoid complicated computations with Fourier integral. More than that, this function is even, equal to one in infinity and easy to factorize since it is in form of

$$\frac{x^2+m^2}{x^2+(m+1)^2} \qquad ./$$

For the equation (23) the corresponding approximate equation will be

$$\widetilde{\Phi}(x) + \frac{x^2+25}{x^2+36}\widetilde{\Phi}(x+i) = 2\frac{x^2+25}{x^2+36}H(x) - H(x) \qquad . \tag{25}$$

In order that the equation (25) has the unique solution it is sufficient /see [2], § 15/ that the following conditions are fulfiled

$$\begin{array}{llll} & 1+\widetilde{D}(x) \neq 0 \quad , & \widetilde{D}(x) \text{ - continuous} & \\ (26) & \widetilde{D}(\pm\infty) = 0 \quad . & \mathrm{Ind}\,[1+\widetilde{D}(x)] = 0 & . \end{array}$$

It is easy to check that all of these conditions hold. Coefficient by $\tilde{\Phi}$ $(x+i)$ can be factorized in the following way

$$(27) \qquad 1+\tilde{D}(x) = \frac{x^2+25}{x^2+36} = \frac{\tilde{N}(x)}{\tilde{N}(x+i)} \quad , \qquad \tilde{N}(x) = \frac{x+5i}{x-6i}$$

where the function $\tilde{N}(z)$ is continuous, bounded and analitic in the belt $0 < \mathrm{Im}\, z < 1$ and nonzero in that domain. Then the boundary condition (25) by the substitution

$$(28) \qquad \tilde{\Phi}(x) = \tilde{N}(x)\cdot\Psi(x)$$

where $\Psi(x)$ is new continuous function, is transformed into

$$(29) \qquad \Psi(x)+\Psi(x+i) = 2\,\frac{H(x)}{\tilde{N}(x+i)} - \frac{H(x)}{\tilde{N}(x)} \quad , \quad -\infty < x < \infty .$$

When we apply the inverse Fourier transformation the equation (29) transformes into

$$(30) \qquad \Psi(x)+e^{-x}\cdot\Psi(x) = h(x)-11\int_{-\infty}^{x}\left[2e^{-6(x-s)}-e^{-5(x-s)}h(s)\,ds\right. \quad .$$

Equation (30) gives us the function $\Psi(x)$. For determining of function $\tilde{\mathcal{C}}(x)$ we use relation

$$(31) \qquad \tilde{\Phi}(x) = \Psi(x)\tilde{N}(x) = \Psi(x)\,\frac{x+5i}{x-6i} = \Psi(x)+\frac{11i}{x-6i}\cdot\Psi(x)$$

When we apply the inverse Fourier transformation on (31) we get

$$(32) \qquad \tilde{\mathcal{C}}(x) = \Psi(x)-11\int_{x}^{\infty} e^{6(x-s)}\Psi(s)\,ds \quad .$$

From formulae (30) and (32) function $\tilde{\mathcal{C}}(x)$ may be determined by $h(x)$. Now the approximate solution $\tilde{u}(x,y)$ of problem (1)-(2)-(3) reduces to the following simplier problem

$$(33) \qquad \Delta^2\tilde{u} = 0$$

$$(34) \qquad \tilde{u}(x,0) = 0$$

$$35)\qquad \tilde{u}_y(x,0) - \frac{1}{2}\,\tilde{u}_{yy}(x,0) = h(x) + \tilde{\varphi}(x)$$

$$\tilde{u}(x,\infty) \text{ - bounded.}$$

Problem (33)-(34)-(35) can be easily solved if we apply the operational calculus/see for example [3] /.Applying Fourier transformation we easily find out that

$$\tilde{U}(x,y) = yC(x)e^{-|x|v}$$

and

$$C(x)(1+|x|) = H(x) + \tilde{\Phi}(x)$$

which gives

$$\tilde{U}(x,y) = ye^{-|x|y}.\frac{H(x)+\tilde{\Phi}(x)}{1+|x|}$$

and

$$(36)\qquad \tilde{u}(x,y) = \mathcal{F}^{-1}\left\{ ye^{-|x|y}\,\frac{H(x)+\tilde{\Phi}(x)}{1+|x|}\right\} .$$

We also have that

$$M = \max\left\{\max_x |\tilde{N}(x)| \ ,\ \max_x |\tilde{N}(x+i)|\right\} = \frac{6}{5}$$

and

$$(37)\qquad \max_x\left|\frac{D(x)-\tilde{D}(x)}{\tilde{N}(x)}\right| \leqslant 0{,}23 \quad .$$

In that way we see that all conditions of the theorem T are fulfiled and that exact solution of the approximate Carlemann's problem (25) can be taken as the approximate solution of the basic Carlemann's problem (23) .

At the end,using the Parcevalle equality and inequalities (21) and (37) we can make the following estimation

$$(38)\quad \|\varphi(x) - \tilde{\varphi}(x)\|_{L_2} = \|\Phi(x) - \Phi(x)\|_{L_2} \leqslant$$

$$\leqslant \frac{1}{1-M\cdot 0{,}23}\,\|\tilde{K}^{-1}(K\tilde{\Phi} - G(x)\|$$

where the operator $\tilde{K}^{-1}$ is determined by formula (22).We have further that

$$(39)\qquad \|\widetilde{K}^{-1}(K\widetilde{\Phi}(x)-G(x)\| = \max\Big\{\Big(\int_{-\infty}^{\infty}|\widetilde{N}(x)\mathcal{F}\Big[\frac{1}{1+e^{-t}}\mathcal{F}^{-1}\ \frac{K\widetilde{\Phi}-G}{\widetilde{N}}\Big]|^2\ dx\Big)^{1/2},\Big(\int_{-\infty}^{\infty}|\widetilde{N}(x+i)\mathcal{F}\Big[\frac{e^{-t}}{1+e^{-t}}\mathcal{F}^{-1}\Big(\Big(\frac{K\widetilde{\Phi}-G}{\widetilde{N}}\Big)\Big]|^2\ dx\Big)^{1/2}\Big\}\leq M\Big(\int_{-\infty}^{\infty}\Big|\frac{K\widetilde{\Phi}(x)-G(x)}{\widetilde{N}(x)}\Big|^2\ dx\Big)^{1/2}\quad .$$

Let us estimate the last integral.

$$(40)\quad \int_{-\infty}^{\infty}\Big|\frac{K\widetilde{\Phi}(x)-G(x)}{\widetilde{N}(x)}\Big|^2\ dx = \int_{-\infty}^{\infty}|\widetilde{N}^{-1}(x)\Big[\widetilde{\Phi}(x)+\frac{1+|x|}{2+|x|}\widetilde{\Phi}(x+i)-$$

$$-2\,\frac{1+|x|}{2+|x|}\ H(x)+H(x)\Big]|^2dx = \int_{-\infty}^{\infty}|\widetilde{N}^{-1}(x)\Big[\Big(\frac{1+|x|}{2+|x|}-\frac{x^2+25}{x^2+36}\Big)\widetilde{\Phi}(x+i$$

$$+2\Big(\frac{x^2+25}{x^2+36}-\frac{1+|x|}{2+|x|}\Big)H(x)\Big]|^2\ dx\leq(0,23)^2\int_{-\infty}^{\infty}|\widetilde{\Phi}(x+i)+2H(x)|^2$$

$$=(0,23)^2\int_{-\infty}^{\infty}|\ e^{-t}\,\widetilde{\varphi}(t)+2h(t)|^2dt\quad .$$

Using (38),(39) and (40) we get inequality

$$(41)\qquad \|\varphi(x)-\widetilde{\varphi}(x)\|_{L_2}\leq 0,38\,\|e^{-x}\,\widetilde{\varphi}(x)+2h(x)\|_{L_2}\quad .$$

In order to make functions φ and $\widetilde{\varphi}$ even closer to each other, instead of approximate problem (25) we can take th approximate problem in the following form

$$(42)\quad \widetilde{K}\widetilde{\Phi}\ \equiv\ \widetilde{\Phi}(x)+\widetilde{\Phi}(x+i)\prod_{k=1}^{n}\frac{x^2+a_k^2}{x^2+(a_k+1)^2} =$$

$$= 2H(x)\prod_{k=1}^{n}\frac{x^2+a_k^2}{x^2+(a_k+1)^2}\ -\ H(x)$$

with conveniently choosen values for n and a_k .

REFERENCES

1. GAHOV F.D.:"Kraevie zadači","Nauka",Moskva,1977.

2. GAHOV F.D.,ČERSKII J.I.:"Uravnenija tipa svertki",Moskva, "Nauka",1978,156-158.

3. SNEDDON I.:"Preobrazovanija Furje",Moskva,IL,1955.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

SOLUTION OF POTENTIAL PROBLEMS WITH INTERNAL SOURCES BY BOUNDARY ELEMENT METHOD

Josip, E. Pečarić, Miodrag M. Radojković

BSTRACT
he paper presents an alternative proof of the boundary integral formulaion for two-dimensional potential problem with internal sources. This roof appeared to be much simpler than one derived by same authors in [4] nd thus is easier to extend to more complex cases (i.e. three-dimensio..al problems). Accuracy of the method is illustrated by an example.

REŠENJE POTENCIJALNOG PROBLEMA SA UNUTRAŠNJIM IZVORIMA PRIMENOM METODE GRANIČNIH ELEMENATA· U radu je prikazan alternativni dokaz integralne granične formulacije za ravanski potencijalni problem sa unutrašnjim izvorima. Ovaj dokaz je jednostavniji od dokaza koji su isti autori izveli u [4] i stoga ga je lakše proširiti na složenije slučajeve (na primer, prostorni problem). Tačnost metode je ilustrovana jednim primerom.

1. INTRODUCTION

Finite difference and finite element techniques were almost exclusively used to solve numericaly the equations governing the potential problems. Recently it was shown that boundary element method (BEM) can be also applied successfully [1].

In order to satisfy requirements that usualy arise in practice of solving potential problems, BEM solution procedure must incorporate the solutions of the following problems:

- modelling of sources with finite radii [4]
- modelling of coupled subregions with constant material properties [3].

In this paper given is a new simpler proof for the result from [4] concerning modelling of internal sources.

2. BASIC THEORY

In [4] the method was developed so that potential in a source can be computed for given flux and vice versa (note that only the first possibility exists in [2]). A potential problem for two-dimensional domain Ω from Fig. 1-a was considered, where Γ, $S_1, S_2, \ldots, S_n$ are its boundaries ($S_1, S_2, \ldots, S_n$ are circles with radii r_{o1}, r_{o2}, $\ldots$, r_{on} representing the system of internal sources).

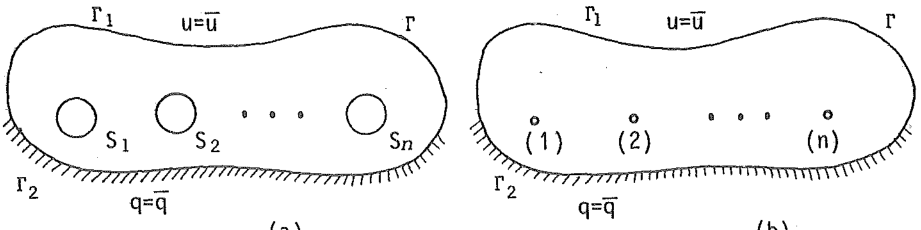


Fig. 1.

Equation that governs the problem reads:

$$(1) \qquad \nabla(K\nabla u) = -\sum_{K=1}^{n} Q'_k\, \delta(\vec{x}-\vec{x}_k)$$

where u is the potential, n is the number of sources, Q'_k is the volume rate of flux (positive for a source, negative for a sink) for the k-th source, $\vec{x}_k$ is the coordinate of the k-th source and δ is Dirac delta function. For a homogeneous medium (K=Const) (1) becomes:

$$(2) \qquad \nabla^2 u = -\sum_{K=1}^{n} Q_k\, \delta(\vec{x}-\vec{x}_k)$$

where $Q_k = Q'_k/K$ ($K = 1,2,\ldots,n$).

The boundary conditions are:

$$u = \bar{u} \text{ on } \Gamma_1,\; q = \frac{\partial u}{\partial n} = \bar{q} \text{ on } \Gamma_2 \qquad (\Gamma = \Gamma_1 + \Gamma_2).$$

Since (2) is Poisson's equation one can start from its well known boundary integral form (see for example [1, pp 45-47])

$$(3) \qquad C^i u_i + \int_\Omega bu^* d\Omega + \int_\Gamma uq^* d\Gamma = \int_\Gamma qu^* d\Gamma$$

where u* is the fundamental solution of the Laplace equation (solution for concentrated potential acting at point "i"):

$$(4) \qquad u^* = \frac{1}{2\pi} \ln \frac{1}{r}$$

where r is the distance from the point "i" to any point, C^i is a constant from [0,1] (C^i = 1 for an internal point, C^i= 0 for an external point and C^i = 0,5 for a point on the smooth boundary) and b is known function from Poisson equation. In the case considered:

$$(5) \qquad b = - \sum_{K=1}^{n} Q_k \, \delta(\vec{x}-\vec{s}_k)$$

Than (3) becomes:

$$(6) \qquad C^i u_i - \sum_{j=1}^{n} u_j^{i*} Q_j + \int_\Gamma u q^* d\Gamma = \int_\Gamma q u^* d\Gamma$$

i.e.

$$(7) \qquad C^i u_i - \sum_{j=1}^{n} u_j^{i*} Q_j + \sum_{K=1}^{N} \int_{\Gamma_k} u q^* d\Gamma = \sum_{K=1}^{N} \int_{\Gamma_k} q u^* d\Gamma$$

where N is the number of segments used to devide the boundary (boundary elements).

In the case when all Q_j's are known one has the case from [2,p.49] where the method of superposition was used to solve the problem.

But the case from [4] where all Q_j are not known can be also obtained using (7). If the potential u_{N+i} on S_i is known (see Fig.1-a) one can assume its value on the distance r_{o1} from source and put point "i" on this distance from the source (see Fig.1-b). Now, this is a point inside the domain and C^i = 1 so that (7) becomes:

$$(8) \qquad u_{N+i} - \sum_{j=1}^{n} u_j^{i*} Q_j + \sum_{K=1}^{N} \int_{\Gamma_k} u q^* d = \sum_{K=1}^{N} \int_{\Gamma_k} q u^* d$$

For $j=i$ one has:

$$(9) \qquad u_i^{i*} = \frac{1}{2\pi} \ln \frac{1}{r_{oi}}$$

For $j \neq i$ one can suppose $r_{ji} \gg r_{oi} + r_{oj}$($r_{ij}$ is the distance between sources i and j) and put:

$$(10) \qquad u_j^{i*} \approx \frac{1}{2\pi} \ln \frac{1}{r_{ij}}$$

Equations (7) and (8), after selection of the boundary element type (constant, linear, quadratic, etc. or mixed) [3] can be solved for all unknowns u's and q's on the boundary Γ and all unknown u's and Q's for sources (sinks) by solving corresponding system of linear algebraic equations.

Furthermore, using these values, one can compute the values of u's and q's at any internal point.

Note that the same result was obtained in [4] but the proof given in this paper is much simpler. The similar procedure can be extended to threedimensional case without any difficulties (at least from the theoretical point of x view).

3. AN EXAMPLE

The procedure outlined above was incorporated in the BEM computer program currently in use at Civil Engineering Faculty in Belgrade to solve two-dimensional potential problems. One example used for verification of the method is given in Fig. 2.

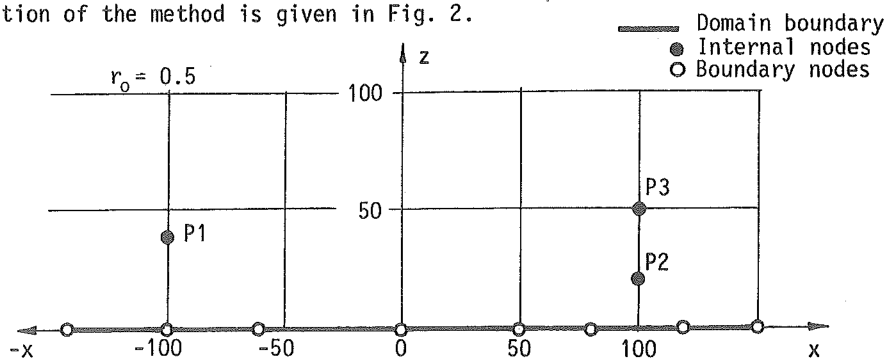


Fig. 2.

The problem is to find the potential distribution in the half plane with straight boundary (y=0) along which the constant potential u = 0 is given. Three internal sources with radius r_s= 0,5 are located at points P_1 (-100,40), P_2 (100,20), P_3 (100,50) with potentials given. Numerically computed fluxes in each source are compared with analytically computed ones in table 1.

Table 1.

	P_1	P_2	P_3	Method
u (potential)	4.97	4.83	-4.94	
Q (flux)	6.025	5.990	4.785	Analytic solution
Q (flux)	6.000	5.975	4.780	Boundary elements

Potentials inside the domain in different cross sections of the half plain are compared in Fig. 3.

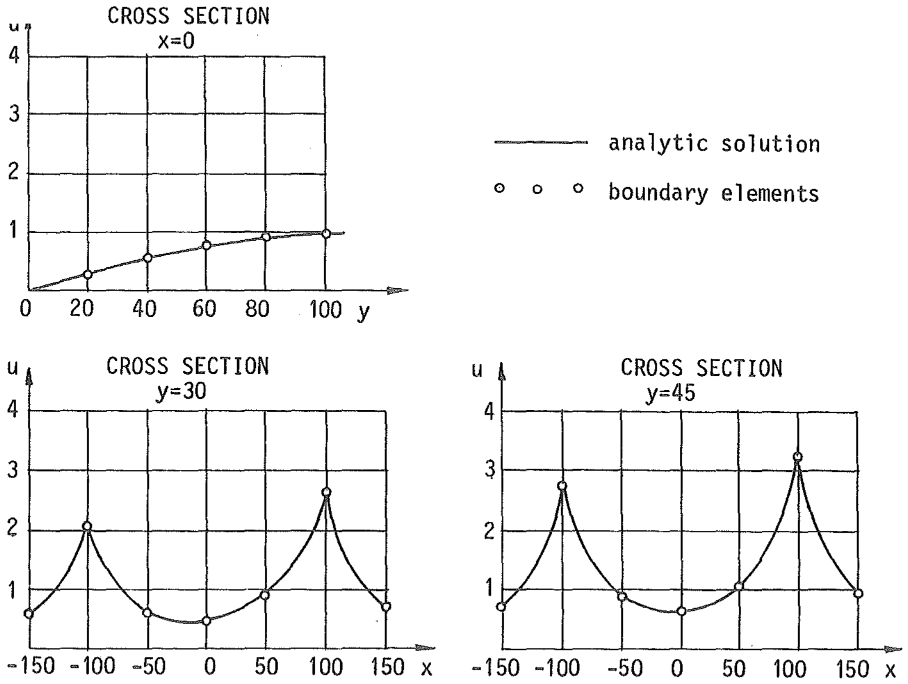


Fig. 3.

It is seen that the numerical solution of the problem considered is very accurate although the boundary discretisation was rather rough. Note that solution of the same problem with finite element method would require very fine discretisation in the vicinity of sources to achieve the same accuracy.

4. REFERENCES

1. BREBBIA C.A.,WALKER S.: Boundary Element Techniques in Engineering. Newnes-Buttherworths,London-Boston-Sydney-Wellington-Durban-Toronto 1980.
2. LIGGET J.A.,LIU P.L.F.: The Boundary Integral Equation Method for Porous Media Flow. George Allen & Unwin, London-Boston-Sydney 1983.
3. PEČARIĆ J.E., RADOJKOVIĆ M.M.: Mixed Boundary Element for Potential Problems; to appear.
4. RADOJKOVIĆ M.M., PEČARIĆ J.E.: Boundary Element Analysis of flow in Aquifers. 5-th International Conference on Finite Elements in Water Resources, Burlington 1984.

Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

A POSSIBILITY FOR CALCULATING PRESSURE GRADIENT FORCE IN SIGMA COORDINATE SYSTEM

Dragutin T. Mihailović

BSTRACT:

new scheme for the calculation of pressure gradient force n the sigma coordinate system is proposed. An approximation or the ωα term in the thermodynamics equation is considered oo. The proposed method and an earlier approach {2} is comared by time-integrations of the atmosphere at rest.

EDNA MOGUĆNOST ZA IZRAČUNAVANJE SILE GRADIJENTA PRITISKA U IGMA KOORDINATNOM SISTEMU. *Predložena je nova šema za izračunavanje sile gradijenta pritiska u sigma koordinatnom sistemu. Razmatrena je jedna aproksimacija za ωα član u jednačini termodinamike. Predloženi metod i jedna ranije predložena aproksimacija {2} pordjene su pomoću vremenskih integracija za slučaj mirne atmosfere.*

1. INTRODUCTION

The problem of calculating pressure gradient force in the sigma coordinate system is well known. It is related to the appearance of two terms in the expressions for the pressure gradient force. For example, with the original sigma coordinate {8} over a sloping terrain the two terms in the expressions of the pressure gradient force tend to be large in absolute value and have opposite signs. If, say, they are individually ten times greater then their sum, a 1% error in temperature (2-3^{O}C) will result in a 10% error in the pressure gradient force {9}. To overcome this problem, a number of difference analogues of the pressure gradient force in the sigma coordinate system have been developed {2},{1},{3}. A problem encountred by some of these analogues when geopotential is initially specified rather than temperature has recently been discussed by Mesinger {5} and compared by a numerical example of the atmosphere at rest {6}.

In this paper we shall concentrate our attention to the possibility of calculating the pressure gradient force in the sigma coordinate system by means of an interpolation procedure. In addition, we shall try to approximate in finite-difference form, the $\omega\alpha$ term that provides consistent transformation from potential to kinetic energy. Finally, the proposed schemes was compared with an earlier one {2}.

2. METHOD OF CONSTRUCTION

Notation

c_p	specific heat at constant pressure
k	suffix indicating level of the model
p	pressure
R	gas constant
s	suffix indicating surface value
t	suffix indicating atmosphere top value
T	temperature
$\mathbf{V}$	lateral vector wind
u,v	components of $\mathbf{V}$
α	specific volume
σ	$(p-p_t)/(p_s-p_t)$ the vertical coordinate
$\dot{\sigma}$	$d\sigma/dt$
π	p_s-p_t
ω	dp/dt
ϕ	geopotential
∇_σ	lateral del operator in "sigma" surfaces
∇_p	lateral del operator in "pressure" surface

In the sigma coordinate system, the differential form of the pressure gradient force has the form

$$- \nabla_p \phi = - \nabla_\sigma \phi - RT\nabla\pi \tag{1}$$

Starting from this expression, Kurihara {4} proposed a technique for calculating the pressure gradient force in the sigma coordinate system. Namely, it is possible to minimize the error in the calculation of pressure gradient force by interpolating geopotential from the nearest sigma surfaces to constant pressure surface. This idea was applied to vertically non-staggered grid with velocity components, temperature and geopotential defined in the middle of the layers.

Kurihara's idea can also be applied in the case of the staggered grid in the vertical, with geopotential located at the interfaces of the layers. This decision seems more reasonable since the latter grid is a better choice then the former one {10}. In our case we used quadratic interpolation in accordance with the hydrostatic equation in the form

$$\frac{\partial \phi}{\partial p} = - \frac{RT}{p} \quad . \tag{2}$$

Let us add that in the case of a more realistic atmosphere (inversion) we can apply the spline method of interpolation using all levels of the atmosphere model.

For a number of pressure gradient force schemes an associated procedure for calculation the $\omega\alpha$ term of the thermodynamic equation,

$$\frac{\partial}{\partial t}(\pi c_p T) + \nabla_\sigma (\pi V c_p T) + \frac{\partial}{\partial \sigma}(\pi \dot{\sigma} c_p T) = \pi \omega \alpha \quad , \tag{3}$$

ensuring consistency of the transformation between kinetic and potential energy, has been developed. Experiance has shown that it is desirable to preserve the consistency even in numerical models designed for short-range simulations. Otherwise, numerical instability may be encountred in less than a day of simulation time, in the presence of steep topography especially.

The contribution to the generation of kinetic energy by the pressure gradient force can be written in the form

$$\pi \omega \alpha = - \frac{\partial}{\partial \sigma}(\pi \, \phi \, \mathring{\sigma}) - \nabla_{\mathring{\sigma}} (\pi \phi V) - \frac{\partial (\phi \sigma)}{\partial \sigma} \frac{\partial \pi}{\partial t} - V \cdot (-\pi \nabla_p \phi) \tag{4}$$

It was stated already that the exact cancelation, in the finite-difference form should be provided between the $\omega\alpha$ terms in (3) and (4).

Taking into account the continuity equation, hydrostatic equation, and ω, we arrive at

$$(5) \qquad \pi\omega\alpha = -\frac{\partial}{\partial\sigma}(\pi\phi\dot{\sigma}) - \frac{\partial}{\partial\sigma}(\phi\sigma)\frac{\partial\pi}{\partial t} - \phi\nabla_\sigma\cdot\pi V + RTp\frac{\partial \, lnp}{\partial p}V\cdot\nabla_\sigma\pi \ .$$

Comparing the right sides of the expressions (4) and (5), that must be equal and in the finite-difference form, we find that the divergence of the surface pressure should be calculated via the expression

$$(6) \qquad \nabla\pi = \frac{\pi(\nabla_p\phi - \nabla_\sigma\phi)}{RT\frac{\partial \, lnp}{\partial p}} \ .$$

In this way we cancel the $\omega\alpha$ terms in (3) and (4) in the expression for total energy.

Using the thermodynamics equation (3), hydrostatic equation (2) and definitio of ω, we can write

$$(7) \qquad \omega\alpha = -\frac{1}{c_p}\{\dot{\sigma}\frac{\partial\phi}{\partial\sigma} + \sigma\frac{1}{\pi}\frac{\partial\pi}{\partial t}\frac{\partial\phi}{\partial\sigma} + \frac{V\cdot|\pi(\nabla_p\phi - \nabla_\sigma\phi)|}{RT\frac{\partial \, lnp}{\partial p}}\frac{\partial\phi}{\partial\sigma}\} \ .$$

In the finite-difference form, the last expression, for the case of horizontally staggered variables, in the x direction, has the form

$$(8) \qquad (\omega\alpha)_0 = -\frac{1}{c_p}\{|\overline{\dot{\sigma}}^\sigma\delta_\sigma\phi + \frac{1}{\pi}\overline{\frac{\partial\pi}{\partial t}\sigma}^\sigma\delta_\sigma\phi|_0 + Z_0 V\cdot\overline{|\pi_*(\delta_{x,p}\phi - \delta_{x,\sigma})|}^x\}$$

where

$$(9) \qquad Z_0 \equiv \frac{\delta_\sigma\phi}{RT\delta \, lnp}$$

and $\delta_{x,p}$ and $\delta_{x,\sigma}$ are operators of divergence in the finite-difference form in x direction for p=const. and σ=const. The subscript 0 denotes the point where the contribution of the $\omega\alpha$ term is calculated; π_* denotes the value of π in the point in which velocity components are not defined.

3. A NUMERICAL EXAMPLE

We copmared the proposed scheme with an earlier approach {2} which includes a non-staggered distribution of variables in the vertical.

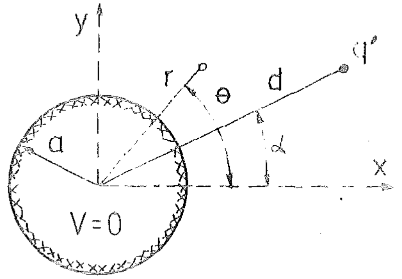
The experiment consisted of time-integrations with an atmosphere in a hydrostatic equilibrium; motions generated are thus a consequence of the pressure gradient force error (Blumber, personal comunication). The integrations were performed in a two dimensional domain (x,σ) with constant boundary conditions specified at the western boundary $(x=0)$. At the eastern boundary (x=12000 km) the radiation boundary condition was used {7}. Atriangular mountain with 500 km width and the maximum height of 2 km was defined in the midlle of the domain. The atmosphere was devided into nine layers in the vertical. The initial surface pressure was 1000 mb away from the mountain. The top of the model atmosphere was at 200 mb. A strong inversion up to 900 mb was located on the left of the mountain; the temperature was 0°C at 900 mb and -10°C at 1000 mb. Otherwise a temperature profile linear in *lnp* was assumed, the temperature taking on the value 3.5°C at 1000 mb and 0°C at 900 mb. The exact initial geopotential was calculated integrating the given temperature profiles. The grid size was 250 km, time step 10 min, and Coriolis parameter 0.0001 s^{-1}.

days	0	1	2	3	4	5	6	7	8	9	10	11
C	18.49	6.08	5.61	5.49	5.55	5.42	5.42	5.40	5.47	5.48	5.53	5.51
I	0.39	0.39	0.41	0.42	0.43	0.44	0.46	0.47	0.48	0.50	0.51	0.53

days	0	1	2	3	4	5	6	7	8	9	10	11
C	8.92	5.40	5.12	4.84	4.68	4.60	4.41	4.32	4.20	4.10	3.99	3.93
I	0.32	0.32	0.33	0.34	0.35	0.36	0.38	0.39	0.40	0.42	0.43	0.44

Table 1. *RMS pressure gradient force error, in terms of geostrophic wind, for different schemes, and for the wind point nearest the mountain at its "inversion" side (above) and its "no inversion" side (below).*

(because of the system symmetry $V=V(r,\theta)$, $V\neq V(z)$ and $\partial V/\partial z=0$) and boundary condition $V=0$ for $r=a$, and may be obtained by conventional application of the image theorem.



Following this theorem the equivalent electrostatic system (the original line charge q' and its image $-q'$, located at the directio $r=a^2/d$, $\theta=\alpha$) can be used.

So the electric scalar potential is

(2) $\quad V = q'G$, where

$$(3) \qquad G = \frac{1}{4\pi\varepsilon} \ln \frac{r^2d^2+a^4-2a^2rd\ \cos(\theta-\alpha)}{a^2\left[r^2+d^2-2rd\ \cos(\theta-\alpha)\right]}$$

is so called Green's function, r,θ and z are cylindrical coordi tes, ε is the electric permitivity and $\delta(r-d)$ and $\delta(\theta-\alpha)$ are Dirac's δ-functions defined for $r=d$ and $\theta=\alpha$, respectively.

3. A NEW INTERPRETATION OF THE IMAGE THEOREM

The equation (1) can be also solved by following variant of integral transform method [2]:

Consider Laplace's equation $\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial V}{\partial r})+\frac{1}{r^2}\frac{\partial^2 V}{\partial \theta^2} = 0$

and assume the potential solution in the separable variable fo $V = R(r)\ F(\theta)$, where

$$(4) \qquad F'' = k^2F , \quad r^2R''+rR'+k^2R^2 = 0$$

and k is the separable constant to be determined. The solutions diferential equations (4) are $e^{k\theta}$ and $e^{-k\theta}$, and $cl(r)=\cos(k\ln\frac{r}{a})$ and $sl(r)=\sin(k\ln\frac{r}{a})$, respectively. Puting $k=nk_o$ (n is an intege and determining k_o so the functions $cl_n(r)=\cos(nk_o\ln\frac{r}{a})$ and $sl_n(r)=\sin(nk_o\ln\frac{r}{a})$ satisfied conditions

$$(5) \qquad \int_a^b cl_n(r)cl_m(r)\frac{dr}{r} = \begin{cases} 0, & \text{for } n\neq m \\ \frac{\pi}{2k_o}(1+\delta_{no}), & \text{for } n=m \end{cases} \qquad \text{and}$$

$$(6) \qquad \int_a^b sl_n(r)sl_m(r)\frac{dr}{r} = \begin{cases} 0, & \text{for } n\neq m \text{ and } n=0 \text{ or } m=0 \\ \frac{\pi}{2k_o} , & \text{for } n=m\neq 0 \end{cases} ,$$

where $b > a$, we have $k_o = \pi/\ln(b/a)$.

expending Dirac's δ-function $\delta(r-d)$ in the series

$$(7)\qquad r\,\delta(r-d) = \sum_{n=1}^{\infty} \frac{2sl_n(r)sl_n(d)}{\ln(b/a)}\ ,\ \text{for } a<d<b\ ,$$

we have in the case when $b\to\infty$

$$(8)\qquad r\delta(r-d) = \frac{2}{\pi}\int_0^{\infty} \sin(p\ln\frac{r}{a})\sin(p\ln\frac{d}{a})dp\ .$$

Because of the obtained Dirac's δ-function integral transformation the solution of Poisson's equation (1) can be written as

$$(9)\qquad V = \frac{q'}{\pi\varepsilon}\int_0^{\infty} \frac{\sin(p\ln\frac{r}{a})\sin(p\ln\frac{d}{a})}{p\ \mathrm{sh}(p\pi)}\,\mathrm{ch}\big[(\theta-\alpha\pm\pi)p\big]dp\ ,$$

wher "+" is for $0\leqslant\theta\leqslant\alpha$ and "-" is for $\alpha\leqslant\theta\leqslant 2\pi$.

3. APPLICATION OF PRESENT RESULTS, EXEMPLES AND CONCLUSION

The physical solution of the consider problem is independent to the mathematical approach. So the solutions (2) and (9) are equal and we have the following expression: V(from formula 2) V(from formula 9). Using series

$$(10)\qquad \frac{\mathrm{ch}(\theta-\alpha\pm\pi)p}{\mathrm{sh}(p\pi)} = \sum_{m=0}^{\infty} e^{p[\theta-\alpha\pm\pi-(2m+1)\pi]} + e^{-p[\theta-\alpha\pm\pi+(2m+1)\pi]}$$

and integral $\int_0^{\infty}\frac{1-\cos pA}{p}e^{-pB}dp = \frac{1}{2}\ln\frac{A^2+B^2}{B^2}$, for $B>0$, we have fro last expression

$$(11)\qquad \prod_{m=-\infty}^{\infty} \frac{\ln^2(\frac{rd}{a^2}) + [\theta-\alpha\pm\pi-(2m+1)\pi]^2}{\ln^2(\frac{r}{d}) + [\theta-\alpha\pm\pi-(2m+1)\pi]^2} = \frac{d^2r^2+a^4-2a^2rd\ \cos(\theta-\alpha)}{a^2[r^2+d^2-2rd\ \cos(\theta-\alpha)]}\ ,$$

where $a\leqslant r<\infty$, $a\leqslant d$ and with "+" for $0\leqslant\theta\leqslant\alpha$ and "-" for $\alpha\leqslant\theta\leqslant 2\pi$. Puting in (11) $\Psi=\theta-\alpha$, $R=r/a\geqslant 1$, $D=d/a\geqslant 1$ or $RD=e^A$, $A\geqslant 0$, and $R/D=e^B$, B is optional, we have:

$$(12)\qquad P=\prod_{m=-\infty}^{\infty} \frac{\ln^2(RD)+(\Psi-2m\pi)^2}{\ln^2(R/D)+(\Psi-2m\pi)^2} = \frac{R^2D^2+1-2RD\cos\Psi}{R^2+D^2-2RD\cos\Psi} \qquad \text{and}$$

$$(13)\qquad P_o=\prod_{m=1}^{\infty} \frac{A^2+(\Psi-2m\pi)^2}{B^2+(\Psi-2m\pi)^2} = \frac{\mathrm{ch}A-\cos\Psi}{\mathrm{ch}B-\cos\Psi}\ .$$

Separately, we have:

$$(14)\qquad P_1=\prod_{m=1}^{\infty} \frac{A^2+4m^2\pi^2}{B^2+4m^2\pi^2} = \frac{B\ \mathrm{sh}(A/2)}{A\ \mathrm{sh}(B/2)}\ ,\ \text{for } \Psi=0,$$

$$(15)\qquad P_2=\prod_{m=1}^{\infty} \frac{C^2+m^2}{m^2} = \frac{\mathrm{sh}(\pi C)}{C\pi}\ ,\ \text{for } \Psi=0,\ A=2\pi C \text{ and } B=0 \text{ and}$$

$$(16)\qquad P_3=\prod_{m=1}^{\infty} \frac{C^2+(2m-1)^2}{D^2+(2m-1)^2} = \frac{\mathrm{ch}(\pi C/2)}{\mathrm{ch}(\pi D/2)}\ ,\ \text{for } \Psi=\pi/2,\ A=\pi C/2 \text{ and } B=\pi D/2\ .$$

The obtained formulas are very useful, because of the slovly convergence of present infinite products. The convergence of products shown folloving numerical results (Approximate values are calculated by multipling 65 membres of infinite products. The exac values are in breckets.):

For R=2, D=2 and $\Psi=-\pi$ P = 1.55783 (exact P = 1.5625).
For R=2, D=4 and $\Psi=-\pi$ P = 2.24039 (exact P = 2.2500).
For R=2, D=2 and $\Psi=-\pi/6$ P = 9.36901 (exact P = 9.93971143).
For R=2, D=2 and $\Psi=-2\pi/3$ P = 1.74477 (exact P = 1.7500).

In the following table the exact values of P_2 for different C are shown.

C	P_2	C	P_2
0.	1.000000000	2.	4.2612923E+01
0.01	1.000164501	5.	2.1121847E+05
0.1	1.016530706	10.	7.0078318E+11
0.5	1.465052383	20.	1.5428269E+25
1.	3.676077910	50.	5.2682675E+65

Combinig the results (16) for C=1 with known formula

$P_3 = \prod_{m=2}^{\infty} (1-\frac{1}{m^2}) = \frac{1}{2}$, we have $P_4 = \prod_{m=2}^{\infty}(1-\frac{1}{m^4}) = \frac{sh\pi}{4\pi} =$

$= 0.9190194776.$

REFERENCES

1. SURUTKA J.V.: Elektromagnetika, Građevinska knjiga, Beograd 1965.
2. VELIČKOVIĆ D.M.: Metodi za proračun elektrostatičkih polja, Stil, Podvis, 1982.
3. GRADŠTAJN I.S., RIŽIK I.M.: Tablici integralov, sum, rjadov proizvedenia, Izdatelstvo "Nauka", Moskva 1971.

Numerical Methods and
Approximation Theory
Niš, September 26 - 28, 1984

VALUATION OF SEVERAL SINGULAR INTEGRALS USING ELECTROSTATIC FIELDS LOWS

Dragutin M. Veličković

BSTRACT:

sing electrostatic fields lows a special approach to the several singular integrals evaluating is presented. The obtained reılts are useful in numerical solution of electrostatic problems ᴗy integral equations technique.

ODREĐIVANJE NEKIH SINGULARNIH INTEGRALA POMOĆU ZAKONITOSTI ELEKTROSTATIČKIH POLJA: Korišćenjem određenih elektrostatičkih zakona izvršeno je izračunavanje određenog broja singularnih integrala. Dobijeni rezultati su od koristi u toku približnog numeričkog rešavanja integralnih jednačina elektrostatike.

1. INTRODUCTION

In the applied electromagnetic field theory we have often necessary to compute several kinds of integrals having singular subintegral functions. Special in the case when field points are in the region of electromagnetic field sources. Because of the singularity of subintegral functions conventional numerical quadrature formulas are not useful, except after the singularity extraction. The present paper shown an effective method for evaluating several kinds of singular integrals. In the present method essence is the application of several electrostatic field lows, in the first place conformal maping and logarithmic potential theory. Except general theoretical description, two separate examples are shown. The obtained results are very useful for numerical solution of electrostatic integral equations systems. So we have the excelent numerical results in the theory of stripe lines.

2. EVALUATION OF SINGULAR INTEGRALS

We consider planparallel electrostatic field with known but arbitrary cross section (Fig.1).

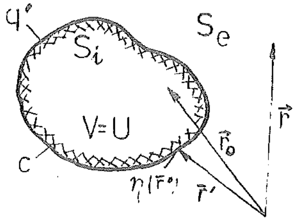
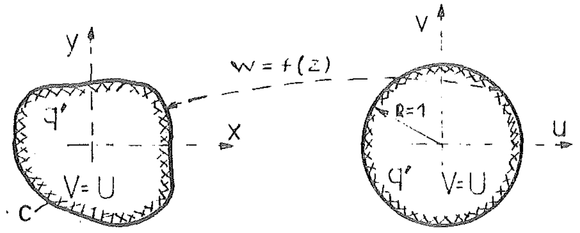


Fig.1



z-plane

w-plane

Fig.2

For electric scalar potential evaluation two general pr
cedure exist:

1.^{o} In the case when the surface charges densities on th electrode, $\eta(\vec r')$, are known, the potential, V, is:

$$(1)\quad V(\vec r) = \begin{cases} U, \text{ for } \vec r \in S_i,\ S_i \text{ is conductor interior} \\ U - \frac{1}{2\pi\varepsilon}\oint\limits_c \eta(\vec r')dl'\ln(|\vec r-\vec r'|/|\vec r_0-\vec r'|), \text{ for } r\in S_e \\ \qquad S_e \text{ is conductor exterior,} \end{cases}$$

where c is the contour of conductor cross section, ε is electri permitivity, U is the conductor potential and $q'=\oint\limits_c \eta(\vec r')dl'$ i the conductor total charge per unite length.

2.^{o} The second way for potential evaluating is based on conformal maping. If we have the complex function $w=Re^{j\Psi}=u+j$ $=f(z=re^{j\Theta}=x+jy)$, $j=\sqrt{-1}$, which map the exterior of conducto z-plane to the unite circular cylinder exterior in w-plane (Fi the complex potential is

$$(2)\quad \Phi = \begin{cases} U \text{ , for } R\leqslant 1 \\ U-\frac{q'}{2\pi\varepsilon}\ln w = V+j\Psi \text{ , for } R\geqslant 1. \end{cases}$$

During conformal maping the electrode potential and total lin charge densities are constant.

The real part of Φ is potential,

$$(3)\quad V = \mathrm{Re}\{\Phi\} = U - \frac{q'}{2\pi\varepsilon}\ln R \text{ .}$$

The electric field on the conductor surface is $E=|d\Phi/dz| =$ $= q'|w_0'|/2\pi\varepsilon$, where $|w_0'|=|dw/dz|$ for R=1. Using boundary con tion $\eta=\varepsilon E$ we have

$$(4)\quad \eta = \frac{q'}{2\pi}|w_0'| \text{ .}$$

uting (4) in (1) the potential is

$$(5)\qquad V(\vec{r}) = U - \frac{q'}{4\pi^2\varepsilon}\oint_C |w'_0|\,dl'\,\ln(|\vec{r}-\vec{r}'|/|\vec{r}_0-\vec{r}'|),\ \text{for } \vec{r}\in S_e\ .$$

Comparing (3) and (5) we have

$$(6)\qquad \oint_C |w'_0|\,dl'\,\ln(|\vec{r}-\vec{r}'|/|\vec{r}_0-\vec{r}'|) = \begin{cases} 0, & \text{for } \vec{r}\in S_i \\ 2\pi\ln R, & \text{for } \vec{r}\in S_e \end{cases}\ .$$

The present integrals have a singular subintegral functions and the results (6) is general formula for evaluating these singular integrals. The value of present integrals is independent of $\vec{r}_0\in S_i$.

3. EXAMPLES

Example I: Consider linear conformal maping $w=z/a$, a is positive constant, which map the circle having radius a in z-plane to the unite circle in w-plane. In this case is: $R=r/a$, $\psi=\theta$, $|w'_0|=1/a$, dl' $ad\theta'$, $|\vec{r}-\vec{r}'|=\sqrt{r^2+a^2-2ar\cos(\theta-\theta')}$, $|\vec{r}_0-\vec{r}'|=\sqrt{r^2+a^2-2ar_0\cos(\theta_0-\theta')}$, for $r_0\leqslant a$, and the expression (6) give:

$$(7)\qquad \int_0^{2\pi}\ln\frac{r^2+a^2-2ar\cos(\theta-\theta')}{r^2+a^2-2ar_0\cos(\theta_0-\theta')}\,d\theta' = \begin{cases} 0, & \text{for } r\leqslant a \\ 4\pi\ln(r/a), & \text{for } r\geqslant a \end{cases}\ .$$

The value of integral (7) is independent of θ_0 and $r_0\leqslant a$.

For $a=1$ and $r_0=0$ we have

$$(8)\qquad \int_0^{2\pi}\ln\left[r^2+1-2r\cos(\theta-\theta')\right]d\theta' = \begin{cases} 0, & \text{for } r\leqslant 1 \\ 4\pi\ln r, & \text{for } r\geqslant 1 \end{cases}\ .$$

Example II: Consider bilinear conformal maping $z=\frac{c}{2}(w+\frac{1}{w})$, c is positive constant, which map exterior of stripe conductor having widht $2c$ in z-plane to the exterior of unite circle in w-plane. Now is: $x=\frac{c}{2}(R+\frac{1}{R})\cos\psi$, $y=\frac{c}{2}(R-\frac{1}{R})\sin\psi$, $|w'_0|=1/\sqrt{c^2-x'^2}$, $|\vec{r}-\vec{r}'|=$ $=\sqrt{(x-x')^2+y^2}$, $|\vec{r}_0-\vec{r}'|=|x_0-x'|$, for $|x_0|\leqslant c$, and

$$(9)\qquad \int_{-c}^{c}\frac{dx'}{\sqrt{c^2-x'^2}}\ln\frac{(x-x')^2+y^2}{(x_0-x')^2} = 2\pi\ln R,\ \text{for } |x_0|\leqslant c,\ \text{where}$$

$$R = \frac{\sqrt{r^2+\sqrt{r^4+c^4-2r^2c^2\cos2\theta}+r\sqrt{2(r^2-c^2\cos2\theta+\sqrt{r^4+c^4-2r^2c^2\cos2\theta})}}}{c}$$

$x=r\cos\theta$, $y=r\sin\theta$, $r^2=x^2+y^2$ and $\cos2\theta=(x^2-y^2)/(x^2+y^2)$.

For different values of x and y we have:

$$\int_{-c}^{c}\frac{dx'}{\sqrt{c^2-x'^2}}\ln\frac{(x-x')^2}{(x_0-x')^2} = \begin{cases} 0, & \text{for } |x|\leqslant c\ ,\ |x_0|\leqslant c \\ \pi\ln\dfrac{2x^2-c^2+2x\sqrt{x^2-c^2}}{c^2}, & \text{for } |x|\geqslant c \end{cases}$$

and

$$\int_{-c}^{c}\frac{dx'}{\sqrt{c^2-x'^2}}\ln\frac{x'^2+y^2}{(x_0-x')^2} = \pi\ln\frac{2y^2+c^2+2y\sqrt{y^2+c^2}}{c^2},\ \text{for } |x_0|\leqslant c.$$