NUMERICAL METHODS AND APPROXIMATION THEORY

Niš, September 26-28, 1984

Edited by G. V. Milovanović
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.Z. Djordjević (Niš)
          P.B. Madić (Beograd)
          I.Z. 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
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

I. Bohte, M. Petkovšek
Gaussian elimination for diagonally dominant matrices 1

M. M. Laban
On some numerical properties of infinite-dimensional simplex 7

Lj. Cvetković, D. Herceg
Some sufficient conditions for convergence of AOR-method 13

M. S. Petkovšek, L. V. Stefanović
Some modified square root iterations for the simultaneous
determination of multiple complex zeros of a polynomial 19

D. V. Slavić
The generalization of ten rational approximations of
iteration functions 27

D. V. Slavić
One-point iteration functions of arbitrary convergence order 31

M. Ćojbašić
On the choice of the initial approximation in solving of the
operator equations by the Newton-Kantorović method 37

T. Slivnik, G. Tomsic
Numerical solution of the Fredholm integral equation of the
first kind with logarithmic singularity in the kernel 43

W. Gautschi, G. V. Milovanović
On a class of complex polynomials having all zeros in a
half disc 49

Lj. D. Petkovšek
On the optimal circular centered form 55

D. Dj. Tošić, D. V. Tošić
Two methods for the curve drawing in the plane 61

B. Cigrovski, M. Lapaine, S. Petrović
Operating with a sparse normal equations matrix 67

S. D. Miloradović
Approximation of $2^m$ - periodic functions by functions of
shorter period 73

V. N. Savić
On the strong summability $(C, \alpha)$ of transformations of simple
and multiple trigonometric Fourier series 79
An estimation for remainder of analytical function in Taylor's series

B.M. Damnjanović
On a method of numerical differentiation

G.V. Milovanović, J.E. Pešarić
On an application of Hermite's interpolation polynomial and some related results

D.V. Slavić, M.J. Stanojević
Classification of formulas for n-dimensional polynomial interpolation

B. Vrdoljak
On approximations of solutions of second order linear differential equations

K. Surla, E. Nikolić, Z. Lozanov
On hypothesis testing in spline regression

I.B. Lacković, Lj.M. Kočić
Approximation in discrete convexity cones

Lj.M. Kočić
Approximation of convex functions by first degree splines

K. Surla
On the spline solutions of boundary value problems of the second order

R. Vulanović
Mesh construction for numerical solution of a type of singular perturbation problems

D. Herceg, Lj. Cveticović
An iterative solution of some discrete analogues of a mildly nonlinear boundary value problem

D.D. Tošić
One way of discretization of Chaplygin's method

B.S. Jovanović, Lav D. Ivanović
On a convergence of the difference schemes for the equation of vibrating string

L.D. Ivanović, B.S. Jovanović
Approximation and regularization of control problem governed by parabolic equation

D.P. Radunović
Solutions of the grid Laplace equation defined in corners
I.S. Canak
A connection between one problem in elasticity theory and the method of approximate solving of Carleman's boundary value problem 173

E. Peđarić, M.M. Radojković
Solution of potential problems with internal sources by boundary element method 181

D.T. Mihailović
A possibility for calculating pressure gradient force in sigma coordinate system 187

D.M. Velišković
A new interpretation for Green's functions of circular conducting cylinder and its applications 193

D.M. Velišković
Evaluation of several singular integrals using electrostatic fields laws 197
GAUSSIAN ELIMINATION FOR DIAGONALLY DOMINANT MATRICES

Zvonimir Bohle, Marko Petkovšek

ABSTRACT:
Wilkinson [1] proved that the property of columnwise diagonal dominance 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.

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

\[ A^{(r)}x = b^{(r)}, \quad r = 1, \ldots, n \]

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

\[
A^{(r)} = \begin{bmatrix}
U_r & X_r \\
\emptyset & A_r
\end{bmatrix}
\]

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

\[
|a_{kk}| > \sum_{i=1}^{n} |a_{ik}|, \quad k = 1, \ldots, n
\]

\[
i \neq k
\]
then the matrix $A_r$ is also columnwise diagonally dominant, i.e.

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

for all $r = 2, \ldots, n-1$. 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) \hspace{1cm} fl(x \cdot y) = (x \cdot y)(1 + e), \quad |e| \leq u

where $x$ and $y$ are any standard floating point numbers and $fl(x \cdot y)$ 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)

$$B(r) = \begin{bmatrix} V_r & Y_r \\ \emptyset & E_r \end{bmatrix}$$
t is assumed that the matrix \( A = B^{(t)} \) is the matrix stored on the computer.

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

\[
\begin{align*}
    r = 1, \ldots, n-1: \\
    i = r+1, \ldots, n: \\
    m_{ir} &= f(t) \frac{b_{ir}^{(r)}}{b_{rr}^{(r)}} \\
    k = r+1, \ldots, n: \\
    b_{ik}^{(r+1)} &= f(t) \left( b_{ik}^{(r)} - f(t) m_{ir} b_{rk}^{(r)} \right) \\
\end{align*}
\]

Let us denote

\[
\begin{align*}
    h_r &= \max |b_{ik}^{(r)}|, \quad i, k = r, \ldots, n \\
    h &= \max h_r, \quad r = 1, \ldots, n \\
\end{align*}
\]

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

\[
\begin{align*}
    m_{ir} &= q_{ir} \left( 1 + x_{ir} \right), \quad i = r+1, \ldots, n \\
\end{align*}
\]

and

\[
\begin{align*}
    b_{ik}^{(r+1)} &= (b_{ik}^{(r)} - m_{ir} b_{rk}^{(r)}) (1 + y_{ik}^{(r)}) (1 + z_{ik}^{(r)}), \quad i, k = r+1, \ldots, n \\
\end{align*}
\]

where

\[
\begin{align*}
    q_{ir} &= b_{ir}^{(r)} / b_{rr}^{(r)} \\
    x_{ir}, \ y_{ik}^{(r)}, \ z_{ik}^{(r)} &\leq u \\
\end{align*}
\]

Let us suppose that

\[
\begin{align*}
    |x_{ir}|, \ |y_{ik}^{(r)}|, \ |z_{ik}^{(r)}| &\leq u \\
\end{align*}
\]

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

\[
\begin{align*}
    m_{ir} &= q_{ir} \left( 1 + x_{ir} \right), \quad i = r+1, \ldots, n \\
\end{align*}
\]

and

\[
\begin{align*}
    b_{ik}^{(r+1)} &= (b_{ik}^{(r)} - m_{ir} b_{rk}^{(r)}) (1 + y_{ik}^{(r)}) (1 + z_{ik}^{(r)}), \quad i, k = r+1, \ldots, n \\
\end{align*}
\]

We can write equation (7) in the form

\[
\begin{align*}
    b_{ik}^{(r+1)} &= b_{ik}^{(r)} - q_{ir} b_{rk}^{(r)} + d_{ik}^{(r)}), \quad i, k = r+1, \ldots, n \\
\end{align*}
\]

where

\[
\begin{align*}
    d_{ik}^{(r)} &= -q_{ir} b_{rk}^{(r)} \left( x_{ir} + y_{ik}^{(r)} + x_{ir} y_{ik}^{(r)} \right) (1 + z_{ik}^{(r)}) + \\
    &+ (b_{ik}^{(r)} - q_{ir} b_{rk}^{(r)} z_{ik}^{(r)} \\
\end{align*}
\]

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

\[
\begin{align*}
    |d_{ik}^{(r)}| &\leq h_r (2u + u^2)(1 + u) + 2uh_r = (4 + 3u + u^2)uh_r \\
\end{align*}
\]

Now, we can formulate the theorem.
3. THE THEOREM

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

\[ |a_{kk}| > \sum_{i=1}^{n} |a_{ik}| + \text{cun}(n-1)|a_{kk}|, \quad k = 1, \ldots, n \]

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,

\[ |b_{kk}^{(r)}| > \sum_{i=r}^{n} |b_{ik}^{(r)}| + \text{cun}(n-r+1)(n-r)|a_{kk}|, \quad k = r, \ldots, n \]

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

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

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 \( \text{cun}(n-1) < 1 \). Therefore, (ii) and (iii) hold trivially for \( r = 1 \).

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

\[ \sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| \leq |b_{rk}^{(r)}| + |b_{rr}^{(r)}| + \sum_{i=r+1}^{n} |b_{ir}^{(r)}| + \]

\[ + \sum_{i=r+1}^{n} |b_{ik}^{(r)}| + \sum_{i=r+1}^{n} |d_{ik}^{(r)}| \]

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

\[ \sum_{i=r+1}^{n} |b_{ir}^{(r)}| < |b_{rr}^{(r)}| - |b_{kr}| \]

\[ i \neq k \]
From (14) we have
\[
\sum_{i=r+1}^{n} |b_{ik}| \leq |b_{rk}| + |b_{kr}| - |b_{kr}|/|b_{rr}| + \sum_{i=r+1}^{n} |b_{ik}| + cuh_r(n-r-1) =
\sum_{i=r}^{n} |b_{ik}| - |q_{kr}| |b_{rk}| + cuh_r(n-r-1)
\]

Finally, from (i), (iii), (11) and (12) it follows
\[
\sum_{i=r+1}^{n} |b_{ik}| < |b_{rk}| - cu(n-r+1)(n-r)|a_{kk}| - \sum_{i=r+1}^{n} |q_{ir}| |b_{rk}| \leq |b_{kk}| - d_{kk}^{(r)} - cu((n-r)(n-r-1) + 2)|a_{kk}| \leq |b_{kk}| + 2cu|a_{kk}| - cu((n-r)(n-r-1) + 2)|a_{kk}| \leq |b_{kk}^{(r+1)}| - cu(n-r)(n-r-1)|a_{kk}|
\]

which proves (i).

To prove (ii), note that
\[
\sum_{i=r+1}^{n} |q_{ir}| < 1
\]
because $B_r$ is columnwise strictly diagonally dominant. Therefore, (11), (12), (15) and (iii) imply that
\[
\sum_{i=r+1}^{n} |b_{ik}| \leq \sum_{i=r+1}^{n} |b_{ik}^{(r)}| + \sum_{i=r+1}^{n} |q_{ir}| + \sum_{i=r+1}^{n} |d_{ik}^{(r)}| \leq \sum_{i=r}^{n} |b_{ik}| + 2cu(n-r)|a_{kk}|
\]

Then, using (ii) it follows
\[
\sum_{i=r+1}^{n} |b_{ik}| \leq \sum_{i=1}^{n} |a_{ik}| + cu(2n-r)(r-1)|a_{kk}| + 2cu(n-r)|a_{kk}| = \sum_{i=1}^{n} |a_{ik}| + cu(2n-r-1)r|a_{kk}|
\]

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
\[
\sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| \leq 2|a_{kk}| - c(u(n-1))|a_{kk}| + cu(2n-r-1)r|a_{kk}| =
\]
\[
= 2|a_{kk}| - cu(n-r)(n-r-1)|a_{kk}|
\]
Therefore, for each pair \(i, k = r+1, \ldots, n\)
\[
|b_{ik}^{(r+1)}| \leq (2 - cu(n-r)(n-r-1))|a_{kk}|
\]
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
\[
|a_{kk}| \geq d \sum_{i=1}^{n} |a_{ik}| , \; k = 1, \ldots, n
\]

The following table shows the minimal length of the mantissa in dependence on \(d\) and \(n\) with rounding in base 10.

<table>
<thead>
<tr>
<th>minimal length of the mantissa</th>
<th>(d)</th>
<th>(n = 5)</th>
<th>(n = 10)</th>
<th>(n = 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.001</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1.01</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES:
IN SOME NUMERICAL PROPERTIES OF INFINITE-DIMENSIONAL SIMPLEX

Miloš M. Laban

ABSTRACT:

Starting from an analytic model of infinite-dimensional simplex in Banach space, the possibility of its good approximation by one of its finite-dimensional subsimplexes is observed. The class of simplexes, where such an approximation is possible to either make or not are established by a sequence of theorems. Herefrom, the members of the class of limited infinite-dimensional simplexes with vertices making the orthogonal system, could not be approximated on such a way.

A finite-dimensional simplex in mathematics and applications is widely treated notion. There exists a great number of articles on analytic-geometrical properties of an 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 Naserick in [4], Phelps in [5], Lau in [3] and Kielein in [2]. For a difference of such a direction, we shall deal with the analytic-geometrical approach to this notion, keeping in 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 $x_0, x_1, \ldots, x_n, \ldots$ be such a vectors in $X$ that $\{x_1-x_0, \ldots, x_n-x_0, \ldots\}$ is the infinite unconditional set of linearly independent vectors. Let us denote

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

$$T = \left\{ \sum_{n=0}^{k} \alpha_n x_n \left| \sum_{n=0}^{k} \alpha_n = 1 \text{; } \alpha_0, \ldots, \alpha_k \geq 0 \text{; } \{i_0, \ldots, i_k\} \subset \{0, 1, 2, \ldots\} \right. \right\}$$

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

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

$$\lim_{j \to +\infty} y_j = y \quad (y_j = \sum_{n=0}^{+\infty} \beta_n x_n \text{; } \sum_{n=0}^{+\infty} \beta_n = 1 \text{; } \beta_n \geq 0)$$

states. Let us denote

$$y_j^* = \sum_{n=0}^{j-1} \beta_n x_n + (1 - \sum_{n=0}^{j-1} \beta_n) x_j \quad (j=2, 3, \ldots)$$

It is easy to verify that

$$y_j^* \in T \quad (j=2, 3, \ldots)$$

Further on, we have

$$\|y_j - y_j^*\| = \| (1 - \sum_{n=0}^{j} \beta_n) x_j - \sum_{n=0}^{+\infty} \beta_n x_n \| \leq (1 - \sum_{n=0}^{j} \beta_n) \| x_j \| + \sum_{n=j+1}^{+\infty} \beta_n \| x_n \|$$

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

$$\lim_{j \to +\infty} (1 - \sum_{n=0}^{j} \beta_n) = 0$$

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

$$\lim_{j \to +\infty} \| \sum_{n=j+1}^{+\infty} \beta_n x_n \| = 0$$

If we now let $j \to +\infty$ in (3), then accordingly to (4) and (5) we have
\[
\lim_{j \to \infty} \| y_j - y_j \| = 0 ,
\]
wherefrom and (1) it follows
\[
\lim_{j \to \infty} y_j = y .
\]
Herefrom, with the regard to (2), we obtain \( y \in \mathbb{T} \). Consequently, \( \mathbb{S} \subseteq \mathbb{T} \).

2° Let \( z \) be an arbitrary vector from \( \mathbb{T} \). Then there exists sequence \( (z_j) (j=1,2,\ldots) \) of vectors from \( \mathbb{T} \) such that
\[
\lim_{j \to \infty} z_j = z \quad (z_j = \sum_{n=0}^{k_j} \eta_n^j \cdot x_n) ; \quad \sum_{n=0}^{k_j} \eta_n^j = 1 ; \quad \eta_0^j, \ldots, \eta_{k_j}^j > 0 ;
\]
\( \{j_{x_0}, \ldots, j_{x_{k_j}} \} \subset \{x_0, x_1, \ldots\} \)
states. Let us denote
\[
z_j^* = \sum_{n=0}^{\infty} \eta_n^j \cdot x_n ,
\]
where
\[
\eta_n^j = \begin{cases} 
\eta_n^j , & x_n = j_{x_1} \\
0 , & x_n \notin \{j_{x_0}, \ldots, j_{x_{k_j}}\}
\end{cases}
\]
It is obvious that \( z_j^* \in \mathbb{S} (j=1,2,\ldots). \) Since \( z_j^* = z_j \), it follows \( \lim_{j \to \infty} z_j = z \), hence \( z \in \mathbb{S} \). Consequently \( \mathbb{T} \subseteq \mathbb{S} \) and the proof is completed.

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

**Definition 1:** Set \( \mathbb{S} \) we shall call the **infinite-dimensional simplex** (IDS in the further text) with vertices \( x_0, x_1, x_2, \ldots \) and denote \( S(x_0, x_1, x_2, \ldots) \). At the same time set
\[
p(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) \equiv x_{j_1} + L(x_{j_2} - x_{j_1}, \ldots, x_{j_k} - x_{j_1}, \ldots)
\]
(\( L \) denotes lineal) we shall call the face of \( S(x_0, x_1, x_2, \ldots) \), if \( \{x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots\} \) is the finite (or infinite) set of different vectors which is subset of \( \{x_0, x_1, x_2, \ldots\} \).

The following theorem (obviously true) points out that such a notion of IDS keeps a number of very important and for application rather useful properties of its 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_0, x_1, \ldots\}. Then:

1° \( S(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) \subseteq S(x_0, x_1, x_2, \ldots) \);

2° \( S(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) \subseteq p(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) \);

3° \( p(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) = p(x_{j_2}, x_{j_1}, \ldots, x_{j_k}, \ldots) \);

4° \( p(x_{j_1}, x_{j_2}, \ldots, x_{j_k}, \ldots) = x_{j_1} + \frac{1}{2}(x_{j_2} - x_{j_1}, \ldots, x_{j_k} - x_{j_1}) \).

2. APPROXIMATION

Let \( S(x_0, x_1, \ldots) \) 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 \( S(x_0, x_1, \ldots) \) is unlimited set and, consequently, it is not possible to replace it with an FDs which is necessarily a limited set. If \( \sup \|x_n\| < +\infty \), then we have the following results:

Theorem 3: Let \( \{x_0, x_1, \ldots\} \) be a orthogonal set and \( \inf \|x_n\| = \lambda > 0 \). 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° \( Y(y) \subseteq S(x_0, x_1, \ldots) \);

2° \( Y(y) \) is itself an IDS;

3° \((\forall x)(x \in Y(y))(\|x - y\| \leq \frac{\lambda}{2})\).

Proof: Without losing the generality in proof, we can observe FDs \( S(x_0, x_1, \ldots, x_k) \) and such \( y \in S(x_0, x_1, \ldots, x_k) \) that

\[ y = \sum_{n=0}^{k} \varphi_n x_n \quad (\sum_{n=0}^{k} \varphi_n = 1 \; \varphi_n > 0 \; (n=0,1,\ldots,k)) \]

where \( \varphi = \varphi_k = \max\{\varphi_n | n=0,1,\ldots,k\} \).

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
\[ \| x - y \|_2^2 = \| x \|_2^2 + \| y \|_2^2 > \| y \|_2^2 > \| x_k \|_2^2 > \frac{1}{4} \delta^2, \]

follows 3°.

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

\[ x^* = \sum_{n=k+1}^{+\infty} \varphi_n ((1-\eta)x_{k+1} + \varphi x_{n+1}), \]

where \( \sum_{n=k+1}^{+\infty} \varphi_n = 1, \varphi_n > 0 \) (\( n=k+1, k+2, \ldots \)). Since

\[ x^* = (\sum_{n=k+1}^{+\infty} \varphi_n (1-\eta)x_{k+1} + \sum_{n=k+1}^{+\infty} \varphi_n \varphi x_{n+1} = (1-\eta)x_{k+1} + \sum_{n=k+1}^{+\infty} \varphi_n \varphi x_{n+1} \]

it follows \( x^* \in S(x_0, x_1, \ldots) \), because

\[ (1-\eta) + \sum_{n=k+1}^{+\infty} \varphi_n \varphi = 1 - \varphi + \varphi = 1. \]

In the base of definition 1 we can now conclude that the condition 2° is fulfilled. Further on, we have

\[ \| x^* - y \|_2^2 = \| x \|_2^2 + \| y \|_2^2 > \| x \|_2^2 > (1-\eta)^2 \| x_{k+1} \|_2^2 > \frac{\delta^2}{4}, \]

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

\[ x_j^* = \sum_{n=k+1}^{+\infty} \varphi_n (j) ((1-\eta)x_{k+1} + \varphi x_{n+1}) (j = 1, 2, \ldots) \]

such that \( x = \lim_{j \to +\infty} x_j^* \).

Since \( \| x_j^* - y \| > \frac{\delta}{2} \) (\( j = 1, 2, \ldots \)), there exists such a natural number \( j_0 \) that

\[ \| x_{j_0}^* - y \| - \| x_{j_0}^* - x \| > \frac{\delta}{2}. \]

States, hence 3° 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 its FDS is not possible in that case, in spite of the fact that such a IDS is limited set. Therefore, it makes a sense to develop 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 \( \varepsilon \) be a arbitrary real number. Let, further, \( y = \sum_{n=0}^{+\infty} Q_n x_n \) (\( \sum_{n=0}^{+\infty} Q_n = 1; Q_n > 0 \) (\( n = 0, 1, \ldots \))) be such a vector that

\[ \sum_{n=0}^{k} Q_n > I - \frac{\varepsilon}{4 \sup \| x_n \|} \]

states. Then for each \( x \in S(x_0, x_1, \ldots) \cap K(y, \frac{\varepsilon}{2}) \), there exists \( y' \in S(x_0, x_1, \ldots, x_k) \) such that \( \| x - y' \| < \varepsilon \) is valid, where
\(K(y, \frac{\varepsilon}{2})\), as usual, denotes \(\{x|\|x-y\| < \frac{\varepsilon}{2}\}\frac{k-1}{k-1}\sum_{n=0}^{k-1} Q_n x_n\)

**Proof:** We shall demonstrate that \(y' = (1 - \sum_{n=0}^{k} Q_n) x_n + \sum_{n=k+1}^{k} \alpha_n x_n\)

satisfies the proposition. Really,

\[
\|x - y'\| \leq \|x - y\| + \|y - y'\| < \frac{\varepsilon}{2} + \left(1 - \sum_{n=0}^{k} Q_n\right)\|x_k\| + \sum_{n=k+1}^{k} \alpha_n \|x_n\| <
\]

\[
\frac{\varepsilon}{2} + \sum_{n=o}^{\sup \|x_n\|} \sup \|x_n\| \left(1 - \sum_{n=0}^{k} Q_n\right) < \frac{\varepsilon}{2} + \sum_{n=0}^{\sup \|x_n\|} \frac{\varepsilon}{n} = \varepsilon
\]

and the proof is completed.

The sufficient conditions when the absolute error made in replacing the IDS by its FDS is lower than given \(\varepsilon > 0\), followed in the next two theorems:

**Theorem 5:** If \(\|x_n\| < \frac{\varepsilon}{2}\) (\(n > k\)), then for each \(y \in S(x_0, x_1, \ldots)\)

there exists \(y' \in S(x_0, x_1, \ldots, x_k)\) such that \(\|y - y'\| < \varepsilon\).

**Proof:** Let \(y = \lim_{j \to +\infty} y_j\), where \(y_j = \sum_{n=0}^{n=0} Q_n x_n\) and \(\sum_{n=0}^{n=0} Q_n = 1\) and \(Q_n > 0\) (\(n = 0, 1, \ldots\)). Let us, further, denote

\[
y' = \sum_{n=0}^{k} r Q_n x_n + \left(1 - \sum_{n=0}^{k} r Q_n\right) x_k, \text{ where } \|y - y'\| < \frac{\varepsilon}{2}
\]

Now we have

\[
\|y - y'\| < \|y - y_x\| + \|y_x - y'\| < \frac{\varepsilon}{2} + \left(1 - \sum_{n=0}^{k} r Q_n\right)\|x_k\| + \sum_{n=k+1}^{k} r Q_n \|x_n\| <
\]

\[
\frac{\varepsilon}{2} + \left(1 - \sum_{n=0}^{k} r Q_n\right)\frac{\varepsilon}{2} + \left(1 - \sum_{n=0}^{k} r Q_n\right)\frac{\varepsilon}{2} < \varepsilon
\]

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 \(\varepsilon\)-approximation of IDS \(S(x_0 - a, x_1 - a, \ldots)\).

**References:**

SOME SUFFICIENT CONDITIONS FOR
CONVERGENCE OF AOR-METHOD

Ljiljana Cvetković, Dragoslav Herceg

ABSTRACT
We 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 $\sigma$ and $\omega$. We also consider the convergence intervals for some matrices, which are not strictly diagonally dominant.

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 $D = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ 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) \((E-\sigma L)x^{k+1} = ((1-\omega)E+(\omega-\sigma)L+wU)x^k + wc, \ k = 0, 1, \ldots\),

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

\[ M_{\sigma, \omega} = (E-\sigma L)^{-1}((1-\omega)E+(\omega-\sigma)L+wU). \]

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

For \(A = [a_{ij}] \in \mathbb{C}^{n \times n}\) (= set of complex nxn matrices) we define for \(i = 1, 2, \ldots, n\)

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

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

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

2. CONVERGENCE OF THE AOR METHOD

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

\[ \min_{1 \leq i \leq n} \frac{|1-\omega| - |\omega-\sigma|e_{\alpha, i} - |\omega|f_{\alpha, i}}{1+|\sigma|e_{\alpha, i}} \leq \rho(M_{\sigma, \omega}) \leq \max_{1 \leq i \leq n} \frac{|1-\omega| + |\omega-\sigma|e_{\alpha, i} + |\omega|f_{\alpha, i}}{1-|\sigma|e_{\alpha, i}}. \]
Proof. We prove the upper bound for \( \rho(M_{\sigma, \omega}) \). Let \( \lambda \) be any eigenvalue of \( M_{\sigma, \omega} \) and suppose that

\[
|\lambda| > \frac{|1-\omega| + |\omega-\sigma|e_{a_i}^T + |\omega|f_{a_i}}{1-|\sigma|e_{a_i}^T}, \quad i=1,2,\ldots,n.
\]

After some manipulations we have

\[
|\lambda + \omega - 1| > |\omega + \sigma (\lambda - 1)|e_{a_i}^T + |\omega|f_{a_i}, \quad i=1,2,\ldots,n,
\]

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

where \( B = [b_{ij}] \in \mathbb{C}^{n, n}, B = (\lambda + \omega - 1)E - (\omega + \sigma (\lambda - 1))L - \omega U \). Then theorem 2.5.2 from [2] shows that \( \det B \neq 0 \). Since \( (E - \alpha L) \cdot (\lambda E - M_{\sigma, \omega}) = B \) 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 \( A = [a_{ij}] \in \mathbb{C}^{n, n}, a_{ii} \neq 0, i=1,2,\ldots,n \). Then for \( \omega, \sigma \in \mathbb{R}, \omega \neq 0, |\sigma|(e_i^T + e_j) < 2, i \neq j, i,j = 1,2,\ldots,n, \rho(M_{\sigma, \omega}) \) satisfies the following:

\[
\min_{i \neq j} \frac{2|1-\omega| - |\omega - \sigma|(e_i^T + e_j) - |\omega|(f_i + f_j)}{2 - |\sigma|(e_i^T + e_j)} \leq \rho(M_{\sigma, \omega}) \leq \max_{i \neq j} \frac{2|1-\omega| + |\omega - \sigma|(e_i^T + e_j) + |\omega|(f_i + f_j)}{2 - |\sigma|(e_i^T + e_j)}
\]

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 \( \lambda \) with

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

\[
i,j = 1,2,\ldots,n.
\]

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

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

we have now

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

But then, theorem 2.4.1 from [2] shows that $\det B \neq 0$. This contradicts the singularity of $\lambda E - M_0, \omega$.

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 $A = [a_{ij}] \in \mathbb{C}^{n \times n}$, $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$ and $\alpha \in [0, 1]$.

Then the AOR method converges for

1. $\max \left( e_i + e_j \right) < 1$, $0 < \omega < \min \left( \frac{2}{1 + e_i + e_j} \right)$,

or

2. $\max \left( e_i + e_j \right) < 2$, $0 < \omega < \min \left( \frac{4}{2 + e_i + e_j} \right)$.

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 $i=1,2,\ldots,n$ holds
\[
e_{\alpha,i} + f_{\alpha,i} < 1, \quad 0 < \omega < \frac{2}{1 + e_{\alpha,i} + f_{\alpha,i}}
\]
\[
to\]
\[
\begin{align*}
2e_{\alpha,i} & < 2 \min(0, 1 - \omega) + \omega(1 + e_{\alpha,i} - f_{\alpha,i}) + 2 \max(0, \omega - 1) \\
2e_{\alpha,i} & < 2e_{\alpha,i}
\end{align*}
\]
\[
|1 - \omega| + \frac{|\omega - \sigma| e_{\alpha,i} + |\omega| f_{\alpha,i}}{1 - |\sigma| e_{\alpha,i}} < 1.
\]
ince for $\sigma$ and $\omega$ from (a) we have $|\sigma| e_{\alpha,i} < 1$, theorem 1 nd (3) show that $p(M_{\sigma,\omega}) < 1$.

o prove implication (2) $\implies$ (3) we consider the next cases.
ase I: $0 < \omega < 1$, $- \omega(1 - e_{\alpha,i} - f_{\alpha,i}) < \sigma < 0$.
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).
ase II: $0 < \omega < 1$, $0 < \sigma < \omega$.
hen $1 - \omega + \omega e_{\alpha,i} - \sigma e_{\alpha,i} + \omega f_{\alpha,i} < 1 - \sigma e_{\alpha,i}$, since $e_{\alpha,i} + f_{\alpha,i} < 1$.
ase III: $0 < \omega < 1$, $\omega < \sigma < \frac{\omega(1 + e_{\alpha,i} - f_{\alpha,i})}{2e_{\alpha,i}}$.
hen $1 - \omega + \sigma e_{\alpha,i} - \omega e_{\alpha,i} + \omega f_{\alpha,i} < 1 - \sigma e_{\alpha,i}$.
ase IV: $1 < \omega < \frac{2}{1 + e_{\alpha,i} + f_{\alpha,i}}$, $\omega + \omega e_{\alpha,i} + \omega f_{\alpha,i} - 2 < \sigma < 0$.
hen $\omega - 1 + \omega e_{\alpha,i} - \sigma e_{\alpha,i} + \omega f_{\alpha,i} < 1 + \sigma e_{\alpha,i}$.
ase V: $1 < \omega < \frac{2}{1 + e_{\alpha,i} + f_{\alpha,i}}$, $0 < \sigma < \omega$.
hen $\omega - 1 + \omega e_{\alpha,i} - \sigma e_{\alpha,i} + \omega f_{\alpha,i} < 1 - \sigma e_{\alpha,i}$.
ase VI: $1 < \omega < \frac{2}{1 + e_{\alpha,i} + f_{\alpha,i}}$, $\omega < \sigma < \frac{-\omega + \omega e_{\alpha,i} - \omega f_{\alpha,i} + 2}{2e_{\alpha,i}}$.
hen $\omega - 1 + \sigma e_{\alpha,i} - \omega e_{\alpha,i} + \omega f_{\alpha,i} < 1 - \sigma e_{\alpha,i}$.
Remark. If in case (a) of theorem 3 we assume $\alpha = 1$, then for strictly diagonally dominant matrices AOR method converges if

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

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

This convergence intervals for $\omega$ and $\sigma$ are larger than the corresponding intervals from theorem 3 of [5].

REFERENCES

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 methods, 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 numerically in the example of an algebraic equation.

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. Farnen and G. Lousi [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 reasonable 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 $n \geq 3$

$$P(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0 = \sum_{j=1}^{k} \left(z - r_j \right)^{m_j} \quad (a_i \in \mathbb{C})$$

with real or complex zeros $r_1, \ldots, r_k$ having the order of multiplicity $m_1, \ldots, m_k$ respectively, where $m_1 + \cdots + m_k = n$. Let $z_1, \ldots, z_k$ be distinct reasonably good approximations to these zeros and $z'_1$ be the next approximation to $r_1$ using some iterative scheme.

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

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

we define

(1) \quad $G(z) = u(z) [u(z) - v(z)]$ \quad (Ostrowski’s function),

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

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

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

(4) \quad $z'_1 = z_1 + N(z_1)$ \quad (Shröder’s modification of Newton’s method for multiple zeros, see [17]),

(5) \quad $z'_1 = z_1 + H(z_1)$ \quad (modification of Halley’s method, introduced by Hansen and Patrick [8] for multiple zeros),
the convergence order two and three respectively. We note that the order of multiplicity in the iterative formulas (4) and (5) take the values \( m = m_i \) \( (i = 1, \ldots, k) \).

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

\[
(6) \quad r_i = z_i - \sqrt[m_i]{G(z_i) - \sum_{j \neq i} m_j (z_i - r_j)^{-2}}^{-1/2} \quad (i = 1, \ldots, k).
\]

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) \( P'(z) \).

Setting \( r_i \neq z_i \) 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 \( r_1, \ldots, r_k \) are polynomial zeros) will be denoted by \( O_R(\text{IP}, \mathbf{r}) \).

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

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

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):

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

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

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):

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

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

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):

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

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

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

\[
O_R((10), \alpha) \geq 3 + \tau_k \in (3, 7)
\]

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 correction (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):

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

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

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):

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

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

\[
O_R((12), p) \geq 3(1 + \sigma_k) \in (6, 8),
\]

where \( \sigma_k \in (1, 5) \) is the unique positive root of equation \( \sigma^k - \sigma - 1 = 0 \) \((k \geq 2)\).

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])

\[ m_i = \lim_{z \to r_i} u'(z) \]

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 \( u(z^{(i)}(\lambda)) \) and \( v(z^{(i)}(\lambda)) \) (\( \lambda = 1, 2, \ldots \) 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

\[ P(z) = z^9 - 7z^8 + 20z^7 - 28z^6 + 10z^5 - 18z^4 + 110z^3 - 44z^2 + 345z + 225 \]

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:

\[ z_1^{(0)} = 1.8 + 2.7i, \quad z_2^{(0)} = 1.8 - 2.7i, \quad z_3^{(0)} = -0.3 - 0.8i, \quad z_4^{(0)} = 2.3 - 0.7i. \]
... site of crude initial approximations, the presented iterative methods demonstrate very fast convergence. Numerical results, obtained in the second iteration, are displayed in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>( \text{Re} { z_i^{(2)} } )</th>
<th>( \text{Im} { z_i^{(2)} } )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TS</strong></td>
<td>( 0.999999853800923892 )</td>
<td>( 2.0000000112716998844 )</td>
</tr>
<tr>
<td>(7)</td>
<td>( 0.999999826741999847 )</td>
<td>( -2.000000351383949125 )</td>
</tr>
<tr>
<td>3</td>
<td>( -0.99999859207295616 )</td>
<td>( -8.18 \times 10^{-7} )</td>
</tr>
<tr>
<td>4</td>
<td>( 3.000000527270300803 )</td>
<td>( -3.48 \times 10^{-8} )</td>
</tr>
<tr>
<td><strong>SS</strong></td>
<td>( 0.99999939617346251 )</td>
<td>( 1.99999964305993363 )</td>
</tr>
<tr>
<td>(8)</td>
<td>( 1.000000861310650873 )</td>
<td>( -2.000000509862992614 )</td>
</tr>
<tr>
<td>3</td>
<td>( 0.999999709498985 )</td>
<td>( 1.35 \times 10^{-9} )</td>
</tr>
<tr>
<td>4</td>
<td>( 3.000000000000030662 )</td>
<td>( 7.16 \times 10^{-14} )</td>
</tr>
<tr>
<td><strong>TSN</strong></td>
<td>( 0.999999455077856744 )</td>
<td>( 2.000000021296109474 )</td>
</tr>
<tr>
<td>(9)</td>
<td>( 1.00000001847137107 )</td>
<td>( -2.00000006835695135 )</td>
</tr>
<tr>
<td>3</td>
<td>( 0.99999974528732211 )</td>
<td>( 3.43 \times 10^{-8} )</td>
</tr>
<tr>
<td>4</td>
<td>( 3.0000007270680682 )</td>
<td>( -9.58 \times 10^{-8} )</td>
</tr>
<tr>
<td><strong>SSN</strong></td>
<td>( 0.999999894886117145 )</td>
<td>( 2.000000042747320793 )</td>
</tr>
<tr>
<td>(10)</td>
<td>( 1.00000001847137107 )</td>
<td>( -2.000000000709903145 )</td>
</tr>
<tr>
<td>3</td>
<td>( -1.00000000007845003 )</td>
<td>( 3.82 \times 10^{-11} )</td>
</tr>
<tr>
<td>4</td>
<td>( 2.999999999999997525 )</td>
<td>( -6.58 \times 10^{-15} )</td>
</tr>
<tr>
<td><strong>TSH</strong></td>
<td>( 1.000000000098386276 )</td>
<td>( 1.9999999999890580897 )</td>
</tr>
<tr>
<td>(11)</td>
<td>( 1.0000000000450329186 )</td>
<td>( -2.000000000521585396 )</td>
</tr>
<tr>
<td>3</td>
<td>( -0.99999999986166747 )</td>
<td>( -2.93 \times 10^{-12} )</td>
</tr>
<tr>
<td>4</td>
<td>( 3.000000000036870406 )</td>
<td>( -6.92 \times 10^{-10} )</td>
</tr>
<tr>
<td><strong>SSH</strong></td>
<td>( 1.000000000032764666 )</td>
<td>( 2.00000000002146278 )</td>
</tr>
<tr>
<td>(12)</td>
<td>( 1.0000000000674921 )</td>
<td>( -1.99999999997025086 )</td>
</tr>
<tr>
<td>3</td>
<td>( -1.000000000001322 )</td>
<td>( 2.74 \times 10^{-15} )</td>
</tr>
<tr>
<td>4</td>
<td>( 3.0000000000000383 )</td>
<td>( -2.01 \times 10^{-16} )</td>
</tr>
</tbody>
</table>

Table 1

We note that the lower bounds of the suggested serial methods in our example \( (k=4) \) are \( O_{\mathbb{R}}((8),r) \geq 4.453 \), \( O_{\mathbb{R}}((10),r) \geq 5.586 \) and \( O_{\mathbb{R}}((12),r) \geq 6.662 \).
REFERENCES


15. PETKOVIĆ M.S.: Generalised root iterations for the simultaneous determination of multiple complex zeros. ZAMM 62 (1982), 627-630.


THE GENERALIZATION OF TEN RATIONAL APPROXIMATIONS OF ITERATION FUNCTIONS

Dušan V. Slavić

ABSTRACT:
I. Newton (1676), E. Halley (1694), P.L. Čebišev (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 authorship priorities are considered.


Let \( u, A, B, C \) be defined by
\[
\begin{align*}
  u &= f/f', \\
  A &= f''/(2f'), \\
  B &= f''/(6f'), \\
  C &= f''/(24f'),
\end{align*}
\]
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) \( y_2 = x - u \) \hspace{1cm} \text{Newton}
(2) \( y_3 = x - u/(1 - Au) \) \hspace{1cm} \text{Halley}
(3) \( y_3 = x - u - Au^2 \) \hspace{1cm} \text{Čebišev}
(4) \( y_4 = x - u - Au^2 - (2A^2 - B)u^3 \) \hspace{1cm} \text{Čebišev}
(5) \( y_4 = x - u (1 - Au)/(1 - 2Au + Bu^2) \) \hspace{1cm} \text{Whittaker}
(6) \( y_2 = x - u/(1 - 2Au) \) \hspace{1cm} \text{Durand}
(7) \( y_2 = x - u (1 - 2Au)/(1 - 3Au + 3Bu^2) \) \hspace{1cm} \text{Durand}
(8) \[ y_4 = x - u(1-3Au+3Bu^2)/(1-4Au+(2A^2+4Bu^2)u^2-4Cu^3) \] Durand

(9) \[ y_4 = x - u (A - (A^2-B)u) / (A - (2A^2-B)u) \] Durand

(10) \[ y_4 = x - u / (1 - Au - (A^2-B)u^2) \] Traub

The literature is full of disagreements concerning 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), P.Viètes (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 hyperbolae (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.Cebyšev (1838), E.Schröder (1870), E.Bodewig (1935), ...


The uniform and simple manner of writing the iteration functions enables one to see more easily the iterations between 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

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

wherefrom, upon neglecting the terms with \( u^3 \) and \( u^4 \), we get
\[ y_4 = x - u \frac{1 + aAu + (bA^2 + cB)u^2}{1 + (a-1)Au + ((b-a-1)A^2 + (c+1)B)u^2} \].

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

\[ y_3 = x - u \frac{(1 + Au)}{1 + (a-1)Au} \).

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)

\[ y_3 = x - u \frac{(1 - Au)}{(1 - 2Au)} \).

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

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

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

\[ y_4 = x - u \frac{(1 - 2Au)}{(1 - 3Au + (A^2 + 3B)u^2)} \),

\[ y_4 = x - u \frac{(1 - 2Au - (A^2 - 2B)u^2)}{(1 - 3Au + 3B u^2)} \).

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

\[ y_4 = x - u \frac{(1 - 3Au + 3B u^2)}{(1 - 4Au + (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).

\[ \star \]

A. Đorđević, N.Z. Klem, G.V. Milovanović, D.S. Mitrović, 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


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.

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

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

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

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

where

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

The expansion (1) is equivalent to the power series of the inverse function.
E.T. Whittaker gave the formula

\[
\gamma_r = x - u - \frac{A u^2}{1_1 A} - \frac{A B u^3}{1_1 A u_1} - \frac{A B C u^4}{1_1 A u_1 A u_1} - \ldots
\]

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

\[
\gamma_3 = x - u / (1 - Au)
\]

E.T. Hamilton provided the method

(3) \[ \gamma_r = x - u \frac{R_{r-1}}{R_r} \]

where

\[
R_1 = R_2 = 1, \quad R_3 = 1 - Au, \quad R_4 = 1 - 2Au + Bu^2, \\
R_5 = 1 - 3Au + (A^2 + 2B)u^2 - Cu^3, \quad \ldots
\]

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

E.D. Durand gave an analogous result:

(4) \[ \gamma_r = x - u \frac{T_{r-1}}{T_r} \]

where

\[
T_1 = 1, \quad T_2 = 1 - 2Au, \quad T_3 = 1 - 3Au + 3Bu^2, \\
T_4 = 1 - 4Au + (2A^2 + 4B)u^2 - 4Cu^3, \\
T_5 = 1 - 5Au + (6A^2 + 5B)u^2 - (5AB + 5C)u^3 + 5Du^4, \quad \ldots
\]

Starting from (1), by means of the formula

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

we get the formula

(6) \[ \gamma_r = x - u \left/ \left( \sum_{k=0}^{r-2} q_k u^k \right) \right. \]

with the coefficients:

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

Equation (6) contains the Traub formula

\[
\gamma_4 = x - u / (1 - Au - (A^2 - B)u^2).
\]
Starting from (6), by means of equation (5), we get the continued fraction

\[ \mathcal{y}_r = x - \frac{u}{s_2} - \frac{u}{s_3} - \ldots - \frac{u}{s_{r-1}} - \frac{u}{s_r} \]

where

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

Formula (7) contains the Halley formula

\[ \mathcal{y}_3 = x - \frac{u}{1 - Au} \]

as well as the Durand formula

\[ \mathcal{y}_4 = x - \frac{u (A - (A^2 - B)u)}{(A - (2A^2 - B)u)} \]

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

\[ 1 + aAu + (bA^2 + cB)u^2 + (dA^3 + eAB + gC)u^3 + \ldots, \]

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

\[ (9) \quad \mathcal{y}_r = x - u \left( \frac{\sum_{k=0}^{r-2} v_k u^k}{\sum_{k=0}^{r-2} w_k u^k} \right) \]

where

\[ v_0 = 1, \quad w_0 = 1, \]

\[ v_1 = aA, \quad w_1 = (a-1)A, \]

\[ v_2 = bA^2 + cB, \quad w_2 = (b-a-1)A^2 + (c+1)B, \]

\[ v_3 = dA^3 + eAB + gC, \quad w_3 = (d-b-a-2)A^3 + (e-c+a+3)AB + (g-1)C, \ldots \]

Formula (9) contains Slavić's formulas

\[ \mathcal{y}_3 = x - u \frac{(1+aAu)}{(1+(a-1)Au)} \]

\[ \mathcal{y}_4 = x - u \frac{1 + aAu + (bA^2 + cB)u^2}{1 + (a-1)Au + ((b-a-1)A^2 + (c+1)B)u^2} \]

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
\[ y_r = x - \frac{t u}{(t-1 + \left(1 + \sum_{k=1}^{r-2} t q_k u^k\right))^n} \]

Upon squaring the expression in brackets we get

\[ (10) \ y_r = x - \frac{t u}{\left(t-1 + \left(\sum_{k=0}^{r-2} h_k u^k\right)^{1/2}\right)} \]

where

\[ h_0 = 1, \quad h_1 = -2tA, \quad h_2 = t(t-2)A^2 + 2tB, \]
\[ h_3 = 2t(t-2)A^3 - 2t(t-3)AB - 2tC, \ldots \]

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

\[ y_3 = x - \frac{2u}{(1 - (1-4Au)^{1/2})} \]

for \( t = 1 \) the Ostrowski formula or the Durand formula

\[ y_3 = x - \frac{u}{(1 - 2Au)^{1/2}} \]

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

\[ y_3 = x - \frac{n u}{(1 + ((n-1)^2 - 2n(n-1)Au)^{1/2})} \]

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

\[ y_3 = x - \frac{t u}{(t-1 + (1-2tAu)^{1/2})} \]

and for \( t = 2 \) the Traub formula

\[ y_4 = x - \frac{2u}{(1 + (1-4Au + 4Bu^2)^{1/2})} \]

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

\[ \]

A.Dorđević, 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 suggestions.

REFERENCES


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


In the Choice of the Initial Approximation in Solving of the Operator Equations by the Newton-Kantorović Method

Milenko Cojbašić

Abstract:

The iterative procedure (see [2]) for the choice of the initial approximation is generalized for the case of solving the equation \( P(x) = 0 \), where \( P \) is a Frechet differentiable operator in a Banach space \( X \). Separately, we consider the case when \( P \) is an integral operator. A numerical example is given.

1. Introduction

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

\[ P(x) = 0, \]

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

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

starting from some suitable chosen \( x_0 \in X \). The sufficient conditions for the success of this procedure are given by the famous theorem of L.V. Kantorović [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 \( \|P_0\| \leq B_0 \)

2) \( \|P(x_0)\| \leq \eta_0 \)

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

4) The constants \( B_0, \eta_0, K \) satisfy the inequality

\[ h = \frac{B_0^2 \eta_0 K}{2}. \]

Then the equation (1) has the solution \( x^* \), which can be found in the ball defined by
and the successive approximants $x_n$ of the iterative procedure (2) converge to $x^*$. For the rapidity of convergence is valid
\[ ||x_n-x^*|| \leq \frac{1}{2^{n-1}} (2h_0)^{2n-1} \cdot n_0. \]

Now let the operator $P$ be integral operator defined by
\[ y(s) = x(s) - \int K(s,t,x(t))dt; \]
and the sequence $x_n(s)$ is formed in the next way: the initial approxima­tion $x_0(s)$ is given. The next approximation $x_1(s)$ is defined from the linear integral equation
\[ x_1(s) - x_0(s) - \int K_x(s,t,x_0(t))(x_1(t) - x_0(t))dt = \epsilon_0(s), \]
where
\[ \epsilon_0(s) = \int K(s,t,x_0(t))dt - x_0(s). \]
The inequality (3) in this case becomes
\[ h = (B+1)^2 \cdot n.K \leq \frac{1}{2}, \]
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
\[ \int G(s,t)dt \leq B; \quad 0 \leq s \leq 1, \]
where $n.K$ 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
\[ x_{n+1} = x_n - [P^*(x_n)]^{-1} [P(x_n) - \alpha_n P(x_0)], \quad (n=0,1,...), \]
where
\[ \alpha_n = \max[0,1 - \frac{1}{2K ||P(x)\|} \left( \frac{1}{||P^*(x_n)\|^{-1}} \right)^2 + \frac{3}{4} \sum_{i<n} \frac{1}{||P^*(x_i)\|^{-1}} \right]. \]

The equation (7) can be taken as the realization of the Newton-Kantorovič method for the equation
\[ P(x)\cdot \alpha_n P(x_0) = 0, \quad \alpha_n \in [0,1]. \]

**LEMMA 1.** If the operator $[P^*(x_0)]^{-1}$ exists then:
(a) The condition (3) is satisfied for each $x_n$, which is obtained by the Newton-Kantorovič method for the equation (9); i.e. exists $\left[ P'(x_n) \right]^{-1}$ and

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

(b) $\alpha_n$ is non increasing sequence; 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

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

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

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

Using the Banach theorem we conclude that the operator

$$H = (I - \left[ P'(x_{n+1}) \right]^{-1} \cdot P'(x_{n+1}))$$

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

$$\| H^{-1} \| = \| \left[ P'(x_{n+1}) \right]^{-1} \cdot P'(x_{n}) \| < 2,$$

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

$$\| \left[ P'(x_{n+1}) \right]^{-1} \| \leq 2 \| \left[ P'(x_{n}) \right]^{-1} \| \quad (13)$$

Using (13) we get

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

Now, using the analogous Taylor's formulae (see [2]) for differentiable operators we find

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

Finally using (11) and (14) we prove that the inequality (10) 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 $\left[ P'(x) \right]^{-1}$ 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
The Newton-Kantorovič method is satisfied, and \( \alpha_n = 0 \), for \( n > n_0 \).

**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

\[ \alpha_n \in [0, 1], \]

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 \( P(x^*) = \in [0, 1] \). Therefore is \( \alpha P(x_0) \in P(G) \). But then

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

and \( x_n \to x \). So

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

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

\[ n \geq n_0 \geq \frac{2KL^2}{P(x_0)^2}, \]

\( \alpha_n = 0 \) 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

\[ \Delta x_n(s) = \int_0^1 K(s,t,x_n(t)) \Delta x_n(t) dt = e_n(s) - \alpha_n e_0(s), \]

where

\[ e_n(s) = \int_0^1 K(s,t,x_n(t)) dt - x_n(s). \]

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

\[ \Delta x_n(s) = e_n(s) - \alpha_n e_0(s) - \int_0^1 G_n(s,t)(e_n(s) - e_0(s)) dt, \]

and \( G_n(s,t) \) is the resolvent of the integral equation with the kernel \( K_n(s,t,x_n(t)) \). 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

\[ x(s) = 1 - 0.4854 s + s^2 + \int_0^s t \arctan(t) \, dt, \]

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

\[ k(s,t) = K_n(s,t,x_0(t)) = \frac{st}{3}, \]

whence
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 $B, n, K, h_0$

$$B = \max_{s} \int |G(s,t)|dt = \frac{3}{10}, n = \max_{s} |G(s)| = 0.9073,$$

$$K = \max_{s,t,u} |K_n(s,t,u)| = 0.6495, h_0 = (B+1)^2 nK = 0.9959 \frac{1}{2}.$$

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

$$\Delta x_0(s) = 0.5021 s^2 + 0.0195 s; x_1(s) = x_0(s) + \Delta x_0(s) = 1 + 0.0195 s + 0.5021 s^2.$$

Likely for $x_1(s)$ we define the constants $n_1$ and $B_1$: $n_1 = 0.4416$, $B_1 = 0.2222$ (see [5]). Now it is

$$h_1 = (B_1+1)^2 n_1 K = 0.4284 < \frac{1}{2}.$$

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

$$\max_{s} |x^*(s) - x_2(s)| = \max_{s} |0.0060| = 0.06 < 10^{-2}.$$

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 immediately. Here $x_1(s) = s^2 + 0.0067 s + 1$.

REFERENCES

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 convergence.

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
\begin{equation}
1 = \int_{-1}^{1} \sigma(y) G(x,y) \, dy \quad -1 < x < 1
\end{equation}

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

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 improvement of the convergence the Richardson extrapolation technique can be used.

\section{Statement of the Problem}

We are trying to find
\begin{equation}
Q = \int_{-1}^{1} \sigma(y) \, dy
\end{equation}

where $\sigma(y)$ is the solution of the equation (1). The kernel $G(x,y)$ has a logarithmic singularity
\begin{equation}
G(x,y) = C \ln|x-y| + K(x,y)
\end{equation}

where $K(x,y)$ is a continuous function. It is known that the
s\text{olution has singularities at the both ends of the interval } [1, 1] \text{ and } a(y) \text{ can be represented as (3)}

\[ a(y) = \frac{f(y)}{\sqrt{1 - y^2}} \]

where \( f(y) \) is the continuous 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

\[ 1 = \int_0^\pi S(a) \left[ \ln |\cos x - \cos y| + H(\alpha, \beta) \right] \, d\alpha \]

where \( S(\alpha) = a(\cos x) \sin x \) and \( H(\alpha, \beta) \) are continuous functions. The kernel can be rewritten

\[ \ln |\cos x - \cos y| = \ln \left| \frac{\sin x - \beta}{\alpha - x} \right| + \ln \left| \frac{\sin x + \beta}{2 - (\omega - x)(\omega + x)} \right| + \ln |\omega - \beta| + \ln |\omega + \beta| + \ln |2\pi - \omega| \]

where the first two terms are continuous, the last three terms are singular. Continuous 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
\[
\int_0^\pi \ln|\cos \beta_i - \cos \alpha| \, d\alpha = \sum_{j=1}^{n} a_{ij} S(\beta_j)
\]

where

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

and

\[a_{ij} = \frac{\sin \frac{\beta_i - \beta_j}{2}}{(2\pi - \beta_i - \beta_j)(\beta_i + \beta_j)} \ln \frac{\sin \frac{\beta_i + \beta_j}{2}}{\beta - \beta_i} + \frac{3h \ln h + h|\phi_0(i-j) + \phi_0(1-i-j) + \phi_0(2n-i-j+1)|}{2} + 3h \ln h + h|\phi_0(i-j) + \phi_0(1-i-j) + \phi_0(2n-i-j+1)|\]

\[\phi_0(x) = \int_0^1 \ln|1+\frac{1}{2} - u| \, du = (1+\frac{1}{2})\ln|1+\frac{1}{2}| - (1-\frac{1}{2})\ln|1-\frac{1}{2}| - 1\]

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

\[(4) \quad Q = Q(h) + A h^\alpha + B h^{\alpha+1} + ...\]

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

\[Q \approx Q(h) + A h^\alpha\]

\[Q \approx Q(\frac{h}{2}) + A(\frac{h}{2})^\alpha\]

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 \(\alpha\).
Nevertheless we can compute experimently for some cases, for which 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

$$\int_{-1}^{1} f(y) dy = -\frac{1}{\ln 2} = -1.442695$$

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

<table>
<thead>
<tr>
<th>n</th>
<th>Q</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.56525</td>
<td>4.6</td>
</tr>
<tr>
<td>2</td>
<td>-1.47283</td>
<td>4.01</td>
</tr>
<tr>
<td>4</td>
<td>-1.45021</td>
<td>4.008</td>
</tr>
<tr>
<td>8</td>
<td>-1.44457</td>
<td>4.03</td>
</tr>
<tr>
<td>16</td>
<td>-1.44316</td>
<td></td>
</tr>
</tbody>
</table>

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

The well-known Richardson's elimination gives the table

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.56525</td>
<td>-1.44202</td>
<td></td>
</tr>
<tr>
<td>-1.47283</td>
<td>-1.44267</td>
<td></td>
</tr>
<tr>
<td>-1.45021</td>
<td>-1.44269</td>
<td></td>
</tr>
<tr>
<td>-1.44457</td>
<td>-1.44269</td>
<td></td>
</tr>
<tr>
<td>-1.44316</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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


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

WALTER GAUTSCHI AND GRADIMIR V. MILOVANOVIĆ

ABSTRACT:

We 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 \( \theta_{n-1} \) is a real constant. We show that all zeros of \( p_n \) lie in the upper half disc \( |z| < a \wedge \Im z > 0 \), if \( 0 < \theta_{n-1} < \pi_n(a)/\pi_{n-1}(a) \), and in the lower half disc \( |z| < a \wedge \Im z < 0 \), if \( -\pi_n(a)/\pi_{n-1}(a) < \theta_{n-1} < 0 \). The ultraspherical weight function is considered as an example.

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. \textbf{Let}

\begin{align*}
    f(x) &= (x-x_1)(x-x_2)\ldots(x-x_n), \\
    g(x) &= (x-y_1)(x-y_2)\ldots(x-y_n),
\end{align*}

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

Here we consider special linear combinations of the form

\begin{equation}
    p_n(z) = \pi_n(z) - i\theta_{n-1}\pi_{n-1}(z),
\end{equation}

where \(\{\pi_k\}\) is a system of monic polynomials orthogonal with respect to an even weight function on \((-a,a), 0 < a < \infty\), and \(\theta_{n-1}\) 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 \(\theta_{n-1}\), 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{equation}
    \begin{cases}
    \pi_{k+1}(z) = z\pi_k(z) - \beta_k\pi_{k-1}(z), & k=0,1,\ldots, \\
    \pi_{-1}(z) = 0, \pi_0(z) = 1,
    \end{cases}
\end{equation}

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

\[ p_n(z) = z^n - i\theta_{n-1}z^{n-1} + \ldots, \]

so that

\[ \sum_{k=1}^{n} \xi_k = i\theta_{n-1}, \]

hence

\[ \sum_{k=1}^{n} \text{Im } \xi_k = \theta_{n-1}, \]
here $\xi_1, \xi_2, \ldots, \xi_n$ 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

$$\text{Im } z > 0, \quad -a < \text{Re } z < a \quad \text{if } \theta_{n-1} > 0,$$

or

$$\text{Im } z < 0, \quad -a < \text{Re } z < a \quad \text{if } \theta_{n-1} < 0,$$

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 $\partial D_a$ its boundary. We first prove the following auxiliary result.

**Lemma.** For each $z \in \partial D_a$ one has

$$|\pi_k(z)| > |\pi_k(a)|, \quad k = 1, 2, \ldots.$$

**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 $z \in \partial D_a$,

$$|r_k(z)| \geq r_k, \quad z \in \partial D_a.$$

From the recurrence relation (2.1) there follows

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

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

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

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

$$r_k = a - \frac{\beta_{k-1}}{a - \frac{\beta_{k-2}}{a - \frac{\beta_{k-3}}{a - \ldots}}}, \quad k \geq 1.$$

It is easily seen that $r_k = \pi_k(a)/\pi_{k-1}(a)$, $k \geq 1$. Indeed, using (2.1),
\[ r_k = a - \frac{\beta_{k-1}}{r_{k-1}} = a - \beta_{k-1} \frac{\pi_{k-2}(a)}{\pi_{k-1}(a)} = \frac{\pi_k(a)}{\pi_{k-1}(a)} . \]

By a similar argument one could show that
\[ 2a - \frac{\pi_k(a)}{\pi_{k-1}(a)} \geq \left| \frac{\pi_k(z)}{\pi_{k-1}(z)} \right| , \quad z \in \partial D_a', \]

but this will not be needed in the following.

**Theorem.** If the constant \( \theta_{n-1} \) 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
\[ |z| < a \land \text{Im} \, z > 0 . \]

If \( -\pi_n(a)/\pi_{n-1}(a) < \theta_{n-1} < 0 \), then all zeros of (1.1) are in the lower half disc
\[ |z| < a \land \text{Im} \, z < 0 . \]

**Proof.** By (2.4) we have
\[ \left| \frac{\pi_n(z)}{\pi_{n-1}(z)} \right| \geq \frac{\pi_n(a)}{\pi_{n-1}(a)} , \quad z \in \partial D_a' , \]

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 Rouche'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
\[ p_k(z) = \frac{k!}{2^k(k!)^2} C_k^\lambda(x) , \]

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

\[
\frac{\pi_k^{(1)}}{\pi_{k-1}^{(1)}} = \frac{k}{2(\lambda k-1)}, \quad \frac{C_k^{(1)}}{C_{k-1}^{(1)}} = \frac{2\lambda+k-1}{2(\lambda k-1)}
\]

and

\[
D^m C_k^\lambda(x) = 2^m (\lambda) C_{k-m}^{\lambda+m}(x), \quad m \leq k,
\]

where D is the differentiation operator, our theorem implies the following.

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

The upper bound \( \frac{2\lambda+n-1}{2(\lambda+n-1)} \) 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**


ON THE OPTIMAL CIRCULAR CENTERED FORM

Ljiljana D. Petković

ABSTRACT:

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

\[
R = R_0 = \max_{z \in Z} |f(z) - f(c)| .
\]

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 (\( e^z \), \( \ln z \), \( \arctan z \), \( \sin z \), \( \cos z \), \( z^n \), \( z^{1/n} \)). 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 \( |f(\xi)| \leq |f(w)| \) for each \( \xi \in L \) then

\[
\max_{\xi \in L} |f(\xi)| = |f(w)| .
\]

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

\[
|e^z - 1| \leq |z| - 1
\]

is valid for arbitrary \( z \in \mathbb{C} \).

**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

\[
\left| \int_a^b f(t) \, dt \right| \leq \int_a^b |f(t)| \, dt .
\]

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

\[
\left| \sum_{k=0}^{+\infty} a_k z^k \right| \leq \sum_{k=0}^{+\infty} a_k |z|^k \quad \text{for} \quad |z| < A .
\]

Let \( r \) be the boundary of the disk \( Z = \{ c; r \} \), \( c = |c|e^{i\pi} \) 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.

$\hat{z}(z) = e^z$.

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

$$|\ln(1 + w)| = \left| \int_0^p \frac{e^{iw}}{1 + te^{iw}} dt \right| = \left| \int_0^p \frac{dt}{1 + te^{iw}} \right| \leq \left| \int_0^p \frac{1}{1 + te^{iw}} dt \right| = \frac{\pi}{2\sin(\omega)} \leq -\ln(1 - p).$$

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

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

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

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 \( Z = \{ c ; r \} \in U \subseteq U \). Then, for the closed united extension \( f^*(Z) \) we have

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

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 \( \Gamma \) 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 \( Z \in U \subseteq K(C) \), 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, \( F_0(Z) \subseteq F_T(Z) \) (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

$$F_p(Z) = \left\{ \sum_{k=1}^{+\infty} \frac{|f^{(k)}(0)|}{k!} (|c| + r)^k - |c|^k \right\} \tag{5}$$

or

$$F_p(Z) = (f(c); R_p) \supseteq (f(z); z \in Z) = f^*(Z),$$

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 $L$-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.

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

**Example 1.** Let

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

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

$$Q_T(Z) = \left\{ q(c); \sum_{k=1}^{n} \frac{|q^{(k)}(c)| r^k}{k!} \right\}$$
\[ Q_p(Z) = \{ q(c) ; \sum_{k=1}^{n} |a_k|((|c| + r)^k - |c|^k) \}. \]

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 conjecture (7) remains as an open problem.

**References**

TWO METHODS FOR THE CURVE DRAWING IN THE PLANE

Dobrilu D. 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 \( y' = -\frac{F_x}{F_y} \), which enables 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 realized in BASIC.

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 asymptotes, number of branches, etc. Thus we can draw a sketch 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 \( M_0(x_0, y_0) \) 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 \( F_x(x_0, y_0) \) and \( F_y(x_0, y_0) \). If \( |F_x(x_0, y_0)| \leq |F_y(x_0, y_0)| \), 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

\[
x_1 = x_0 + S_x h, \quad y_1 = y_0 + S_y h \frac{F_x(x_0, y_0)}{F_y(x_0, y_0)}.
\]

If \( |F_x(x_0, y_0)| > |F_y(x_0, y_0)| \), i.e. \( |y'| > 1 \), then

\[
x_1 = x_0 + S_y h \frac{F_y(x_0, y_0)}{F_x(x_0, y_0)}, \quad y_1 = y_0 + S_y h.
\]

The position of the predicted point \( M_1 \) can be corrected by the Newton-Raphson method. If
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 $\text{sgn}(x_1 - x_0) > 0$, then $S_x = 1$, and if $\text{sgn}(x_1 - x_0) \leq 0$, 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 at 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 successfully 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 TPLIT (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.

\[ 1^o \quad x^5 - x^3 y - x y^2 + y^3 = 0. \]

Random \hspace{2cm} Implicit

\[ 2^o \quad (y^2 - x^2)(x - 1)(x - \frac{3}{2}) = 2(y^2 + x^2 - 2x)^2. \]

Random \hspace{2cm} Implicit
OPERATING WITH A SPARSE NORMAL EQUATIONS MATRIX

Brankica Cigrovski, Miljenko Lapaine, Svetozar Petrović

ABSTRACT:
In natural and applied sciences, especially in geodesy, symmetric normal equations matrices $Q = A^T A$ containing relatively few nonzero elements occur quite often. Then, one frequently rearranges the matrix $Q$ (interchanging columns and rows simultaneously) to bring it to a form more convenient for further treatment. The authors of the present paper have come to see that in such cases it would make sense to rearrange first the original matrix $A$, and only then to form $Q = A^T A$. So they have elaborated one such procedure and tested it.

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
B

\[ M_p = \pm \sqrt{\frac{f^t Q^{-1} f}{n}} \]

(e.g. [7] p. 132, [1] p. 257, [11] p. 3). The symmetric n\times n-matrix \( Q=A^t A \) is the so-called correlation matrix (the normal equations matrix), the m\times n-matrix \( A \) being a condition equations matrix, and \( f \) is a n\times 1-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 \( f^t Q^{-1} f \) 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 \( f^t Q^{-1} f \) 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 matrix \( Q \).

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 special form. Its dimension was up to $666 \times 180$, but each column contained only 6 nonzero elements, each of them being either $+1$ or $-1$.

\[
\begin{array}{cccccc}
-2 & -10 & 135 & -6 & -26 & \ldots \\
4 & 13 & 162 & 8 & 28 & \ldots \\
9 & 91 & -148 & 128 & -172 & \ldots \\
-13 & -97 & -165 & 29 & 174 & \ldots \\
107 & -107 & -662 & -170 & -190 & \ldots \\
-118 & 110 & 666 & 172 & 193 & \ldots \\
\end{array}
\]

Fig. 1 Storing the matrix $A$

Thus it was possible to accommodate that matrix into an array not greater than $6 \times 180$. Fig. 1 presents a 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 $Q=AA^t$ 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.

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 accommodate 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=AT A. 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=AT A) 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 preceding 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_{ij} of the matrix Q i-th column are zero for \( i<j \) and for \( i>k \), i.e. with A becoming "more diagonal", Q also becomes such and its profile reduces.

![Fig.3 a. the original matrix A](image)

![Fig.3 b. A after completion of the first step (row interchanges)](image)

![Fig.3 c. A after second step (column interchanges)](image)

![Fig.3 d. i-th column of the matrix Q formed from rearranged A (c.)](image)

(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 non-zero 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 preceding 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.

<table>
<thead>
<tr>
<th>n</th>
<th>number of elements of the upper triangle total</th>
<th>share of nonzeros</th>
<th>upper trian. diagonals with nonzer.</th>
<th>profile orig.</th>
<th>rearr. profile</th>
<th>ratio of profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>190</td>
<td>56</td>
<td>29%</td>
<td>18</td>
<td>9</td>
<td>123</td>
</tr>
<tr>
<td>152</td>
<td>11628</td>
<td>561</td>
<td>4.8%</td>
<td>145</td>
<td>24</td>
<td>5170</td>
</tr>
<tr>
<td>162</td>
<td>13203</td>
<td>755</td>
<td>5.7%</td>
<td>160</td>
<td>35</td>
<td>6269</td>
</tr>
<tr>
<td>173</td>
<td>15051</td>
<td>773</td>
<td>5.1%</td>
<td>171</td>
<td>34</td>
<td>6863</td>
</tr>
<tr>
<td>180</td>
<td>16290</td>
<td>631</td>
<td>3.9%</td>
<td>173</td>
<td>24</td>
<td>6852</td>
</tr>
</tbody>
</table>

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


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)


7. MIHAILOVIć K.: Geodenzija II (I part), (Serbocroatian), Gradjevinska knjiga, Beograd 1974


APPROXIMATION OF $2^n$ - PERIODIC FUNCTIONS BY FUNCTIONS OF SHORTER PERIOD

Slobodan D. Miloradović

ABSTRACT:

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

Let $C[a,b]$, as usual 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 $(-\infty, +\infty)$, whose period is $2\pi$ 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 $\Omega_f$. For example, if $f$ is a $2\pi$ - periodic function, we shall write $2\pi \in \Omega_f$.

In this paper we shall find the value of the exact approximation of a fixed $2\pi$ - periodical function $f$, by a functions $\psi$ of period $2\pi m/k$, 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 interval in which the periods of the functions are contained a whole number of times. Thus for example, if $2\pi \in \Omega_f$ and $2\pi/\ell \in \Omega_\Psi$ then we have

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

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

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

$$f_k(x) = \min_{0<s<k-1} f(x + \frac{2\pi s}{k}),$$

Then $f_k$ and $f_k$ are continuous and $2\pi/\ell \in \Omega_\Psi$ and $2\pi/\ell \in \Omega_\Psi$. 

**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

$$|x'-x''| < \delta(\varepsilon) = |f(x') - f(x'')| < \varepsilon$$

According to equality

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

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

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

that is

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

hence

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

that is

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

which means that $f_k$ is continuous. Accordingly it is proved in the same way that $f_k$ is continuous. Since $2\pi \in \Omega_f$, it follows that

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

that is $2\pi/\ell \in \Omega_\Psi$. Similarly, $2\pi/\ell \in \Omega_\Psi$.

**Theorem 2.** If $f \in C$, $2\pi \in \Omega_f$, $d_k = \frac{f_k - f_k}{2}$, then

$$\inf_{\psi,2\pi/\ell \in \Omega_\Psi} ||f - \psi||_{C[0,2\pi]} = ||d_k||_{C[0,2\pi/\ell]}.$$

Proof. According to lemma 1 the functions
\[ d_k = \frac{T_k - f_k}{2} \] and \[ \psi_k = \frac{T_k + f_k}{2} \]
are from C and \( 2\pi/k \in \Omega_{d_k} \), \( 2\pi/k \in \Omega_{\psi_k} \). Since for \( x \in [0,2\pi] \)
\[ f(x) - \psi_k(x) < T_k(x) - \psi_k(x) = d_k(x), \]
\[ f(x) - \psi_k(x) > f_k(x) - \psi_k(x) = -d_k(x), \]
then
\[ |f(x) - \psi_k(x)| \leq d_k(x) \]
and
\[ ||f - \psi_k||_{C[0,2\pi]} \leq ||d_k||_{C[0,2\pi/k]} \] (4)
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 \( \psi \in \mathcal{C}, \frac{2\pi}{k} \in \Omega_{\psi} \), we have
\[ ||f - \psi||_{C[0,2\pi]} \geq \max_{0 \leq s < k-1} |f(x_0 + \frac{2\pi s}{k}) - \psi(x_0 + 2\pi s/k)| = \]
\[ \max_{0 \leq s < k-1} |f(x_0 + \frac{2\pi s}{k}) - \psi(x_0)| = \]
\[ \max_{0 \leq s < k-1} \left( \max \left( |f(x_0 + \frac{2\pi s}{k}) - \psi(x_0)|, \psi(x_0) - f(x_0 + \frac{2\pi s}{k}) \right) \right) = \]
\[ \max_{0 \leq s < k-1} \left( \max \left( f(x_0 + \frac{2\pi s}{k}) - \psi(x_0), \psi(x_0) - f(x_0 + \frac{2\pi s}{k}) \right) \right) = \]
\[ \max \left( T_k(x_0) - \psi(x_0), \psi(x_0) - f_k(x_0) \right) \geq d_k(x_0) = ||d_k||_{C[0,\frac{2\pi}{k}]} \]
that is
\[ ||f - \psi||_{C[0,2\pi]} \geq ||d_k||_{C[0,\frac{2\pi}{k}]} \] (5)
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 \( f \in \mathcal{C}, \frac{2\pi}{k} \in \Omega_{\mathcal{F}}, M = \max f(x), m = \min f(x) \), then
\[ \inf_{\psi, \frac{2\pi}{k} \in \Omega_{\psi}} ||f - \psi|| \leq \frac{M-m}{2}. \] (6)
We have equality in (6) for such function \( f \) for which there is a point \( x_0 \) in which \( M = T_k(x_0) \) and \( m = f_k(x_0) \).
Corollary 4. If \( f \in C, 2\pi \in \Omega_f \), then

\[
\lim_{k \to \infty} \psi_2 \frac{2\pi}{k} \in \Omega_{\psi} = \frac{M-m}{2}.
\]

The sequence \( \left( \inf_{\psi} \frac{2\pi}{k} \in \Omega_{\psi} \right) \) 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

\[
f_k(x) = \max_{0 \leq s \leq k-1} f(x + 2\pi ms/k),
\]

\[
f_k(x) = \min_{0 \leq s \leq k-1} f(x + 2\pi ms/k),
\]

\[
d_k = \frac{f_k - f_k}{2},
\]

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

Theorem 5. If \( f \in C, 2\pi \in \Omega_f \), then

\[
\inf_{\psi} \frac{2\pi m}{k} \in \Omega_{\psi} = \left\| d_k \right\|.
\]

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

\[
\left\{ \psi : \frac{2\pi}{k} \in \Omega_{\psi} \right\} \subseteq \left\{ \psi : \frac{2\pi m}{k} \in \Omega_{\psi} \right\},
\]

and for \( f \in C, 2\pi \in \Omega_f \),

\[
\left\| f - \psi \right\|_{C[0,2\pi]} \geq \left\| f - \psi \right\|_{C[0,2\pi]}
\]

It could be expected that strict that strict inequality is not excluded in (8). We shall prove, however, 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

\[
m, 2m, \ldots, (k-1)m
\]

by \( k \), we obtain the sequence of remainders

\[
r_1, r_2, \ldots, r_{k-1}
\]
Then (10) is a permutation of the sequence

(11) 1, 2, $\ldots$, 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 $l_m$ and $n_m$, $1 < n < k-1$, which when divided by k give the same remainder:

\[ l_m = kp + r_n, \]
\[ n_m = kq + r_n, \quad p, q \in \mathbb{N} \cup \{0\}, \quad r_n < k, \quad r_n \in \mathbb{N}. \]

By subtracting the first equality from the second equality we get

\[ (n-1)m = k(g-p). \]

Hence, considering the fact that numbers m and k are mutually it follows that the number $n-1$ is divisible by k, and that does not agree with inequality $0 < n-1 < k$.

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

\[ \left\| f - \psi \right\|_{C[0, 2\pi]} = \max_{0 < s < k-1} f(x + \frac{2\pi ms}{k}). \]

**Proof.** Due to lemma 6 $m \cdot s = g_s \cdot k + r_s$, $g_s \in \mathbb{N} \cup \{0\}$, 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

\[ f_k(x) = \max_{0 < s < k-1} f(x + \frac{2\pi ms}{k}). \]

Also $f_k(x) = f_k(x)$ and $\frac{2\pi}{k} \in \Omega_k$. 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

\[ \left\| d_k \right\|_{C[0, \frac{2\pi}{k}]} = \left\| d_k \right\|_{C[0, \frac{2\pi m}{k}]} = \left\| d_k \right\|_{C[0, \frac{2\pi m}{k}]}.

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

If $m \geq k$ 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 $2\pi$, also. But theorem 7 confirms that approximation of $2\pi$ - periodic function whose periods are larger than $2\pi$ are equal with approximations
of functions whose periods are smaller than $2\pi$ which is in accord with the title of this paper.

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

\[ \inf_{\psi} \left\| f - \psi \right\|_{C[0, 2\pi]} = \begin{cases} 1, & n = 2s, \ s \in \mathbb{N} \\ 0, & n = 2s+1, \ s \in \mathbb{N} \end{cases} \]

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

\[ f_k(x) = \begin{cases} \cos x, & x \in [0, \frac{\pi}{2s}] \\ \cos (x + \frac{2s-1}{s} \pi), & x \in [\frac{\pi}{2s}, \frac{\pi}{s}] \end{cases} \]

\[ f_k(x) = \begin{cases} \cos (x + \pi), & x \in [0, \frac{\pi}{2s}] \\ \cos (x + \frac{3s-1}{s} \pi), & x \in [\frac{\pi}{2s}, \frac{\pi}{s}] \end{cases} \]

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

and $\left\| d_k \right\|_1 = 1$. According to theorem 2 we get the first part of equality (14). If $k = 2s+1$. Then we get

\[ f_k(x) = \cos (x + \frac{2s}{2s+1}), \ x \in [0, \frac{2\pi}{2s+1}] \]

\[ d_k(x) = \begin{cases} \sin (x + \frac{\pi s}{2s+1}) \cos \frac{\pi}{2(2s+1)}, & x \in [0, \frac{\pi}{2s+1}] \\ \sin (x + \frac{\pi (s-1)}{2s+1}) \cos \frac{\pi}{2(2s+1)}, & x \in [\frac{\pi}{2s+1}, \frac{2\pi}{2s+1}] \end{cases} \]

and $\left\| d_k \right\|_1 = \cos \frac{\pi}{2(2s+1)}$, which proves the other part of equality (14).

References

[1] МИЛОРАДОВИЧ С. О коэффициентах Фурье класса $W^1 H[\delta_0]$. - Publications de l'institut mathématique, Beograd, tome
ON THE STRONG SUMMABILITY \((C, \alpha)\) 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 \((C, \alpha)\) of transformations of simple and multiple trigonometric Fourier series.

JAKOJ SUMABILNOSTI \((C, \alpha)\) TRANSFORMACIJA PROSTIH I VIŠESTRUKIH TRIGONOMETRIJSKIH FURIJEOVIH REDOVA. U radu su dokazane nejednakosti iz kojih sleduje niz rezultata o jakoj sumabilnosti \((C, \alpha)\) transformacija prostih i višestrukih trigonometrijskih Furijeovih redova.

Definition 1. Let \(f \in L[-\pi, \pi]\),

\[
\frac{a_0}{2} + \sum_{k=1}^{\infty} \left[ \alpha_k \cos kx + \beta_k \sin kx \right]
\]

Fourier series of \(f\), \(s_n(x, y)\) the sequence of its partial sums, \(\alpha > 0\), \(n\) natural number or zero,

\[
A^\alpha_n = \left( \frac{n+\alpha}{n} \right),
\]

\[
(2) \quad a_n^\alpha(x, y) = \frac{1}{A_n^\alpha} \sum_{l=0}^{n} A_{n-l}^\alpha s_l\left( \frac{\alpha}{\pi} x + \frac{\beta}{\pi} y \right) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x+t) + f(x-t) K^\alpha(t) dt
\]

\((C, \alpha)\) transformation of the series (1), \(T = (\alpha_n, \beta)\) regular matrix and \(p > 0\). If

\[
\lim_{n \to \infty} \sum_{k=0}^{\infty} |a_n, k| \left| g^\alpha_{n, k}(x, y) - f(x) \right|^p = 0
\]

we say that series (1) \((H, p, T, \alpha)\) (or strongly summable in point \(x\) towards \(f(x)\)). In addition to the above-stated, if \(f \in C[-\pi, \pi]\) and

\[
\lim_{n \to \infty} \left\| \sum_{k=0}^{\infty} |a_n, k| \left| g^\alpha_{n, k}(x, y) - f(x) \right| \right\|_p = 0
\]

we say that series (1) \((H, p, T, \alpha)\) (or strong) is summable uniformly at \([-\pi, \pi]\) towards the \(f\) function.
Theorem 1. Let \( \varepsilon \in [1-K, 1] \) and \( \| \varphi \|_{\infty} < 1 \), then for each \( \rho > 0 \), \( K > 1 \), and \( \lambda > 0 \)

\[
\frac{1}{K+1} \sum_{k=1}^{2^{n+1}-1} \left| \Phi_{\varepsilon}(x, y) \right|^p \leq C(\rho, \lambda) \| \varphi \|_{\infty}^p
\]

where \( C(\rho, \lambda) \) is a positive constant depending only on \( \rho \) and \( \lambda \).

Proof. It is sufficient to examine the case \( 0 < \lambda < 1 \). From (2), putting it as

\[
\Phi_{\varepsilon}(x, y) = \frac{A}{K} \left( \int_{0}^{A} \int_{0}^{A} \right),
\]

and from

\[
\left| \Phi_{\varepsilon}(x, y) \right| < A(\lambda) \nu - \delta \nu - (\lambda + 1)
\]

for \( \lambda = 1, 2, 3, \ldots \); \( 0 < \nu < \frac{1}{2} \) there follows

(3) \[
\left| \frac{A}{K} \int_{0}^{A} \int_{0}^{A} \right| < 2 C, \quad \left| \frac{A}{K} \int_{0}^{A} \int_{0}^{A} \right| < A(\omega) \nu
\]

where \( A(\lambda), A_1(\lambda) \) are the positive constants depending only on \( \lambda \).

From inequality

\[
|\alpha + \beta|^p \leq 2^p (|\alpha|^p + |\beta|^p) \quad (\forall \alpha, \beta \in \mathbb{R}) (\forall \rho > 0)
\]

and (3) there follows

\[
-A \sum_{n=2^{n+1}-1}^{2^{n+1}} \left| \Phi_{\varepsilon}(x, y) \right| ^p \leq 2^p (2^1 + A_1(\lambda) \nu) M.
\]

Theorem 2. Let \( \varepsilon \in \mathbb{C}[K, K] \) then for each \( \rho > \varepsilon, K > 1, \lambda > 0 \)

\[
\sum_{n=2^{n+1}-1}^{2^{n+1}} \left| \Phi_{\varepsilon}(x, y) - \Phi(\zeta) \right| ^p \leq C_1(\rho, \lambda) \sum_{n=1}^{\infty} \left| E_\nu(\xi) \right| ^p \lambda
\]

where \( \Phi_{\varepsilon}(x, y) \) is any non-negative, not-increasing numbers, \( E_\nu(\xi) \)

the best approximation of the \( f \) function to the trigonometric polynomials of the degree \( \lambda \) in distance of the space \( \mathbb{C}[K, K] \), and \( C_1(\rho, \lambda) \) a positive constant depending only on \( \rho \) and \( \lambda \).

Proof. According to Theorem 1 there follows

\[
\sum_{n=2^{n+1}-1}^{2^{n+1}} \left| \Phi_{\varepsilon}(x, y) - \Phi(\zeta) \right| ^p = \sum_{n=1}^{\infty} \sum_{\nu=2^{n+1}-1}^{2^{n+1}} \left| \Phi_{\varepsilon}(x, y) - \Phi(\zeta) \right| ^p
\]

\[
\leq 2^p C(\rho, \lambda) \sum_{n=1}^{\infty} \sum_{\nu=2^{n+1}-1}^{2^{n+1}} \left[ E_\nu(\xi) \right] ^p \lambda
\]

\[
\leq 2^p C(\rho, \lambda) \sum_{n=1}^{\infty} \sum_{\nu=2^{n+1}-1}^{2^{n+1}} \left[ E_\nu(\xi) \right] ^p \lambda
\]

\[
= C_1(\rho, \lambda) \sum_{\nu=2^{n+1}-1}^{2^{n+1}} \left[ E_\nu(\xi) \right] ^p \lambda
\]
Theorem 2. Let $F_n \downarrow 0$ any sequence and $C_F$ collection of
the function $\beta \in C[F_{\lambda}, F_{\lambda}]$ for which $E_n(\beta) \leq F_n$.

Theorem 3. Let $\beta \in C_F$ and $T=(\alpha_n \beta)$ a non-negative regular
matrix the elements of which satisfy the condition
$$\alpha_{n,k+1} \leq \alpha_{n,k} \quad (\forall n = 0,1,2, \ldots)$$
Then, there exists the sequence $\rho(n) \uparrow +\infty$ such that
$$\lim_{n \to \infty} \left\| \left\{ \sum_{\beta=0}^{\infty} \alpha_{n,\beta} \rho(\beta)(\beta, x) - \beta(x) \right\} \right\| = C.$$

Theorem is proved as in (2).

Theorem 4. If $\beta \in C_F$ and $\mathcal{L}$ is any sequence of non-negative,
not increasing numbers, then for each $p > 0, \lambda > 1$ and $s > 0$ the fol-
lowing inequality is correct:

$$\sum_{\nu=2}^{\infty} F^p_{\nu} \lambda_{\nu} \leq \sum_{\beta \in C_F} \left\| \sum_{\nu=2}^{\infty} \left| \beta(\nu, x) - \beta(x) \right| \right\|^{p} \lambda_{\nu} \leq \sum_{\nu=2}^{\infty} F^p_{\nu} \lambda_{\nu}$$

Proof. The right hand inequality results from Theorem 2.

Let $\beta_{3o}(x) = \sum_{n=1}^{\infty} (F_{n-1} - F_n) \cos \frac{\pi x}{n}.$

Then $\beta_{3o}(t) - \beta(0, \beta_{3o}) = F_{\gamma}, \beta_{3o} \in C_F$
and
$$\sum_{\nu=2}^{\infty} \left\| \sum_{\nu=2}^{\infty} \left| \beta_{3o}(\nu, x) - \beta_{3o}(x) \right| \right\|^{p} \lambda_{\nu} \geq \sum_{\nu=2}^{\infty} F^p_{\nu} \lambda_{\nu}$$
by which Theorem 4 is proved.

From inequality (4) it results that the approximation
rate to the strong means, $(C, \mathcal{L}) (\alpha \beta > 0)$ transformations of
trigonometric Fourier series of the function $\beta \in C_F$ (for the whole
$C_F$ collection) cannot be improved.

Similarly, we testify and prove the theorems relating
to $(C, \mathcal{L}) (\lambda > 0)$ transformations of multiple trigonometric
Fourier series.

REFERENCES


2. Gogoladze L. D.: O jakom sumiranju prostih i višestrukih
trigonometrijskih Furijeovih redova, Zbornik radova: "Neka
pitanja teorije funkcija", tom 2. izdanje Tbilisiskog
univerziteta, Tbilisi 1981. str. 5.-30.
ESTIMATION FOR REMAINDER OF ANALYTICAL FUNCTION IN TAYLOR’S SERIES

Petar M. Vasić, Igor Ž. Milovanović and Josip E. Pečarić

ABSTRACT:
In this paper some estimations of difference modul of analytical functions and referred Taylor's polynomials are given. The obtained results are illustrated on a certain concrete f cases, i.e. on exponential, sinhyperbolic and cosinhypertbolic function.

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 \( z \mapsto f^{(k)}(z), k \in \mathbb{N}_0 \), map real axis in real axis. If a natural number \( r \) exist, \( 1 \leq r \leq n \), so that
\[
0 \leq f^{(n+k)}(0) \leq f^{(r+k)}(0), \quad k \in \mathbb{N}
\]
the inequality
\[
(1) \quad \left| \frac{z^{r+1}}{(r+1)!} f(z) - \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} z^k \right| \leq \left( \frac{z^{n+1}}{(n+1)!} f(z) - \sum_{k=0}^{r} \frac{f^{(k)}(0)}{k!} z^k \right)
\]
holds.

Proof. Assume that all conditions given in theorem 1 are fulfilled. Then
\[
|f(z) - \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} z^k| = \left| \sum_{k=n+1}^{+\infty} \frac{f^{(k)}(0)}{k!} z^k \right| \leq \sum_{k=n+1}^{+\infty} \frac{f^{(k)}(0)}{k!} |z|^k
\]
wherefrom the inequality (1) is obtained.

In the similar way the following theorem can be proved.

**THEOREM 2.** Let \( z \mapsto f(z) \) and \( z \mapsto g(z) \) be analytical functions in the circle \( |z| < R \). Let functions \( z \mapsto f^{(k)}(z) \) and \( z \mapsto g^{(k)}(z) \), \( k \in \mathbb{N}_0 \), map real axis in real axis. If a natural number \( r \) exist, \( 1 < r < n \), so that \( 0 < f^{(n+k)}(0) \leq g^{(n+k)}(0) \), \( k \in \mathbb{N} \), then the inequality

\[
\begin{equation}
\left| \frac{z}{(n+1) \ldots (r+2)} \right| f^{(n+1)}(0) + \left| \frac{z}{(r+2)!} \right| f^{(n+2)}(0) + \ldots \end{equation}
\]

\[
\begin{equation}
\left| \frac{z}{(n+1) \ldots (r+2)} \right| f^{(r+1)}(0) + \left| \frac{z}{(r+2)!} \right| f^{(r+2)}(0) + \ldots \end{equation}
\]

\[
\begin{equation}
\left| \frac{z}{(n+1) \ldots (r+2)} \right| (f(|z|) - \sum_{k=0}^{r} \frac{f^{(k)}(0)}{k!} |z|^k) = \sum_{k=0}^{r} \frac{f^{(k)}(0)}{k!} |z|^k,
\end{equation}
\]

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 \( f^{(k)}(0) = 1, k \in \mathbb{N}_0 \), on the basis of the inequality (1) we obtain

\[
\begin{equation}
\left| \frac{z}{(r+1)!} \right| e^z - \sum_{k=0}^{n} \frac{z^k}{k!} \leq \left| \frac{z}{(n+1)!} \right| (e |z| - \sum_{k=0}^{r} \frac{z^k}{k!}) = e^z - \sum_{k=0}^{n} \frac{z^k}{k!}
\end{equation}
\]

The inequality (3) (see [1]) is a generalization of Garnir's inequality (see for example [2, p. 323])

\[
|e^z - (1 + \frac{z}{1!} + \ldots + \frac{z^n}{n!})| \leq \frac{|z|^{n+1}}{(n+1)!} e |z|.
\]
If we put that
\[ T_{2n-1}(z) = z + \frac{z^2}{2!} + \cdots + \frac{z^{2n-1}}{(2n-1)!}, \quad T_{-1}(z) = 0, \]
and
\[ T_{2n-2}(z) = 1 + \frac{z^2}{2!} + \cdots + \frac{z^{2n-2}}{(2n-2)!}, \quad T_{-2}(z) = 0, \]
then for \( f(z) = \text{sh} \, z \) and \( f(z) = \text{ch} \, z \) we obtain inequalities
\[ |\text{sh} \, z - T_{2n-1}(z)| \leq \frac{(2n-2m+1)!}{(2n+1)!} |z|^{2m} (|\text{sh} \, z| - T_{2n-2m-1}(|z|)) \]
and
\[ |\text{ch} \, z - T_{2n-2}(z)| \leq \frac{(2n-2m)!}{(2n)!} |z|^{2m} (|\text{ch} \, z| - T_{2n-2m-2}(|z|)) \]
for \( m = 1, \ldots, n \), respectively.

In the same way on the basis of the inequality (2) we obtain inequalities
\[ |\text{sh} \, z - T_{2n-1}(z)| \leq \frac{(2n-2m+1)!}{(2n+1)!} |z|^{2m+1} (|\text{ch} \, z| - T_{2n-2m-2}(|z|)) \]
and
\[ |\text{ch} \, z - T_{2n-2}(z)| \leq \frac{(2n-2m)!}{(2n)!} |z|^{2m-1} (|\text{sh} \, z| - T_{2n-2m-1}(|z|)) \]
for \( m = 0, 1, \ldots, n \).

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
\[ |\text{sh} \, z - T_{2n-1}(z)| \leq \frac{(2n+2)(2n+3)}{(2n+1)!} \frac{|z|^{2n+1}}{(2n+2)(2n+3) - |z|^2} \]
holds.
Proof. As

\[ |\text{sh} z - T_{2n-1}(z)| \leq \frac{|z|^{2n+1}}{(2n+1)!} (1 + \frac{|z|^2}{(2n+2)(2n+3)} + \cdots ) \]

\[ \leq \frac{|z|^{2n+1}}{(2n+1)!} (1 + \frac{|z|^2}{(2n+2)(2n+3)} + \frac{|z|^4}{(2n+2)^2(2n+3)^2} + \cdots ) \]

\[ = \frac{|z|^{2n+1}}{(2n+1)!} \frac{(2n+2)(2n+3)}{(2n+2)(2n+3) - |z|^2} \]

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

\[ |\text{ch} z - T_{2n-2}(z)| \leq \frac{(2n+1)(2n+2)}{(2n)!} \frac{|z|^{2n}}{(2n+1)(2n+2) - |z|^2} \]

holds.

**REFERENCES**

ON A METHOD OF NUMERICAL DIFFERENTIATION

Bogdan M. Damnjanović

ABSTRACT:
A method for numerical differentiation of a function assigned tabularly is described in the paper. The orthogonal system
\[ \pi^{-1} \cos x, \pi^{-1} \sin x, \pi^{-1} \cos 2x, \pi^{-1} \sin 2x, \ldots \]
is used.

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 \( f_\beta \). 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

\[ \int_{-\pi}^{\pi} (f''(x))^2 \, dx < \infty, \]

\[ \|f - f_\beta\| = \left( \int_{-\pi}^{\pi} [f(x) - f_\beta(x)]^2 \, dx \right)^{1/2} \leq \beta, \]

where \( \beta > 0 \) is assigned.

On the basis of the known function \( f_\beta \), 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 $f_\beta$ is approximated by a polynomial $q_{n+1}(x)$ of degree $n+1$, such one that the inequality

\[ \|f_\beta - q_{n+1}\| \leq \beta \]

holds. The polynomial $q_{n+1}$ is obtained developing $f_\beta$ by Fourier' series using the orthogonal system

\[ \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \]

namely,

\[ q_{n+1}(x) = \frac{1}{\pi} \sum_{k=1}^{n+1} (a_k \cos kx + b_k \sin kx), \]

where

\[ a_k = \int_{-\pi}^{\pi} f_\beta(x) \cos kx \, dx, \quad b_k = \int_{-\pi}^{\pi} f_\beta(x) \sin kx \, dx, \]

and $n$ is chosen so that (3) is valid.

According to (2) and (3) it follows

\[ \|f - q_{n+1}\| \leq \|f - f_\beta\| + \|f_\beta - q_{n+1}\| \leq 2\beta. \]

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

\[ \int_{-\pi}^{\pi} u(t) \, dt = f(x). \]

Let (7) be written in the operator form

\[ Au = f. \]

Introducing the functionals

\[ I[u, q_{n+1}] = \|Au - q_{n+1}\|^2 \]

and

\[ K[u] = \|u'\|^2, \]

the approximate solution of the equation (7) (i.e. (8)) can be found by minimizing the functional $K[u]$ under the condition

\[ I[u, q_{n+1}] \leq 4\beta^2 \]

(according to (6)). Denote a such minimum by $u_\beta(x)$. Then, it is easy to prove that

\[ I[u_\beta, q_{n+1}] = 4\beta^2. \]
In the following, we will show that \( u_n(x) \) is the required minimum if the condition
\[ 13) \quad \alpha < \frac{1}{2} \| q_{n+1} \| \]
is satisfied.

Let
\[ A \mu = v \]
be Fourier's development of \( v \) using the orthogonal system \((\ast)\), 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
\[ u(x) \sim \frac{1}{\pi} \sum_{k=1}^{\infty} [c_k(-k \sin kx) + d_k k \cos kx], \]
and
\[ u''(x) \sim -\frac{1}{\pi} \sum_{k=1}^{\infty} k^2 (c_k \cos kx + d_k \sin kx), \]
where
\[ c_k = -k^{-2} \int_0^\pi u'(x) \cos kx \, dx, \]
\[ d_k = -k^{-2} \int_0^\pi u'(x) \sin kx \, dx. \]

Now, we have
\[ K[u] = \int_{-\pi}^{\pi} [u'(x)]^2 \, dx = \int_{-\pi}^{\pi} \left[ -\frac{1}{\pi} \sum_{k=1}^{\infty} k^2 (c_k \cos kx + d_k \sin kx) \right]^2 \, dx \]
\[ = \frac{1}{\pi} \sum_{k=1}^{\infty} k^4 (c_k^2 + d_k^2) \]

and
\[ I[u, q_{n+1}] = \int_{-\pi}^{\pi} [u(x) - q_{n+1}(x)]^2 \, dx = \int_{-\pi}^{\pi} [v(x) - q_{n+1}(x)]^2 \, dx \]
\[ = \int_{-\pi}^{\pi} \left[ \frac{1}{\pi} \sum_{k=1}^{\infty} (c_k \cos kx + d_k \sin kx) - \frac{1}{\pi} \sum_{k=1}^{n+1} (a_k \cos kx + b_k \sin kx) \right]^2 \, dx \]
\[ = \frac{1}{\pi} \sum_{k=1}^{n+1} (c_k^2 + d_k^2) + \sum_{k=n+2}^{\infty} (c_k^2 + d_k^2). \]

The coefficients \( c_k \) and \( d_k \) are defined by looking for the minimum of the functional \( K[u] \) under the condition \( I[u, q_{n+1}] = 4\beta^2 \) using the method of regulation:
\[ \phi[u, q_{n+1}] = K[u] + nI[u, q_{n+1}], \quad I[u, q_{n+1}] = 4\beta^2. \]

If we put \( \lambda = \frac{1}{n} > 0 \), then \( \phi[u, q_{n+1}] \) becomes

\[ \phi[u, q_{n+1}] = \frac{1}{n} \sum_{k=1}^{n+1} \left[ (c_k - a_k)^2 + (d_k - b_k)^2 \right] \]

\[ + \sum_{k=n+2}^{\infty} \left( c_k^2 + b_k^2 \right) + \frac{\lambda}{\pi} \sum_{k=1}^{\infty} k^4 (c_k^2 + b_k^2) \]

\[ = \frac{1}{n} \sum_{k=1}^{n+1} \left[ (c_k^2 - a_k^2) + (d_k - b_k)^2 + \frac{\lambda}{\pi} k^4 (c_k^2 + b_k^2) \right] \]

\[ + \sum_{k=n+2}^{\infty} \left( \frac{1}{\pi} + \frac{\lambda}{\pi} k^4 \right) (c_k^2 + d_k^2). \]

Since the second sum is nonnegative, for the minimum is suppose that \( c_k \to 0 \) and \( d_k \to 0 \) for \( k \geq n+2 \). Then, according to

\[ \frac{\partial \phi}{\partial c_k} = 0 \quad \text{and} \quad \frac{\partial \phi}{\partial d_k} = 0, \]

we find

\[ c_k = \frac{\pi a_k}{1 + \lambda k^4}, \quad d_k = \frac{\pi b_k}{1 + \lambda k^4}. \]

On the basis of (14) we obtain

\[ u_{\beta}(x) = \frac{1}{\pi} \sum_{k=1}^{n+1} \frac{a_k \cos kx + b_k \sin kx}{1 + \lambda k^4}, \]

where \( \lambda \) is determined from

\[ \| u_{\beta} - q_{n+1} \|^2 = 4\beta^2, \]

that is

\[ \frac{1}{\pi} \sum_{k=1}^{n+1} \lambda^2 k^8 (a_k^2 + b_k^2) (1 + \lambda k^4)^{-2} = 4\beta^2. \]

The left-hand side of (18) is monotonically increasing function of \( \lambda \) 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

\[ \frac{1}{\pi} \sum_{k=1}^{n+1} (a_k^2 + b_k^2) > 4\beta^2, \]
... from we immediately (13). Therefore, \( u_{n}(x) \) will be minimum of \( \psi[u, q_{n+1}] \) if (13) is valid.

It is easy to prove that \( u_{n}(x) \) is minimum of \( \psi[u, q_{n+1}] \) for each \( x \in (-\pi, \pi) \) and \( \lim_{\beta \to 0} u_{n}(x) = u_{0}(x) = f'(x) \) 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 \( \lambda \) are defined by (5) and (18).

REFERENCES

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.

1. INTRODUCTION
In the journal Amer. Math. Monthly the following two problems ([1],[2]) are posed:

1° Suppose \( f(x) \) has a continuous \((2m)\)-th derivative on \( a \leq x \leq b \), that \( \left| f^{(2m)}(x) \right| \leq M \), and that \( f^{(r)}(a) = f^{(r)}(b) = 0 \) for \( r = 0, 1, \ldots, m-1 \). Show that

\[
\int_a^b f(x) \, dx \leq \frac{(m!)^2 M}{(2m)! (2m+1)!} (b-a)^{2m+1}.
\]

2° Let \( f: [a, b] \to \mathbb{R} \) be a continuous function which is twice differentiable in \((a, b)\) and satisfies \( f(a) = f(b) = 0 \). Prove that

\[
\int_a^b |f(x)| \, dx \leq \frac{1}{12} M (b-a)^3,
\]
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 \([4, \text{ pp. 297-298}]\).\(^1\)

In this paper we shall prove some inequalities which generalize (1) and (2) in many senses.

Let us define the two-parameter class of polynomials \(P_n^{(m,k)}(0 \leq m, k < n; m, k, n \in \mathbb{N})\) by means of

\[
P_n^{(m,k)}(x) = P_n^{(m,k)}(x; a, b) = \sum_{i=0}^{k} \binom{k}{i} (x-a)^m (x-b)^{n-k-1} \int_{b}^{x} \frac{(-1)^{k-i}(n-m)!}{m!(k-m)!(n-k-1)!} \, dx
\]

where \(a\) and \(b\) are real parameters and

\[
C_n^{(m,k)}(a, b) = \frac{(-1)^{n-k}(n-m)!}{m!(k-m)!(n-k)!} (b-a)^{m-n}.
\]

For this polynomials the following relations hold:

\[
\frac{d^i}{dx^i} P_n^{(m,k)}(x) \bigg|_{x=a} = \delta_{im} (i=0,1,\ldots,k; \delta_{im} \text{ is the CRONECKER symbol}),
\]

\[
\frac{d^i}{dx^i} P_n^{(m,k)}(x) \bigg|_{x=b} = 0 (i=0,1,\ldots,n-k-1),
\]

\[
P_n^{(m,k)}(x) = C_n^{(m,k)}(a, b) \sum_{i=0}^{k-m} \frac{(b-a)^i}{n-m+i} \binom{k-m}{i} (x-a)^m (x-b)^{n-m-i} \int_{b}^{x} \frac{(-1)^{k-i}(n-m)!}{m!(k-m)!(n-k-1)!} \, dx
\]

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:

\[
S_n^{(m,k)}(x) = \sum_{m=0}^{k-1} P_n^{(m,k-1)}(x; a, b) f^{(m)}(a) + \sum_{m=0}^{n-k-1} P_n^{(m,n-k-1)}(x; b, a) f^{(m)}(b).
\]

\(^1\) On some generalizations IYENGAR's inequality see [5-7].
.. 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 \( x \mapsto f(x) \) be a \( n \)-times differentiable function such that \(|f^{(n)}(x)| \leq M\) \((\forall x \in (a,b))\). If \( x \mapsto p(x) \) is an integrable function on \((a,b)\) such that

\[ 0 < c \leq p(x) \leq \lambda c \ (\lambda \geq 1, x \in [a,b]), \]

the following inequality holds:

\[ M^r(g;p) \leq \frac{M_0 b-a^n}{n!} \frac{\lambda B(rk+1,r(n-k)+1)}{C^r(\lambda-1)B(rk+1,r(n-k)+1)} \left( \frac{\lambda N_{r\mu}}{N+(\lambda-1)\mu} \right)^{1/r}, \]

where \( B \) is beta function and \( C = k^{r(k-1)}n^{-k}n^k \).

**Proof.** Since \(|f^{(n)}(x)| \leq M\), the inequality

\[ |f(x) - S_{n,k}(x)| \leq \frac{M}{n!} |(x-a)_k(x-b)^{n-k}| \]

is valid, wherefrom (for \( r > 0 \))

\[ M^r(g;p) = \left( \frac{\int_a^b p(x)(x-a)^{rk}(b-x)^r(n-k) \, dx}{\int_a^b p(x) \, dx} \right)^{1/r}, \]

According to J. KARAMATA's inequality [8] (see also [5]) we have

\[ \frac{\int_a^b p(x)(x-a)^{rk}(b-x)^r(n-k) \, dx}{\int_a^b p(x) \, dx} \leq \frac{\lambda N_{r\mu}}{N+(\lambda-1)\mu}, \]

where

\[ N = C^r(b-a)^n \text{ and } \mu = (b-a)^{n-r}B(rk+1,r(n-k)+1), \]

which combined with (4) gives (3).

From Theorem 1, we directly get the following result:
COROLLARY 1. Let $x\rightarrow f(x)$ be a $n$-times differentiable function such that $|f^n(x)| \leq M (\forall x \in (a,b))$ 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

$$
\left(\frac{b-a}{b-a}\right) \frac{b}{|f(x)|^{1/r}} \leq \frac{M(b-a)^n}{n!} B(rk+1,r(n-k)+1)^{1/r} (r>0).
$$

For $n=2m$, $k=m$, $r=1$, inequality (5) reduces to

$$
\frac{b}{a} \int_{a}^{b} f(x) dx \leq \frac{M(b-a)^{2m+1}(m!)^2}{(2m)! (2m+1)!}
$$

which generalize (2), and which is evidently stronger than the inequality (1).

COROLLARY 2. Let function $x\rightarrow f(x)$ satisfy the conditions as in Corollary 1. If $x\rightarrow p(x)$ is arbitrary nonnegative function, then

$$
M[x] (f;p) \leq \frac{M^k(n-k)^{n-k}}{n!^n} (b-a)^n (r>0).
$$

REMARK 1. Corollary 2 can be formally obtained from Theorem 1 putting $\lambda \rightarrow +\infty$. Using N. J. ÖSTlund's result (9), the inequality (7) can be substituted by a somewhat simpler but weaker inequality

$$
M[x] (f;p) \leq \frac{M}{n!^k} (n-1) (b-a)^n (r>0).
$$

3. SOME SIMILAR RESULTS

According to the results from the previous section and the inequality $
\int_{a}^{b} h(x) dx \leq \int_{a}^{b} |h(x)| dx$, we obtain the following inequality

$$
\int_{a}^{b} f(x) dx - \sum_{k=1}^{m} \frac{(2m-k)!}{(2m)!} \frac{m}{K} (b-a)^k f^{(k-1)}(a) - (-1)^k f^{(k-1)}(b) \leq \frac{M(m!)^2 (b-a)^{2m+1}}{(2m)! (2m+1)!}.
$$

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 $[0,1]$. If $x \mapsto f(x)$ is $n$-times differentiable function such that $|f^{(n)}(x)| \leq M \ (\forall x \in (a,b))$, then

\[
\left| \int_a^b f(x) \, dx - \sum_{k=0}^{n-1} (-1)^k (b-a)^k (P_k(0) f^{(k-1)}(a) - P_k(1) f^{(k-1)}(b)) \right| \\
\leq M (b-a)^n + \frac{1}{n} \left| \int_0^n P_n(t) \, dt. \right|
\]

Proof. If $h(t) = f(a+t(b-a))$ we have $\int_a^b f(x) \, dx = \int_a^b h(t) \, dt$, wherefrom, applying integration by parts on the last integral, we obtain

\[
\int_0^1 h(t) \, dt = h(1) - \int_0^1 t h'(t) \, dt.
\]

Since $P_1(t) = P_0 t + P_1(0)$, 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 successive 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
\[
\left| \int_a^b f(x) \, dx \right| \leq \frac{M(b-a)^{n+1}}{2^n(n+1)!}.
\]

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**

CLASSIFICATION OF FORMULAS
FOR N-DIMENSIONAL POLYNOMIAL INTERPOLATION

Dušan V. Slavić, Milorad J. Stanojević

ABSTRACT:

The increased significance of interpolation in complicated computer calculation of the function values for either one or several variables imposes the need for evaluation of the possibilities and the efficiency of some interpolation formulas. The paper gives a classification of the formulas for polynomial interpolation in the n-dimensional space (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

\[ f(x) = \sum_{j=1}^{n} \left( \sum_{k=1}^{n} \frac{x - x_k}{x_j - x_k} \right) f(x_j), \]

usually referred to as Lagrange formula containing the following four known specific formulas, for example in [1].

\[ f(x) = a \]
\[ f(x) = f_0 \]
\[ f(x) = a + bx \]
\[ f(x) = (1 - x)f_0 + xf_1 \]
\[ f(x) = a + bx + cx^2 \]
\[ f(x) = \frac{1}{2}x(x - 1)f_{-1} + (1 - x^2)f_0 + \frac{1}{2}x(x + 1)f_1 \]
\[ f(x) = a + bx + cx^2 + dx^3 \]
\[ f(x) = -\frac{1}{6}x(x - 1)(x - 2)f_{-1} + \frac{1}{2}(x^2 - 1)(x - 2)f_0 + \frac{1}{6}x(x + 1)(x - 2)f_1 \]

TWO-DIMENSIONAL INTERPOLATION

In the case of the two-dimensional interpolation three-point formulas are presented first. These formulas are function approximation by means of the plane \( z = A + Bx + Cy \).

\[ f(x,y) = A + Bx + Cy \]
\[ f(x,y) = (1 - x - y)f_{0,0} + xf_{1,0} + yf_{0,1} \]
\[ f(x,y) = (1 - x - 2y)f_{0,0} + (x + y)f_{1,0} + yf_{-1,1} \]
\[ f(x,y) = \frac{1}{2}(1 - x - y)f_{-1,0} + yf_{0,1} + \frac{1}{2}(1 + x - y)f_{1,0} \]
The general three-point formula has been considered, for example, Young and Gregory and it reads:

\[ f(x_1, y_1) = f_1 \]
\[ f(x, y) = (a_1 f_1 + a_2 f_2 + a_3 f_3)/d \]
\[ a_1 = (x - x_2)(y - y_3) - (x - x_3)(y - y_2) \]
\[ a_2 = (x - x_3)(y - y_1) - (x - x_1)(y - y_3) \]
\[ a_3 = (x - x_1)(y - y_2) - (x - x_2)(y - y_1) \]
\[ d = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1). \]

The formula (2) is equivalent to the following formula given by erezin - Žitkov (r and R are vectors)

\[ r_k = (x - x_k)i + (y - y_k)j \]
\[ r_{k\ell} = (x_{k'} - x_{\ell'})i + (y_{k'} - y_{\ell'})j \]
\[ R_{k\ell} = (y_{k'} - y_{\ell'})i - (x_{k'} - x_{\ell'})j \]
\[ f(x, y) = \frac{(r_2 R_{23})}{(r_{12} R_{23})} f(x_1, y_1) + \frac{(r_3 R_{31})}{(r_{23} R_{31})} f(x_2, y_2) + \frac{(r_1 R_{12})}{(r_{31} R_{12})} f(x_3, y_3). \]

This formula may be written in a shorter way, as follows

\[ f(x, y) = \sum_3 \frac{(r_2 R_{23})}{(r_{12} R_{23})} f(x_1, y_1), \]

where 3 represents the number of cyclic permutation.

Four-point formulas which include the term with \( xy \) are presented here

\[ f(x, y) = A + Bx + Cy + Exy \]
\[ f(x, y) = (1 - x)(1 - y)f_{0,0} + x(1 - y)f_{1,0} + (1 - x)yf_{0,1} + xyf_{1,1} \]
\[ f(x, y) = \frac{1}{2} x(x - 1)f_{-1,0} + (1 - x^2 - y)f_{0,0} + \frac{1}{2} x(x + 1)f_{1,0} + xyf_{0,1} \]

\[ f(x, y) = \frac{1}{2}x(x - 1)f_{-1,0} + (1 - x^2)f_{0,0} + \left(\frac{1}{2}x(x + 1) - y\right)f_{1,0} + yf_{1,1}. \]

The following five-point formula instead of the term with \( xy \) contains the terms with \( x^2 \) and \( y^2 \)
\[ f(x, y) = A + Bx + Cy + Dx^2 + Ey^2 \]
\[ f(x, y) = (1 - x^2 - y^2)f_{0,0} + \frac{1}{2}x(x + 1)f_{1,0} \]
\[ + \frac{1}{2}y(y + 1)f_{0,1} + \frac{1}{2}x(x - 1)f_{-1,0} + \frac{1}{2}y(y - 1)f_{0,-1}. \]

The following six-point formulas contain all the terms mentioned earlier
\[ f(x, y) = A + Bx + Cy + Dx^2 + Exy + Fy^2 \]
\[ f(x, y) = \frac{1}{2}(1 - x - y)(2 - x - y)f_{0,0} + x(2 - x - y)f_{1,0} \]
\[ + y(2 - x - y)f_{0,1} + \frac{1}{2}x(x - 1)f_{2,0} \]
\[ + xyf_{1,1} + \frac{1}{2}y(y - 1)f_{0,2}. \]
\[ f(x, y) = (1 - x)(1 - y)f_{0,0} + (x - \frac{1}{2}y)(1 - y)f_{1,0} \]
\[ + (y - \frac{1}{2}x)(1 - x)f_{0,1} + \frac{1}{2}x(x - 1)f_{2,1} \]
\[ + (x(1-x) + y(1-y) + xy)f_{1,1} + \frac{1}{2}y(y - 1)f_{1,2}. \]

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
\[ f(x, y) = \sum_{6} \left| \begin{array}{ccc} \left( r_{23} \right) & \left( r_{45} \right) & \left( r_{62} \right) \left( r_{23} \right) \left( r_{64} \right) \left( r_{45} \right) \\ \left( r_{24} \right) & \left( r_{35} \right) & \left( r_{62} \right) \left( r_{24} \right) \left( r_{63} \right) \left( r_{35} \right) \\ \left( r_{12} \right) \left( r_{23} \right) \left( r_{14} \right) & \left( r_{62} \right) \left( r_{23} \right) \left( r_{64} \right) \left( r_{45} \right) \\ \left( r_{12} \right) \left( r_{24} \right) & \left( r_{13} \right) & \left( r_{62} \right) \left( r_{24} \right) \left( r_{63} \right) \left( r_{35} \right) \end{array} \right| \cdot f(x_1, y_1). \]
The following nine-point formula has the same order of accuracy.

\[
f(x,y) = (1 - x^2)(1 - y^2)f_{0,0} + \frac{1}{2}x(1 + x - xy)f_{1,0} \quad \bullet \quad \bullet
\]
\[
+ \frac{1}{2}y(1 + y - x^2)f_{0,1} - \frac{1}{2}x(1 - x + xy)f_{-1,0} \quad \bullet \quad \bullet
\]
\[
- \frac{1}{2}y(1 - y + x^2)f_{0,-1} + \frac{1}{4}xy(1 + xy)f_{1,1} \quad \bullet \quad \bullet
\]
\[
- \frac{1}{4}xy(1 - xy)f_{1,-1} + \frac{1}{4}xy(1 - xy)f_{-1,-1} + \frac{1}{4}xy(1 + xy)f_{-1,1}.
\]

J. H. Lambert [1] gave the formula which can be concisely written as

\[
f(x,y) = \sum_{m=0}^{n} \sum_{k=0}^{m} \left( x\text{ }_{m-k} y\text{ }_{k} \right) \Delta_{m-k,k} f_{0,0},
\]

where:

\[
\Delta_{0,0}^{f} p,q = f_{p,q}
\]
\[
\Delta_{1,0}^{f} p,q = f_{p+1,q} - f_{p,q}
\]
\[
\Delta_{0,1}^{f} p,q = f_{p,q+1} - f_{p,q}
\]
\[
\Delta_{n+1,m}^{f} p,q = \Delta_{n,m}^{f} 1,0 p,q
\]
\[
\Delta_{n,m+1}^{f} p,q = \Delta_{n,m}^{f} 0,1 p,q
\]

A specific case of the ten-point formula (4) is

\[
f(x,y) = A + Bx + Cy + Dx^2 + Exy + Fy^2 + Gx^3 + Hx^2y + Ixy^2 + Jy^3
\]

\[
f(x,y) = \frac{1}{6}(1 - x - y)(2 - x - y)(3 - x - y)f_{0,0} \quad \bullet \quad \bullet
\]
\[
+ \frac{1}{2}x(2 - x - y)(3 - x - y)f_{1,0} \quad \bullet \quad \bullet
\]
\[
+ \frac{1}{2}y(2 - x - y)(3 - x - y)f_{0,1} \quad \bullet \quad \bullet
\]
\[
+ \frac{1}{2}y(y - 1)(3 - x - y)f_{2,0} \quad \bullet \quad \bullet
\]
\[
+ xy(3 - x - y)f_{1,1} + \frac{1}{2}x(x - 1)(3 - x - y)f_{0,2} \quad \bullet \quad \bullet
\]
\[
+ \frac{1}{6}x(x - 1)(x - 2)f_{3,0} + \frac{1}{2}xy(x - 1)f_{2,1} \quad \bullet \quad \bullet
\]
\[
+ \frac{1}{2}xy(y - 1)f_{1,2} + \frac{1}{6}y(y - 1)(y - 2)f_{0,3} \quad \bullet \quad \bullet
\]

\[
F. B. Hildebrand gives the following formula

\[
f(x,y) = A + Bx + Cy + Dx^2 + Exy + Fy^2 + Gx^3 + Hx^2y + Ixy^2 + Jy^3 + Lx^3 + Nxy^3
\]

\[
f(x,y) = \frac{1}{2}(1-x)(1-y)[2+x(1-x)+y(1-y)] f_{0,0} + \frac{1}{2}x(x-y)[2+x(1-x)+y(1-y)] f_{1,0} + \frac{1}{2}xy[2+x(1-x)+y(1-y)] f_{1,1} + \frac{1}{2}(1-x)y[2+x(1-x)+y(1-y)] f_{0,1} + \frac{1}{6}x(x-1)(x-2)(y-1)f_{-1,0} + \frac{1}{6}(x-1)y(y-1)(y-2)f_{0,-1} + \frac{1}{6}xy(y-1)(2-y)f_{1,-1} + \frac{1}{6}x(1-x^2)(y-1)f_{2,0} + \frac{1}{6}(x^2-1)yg_{2,1} + \frac{1}{6}xy(y^2-1)f_{1,2} + \frac{1}{6}(x-1)y(1-y^2)g_{0,2} + \frac{1}{6}x(x-1)(2-x)yf_{-1,1}.
\]

THREE-DIMENSIONAL INTERPOLATION

For interpolation in the three-dimensional space the following formulas are given

\[
f(x,y,z) = a
\]

\[
f(x,y,z) = f_{0,0,0}
\]

\[
f(x,y,z) = \alpha + \beta x + \gamma y + \delta z
\]

\[
f(x,y,z) = (1-x-y-z)f_{0,0,0} + xf_{1,0,0} + yf_{0,1,0} + zf_{0,0,1}
\]
\[ f(x, y, z) = a + bx + cy + dz + ex^2 + fxy + gyz + hz^2 + kxz \]

\[ f(x, y, z) = \frac{1}{2}(1 - x - y - z)(2 - x - y - z)f_{0,0,0} \]
\[ + x(2 - x - y - z)f_{1,0,0} \]
\[ + y(2 - x - y - z)f_{0,1,0} \]
\[ + z(2 - x - y - z)f_{0,0,1} \]
\[ + \frac{1}{2}x(x - 1)f_{2,0,0} + xyf_{1,1,0} \]
\[ + \frac{1}{2}y(y - 1)f_{0,2,0} + yzf_{0,1,1} \]
\[ + \frac{1}{2}z(z - 1)f_{0,0,2} + xzf_{1,0,1} \]

**N-DIMENSIONAL INTERPOLATION**

Suppose that the four-dimensional-space points are provided

\[ f(a_i, b_j, c_k, d_{\ell}) \]

where:

\[ i = 1(1)n_a, \quad j = 1(1)n_b, \quad k = 1(1)n_c, \quad \ell = 1(1)n_d \]

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

\[ f(a_i, b_j, c_k, d_{\ell}) = \sum_{\ell=1}^{n_d} \left( \prod_{m=1, m \neq \ell}^{n_d} \frac{d_{\ell} - d_m}{d_{\ell} - d_m} \right) f(a_i, b_j, c_k, d_{\ell}) \]

\[ i = 1(1)n_a, \quad j = 1(1)n_b, \quad k = 1(1)n_c \]

By intersecting the planes the following formula is obtained

\[ f(a_i, b_j, c_k, d) = \sum_{k=1}^{n_c} \left( \prod_{m=1, m \neq k}^{n_c} \frac{c_{k} - c_m}{c_{k} - c_m} \right) f(a_i, b_j, c_k, d) \]

\[ i = 1(1)n_a, \quad j = 1(1)n_b \]
By intersecting straight lines the following formula is obtained

\[ f(a_i, b, c, d) = \sum_{j=1}^{n_b} \left( \prod_{m=1}^{n_b} \frac{b - b_m}{b_j - b_m} \right) f(a_i, b_j, c, d), \quad i=1(1)n_a. \]

By applying the mentioned Lagrange formula (1) it can be obtained

\[ f(a, b, c, d) = \sum_{i=1}^{n_a} \left( \prod_{m=1}^{n_a} \frac{a - a_m}{a_i - a_m} \right) f(a_i, b, c, d). \]

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. Alikalić, 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

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.

Let us consider the equation

\[ y'' + p(t)y' + q(t)y = f(t), \]

where \( p, q, f \in C(I), I = (-\infty, \infty). \) Let \( y = \Psi(t), \Psi \in C(I) \) be an arbitrary solution of equation (1). We shall use functions \( \beta, \rho \in C(I), \rho(t) > 0 \) on \( I \) and notations \( p_0 = p(t_0), \beta_0 = \beta(t_0), \rho_0 = \rho(t_0), \Psi_0 = \Psi(t_0), y_0 = y(t_0), y'_0 = y'(t_0). \)

THEOREM 1. Let us take functions \( \beta \) and \( \rho \) such that

\[ (\beta' + \beta_0^2 + \beta_0 \rho' + q - 1)^2 < 4(-\beta - p - \rho' / \rho)(\beta - \rho' / \rho) \text{ on } I. \]

(a) If

\[ \beta - \rho' / \rho < 0 \text{ on } I, \]

then all solutions \( y = y(t) \) of equation (1) satisfying initial condition

\[ (y_0 - \Psi_0)^2 + (y'_0 - \beta_0 y_0)^2 \leq \rho_0^2, \quad t_0 \in I \]
satisfy also condition

\[ |y(t) - \Psi(t)| < \rho(t) \quad \text{for every } t > t_0. \]

(b) If

\[ \beta - \rho'/\rho > 0 \quad \text{on } I, \]

then problem (1)-(4) has at least one solution satisfying condition (5).

Proof. For equation (1) let us introduce the substitute

\[ y = x + \beta(t)y, \]

where \( x = x(t) \) is a new unknown function. Equation (1) is transformed to the system of equations

\[ \begin{align*}
x' &= -\beta - p x - (\beta^2 + \beta p + q) y + f \\
y' &= x + \beta y.
\end{align*} \]

Let \( K = \{ (x, y, t) : x = \Psi(t), y = \Psi(t), t \in I \} \), where \( \Psi \in C^1(I), \Psi(t_0) = 0 \) is an integral curve of system (8). Let \( \varOmega = \mathbb{R}^2 \times I \) and

\[ \omega = \left\{ (x, y, t) \in \varOmega : \rho(t) \left[ (x - \Psi(t))^2 + (y - \Psi(t))^2 \right] < 1 \right\} \]

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 \( \partial \omega \) \( (\partial \omega = \partial t - \omega) \), and let \( \nu(t) \) be vector of external normal on surface \( \partial \omega \), i.e.

\[ \begin{align*}
\tau(t) &= (x(t), y(t), 1), \\
\nu(t) &= ((x - \Psi)^2 \rho^2, (y - \Psi)^2 \rho^2, \\
&\quad - \left[ (x - \Psi)^2 \rho^2 + (x - \Psi) \rho' \psi + (y - \Psi)^2 \rho^4 + (y - \Psi) \rho' \psi' \right] \rho^3).
\end{align*} \]

Let us consider now the scalar product \( P(t) = \langle \tau, \nu \rangle \) in the points of surface \( \partial \omega \). We have

\[ \begin{align*}
P(t) &= (-\beta - p - \rho'/\rho)(x - \Psi)^2 \rho^2 + (\beta - \rho'/\rho)(y - \Psi)^2 \rho^2 + \\
&\quad + (1 - \beta^2 - \beta p - q)(x - \Psi) \rho^4 (y - \Psi) \rho^4.
\end{align*} \]

Let us note that \( P(t) \) is a quadratic symmetric form

\[ P(t) = a_{11} x^2 + 2a_{12} x y + a_{22} y^2, \]

where

\[ \begin{align*}
a_{11} &= -\beta - p - \rho'/\rho, \\
a_{12} &= a_{21} = (1 - \beta^2 - \beta p - q)/2, \\
a_{22} &= \beta - \rho'/\rho,
\end{align*} \]

\[ \begin{align*}
x = (x - \Psi) \rho^4, \\
y = (y - \Psi) \rho^4.
\end{align*} \]

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 $\omega$ and $\Omega$. Consequently, all solutions of system (8) satisfying initial condition

$$x_0^2 + (y_0 - \Psi_0)^2 \leq \rho_0^2, \quad t_0 \in I$$

$$x_0 = x(t_0)$$

satisfy also condition

$$(x(t) - \varphi(t))^2 + (y(t) - \Psi(t))^2 < \rho^2(t) \quad \text{for every } t > t_0.$$  

Since, in view of (7), $x_0 = y_0' - \beta_0 y_0$ 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 $\exists \omega$ is a set of points of strict exit of integral curves of system (8) with respect to sets $\omega$ and $\Omega$. Hence, according to retraction method (14), there exists at least one integral curve of system (8) which belongs to set $\omega$ 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 $\beta$ and $\rho$ 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

$$(y_0 - \Psi_0)^2 + (y_0' + y_0)^2 \leq r^2,$$

where $r$ is a positive constant, satisfy also condition

(11) $$|y(t) - \Psi(t)| < r \quad \text{for every } t > t_0$$

(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

$$(y_0 - \Psi_0)^2 + (y_0' - y_0)^2 \leq r^2$$

satisfies also condition (11).

**COROLLARY 2** ($\beta(t) \equiv 0$). Let

$$(q - 1)^2 < 4(p + p'/p)p'/p \quad \text{on } I.$$  

(a) If $p' > 0$ on $I$, then all solutions of equation (1) which satisfy initial condition

(12) $$(y_0 - \Psi_0)^2 + y_0'^2 < \rho_0^2 \quad \text{at } t_0 \in I$$

satisfy also condition (5).

(b) If $p' < 0$ on $I$, then problem (1)-(12) has at least one solution satisfying condition (5).

**COROLLARY 3** ($f_3(t) = -p(t)$). Let

$$(q-p'-1)^2 < 4(p+p'/p)p'/p$$ on $I$.

(a) If $p' > 0$ on $I$, then all solutions of equation (1) satisfying initial condition

$$(\psi - \psi_0)^2 + (\psi' + p_0 \psi_0)^2 \leq p_0^2$$

satisfy also condition (5).

(b) If $p' > 0$ on $I$, then problem (1)-(13) has at least one solution satisfying condition (5).

**THEOREM 2.** (a) If there exist functions $\beta$ and $\rho$ such that

$$(14) \quad 2\beta + p \leq 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2(\beta + \rho'/\rho)$$
or

$$(15) \quad 2\beta + p > 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2(-\beta + \rho'/\rho)$$
on $I$, then statement (a) of Theorem 1 holds true.

(b) If there exist functions $\beta$ and $\rho$ such that

$$(2\beta + p > 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2(-\beta - \rho'/\rho)$$
or

$$2\beta + p \leq 0, \quad |\beta' + \beta^2 + \beta p + q - 1| < 2(\beta - \rho'/\rho)$$
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 $ab \leq (a^2 + b^2)/2$ for every $a, b \in \mathbb{R}$, on $\Omega \omega$ it is valid

$$P(t) \leq (-\beta - p - \rho'/\rho) X^2 + (\beta - \rho'/\rho) Y^2 + |1 - \beta - \beta^2 - \beta p - q| - (X^2 + Y^2)/2 \leq \bar{P}(t).$$

In view of (14) on $\Omega \omega$ 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 $\Omega \omega$ it is valid

$$\bar{P}(t) = (\beta - \rho'/\rho + |1 - \beta - \beta^2 - \beta p - q|/2) - (2\beta + p) X^2 < 0.$$
Here it should be noted that on \( \partial \omega \)
\[
P(t) > (\beta - p - p'/p)X^2 + (\beta - p'/p)Y^2 - \left(1 - \beta^2 - \beta p - p_q \right)(X^2 + Y^2)/2 \equiv P(t)
\]
and that \( P(t) > 0 \).

Conditions of Theorem 2 are simplified if functions \( \beta \) and \( p \) are taken in a special form. For example \( \beta(t) = 0, -1, 1 \); \( p(t) = p_0 e^{s(t-t_0)} \), \( s \in \mathbb{R} \).

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 \( p \) 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 \( p_0(t) \). If \( p(t) \to 0 \), \( t \to \infty \), we have asymptotic stability of all solutions with the function of stability \( p \).

2. If \( p(t) \to \infty \), \( t \to \infty \), 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 \( p \) are very significant. Approximation is particularly good when \( p(t) \to 0 \), \( t \to \infty \). 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 \( p(t) \) we take function \( c_p(t) \), where \( c > 0 \) is an arbitrary constant.

In the case of a homogeneous equation \( (f(t)) = 0 \) it is interesting to consider the behaviour of solutions in the neighbourhood of a trivial solution \( y(t) = 0 \) (case \( \Psi(t) = 0 \)).

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 \( \mathbb{R} \times I \), 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
\[
t^2 y'' + ty' + (t^2 - \lambda^2) y = 0, \ \lambda \in \mathbb{R}
\]
which satisfy the initial condition
\[
(y_0 - \Psi_0)^2 + (y_0'/\Psi_0 + 2t_0')^2 \leq \rho_0^2, \ \ t_0 \in I,
\]
where \( t > \sqrt{\lambda - 1/4} / (1-2s) \), \( s \in \mathbb{R} \), \( 0 < s < 1/2 \), then also satisfy condition

\[
|y(t) - \psi(t)| < \rho_o (t_o / t)^s \quad \text{for every } t > t_o.
\]

For example, for \( \lambda = 1/2 \), \( \psi(t) \equiv 0 \) all solutions of the Bessel's equation satisfying initial conditions \( y(0) = 0 \), \( y'(0) \leq \rho_o \), \( t_o > 0 \) satisfy also condition

\[
|y(t)| < \rho_o (t_o / t)^s \quad \text{for every } t > t_o.
\]

Let us note that for \( \lambda = 1/2 \) Bessel's equation has a general solution

\[ y(t) = (c_1 \cos t + c_2 \sin t) / \sqrt{t} \]

The proof of this result follows from Theorem 1 when \( \beta(t) \equiv -1/2 t \), \( \rho(t) \equiv \rho_o (t_o / t)^s \).

Example 2. Let us take equation

\[
y'' + p(t) (y' + y) + q(t) y = f(t),
\]

where \( p(t) \geq 2 \), \( |q(t)| < 2 (1-s) \), \( s \in \mathbb{R} \), \( 0 \leq s < 1 \), on \( I \), \( t > 0 \). All solutions of this equation which satisfy the initial condition

\[
(y_o - \psi_o)^2 + (y'_o + \psi'_o)^2 \leq \rho_o^2 \quad \text{at } t_o \in I
\]

also satisfy condition

\[
|y(t) - \psi(t)| < \rho_o e^{-s(t-t_o)} \quad \text{for every } t > t_o.
\]

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.


Faculty of Civil Engineering, University of Split,
V.Maslešće bb, P.O. Box 389, 58000 Split, Yugoslavia
ON HYPOTHESIS TESTING IN SPLINE REGRESSION

K. Surla, E. Nikolić, Z. Lozanov

ABSTRACT:
Algorithm given in [8] for determining 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 connected with testing continuity of the third derivative are proved.

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 spline 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 sense 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 observations, 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 apply the results obtained in [2] to B-spline regression, and to modify 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,...,m, x_q<x_{q+1}, consider the model

\[ y = X\beta + \epsilon \]

where

\[ X = \begin{bmatrix} B_{-3}(x_1) & \ldots & B_g(x_1) \\ \vdots & \ddots & \vdots \\ B_{-3}(x_m) & \ldots & B_g(x_m) \end{bmatrix}, \quad C = \begin{bmatrix} C_{-3} \\ \vdots \\ C_g \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{bmatrix} \]
$e \sim N(0, \sigma^2 I)$, $B_i(x)$ are B-splines functions of the third degree on the grid $a = \lambda_0 < \lambda_1 < \cdots < \lambda_{g+1} = b$, with additional knots $\lambda_{-3} = \lambda_{-2} = \cdots = \lambda_{-1} = a$, $b = \lambda_{g+2} = \lambda_{g+3} = \lambda_{g+4}$. 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

$$x_{q_i} - h \lambda_1 \preceq x_{q_i}$$

(condition Schoenberg and Whitney).

The approximation criterion is as in [2]

$$\text{Minimize} \sum_{i=1}^g \left( \sum_{j=-3}^{g-1} c_i a_{i,j} \right)^2,$$

subject to the constraint

$$\sum_{i=1}^g (y_q - \sum_{j=-3}^{g} c_i B_i(x_q))^2 < S.$$  

where $S$ is given, nonnegative constant (smoothing factor).

We remark that our spline function $S(x) = \sum_{i=-3}^{g-1} c_i B_i(x)$ becomes a single polynomial on $[x, b]$ if $\sum_{i=-3}^{g-1} a_{i,j} = 0$, $(j=q+1, \ldots, r-1)$, $a_{q,r} x \leq b$.

In [2] it is shown that problem (3) has a solution and that the algorithm given there leads to the number of knots which

$$F(p) = \sum_{i=1}^g \left( y_q - \sum_{j=-3}^{g} c_i B_i(x_q) \right)^2 \leq S,$$

where $p$ 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 $\lambda_j$ we test the following hypothesis (see [8]):

$$H_0 : \sum_{i=-3}^{g} c_i B_i(\lambda_j+0) - B_i(\lambda_j-0) = 0$$

We use the statistics

$$F = \frac{(1-C)^2 (1-XX')^{-1} \hat{\Sigma}}{y(1-XX')^{-1}X'y} (m-g-4) \quad \text{or} \quad F = \frac{(1-C)^2}{\hat{\Sigma}(1-C)}$$

where $\hat{\Sigma}$ is estimated variance of linear combination $1'c$, $1=(a_{-3,j} \cdots, a_{r,j})$. Under null hypothesis $F$ 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 $\lambda_j$ is statistically not significant. Statistically not significant knots we shall denote by $\bar{x}_j$, $(j=1, \ldots, g)$.

4. Denote the spline function with knots $\lambda_j (j=0, \ldots, g+1)$ by $S_0(x)$.

The idea of our algorithm 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
\[
F_g_j = \frac{m}{q} \sum_{q=1}^{m} (y_q - S_g_j(x_q))^2 \leq \delta
\]
isfied. The additional number of knots \( \Delta g_j \) and their positions in each
eration is determined according to algorithm I (see [2]):
\[
\Delta g_j = \left\{ \begin{array}{cl}
1 & j=0 \\
\min\{\Delta 1, \Delta 2, \max\{1, \Delta 3, \Delta 4\}\} & j=1,2,\ldots 
\end{array} \right. 
\]
with
\[
=2 \Delta g_{j-1}, \Delta 2=2m-4-g_j, \Delta 3=[\Delta g_{j-1}/2], \Delta 4=[(F_g_{j-1}-S)g_j/(F_{g_{j-1}}-F_g_j)]
\]
additional knots are then located inside the intervals \([\lambda_{j}, \lambda_{j+1}]\) with
rgest partial sum of squares of residuals. For details see [2].

Let knots \( \lambda_j \) (\( j=0,\ldots,g_{j+1} \)) be determined. As the next iteration we:

1. Determine spline \( S_g_j(x) \).

2. Test all knots \( \lambda_j \) (\( i=1,\ldots,g_j \)) using the given statistical test in 3.

3. If \( F_g_j \leq S \), \( S_g_j(x) \) is the solution to our problem. If not we go to (iv).

4. The additional knots are determined according to the algorithm I.

5. New set of knots \( \lambda_j \) (\( i=0,\ldots,g_{j+1}+1 \)) is formed taking statistically
significant knots from (iii) and additional knots from (iv).

6. Put \( g_j = g_{j+1} \) and go to (i).

From \( g_j < g_{j+1} \) and \( \bar{g}_j > g_j \) it follows \( F_g_j > F_g_{j+1} \) and \( F_g_j > F_g_j \). Suppo-
sing that the sum of squares of residuals will not be significantly changed by substituteing knots \( \lambda_j \) (\( i=1,\ldots,g_j \)) by \( \bar{\lambda}_j \) (\( i=1,\ldots,\bar{g}_j \)) we can con-
clude that the relation (5) will be satisfied after finite number of iter-
ations. Namely, for maximal number of knots \( g=m-4 \) we get a interpolating spline, i.e. \( F_g_{m-4}=0 \). As the values \( F_g_j \) and \( F_g_{j+1} \) are evaluated the rela-
tion \( F_{g_{j+1}} > F_g_j \) can be checked. If significant deviation occurs, it is
possible to take \( g_j \) instead \( \bar{g}_j \). This algorithm, compared with algorithm in
[2] changes the position of knots, and we get a curve with statistically
significant knots.

5. Before proving several statements, we shall introduce some defi-
nitions and notations. Denote by \( P_i(x)=a_i x^3+b_i x^2+c_i x+d_i \) the restriction of
function \( S(x) \) on interval \( [\lambda_{i-1}, \lambda_i] \) and put \( P=(P_i(x), i=0,\ldots,g_i) \). The
relations between coefficients of \( P_i(x) \) and coefficients of \( S(x) \) are
given in [9].

**DEFINITION 1.** Polynomials \( P_i(x) \) and \( P_{i-1}(x) \) are not statistically different
at a prescribed level of significance \( \alpha \), if their corresponding coefficients
are not statistically different at the level \( \alpha \).

**DEFINITION 2.** Function \( \bar{S}(x) \) is statistical equivalent of spline \( S(x) \) at
the prescribed level of significance \( \alpha \) if
(i) \( \overline{S}(x) = Q_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i, \ x \in [\mu_i, \mu_{i+1}], \ M \supseteq \Lambda, \) where

\[ M = \{ \mu_i, i=0, \ldots, r+1 \}, \ \Lambda = \{ \lambda_i, i=0, \ldots, g+1 \}, \) \quad and \quad Q = \{ Q_i(x), i=0, 1, \ldots, r \}.

(ii) for fixed values \( x, S(x) \) and \( \overline{S}(x) \) are evaluated using the polynomials which are not statistically different at level \( \alpha \).

(iii) Discontinuities of function \( \overline{S}(x) \), its first and second derivatives at points \( \mu_i \) are not statistically significant at the level \( \alpha \).

Lemma 1. If hypothesis \( H_0 \) is accepted at the level \( \alpha \), then polynomials \( P_{j-1}(x) \) and \( P_{j}(x) \) are not statistically different.

Proof. Hypothesis \( H_0 \) is equivalent to the hypothesis

\[ H_0^* : 6(a_j - a_{j-1}) = 0 \quad \text{and} \quad H_0^* : A(a_j - a_{j-1}) = 0 \]

where \( A \) is constant. Spline \( S(x) \) and its first and second derivative is continuous at point \( \lambda_j \), 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^2 (a_j - a_{j-1}) = d_j - d_{j-1}.

It follows that hypothesis of equality of corresponding coefficients of polynomials \( P_{j-1}(x) \) and \( P_{j}(x) \) are accepted at level \( \alpha \).

Theorem 1. Let there exists at most one statistically not significant knot \( \lambda_j \) between two statistically significant knots of spline \( S(x) \).

Function \( \overline{S}(x) \) where \( \overline{S}(x) = P_{j-1}(x) \) for \( x \in [\lambda_{j-1}, \lambda_{j+1}] \) and \( S(x) = \overline{S}(x) \) otherwise is statistical equivalent of spline \( S(x) \) at level \( \alpha \).

Proof. We know that

\[ P_{j-1}(\lambda_{j+1}) - P_{j+1}(\lambda_{j+1}) = (\lambda_{j+1} - \lambda_j) \lambda_j (\lambda_{j+1} - \lambda_j) (a_j - a_{j-1}), \]

\[ P_{j-1}(\lambda_{j+1}) - P_{j+1}(\lambda_{j+1}) = (\lambda_{j+1} - \lambda_j) \lambda_j (\lambda_{j+1} - \lambda_j) (a_j - a_{j-1}), \]

\[ P_{j-1}(\lambda_{j+1}) - P_{j+1}(\lambda_{j+1}) = (6\lambda_{j+1} - 6\lambda_j) (a_j - a_{j-1}). \]

So statement follows from Lemma 1.

Theorem 2. Let \( \Lambda_j = \{ \lambda_i, i=0, \ldots, g_j \} \) 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:

\[ F = \frac{F_{g_2} - F_{g_1}}{F_{g_1}} \frac{m-g_{1}-4}{g_{1}-g_{2}} \]

\[ \frac{F_{g_{2}}}{g_{2}} < \frac{F_{g_{1}}}{g_{1}} \]

\[ \frac{F_{g_{2}}}{g_{2}} > \frac{F_{g_{1}}}{g_{1}} \]

\[ \frac{F_{g_{2}}}{g_{2}} \]

From \( \frac{F_{g_{2}}}{g_{2}} \) it follows that \( \frac{F_{g_{2}}}{g_{2}} \) are independent 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 function with smaller number of knots, when positions of knots is not changed. For prescribed
level of significance $\alpha$ value $F_{\alpha}$ can be found such that

$$P(0 < F < F_{\alpha}) = 1 - \alpha.$$ 

then

$$0 < F - F_0 < F_{\alpha} \frac{g_1 - g_2}{g_2} \alpha \frac{g_1 - g_2}{m - g_1 - 4}$$ 

with probability $\alpha$.

REFERENCES


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 limit (in $d_2$ 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].

1. ALGEBRA

In this paper, the following denotations will be used: $N = \{1, 2, 3, \ldots \}$, $N_0 = N \cup \{0\}$, $x = (x_0, x_1, \ldots) = (x_k)_{k \in N_0}$, $S$ - the set of all sequences $x$. Further, the sequences $e_n \in S$ $(n \in N_0)$ is defined by

\[ e_n = (\delta_{nk})(k \in N_0), \]

where $\delta_{nk}$ 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 $k \in N_0$, and $x + y = (x_k + y_k)(k \in N_0)$. If $\lambda \in \mathbb{R}$ then $\lambda x$ means $(\lambda x_k)_{k \in N_0}$. Thus, $S$ together with defined operations consist a linear space over the field $\mathbb{R}$, with the sequences $e_n(n \in N_0)$ as a base.

The sequence $x = (x_k)(k \in N_0)$ is convex if $2x_{n+1} - 2x_n + x_{n-1} \geq 0 (n \in N_0)$. The set of all convex sequences will be denoted by $K$. It is known, that $K$ is a cone in $S$.

Further, let $D \subseteq \mathbb{R}$ be a nonempty set. With $F(D)$ we will denote the set of all functions $f: D \to \mathbb{R}$. For the operator $A: S \to F(D)$ we say that it is linear if for every $x, y \in S$ and
the equality $A(\lambda x + \mu y) = \lambda Ax + \mu Ay$, holds.

If for $x \in S$ and $f \in F(D)$ we have $f_x = Ax$, then we write $Ax \geq 0$ if $f_x(t) \geq 0$, for every $t \in D$. Similarly, $Ax = 0$ if $f_x(t) = 0$ for every $t \in D$.

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

$$(2) \quad d_S(x, y) = \sum_{k=0}^{+\infty} 2^{-k} \frac{|x_k - y_k|}{1 + |x_k - y_k|}.$$ 

Now, $(S, d_S)$ is a metric space with finite metric: $d_S(x, y) < +\infty$, for every $x, y \in S$. The sequence $x^{(n)} \in S$ converges to $x \in S$ 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 $\lim n x^{(n)} = x$ in metric $d_S$, and say that $x$ is a $d_S$-limit for $x^{(n)}$.

Now we have:

Lemma 1. Every sequence $u = (u_k)(k \in \mathbb{N})$ from $S$ is a $d_S$-limit of the sequences $u(n)(n \in \mathbb{N})$, having the form

$$(3) \quad u(n) = \sum_{k=0}^{n} u_k e_k.$$ 

Proof. We have

$$d_S(u(n), u) = \sum_{i=n+1}^{+\infty} 2^{-i} \frac{|u_i|}{1 + |u_i|} \leq \sum_{i=n+1}^{+\infty} 2^{-i} = 2^{-n},$$

wherefrom $d_S(u(n), u) \to 0$ when $n \to \infty$. Consequently,

$$u = \lim_{n} u(n) = \lim_{n} \left( \sum_{k=0}^{n} u_k e_k \right) = \sum_{k=0}^{+\infty} u_k e_k$$

(in $d_S$-metric)

for every $u \in S$.

3. REPRESENTATIONS

From the lemma 1 we have that the representation

$$(4) \quad u = \sum_{k=0}^{+\infty} u_k e_k.$$
n for every \( u = (u_k) \in S \). However, we need the following statement concerning representations.

Let the sequences \( E_0, E_1 \) and \( W_k \) be defined with

\[
E_0 = \sum_{k=0}^{+\infty} e_k, \quad E_1 = \sum_{k=0}^{+\infty} k e_k, \quad W_k = \sum_{i=k+2}^{+\infty} (i-k-1)e_i \quad (k \in \mathbb{N}_0).
\]

Now, we have

**Theorem 1.** (a) Every sequence of the form

\[
u(n) = \lambda(n)E_0 + \mu(n)E_1 + \sum_{k=0}^{n} c_k(n)W_k \quad (n \in \mathbb{N}_0),
\]

where \( \lambda(n), \mu(n) \in \mathbb{R}, \ c_k(n) > 0 \ (k \in \mathbb{N}_0) \) for fixed \( n \), is convex, i.e. depends to the cone \( K \).

(b) Every sequence \( u \in K \) is a limit (in \( d_s \) metric) of sequences \( u(n) \) given by (6).

**Proof.** (a) It is obvious that \( E_0 = (1,1,1,\ldots) \) and \( E_1 = (0,1,2,3,\ldots) \), i.e. \( \Delta^2 E_{0k} = \Delta^2 E_{1k} = 0 \ (k \in \mathbb{N}_0) \). From (5) we also have \( W_{ki} = \{(i+1)^2, \ i \geq k+2 \} \) which gives \( \Delta^2 W_{ki} = \{1, \ i \geq k-2 \} \), i.e. \( \Delta^2 W_{ki} \geq 0 \) for every \( i \in \mathbb{N}_0 \) and \( k \in \mathbb{N}_0 \). In virtue of nonnegativity of \( c_k(n) \) we have from (6) \( \Delta^2 u(n) \geq 0 \). Accordingly, \( u(n) \) is convex for every \( n \in \mathbb{N}_0 \).

(b) Substituting the obvious identity

\[u = u_0 + k \Delta u_0 + \sum_{i=0}^{k-2} (k-i-1) \Delta^2 u_i\]

into (4) we have

\[
u = u_0 e_0 + u_1 e_1 + \sum_{k=2}^{+\infty} \sum_{i=0}^{k-2} (u_0 + k \Delta u_0 + \sum_{i=0}^{k-2} (k-i-1) \Delta^2 u_i) e_k.
\]

After some transformations (7) get the form

\[
u = u_0 \left( \sum_{k=0}^{+\infty} e_k \right) + \left( \sum_{k=0}^{+\infty} k e_k \right) \left( \sum_{i=k+2}^{+\infty} (i-k-1)e_i \right),
\]

which is a \( d_s \)-limit of the sequence.
\[ u(n) = \sum_{k=0}^{\infty} u_k e_k + \Delta u o \sum_{k=0}^{\infty} k e_k + \sum_{i=0}^{\infty} (\Delta^2 u_k)(\Sigma_{i=0}^{\infty} (i-k-1)e_i) \], or by notation introduced by (5) \[ u(n) = u_o \mathcal{E}_0 + (\Delta u_o) \mathcal{E}_1 + \sum_{k=0}^{n} (\Delta^2 u_k)W_k. \] Thus, we have \[ d_s(u(n), u) = \sum_{k=n+2}^{\infty} 2^{-k} \frac{|u_k - v_k|}{1 + |u_k - v_k|} 2^{-n}, \] for every fixed \( n \), where \( v_k \) \((k \geq n+2)\) 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 \( A:S \to \mathbb{F}(D) \) be linear and continuous over the sequences in \( S \). Then, for every \( u \in S \), the implication

\[ u \in \mathcal{K} \iff A(u) \geq 0 \]

holds if and only if

\[ A \mathcal{E}_o = A \mathcal{E}_1 = 0, \]
\[ A W_k \geq 0 \quad (k \in \mathbb{N}_0). \]

**Proof.** i) Suppose that (8) holds for every \( u \in S \). Then, if we choose \( u = \mathcal{E}_o (u = \mathcal{E}_1) \) we have that \( u \in \mathcal{K} \) which imply \( A(u) \geq 0 \), i.e. \( A \mathcal{E}_o \geq 0 \) (\( A \mathcal{E}_1 \geq 0 \)). But \( -u \in \mathcal{K} \) too. Thus, \( A(-\mathcal{E}_o) \geq 0 \) or \( A(-\mathcal{E}_1) \geq 0 \) from which \( A \mathcal{E}_o = A \mathcal{E}_1 = 0 \). By theorem 1-a \( W_k \in \mathcal{K} \) \((k \in \mathbb{N}_0)\), so, in virtue of (8) we have \( A W_k \geq 0 \) for every \( k \in \mathbb{N}_0 \).

ii) Suppose now that (9) and (10) holds. Then, on the basis of theorem 1-b, every sequence \( u \in \mathcal{K} \) is \( d_s \)-limit of the sequence \( (u(n))(n \in \mathbb{N}_0) \) given by (6). This means that \( u = \lim u(n) \), wherefrom, accordingly with continuity of \( A \) over the sequences in \( S \), we have

\[ A u = A \left( \lim_n u(n) \right) = \lim_n \left( A u(n) \right), \]

and, in virtue of linearity of \( A \) over the sequences in \( S \), we get

\[ A u = \lim_n \left( \lambda(n)(A \mathcal{E}_o) + u(n)(A \mathcal{E}_1) + \sum_{k=0}^{n} c_k(n)(A W_k) \right) \]

and, if (9) and (10) holds, and keeping in the mind that \( c_k(n) \geq 0 \),
Remark 1. It is easy to see that our theorem 2 generalizes the result of theorem 4 in [11]. 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 [21]. The sequences $W_k$ we can call a discrete splines.

REFERENCES


Approximation of Convex Functions by First Degree Splines

Ljubiša M. Kocić

Abstract:
A method for approximation of functions convex on a finite interval by piecewise affine function is developed. For the segments of approximating function we used the support affine function 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, udeo po deo afinom funkcijom. Segmenti aproksimirajuće funkcije su potporne affine 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 (piecewise affine functions, polygonal lines) is minutely studied from many authors, and for classes \( C[a, b] \) and \( C^2[a, b] \), as well as the interpolated classes \( H_\omega \) and \( WH_\omega \), see, for example, [1] and [3]. A lot of results are oriented to applications on computers [7]. In all appearance, this kind of approximation is especially important for convex functions. Early results was obtained by K. Toda [6] and T. Popoviciu [4]. They were shown the following theorem:

Theorem 1. The first degree spline function

\[
(S_n f)(x) = p x + q + \sum_{k=0}^{n} c_k (x - x_k)_+^+, \quad x \in [a, b], \quad n \in \mathbb{N},
\]
where \( p, q \in \mathbb{R}, c_k \geq 0, x_k \in [a, b] \) \((k = 0, 1, \ldots, n)\) 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 \( x_k \in [a, b], p = [x_0, x_1; f], q = f(x_0) - x_0 [x_0, x_1; f] \) 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 knots \( x_k \) \([a, b]\), and if we introduce for \( x \in [a, b] \)

\[
(2) \quad \text{epi} f = \left\{ (x, y) \in \mathbb{R}^2 \mid y \geq f(x) \right\},
\]

\[
(3) \quad \text{hyp} f = \left\{ (x, y) \in \mathbb{R}^2 \mid y \leq f(x) \right\},
\]

it is easy to see that \( (S_n f)(x) \in \text{epi} f \), for \( n \in \mathbb{N} \), and \( x \in [a, b] \).

On the basis of theorem 1, P. M. VASIĆ and I. B. LAKOVIĆ were 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 \( (x, y) \rightarrow f(x, y) \). 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 convenient. Thus, we shall develope this kind of approximation for one variable function. This circumscribed spline will 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, \( (S_n f)(x) \in \text{hyp} f \).

2. PRELIMINARY LEMAS

Function \( f: [a, b] \rightarrow \mathbb{R} \) is convex on \([a, b]\) if the inequality \( f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda)f(v) \) holds for every \( u, v \)
The function $f$ is called strictly convex if above inequality is strict. Let $K[a, b]$ and $K^+[a, b]$ denote the cones of convex and strictly convex functions on $[a, b]$, continuous from the left (right) at the edge point $a$ ($b$). The function $x \mapsto p(t, x)$ is called support function of convex function $f$ if the following conditions are fulfilled:

$$f(x) \geq p(t, x), \quad x \in [a, b], \quad x \neq t \quad \text{and} \quad f(t) = p(t, t).$$

For strictly convex functions the following lemma takes place see for ex. [51]:

**Lemma 1.** Let $f: [a, b] \to \mathbb{R}$ be strictly convex on $[a, b]$, and let $f'(x)$ be the right derivative of $f$ in $x$. Then

- (a) $f'_+$ is increasing function on $[a, b]$,
- (b) for every $a \leq u < v \leq b$ holds

$$f'_+(u) < \frac{f(v) - f(u)}{v - u} < f'_+(v).$$

Let $f \in K^+[a, b]$, and $x_1 < x_2 < \ldots < x_n$ be a set of knots from $(a, b)$. The first order spline, circumscribed around the graph of $f$ is given by

$$s_f(x) = \sup_{1 \leq i \leq n} \left\{ p(x_i, x) \right\}, \quad x \in [a, b],$$

where $p(x_i, x)$ is affine support function of $f$ in the point $x_i$:

$$p(x_i, x) = f(x_i) + f'(x_i)(x - x_i).$$

So we have

**Lemma 2.** a) $s_f$ interpolates $f$ in the knots $x_1, \ldots, x_n$,

- (b) $s_f$ is convex on $[a, b]$.

**Proof.** a) Let $1 \leq k \leq n$ be a fixed number. Then, $\sup_{1 \leq i \leq n} \left\{ p(x_i, x_k) \right\} = p(x_k, x_k) = f(x_k)$ which is an interpolate property.

- (b) As every function $x \mapsto p(x_i, x)$ ($i = 1, 2, \ldots, n$) is
convex, so \{ p(x_i, x) \mid 1 \leq i \leq n \} is a family of convex functions. But, it is known ([51]) that \( \sup_{1 \leq i \leq n} p(x_i, x) \) is also convex.

Now, let \( x_k \) and \( x_{k+1} \) be two adjacent knots. Corresponding 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

\[
(6) \quad t_k = \frac{f'(x_{k+1})x_{k+1} - f'(x_{k})x_{k} + f(x_{k}) - f(x_{k+1})}{f'(x_{k+1}) - f'(x_{k})}
\]

if \( f'(x_{k+1}) \neq f'(x_{k}) \) i.e. if \( f \in K^+[a,b] \). The points \( t_k \) (\( k = 1, 2, \ldots, n-1 \)) are the abscissae 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

\[
(7) \quad x_k < t_k < x_{k+1}
\]

holds for every \( k = 1, 2, \ldots, n-1 \).

Proof. From lemma 1 - b), we have that \( f'(x_k) < \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} \)

i.e. from \( x_{k+1} - x_k > 0 \) follows \( (x_{k+1} - x_k)f'(x_k)f'(x_{k+1}) - f(x_k) \)

wherefrom we get \( x_{k+1}f'(x_{k+1}) - x_kf'(x_k) + f(x_k) - f(x_{k+1}) < x_k \)

\( *(f'(x_{k+1}) - f'(x_k)) \) 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 sequel, we introduce a v-shaped function \( v_k \) w:

\[
v_k(x) = \sup \left\{ p(x_k, x), p(x_{k+1}, x) \right\}, \quad x_k < x \leq x_{k+1}
\]

which approximates \( f \) on \([x_k, x_{k+1}]\). Let \( E_k \) be defined w: \( E_k = f - v_k \), 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

\[
(8) \quad E(x) = \begin{cases} f(x) - p(x_k, x), & x \in [x_k, t_k), \\ f(x) - p(x_{k+1}, x), & x \in [t_k, x_{k+1}], \end{cases}
\]
We shall prove that \( E(x) \) monotonely increasing on \((x_k, t_k)\). Let \( x_k < x < y < t_k \). Then, we can find \( \lambda \in (0, 1) \) so that \( x = \lambda x_k + (1 - \lambda)y \), which, with strict convexity of the function \( f \) gives

\[
f(x) < \lambda f(x_k) + (1 - \lambda)f(y).
\]

From (8), (5) and (9) we have \( E(x) = f(x) - f(x_k) - f'(x_k) (x-x_k) \)

\[
\begin{align*}
\lambda f(x_k) + (1 - \lambda)f(y) - f(x_k) - f'(x_k) (x-x_k) &= (1-\lambda)[f(y)-f(x_k)] \\
- f'(x_k) (\lambda x_k + (1-\lambda)y - x_k) &= (1-\lambda)[f(y)-f(x_k)] \\
- f'(x_k) (y-x_k) &= (1-\lambda)E(y) for x \in [x_k, t_k], \text{ which means that } E(x) < E(y),
\end{align*}
\]

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. \( E_k = \sup \{ E(x) \} = E(t_k) = f(t_k) - v_k(t_k) \).

3. APPROXIMATION

On the basis of previous lemmas we can state

**Theorem 2.** a) The spline \( s_n \) have explicit form

\[
(s_n f)(x) = Ax + B + \sum_{k=1}^{n-1} d_k(x-t_k), \quad x \in [a, b],
\]

where \( A = (T_1 - T_0)/(t_1 - t_0), \ B = (t_1 T_0 - t_0 T_1)/(t_1 - t_0), \ T_k = p(x_k, t_k), \ d_k = f'_+(x_{k+1}) - f'_+(x_k), \ t_k \) is given by (6) and \( t_0 = a. \)

b) For every \( f \in K^+[a, b] \), \((s_n f)(x)\) approximates \( f \) uniformly on \([a, b]\) when \( n \to \infty \) and \( \max(x_{k+1} - x_k) \to 0. \)

**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, \( d_k \geq 0 \) \( k=1, \ldots, n-1. \) The
proof of b) follows from lemma 4. Namely, for \( x \in [x_k, x_{k+1}] \), we have
\[
|f(x) - v_k(x)| \leq E(t_k) = f(t_k) - f(x_k) - f'_+(x_k)(t_k - x_k),
\]
and if we put \( h = x_{k+1} - x_k \), then
\[
|f(t_k) - f(x_k)| \leq \omega(f,h),
\]
and also
\[
t_k - x_k < h,
\]
wherefrom
\[
|f(x) - v_k(x)| < \omega(f,h) + h \to 0,
\]
and according ly, \( f(x) - s_n(x) \to 0 \), when \( n \to \infty \) and \( \max(x_{k+1} - x_k) \to 0 \).

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 remained graph will be a strictly convex function. Now, we underline that the form of the spline does not new. There is no difference, in formal sense, between \( S_n f \) and \( s_n f \). However, we have \( s_n f \not\leq f \not\leq S_n f \) on \([a,b]\), and from this reason, we call \( S_n f \) an upper spline and \( s_n f \) a lower spline of the convex function \( f \). There is, also, a middle spline, which have been studying by M. GAVRILOVIĆ in [2] and provid the mini-max approximation. The spline \( S_n f \) exists for every continuous function. But, the middle and the lower spline do exist only for convex functions.

REFERENCES

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 \( \sigma_i = \frac{2\sinh^2(h_i \sqrt{q_i}/2)}{h_i^2 q_i} \) 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 \( (\varepsilon_a = 0, h_0 = Mh^2) \).

Consider the problem
\[
\begin{align*}
-\gamma'' + q(x)\gamma &= f(x), \quad q(x) \geq 0, \quad x \in [a, b], \quad (a, b \in \mathbb{R}), \\
\alpha_0 y(a) + \beta_0 y'(a) &= \gamma_a, \quad |\alpha_0| + |\beta_0| \neq 0, \\
\alpha_1 y(b) + \beta_1 y'(b) &= \gamma_b, \quad |\alpha_1| + |\beta_1| \neq 0.
\end{align*}
\]

The approximate solution of the problem (1) we want to obtain in the form of the cubic spline \(v(x) \in C^2[a,b]\) on the grid

\[a = x_0 < x_1 < \ldots < x_{n+1} = b\]

The restriction \(v(x)\) on \([x_i, x_{i+1}]\) is \(v_i(x), v_i(x) = v_i^{(0)} + v_i^{(1)}(x-x_i) + \frac{1}{2} v_i^{(2)}(x-x_i)^2 + \frac{1}{6} v_i^{(3)}(x-x_i)^3\) \((i=0,1,\ldots,n)\)

where \(v_i^{(k)}\) are constants which approximating \(y_i^{(k)} = y^{(k)}(x_i)\).

Using the equations

\[-q_i v_i^{(2)} + q_i v_i^{(0)} = f_i \quad (i=0,1,\ldots,n)\]

\[-\sigma_{n+1} v_{n+1}^{(2)} + \sigma_n v_n^{(3)} + q_{n+1} v_{n+1}^{(0)} + v_n^{(1)} = f_n + \frac{h^2}{6} \quad (i=n+1)\]

and the suppositions on the continuity we obtain

\[(2) \quad \sum_{i=0}^{i=0} h v_i^{(0)} = k_i v_{i-1}^{(0)} + m_i v_i^{(0)} - m_i v_{i+1}^{(0)} = R_i, \quad (i=0,1,\ldots,n)\]

where \(h_i = x_{i+1} - x_i, \quad h = \max h_i, \quad h_{n+1} = h_n\)

\[k_i = (1 - \frac{h_{i-1} q_{i-1}}{6 \sigma_{i-1}}) \cdot \frac{1}{h_{i-1}}, \quad m_i = (1 - \frac{h_{i+1} q_{i+1}}{6 \sigma_{i+1}}) \cdot \frac{1}{h_{i+1}}\]

\[R_i = \frac{h_{i-1} f_{i-1}}{6 \sigma_{i-1}} + \frac{f_i}{3 \sigma_i} (h_{i-1} + h_i) + \frac{f_{i+1} h_i}{6 \sigma_{i+1}}, \quad (i=1,\ldots,n-1)\]

\[k_o = 0, \quad \ell_o = \frac{3 \sigma_o + h_o^2 q_o}{3 \sigma o} \cdot \frac{\beta}{\sigma_o} - \frac{\alpha}{3 \sigma_o}, \quad m_o = \frac{(6 \sigma_1 - h_o q_1) \beta_a}{6 \sigma_1 h_o}\]

\[R_o = -\gamma_a - \beta_a \frac{h_o}{6} (\frac{f_o}{\sigma_o} + \frac{f_1}{\sigma_1})\]
\[ c_n = (1 - \frac{h_{n-1}^2 q_{n-1}}{6 \sigma_{n-1}^2}) \cdot \frac{1}{h_{n-1}^2}, \quad m_n = 0 \]

\[ g_n = \frac{1}{h_{n-1}^2} (1 + \frac{h_{n-1}^2 q_n}{3 \sigma_n^2}) + 6F, \quad V = q_{n+1} - A(\frac{q_n}{\sigma_n} + \frac{6a_B}{h_n B}) + \frac{q_n}{\sigma_n} E \]

\[ F = h_n q_{n+1} - A \cdot D, \quad A = -\sigma_{n+1} + \frac{q_{n+1} h_n^2}{6} \]

\[ D = (6a_B h_n + 6\beta_B)/(h_n B), \quad B = a_B h_n + 3\beta_B \]

\[ C = \frac{3a_B h_n + 6\beta_B}{a_B h_n + 3\beta_B}, \quad E = -\sigma_{n+1} + \frac{q_{n+1} h_n^2}{2} \]

\[ R_n = h_n (-\frac{q_{n-1}^2}{6 \sigma_{n-1}^2} + \frac{1}{F}) \left[ \frac{f_n + f_{n-1}}{\sigma_n} - A(\frac{q_n}{\sigma_n} + \frac{6y_B}{h_n B}) \right] \]

\[ \sigma_i = \frac{2sh^2(h_i \sqrt{q_i})}{h_i q_i} \]

The constants \( v_i^{(k)}, k=1,2,3 \) we obtain from the relations

\[ a_{v_o} v^{(0)} + \beta_{v_o} v^{(1)} = \gamma_a \]

\[ v_1^{(1)} = (\alpha_{v_1} v_1^{(0)} - \beta_{v_1} v_1^{(0)} - s_1) \cdot h_1^{-1}, \quad (i=1,n) \]

\[ a_i = 1 - \frac{h_{i-1}^2 q_i}{6 \sigma_i}, \quad \beta_i = 1 + \frac{h_{i-1}^2 q_{i-1}}{3 \sigma_{i-1}^2}, \quad s_i = -\frac{h_{i-1}^2}{2} \left( \frac{2f_{i-1}}{\sigma_{i-1}^2} + \frac{f_i}{\sigma_i} \right) \]

\[ v_1^{(1)} = v_1^{(1)} + r_n \cdot h_{n-1} \left( \frac{q_n v_1^{(0)}}{\sigma_n} + \frac{q_{n-1} v_{n-1}^{(0)}}{\sigma_{n-1}} \right) \]

\[ r_n = -\frac{h_{n-1}}{2} \left( \frac{f_{n-1}}{\sigma_{n-1}^2} + \frac{f_n}{\sigma_n} \right) \]

\[ -\sigma_i v_1^{(2)} + q_i v_1^{(0)} = f_1, \quad v_1^{(2)} = v_1^{(2)} + h_{n-1} v_{i-1}^{(3)}, \quad (i=1,2,\ldots,n) \]

\[ -\sigma_{n+1} v_n^{(2)} + h_n v_n^{(3)} + q_{n+1} (v_n^{(0)} + v_1^{(1)} h_n + h_n^2 v_n^{(2)})/2 + h_n^3 v_n^{(3)}/6 = f_{n+1} \]

Similar to [2] it can be shown that \( z_i^{(k)} = v_i^{(k)} - v_1^{(k)} \) satisfy the equations.
\begin{align*}
(3) \quad l^h z_i^{(0)} &= -k_i z_{i-1}^{(0)} + \psi_i z_i^{(0)} - m_i z_{i+1}^{(0)} = -\psi_i, \quad (i=0, \ldots, n) \\
\psi_i &= \phi_i^{(2)} h_i^{-1} - \phi_i^{(2)} h_{i-1}^{-1} + \phi_i^{(1)}, \quad (i=1, \ldots, n-1) \\
\psi_0 &= \phi_i^{(2)} \beta h_0^{-1} \\
\psi_n &= -\phi_n h_n^{-1} \phi_i^{(1)} - \frac{1}{\sigma} [\eta n - \psi - \frac{6 \psi A - \eta n}{\sigma} - \eta] (E - A \cdot C) \\
\eta_i &= \psi_i^{(2)} (\sigma - 1), \quad \phi_i^{(2)} = \psi_i^{(0)} + h_i^2 (\eta n + \eta i - \frac{\psi_i^{(2)}}{6}) \\
\phi_i^{(1)} &= \psi_i^{(1)} - \frac{h_i - 1}{2} \psi_i^{(2)} h_i^{-1} - \frac{h_i - 1}{2} (\eta i - \eta i + \eta_i) \\
\psi &= -\sigma n + \psi_i^{(2)} + q_{n+1} \psi_{n+1}, \quad \psi_n = -\beta \psi_{n+1} - \beta \psi_{n+1} \\
\psi^{(k)} &= \frac{y^{IV}(\theta) h_i^{-1}}{(4-k) i} \psi_i^{(k)} - \psi_i^{(k)} (k=0, 1, 2), \quad y_i < \theta_i < y_{i+1} \\
\alpha_1 z_0^{(o)} + \beta_1 z_0^{(1)} &= 0 \\
z_i^{(1)} &= z_i^{(1)} - \frac{h_i - 1}{2} + \frac{q_{i-1} z_i^{(1)} - q_i z_i^{(0)}}{\sigma_i - 1} + \phi_i^{(1)}, \quad (i=1, \ldots, n) \\
\alpha_1 z_i^{(o)} &= \beta_1 z_i^{(o)} + h_i z_i^{(1)} + \psi_i^{(1)}, \quad (i=1, \ldots, n) \\
z_i^{(2)} &= \psi_i^{(1)} (\eta_i + q_i z_i^{(0)})^{(1)} \\
z_{i-1}^{(3)} &= (z_i^{(2)} - z_i^{(2)} - \psi_i^{(1)}) h_i^{-1} \\
-\sigma n + (z_i^{(2)} + h_i z_i^{(3)}) + q_{n+1} (z_i^{(o)} + h_k z_i^{(1)} + h_n z_i^{(2)}) + \frac{h_i^2 z_i^{(2)}}{6} + \\
-\sigma n + (z_i^{(2)} + h_i z_i^{(3)}) + q_{n+1} (z_i^{(o)} + h_k z_i^{(1)} + h_n z_i^{(2)}) + \frac{h_i^2 z_i^{(2)}}{6} &= -\sigma n - \psi
\end{align*}

**THEOREM 1.** Let $6z_i^{(h_2 - h_{i-1})} q_i^{(h_1 - h_{i-1})} > 0$, $\alpha_i^2 A - \beta_i \cdot B > 0$ and at least one of $q_k (k=i-1, i, i+1)(q_{i-1} = 0)(i=0, 1, \ldots, n)$ is different from zero. Then the matrix of the system (2) is inverse monotone.

**COROLLARY 1.** The condition $6z_i^{(h_2 - h_{i-1})} q_i^{(h_1 - h_{i-1})} > 0$ in theorem 1 we can be replaced by $q_i^{(h_2 - h_{i-1})} + 6 > 0$. 
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 \( \beta_a = 0 \)

\[
|z_{i}^{(k)}| \leq Mh \quad (k=0,1,2), \quad |z_{i}^{(3)}| \leq M ,\text{ where } M \text{ denotes different constants, independent on } h.
\]

Proof.

(10) \( \Delta_{i} = -k_{i}^{2} - m_{i} = \frac{h_{i-1} \sigma_{i} + q_{i} \sigma_{i-1} + Q_{i+1} \sigma_{i+1}}{6 \sigma_{i-1} \sigma_{i} \sigma_{i+1}} > Mh \)

\( (i=1,\ldots,n) \)

(11) \( \Delta_{o} = \frac{\beta_{o} (\beta_{1} - \sigma_{1})}{h_{o}} - \sigma_{a} > Mh \)

(12) \( \Delta_{n} = \frac{1}{h_{n}} (\beta - \alpha_{n}) + \frac{h_{n-1} \sigma_{n} - q_{n} \sigma_{n-1} + Q_{n+1}}{2 \sigma_{n} + \sigma_{n-1}} + V^{-1} > Mh, \quad M > 0 \)

From (10), (11) and (12) we obtain that \( ||A^{-1}|| \leq \frac{1}{\max_{i} \Delta_{i}} < Mh^{-1} \).

\( A \) 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 theorem 1 are fulfilled. Then

\[
|z_{i}^{(o)}| \leq Mh^3; \quad |z_{i}^{(k)}| \leq Mh^2; \quad |z_{i}^{(3)}| \leq Mh , \quad (k=1,2)
\]

Proof. A simple calculation shows that \( \tilde{A} > h^{-1}B, \)

where \( \tilde{A} \) is matrix determined by (3) for \( i=1,\ldots,n \), and \( z_{o}^{(i)} = 0, B = \{b_{ij}\} (i,j=1,\ldots,n) \) 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
\( B u = \omega, \ omega = \max_{i} \psi_{i} h \) has the form \( u_{i} = \frac{i(n+1-i)}{2} \omega \). Since \( \left| z_{i}^{(1)} \right| \leq u_{i} \) we have \( \left| z_{i}^{(1)} \right| \leq u_{1} = o(h^{3}), \left| z_{n}^{(1)} \right| \leq u_{n} = o(h^{3}). \)

Then form (3) for \( i = 2 \) we obtain \( z_{2}^{(1)} = o(h^{3}) \) and then by induction we can conclude that \( \left| z_{i}^{(1)} \right| = o(h^{3}) \) (\( i = 2, \ldots, n \)). The estimates for derivatives we get from (5)-(9).

**Theorem 3.** Let \( \beta = 0 \) and \( h_{o} = M h^{2} \). Then
\[
\left| z_{i}^{(k)} \right| \leq M h^{4-k} \quad (k = 0, 1, 2, 3).
\]

**Proof.** See [4].

**References**

MESH CONSTRUCTION FOR NUMERICAL SOLUTION
OF A TYPE OF SINGULAR PERTURBATION PROBLEMS

Relja Vulanović

ABSTRACT:

The singular perturbation problem (1) is considered. It is solved numerically by classical difference schemes on a non-uniform mesh. The discretization mesh is constructed in a special way, which gives linear convergence uniform in small perturbation parameter.

1. INTRODUCTION

We consider the two point boundary value problem

(1a) \( L_{\epsilon} u := \epsilon^2 u'' + x b(x) u' - c(x) u = f(x), \ x \in I = [0,1] \)

(1b) \( u(0) = U_0, \ u(1) = U_1, \)

with basic assumptions

\[ b, c, f \in C^2(I) \]

\[ b(x) > 2\beta > 0, \ c(x) > \gamma > 0, \ 2b(0) < c(0), \]

\[ 0 < \epsilon < \epsilon_0. \]

This problem was solved in [2] by a special method which gives linear convergence uniform in small perturbation parameter \( \epsilon \). Our method seems to be somewhat simpler. It is based on the idea of Bahvalov, [1], that was genera-
lized in [3] and uses a special mesh construction. We also achieve linear convergence uniform in \( \epsilon \) 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 \( u_\epsilon \in C^4(I) \) to the problem (1). We use the result from Theorem 2. from [2]. Each positive constant independent of \( \epsilon \) and of discretization mesh will be denoted by \( M \).

**THEOREM 1.** For the solution \( u_\epsilon \) to the problem (1) the following estimates hold:

\[
(2) \quad |u_\epsilon^{(i)}(x)| \leq M(\epsilon^{-1} + \epsilon^{-1} \exp(-\beta(\frac{x}{\epsilon})^2)) ,
\]

\( i=0,1,2,3 \),

where \( \epsilon^- = \text{max}(0, i-2) \).

**Proof.** Using the same proof as in [2], we obtain

\[
(3) \quad |u_\epsilon^{(i)}(x)| \leq M(1 + \epsilon^{-1} \exp(-h(x)/\epsilon^2)) ,
\]

\( i=0,1,2 \), \( h(x) = \int_0^x tb(t)dt \). We use this inequality to get (2) for \( i=0,1,2 \). This part of the proof needs the assumption \( u_\epsilon \in C^4(I) \) and \( 2b(0) < c(0) \).

Let us now obtain the estimate (2) for \( i=3 \). Differentiating both sides of (1a) twice, we have

\[
\epsilon^2 u^{IV} + xb(x) u''' = g(x) ,
\]

where, according to (3):

\[
|g(x)| \leq M(1 + \epsilon^{-2} \exp(-h(x)/\epsilon^2)) .
\]

Now it follows

\[
u^{''''}(x) = \exp(-h(x)/\epsilon^2)[u^{''''}(0) + \epsilon^{-2} \int_0^x g(t) \exp(h(t)/\epsilon^2) dt] .
\]

Since from [2] we have \( |u^{''''}(0)| \leq M \epsilon^{-3} \), we get:

\[
|u^{''''}(x)| \leq M(A + B + C) ,
\]

with

\[
A = \epsilon^{-3} \exp(-h(x)/\epsilon^2) ,
\]

and

\[
B = \epsilon^{-3} \exp(-h(x)/\epsilon^2) ,
\]

\[
C = \epsilon^{-3} \exp(-h(x)/\epsilon^2) .
\]
\[ B = \varepsilon^{-2} \int_0^x \exp\left(\frac{(h(t) - h(x))/\varepsilon^2}{\varepsilon^2}\right) dt, \]
\[ C = \varepsilon^{-4} x \exp\left(-\frac{h(x)}{\varepsilon^2}\right). \]

Now we have
\[ A, C \leq M \varepsilon^{-3} \exp\left(-\beta \frac{x}{\varepsilon}\right) \]
and, see [2],
\[ B \leq \frac{1}{\varepsilon \beta y} \left(1 - \exp\left(-\beta y^2\right)\right), \ y = x/\varepsilon. \]

Hence,
\[ B \leq M \varepsilon^{-1} \]
and the theorem is proved.

2. THE MESH CONSTRUCTION

Let us denote by \( q \) a fixed number, \( q \in (0,1) \), independent of \( \varepsilon \) 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
\[ x_i = \lambda(t_i), \ t_i = i/n, \ i=0,1,\ldots,n, \]
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 \( \alpha \) we can get:
\[ \alpha = (q-\alpha q (1-q+a \varepsilon))^{1/2}/(1+a \varepsilon). \]
ote: 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 $b(x)=0$.

On this mesh we form the discretization of the problem (1):

$$
\begin{align*}
L_h u_i &= \varepsilon^2 D'' u_i + x_i b(x_i) D' u_i - c(x_i) u_i = f(x_i), \quad i=1,2,\ldots,n-1, \\
\begin{align*}
u_n &= U_1,
\end{align*}
\end{align*}
$$

where

$$
\begin{align*}
D'' u_i &= 2(h_{i+1} u_i - (h_i + h_{i+1}) u_i + h_i u_{i+1})/(h_i h_{i+1} (h_i + h_{i+1})), \\
D' u_i &= (u_{i+1} - u_i)/h_{i+1}, \\
h_i &= x_i - x_{i-1}, \quad i=1,2,\ldots,n.
\end{align*}
$$

3. CONVERGENCE UNIFORM IN $\varepsilon$

Because of $c(x) \geq \gamma > 0$ we can easily get that the scheme (4) is stable uniformly in $\varepsilon$, see [3], for instance.

Now we shall state our main result.

THEOREM 2. For the solution $u_\varepsilon$ to the problem (1) and for the solution $u_i$ to (4) we have

$$
|u(x_i) - u_i| < M \frac{1}{n}, \quad i=0,1,\ldots,n.
$$

Proof. We only have to prove consistency uniform in $\varepsilon$, i.e.

$$
|r_i| < M \frac{1}{n}, \quad i=1,2,\ldots,n-1,
$$

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
Another estimate for $r_i$ is

$$
| r_i | \leq M (\varepsilon^2 + v_{i-1} + x_i + \frac{x_i}{\varepsilon} v_i ) .
$$

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

$$
v_i \leq v_{i-1} \leq \exp(-b \varepsilon^2 \phi(\alpha)) .
$$

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 \prime(t_{i+1}) v_{i-1} \leq

\leq M (q - t_{i-1})^{-2} \exp(-b \varepsilon^2 \left( \frac{t_{i-1}}{q - t_{i-1}} \right)^2) \leq M .
$$

Similarly, from (6c) we have

$$
Q_i \leq a^2 \phi(t_i) \phi \prime(t_{i+1}) v_i \leq M
$$

and (6a) give us (5).

3° The last case is

$$
q - \frac{4}{n} < t_{i-1} < \alpha .
$$

From this inequality it follows

$$
q - \alpha \leq \frac{4}{n}
$$

and
(8) \[ \sqrt{e} < M \frac{1}{n}, \]

because

\[ q - a > \frac{\sqrt{(1-q)ag \varepsilon}}{1+q} \]

Now \( x_{i-1} > a \phi (q - \frac{4}{n}) \), (notice \( q - \frac{4}{n} > 0 \)), and we get

(9) \[ v_{i-1} \leq M \frac{1}{n}. \]

Similarly:

(10) \[ \frac{x_i}{e} v_i < M \sqrt{v_i} \leq M \frac{1}{n}. \]

For \( x_i \) we have

\[ x_i = \lambda(t_i) < \lambda(t_{i-1}) + M \frac{1}{n}, \]

and

\[ \lambda(t_{i-1}) < \lambda(a) < M \sqrt{e}, \]

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


AN ITERATIVE SOLUTION OF SOME DISCRETE ANALOGUES OF A MILDLY NONLINEAR BOUNDARY VALUE PROBLEM

Dragoslav Herceg, Ljiljana Cvetković

ABSTRACT:

In this paper we consider numerical solution of the system of nonlinear equations $A(x)x = BFx$ by the iteration $x^0 \in \mathbb{R}^n, x^{k+1} = A(x^k)^{-1}BFx^k, k=0,1,\ldots$. 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.

1. INTRODUCTION

We shall consider a system of nonlinear equations

$$A(x)x = BFx,$$

where $A(x), B \in \mathbb{R}^{n \times n}$ (= set of all $n \times n$ 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 abbreviate this as

\[
(A(x))_{11}, \ldots, (A(x))_{ij}, \ldots, (A(x))_{in} = (B_{11}, \ldots, B_{ij}, \ldots, B_{in}),
\]

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 parentheses. The diagonal elements are underlined.

The iteration which we shall consider for the solution of (1) is

\[
(2) \quad x^0 \in \mathbb{R}^n, \quad A(x)^k x^{k+1} = B F x^k, \quad k = 0, 1, \ldots.
\]

If \(A(x)\) is a regular matrix for all \(x \in \mathbb{R}^n\), the iteration (2) can be written in the form

\[
(3) \quad x^0 \in \mathbb{R}^n, \quad x^{k+1} = T x^k, \quad k = 0, 1, \ldots,
\]

where \(T x = (A(x))^{-1} B F x\).

In the 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 mildly 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 scheme (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 \(I_h = \{t_i = (i-1)h : i = 1, 2, \ldots, n\}\). For the numerical solution of problem

\[
(4) \quad -u'' + q(u)u = f(t, u), \quad t \in [0, 1]
\]

\[\quad u(0) = u(1) = 0,\]

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 \(F x \in \mathbb{R}^n\) whose \(i\)-th component is given via

\[\quad (F x)_i = f(t_i, x_i), \quad i = 1, 2, \ldots, n.\]

The matrices \(A(x)\) and \(B\) are defined by
(5) $h^{-2}(-1, 2 + h^2 q(x_i), -1) = (1)$ for $i = 2, 3, ..., n-1$, (second order approximation),

(6) $\frac{h^{-2}}{12} (1, -16, 30 + 12h^2 q(x_i), -16, 1) = (1)$, for $i = 3, 4, ..., n-2$, (fourth order approximation), and second order approximation for $i = 2, n-1$ as in (5),

(7) $\frac{h^{-2}}{180} (-2, 27, -270, 490 + 180h^2 q(x_i), -270, 27, -2) = (1)$ for $i = 4, 5, ..., n-3$, (sixth order approximation), and fourth order approximation for $i = 3, n-2$ as in (6), and a fourth order unsymmetric approximation

$$\frac{h^{-2}}{12} (-10, 15 + 12h^2 q(x_i), 4, -14, 6, -1) = (1)$$ for $i = 2$, $\frac{h^{-2}}{12} (-1, 6, -14, 4, 15 + 12h^2 q(x_i), -10) = (1)$ for $i = n-1$.

In (5), (6) and (7) we have $\lambda_i = (0)$ for $i = 1, n$.

The solution $x = [x_1, x_2, ..., x_n]^T \in \mathbb{R}^n$ of (1) is the numerical solution of the boundary value problem (4), i.e. $x_i = u(t_i)$, $i = 1, 2, ..., n$.

THE CONVERGENCE ANALYSIS

Theorem. Let $A(x) = [a_{ij}(x)] \in \mathbb{R}^{n \times n}$ is inverse-monotone matrix for all $x \in \mathbb{R}^n$ and let $BF$ is Frechet-differentiable in $\mathbb{R}^n$. Suppose that

$$\max_{1 \leq i \leq n} \sum_{j=1}^{n} \left| \frac{\partial a_{ij}(x)}{\partial x_j} \right| < M, \quad \| BFx \|_\infty \leq M_0, \quad \| (BF)'(x) \|_\infty \leq M_1,$$

$$\| A^{-1}(x) \|_\infty \leq \alpha, \quad x \in \mathbb{R}^n, \quad \alpha^2 M_0 M + \alpha M_1 < 1.$$

Then the equation (1) has a unique solution and the sequence $x^0, x^1, x^2, ..., \text{generated by (3)}$ 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 $\|C(y)\|_\infty \leq \alpha$, $\| (BF)'(y) \|_\infty \leq M_1$ it follows

$$\|C(y)(BF)'(y)\|_\infty \leq \alpha M_1.$$ 

Let $G = [g_{ij}] = (C(x)BFy)'(y)$, $H_p = [\frac{\partial a_{ij}}{\partial x_p}] \in \mathbb{R}^{n,n}$, $p=1,2,\ldots,n$.

Then

$$g_{ij} = \sum_{k=1}^{n} \frac{\partial C_{ik}(y)}{\partial x_j} (BFy)_k = \sum_{k=1}^{n} (CH_jC)_{ik}(BFy)_k = (CH_jCBFy)_i.$$ 

Since $C(y) \geq 0$, it follows,

$$\max_{1 \leq i \leq n} \sum_{j=1}^{n} |g_{ij}| \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |(CH_jC)(BFy)_i| = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |(CH_jC)(BFy)|_i.$$ 

$$\max_{1 \leq j \leq n} \sum_{i=1}^{n} |(CH_jC)(BFy)|_i = \|C\|_\infty \|\sum_{j=1}^{n} |H_j|_C|BFy|\|_\infty \leq \|C\|_\infty \|H_j\|_C \|BFy\|_\infty \leq \alpha^2 M_0.$$ 

Now we have

$$\|T'(y)\|_\infty \leq \alpha^2 M_0 + \alpha M_1 < 1.$$ 

Theorem is proved.

APPLICATION TO THE PROBLEM (4)

We apply our theorem on discrete analogues for (4), which are defined by (5), (5) and (7). First we summarize some properties of the matrices $A(x)$ and $B = \text{diag}(0,1,\ldots,1,0)$ as defined by the schemes (5)-(7). The functions $q(t)$ and $f(t,u)$ are assumed to satisfy the conditions

$$|q(t)| \leq M, t \in \mathbb{R}, \quad |f(t,u)| \leq M_0, \quad \frac{\partial f}{\partial x}(t,u) \leq M_1, (t,u) \in \mathbb{R} \times \mathbb{R}$$

$$-\lambda \leq q(t) \leq \lambda, \quad -\lambda \leq \lambda q(t) \leq h^{-2} q_i, \quad t \in \mathbb{R},$$

for some real $\mu$, where $\lambda$ and $q_i$ depend upon the scheme as follows. Let $A_0$ is the matrix $A(x)$ for $q(t) \equiv 0$, $t \in \mathbb{R}$. Then $A(x) = A_0 + Q(x)$, where $A_0 \in \mathbb{R}^{n,n}$ is independent of $x$ and $Q(x)$ is defined by (5), $q(x) \equiv 0$. The matrix $A_0$ is inverse-monotone, [1], and there exists the smallest positive eigenvalue $\lambda$ to the eigenvalue problem $A_0 x = \lambda B x$. From [1] we have that
is inverse-monotone for any diagonal matrix $D$ whose diagonal elements are all in $(-\lambda, h^{-2}q_+^-)$. Next table shows a type of this kind where $q_+^- = \infty$ means that $(-\lambda, h^{-2}q_+^-) = (-\lambda, \infty)$.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_+^-$</td>
<td>$\infty$</td>
<td>$3$</td>
<td>$1/18$</td>
</tr>
</tbody>
</table>

Now, from (8) follows that $A(x)$ is inverse-monotone and

$$0 \leq A(x)^{-1} \leq (A_0 + uB)^{-1} \quad \text{for } x \in \mathbb{R}^n,$$

$$\|A(x)^{-1}\|_\infty \leq \alpha,$$

here

$$\alpha = \| (A_0 + uB)^{-1} \|_\infty$$

depends on the scheme. Since

$$\max_{1 \leq i \leq n} \sum_{k=1}^{n} \left| \frac{\partial a_{ik}}{\partial x_j} (x) \right| = \max_{1 \leq i \leq n} \left| \frac{\partial a_{ii}}{\partial x_i} (x) \right| = \max_{1 \leq i \leq n} \left| q^- (x_i) \right| \leq M,$$

the assumptions of our theorem are satisfied. Then for any of the schemes (5)-(7) there exists the unique solution and the sequence $x^0, x^1, x^2, \ldots$ 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 $-\lambda < u$ is satisfied if $u > -\lambda(h)$. Now we can give easily computed estimates for $h > 0$ such that the condition $u > -\lambda(h)$ 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 $\lambda(h) > 8$ for $h \in [0, 0.5]$. So, if $\mu > -8$, we have $u > -\lambda(h)$ for all $h \in [0, 0.5]$. The restriction on $h$ are described by

$$-\lambda(h) < u < h^{-2}q_+^-.$$ The computable bounds of $\|(A_0 + uB)^{-1}\|_\infty$ are given in $|1|$. So we have

$$\| (A_0 + uB)^{-1} \| \leq \begin{cases} 1/8 & \text{for } \mu = 0 \\ d - \mu^{-1} & \text{for } h^2 < \mu < 0, \end{cases}$$

where
\[ d = (\mu \cos(0.5\beta))^{-1}, \quad 0 < \beta < \pi, \quad \mu h^2 = 2(1-\cos\beta h). \]

Now we can easily see that the condition \( \alpha_2 M_0 + \alpha M_1 < 1 \) reduces to the case when \( \alpha = \frac{1}{8} \) for \( \mu = 0 \) and \( \alpha = d - \mu^{-1} \) for \( -\pi^2 < \mu < 0 \).

REFERENCES


ONE WAY OF DISCRETIZATION OF CHAPLYGIN'S METHOD

Dušan D. Tošić

ABSTRACT:

Chaplygin's method (described in [1] and [3]) is an analytic and iterative method for two-sided approximation to the solution of ordinary differential equations. This method is difficult for practical applications in analytic form. In this work one way of discretization of Chaplygin's method is proposed. Chaplygin's approximations are calculated by using the interpolation and the numerical integration. Some examples with the cubic spline interpolation and Simpson's rule are presented.

1. INTRODUCTION

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

\[ G_h = \{ x_i | x_i = a + ih, \ i = 0, \ldots, n, \ b-a = nh \}. \]

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 (see [1]):

(2) \[ \max_{x \in [a,b]} |u_k(x) - v_k(x)| < \frac{c}{2^{2k}}, \quad (0 \in \mathbb{R}^+). \]

In [4] and [6] some shortcomings and problems related to Chaplin'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:

\[ p_k(x) = \frac{\partial f(x, u_k(x))}{\partial y} \]

\[ q_k(x) = \frac{f(x, v_k(x)) - f(x, u_k(x))}{v_k(x) - u_k(x)} \]

and

\[ I(a(x), b(x)) = \exp(-\int_{x_0}^{x} a(t) dt)(v_0 + \int_{x_0}^{x} (f(t, b(t)) + a(t)b(t)) \exp(\int_{x_0}^{x} \alpha(z) dz) dt), \]

then we have:

\[ u_{k+1}(x) = I(p_k(x), u_k(x)) \]

\[ v_{k+1}(x) = I(q_k(x), v_k(x)) \]

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 \( P_k(x) \) 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 \( O(h^6) \), as \( h \to 0 \),

(b) each integral in (5) is calculated with the truncation error not greater than \( O(h^8) \), 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:

\[ \int_{x_0}^{x_i} a(t) dt = S_1^h + R_1 + R_2 \]
\( R_1 \) is the roundoff error and \( R_2 \) is the truncation error of numerical integration. According to (b) we have:

\[
\int_{x_0}^{x_i} a(x) \, dx = S^h_1 + 0(h^r)
\]

where \( r = \min(e, s) \). From (9) we get:

\[
\exp(- \int_{x_0}^{x_i} a(x) \, dx) = C^h + 0(h^r), \quad \text{as } h \to 0.
\]

Let be:

\[
(11) \quad g(x) = (f(x, b(x)) + a(x)b(x)) \exp(\int_{x_0}^{x} a(t) \, dt).
\]

For the expression (5) we may write:

\[
(12) \quad I(a(x), b(x)) = \exp(- \int_{x_0}^{x} a(t) \, dt) \left( y_0 + \int_{x_0}^{x} g(t) \, dt \right).
\]

By using (10) and (11) (for \( x_i \in G_h \)) we get:

\[
g(x_j) = (f(x_j, b(x_j)) + a(x_j)b(x_j))(C^h_j + 0(h^r))
\]

and according to (a):

\[
g(x_j) = g^h_j + 0(h^r).
\]

Now we obtain (like as in (9)):

\[
(13) \quad \int_{x_0}^{x_i} g(x) \, dx = S^h_2 + R_1 + R_2
\]

\[
= S^h_2 + 0(h^r).
\]

Finally from (12), (10) and (13) it follows that:

\[
(14) \quad I(a(x_i), b(x_i)) = C^h y_0 + C^h S^h_2 + 0(h^r)
\]

\[
= I^h + 0(h^r)
\]

for \( x_i \in G_h \), as \( h \to 0 \) and the theorem is proved.

Denoting by \( u^h_{ki} \) and \( v^h_{ki} \) discrete Chaplygin's approximations in the point \( x_i \in G_h \), from (6), (7) and (14) we have:

\[
(15) \quad u_k(x_i) = u^h_{ki} + 0(h^r)
\]

\[
(16) \quad v_k(x_i) = v^h_{ki} + 0(h^r)
\]

for k-th iteration. By using (2), (15) and (16) we make the following estimation:

\[
|u_k(x_i) - v_k(x_i)| < |u^h_{ki} - v^h_{ki}| + 0(h^r) + \frac{c}{2^{2k}}
\]
The error estimation for the solution $y(x)$ of (1) is based on the inequalities:

$$(17) \quad u_k(x) < y(x) < v_k(x)$$

where $x \in [a, b]$ and $k = 0, 1, \ldots$.

3. NUMERICAL EXAMPLES

In the following examples we apply the results of the previous section. As polynomial $P_{uk}(x)$ (i.e. $P_{vk}(x)$) 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

$$(18) \quad y' = y^2 - y \sin x + \cos x, \quad y(0) = 0,$$

by the previous method, using a step length $h = 0.1$ on the interval $[0, 1]$. As the initial approximations we choose

$$u_0(x_i) = \sin x_i - 0.01i \quad (i = 0, 1, \ldots, 10)$$

$$v_0(x_i) = \sin x_i + 0.01i.$$

(Similar results are obtained for $u_0(x_i) = \sin x_i - 0.1i$ and $v_0(x_i) = \sin x_i + 0.1i$, $i = 0, 1, \ldots, 10$.) The results are presented in the table 1.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$u_{1i}^h$</th>
<th>$v_{1i}^h$</th>
<th>$u_{2i}^h$</th>
<th>$v_{2i}^h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.09983016</td>
<td>0.09983686</td>
<td>0.09983352</td>
<td>0.09983291</td>
</tr>
<tr>
<td>0.2</td>
<td>0.19864265</td>
<td>0.19869461</td>
<td>0.19866954</td>
<td>0.1986895</td>
</tr>
<tr>
<td>0.3</td>
<td>0.29542913</td>
<td>0.29561211</td>
<td>0.29552049</td>
<td>0.29551989</td>
</tr>
<tr>
<td>0.4</td>
<td>0.38919984</td>
<td>0.38963886</td>
<td>0.38941868</td>
<td>0.38941806</td>
</tr>
<tr>
<td>0.5</td>
<td>0.47899265</td>
<td>0.4798601</td>
<td>0.47963897</td>
<td>0.47942527</td>
</tr>
<tr>
<td>0.6</td>
<td>0.56388187</td>
<td>0.56541504</td>
<td>0.56462478</td>
<td>0.5646220</td>
</tr>
<tr>
<td>0.7</td>
<td>0.64298689</td>
<td>0.64547439</td>
<td>0.64421786</td>
<td>0.64421744</td>
</tr>
<tr>
<td>0.8</td>
<td>0.71547964</td>
<td>0.71928375</td>
<td>0.71735594</td>
<td>0.71735601</td>
</tr>
<tr>
<td>0.9</td>
<td>0.78059514</td>
<td>0.78615534</td>
<td>0.78332613</td>
<td>0.78332729</td>
</tr>
<tr>
<td>1.0</td>
<td>0.83762731</td>
<td>0.84547896</td>
<td>0.84146881</td>
<td>0.84147252</td>
</tr>
</tbody>
</table>

Table 1

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. Consider the initial value problem:

$$y' = x^2 + y^2 - 32.1, \quad y(0) = 0.1,$$

in the interval $[0,1]$. Let be:

$$G_h = \{ x_i \mid x_i = 0.2i, \ i=0,\ldots,5 \}.$$

As the initial approximations we take:

$$u_0(x_i) = -2 \quad \text{and} \quad v_0(x_i) = -6. \quad (i=1,\ldots,5)$$

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>
<thead>
<tr>
<th>$x_i$</th>
<th>$u_{5i}$</th>
<th>$v_{5i}$</th>
<th>Runge-Kutta</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>-4.6109031</td>
<td>-4.6105720</td>
<td>-4.39460...</td>
</tr>
<tr>
<td>0.4</td>
<td>-5.5755214</td>
<td>-5.5754871</td>
<td>-5.11619...</td>
</tr>
<tr>
<td>0.6</td>
<td>-5.6738443</td>
<td>-5.6738269</td>
<td>-5.39515...</td>
</tr>
<tr>
<td>0.8</td>
<td>-5.6625736</td>
<td>-5.6625541</td>
<td>-5.50713...</td>
</tr>
<tr>
<td>1.0</td>
<td>-5.6339439</td>
<td>-5.6339192</td>
<td>-5.53980...</td>
</tr>
</tbody>
</table>

The numerical results presented in this paper (and a lot of others numerical results) are obtained on the microcomputer COMODORE 64.

REFERENCES

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:

\[\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad (x, t) \in Q = (0,1) \times (0,T),\]

(1) \[u(0,t) = u(1,t) = 0, \quad t \in (0,T),\]

\[u(x,0) = \frac{\partial u(x,0)}{\partial t} = 0, \quad x \in [0,1].\]

Throught the note we will assume that the generalized solution of (1) belongs to a Sobolev-Slobodetsky space \(W^s_2(Q)\), \(1 \leq s \leq 4\), [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 \(W^s_2(Q)\).

Pick a nonnegative integer \(n\) and let \(h = 1/n\). We define a uniform grid \(\omega_n\) with the step \(h\) over \((0,1)\). In the same way we define a uniform grid \(\omega_T\) with the step...
\( \tau = T/(m+0.5) \) over \((-0.5, T]\) and put \( Q_{h \tau} = \omega_h \times \omega_\tau \).

We will assume that \( c_1 h \leq \tau \leq c_2 h \), \( c_1, c_2 = \text{const} > 0 \).

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 \( v_x, v_{xx}, v_t \)

and \( v_x^\prime \) 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 \( \omega_h \), and which are equal to zero over the boundary knots the operator:

\[ \Lambda v = -v_{xx} \]

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:

\[ \| v \|_{2,\infty, h}^{(1)} = \max_k (\| v_t^k \|_h + 0.5 \| v_x^k + v_{xx}^{k+1} \|_h) \]

\[ \| v \|_{2,\infty, h}^{(0)} = \max_k (\| v_t^k \|_{\Lambda^{-1}} + 0.5 \| v^k + v_{xx}^{k+1} \|_h) \].

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:

\[ T^0 g(x) = g(x), \quad T^k g(x) = T(T^{k-1} g(x)), \quad k = 1, 2, \ldots \]

and \( T^0 g(x) = g(x) \), \( T^0 g(x) = T(T^{k-1} g(x)) \), \( k = 1, 2, \ldots \)

By \( T^k \) we will denote the product of \( T^k \) over \( x \), and \( T^2 \) over \( t \).

Then we will approximate (1) by a weighted difference scheme (a = const > 0):

\[ v^j_{tt} = a v^{j+1}_x x + (1 - 2a) v^j_x x + a v^{j-1}_x x + T^2 v^j, \]

\[ v^j_t = 0 \quad \text{for} \quad x = 0 \quad \text{and} \quad x = 1, \]

\[ v^0 = 1 \]

When \( u \in W^{s}(Q), \quad 2 \leq s \leq 4 \), we will denote \( z = u - v \). Function \( z \) is well defined over \( Q_{h \tau} \) satisfying:
\[ z_{tt}^j = a z_{xx}^{j+1} + (1 - 2a) z_{xx}^j + a z_{xx}^{j-1} + \phi_{xx}^j + \psi_{xx}^j, \]
\[ z_0^j = 0 \quad \text{for} \quad x = 0 \quad \text{and} \quad x = 1, \]
\[ z_0^1 = u(x, -0.5 \tau), \]
\[ z_1^1 = u(x, 0.5 \tau), \]

where
\[ \phi^j = T_{0, 2}^j u^j - a u_{j+1}^j + (1 - 2a) u_j^j + a u_{j-1}^j, \]
\[ \psi^j = u_j^j - T_{2, 0}^j u_j^j. \]

When \( u \in W_2^s(Q), 1 \leq s \leq 3 \), we will denote \( z = T_{2, 0}^j u - v \). 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 \( z \) is defined over the grid \( Q_{h \tau} \) satisfying:
\[ z_{tt}^j = a z_{xx}^{j+1} + (1 - 2a) z_{xx}^j + a z_{xx}^{j-1} + \phi_{xx}^j, \]
\[ z_0^j = 0 \quad \text{for} \quad x = 0 \quad \text{and} \quad x = 1, \]
\[ z_0^1 = T_{2, 0}^j u(x, -0.5 \tau), \]
\[ z_1^1 = T_{2, 0}^j u(x, 0.5 \tau), \]

where
\[ \phi^j = T_{0, 2}^j u^j - T_{2, 0}^j [a u_{j+1}^j + (1 - 2a) u_j^j + a u_{j-1}^j]. \]

Using the method of energy inequalities [6] one can prove the following a priori estimates:

\[ (3) \| z \|_{2, 1, \infty, h} \leq C [\| z_0^0 \|_{h} + \| z_0^1 + z_1^1 \|_{h} + \tau \sum_{j=1}^{m} (\| \phi_{xx}^j \|_{h} + \| \psi_{xx}^j \|_{h})] \]

\[ (4) \| \bar{z} \|_{2, 1, \infty, h} \leq C \left( \| z_0^0 \|_{L_1 - 1} + \| z_0^1 + z_1^1 \|_{h} + \tau \sum_{j=1}^{m} \| \phi_{xx}^j \|_{h} \right). \]

The convergence rate estimates in this note are based on the following generalization of the Bramble-Hilbert lemma.

**Lemma:** Let \( [s] \) be a nonnegative integer, \( [s] < s \leq [s] + 1 \) and let \( P_{[s]} \) be the set of polynomials of degree \( \leq [s] \). If \( \eta = \eta(u) \) is a bounded, linear functional over \( W_2^s(Q) \) such that \( P_{[s]} \subset \text{Kernel}(\eta) \), then for
For every $u \in W^2_s(Q)$ the following inequality is valid:
\[ |\eta(u)| \leq C |u|_{s,Q}, \quad C = C(Q,s) = \text{const}. \]

The proof of the lemma follows from the Dupont-Scott theorem [1].

Functionals $\varphi_{xX}$, $\psi_{tT}$, $z^0_t$ and $(z^0_t + z^1_t)$ are bounded and linear over $W^2_s(Q)$ for $s > 2$ while $P_3 \subset \text{Kernel}(\varphi_{xX})$, $P_3 \subset \text{Kernel}(\psi_{tT})$, $P_2 \subset \text{Kernel}(z^0_t)$, $P_2 \subset \text{Kernel}(z^0_t + z^1_t)$. Using the lemma one obtains the following estimates:

\[ \sum_{j=1}^m (\|\varphi_{xX}^j\|_{h} + \|\psi_{tT}^j\|_{h}) \leq C h^{s-2} |u|_{s,Q}, \quad 2 < s \leq 4 \]  
(5)

\[ \|z^0_t\|_h + \|z^0_t + z^1_t\|_h \leq C h^{s-1.5} |u|_{s,Q}, \quad 2 < s \leq 3, \]

where $Q_\tau = (0,1) \times (-0.5 \tau, 0.5 \tau)$. Using (3), (5), (6) and (7)

\[ |u|_{k,Q_\tau} \leq C \tau^{k} \|w(h,a)\|_{k+a,Q_\tau}, \quad k = 0,1,2, \ldots, \]

where

\[ F(h,a) = \begin{cases} h^a & \text{if } 0 \leq a < 0.5, \\ h^{0.5} \ln h & \text{if } a = 0.5, \\ h^{0.5} & \text{if } 0.5 < a \leq 1 \end{cases} \]

(see [5]) one obtains the following convergence rate estimate for the difference scheme (2):

\[ \|z\|_{2,\infty,h} \leq C h^{s-2} |u|_{s,Q}, \quad 2 < s \leq 4. \]

Similarly from the lemma and from (7) one obtains:

\[ \|z^0 + z^1\|_h + \tau \sum_{j=1}^m \|\varphi_{xX}^j\|_h \leq C h^{s-1} |u|_{s,Q}, \quad 1 < s \leq 3, \]

(8)

\[ \|z^0_t\|_{h^-1} \leq C h^{s-1.5} |w|_{s,Q_\tau}, \quad 2 < s \leq 3, \]

(9)

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
\[ \| z \|_{2,\infty,h} \leq c h^{s-1} (\| u \|_{s,Q} + f \|_{p(s),s-1,Q}) , \]

where \( p(s) = \max\{0, s - 2\} \) and \( f \|_{p,q,Q} \)

the norm of the anisotropic Sobolev-Slobodetsky space

\[ \mathcal{W}^{p,q}(Q) = L^2(0,T; \mathcal{W}^{2}(0,1)) \cap \mathcal{W}^{2}(0,T; L^2(0,1)) \] (see [4]).

REFERENCES


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 converges to the original solution. Also, we construct a minimizing sequence which converges to the optimal control.

We shall consider the following optimal control problem:

(1) \[ J(v) = \int_{Q_T} f(u(x,t), v_0 (x), v_1 (x)) \, dx \, dt \rightarrow \inf_U \]

(2) \[ u_t = \Delta u + v_0 , \quad (x,t) \in Q_T = (0,1)^2 \times (0,T] \]

(3) \[ u(x,t) = 0 , \quad (x,t) \in \Omega \times (0,T] , \quad \Omega = (0,1)^2 \]

(4) \[ u(x,0) = v_1 (x) , \quad x \in \Omega \]

It is well known \[5\] that if \( v_0, v_1 \in H^{2r}, H^{2r+1}(Q_T) \) than exists a unique solution of (2), (3), (4) for \( r \geq 0 \) and the following estimate
valid:

\[ \|u\|_{H^{2r+2},r+1}(\Omega) \leq C \|v\|_{H^{2r},r}(\Omega) + \|v_{1}\|_{H^{2r+1}(\Omega)}. \]

We shall denote \( v=(v_{0},v_{1}) \) and \( U=\left\{ v \in X_{r} \mid H^{2r},r(\Omega) \right\} \times H^{2r+1}(\Omega) : \|v\|_{X_{r}} \leq R \). Throughout the note we shall assume that \( f \) is a convex function and

\[ |f(a_{0},a_{1},a_{2})-f(b_{0},b_{1},b_{2})| \leq g(a_{0},b_{0}) \sum_{i=0}^{2}|a_{i}-b_{i}| \]

where \( g(a_{0},b_{0}) \) 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 \( \Omega_{h} \) be a uniform grid with the step \( h=1/n \) over \( \Omega \) and let \( \Omega_{h} \) be a uniform grid over \((0,T]\) with the step \( \tau=T/m \). In this note we shall assume that constants \( c_{1},c_{2} \geq 0 \) exists such that \( a_{1}h^{2} \leq \tau \leq c_{2}h^{2} \).

In the set of discrete functions over \( Q_{h}\tau = \Omega_{h} \times (0,\tau) \), we shall introduce the following norms:

\[ \|y\|_{0,h} = (\sum_{j=0}^{m-1} 0.25 \|y_{j+1}^{1} + y_{j}^{1}\|_{h}^{2})^{1/2} \]
\[ \|y\|_{1,h} = (\sum_{j=0}^{m-1} \sum_{i=1}^{2} 0.25 \|(y_{j+1}^{1} + y_{j}^{1})_{x_{i}}\|_{h}^{2})^{1/2} \]
\[ \|y\|_{2,h} = (\sum_{j=0}^{m-1} \|y_{j}^{2}\|_{h}^{2} + 0.25 \sum_{i=1}^{2} \|(y_{j+1}^{1} + y_{j}^{1})_{x_{i}}\|_{h}^{2})^{1/2} \]
\[ + 0.25 \|(y_{j+1}^{1} + y_{j}^{1})_{x_{1}x_{2}}\|_{h}^{2})^{1/2} \]
Let $T$ be a Steklov mollifier \cite{4}, \cite{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 \cite{4}, \cite{6}:

\begin{align}
L_1 y = (y^{j+0.5} - y^{j})/0.5 - y^{j+0.5}_{x_1 x_1} - y^{j}_{x_2 x_2} = \phi^{j+0.5}_0 \\
L_2 y = (y^{j+1} - y^{j+0.5})/0.5 - y^{j+0.5}_{x_1 x_1} - y^{j+1}_{x_2 x_2} = \phi^{j+0.5}_0 \\
y^0 = \phi_1
\end{align}

where $y^{j+0.5}$ denotes the value of $y$ on the auxiliary time slice $t=(j+0.5) \tau$, $\phi_0^{j+0.5} = (T^{2,2,1} v_0)^{(j+0.5)}$ and $\phi_1 = T^{2,2} v_1$.

Using (5) and the discrete solution estimates \cite{4} from Dupont-Scott theorem \cite{2} follows:

\begin{align}
\|z\|_{2,1,h} & \leq C h^{2r} \|v\|_{\chi_r} & 0 < r \leq 1 \\
\|z\|_{1,0,h} & \leq C h^{2r+1} \|v\|_{\chi_r} & 0 < r \leq 0.5 \\
\|z\|_{0,0,h} & \leq C h^{2} \|v\|_{\chi_0} & \cdot
\end{align}

The cost function $J(u)$ will be approximated by the following equation:

$$I_n(y) = \sum_{y \in \zeta} \chi h^2 \zeta f(y, y_{x_1}, y_{x_2})$$

Now we can formulate the sequence of finite dimensional problems:
\[(13) \quad I_n \rightarrow \inf_{w \in WP^{p,q,s}} \]

\[(14) \quad L_1 y = w_0 \quad , \quad L_2 y = w_0 \quad , \quad y^0 = w_1\]

where the operators $L_1, L_2$ are defined by (7), (8) and the set $WP^{p,q,s}_n = \left\{ w = (w_0, w_1) \in Y_n \in H^{p,q}_n, H^{s}_n : \|w\|_{Y_n} \leq R \right\}$.

We denoted by $H^{p,q}_n, H^{s}_n$ the discrete analogues of the Sobolev spaces $H^{p,q}(QT), H^{s}(\Omega)$. 

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 $Q_n : X_n \rightarrow Y_n$ and $P_n : Y_n \rightarrow X_n$ by

\[(15) \quad Q_n(v) = (Q^0_n(v), Q^1_n(v_1)) = (T^{2,2}_n v_0, T^{2,2}_n v_1)\]

\[(16) \quad P_n(w) = \left( \frac{w_0}{w_1}, w_1 \right)\]

where $w_0 \in H^{2,1}(QT)$ and $w_1 \in H^{3}(\Omega)$ are the interpolants defined in [1].

Theorem 1. If the above assumptions are valid then

\[\lim_{n \rightarrow \infty} I^*_n = J^*_u = \inf_{u \in U} J(u) \quad \text{and for} \quad 0 < r < 0.5\]

\[(17) \quad \|I^*_n - J^*_u\| \leq C h^{2r+1} .\]

Proof. Using 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:

\[T_n(w) = I_n(w) + \lambda \|w\|_{Y_n}^2 .\]
Let \( w_n^* \) be a sequence of discrete controls such that:
\[
\inf_{w_n} T_n(w) \leq T_n(w_n^*) \leq T_n^* + M_n.
\]
Sequences \( \lambda_n \) and \( M_n \) are positive and \( \lim_{n \to \infty} \lambda_n = \lim_{n \to \infty} M_n = 0 \).

Theorem 2. If the theorem 1. is valid and if
\[
\lim_{n \to \infty} (h_n^{2r+1} + M_n) / \lambda_n = 0 \quad \text{for } 0 < r \leq 0.5
\]
then
\[
\lim_{n \to \infty} J(P_n(w_n^*)) = J^* \quad \text{and} \quad \lim_{n \to \infty} \| P_n(w_n^*) - v^* \|_{X_r} = 0
\]
where \( J(v^*) = J^* \).

Proof. Using the same technique as in [7] from the estimate (11) follows the theorem.

Remark. If the cost functional is of the form
\[
J(u) = \int_Q f(u, u_x, u_{xx}, u_t) dx dt
\]
than in (17), (18) \( 2r+1 \) must be replaced by \( 2r \). If the cost functional is
\[
J(u) = \int_Q f(u) dx dt
\]
and if \( r=0 \) than in (17), (13) \( 2r+1 \) must be replaced by \( 2 \).
REFERENCES


Department of Mathematics
Faculty of Sciences
11 000 Belgrade, Studentski trg 16
Yugoslavia
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 $\pi$ are obtained. From their asymptotic expansions it can be seen that the orders of errors are $O(h^{3/2}/|z|^{1/v}), v=3/2, 2, 3, 4$, and that in some cases the accuracy can be improved by the appropriate choice of the grid parameters.

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

$$\Omega = \{(x,y) \mid x=mh, y=nh', m,n \in \mathbb{Z}, h'=eh, h,e>0\}$$

following problems: to find the continuous function $v$, not identically equal to zero, harmonic in the corner $0<\phi<\pi$, $0<r<\infty ((r,\phi)$ polar coordinates), $v=3/2, 2, 3, 4$, equal to zero on the positive part of the x-axis and on the line $v\pi$, and which does not grow too rapidly at the infinity, i.e.

$$\lim_{r \to \infty} \frac{v}{r^{3/2}} = 0.$$

Solutions of these problems, for corresponding $v$, are (from [2])

$$v = C \operatorname{Im} z^{1/v} = C r^{1/v} \sin \frac{\phi}{v}, \quad C = \text{const}, \quad z = r e^{i\phi},$$

so, their first derivatives have integrable singularities at the origin. For $v=3$ the initial problem can be replaced by the equivalent one,
defined in the corner \( 0 \leq \varphi \leq 3\pi/2, 0 \leq r < \infty \) with the boundary condition \( \frac{\partial v}{\partial x} = 0 \) on the line \( \varphi = 3\pi/2 \). Similarly, for \( v=4 \) 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 \([3]\))

\[
\Lambda_{\alpha} u \equiv u_{xx} + u_{yy} - \alpha h^2 \frac{1+\theta^2}{2} u_{xxyy}, \quad \alpha > -1/2,
\]

and difference operator families in order to approximate the boundary conditions of the second type

\[
\lambda_{3/2} u \equiv u_x - \frac{h}{2} u_{yy} - \alpha h^2 \frac{1+\theta^2}{2} u_{xxyy} \quad \text{on the line } 3\pi/2,
\]

and

\[
\lambda_{1/2} u \equiv u_y - \frac{\theta h}{2} u_{xx} - \alpha h^2 \frac{1+\theta^2}{2} u_{xxyy} \quad \text{on the line } 2\pi.
\]

(for \( \alpha > -1/2 \) \( \Lambda_{\alpha} \) is the elliptic operator, \([4]\)). Let us denote

\[
\Omega_{\varphi} = \{(x,y) \mid (x,y) \in \Omega, 0 < \varphi < \pi/2, 0 < r < \infty\},
\]

\[
\Gamma_{\varphi} = \{(x,y) \mid (x,y) \in \Omega, \varphi = \pi/2, 0 < r < \infty\},
\]

\[
\Gamma_0 = \{(x,y) \mid (x,y) \in \Omega, x>0, y<0\}, \quad \Gamma_0 = \Gamma_0 \cup \{(0,0)\},
\]

\[
\Gamma_2 = \{(x,y) \mid (x,y) \in \Omega, x>0, y=0\}.
\]

Solutions of the following problems will be determined:

**PROBLEM 1.** (\( v=3/2 \)) \( \Lambda_{\alpha} u = 0, \quad (x,y) \in \Omega_{3/2} \), \( \lim_{r \to 0} u/r^{3/2} = 0 \),

\( u = 0, \quad (x,y) \in \Gamma_0 \), \( u = 0, \quad (x,y) \in \Gamma_{3/2} \), \( u(-h,0) = A \).

**PROBLEM 2.** (\( v=2 \)) \( \Lambda_{\alpha} u = 0, \quad (x,y) \in \Omega_2 \), \( \lim_{r \to 0} u/r = 0 \),

\( u = 0, \quad (x,y) \in \Gamma_0 \), \( u = 0, \quad (x,y) \in \Gamma_2 \), \( u(-h,0) = A \).

**PROBLEM 3.** (\( v=3 \)) \( \Lambda_{\alpha} u = 0, \quad (x,y) \in \Omega_3 \), \( \lim_{r \to 0} u/r^{2/3} = 0 \),

\( u = 0, \quad (x,y) \in \Gamma_0 \), \( \lambda_{3/2} u = 0, \quad (x,y) \in \Gamma_{3/2} \), \( u(-h,0) = A \).

**PROBLEM 4.** (\( v=4 \)) \( \Lambda_{\alpha} u = 0, \quad (x,y) \in \Omega_4 \), \( \lim_{r \to 0} u/r^{1/2} = 0 \),

\( u = 0, \quad (x,y) \in \Gamma_0 \), \( \lambda_2 u = 0, \quad (x,y) \in \Gamma_2 \), \( u(-h,0) = A \).

**THEOREM 1.** The solutions of the problems 1-4, for corresponding \( v \), are
\[ u(x_m, y_n) = \begin{cases} 
\sum_{j=0}^{\infty} \frac{A}{2\pi j} \int_{0}^{2\pi} \frac{e^{ij}e^{im\xi}}{(1 - e^{ij})^{1/2}} \frac{F\left(\frac{e^{ij} - z_j}{1 - z_j e^{ij}}\right)}{d\xi}, & n > 0, \\
u(x_m, y_n) + 2\cos\frac{\pi}{\nu} u(x_m, y_n), & n < 0, 
\end{cases} \]

where \( m = 0, \pm 1, \pm 2, \ldots, \)

and

\[ F(z) = \frac{1}{2} \text{F}_1\left(\begin{array}{c} \frac{1}{2} - 1, 1 - \frac{1}{\nu}; \frac{3}{2}; z \end{array}\right) \]

a hypergeometric function,

\[ z_1 = \left( \sqrt{((1 + \theta^2)(1 + 2\alpha) - 1)/\sqrt{(1 + \theta^2)(1 + 2\alpha) + 1}} \right), \]

and

\[ q(\xi) = \frac{\sqrt{\cos^2 \frac{\xi}{2} + (1 + \theta^2)(1 + 2\alpha) \sin^2 \frac{\xi}{2}} - \theta \sin \frac{\xi}{2}}{\sqrt{\cos^2 \frac{\xi}{2} + (1 + \theta^2)(1 + 2\alpha) \sin^2 \frac{\xi}{2}} + \theta \sin \frac{\xi}{2}}. \]

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 \( 3\pi/2 < \phi < 2\pi \) 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

\[ u(x_m, y_n) = \frac{1}{2\pi} \sum_{j=0}^{\infty} \int_{0}^{2\pi} q_j(\xi) [q_n e^{im\xi} - 1] d\xi, \quad m = 0, \pm 1, \pm 2, \ldots, \]

(5) \[ u(x_m, y_n) = \frac{1}{2\pi} \sum_{j=0}^{\infty} \int_{0}^{2\pi} q_j(\xi) [q_n e^{im\xi} - 1] d\xi, \quad m = 0, \pm 1, \pm 2, \ldots, \]

(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 analytic functions on the unit circle in the complex plane. This problem can be reduced to the singular integral equation

\[ \int_{0}^{1} \frac{\psi(t)}{t - x} dt + \cos \frac{\pi}{\nu} \int_{0}^{1} \frac{\psi(t)}{1 - x t} dt = 0, \]

and its solution is

\[ \psi(x) = C x^{-1/3} (1 - x)^{-\nu/2} F\left(\begin{array}{c} 1/3 - 1, 1 - \frac{1}{\nu}; \frac{3}{2}; x \end{array}\right), \quad C = \text{const.} \]
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 \(\nu\), are

\[
 u(x_m, y_n) = - \frac{A \nu (1+z_1)^{1/\nu}}{\eta(1-z_1) F(-z_1) h^{1/\nu}} \text{Im} \int \frac{(1-z)^{-i\nu} d[F(z) q^n(z) (z+z_1)^m]}{|z|^2},
\]

where \(F(z)\) and \(z_1\) are given by (2) and (3), and \(z = x_m + i y_n = h\beta\).

For the proof of this theorem, we obtain more convenient expression of the solution (1) with substitutions \(z = e^{it}\) and \(z = (z_1 - z)/(1 - z_1)\)

\[
 u(x_m, y_n) = - \frac{A \nu (1+z_1)^{1/\nu}}{\eta(1-z_1) F(-z_1) h^{1/\nu}} \text{Im} \int \frac{(1-z)^{-i\nu} d[F(z) q^n(z) (z+z_1)^m]}{|z=1|^2},
\]

where \(q(z) = q(\xi) = q(\xi)\) and \(F(z) = F([e^{it} - z_1]/(1 - z_1 e^{it}))\). 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 distorted by estimating the integral for separate parts of the contour, we obtain

\[
 u(x_m, y_n) = - \frac{A \nu (1+z)^{1/\nu}}{\eta(1-z_1) F(-z_1) h^{1/\nu}} \int \frac{(1-z)^{-i\nu} d[F(z) e^{-\nu(z)\beta}] + O(\beta^{-N})}{C},
\]

for any \(N>0\). \(v(z) = w(z)\beta\) is defined by the expression

\[
 e^{-v(z)} = q^m(z) \left( \frac{z+z_1}{1+z_1} \right)^m.
\]

The curve \(C\) is \(C = \{z \mid \text{Im} v(z) = 0\}\) for \(z \in \{z \mid |z - 1| < \delta, \text{Im} z > 0\}\), 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 analytically. With regard to the features of the function \(w\), its inverse function \(z = z(w)\) for \(|w| < \delta\) exists. So, in the integral (8) we can use the new argument of
We suppose some series representations as functions of \( w \) or \( (1-z)^{-1/\nu} \) and \( F(z) \), and after some estimates we have

\[
 u(x_m,y_n) \approx \frac{A_{\nu}(1+z_1)^{1/\nu}}{F(1-z_1)^{1/\nu}} \text{Im} \sum_{k=0}^{\infty} \left\{ p_k(v) r(k-1)\nu \beta^{1/\nu} - p_k(-v) r(k+1)\nu \beta^{-1/\nu} \right\} b_k,
\]

where the coefficients \( p_k(v) \) are determined by the series expansion of the function

\[
 p(w, \nu) = w^{1+\nu} \frac{1}{4(1-z)^{\nu/2}} \left( \frac{1+z}{1-z} \right)^{1/\nu} \frac{d z}{dz} = \sum_{k=0}^{\infty} p_k w^k.
\]

If we express \( z \) by \( w \) and obtain

\[
 p_0(v) = -2^{-z(1-\nu)} \frac{1-z_1}{1+z_1}^{1/\nu},
\]

\[
 p_2(v) = -2^{-z(1-\nu)} \frac{1-z_1}{1+z_1}^{1/\nu} \left( 2 - \frac{1}{\nu} \right) \frac{1}{3} [(1+\theta^2)(1+6\alpha) \beta \beta + 2(\theta^2-1)],
\]

\[
p_{2k+1}(v) = 0, \ k=0,1,2,\ldots.
\]

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

\[
 \Omega_\epsilon = \{(x,y) \mid x = (m+\epsilon)h, y = nh', m, n \in \mathbb{Z}, h' = \epsilon h, h, \epsilon > 0\}.
\]

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 \( \Omega_\epsilon \), is

\[
 u(x_m,y_n) = \frac{2A}{\sqrt{[F(1-z_1)h]}} \text{Im} \left\{ z^{1/2} - \frac{1}{2} \left( e - \sqrt{[1+\theta^2][1+2\alpha]} \right) \frac{h}{z^{1/2}} \right\} - \frac{1}{2} \left[ \frac{\beta \epsilon^3 - 4\epsilon \sqrt{[1+\theta^2][1+2\alpha]} \epsilon \frac{\beta}{2\beta^2 + \theta^2 - 1]}{\left[ \frac{h}{z^{1/2}} \right]} + O\left( \frac{h^3}{z^{1/2}} \right) \right],
\]

where \( z = h(m+\epsilon+in) = h \beta_\epsilon \).

For the proof of the theorem we put \( \nu = 2 \) and replace \( \beta \) with \( \beta_\epsilon \) in the expression (6), where \( \beta = \beta_\epsilon (1-\epsilon/\beta_\epsilon) \).
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 choice of α, θ and ε such as ε = \sqrt{(1+θ^2)(1+2α)/4}, the problem 2. can be approximated in such a way that the order of an error is O(h²) for |z| = O(1). For the scheme with α = -1/6, the choice θ = \sqrt{2} and ε = 1/(2\cdot2) provides the order of the accuracy O(h³).

REFERENCES:
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 analytic 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.

In this paper we consider the following problem:

Find such solution of biharmonic equation

\[ \Delta^2 u = 0 , \quad y > 0 \]

which satisfies the following conditions

2. \[ u(x,0) = 0 , \quad -\infty < x < \infty \]

3. \[ \frac{\partial u}{\partial y}(x,0) - 6'(x)u_{yy}(x,0) = h(x) , \quad -\infty < x < \infty \]
and which is bounded when \( y \to \infty \), if \( h(x) \) is a given continuous function and

\[
\mathcal{S}'(x) = \frac{ae^{-x} + b}{e^{-x} + c}, \quad /a, b, c - \text{const.}/
\]

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

\[
(4) \quad u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0
\]

the Fourier transformation

\[
(5) \quad \mathcal{F}\left\{ \frac{\partial^p}{\partial x^p} \frac{\partial^q}{\partial y^q} u \right\} = (-ix)^p \frac{a^q}{dy^q} U(x,y)
\]

we get the ordinary differential equation

\[
(6) \quad U_{yyyy} - 2x^2 U_{yy} + x^4 U = 0
\]

with general solution

\[
(7) \quad U(x,y) = c_1(x)e^{xy} + c_2(x)ye^{xy} + c_3(x)e^{-xy} + c_4(x)ye^{-xy}.
\]

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

\[
(8) \quad U(x,y) = yC(x)e^{-x|y|}.
\]

\[
C(x) = \begin{cases} 
  c_4(x), & x > 0 \\
  c_2(x), & x < 0 
\end{cases} \quad c_2(0) + c_4(0) = 0.
\]

If we substitute the value of \( \mathcal{S}'(x) \) 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) \( u_y(x,0) - au_{yy}(x,0) - h(x) = \xi(x) \);

in the condensed form

\[ (10) \quad cu_y(x,0) - bu_{yy}(x,0) - ch(x) = -e^{-x}\xi(x). \]

When we apply the Fourier transformation (5) on the equation (10) it transforms into

\[ (11) \quad cu_y(x,0) - bu_{yy}(x,0) - ch(x) = -\Phi(x+i). \]

Since \( u_y(x,0) = C(x) \) and \( u_{yy}(x,0) = -2|x|C(x) \), substituting these values into (11) we have

\[ (12) \quad C(x)(c+2b|x|) = ch(x) - \Phi(x+i). \]

Application of Fourier transformation (5) on the relation (9) gives

\[ (13) \quad C(x)(1+2a|x|) = H(x) + \Phi(x). \]

Elimination of \( C(x) \) from (12) and (13) gives

\[ (14) \quad \Phi(x)v = \frac{1+2a|x|}{c+2b|x|} \Phi(x+i) + cH(x) \frac{1+2a|x|}{c+2b|x|} - H(x). \]

Relation (14) represents, so called, Carleman's boundary value problem for determining analytic function \( \Phi(z) \).

The most of the theory of Carleman's boundary value problem is developed by soviet authors and analytic solving methods are given in details in [1].

Nevertheless, in many cases it is more convenient to apply approximate solving methods. In monograph [2] pp 156-158, the following theorem, which enables approximate solution of problem (14), is formulated and proved.

**Theorem T:** Given the Carleman's boundary value problem

\[ (15) \quad K\Phi \equiv \Phi(x) + [1 + D(x)] \Phi(x+i) = G(x), \quad -\infty < x < \infty \]

and corresponding approximate problem

\[ (16) \quad \tilde{K}\tilde{\Phi} \equiv \tilde{\Phi}(x) + [1 + \tilde{D}(x)] \tilde{\Phi}(x+i) = \tilde{G}(x), \quad -\infty < x < \infty. \]
Suppose that for coefficient by \( \phi(x+i) \) the following conditions are fulfilled

\[
\begin{align*}
1 + \tilde{D}(x) & \neq 0, \quad \tilde{D}(x) - \text{continuous}, \quad \tilde{D}(\pm \infty) = 0, \\
\text{Ind}[1 + \tilde{D}(x)] &= 0
\end{align*}
\]

and that it may be factorized in the following way

\[
1 + \tilde{D}(x) = \frac{\tilde{N}(x)}{\tilde{N}(x+i)}
\]

where function \( \tilde{N}(z) \) is continuous, bounded and analytic in the belt \( 0 < \text{Im } z < 1 \) and nonzero in that domain.

Let us introduce the notation

\[
M = \max \left\{ \max_x |\tilde{N}(x)|, \max_x |\tilde{N}(x+i)| \right\}
\]

Let, further, in Carleman's boundary value problem (15) function \( D(x) \) be bounded and such that

\[
|D(x) - \tilde{D}(x)| \leq M \max_x \left| \frac{\tilde{N}(x)}{\tilde{N}(x+i)} \right|
\]

Then for any right-hand side \( G(x) \) from \( L_2(-\infty, \infty) \), problem (15) has the unique solution in belt \( 0 < \text{Im } z < 1 \). That solution is determined by formula

\[
\phi = \tilde{\phi} + \left[ I + \tilde{K}^{-1} \left( K - \tilde{K} \right) \right]^{-1} \tilde{K}^{-1} \left( G(x) - K \tilde{\phi} \right)
\]

where \( \tilde{\phi} = \tilde{\phi}(x) \) is the solution of approximate Carleman's problem (16). The difference between solutions is estimated by the inequality

\[
\| \phi - \tilde{\phi} \|_{L_2} \leq \| \tilde{K}^{-1} (G(x) - K \tilde{\phi}) \|_{L_2} - \| \tilde{K}^{-1} (K - \tilde{K}) \|_{L_2}
\]

where the inverse operator \( \tilde{K}^{-1} \) is determined by formula

\[
\tilde{K}^{-1} = \tilde{N}(x) \tilde{F} \left[ \frac{1}{1 + e^{-t}} \tilde{C}^{-1} \left( \frac{G}{\tilde{N}} \right) \right]
\]
Let us now apply the mentioned theorem on the approximate solving of Carlemann's problem (14). Instead of the whole upper halfplane we'll obtain only the belt $0 < \text{Im } z < 1$ and choose, for sake of easier computation, that $a = 1/2$, $b = 1/2$ and $c = 2$. Then Carlemann's boundary value problem (14) appears to be

\[
\phi(x) + \frac{1 + |x|}{2 + |x|} \phi(x + i) = 2H(x) \cdot \frac{1 + |x|}{2 + |x|} - H(x),
\]

here function $\phi(x)$ has to be analytic in the belt $0 < \text{Im } z < 1$ and for every $y \in [0, 1]$ satisfies the inequality

\[
\int_{-\infty}^{\infty} |\phi(x + iy)|^2 \, dx \leq C.
\]

The free term $G(x) = 2H(x) \cdot \frac{1 + |x|}{2 + |x|} - H(x)$ is given in

$\mathbb{R}(-\infty, \infty)$. Let us choose in our case that $1 + \tilde{D}(x) = \frac{x^2 + 25}{x^2 + 36}$.

Coefficient $1 + \tilde{D}(x)$ of the approximate problem is chosen 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

\[
\frac{x^2 + m^2}{x^2 + (m + 1)^2}.
\]

For the equation (23) the corresponding approximate equation will be

\[
\tilde{\phi}(x) + \frac{x^2 + 25}{x^2 + 36} \tilde{\phi}(x + i) = 2 \frac{x^2 + 25}{x^2 + 36} H(x) - H(x).
\]

In order that the equation (25) has the unique solution it is sufficient /see [2], § 15/ that the following conditions are fulfilled

\[
1 + \tilde{D}(x) \neq 0 \quad \text{and} \quad \tilde{D}(x) - \text{continuous}.
\]

\[
\tilde{D}(\pm \infty) = 0 \quad \text{and} \quad \text{Ind} \left[ 1 + \tilde{D}(x) \right] = 0.
\]
It is easy to check that all of these conditions hold. Coefficient by $\mathcal{F}(x+i)$ can be factorized in the following way

\[
1 + \mathcal{F}(x) = \frac{x^2 + 25}{x^2 + 36} = \frac{\mathcal{N}(x)}{\mathcal{N}(x+i)}, \quad \mathcal{N}(x) = \frac{x + 5i}{x - 6i}
\]

where the function $\mathcal{N}(z)$ is continuous, bounded and analytic in the belt $0 < \text{Im} z < 1$ and nonzero in that domain. Then the boundary condition (25) by the substitution

\[
\mathcal{F}(x) = \mathcal{N}(x) \psi(x)
\]

where $\psi(x)$ is new continuous function, is transformed into

\[
\psi(x) + \psi(x+i) = 2 \frac{H(x)}{\mathcal{N}(x+i)} - \frac{H(x)}{\mathcal{N}(x)}, \quad -\infty < x < \infty.
\]

When we apply the inverse Fourier transformation the equation (29) transforms into

\[
\psi(x) + e^{-x} \psi(x) = h(x) - 11 \int_{-\infty}^{x} 2e^{-6(x-s)}e^{-5(x-s)} h(s) ds.
\]

Equation (30) gives us the function $\psi(x)$. For determining of function $\mathcal{E}(x)$ we use relation

\[
\mathcal{F}(x) = \psi(x) \mathcal{N}(x) = \psi(x) \frac{x + 5i}{x - 6i} = \psi(x) + \frac{11i}{x - 6i} \psi(x)
\]

When we apply the inverse Fourier transformation on (31) we get

\[
\mathcal{E}(x) = \psi(x) - 11 \int_{-\infty}^{x} e^{6(x-s)} \psi(s) ds.
\]

From formulae (30) and (32) function $\mathcal{E}(x)$ may be determined by $h(x)$. Now the approximate solution $\bar{u}(x,y)$ of problem (1) - (2) - (3) reduces to the following simpler problem

\[
\Delta^2 \bar{u} = 0
\]

\[
\bar{u}(x,0) = 0
\]
Problem (33)-(34)-(35) can be easily solved if we apply the operational calculus/see for example [3]. An applying Fourier transformation we easily find out that

\[ \tilde{u}(x, y) = y C(x) e^{-|x| y} \]

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

\[ (36) \]

\[ \tilde{u}(x, y) = \mathcal{F}^{-1} \left\{ ye^{-|x| y} \frac{H(x) + \tilde{\Phi}(x)}{1 + |x|} \right\} \]

We also have that

\[ M = \max_x \left\{ \max_x |\tilde{N}(x)| , \max_x |\tilde{N}(x + 1)| \right\} = \frac{6}{5} \]

and

\[ (37) \]

\[ \max_x \left| \frac{D(x) - \tilde{D}(x)}{N(x)} \right| \leq 0.23 \]

In that way we see that all conditions of the theorem T are fulfilled and that exact solution of the approximate Carleman's problem (25) can be taken as the approximate solution of the basic Carleman's problem (23).

At the end, using the Parseval equality and inequalities (21) and (37) we can make the following estimation

\[ (38) \]

\[ \| \mathcal{C}(x) - \widetilde{\mathcal{C}}(x) \|_L^2 = \| \Phi(x) - \tilde{\Phi}(x) \|_L^2 \leq \]

\[ \leq \frac{1}{1 - M \cdot 0.23} \| \mathcal{K}^{-1}(k \tilde{\Phi} - G(x)) \| \]

where the operator \( \mathcal{K}^{-1} \) is determined by formula (22). We have further that
Let us estimate the last integral.

C39) \[ \| N^{-1}(K \Phi(x) - \tilde{g}(x)) \| = \max \left\{ \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} N(x) \int_{-\infty}^{\infty} \frac{1}{1 + e^{-t}} \right| \right\} \]

C40) \[ \int_{-\infty}^{\infty} \left\| \frac{K \Phi(x) - \tilde{g}(x)}{N(x)} \right\|^2 dx \leq M \left( \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} K \Phi(x) - \tilde{g}(x) \right|^2 dx \right)^{1/2} \]

Using (38), (39) and (40) we get inequality (41)

\[ \| \psi(x) - \tilde{\psi}(x) \|_2 \leq (0,23)^2 \int_{-\infty}^{\infty} \tilde{\psi}(x + i + 2H(x))^2 \]

In order to make functions \( \psi \) and \( \tilde{\psi} \) even closer to each other, instead of approximate problem (25) we can take it approximate problem in the following form

C42) \[ \tilde{K} \Phi(x) + \tilde{K} \Phi(x + i) \prod_{k=1}^{n} \frac{x^2 + a_k^2}{x^2 + (a_k + 1)^2} = \]

\[ = 2H(x) \prod_{k=1}^{n} \frac{x^2 + a_k^2}{x^2 + (a_k + 1)^2} \]

with conveniently chosen values for \( n \) and \( a_k \).

REFERENCES
SOLUTION OF POTENTIAL PROBLEMS WITH INTERNAL SOURCES BY BOUNDARY ELEMENT METHOD

Josip, E. Pečarić, Miodrag M. Radojković

ABSTRACT

The paper presents an alternative proof of the boundary integral formulation for two-dimensional potential problem with internal sources. This proof appeared to be much simpler than one derived by same authors in [4] and thus is easier to extend to more complex cases (i.e. three-dimensional problems). Accuracy of the method is illustrated by an example.

1. INTRODUCTION

Finite difference and finite element techniques were almost exclusively used to solve numerically 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 usually 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 from Fig. 1-a was considered, where \( r, S_1, S_2, \ldots, S_n \) are its boundaries \((S_1, S_2, \ldots, S_n \) are circles with radii \( r_01, r_02, \ldots, r_0n \) representing the system of internal sources).

\[ \n \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
\[ \]
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:

\begin{equation}
(5) \quad b = - \sum_{K=1}^{n} Q_k \delta(x - \hat{x}_k)
\end{equation}

Then (3) becomes:

\begin{equation}
(6) \quad c^i u_i - \sum_{j=1}^{n} u_j^{i\ast} Q_j + \int_{\Gamma} u q^{\ast} d\Gamma = \int_{\Gamma} \sum_{K=1}^{N} f q u^{\ast} d\Gamma
\end{equation}

i.e.

\begin{equation}
(7) \quad c^i u_i - \sum_{j=1}^{n} u_j^{i\ast} Q_j + \sum_{K=1}^{N} \int_{\Gamma_k} f q u^{\ast} d\Gamma = \sum_{K=1}^{N} \int_{\Gamma_k} f q u^{\ast} d\Gamma
\end{equation}

where $N$ is the number of segments used to divide the boundary (boundary elements).

In the case when all $Q_j$ 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 $r_{01}$ 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:

\begin{equation}
(8) \quad u_{N+i} - \sum_{j=1}^{n} u_j^{i\ast} Q_j + \int_{\Gamma} u q^{\ast} d\Gamma = \int_{\Gamma} \sum_{K=1}^{N} f q u^{\ast} d\Gamma
\end{equation}

For $j=1$ one has:

\begin{equation}
(9) \quad u_j^{i\ast} = \frac{1}{2\pi} \ln \frac{1}{r_{0i}}
\end{equation}

For $j \neq i$ one can suppose $r_{ij} >> r_{0i} + r_{0j}$ ($r_{ij}$ is the distance between sources i and j) and put:

\begin{equation}
(10) \quad u_j^{i\ast} \approx \frac{1}{2\pi} \ln \frac{1}{r_{ij}}
\end{equation}

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 $\Gamma$ 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 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.

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 \(r_3 = 0.5\) 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>
<thead>
<tr>
<th>P_1</th>
<th>P_2</th>
<th>P_3</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.97</td>
<td>4.83</td>
<td>-4.94</td>
<td>Analytic solution</td>
</tr>
<tr>
<td>6.025</td>
<td>5.990</td>
<td>4.785</td>
<td>Boundary elements</td>
</tr>
</tbody>
</table>

Table 1.

Potentials inside the domain in different cross sections of the half plain are compared in 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


A POSSIBILITY FOR CALCULATING PRESSURE GRADIENT FORCE
IN SIGMA COORDINATE SYSTEM

Dragutin T. Mihailović

ABSTRACT:

A new scheme for the calculation of pressure gradient force in the sigma coordinate system is proposed. An approximation or the wa term in the thermodynamics equation is considered 00. The proposed method and an earlier approach {2} is compared by time-integrations of the atmosphere at rest.

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°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 encountered by some of these analogues when geopotential is initially specified rather than temperature has recently been discussed by Mesinger {5} and compared by a numeri-
cal example of the atmosphere at rest \cite{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 \( \omega a \) term that provides consistent transformation from potential to kinetic energy. Finally, the proposed schemes was compared with an earlier one \cite{2}.

2. METHOD OF CONSTRUCTION

Notation

- \( c_p \): specific heat at constant pressure
- \( p_k \): suffix indicating level of the model
- \( p \): pressure
- \( R \): gas constant
- \( s \): suffix indicating surface value
- \( t \): suffix indicating atmosphere top value
- \( T \): temperature
- \( V \): lateral vector wind
- \( u, v \): components of \( V \)
- \( \alpha \): specific volume
- \( \sigma \): \((p - p_t)/(p_{\sigma} - p_t)\) the vertical coordinate
- \( \delta \): \(\partial \sigma/\partial t\)
- \( \pi \): \(p_{\sigma} - p_t\)
- \( \omega \): \(\partial p/\partial t\)
- \( \phi \): geopotential
- \( \nabla_{\sigma} \): lateral del operator in "sigma" surfaces
- \( \nabla_p \): lateral del operator in "pressure" surface

In the sigma coordinate system, the differential form of the pressure gradient force has the form

\[
- \nabla_p \phi = - \nabla_{\sigma} \phi - RT \nabla \pi
\]

Starting from this expression, Kurihara \cite{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 than the former one \(^{10}\). In our case we used quadratic interpolation in accordance with the hydrostatic equation in the form

\[
\frac{\partial \phi}{\partial p} = -\frac{RT}{\rho}.
\]

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 \(\omega a\) term of the thermodynamic equation,

\[
\frac{\partial}{\partial \sigma} (\pi \sigma_p T) + \nabla_\delta (\pi \nabla_\delta T) + \frac{\partial}{\partial \sigma} (\pi \delta c_p T) = \pi \omega a,
\]

ensuring consistency of the transformation between kinetic and potential energy, has been developed. Experience has shown that it is desirable to preserve the consistency even in numerical models designed for short-range simulations. Otherwise, numerical instability may be encountered 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

\[
\pi \omega a = -\frac{\partial}{\partial \sigma} (\pi \phi \delta) - \nabla_\delta (\pi \phi V) - \frac{\partial}{\partial \sigma} (\pi \delta \sigma_p \delta) - \nabla_\delta (\pi \nabla_\delta \phi).
\]

It was stated already that the exact cancelation, in the finite-difference form should be provided between the \(\omega a\) terms in (3) and (4).
Taking into account the continuity equation, hydrostatic equation, and \( \omega \), we arrive at

\[
\pi \omega = -\frac{3}{\sigma} (\pi \phi) - \frac{3}{\sigma^2} (\phi \omega) \frac{\partial \pi}{\partial \xi} - \phi \nabla \phi \cdot \omega + RT \frac{\partial \ln \rho}{\partial \rho} \omega \nabla \phi \pi.
\]

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

\[
\nabla \pi = \frac{\pi (\nabla \phi - \nabla \phi)}{RT \frac{\partial \ln \rho}{\partial \rho}}.
\]

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 definition of \( \omega \), we can write

\[
\omega \alpha = -\frac{1}{\sigma^2} \left[ (\delta_{\phi} \phi) + \frac{1}{\sigma} \frac{\partial \pi}{\partial \xi} \delta_{\sigma} \phi + \frac{\nabla \phi}{RT \frac{\partial \ln \rho}{\partial \rho}} \frac{\partial \phi}{\partial \sigma} \right].
\]

In the finite-difference form, the last expression, for the case of horizontally staggered variables, in the \( x \) direction, has the form

\[
(\omega \alpha)_0 = -\frac{1}{\sigma^2} \left[ (\delta_{\phi} \phi) + \frac{1}{\sigma} \frac{\partial \pi}{\partial \xi} \delta_{\sigma} \phi \right]_0 + Z_0 \nabla \pi \left[ \pi \sigma (\delta_{\sigma} \phi - \delta_{\sigma} \phi) \right]
\]

where

\[
Z_0 = \frac{\delta_{\sigma} \phi}{RT \frac{\partial \ln \rho}{\partial \rho}}.
\]

and \( \delta_{x,\rho} \) and \( \delta_{x,\sigma} \) are operators of divergence in the finite-difference form in \( x \) direction for \( p=\text{const.} \) and \( \sigma=\text{const.} \). The subscript 0 denotes the point where the contribution of the \( \omega \alpha \) term is calculated; \( \pi \sigma \) denotes the value of \( \pi \) in the point in which velocity components are not defined.
3. A NUMERICAL EXAMPLE

We compared 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 communication). The integrations were performed in a two dimensional domain \((x, \sigma)\) with constant boundary conditions specified at the western boundary \((x=0)\). At the eastern boundary \((x=12000 \text{ km})\) the radiation boundary condition was used [7]. A triangular mountain with 500 km width and the maximum height of 2 km was defined in the middle of the domain. The atmosphere was divided 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 \(\ln p\) 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\(^{-1}\).

<table>
<thead>
<tr>
<th>days</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>18.49</td>
<td>6.08</td>
<td>5.61</td>
<td>5.49</td>
<td>5.55</td>
<td>5.42</td>
<td>5.42</td>
<td>5.40</td>
<td>5.47</td>
<td>5.48</td>
<td>5.53</td>
<td>5.51</td>
</tr>
<tr>
<td>I</td>
<td>0.39</td>
<td>0.39</td>
<td>0.41</td>
<td>0.42</td>
<td>0.43</td>
<td>0.44</td>
<td>0.46</td>
<td>0.47</td>
<td>0.48</td>
<td>0.50</td>
<td>0.51</td>
<td>0.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>days</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>8.92</td>
<td>5.40</td>
<td>5.12</td>
<td>4.84</td>
<td>4.68</td>
<td>4.60</td>
<td>4.41</td>
<td>4.32</td>
<td>4.20</td>
<td>4.10</td>
<td>3.99</td>
<td>3.93</td>
</tr>
<tr>
<td>I</td>
<td>0.32</td>
<td>0.32</td>
<td>0.33</td>
<td>0.34</td>
<td>0.35</td>
<td>0.36</td>
<td>0.38</td>
<td>0.39</td>
<td>0.40</td>
<td>0.42</td>
<td>0.43</td>
<td>0.44</td>
</tr>
</tbody>
</table>

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 $V=V(r,\theta)$, $V=V(z)$ and $\delta V/\delta z=0$) 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 direction $r=a^2/d$, $\phi=\lambda$) can be used.

So the electric scalar potential is

$$V = q'G,$$

where

$$G = \frac{1}{4\pi \varepsilon} \ln \frac{r^2 d^2 + a^4 - 2ad \cos(\phi - \lambda)}{a^2 \left[ (r^2 + d^2 - 2rd \cos(\phi - \lambda)) \right]}$$

is so-called Green's function, $r$, $\theta$ and $\phi$ are cylindrical coordinates, $\varepsilon$ is the electric permittivity and $\delta(r-d)$ and $\delta(\phi-\lambda)$ are Dirac's $\delta$-functions defined for $r=d$ and $\phi=\lambda$, 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

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial V}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 V}{\partial \phi^2} = 0$$

and assume the potential solution in the separable variable for $V = R(r) F(\phi)$, where

$$F'' = k^2 F, \quad r^2 R'' + r R' + \lambda^2 R = 0$$

and $\lambda$ is the separable constant to be determined. The solutions differential equations (4) are $e^{\lambda r}$ and $e^{-\lambda r}$, and $cl(r) = \cos(k \ln \frac{r}{a})$ and $sl(r) = \sin(k \ln \frac{r}{a})$, respectively. Puting $k = \lambda k_0$ ($\lambda$ is an integer) and determining $\lambda_0$ so the functions $cl_n(r) = \cos(n \lambda k_0 \ln \frac{r}{a})$ and

$$sl_n(r) = \sin(n \lambda k_0 \ln \frac{r}{a})$$

satisfied conditions

$$\int_a^b \frac{cl_n(r) cl_m(r) dr}{r} = \begin{cases} 0, & \text{for } n \neq m \\ \frac{1}{2\lambda k_0} (1 + \delta_{nm}), & \text{for } n = m \end{cases}$$

and

$$\int_a^b \frac{sl_n(r) sl_m(r) dr}{r} = \begin{cases} 0, & \text{for } n \neq m \text{ and } n=0 \text{ or } m=0 \\ \frac{1}{2\lambda k_0}, & \text{for } n=0 \text{ or } m=0 \end{cases},$$

where $b > a$, we have $k_0 = \pi / \ln(b/a)$. 
expanding Dirac's $\delta$-function $\delta(r-d)$ in the series

$$ r \delta(r-d) = \sum_{n=1}^{\infty} \frac{2 \text{sn}_n(r) \text{sn}_n(d)}{\ln(b/a)} , \text{ for } a < d < b , $$

we have in the case when $b = \infty$

$$ r \delta(r-d) = \frac{2}{\pi} \int_{0}^{\infty} \frac{\sin(\text{pln}_a^2) \sin(\text{pln}_a^d)}{p \text{sh}(pR)} \text{ch}[\psi - \delta(R)] p \, dp . $$

Because of the obtained Dirac's $\delta$-function integral transformation the solution of Poisson's equation (1) can be written as

$$ V = \frac{2}{\pi \epsilon} \int_{0}^{\infty} \frac{\sin(\text{pln}_a^2) \sin(\text{pln}_a^d)}{p \text{sh}(pR)} \text{ch}[\psi - \delta(R)] p \, dp , $$

where "+" is for $0 < \phi < A$ and "-" is for $A < \phi < 2\pi$.

3. APPLICATION OF PRESENT RESULTS, EXAMPLES 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: $V(\text{from formula } 2) = V(\text{from formula } 9)$. Using series

$$ \text{ch}(e^{-\delta(R)} p) = \sum_{m=0}^{\infty} e^{p[e^{-\delta(R)} - (2m+1)\pi]} + e^{-p[e^{-\delta(R)} + (2m+1)\pi]} $$

and integral

$$ \int_{0}^{\infty} \frac{1-\cos pA}{p} e^{-pB} dp = \frac{1}{2} \ln \frac{A^2+B^2}{B^2} , \text{ for } B > 0 , $$

we have from last expression

$$ \sum_{m=-\infty}^{\infty} \ln \left( \frac{R^2}{(m+1)^2} \right)^2 + \left[ e^{-\delta(R)} - (2m+1)\pi \right]^2 \frac{d^2 x^2 + a^4 - 2a^2 x^d \cos(e-d)}{a^2 \left[ x^2 + d^2 - 2x^d \cos(e-d) \right]^2} , $$

where $a < x < \infty$, $a < d$ and with "+" for $0 < \phi < A$ and "-" for $A < \phi < 2\pi$.

Putting in (11) $\psi = \phi - \delta$, $R = r/a > 1$, $D = d/a > 1$ or $R = a^A$, $A > 0$, and $R/D = e^B$, $B$ is optional, we have:

$$ P = \sum_{m=-\infty}^{\infty} \frac{\ln^2(RD) + (\psi - 2m\pi)^2}{\ln^2(R/D) + (\psi - 2m\pi)^2} = \frac{R^2(D^2+1-2RD\cos\psi)}{D^2+2-2RD\cos\psi} $$

and

$$ P_0 = \sum_{m=1}^{\infty} \frac{A^2 + (\psi - 2m\pi)^2}{B^2 + (\psi - 2m\pi)^2} = \frac{\text{chA} - \cos\psi}{\text{chB} - \cos\psi} . $$

Separately, we have:

$$ P_1 = \sum_{m=1}^{\infty} \frac{A^2 + 4m^2\pi^2}{B^2 + 4m^2\pi^2} = \frac{B}{A} \frac{\text{sh}(A/2)}{\text{sh}(B/2)} , \text{ for } \psi = 0 , $$

$$ P_2 = \sum_{m=1}^{\infty} \frac{C^2 + m^2}{m^2} = \frac{\text{sh}(B)}{C^2} , \text{ for } \psi = 0, A = 2\pi C \text{ and } B = 0 \text{ and }$$

$$ P_3 = \sum_{m=1}^{\infty} \frac{C^2 + (2m-1)^2}{D^2 + (2m-1)^2} = \frac{\text{sh}(C/2)}{\text{ch}(D/2)} , \text{ for } \psi = \pi/2, A = \pi C/2 \text{ and } B = \pi D/2 . $$
The obtained formulas are very useful, because of the slow convergence of present infinite products. The convergence of products shown following numerical results (Approximate values are calculated by multiplying 65 members of infinite products. The exact values are in brackets):)

For R=2, D=2 and Ψ=-R  \( P = 1.55783 \) (exact \( P = 1.5625 \)).
For R=2, D=4 and Ψ=-R  \( P = 2.24039 \) (exact \( P = 2.2500 \)).
For R=2, D=2 and Ψ=-R/6  \( P = 9.36901 \) (exact \( P = 9.93971143 \)).
For R=2, D=2 and Ψ=-2R/3  \( P = 1.74477 \) (exact \( P = 1.7500 \)).

In the following table the exact values of \( P_2 \) for different \( C \) are shown.

<table>
<thead>
<tr>
<th>C</th>
<th>( P_2 )</th>
<th>C</th>
<th>( P_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0000000000</td>
<td>2.0</td>
<td>4.2612923E+01</td>
</tr>
<tr>
<td>0.01</td>
<td>1.000164501</td>
<td>5.0</td>
<td>2.1121847E+05</td>
</tr>
<tr>
<td>0.1</td>
<td>1.016530706</td>
<td>10.0</td>
<td>7.0078318E+11</td>
</tr>
<tr>
<td>0.5</td>
<td>1.451052383</td>
<td>20.0</td>
<td>1.5428269E+25</td>
</tr>
<tr>
<td>1.0</td>
<td>3.676077910</td>
<td>50.0</td>
<td>5.2682675E+65</td>
</tr>
</tbody>
</table>

Combining the results (16) for \( C=1 \) with known formula

\[
P_3 = \lim_{n \to \infty} \left( 1 - \frac{1}{m^2} \right) = \frac{1}{2},
\]
we have

\[
P_4 = \lim_{m \to \infty} \left( 1 - \frac{1}{m^4} \right) = \frac{\sinh R}{4R} = 0.9190194776.
\]

REFERENCES

VALUATION OF SEVERAL SINGULAR INTEGRALS USING ELECTROSTATIC FIELDS LOWS

Dragutin M. Velicković

ABSTRACT:
Using electrostatic fields lows a special approach to the several singular integrals evaluating is presented. The obtained results are useful in numerical solution of electrostatic problems by integral equations technique.

ODREĐIVANJE NEKIH SINGULARNIH INTEGRALI POMOĆU ZAKONITOSTI ELEKTROSTATIČKIH POLJI: 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 mapping 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 excellent 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).

For electric scalar potential evaluation two general procedure exist:
1. In the case when the surface charges densities on the electrode, \( \eta'(\vec{r}') \), are known, the potential, \( V \), is:

\[
V(\vec{r}) = \begin{cases} 
U, & \text{for } \vec{r} \in S_1, S_1 \text{ is conductor interior} \\
U - \frac{1}{2\pi \epsilon} \oint \eta(\vec{r}') d\vec{r}' \ln(|\vec{r}-\vec{r}'|/|\vec{r}_0-\vec{r}'|), & \text{for } \vec{r} \in S_e, S_e \text{ is conductor exterior},
\end{cases}
\]

where \( c \) is the contour of conductor cross section, \( \epsilon \) is electric permittivity, \( U \) is the conductor potential and \( q' = \oint \eta(\vec{r}') d\vec{r}' \) is the conductor total charge per unite length.

2. The second way for potential evaluating is based on conformal mapping. If we have the complex function \( w = Re^{j\phi} = u+jv = f(z=re^{j\theta}=x+jy) \), \( j = \sqrt{-1} \), which map the exterior of conductor: \( z \)-plane to the unite circular cylinder exterior in \( w \)-plane (Fig.2), the complex potential is

\[
\Phi = \begin{cases} 
U, & \text{for } R < 1 \\
U - \frac{q}{2\pi \epsilon} \ln w = V+j\phi, & \text{for } R > 1.
\end{cases}
\]

During conformal mapping the electrode potential and total line charge densities are constant.

The real part of \( \phi \) is potential,

\[
V = Re \{\phi\} = U - \frac{q}{2\pi \epsilon} \ln R.
\]

The electric field on the conductor surface is \( E = |d\phi/dz| = q' |w'||2\pi \epsilon \), where \( |w'||=|dw/dz| \) for \( R=1 \). Using boundary condition \( \eta = \epsilon E \) we have

\[
\eta = \frac{q'}{2\pi} |w'_0|.
\]
uting (4) in (1) the potential is
\[ V(\tilde{r}) = U - \frac{Q}{4\pi\epsilon} \int \frac{|w'| \, dl'}{c} \ln\left(\frac{|\tilde{r} - \tilde{r}'|}{|\tilde{r}_0 - \tilde{r}'|}\right), \text{ for } \tilde{r} \in S_e. \]

Comparing (3) and (5) we have
\[ \int \frac{|w'| \, dl'}{c} \ln\left(\frac{|\tilde{r} - \tilde{r}'|}{|\tilde{r}_0 - \tilde{r}'|}\right) = \begin{cases} 0, & \text{for } \tilde{r} \in S_1^\prime, \\ 2\pi \ln R, & \text{for } \tilde{r} \in S_e. \end{cases} \]

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 \( \tilde{r}_0 \in S_1^\prime. \)

3. EXAMPLES

**Example I**: Consider linear conformal mapping \( w = z/a, \) 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: \( R = y/b, z = \phi \) \( |w'| = 1/a, \) \( \int \text{d}E, \) \( |\tilde{r} - \tilde{r}'| = \sqrt{a^2 + 2ar_0 \cos(\phi - \phi')}, |\tilde{r}_0 - \tilde{r}'| = \sqrt{r^2 + 2r_0 \cos\phi}, \)

for \( r_0 \leq a, \) and the expression (6) give:
\[ \int \frac{|w'|}{c} \ln\left(\frac{|\tilde{r} - \tilde{r}'|}{|\tilde{r}_0 - \tilde{r}'|}\right) = \begin{cases} 0, & \text{for } r \leq a \\ 4\pi \ln(r/a), & \text{for } r > a. \end{cases} \]

The value of integral (7) is independent of \( \phi_0 \) and \( r_0 \leq a. \)

For \( a = 1 \) and \( r_0 = 0 \) we have
\[ \int \frac{|w'|}{c} \ln\left(\frac{|\tilde{r} - \tilde{r}'|}{|\tilde{r}_0 - \tilde{r}'|}\right) = \begin{cases} 0, & \text{for } r \leq 1 \\ 4\pi \ln r, & \text{for } r > 1. \end{cases} \]

**Example II**: Consider bilinear conformal mapping \( z = \frac{w - 1}{w}, \) c is positive constant, which map exterior of strip conductor having width \( 2c \) in \( z\)-plane to the exterior of unite circle in \( w\)-plane.

Now is: \( x = x_0 + 1 \) \( \frac{1}{R} \cos\gamma, \) \( y = \frac{1}{R} \sin\gamma, \) \( |w'| = 1/\sqrt{c^2 - x^2}, \)

\[ |\tilde{r} - \tilde{r}'| = \sqrt{(x-x')^2 + y^2}, |\tilde{r}_0 - \tilde{r}'| = |x_0 - x'|, \text{ for } |x_0| \leq c, \text{ and } \]
\[ c \]
\[ \int_{-c}^{c} \frac{d^2x}{c^2 - x^2} \ln\left(\frac{(x-x')^2 + y^2}{(x_0-x')^2}\right) = 2\pi \ln R, \text{ for } |x_0| \leq c, \text{ where } \]
\[ R = \sqrt{c^2 + 2c^4 - 2c^2 \cos^2 \phi \pm \sqrt{2c^4 - 4c^2 \cos^2 \phi}} \]
\[ x = r \cos \phi, \quad y = r \sin \phi, \quad x^2 + y^2, \quad \text{and} \quad \cos 2\phi = (x^2 - y^2)/(x^2 + y^2). \]

For different values of \( x \) and \( y \) we have:
\[ \int_{-c}^{c} \frac{d^2x}{c^2 - x^2} \ln\left(\frac{(x-x')^2}{(x_0-x')^2}\right) = \begin{cases} 0, & \text{for } |x| \leq c, |x_0| \leq c \\ \pi \ln \frac{2x^2 - c^2 + 2x^2 \sqrt{x^2^2 - c^2}}{c^2}, & \text{for } |x| > c \end{cases} \]
and
\[ \int_{-c}^{c} \frac{d^2x}{c^2 - x^2} \ln\left(\frac{x^2 + y^2}{(x_0-x')^2}\right) = \pi \ln \frac{2x^2 + c^2 + 2x^2 \sqrt{x^2 + c^2}}{c^2}, \text{ for } |x_0| \leq c. \]