UNIVERSITY OF NIŠ
FACULTY OF ELECTRONIC ENGINEERING

NUMERICAL METHODS AND APPROXIMATION THEORY III

Edited by G. V. Milovanović

N i š , 1988

UNIVERSITY OF NIŠ
FACULTY OF ELECTRONIC ENGINEERING

NUMERICAL METHODS AND APPROXIMATION THEORY III

Niš, August, 18 - 21, 1987

Edited by G. V. Milovanović

N i š , 1988

Numerical Methods and Approximation Theory III

This publication was in part supported by The Regional Science Fondations of Niš.

Published by. Faculty of Electronic Engineering, University of Niš, P. O. Box 73, 18000 Niš, Yugoslavia

Technical support: Lj. M. Kocić and S. Zinovijev

Printed by: Prosveta, Niš

Numbers of copies: 500

PREFACE

The third conference on Numerical Methods and Approximation Theory was held in Niš at the Faculty of Electronic Engineering, University of Niš, August 18–21, 1987. It was attended by 140 participants from 20 countries. There were 85 papers presented in three sections.

Previous conferences were held in Niš (1984) and Novi Sad (1985) with 55 and 68 participants, respectively.

Two types of selected and refereed papers appear in this Proceedings: four long survey papers, based on 45–minute invited lectures, and 31 shorter research papers, presented at the thirty– and fifteen–minute talks. The papers were submitted in the prescribed form ready for copying. In both parts, Invited papers and Contributed papers, they are published in the alphabetic order of the surnames of the first authors.

I wish to thank the members of the Organizing Committee and all the referees for their voluntary work.

G. V. Milovanović

C O N T E N T S

LIST OF AUTHORS

ALIĆ, MLADEN
Dept. of Mathematics, Univ. of Zagreb, p.o.box 173, 41001 Zagreb

AŠIĆ, MIROSLAV D.
Dept. of Mathematics, Faculty of Natural Sciences and Mathematic
Studentski Trg 16, 11000 Belgrade, YU

ARAKELIAN, ARAM H.
Academy of Sciences of Aramenian SSR, P.Sevaka 1, 375044Yerevan,

BIDKHAM, MOHAMMAD
Dept. of Mathematics, Faculty of Natural Science and Technology,
mia Millia Islamia, 110025 New Delhi, INDIA

BOHTE, ZVONIMIR
Institute of Mathematics, Physics and Mechanics, Jadranska 19,
61000 Ljubljana, YU

BUTZER, PAUL L.
Lehrstuhl A für Mathematik, R.W.T.H., Templergraben 55, 5100 Aac
FRG

CRISCUOLO, GIULIANA
Istituto per Applicazioni della Matematica - C.N.R., via P. Cas-
llino 111, 80131 Napoli, ITALY

CVETKOVIĆ, LJILJANA
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad,

DANKOVIĆ, BRATISLAV
Dept. of Automatics, Faculty of Electronic Engineering, p.o.box
18000 Niš, YU

DA SILVA, MANUEL
Grupo de Matemática Aplicada, Faculdade de Ciências, Universida
do Porto, 4000 Porto, PORTUGAL

DELLA VECCHIA, BIANCAMARIA
Istituto per Applicazioni della Matematica - C.N.R., via P. Cas
llino 111, 80131 Napoli, ITALY

DELVOS, FRANZ-JÜRGEN
Lehrstuhl für Mathematik I, Univ. of Siegen, Hölderlin Str. 3,
5900 Siegen, FRG

DEWAN, KUM KUM
Dept. of Mathematics, Faculty of Natural Science and Technology
mia Millia Islamia, 110025 New Delhi, INDIA

DEY, CHARLIE
Charleston High School, Charleston, IL 61920, USA

, SUHRIT KUMMAR
Dept. of Mathematics, Eastern Illinois Univ.,Charleston,IL 61920,USA

DJURIĆ, MIRJANA
Fakultet zaštite na radu, Čarnojevićeva 10a, 18000 Niš, YU

FISCHER, HERBERT
Institut für Angewandte Mathematik und Statistik, Technische Universität München, Arcisstrasse 21, 8000 München 2, FRG

GATTESCHI, LUIGI
Dipartimento di Matematica dell'Univesità,Via Carlo Alberto 10,10123 Torino, ITALY

GAUTSCHI, WALTER
Dept. of Computer Sci.,Purdue Univ.,West Lafayette,IN 47907, USA

GHELARDONI, PAOLO
Istituto di Matematiche Applicate "U. Dini", Facoltà di Ingegneria, Università di Pisa, Via Bonanno 25B, 56100 Pisa, ITALY

GHERI, GIOVANNI
Istituto di Matematiche Applicate "U. DINI", Facoltà di Ingegneria, Università di Pisa, Via Bonanno 25B, 56100 Pisa, ITALY

GUESSAB, ALLAL
Département de Mathématiques,Ave. de l'Université,64000 Pau, FRANCE

HERCEG, DRAGOSLAV
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad, YU

HERZBERGER, JÜRGEN
Fachbereich Mathematik, Universität Oldenburg, 2900 Oldenburg, FRG

KOCIĆ, LJUBIŠA
Dept. of Mathematics, Faculty of Elactronic Engineering, p.o.box 73 18000 Niš, YU

KOVAČEVIĆ-VUJČIĆ, VERA V.
Faculty of Organizational Sciences, Jove Ilića 154,11040 Belgrade,

KOZAK, JERNEJ
Dept. of Mathematics and Mechanics, E.K. University of Ljubljana,Jaranska 19, 61111 Ljubljana, YU

LOKAR, MATIJA
Dept. of Mathematics and Mechanics, E.K. University of Ljubljana,Jaranska 19, 61111 Ljubljana, YU

MANGER, ROBERT
Rade Končar Institute, Baštijanova bb, 41000 Zagreb, YU

MARKOWICH, P. A.
Institut für Angewandte und Numerische Mathematik, Wiedner Hauptstr 8-10/115, 1040 Wien, AUSTRIA

MARZULLI, PIETRO
Istituto di Matematiche Applicate "U. DINI", Facoltà di Ingegneria, Università di Pisa, Via Bonanno 25B, 56100 Pisa, ITALY

MASTROIANNI, GIUSEPPE
Università degli Studi della Basilicata, Via N.Sauro, Potenza, ITA

MIJALKOVIĆ, SLOBODAN
Dept. of Microelectronics, Faculty of Electronic Engineering, p.o box 73, 18000 Niš, YU

MILOVANOVIĆ, GRADIMIR
Dept. of Mathematics, Faculty of Electronic Engineering, p.o.box 18000 Niš, YU

PETKOVIĆ, DEJAN
Fakultet zaštite na radu, Čarnojevićeva 10a, 18000 Niš, YU

PETKOVIĆ, MIODRAG
Dept. of Mathematics, Faculty of Electronic Engineering, p.o.box 18000 Niš, YU

POGÁNY, TIBOR
Tehnički fakultet Bor, JNA 12, 19210 Bor, YU

RASSIAS, THEMISTOCLES
Dept. of Mathematics, Univ. of LaVerne, p.o.box 51105, Kifissia, Athens, GRECE 145 10

SAVIĆ, VLADIMIR
PMF Kragujevac, R. Domanovića 12, 34000 Kragujevac, YU

SCHEMPP, WALTER
Lehrstuhl für Mathematik 1, Univ. of Siegen, 5900 Siegen, FRG

SCHMEISER, CHRISTIAN
Institut für Angewandte und Numerische Mathematik, Wiedner Haupt 8-10/115, 1040 Wien, AUSTRIA

SELBERHERR, S.
Institut für allgemeine Elektrotechnik, TU Wien, AUSTRIA

STANKOVIĆ, MIOMIR
Fakultet zaštite na radu, Čarnojevićeva 10a, 18000 Niš, YU

STEFANOVIĆ, LIDIJA
Dept. of Mathematics, Faculty of Electronic Engineering, p.o.bo: 18000 Niš, YU

STENS, R. L.
Lehrs. A für Mathematik, R.W.T.H.,Templergraben 55, 5100 Aachen

STOJADINOVIĆ, NINOSLAV
Dept. of Microelectronics, Faculty of Electronic Engineering, p box 73, 18000 Niš, YU

SURLA, KATARINA
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad,

TAKAČI, DJURDJICA
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad,

›resent instant. Therefore the question: is it possible to re-
:onstruct a bandlimited function (to begin with) from its
;amples taken exclusively from the past, i.e., taking into
ıccount only those f(t) for which $t < t_o$?

One answer to this question is the following: can one find
:oefficients $a_{kn} \in \mathbb{R}$ such that f can be reconstructed from its
;amples taken at the points t_o-T/W, t_o-2T/W, t_o-3T/W, ... from
:he past, in terms of

$$f(t_o) = \lim_{n\to\infty} \sum_{k=1}^{n} a_{kn}\, f(t_o - \tfrac{kT}{W}) \qquad (1.2)$$

:or each $t_o \in \mathbb{R}$? This would determine the value of f at the
›resent time instance $t = t_o$. It is the question of predicting
from its past samples.

There are two problems in this respect: (i) the role of T, naturally $T \in (0,1]$ - the closer T is to 1 the wider apart can the sampling points t_o-kT/W, $k \in \mathbb{N}$ be - and whether for each $T \in (0,1]$ the existence of the predictor coefficients is guaranteed, (ii) the evaluation of these coefficients, i.e., the construction of prediction formulae (1.2) in dependence on T - the closer T is to 1 the nearer is the sampling rate to that of the classical sampling theorem, namely the Nyquist rate 1/W.

Regarding the first problem, by applying a general result due to G. Szegö (1920) or a more general one due to N. Levinson (1940) one can show that for each T with $0 < T < 1$ there exist predictor coefficients a_{kn} such that (1.2) holds uniformly in $t_o \in \mathbb{R}$.

Regarding the second, Wainstein and Zubakow [25] (1962) showed that (1.2) is valid with $a_{kn} := (-1)^{k+1}\binom{n}{k}$ provided $0 < T < 1/3$; J.L. Brown Jr. [2] (1972) extended T to $T < 1/2$ for the coefficient choice $a_{kn} := (-1)^{k+1}\binom{n}{k}(\cos \pi T)^k$. This result was extended even further by W. Splettstoesser [22,23] (1981/82) who showed that (1.2) holds uniformly in $t_o \in \mathbb{R}$ for $a_{kn} := (-1)^{k+1}\binom{n+k-1}{k} 4^{-k}$ with $0 < T < \pi^{-1}$ arc cos $(-8^{-1}) \approx 0.5399$. Thus a sampling rate (even) larger than half the Nyquist rate

is possible in predicting bandlimited functions with coefficients a_{kn} that are even independent of T. Generally, the closer T is to 1, the more complicated will the coefficients a_{kn} (dependent on T) be.

The coefficients that are best, in the sense that the mean square error is minimized, are the solutions of the linear system

$$\sum_{k=1}^{n} a_{kn}\, si(\pi(k-j)TW) = si(\pi jTW) \qquad (1 \leq j \leq n) \tag{1.3}$$

where $si(x) = \sin x/x$. Since these are difficult to determine, and because they depend on n, the foregoing sub-optimal coefficients are more efficient.

Now it is known that a function being bandlimited is a rather restrictive condition. Such a function cannot be simultaneously duration limited, and it is the latter class of functions which actually occurs in practice. Further, beginning with bandlimited functions $f \in L^2(\mathbb{R})$, then f can be extended to the complex plane as an entire function (so one that is extremely smooth) that is of exponential type πW. The next question therefore is whether prediction can be carried out for functions that are <u>not</u> necessarily bandlimited. In this respect W. Splettstoesser [24] showed that if the (r+1)th derivative $f^{(r+1)} \in C(\mathbb{R})$ (= space of all uniformly continuous and bounded functions on $\mathbb{R}$), then

$$\sup_{t\in\mathbb{R}} \left| f(t) - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} (\cos \pi T)^k\, f\left(t - \frac{kT}{W}\right) \right| = O\left[(1+\cos \pi T)^n\, W^{-r-1} + (\sin \pi T)^n \sqrt{W}\right] \qquad (n, W \to \infty) \tag{1.4}$$

for each $0 < T < 1/2$. Since both terms on the right of (1.4) contain a factor tending to zero and one to infinity for $n, W \to \infty$, one has to choose n in dependence on W (or vice versa) such that both terms still tend to zero. It turns out that all the sample instants accumulate at t for $n, W \to \infty$. The details are to be found in [24].

The disadvantages in the prediction procedure described so far are (i) the sampling rates are just T/W with $0 < T << 1$ instead of the Nyquist rate $1/W$; (ii) the sample points in (1.2) depend on t, thus all the sample values have to be computed or measured anew when the series are to be evaluated for another t; (iii) in the case of prediction of not necessarily bandlimited functions generally the number of samples plus the distance between the sample points has to be regulated appropriately (recall (1.4)); (iv) to improve the approximation of f by the series in (1.2) or (1.4) the number n of samples has to be increased; (v) the sampling series (1.2) does not have the (classical) convolution structure for sums as given by the Shannon series (1.1).

To avoid these disadvantages, let us try to reconstruct functions from its past samples by the convolution series

$$(1.5) \qquad (S_W^{\varphi} f)(t) := \sum_{k=-\infty}^{\infty} f\left(\frac{k}{W}\right) \varphi(W(t - \frac{k}{W}))$$

for $W \to \infty$, where the kernel φ will be assumed to be continuous and have compact support contained in $[T_0, T_1]$ for some $0 < T_0 < T_1$. This means that $\varphi(Wt-k) \neq 0$ only for those $k \in \mathbb{Z}$ for which $k/W \in (t - T_1/W, t - T_0/W)$, so that only a finite number of samples taken from the past will be needed to evaluate (1.5), and this number will be fixed for all f, W and t. Increasing W in the series (1.5) will only mean that the distance between the sample points will decrease. Further, f need not necessarily be bandlimited. Of course, the coefficients $\varphi(Wt - k)$ depend on t, but the evaluation of φ should be simpler than that of the signal f to be sought.

It will be seen that our results enable one to predict or extrapolate the value of a signal even arbitrarily far ahead of the sample values.

The aim of this paper is to present a well-motivated overview of recent results obtained at Aachen in the matter. Most of the details, including the proofs of results stated, are to be

found in [7]. See also Chapter 5 of [6] which deals with prediction theory. Regarding the specific examples of Sections 2.3 and 4, they are treated here for the first time in actual detail.

For a continuation of the above approach of Splett-stoesser in the matter, see especially [23], [18], [19].

Connections of the present study with the basic work of A.N. Kolmogorov [12] (1941), N. Wiener [26] (1949) as well as of M.G. Krein [14] (1954) in the subject will be sketched in Section 6.

Concerning possible applications, one of the main ones is to speech processing, see e.g. [17], including differential pulse-code modulation [10]. Further applications are to economic prediction and forecasting, see e.g. [1], to geophysics and medicine, see e.g. [16].

2. PREDICTION OF DETERMINISTIC SIGNALS

2.1. GENERAL RESULTS

Let us now study sampling series of the form $(S_W^{\varphi} f)(t)$, defined in (1.5), where the si -function has been replaced by a kernel $\varphi \in C_{oo}(\mathbb{R})$ (=those $f \in C(\mathbb{R})$ that have compact support). Firstly, $S_W^{\varphi} f$ defines a family of bounded, linear operators from $C(\mathbb{R})$ into itself, with the operator norm

$$\| S_W^{\varphi} \|_{[C,C]} = m_o(\varphi) \qquad (W > 0),$$

$m_r(\varphi)$ denoting the absolute (sum) moment of φ of order $r \in \mathbb{N}_o$, namely

$$m_r(\varphi) := \sup_{t \in \mathbb{R}} \sum_{k=-\infty}^{\infty} |t-k|^r |\varphi(t-k)| .$$

Denote the Fourier transform of $g \in L^1(\mathbb{R})$ by

$$g^{\wedge}(v) = (1/\sqrt{2\pi}) \int_{-\infty}^{\infty} g(t)\ e^{-ivt} dt \qquad (v \in \mathbb{R}).$$

Proposition 1. Let $\varphi \in C_{OO}(\mathbb{R})$. The following three assertions are equivalent:

(i) $\lim_{W\to\infty} (S_W^{\varphi} f)(t) = f(t)$

for each $f \in C(\mathbb{R})$ and each $t \in \mathbb{R}$;

(ii) $\sum_{k=-\infty}^{\infty} \varphi(t-k) = 1 \qquad$ (each $t \in \mathbb{R}$);

(iii) $\varphi^{\wedge}(2k\pi) = \begin{cases} 1/\sqrt{2\pi}\ , & k = 0 \\ 0\ , & k \in \mathbb{Z} \setminus \{0\}\ . \end{cases}$

Proposition 2. Let $\varphi \in C_{OO}(\mathbb{R})$, $r \in \mathbb{N}$. If, in addition to the properties (i), (ii) or (iii) of Proposition 1, there holds

(ii)* $\sum_{k=-\infty}^{\infty} (t-k)^j \varphi(t-k) = 0 \qquad (j = 1,2,\ldots,r-1;\ t \in \mathbb{R})$

or, equivalently,

(iii)* $\varphi^{\wedge(j)}(2k\pi) = 0 \qquad (j = 1,2,\ldots,r-1;\ k \in \mathbb{Z})$

(for r=1 only one condition of Prop. 1 need hold), then there hold the estimates

(2.1) $$\|S_W^{\varphi} g - g\|_C \leq \frac{m_r(\varphi)}{r!} \|g^{(r)}\|_C\ W^{-r} \qquad (g \in C^r(\mathbb{R});\ W > 0)$$

$$\|S_W^{\varphi} f - f\|_C \leq K\omega_r(W^{-1}; f; C(\mathbb{R})) \qquad (f \in C(\mathbb{R});\ W > 0),$$

the constant K depending only on φ. In particular, if $f^{(r-1)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$, then

$$\|S_W^{\varphi} f - f\|_C = O(W^{-r+1-\alpha})\ , \quad W \to \infty\ .$$

Above, $\omega_r(\delta; f; C(\mathbb{R}))$ stands for the rth modulus of continuity of $f \in C(\mathbb{R})$, and $\mathrm{Lip}(\alpha; C(\mathbb{R}))$ for the Lipschitz class of order α. Regarding the foregoing propositions, see

e.g. Ries and Stens [21], [5]. Conditions of the type (ii)*, (iii)* were already used in connection with finite element approximation in Fix and Strang [9].

2.2. CONSTRUCTION OF KERNELS

Fejér's kernel F, defined by

$$F(t) := \frac{1}{2\pi}\left[\frac{\sin t/2}{t/2}\right]^2 , \quad F^{\wedge}(v) = \frac{1}{\sqrt{2\pi}} \begin{cases} 1-|v|, & |v| \leq 1 \\ 0, & |v| > 1 \end{cases}$$

satisfies property (ii)* for r=1. Likewise does de la Vallée Poussin's kernel. However, these kernels have unbounded support. The best examples of φ having compact support are the so-called central B-splines of order $r \geq 2$, defined by

$$M_r(t) := \frac{1}{(r-1)!} \sum_{k=0}^{r} (-1)^k \binom{r}{k} (t+\frac{r}{2}-k)_+^{r-1}$$

where $t_+^r = \max(t^r, 0)$, their Fourier transforms being simply

$$M_r^{\wedge}(v) = \frac{1}{\sqrt{2\pi}} \left(\frac{\sin v/2}{v/2}\right)^r \qquad (v \in \mathbb{R}).$$

The M_r are piecewise polynomials of degree r-1 having support [-r/2,r/2]. It is compact, but not contained in $(0,\infty)$, as required.

Let us now construct kernels without the latter deficiency for which Proposition 2 holds by taking appropriate linear combinations of translations of the M_r.

Proposition 3. For $\varepsilon_0 \in \mathbb{R}$ and $r \in \mathbb{N}$, $r \geq 2$, let $a_{\mu r}$, $\mu = 0,1,\ldots,r-1$ be the unique solutions of the linear system

$$(2.2) \qquad \sum_{\mu=0}^{r-1} a_{\mu r} \, (-i(\varepsilon_0+\mu))^j = (1/\sqrt{2\pi}\, M_r^{\wedge})^{(j)}(0) \quad (j=0,1,\ldots,r-1)$$

where $i = \sqrt{-1}$. Then

$$\varphi_r(t) := \sum_{\mu=0}^{r-1} a_{\mu r}\, M_r(t-\varepsilon_0-\mu) \qquad (t \in \mathbb{R})$$

is a polynomial spline of order r satisfying conditions (ii) and (ii)*, having support contained in $[T_0, T_1]$ with $T_0 = \varepsilon_0 - r/2$, $T_1 = \varepsilon_0 + 3r/2 - 1$.

Since $M_r^{\wedge}$ is even, the right side of (2.2) vanishes for j odd. So the solutions $a_{\mu r}$ are all real.

Corollary. In regard to $\varphi_r(t)$ there holds for $f^{(r-1)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$,

$$\|S_W^{\varphi_r} f - f\|_C = O(W^{-r+1-\alpha}) \,. \tag{2.3}$$

For a proof of Proposition 3 see Butzer and Stens [7]. In order to solve equation (2.2), one needs to know the derivatives $(1/M_r^{\wedge})^{(j)}(0)$, at least for small values of r. This can be achieved with the aid of the expansion

$$\left(\frac{v/2}{\sin v/2}\right)^r = \sum_{k=0}^{\infty} b_{kr}\, v^{2k} \qquad (|v| < 2\pi),$$

$$b_{kr} := (-1)^k \frac{(2k+r)!}{r!} \sum_{l=0}^{2k} (-1)^l \frac{r}{r+l} \cdot \frac{T(2k+l,l)}{(2k-k)!\,(2k+l)!},$$

where T(k,l) are the central factorial numbers of the second kind.

These derivatives can be taken from the following table which could readily be enlarged.

Table 1: $(1/\sqrt{2\pi}\, M_r^{\wedge})^{(j)}(0)$: r = 2,3,4,5; j = 0,1,2,3,4.

r \ j	0	1	2	3	4
2	1	0	-	-	-
3	1	0	1/4	-	-
4	1	0	1/3	0	-
5	1	0	5/12	0	9/16

2.3. SPECIFIC EXAMPLES

1. Take $r=2$, $\varepsilon_o=2$, so that $\varepsilon_o > r/2$ and $[T_o,T_1] = [1,4]$. The system (2.2) then reads, noting Table 1, $a_{02} + a_{12} = 1$, $a_{02}(-2i) + a_{12}(-3i) = 0$ for which $a_{02} = 3$, $a_{12} = -2$. Hence

$$\varphi_2(t) = 3M_2(t-2) - 2M_2(t-3) \ ;$$

the associated sampling series (1.5) involves only those samples at $k \in \mathbb{Z}$ for which $k/W \in (t-4/W, t-1/W)$. For example, if t would lie in the interval $(1/W, 2/W)$, the series consists of three terms only, namely for $k = 0,-1,-2$ for which $k/W < t-1/W < t$. If $f' \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, then by (2.3),

$$\| f(t) - \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W})\, \varphi_2(Wt-k) \|_C = \mathcal{O}(W^{-1-\alpha}) \ ,$$

enabling one to predict at least $1/W$ units ahead with error $\mathcal{O}(W^{-1-\alpha})$. If $f'' \in C(\mathbb{R})$ with $\|f''\|_C \leq M$, so that $\alpha=1$, then, according to (2.1), the large-$\mathcal{O}$ constant in (2.3) is $M\, m_2(\varphi_2)/2!$, which is bounded by $15M$ (a fact which cannot be derived theoretically but by employing a computer).

If one would take $r=2$ as above, but $\varepsilon_o > r/2$ arbitrary, then $[T_o,T_1] = [\varepsilon_o - 1, \varepsilon_o + 2]$, and

$$\varphi_{2,\varepsilon_o}(t) = (1+\varepsilon_o)\, M_2(t-\varepsilon_o) - \varepsilon_o M_2(t-\varepsilon_o-1) \ .$$

Here the samples are taken at $k \in \mathbb{Z}$: $k/W \in (t-(\varepsilon_o+2)/W, t-(\varepsilon_o-1)/W)$. In particular, if $\varepsilon_o=8$ and $t \in (2/W, 3/W)$, the series consists of three terms at $k = -5,-6,-7$, for which $k/W < t-7/W < t$. Whereas this is at least $7/W$ units to the left of t, the prediction instant, it was only $1/W$ units in the case of the kernel φ_2. Thus the kernel $\varphi_{2,\varepsilon_o}$ allows one to predict much further ahead with the same number of sampled values (the constant $m_2(\varphi_{2,\varepsilon_o})$ will, however, be much larger than 2·15). In fact, this procedure even enables one to predict or extrapolate a signal arbitrarily far ahead.

2. Now take $r=3$, $\varepsilon_o=2$, so that $[T_o,T_1]=[1/2,\ 11/2]$. The system (2.2) now reads

$$a_{03} + a_{13} + a_{23} = 1$$

$$-2i\,a_{03} - 3i\,a_{13} - 4i\,a_{23} = 0$$

$$4\,a_{03} + 9\,a_{13} + 16\,a_{23} = 1/4$$

which has as solutions $a_{03}=47/8$, $a_{13}=-62/8$, $a_{23}=23/8$. Whence

$$\varphi_3(t) = \frac{1}{8}\,[47M_3(t-2) - 62M_3(t-3) + 23M_3(t-4)]\ , \tag{2.4}$$

the sampling series now consisting of those $k\in\mathbb{Z}$ for which $k/W\in(t-11/2W, t-1/2W)$, thus of five terms for which $k/W<t-1/2W<t$.

In particular, if $\|f'''\|_C\leqslant M$, then $\|f-S_W^{\varphi_3}f\|_C\leqslant M\cdot 54W^{-3}$, noting that $m_3(\varphi_3)/3!\leqslant 54$.

3. Let us finally take $r=4$, $\varepsilon_o=3$, so that $[T_o,T_1]=[1,8]$. By solving a system of four equations in four unknowns one can readily show that

$$\varphi_4(t) = \frac{1}{6}\,[115M_4(t-3) - 256M_4(t-4) + 203M_4(t-5) - 56M_4(t-6)]\ .$$

This time the series consists of seven terms (at most), namely those $k\in\mathbb{Z}$ for which $t-10/W<k/W<t-1/W<t$. In particular, if $\|f^{(4)}\|_C\leqslant M$, then the corresponding rate of approximation can, in comparison with Example 1, be improved to $970\cdot M\,W^{-4}$. By enlarging the $\varepsilon_o\,(\geqslant 4)$ one could again achieve that, instead of being able to predict just (at least) $1/W$ units ahead (from k/W $(<t-1/W)$ to t), one could even predict $(\varepsilon_o-2)/W$ units ahead. Then of course the kernel $\varphi_4(t)$ would take on a different form.

In case $r=5$, $\varepsilon_o=3$ so that $[T_o,T_1]=[4/3,19/2]$, then

$$\varphi_5(t) = \frac{1}{1152}\{36767M_5(t-3) - 108188M_5(t-4)$$
$$+ 127914M_5(t-5) + 14927M_5(t-6)\} .$$

Here seven samples will be needed, the order of approximation being $O(W^{-5})$ provided $\|f^{(5)}\|_C \leq M$. The constant in the order is however large; in fact $m_5(\varphi_5)/5! \leq 3400$.

More generally, if $f^{(r)} \in C(\mathbb{R})$ with $\|f^{(r)}\|_C \leq M$, it is possible to construct a kernel $\varphi_r(t)$ such that the number of samples needed in the convolution sum is just 2r-1 and the associated order is $O(W^{-r})$. However, the constant will be correspondingly large. By this method one cannot increase the approximation order by taking more samples without increasing the order r of $\varphi_r(t)$.

Observe that it is an open question whether there exists a closed form of the solutions $a_{\mu r}$, $\mu = 0,1,\ldots,r-1$ of (2.2). So far the construction can be used in actual practice only for smaller values of r. However, as already the simplest Example 1 shows, even the case r=2 gives the pretty good rate $15\,MW^{-2}$, $W \to \infty$, if $\|f''\|_C \leq M$.

3. TIME-JITTER AND AMPLITUDE ERRORS

It is especially easy to treat time-jitter errors in this frame. These arise when the sample instants are not correctly met but might differ from the exact k/W by δ_k, so that the sampled values are now $f(k/W + \delta_k)$. Here one is interested in estimating the error occurring when f(t) is approximated by the series $S^{\varphi}_{W,\delta}f(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{W} + \delta_k)\varphi(Wt-k)$. This error can be split up as

$$|f(t) - S^{\varphi}_{W,\delta}f(t)| \leq |f(t) - S^{\varphi}_{W}f(t)| + (J_\delta f)(t) ,$$

$$(J_\delta f)(t) := \Big|\sum_{k=-\infty}^{\infty} [f(\tfrac{k}{W}) - f(\tfrac{k}{W} + \delta_k)]\,\varphi(Wt-k)\Big|$$

being the so-called total time-jitter error. It can be esti-

mated in terms of the modulus of continuity, assuming $|\delta_k| \leq \delta$, $k \in \mathbb{Z}$, by

$$|(J_\delta f)(t)| \leq \{\sup_{k\in\mathbb{Z}} \|f(\cdot) - f(\cdot+\delta_k)\|_C\} \{\sup_{t\in\mathbb{R}} \sum_{k=-\infty}^{\infty} |\varphi(t-k)|\}$$

$$\leq m_o(\varphi)\cdot\omega_1(\delta;f;C(\mathbb{R})) \qquad (t \in \mathbb{R}).$$

As a consequence we have

Proposition 4. There hold

a) $$\|f(\cdot) - \sum_{k=-\infty}^{\infty} f(\frac{k}{W} + \delta_k)\varphi(W\cdot - k)\|_C$$

$$\leq \|f - S_W^\varphi f\|_C + m_o(\varphi)\cdot\omega_1(\delta;f;C(\mathbb{R})) \qquad (f \in C(\mathbb{R})).$$

b) If $f \in \mathrm{Lip}(\alpha;C(\mathbb{R}))$, $0 < \alpha \leq 1$, then, provided $\delta \leq 1/W$, $W \geq 1$, the order in part a) is given by $O(W^{-\alpha})$.

Note that this order cannot be improved even if f possesses derivatives of arbitrary order. On the other hand, if $W^{-1} \leq \delta$, then the order in part a) is $O(\delta^\alpha)$.

Thus the prediction series $S_W^\varphi f(t)$ exemplifies stability with respect to the sample points, a small error in each of the sample points produces a correspondingly small error in the prediction series.

There is also the amplitude error $(A_\varepsilon f)(t)$, arising if the exact sample values $f(k/W)$ are not at one's disposal but only falsified values $\overline{f}(k/W)$, differing by $\varepsilon_k := f(k/W) - \overline{f}(k/W)$ with $|\varepsilon_k| \leq \varepsilon$, $k \in \mathbb{Z}$, for some $\varepsilon > 0$. This falsification may be due to rounding-off, quantization or noise. The total amplitude error

$$|(A_\varepsilon f)(t)| := |(S_W^\varphi f)(t) - (S_W^\varphi \overline{f})(t)| \leq \varepsilon\, m_o(\varphi) ,$$

so that the error occurring when $f(t)$ is approximated by $S_W^\varphi \overline{f}(t)$ can be estimated by

Proposition 5. There hold

a) $$\| \sum_{k=-\infty}^{\infty} \overline{f}(\tfrac{k}{W})\, \varphi(W\cdot - k) - f(\cdot) \|_C \leq \| S_W^{\varphi} f - f \|_C + \varepsilon\, m_o(\varphi) \ .$$

b) If $f \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$, then the order in part a) is $O(W^{-\alpha})$ provided $\varepsilon \leq W^{-1}$, $W \geq 1$.

Thus the prediction series also illustrates stability with respect to the function values, a uniformly small change in the function values at all of the sample points produces a correspondingly small change in the prediction series.

4. PREDICTION OF DERIVATIVES $f^{(s)}$ BY SAMPLES OF f

Let us now consider the prediction of derivatives $f^{(s)}$ of a signal f by samples of f only, in terms of derivatives of $S_W^{\varphi} f$, i.e., of

$$(S_W^{\varphi})^{(s)} f(t) = (\tfrac{d}{dt})^s (S_W^{\varphi} f)(t) = W^s \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W}) \varphi^{(s)}(Wt-k)$$

$$(s \in \mathbb{N}_o) \ .$$

Proposition 6. Let $\varphi \in C_{OO}^{(s)}(\mathbb{R})$ satisfy (ii), (ii)* for some $r \geq s+1$ with $s \in \mathbb{N}_o$, $r \in \mathbb{N}$. Then $(S_W^{\varphi})^{(s)} f$ defines a family of bounded, linear operators mapping $C^{(s)}(\mathbb{R})$ into $C(\mathbb{R})$, with norm

$$\| S_W^{\varphi(s)} \|_{[C^{(s)}, C]} \leq \frac{m_s(\varphi^{(s)}_{\cdot})}{s!} \qquad (W > 0) \ .$$

Further,

$$\| (S_W^{\varphi})^{(s)} g - g^{(s)} \|_C \leq \frac{m_r(\varphi^{(s)})}{r!} \| g^{(r)} \|_C \ W^{-r+s}$$

$$(g \in C^{(r)}(\mathbb{R});\ W > 0) \ ,$$

$$\| (S_W^{\varphi})^{(s)} f - f^{(s)} \|_C \leq K \omega_{r-s}(W^{-1}; f^{(s)}; C(\mathbb{R}))$$

$$(f \in C^{(s)}(\mathbb{R});\ W > 0).$$

In particular, one has for $f \in C^{(s)}(\mathbb{R})$,

$$\lim_{W \to \infty} \left(\frac{d}{dt}\right)^s (S_W f)(t) = \left(\frac{d}{dt}\right)^s f(t)$$

uniformly in $t \in \mathbb{R}$; if $f^{(r-1)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, then $\| (S_W^{\varphi})^{(s)} f - f^{(s)} \|_C = O(W^{-r+1+s-\alpha})$, $W \to \infty$.

These results would enable one to predict the speed or acceleration of flying objects.

Let us consider an example. For this purpose we begin with example 2 of Section 2.3 where $r=3$, $\varepsilon_o = 2$, $[T_0, T_1] = [1/2, 5\ 1/2]$ and $\varphi_3(t)$ is given by (2.4). Let us apply Proposition 6 to $\varphi_3(t)$ in the case $s=1$. Noting that

$$M_r'(t) = M_{r-1}(t + 1/2) - M_{r-1}(t - 1/2) \qquad (t \in \mathbb{R}),$$

$$\varphi_3'(t) = \frac{1}{8}[47\, M_2(t - 3/2) - 109\, M_2(t - 5/2) + 85\, M_2(t - 7/2)$$

$$- 23\, M_2(t - 9/2)].$$

Here $\varphi_3' \in C_{OO}(\mathbb{R})$. In particular, if $f^{(2)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$, then

$$\| W \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W}) \varphi_3'(Wt - k) - f'(t) \|_C = O(W^{-1-\alpha}) .$$

This result enables one to predict the derivative $f'(t)$ in terms of a series which involves just five samples of f which all lie to the left of $t - 1/2W < t$. If $\| f^{(3)} \|_C \leq M$, then the large-O constant is given by $M\, m_3(\varphi_3')/3!$.

5. PREDICTION OF RANDOM SIGNALS

Signal functions are often of random character, random signals play an important role in signal processing and sampling prediction. For this purpose one often uses stochastic processes which are stationary in the weak sense as a model. Given a probability space (Ω,A,P), a real-valued stochastic (random) process, namely an A-measurable function $X = X(t) = X(t,\omega)$ of $\omega \in \Omega$ for each $t \in \mathbb{R}$, is said to be weak sense stationary (w.s.s.), if its autocorrelation function

$$R_X(t,t+\tau) := \int_\Omega X(t,\omega)X(t+\tau,\omega)dP(\omega)$$

is independent of $t \in \mathbb{R}$, i.e., $R_X(t,t+\tau) = R_X(\tau)$. Here X is assumed to belong to $L^2(\Omega)$, i.e., the norm

$$\|X(t,\cdot)\|_2 := \{\int_\Omega |X(t,\omega)|^2 dP(\omega)\}^{1/2} := \{E[|X(t)|^2]\}^{1/2} \tag{5.1}$$

is finite for all $t \in \mathbb{R}$. Note that $R_X(\tau)$ is even in τ, $\|R_X\|_C = R_X(0)$, and the norm (5.1) is independent of t, equalling $\|R_X\|_C^{1/2}$.

For the prediction of such a process $X \in L^2(\Omega)$ let us consider the prediction series

$$(S_W^\varphi X)(t,\omega) := \sum_{k=-\infty}^{\infty} X(\frac{k}{W},\omega)\varphi(Wt-k) \qquad (t \in \mathbb{R}).$$

It defines a family of bounded, linear operators from $L^2(\Omega)$ into itself, with

$$\|S_W^\varphi X(t,\cdot)\|_2 = \{\sum_{k,\mu=-\infty}^{\infty} R_X(\frac{k-\mu}{W})\varphi(Wt-k)\varphi(Wt-\mu)\}^{1/2}$$

$$\leq R_X(0)^{1/2} m_0(\varphi) = m_0(\varphi)\|X\|_2 .$$

Proposition 7. Let $\varphi \in C_{OO}(\mathbb{R})$ satisfy (ii), (ii)* with r-1 replaced by 2(r-1) for some $r \in \mathbb{N}$. If X is a w.s.s. process with $X^{(r)} \in L^2(\Omega)$, then

$$\{E[|S_W X - X|^2]\}^{1/2} \leq \left\{\frac{(m_0(\varphi)+3)m_{2r}(\varphi)}{2r!}\right\}^{1/2} \cdot$$

$$\cdot \frac{\{E[|X^{(r)}|^2]\}^{1/2}}{W^r} \quad (t \in \mathbb{R};\ W > 0).$$

There exists a constant $K > 0$ such that for any w.s.s. process $X \in L^2(\Omega)$, continuous in the mean,

$$\{E[|S_W X - X|^2]\}^{1/2} \leq K\omega_r(W^{-1}; X; L^2(\Omega)) \quad (t \in \mathbb{R};\ W > 0).$$

Regarding proofs in the case of random processes, one reduces the matter to the deterministic case, namely from assertions dealing with the random process X to those concerned with the deterministic function R_X, by the following basic connections:

i) the rth derivative (in mean) $X^{(r)}$ exists at $t_0 \in \mathbb{R}$ if $R_X \in C^{(2r)}(\mathbb{R})$;

ii) $$\omega_s(\delta; X; L^2(\Omega)) = \{\omega_{2s}(\delta; R_X; C(\mathbb{R}))\}^{1/2};$$

iii) $$E[|S_W^{\varphi} X - X|^2] = R_X(0) - 2\sum_{k=-\infty}^{\infty} R_X(\frac{k}{W} - t)\cdot\varphi(Wt-k) +$$

$$+ \frac{1}{2\pi}\sum_{k,\mu=-\infty}^{\infty} R_X(\frac{k-\mu}{W})\varphi(Wt-k)\cdot\varphi(Wt-\mu)$$

$$\leq (m_0(\varphi)+3)\sup_{u\in\mathbb{R}} |(S_W^{\varphi}\tau_u R_X)(t) - (\tau_u R_X)(t)$$

where $(\tau_u f)(t) = f(t-u)$.

The following table gives the best possible order of approximation according to Proposition 7 for the kernels φ_r of Section 2.3.

Table 2

Kernels	φ_2	φ_3	φ_4
Orders	$O(W^{-1})$	$O(W^{-1})$	$O(W^{-2})$

6. THE APPROACHES OF WIENER AND KREIN IN COMPARISON

Let us finally roughly compare the present approach with the work of Wiener [26] and M.G. Krein [14] (1954) in the matter. For this purpose let us express our convolution sum (1.5), thinking of the commutativity of convolution products, as

$$(6.1) \qquad \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W})\varphi(W(t-\tfrac{k}{W})) \cong \sum_{k=-\infty}^{\infty} f(t-\tfrac{k}{W})\varphi(k) \ .$$

Although the two sums are generally not equal (except under special conditions, see [6]), it is nevertheless also possible to set up our approach to prediction for the right hand one (using parallel arguments, see [7]). If φ has compact support in $[T_0,T_1]$, then the right sum only runs over all k with $T_0 < k < T_1$ so that one can see from it right off where the prediction points lie, namely to the left of t at $\ldots\ t-k/W,\ \ldots,\ t-1/W$.

Now Wiener's aim was to predict the future at time t from the <u>whole</u> past $f(u)$: $-\infty \leqslant u \leqslant t-\varepsilon_0$, $\varepsilon_0 > 0$ prescribed, in a non-discrete setting (where our sum is replaced by an integral). In fact, his aim was to minimize as a function of the kernel Φ the mean-square error

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \left| f(t) - \int_0^{\infty} f(t-\varepsilon_0-u)\Phi(u)du \right|^2 dt \ .$$

He showed that his problem amounts to solving the integral equation

$$R(t) = \int_0^\infty R(t-\varepsilon_o-u)\Phi(u)du \qquad (t \geq \varepsilon_o), \tag{6.2}$$

where R is the auto-correlation function,

$$R(t) := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t+u)f(u)du \ .$$

Now it is to be emphasized that the equation (6.2) only holds for $t \geq \varepsilon_o$, and not for all $t \in \mathbb{R}$. So it is not solvable by routine Fourier methods. The so-called Wiener-Hopf technique (of 1931) has to be employed. In this respect Wiener notes [26; p.65] that there are "limitations and precautions which must be observed" in solving (6.2) and illustrated his method by several examples. In fact, Dym and McKean add [8 , p.92] "it is not clear how to proceed much further in the present direction save by examples". In any case, for a formal derivation as well as exellent coverage of the matter see the treatment in [8] pp. ix, 2-5, 82-96. For good information concerning effective computation see Lee [15], pp. 354-439, Kailath [11], also Noble [20]. For further literature see the extensive reference lists in the commentaries on the work of Wiener by P. Masani, H. Salehi, T. Kailath, P.S. Muhly and G. Kallianpur in [27].

Now the problem treated in this paper is actually that of predicting the future from only a <u>part</u> $f(u) : -T \leq u \leq t-\varepsilon_o$ of the past, in the case of discrete u. Especially in the non-discrete case was this problem solved by Krein [14]; it required even much heavier machinery than that of the (Kolmogorov)-Wiener problem, namely a so-called "method of strings" in the context of operator theory, complex function theory and Hardy functions, wave and spectral functions,all combined with the theory of spaces of entire functions (in the sense of de Branges [3]). This theory was carried out in expert fashion by Dym-McKean [8] pp. 146-278, applied to the actual prediction problem on pp. 279-91; there is an overview on pp. 5-9. However, as these authors write (p. X): "it is hoped that electrical engineers and other people dealing with the practical aspects of prediction will find in it

[our volume] something to interest them too, though it has to be confessed that the computations to which the theory leads are usually difficult to perform and that their statistical content is often obscure; in fact, much remains to be done to clarify the statistical content of the whole subject."

The methods needed to prove the results of this overview, presented in [7] are, in comparison, elementary indeed. Thus Proposition 1 is based upon a simple application of the Poisson summation formula of Fourier analysis, Proposition 2 upon Taylor's formula and elementary approximation theory, while Proposition 3 uses elementary results on B-splines (together with some new results on central factorial numbers). Proposition 7 shows that the treatment of random prediction theory can essentially be reduced to that of the deterministic situation so that no separate approach is necessary.

Most of the results discussed in this overview arose from questions posed by electrical and communication engineers in the course of some seven years of cooperative work. It is to be expected that they can also follow the proofs. The fact that the matter is indeed easy to apply has been demonstrated with the various examples.

LITERATURE

1. G.E.T. BOX and G.M. JENKINS: Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, CA, 1976 (rev. edition).

2. J.L. BROWN: Uniform prediction of bandlimited processes from past samples. IEEE Trans. Inform. Theory IT-18 (1972), 662-664.

3. L. DE BRANGES: Hilbert Spaces of Entire Functions. Prentice Hall, Englewood Cliffs, N.J., 1968.

4. P.L. BUTZER, W. ENGELS, S. RIES and R.L. STENS: The Shannon sampling series and the reconstruction of signals in terms of linear, quadratic and cubic splines. SIAM J. Appl. Math. 46 (1986), 299-323.

5. P.L. BUTZER, S. RIES and R.L. STENS: Approximation of continuous and discontinuous functions by generalized sampling series. J. Approx. Theory 50 (1987), 25-39.

6. P.L. BUTZER, W. SPLETTSTOESSER and R.L. STENS: The sampling theorem and linear prediction in signal analysis. Jahresber. Deutsch. Math.-Verein. 90 (1988), (in print).

7. P.L. BUTZER and R.L. STENS: Prediction of non-band-limited signals from past samples in terms of splines of low degree. Math. Nachr. (in print).

8. H. DYM and H.P. McKEAN: Gaussian Processes, Function Theory, and the Inverse Spectral Problem. Academic Press, New York, 1976, xii +333 pp.

9. G. FIX and G. STRANG: Fourier analysis of the finite element method in Ritz-Galerkin theory. Studies Appl. Math. 48 (1969), 268-273.

10. S. HAYKIN: Introduction to Adaptive Filters. MacMillan, New York and London, 1984, xii +217 pp.

11. T. KAILATH: Lectures on Linear Least-Square Estimation. CISM Courses and Lectures No. 140. Springer, Wien / New York 1976, ii +169 pp.

12. A.N. KOLMOGOROV: Interpolation and Extrapolation von stationären zufälligen Folgen. Isz. Akad. Nauk SSSR. Ser. Math. 5 (1941), 3-14.

13. M.G. KREIN: On a problem of extrapolation of A.N. Kolmogorov. Dokl. Akad. Nauk SSSR 46 (1954), 306-309.

14. M.G. KREIN: On a fundamental approximation problem in the theory of extrapolation and filtration of stationary random processes. Dokl. Akad. Nauk SSSR

94 (1954), 13-16 [Engl. transl.: Amer. Math. Soc. Selected Transl. Math. Statist. Prob. 4 (1964), 127-131].

15. Y.W. LEE: Statistical Theory of Communication. John Wiley, New York, 1960, xvii + 509 pp.

16. J. MAKHOUL: Linear prediction: A tutorial review. Proc. IEEE 63 (1975), 561-580.

17. J.D. MARKEL and H.H. GRAY, JR.: Linear Prediction of Speech. Springer, New York, 1982.

18. D.H. MUGLER and W. SPLETTSTOESSER: Difference methods and round-off error bounds for the prediction of bandlimited functions from past samples. Frequenz 39 (1985), 182-187.

19. D.H. MUGLER and W. SPLETTSTOESSER: Linear prediction from samples of a function and its derivatives. IEEE Trans. Inform. Theory IT-33 (1987), 360-366.

20. B. NOBLE: Methods based on the Wiener-Hopf Technique for the Solution of Partial Differential Equations. Pergamon, London, 1958, X + 246 pp.

21. S. RIES and R.L. STENS: Approximation of generalized sampling series. In: Constructive Theory of Functions (Proc. Conf. Varna, Bulgaria, May 27 - June 2, 1984; Eds. Bl. Sendov et al.). Publ. House Bulg. Acad. Sci., Sofia, 1984, (939 pp.), pp. 746-756.

22. W. SPLETTSTOESSER: Bandbegrenzte und effektiv bandbegrenzte Funktionen und ihre Praediktion aus Abtastwerten. Habilitationsschrift, RWTH Aachen, 1981, 65 pp.

23. W. SPLETTSTOESSER: On the prediction of bandlimited signals from past samples. Information Sci. 28 (1982), 115-130.

24. W. SPLETTSTOESSER: Lineare Praediktion von nicht bandbegrenzten Funktionen. Z. Angew. Math. Mech. 64 (1984), T 939 - T 395.

25. L.A. WAINSTEIN and V.D. ZUBAKOV: Extraction of Signals from Noise. Prentice-Hall, Englewood Cliffs, N.J., 1962.

26. N. WIENER: Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. M.I.T. Press, Cambridge, MA., 1949.

27. N. WIENER: Collected Works, Vol. III. (ed. by P.R. Masani), MIT Press, Cambridge, MA., 1981.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

SOME NEW INEQUALITIES FOR THE ZEROS OF LAGUERRE POLYNOMIALS*

LUIGI GATTESCHI

ABSTRACT: *It is shown that certain approximations for the zeros* $\lambda_{n,k}^{(\alpha)}$ *of the Laguerre polynomials* $L_n^{(\alpha)}(x)$, $\alpha > -1$, *are upper or lower bounds. These bounds involve the zeros of the Bessel function* $J_\alpha(x)$ *or the zeros of the Airy function* Ai (x) *and are obtained by using the Sturm comparison theorem.*

1. INTRODUCTION

In a recent paper [3] we have obtained some inequalities for the zeros of Jacobi polynomials. In this paper we will apply the same technique to derive bounds for the zeros $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, of the Laguerre polynomials $L_n^{(\alpha)}(x)$, $\alpha > -1$.

To this purpose we need the well-known Sturm comparison theorem in the following form given by Szegö [5, p. 19].

THEOREM 1.1 (Sturm's comparison theorem). *Let* $f(x)$ *and* $F(x)$ *be functions continuous in* $x_0 < x < X_0$, *with* $f(x) \leqslant F(x)$. *Let the functions* $y(x)$ *and* $Y(x)$, *both not identically zero, satisfy the differential equations*

$$y'' + f(x)\, y = 0 \quad , \quad Y'' + F(x)Y = 0, \tag{1.1}$$

respectively. Let x' *and* x'', $x' < x''$, *be two consecutive zeros of* $y(x)$. *Then the function* $Y(x)$ *has at least one zero in the interval* $x' < x < x''$ *provided* $f(x) \not\equiv F(x)$ *in* $[x', x'']$.

The statement also holds for $x' = x_0$ $[y(x_0 + 0) = 0]$ *if the additional condition*

* *This work was supported by the Consiglio Nazionale delle Ricerche of Italy and by the Ministero della Pubblica Istruzione of Italy.*

(1.2) $$\lim_{x \to x_0 + 0} [y'(x)\, Y(x) - y(x)\, Y'(x)] = 0$$

is satisfied (similarly for $x'' = X_0$*).*

The differential equations that we shall use as comparison equations are the ones used by Erdélyi [2] in deriving uniform asymptotic approximations for the Laguerre polynomials. Such equations can also be obtained by applying Olver's theory [4] to the asymptotic study of the Laguerre differential equation near the singularity $x = 0$ and near the turning point $x = 4n + 2\alpha + 2$.

Let us recall the following inequalities and asymptotic results.

THEOREM 1.2 (see Szegö [5], p. 127). *Let* $\alpha > -1$ *and let* $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, *be the zeros of* $L_n^{(\alpha)}(x)$ *in increasing order. Then*

(1.3) $$\lambda_{n,k}^{(\alpha)} > \frac{j_{\alpha,k}^2}{\nu}, \quad \nu = 4n + 2\alpha + 2,$$

for $k = 1, 2, \ldots, n$ *and where* $j_{\alpha,k}$ *is the* k*-th positive zero of the Bessel functions* $J_\alpha(x)$. *Furthermore, we have for a fixed* k, *as* $n \to \infty$,

(1.4) $$\lambda_{n,k}^{(\alpha)} = \frac{j_{\alpha,k}^2}{\nu} + O(n^{-2})$$

Tricomi [7] gave an improvement of (1.4), but its validity remains still restricted to the case of a fixed k.

THEOREM 1.3 (see Szegö [5], p. 131). *Let* a_k, $k = 1, 2, \ldots$, *be the zeros in decreasing order* $0 > a_1 > a_2 \ldots$, *of the Airy function* Ai (x).
If $|\alpha| \geqslant 1/4$, $\alpha > -1$, *then*

(1.5) $$\lambda_{n,k}^{(\alpha)} < [\nu^{1/2} + 2^{-1/3}\, \nu^{-1/6}\, a_{n-k+1}]^2,$$

for $k = 1, 2, \ldots, n$ *and where* ν *has the same meaning as in* (1.3). *Furthermore, we have for fixed* $n-k$, *as* $n \to \infty$,

(1.6) $$\lambda_{n,k}^{(\alpha)} = [\nu^{1/2} + 2^{-1/3}\, \nu^{-1/6}\, (a_{n-k+1} + \varepsilon_n)]^2,$$

where $\lim_{n \to \infty} \varepsilon_n = 0$.

Here the notations for the Airy function Ai(x) and for the zeros a_k are different from the ones used by Szegö; he uses $i_k = -3^{1/3}\, a_k$ instead of a_k.

A simplified form of a formula due to Tricomi [8] is given by the following:

THEOREM 1.4. *Let $\alpha > -1$ and let $x_{n,k}^{(\alpha)}$ be the root of the equation*

$$x - \sin x = \frac{4n - 4k + 3}{\nu}\,\pi \tag{1.7}$$

Then we have

$$\lambda_{n,k}^{(\alpha)} = \nu \cos^2 (x_{n,k}^{(\alpha)}/2) + O(n^{-1}), \tag{1.8}$$

for the zeros which belong to the interval $(a\nu, b\nu)$, where a and b, $0 < a < b < 1$, are fixed positive constants.

Recently, Temme [6] has obtained an interesting asymptotic representation of $\lambda_{n,k}^{(\alpha)}$ which involves the zeros of the Hermite polynomial $H_n(x)$. This representation gives good numerical results expecially for large values of the parameter α.

2. AN UPPER BOUND FOR THE ZEROS OF $L_n^{(\alpha)}(x)$

We shall refer throughout this paper to the differential equation

$$\frac{d^2y}{dt^2} + \left[\frac{\nu^2}{4}\left(\frac{1}{t} - 1\right) + \frac{1-\alpha^2}{4t^2}\right] y = 0, \tag{2.1}$$

$$\nu = 4n + 2\alpha + 2,\ \alpha > -1,$$

which is satisfied by

$$y(t) = e^{-\frac{1}{2}\nu t}\,(\nu t)^{\frac{1}{2}(\alpha+1)}\,L_n^{(\alpha)}(\nu t). \tag{2.2}$$

Now we observe that the function

$$z(t) = \left(\frac{f}{f'}\right)^{1/2} J_\alpha [f(t)] \tag{2.3}$$

satisfies the differential equation

$$\frac{d^2z}{dt^2} + F(t)\,z = 0, \tag{2.4}$$

where

$$F(t) = \frac{1}{2}\frac{f'''}{f'} - \frac{3}{4}\left(\frac{f''}{f'}\right)^2 + \left(\frac{1}{4} - \alpha^2\right)\left(\frac{f'}{f}\right)^2 + f'^2. \tag{2.5}$$

The equation (2.4) can be used, by assuming

$$(2.6)\qquad f(t) = \frac{\nu}{2}\,[(t-t^2)^{1/2} + \arcsin t^{1/2}],\ 0 < t < 1,$$

as a comparison equation to derive, by means of Sturm's method, inequalities for the zeros $\lambda_{n,k}^{(\alpha)}, k = 1, 2, \ldots, n$, of $L_n^{(\alpha)}(x)$.

This requires the study of the function

$$(2.7)\qquad G(t,\alpha) = F(t) - \left[\frac{\nu^2}{4}\left(\frac{1}{t} - 1\right) + \frac{1-\alpha^2}{4t^2}\right],$$

for $0 < t \leqslant 1$, or, more simply, of the function

$$(2.8)\qquad G^*(t,\alpha) = \frac{t\,G(t,\alpha)}{1-t},$$

which is analytic at $t = 0$. Indeed, it is easily seen that

$$(2.9)\qquad G^*(t,\alpha) = \frac{1/4 - \alpha^2}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} + \frac{3-8t-4(1-\alpha^2)(1-t)^2}{16t(1-t)^3}$$

and that

$$(2.10)\qquad G^*(t,\alpha) = \frac{\alpha^2 - 1}{6} + \frac{13\alpha^2 - 37}{60}\,t + O(t^2).$$

LEMMA 2.1. *Let* $\alpha^2 = 1$. *Then* $G^*(t, \pm 1) \leqslant 0$ *for* $0 \leqslant t < 1$. *The equality sign holds if and only if* $t = 0$.

We have

$$(2.11)\qquad 4t\,G^*(t, \pm 1) = \frac{-3t}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} + \frac{3-8t}{4(1-t)^3}.$$

First we prove that the property $G^*(t, \pm 1) < 0$, which is trivial for $3/8 \leqslant t < 1$, holds in the interval $1/16 \leqslant t < 1$. Indeed, by observing that the function

$$u(t) = \frac{t}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2}$$

increases in $0 < t < 1$ and that the function

$$v(t) = \frac{3 - 8t}{(1 - t)^3}$$

decreases in $1/16 < t < 1$, we obtain for $1/16 \leqslant t < 1$,

$$4t\, G^*(t, \pm 1) \leqslant -3\, u\left(\frac{1}{16}\right) + \frac{1}{4} v\left(\frac{1}{16}\right) < 0.$$

For the remaining interval $0 < t < 1/16$ we use the inequality

$$(2.12) \qquad \frac{1}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} > \frac{1}{4t}\left(1 + \frac{t}{3}\right), \quad 0 < t < 1,$$

and we set

$$(2.13) \qquad \frac{1}{(1 - t)^3} = 1 + 3t + a(t)\, t^2,$$

where

$$a(t) = \left[\frac{1}{(1-t)^3} - 1 - 3t\right] \frac{1}{t^2}$$

is an increasing function in $0 < t < 1$. Then from (2.11) we obtain

$$4\, G^*(t, \pm 1) < \frac{-3}{4t}\left(1 + \frac{t}{3}\right) + \left(\frac{3}{4t} - 2\right)\, [1+3t+a(t)t^2]$$

$$= 3\left[\frac{a(t)}{4} - 2\right] t - 2\, a(t)\, t^2,$$

i.e.

$$4\, G^*(t, \pm 1) < 3\left[\frac{a(t)}{4} - 2\right] t,$$

which, being $a(t) < a(1/16) = 6.689\ldots$ if $0 < t < 1/16$, completes the proof of the lemma.

LEMMA 2.2. *Let $G(t, \alpha)$ be the function defined by* (2.7). *In the interval $0 < t < 1$, $G(t, \alpha)$ has at least one zero if $\alpha^2 > 1$ and is negative if $\alpha^2 \leqslant 1$.*

For the proof we use the function $G^*(t, \alpha)$, defined by (2.8), which is continuous on $0 \leqslant t < 1$. From (2.9) and (2.10) we obtain, if $\alpha^2 > 1$,

$$\lim_{t \to 1-0} G^*(t, \alpha) = -\infty$$

and

$$\lim_{t \to 0+0} G^*(t, \alpha) = \frac{\alpha^2 - 1}{6} > 0,$$

respectively. Therefore, the first part of the lemma is proved.

We now observe that $G^*(t, \alpha)$ increases with respect to the parameter α^2. Indeed from (2.8) we have

$$\frac{\partial G^*}{\partial(\alpha^2)} = \frac{-1}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} + \frac{1}{4t(1-t)}$$

and setting $t^{1/2} = \sin\vartheta$ we find

$$\frac{\partial G^*}{\partial(\alpha^2)} = \frac{-1}{[\sin\vartheta\cos\vartheta + \vartheta]^2} + \frac{1}{4\sin^2\vartheta\cos^2\vartheta}$$

$$= \frac{-1}{\sin^2 2\vartheta\left[\frac{1}{2} + \frac{1}{2}\frac{2\vartheta}{\sin 2\vartheta}\right]^2} + \frac{1}{\sin^2 2\vartheta} > 0,$$

for $0 < \vartheta < \pi/2$. Hence, by using Lemma 2.1,

$$G^*(t, \alpha) \leqslant G^*(t, \pm 1) < 0, \quad 0 < t < 1,$$

when $\alpha^2 \leqslant 1$.

The property $G(t, \alpha) < 0$, *if* $0 < t < 1$ and $-1 < \alpha \leqslant 1$, established by Lemma 2.2, enables us to compare the zeros of the solution $y(t)$ of the equation (2.1) with the positive zeros of the function $z(t)$ defined by (2.3) and (2.6).

We notice that

$$\left(\frac{f}{f'}\right)^{1/2} = (2t)^{1/2}\left(1 + \frac{1}{6}t + \ldots\right), \quad 0 < t < 1.$$

Therefore, by means of the series representation of $J_\alpha(z)$, we obtain

(2.14) $\quad z(t) = t^{\frac{1}{2}(\alpha+1)}(a_0 + a_1 t + \ldots), \quad 0 < t < 1,$

with $a_0 \neq 0$.

Now, let $-1 < \alpha \leqslant 1$ and let $\tau_{n,k} \equiv \tau_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots$, be the zeros of $z(t)$ in $0 \leqslant t < 1$. We have

(2.15) $\tau_{n,0} = 0, \quad \frac{\nu}{2}\,[(\tau_{n,k} - \tau_{n,k}^2)^{1/2} + \arcsin \tau_{n,k}] = j_{\alpha,k},$

$k = 1, 2, \ldots, n.$

This follows by observing that $f(t)$ is a positive increasing function in $0 \leqslant t \leqslant 1$ varying from 0 to $\nu\pi/4$ and that (see Watson [9], p. 497) the number of the positive zeros of $x^{-\alpha}J_\alpha(x)$ between 0 and $n\pi + \alpha/2 + \pi/4$ is exactly n.

The condition (1.2), which is required when we apply Theorem 1.1 to the interval $[0, \tau_{n,1}]$, is satisfied if $\alpha > -1$ since, from (2.2),

$$y(t) = t^{\frac{1}{2}(\alpha+1)}\,(b_0 + b_1 t + \ldots)$$

and consequently, by using (2.14), we find that

$$y(t)\,z'(t) - z(t)\,y'(t) = 0\,(t^{\alpha+1}), \qquad t \to 0.$$

We may conclude that each interval

$$\tau_{n,k-1} < t < \tau_{n,k}, \quad k = 1, 2, \ldots, n,$$

contains exactly one zero $t_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, of $L_n^{(\alpha)}(\nu t)$.
Or, in other words: for the zeros $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, of $L_n^{(\alpha)}(x)$, if $-1 < \alpha \leqslant 1$, we have

(2.16) $\lambda_{n,k}^{(\alpha)} < \nu\,\tau_{n,k}^{(\alpha)}, \quad k = 1, 2, \ldots, n.$

This is the main result of this section. It can be stated in the following form.

THEOREM 2.1. *Let $-1 < \alpha \leqslant 1$. Let $x_{n,k}^{(\alpha)}$ be the root of the equation*

$$(2.17) \qquad x - \sin x = \pi - \frac{4\,j_{\alpha,k}}{\nu}, \quad \nu = 4n + 2\alpha + 2,$$

where $j_{\alpha,k}$ is the k-th positive zero of $J_\alpha(x)$. Then the k-th zero $\lambda_{n,k}^{(\alpha)}$ of $L_n^{(\alpha)}(x)$ satisfies the inequality

$$(2.18) \qquad \lambda_{n,k}^{(\alpha)} < \nu \cos^2 (x_{n,k}^{(\alpha)}/2), \quad k = 1, 2, \ldots, n.$$

Indeed, by setting $t^{1/2} = \cos\vartheta$, the equation $f(t) = j_{\alpha,k}$ becomes

$$2\vartheta - \sin 2\vartheta = \pi - \frac{4\,j_{\alpha,k}}{\nu}.$$

Thus, for the zeros $\tau_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, defined by (2.15), we have

$$\tau_{n,k}^{(\alpha)} = \cos^2 (x_{n,k}^{(\alpha)}/2)$$

3. INEQUALITIES INVOLVING THE ZEROS OF THE AIRY FUNCTION

We shall use in this section, as comparison equation, the differential equation

$$(3.1) \qquad \frac{d^2u}{dt^2} + H(t)\, u = 0,$$

with

$$(3.2) \qquad H(t) = \frac{1}{2}\frac{h'''}{h'} - \frac{3}{4}\left(\frac{h''}{h'}\right)^2 - h\,h'^2,$$

which is satisfied by

$$(3.3) \qquad u(t) = [h'(t)]^{-1/2}\, \mathrm{Ai}\,[h(t)],$$

where $\mathrm{Ai}(x)$ is the Airy function of first kind.

It will be useful to recall some properties of $\mathrm{Ai}\,(x)$ and their zeros.

The function $\mathrm{Ai}\,(x)$ has no positive zero and infinitely many negative zeros, it is positive for $x > 0$ and $\mathrm{Ai}'(x) \to 0$ as $x \to \infty$. More precisely, we have as $x \to +\infty$.

$$(3.4) \quad \begin{cases} \mathrm{Ai}\,(x) \sim \dfrac{1}{2}\,\pi^{-1/2}\, x^{-1/4} \exp\left(-\dfrac{2}{3}\,x^{3/2}\right), \\ \mathrm{Ai}'(x) \sim \dfrac{1}{2}\,\pi^{-1/2}\, x^{1/4} \exp\left(-\dfrac{2}{3}\,x^{3/2}\right). \end{cases}$$

LEMMA 3.1. *Let a_k, $k = 1, 2, \ldots$, be the zeros in decreasing order of* $\mathrm{Ai}\,(x)$. *Then*

$$(3.5) \qquad -\left[\frac{3}{8}\left(4k - \frac{5}{6}\right)\pi\right]^{2/3} < a_k < -\left[\frac{3}{8}\,(4k-1)\,\pi\right]^{2/3},$$

$$k = 1, 2, \ldots .$$

For the proof we first consider the cylinder function

$$C_\alpha(x) = J_\alpha(x)\cos\varphi - Y_\alpha(x)\sin\varphi,$$

with $0 \leqslant \varphi < \pi$ and where $Y_\alpha(x)$ is the Bessel function of second kind. The positive zeros $c_{\alpha,k}$, $k = 1, 2, \ldots$, of $C_\alpha(x)$ satisfy, when $-1/2 < \alpha \leqslant 1/2$, the inequalities of Schafheitlin [9, p. 490].

$$(3.6)\qquad k\,\pi-\frac{\pi}{4}+\frac{1}{2}\,\alpha\,\pi-\varphi<c_{\alpha,k}<k\,\pi-\frac{\pi}{8}+\frac{1}{4}\,\alpha\,\pi-\varphi,$$

$$-\frac{1}{2}<\alpha\leqslant\frac{1}{2},\quad k=1,2,\ldots.$$

Next, by using the representation of Airy's function in terms of Bessel functions

$$\mathrm{Ai}\,(-x)=\frac{1}{3}\sqrt{x}\,[J_{1/3}(\xi)+J_{-1/3}(\xi)],\ \xi=\frac{2}{3}\,x^{3/2},$$

and the formula

$$J_{-1/3}(z)=J_{1/3}(z)\cos\pi/3-Y_{1/3}(z)\sin\pi/3,$$

we obtain

$$\mathrm{Ai}\,(-x)=\sqrt{\frac{x}{3}}\,[J_{1/3}(\xi)\cos\pi/6-Y_{1/3}(\xi)\sin\pi/6],\ \xi=\frac{2}{3}\,x^{3/2}.$$

Then, (3.6) with $\varphi=\pi/6$ yields

$$\frac{4k-1}{4}\,\pi<\frac{2}{3}\,(-a_k)^{3/2}<\frac{24k-5}{24}\,\pi,\quad k=1,2,\ldots,$$

that is the inequalities (3.5).

In order to compare the equation (3.1) with the Laguerre equation (2.1), we assume in (3.2)

$$(3.7)\qquad h(t)=\begin{cases}-\nu^{2/3}\left[\dfrac{3}{4}\,[\arccos t^{1/2}-(t-t^2)^{1/2}]\right]^{2/3}, & 0\leqslant t\leqslant 1,\\[2ex] +\nu^{2/3}\left[\dfrac{3}{4}\,[(t^2-t)^{1/2}-\operatorname{arccosh} t^{1/2}]\right]^{2/3}, & t>1.\end{cases}$$

We find

$$(3.8)\qquad H(t)=\frac{5}{36}\,\frac{1-t}{t\,\psi(t)}+\frac{3-8t}{16\,t^2\,(1-t)^2}+\frac{\nu^2\,(1-t)}{4t},\quad t>0$$

where

$$(3.9)\qquad \psi(t)=\begin{cases}[\arccos t^{1/2}-(t-t^2)^{1/2}]^2, & 0<t\leqslant 1,\\[2ex] -[(t^2-t)^{1/2}-\operatorname{arccosh} t^{1/2}]^2, & t>1.\end{cases}$$

The following lemma holds.

LEMMA 3.2. *In the interval* $0 < t < +\infty$, *the function*

$$(3.10)\quad Q(t,\alpha) = H(t) - \left[\frac{\nu^2}{4}\left(\frac{1}{t}-1\right) + \frac{1-\alpha^2}{4t^2}\right],$$

with $H(t)$ *defined by* (3.8) *and* (3.9), *is negative if* $\alpha^2 \leqslant 1/4$, *is positive if* $\alpha^2 \geqslant 4/9$ *and has exactly one zero if* $1/4 < \alpha^2 < 4/9$.

For the proof we write $Q(t,\alpha)$ in the form

$$Q(t,\alpha) = \frac{1}{4t^2}\left[q(t) + \alpha^2 - \frac{1}{4}\right],$$

where

$$q(t) = \frac{5}{9}\,\frac{t(1-t)}{\psi(t)} - \frac{t(2+3t)}{4(1-t)^2}$$

with $\psi(t)$ given by (3.9). Then it is easily seen that the function $q(t)$ is negative and decreasing for $0 < t < \infty$. Further, we have

$$q(0) = 0, \qquad \lim_{t\to\infty} q(t) = \frac{-7}{36};$$

whence the lemma readily follows.

We can now derive this final result.

THEOREM 3.1. *Let* $x_{n,k}^{(\alpha)}$ *be the root of the equation*

$$(3.11)\quad x - \sin x = \frac{8}{3\nu}\,(-a_{n-k+1})^{3/2},\quad \nu = 4n+2\alpha+2,$$

where a_j *is the* j*-th zero of the Airy function* Ai (x). *Then, for the* k*-th zero* $\lambda_{n,k}^{(\alpha)}$ *of* $L_n^{(\alpha)}(x)$ *we have*

$$(3.12)\quad \lambda_{n,k}^{(\alpha)} > \nu\cos^2(x_{n,k}^{(\alpha)}/2),\ \text{if}\ -\frac{1}{2} \leqslant \alpha \leqslant \frac{1}{2},$$

$$(3.13)\quad \lambda_{n,k}^{(\alpha)} < \nu\cos^2(x_{n,k}^{(\alpha)}/2),\ \text{if}\ -1 < \alpha \leqslant -\frac{2}{3}\ \text{ or }\ \alpha \geqslant \frac{2}{3},$$

where $k = 1, 2, \ldots, n$.

We give only the essential steps of the proof.

In the case $-1/2 \leqslant \alpha \leqslant 1/2$ the function $u(t)$, defined by (3.3) and (3.7), has exactly n real zeros. Moreover, these zeros belong to the interval $(0,1)$ and can be obtained by solving the equations

(3.14) $$\arccos t^{1/2} - (t-t^2)^{1/2} = \frac{4}{3\nu}(-a_m)^{3/2},\quad \nu = 4n+2\alpha+2,$$

for $m = 1, 2, \ldots, n$, with respect to t. Indeed, the inequalities (3.5) show that

$$0 < \frac{4(-a_m)^{3/2}}{3\nu} < \frac{4n-5/6}{4n+2\alpha+2}\,\frac{\pi}{2} < \frac{\pi}{2}\ ,\quad -\frac{1}{2} \leqslant \alpha \leqslant \frac{1}{2}\ ,$$

for $m = 1, 2, \ldots, n$, while

$$\frac{4(-a_m)^{3/2}}{3\nu} > \frac{\pi}{2}\ ,$$

for $m > n$. Since the function $h(t)$ is negative and decreasing, the statement easily follows.

Now, let

$$u_{n,1}^{(\alpha)} > u_{n,2}^{(\alpha)} > \ldots > u_{n,n}^{(\alpha)}$$

be the zeros, in decreasing order, of $u(t)$ and let $u_{n,0}^{(\alpha)} = +\infty$. According to Lemma 3.2, $Q(t, \alpha)$ is negative for $0 < t < \infty$ if $-1/2 \leqslant \alpha \leqslant 1/2$. Therefore, we can apply Theorem 1.1 to the interval $(0, \infty)$. By using (3.4) we see that the condition (1.2) is satisfied at $t = \infty$ and we conclude that each interval

$$u_{n,n-k+1}^{(\alpha)} < t < u_{n,k}^{(\alpha)}\ ,\ k = 1, 2, \ldots n,$$

contains exactly one zero $t_{n,m}^{(\alpha)}$, $m = 1, 2, \ldots, n$. More precisely, we have

$$t_{n,k}^{(\alpha)} > u_{n,n-k+1}^{(\alpha)},\ k = 1, 2, \ldots, n.$$

By setting $t^{1/2} = \cos(x/2)$ in (3.14) with $m = n - k + 1$ we derive (3.11) and (3.12).

For the proof of (3.13) we use the same interval $(0, \infty)$. Since $Q(t, \alpha)$ is positive if $-1 < \alpha \leqslant -2/3$ or $\alpha \geqslant 2/3$, we find that each interval

$$t_{n,k}^{(\alpha)} < t < t_{n,k+1}^{(\alpha)},\quad k = 1, 2, \ldots, n,$$

where $t_{n,n+1}^{(\alpha)} = +\infty$, contains at least one zero $u_{n,m}^{(\alpha)}$, $m = 1, 2, \ldots$ That is, we have

$$t_{n,k}^{(\alpha)} < u_{n,n-k+1}^{(\alpha)},\quad k = 1, 2, \ldots, n.$$

Whence (3.13) follows.

4. NUMERICAL RESULTS AND CONCLUDING REMARKS

The inequalities given in Theorem 2.1 and in Theorem 3.1 furnish very sharp results, which are generally better than those we can obtain by using previously known inequalities.

TABLE 1 - *Bounds for some zeros* $\lambda_{20,k}^{(0)}$.

k	*Lower bound (1.3)*	*Exact value*	*Upper bound (2.18)*
1	0.070527	0.070540	0.070547
2	0.371601	0.372127	0.372164
10	11.444867	12.038803	12.040338
20	46.951357	66.524416	66.642245

The Table 1, which refers to some few values of k in the case $\alpha = 0$ and $n = 20$, shows, in the first column, the lower bounds given by the old inequality (1.3) and, in the third column, the upper bounds furnished by applying the new inequality (2.18).

The case $-1/2 \leqslant \alpha \leqslant 1/2$ is particularly interesting. Indeed, in this case, Theorem 2.1 and Theorem 3.1 give upper and lower bounds for $\lambda_{n,k}^{(\alpha)}$ respectively and the following corollary holds.

COROLLARY 4.1. *Let* $-1/2 \leqslant \alpha \leqslant 1/2$ *and let* $X(y)$ *be the function that we obtain upon inverting*

$$y = \sin x - x. \tag{4.1}$$

Then

$$\nu\cos^2\left[\frac{1}{2}X\left(\frac{8}{3\nu}(-a_{n-k+1})^{3/2}\right)\right] < \lambda_{n,k}^{(\alpha)} < \nu\cos^2\left[\frac{1}{2}X\left(\pi - \frac{4j_{\alpha,k}}{\nu}\right)\right], \tag{4.2}$$

for $k = 1, 2, \ldots, n$ *and where* ν, a_s *and* $j_{\alpha,s}$ *have the previous meaning.*

The lower and upper bounds in Theorem 4.1 furnish very sharp results when they are used as approximations, say $1_{n,k}^{(\alpha)}$, of $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$. This is shown in Figure 1 where the number of exact significant digits in the approximation of $\lambda_{n,k}^{(\alpha)}$, i.e. the *digits of accuracy* represented by the function

$$r_{n,k}(\alpha) = -\log_{10}\left|\frac{\lambda_{n,k}^{(\alpha)} - l_{n,k}^{(\alpha)}}{\lambda_{n,k}^{(\alpha)}}\right|$$

is plotted for $\alpha = 1/2$ and $n = 20$.

The curves Γ_1 and Γ_2 refer to the approximations

$$l_{20,k}^{(1/2)} = 83\cos^2\left[\frac{1}{2}X\left(\frac{8}{249}(-a_{21-k})^{3/2}\right)\right], \quad \textit{(lower bound)}$$

and

$$l_{20,k}^{(1/2)} = 83\cos^2\left[\frac{1}{2}X\left(\pi - \frac{4j_{1/2,k}}{83}\right)\right], \quad \textit{(upper bound)}$$

respectively.

In the same figure we have plotted (see the dashed curves γ_1 and γ_2) the digits of accuracy corresponding to the approximations

$$l_{20,k}^{(1/2)} = \frac{1}{83}\,j_{1/2,k}^2, \quad \textit{(lower bound)}$$

$$l_{20,k}^{(1/2)} = [(83)^{1/2} + 2^{-1/3}(83)^{-1/6}\,a_{21-k}]^2, \quad \textit{(upper bound)}$$

obtained by using the old bounds (1.3) and (1.5) respectively.

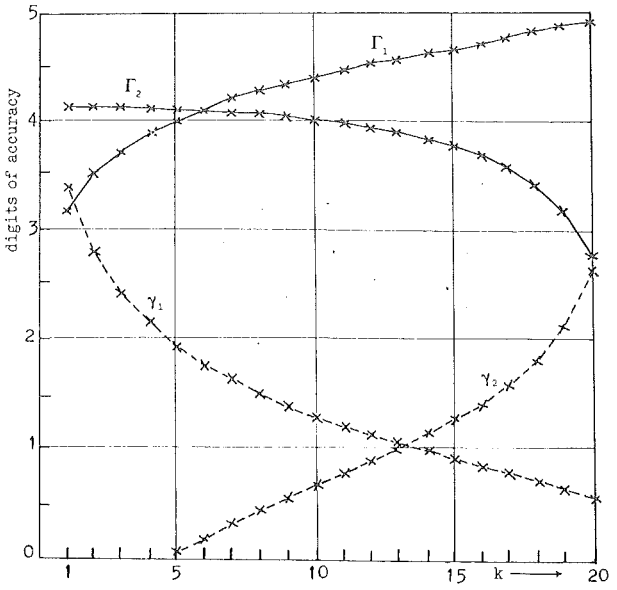


FIG. 1 - $r_{20,k}(1/2)$ versus k = 1(1)20.

The inequalities (4.2) can be used to derive upper and lower bounds for the zeros of the Hermite polynomials $H_n(x)$. Indeed, by taking into account that

$$H_{2m}(x) = (-1)^m 2^{2m} m!\, L_m^{(-1/2)}(x^2);\quad H_{2m+1}(x) = (-1)^m 2^{2m+1} m!\, x\, L_m^{(1/2)}(x^2),$$

and that

$$j_{1/2,k} = k\pi,\quad j_{-1/2,k} = (2k-1)\frac{\pi}{2},\qquad k = 1, 2, \ldots,$$

we obtain the following result:

COROLLARY 4.2. *Let $h_{n,k}$, $k = 1, 2, \ldots, [n/2]$, be the positive zeros in increasing order of the Hermite polynomial $H_n(x)$. Then*

$$\begin{aligned} h_{n,k} &< \sqrt{2n+1}\cos\left[\frac{1}{2}X\left(\frac{2n-4k+3}{2n+1}\pi\right)\right], \text{ if } n \text{ is even},\\ h_{n,k} &< \sqrt{2n+1}\cos\left[\frac{1}{2}X\left(\frac{2n-4k+3}{2n+1}\pi\right)\right], \text{ if } n \text{ is odd}, \end{aligned} \tag{4.3}$$

Furthermore, we have

$$h_{n,k} > \sqrt{2n+1}\cos\left[\frac{1}{2}X\left(\frac{8}{3(2n+1)}(-a_{[n/2]-k+1})^{3/2}\right)\right]. \tag{4.4}$$

Here $X(y)$ has the same meaning as in Corollary 4.1.

We remark that the Tricomi asymptotic formula (1.8) can be written, as $n \to \infty$,

$$\lambda_{n,k}^{(a)} \sim \nu\cos^2\left[\frac{1}{2}X\left(\frac{4n-4k+3}{\nu}\pi\right)\right],$$

for all the zeros $\lambda_{n,k}^{(a)}$, belonging to the interval $(a\nu, b\nu)$ with a and b fixed positive constants, $0 < a < b < 1$.

Now, by using the asymptotic expansions (see Abramowitz and Stegun [1], p. 371 and p. 450)

$$j_{a,s} = \left(s + \frac{\alpha}{2} - \frac{1}{4} \right) \pi \, [1+0(s^{-2})], \quad s \to \infty,$$

$$-a_s = \left[\frac{3\pi}{8} (4s-1) \right]^{2/3} [1+0(s^{-2})], \quad s \to \infty,$$

it is easily seen that

$$\pi - \frac{4j_{a,k}}{\nu} \sim \frac{4n-4k+3}{\nu} \pi, \text{ for } k \to \infty,$$

$$\frac{8}{3\nu}(-a_{n-k+1})^{3/2} \sim \frac{4n-4k+3}{\nu} \pi, \text{ for } n-k \to \infty.$$

Hence, the bounds for $\lambda_{n,k}^{(\alpha)}$ that we have considered in this paper are in fact approximations which coincide with the Tricomi approximation as $n \to \infty$, uniformly for all values of $k = [pn], [pn]+1, \ldots, [qn]$, where $p, q \in (0,1)$, $p < q$. More precisely, taking into account the results of Erdélyi [2] on the asymptotics for Laguerre polynomials, we have

$$\lambda_{n,k}^{(\alpha)} \sim \nu \cos^2 \left| \frac{1}{2} X \left(\pi - \frac{4j^{\alpha,k}}{\nu} \right) \right|, \quad n \to \infty,$$

$k = 1, 2, \ldots [qn],$

and

$$\lambda_{n,k}^{(\alpha)} \sim \nu \cos^2 \left| \frac{1}{2} X \left(\frac{8}{3\nu}(-a_{n-k+1})^{3/2} \right) \right|, \quad n \to \infty,$$

$k = [pn], [pn]+1, \ldots, n-1, n.$

longer assumed to be supported on $\mathbb{R}_+$ and to have constant sign. What, for example, would happen if one took a typical oscillatory measure, like $ds(t) = P_n(t)dt$ on $[-1,1]$, where P_n is the Legendre polynomial of degree n?

In a letter to Hermite, dated November 8, 1894 (in fact, his last letter in the life-long correspondence with Hermite; see Baillaud and Bourget [1905, v.2, pp. 439–441]), Stieltjes indeed looks at (what is now called) Legendre's function of the second kind

$$Q_n(z) = \int_{-1}^{1} \frac{P_n(t)}{z-t}\, dt, \tag{1.2}$$

expands it into descending powers of z (beginning with $z^{-(n+1)}$ by orthogonality of P_n) and then has the fortunate idea of expanding the reciprocal of Q_n,

$$\frac{1}{Q_n(z)} = z^{n+1}(\mu_0^{(n)} + \mu_1^{(n)} z^{-1} + \cdots), \quad \mu_0^{(n)} \neq 0. \tag{1.3}$$

This led him naturally to consider the polynomial part in (1.3),

$$E_{n+1}(z) = z^{n+1}(\mu_0^{(n)} + \mu_1^{(n)} z^{-1} + \cdots + \mu_{n+1}^{(n)} z^{-(n+1)}), \tag{1.4}$$

a polynomial of exact degree $n + 1$, now appropriately called *Stieltjes' polynomial*, and to investigate its properties. By a residue calculation, he first observes that

$$E_{n+1}(t) = \frac{1}{2\pi i} \oint_C \frac{dz}{(z-t)Q_n(z)}, \tag{1.5}$$

where C is a sufficiently large contour, and then goes on to multiply (1.5) by $t^k P_n(t)dt$, $k = 0, 1, \ldots, n$, and to integrate, obtaining

$$\int_{-1}^{1} E_{n+1}(t) t^k P_n(t)dt = \frac{1}{2\pi i} \oint_C \frac{dz}{Q_n(z)} \int_{-1}^{1} \frac{t^k P_n(t)}{z-t}\, dt$$

$$= \frac{1}{2\pi i} \oint_C \frac{dz}{Q_n(z)} \int_{-1}^{1} \frac{z^k - (z^k - t^k)}{z-t} P_n(t)dt$$

$$= \frac{1}{2\pi i} \oint_C \frac{z^k dz}{Q_n(z)} \int_{-1}^{1} \frac{P_n(t)}{z-t}\, dt$$

$$= \frac{1}{2\pi i} \oint_C z^k\, dz = 0,$$

where orthogonality of P_n is used in the third equality. Thus,

$$\int_{-1}^{1} E_{n+1}(t) p(t) P_n(t) dt = 0, \quad \text{all} \quad p \in \mathbb{P}_n, \tag{1.6}$$

that is, Stieltjes' polynomial E_{n+1} *is orthogonal to all lower-degree polynomials relative to the* (sign-variable) *measure* $ds(t) = P_n(t)dt$.

At this point, Stieltjes conjectures (1) that E_{n+1} has $n + 1$ real simple zeros, all contained in $(-1,1)$ and (2) that they separate those of P_n. He presents a numerical example with $n = 4$. He furthermore believes (strongly so in the case of reality and simplicity of the roots, less so for the separation property) that this is a special case of "a much more general theorem".

In his reply (of November 10, 1894), Hermite expressed his delight in the polynomials E_{n+1} and "the beautiful properties" conjectured for it and encouraged Stieltjes to look for a differential equation as a possible key to these properties. Stieltjes may have already been too ill to respond. Neither he, nor anybody else after him was able to give an affirmative answer to Hermite's suggestion. (It has been found, nevertheless, that the Stieltjes polynomials, at least in the realm of Jacobi measures $d\sigma^{(\alpha,\beta)}(t) = (1-t)^{\alpha}(1+t)^{\beta}\, dt$, do *not* satisfy a three-term recurrence relation unless $|\alpha| = |\beta| = \frac{1}{2}$, in which case they do, and in fact also satisfy a differential equation; cf. Monegato [1982].)

Stieltjes' ideas seem to have gone unnoticed for many years. Geronimus in 1930, however, developed similar ideas, considering in place of (1.3) the expansion of $[Q_n(z)\sqrt{z^2-1}]^{-1}$, where $Q_n(z) = \int_{-1}^{1} \pi_n(t\,; wdt) w(t) dt/(z-t)$ and $\pi_n(\cdot\,; wdt)$ is the nth degree orthogonal polynomial associated with the weight function $w(t) = (1-t)^{\alpha}(1+t)^{\beta} h(t)$, h being continuous and positive on $[-1,1]$ (Geronimus [1930]). Although this approach does not lead to a perfect orthogonality result, like the one in (1.6), it nevertheless has relevance to the subject at hand; see the beginning of Subsection 3.5 below.

The first who has taken up Stieltjes' challenge in earnest was Szegö in 1935. He expresses (Szegö [1935]) Stieltjes' polynomial on the circle as a cosine polynomial,

$$E_{n+1}(\cos\theta) = \lambda_0^{(n)} \cos(n+1)\theta + \lambda_1^{(n)} \cos(n-1)\theta + \cdots , \tag{1.7}$$

and relates an extended (infinite) sequence $\lambda_\nu = \lambda_\nu^{(n)}$ to an explicitly known sequence $f_\nu = f_\nu^{(n)}$ via a reciprocity identity for the respective power series. From this he proves $\lambda_0 > 0$ and the negativity of all λ_ν, $\nu \geq 1$, as well as $\sum_{\nu=0}^{\infty} \lambda_\nu = 0$. It follows from this that the polynomial $\lambda_0 z^{n+1} + \lambda_1 z^{n-1} + \cdots$ has all its zeros in $|z| < 1$, which implies, via the argument principle, that (1.7) vanishes at least $2n + 2$ times. This proves Stieltjes' first conjecture. Szegö also proves the second conjecture, but this requires a deeper analysis involving, in particular, Legendre functions on the cut.

Szegö's analysis is not peculiar to Legendre polynomials. Indeed, he himself extends it to Gegenbauer polynomials $P_n^{(\lambda)}$, orthogonal on $[-1,1]$ with respect to the measure $d\sigma(t) = (1-t^2)^{\lambda-1/2}dt$, $\lambda > -1/2$. If $E_{n+1}^{(\lambda)}$ denotes the corresponding Stieltjes polynomial,

$$\int_{-1}^{1} E_{n+1}^{(\lambda)}(t)p(t)P_n^{(\lambda)}(t)(1-t^2)^{\lambda-1/2}dt = 0, \quad \text{all } p \in \mathbb{P}_n, \tag{1.8}$$

which (up to a multiplicative constant) is uniquely defined, then Szegö shows that both conjectures of Stieltjes continue to hold for $0 < \lambda \leq 2$. When $\lambda=0$, two zeros of $E_{n+1}^{(\lambda)}$ move into the endpoints ± 1; they move outside of $[-1,1]$ for $\lambda<0$, as is shown by the example $n=2$. The question of whether the same can happen for $\lambda>2$ is left unanswered by Szegö. (The answer is still unknown today, but, according to Table 3.1 below, is probably "no", at least as long as the interlacing property holds.)

Szegö concludes by considering the Gaussian quadrature formula for the (sign-variable) measure $ds(t) = P_n(t)dt$ and shows that its weights alternate in sign.

This brings us naturally to the work of Kronrod in 1964, which is also concerned with quadrature. Motivated by a desire to economically estimate the error in the classical Gaussian quadrature formula

$$\int_{-1}^{1} f(t)dt \approx \sum_{\nu=1}^{n} \gamma_\nu f(\tau_\nu), \tag{1.9}$$

where $\tau_\nu = \tau_\nu^{(n)}$ are the zeros of the Legendre polynomial P_n and $\gamma_\nu = \gamma_\nu^{(n)}$ the corresponding Christoffel numbers, Kronrod [1964a,b] proposes to extend the n-point formula (1.9) to a $(2n + 1)$-point formula

$$\int_{-1}^{1} f(t)dt = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \tag{1.10}$$

in which the τ_ν are the same as in (1.9), but new nodes τ_μ^* and new weights σ_ν, σ_μ^* have been introduced and chosen to increase the degree of exactness from $2n - 1$ (for (1.9)) to $3n + 1$ (for (1.10)), i.e.,

$$R_n(f) = 0, \quad \text{all } f \in \mathbb{P}_{3n+1}. \tag{1.11}$$

It turns out that the nodes τ_μ^* must be precisely the zeros of Stieltjes' polynomial E_{n+1}. With all nodes τ_ν, τ_μ^* at hand, it is then easy to determine the weights σ_ν, σ_μ^* by interpolation.

In the same manner, one can try to extend the Gauss-Gegenbauer quadrature formula to a formula of the type

$$\int_{-1}^{1} f(t)(1-t^2)^{\lambda-\frac{1}{2}}\, dt = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \quad \lambda > -\tfrac{1}{2}, \tag{1.12}$$

and, more generally, to do the same for an integral with arbitrary (positive) measure $d\sigma$,

$$\int_{-1}^{1} f(t)d\sigma(t) = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \quad R_n(\mathbb{P}_{3n+1}) = 0. \tag{1.13}$$

(The dependence of the nodes and weights on n and $d\sigma$ will from now on be suppressed in our notation.) The new nodes τ_μ^*, similarly as before, are then the zeros of the (unique, monic) polynomial $\pi_{n+1}^*(\cdot) = \pi_{n+1}^*(\cdot\,; d\sigma)$ satisfying the orthogonality property

$$\int_{\mathbb{R}} \pi_{n+1}^*(t)p(t)\pi_n(t)d\sigma(t) = 0, \quad \text{all } p \in \mathbb{P}_n, \tag{1.14}$$

where $\pi_n(\cdot) = \pi_n(\cdot\,; d\sigma)$ is the orthogonal polynomial of degree n associated with the measure $d\sigma$. To be useful in practice, the formulae (1.12), (1.13) should have nodes τ_μ^* which are all contained in the support interval of $d\sigma$ and are different from the τ_ν, and they should have weights σ_ν, σ_μ^* which, if at all possible, are all positive. By Szegö's theory, we know that the former is

true for (1.12), if $0 < \lambda \leq 2$, while the latter has been proven true by Monegato [1978a] if $0 \leq \lambda \leq 1$, hence, in particular, for the original Gauss-Kronrod formula (1.10) (which corresponds to $\lambda = 1/2$).

Soon after Kronrod's work, it has occurred to a number of people (probably first to Patterson [1968a]) that other quadrature rules can be similarly extended, for example, the Gauss-Lobatto rule. In addition, it is not unreasonable to also consider the interpolatory quadrature rule based solely on the nodes τ_μ^* in (1.13). In the case of (1.10), numerical results suggest that these quadrature rules also have all weights positive and enjoy an interlacing property of their own: the zeros of E_{n+1} alternate with those of E_n; cf. Monegato [1982]. Indeed, having three quadrature rules at disposal – the one just mentioned, the Gauss rule (1.9), and (1.10) – with degrees of exactness roughly equal to n, $2n$ and $3n$, respectively, might well be an attractive feature that could be useful in automatic integration schemes (Kahaner [1987]).

Orthogonality with respect to sign-variable measures and related quadrature rules have independently been studied by Struble [1963], who develops a general theory. It might be interesting to explore this theory in the framework of more general indefinite inner product spaces (cf., e.g., Bognár [1974]).

The merit of discovering the connection between Kronrod's work and the earlier work of Stieltjes and Szegö is due to Mysovskih [1964], although it has been noted, independently, in the Western literature, by Barrucand [1970]. The relevance of Geronimus' work to Gauss-Kronrod quadrature is pointed out by Monegato [1982] and Monegato and Palamara Orsi [1985].

Brief accounts of the Kronrod and Patterson methods can be found in Davis and Rabinowitz [1984, pp. 106–109, 426] and Atkinson [1978, pp. 243–248].

2. Extended quadrature formulae. We now give a more systematic treatment of the problem of extending quadrature rules. We begin with a general theorem, which has become part of "folklore" in numerical quadrature and is difficult to attribute to any one in particular. In its key ingredients, it goes back to Jacobi [1826].

Let $d\sigma$ be a nonnegative measure on the real line $\mathbb{R}$, with bounded or unbounded support and with infinitely many points of increase. Assume that all its moments $\mu_k = \int_{\mathbb{R}} t^k d\sigma(t)$ exist and are finite. We consider quadrature rules of the form

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{N} \sigma_\nu f(\tau_\nu) + R_N(f), \tag{2.1}$$

where τ_ν, σ_ν are real and $N \geq 1$ an integer. We say that (2.1) has *degree of exactness* d if $R_N(f) = 0$ for every $f \in \mathbb{P}_d$, the class of polynomials of degree $\leq d$. We associate with (2.1) the polynomial

$$\omega(t) = \prod_{\nu=1}^{N} (t - \tau_\nu) \tag{2.2}$$

and call it the *node polynomial*. The theorem in question then reads as follows.

Theorem. *The quadrature rule* (2.1) *has degree of exactness* $d = N - 1 + k$, $k \geq 0$, *if and only if both of the following conditions are satisfied:*

(i) (2.1) *is interpolatory* (i.e., $d = N-1$);

(ii) $\int_{\mathbb{R}} \omega(t)p(t)d\sigma(t) = 0$ *for all* $p \in \mathbb{P}_{k-1}$.

We remark that polynomial degree of exactness $N-1$ (the case $k=0$ of the theorem) can always be achieved, simply by interpolating at the nodes τ_ν; this is condition (i) of the theorem. To get higher degree of exactness ($k>0$), the node polynomial, according to (ii), has to be orthogonal (relative to the measure $d\sigma$) to sufficiently many polynomials. If we have complete freedom in the choice of τ_ν and σ_ν, we can take k as large as $k=N$, in which case (ii) identifies $\omega(\cdot)$ with the (monic) orthogonal polynomial $\pi_N(\cdot\,; d\sigma)$ of degree N associated with the measure $d\sigma$, and the nodes τ_ν in (2.1) with its zeros. This, of course, is the well-known Gauss-Christoffel quadrature rule (cf., e.g., Gautschi [1981]).

The situation we are going to consider here is somewhat different: We shall assume that some of the nodes are prescribed and the rest variable. Let

$$N = N^\circ + N^*, \tag{2.3}$$

and suppose the prescribed (distinct) nodes are $\tau_1, \tau_2, \ldots, \tau_{N^\circ}$; we denote the remaining ones by

$$\tau_\mu^* = \tau_{N^\circ+\mu}, \quad \mu = 1, 2, \ldots, N^*. \tag{2.4}$$

Correspondingly, we let $\sigma_\mu^* = \sigma_{N^\circ+\mu}$ and write (2.1) in the form

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{N^\circ} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{N^*} \sigma_\mu^* f(\tau_\mu^*) + R_N(f). \tag{2.5}$$

We may interpret (2.5) as an "extension" of some quadrature rule

$$\int_{\mathbb{R}} f(t)d\sigma(t) \approx \sum_{\nu=1}^{N^\circ} \gamma_\nu f(\tau_\nu). \tag{2.6}$$

The degree of exactness of (2.6) is quite irrelevant for what follows, as the weights γ_ν are being discarded.

Putting

$$\pi_{N^\circ}^\circ(t) = \prod_{\nu=1}^{N^\circ} (t - \tau_\nu), \quad \pi_{N^*}^*(t) = \prod_{\mu=1}^{N^*} (t - t_\mu^*), \tag{2.7}$$

the theorem above, since $\omega(t) = \pi_{N^\circ}^\circ(t)\pi_{N^*}^*(t)$, becomes:

Corollary. *The quadrature formula* (2.5) *has degree of exactness* $d = N - 1 + k$, $k \geq 0$, *with N given by* (2.3), *if and only if it is interpolatory and the polynomial* $\pi_{N^*}^*$ *satisfies*

$$\int_{\mathbb{R}} \pi_{N^*}^*(t)p(t)\pi_{N^\circ}^\circ(t)d\sigma(t) = 0, \quad \textit{all} \quad p \in \mathbb{P}_{k-1}. \tag{2.8_k}$$

One expects the maximum degree of exactness to be realized for $k = N^*$ (there are $N + N^*$ degrees of freedom!), in which case (2.8_k) becomes

$$\int_{\mathbb{R}} \pi_{N^*}^*(t)p(t)\pi_{N^\circ}^\circ(t)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_{N^*-1}. \tag{2.8_{N^*}}$$

We call (2.5) an *optimal extension* of (2.6) if $k = N^*$, i.e., if (2.8_{N^*}) holds, and a *nonoptimal* [*interpolatory*] *extension* if (2.8_k) holds with $0 \leq k < N^*$ [k=0]. (We assume $p \equiv 0$ in (2.8_k) if k=0.) Thus, (2.5) is an optimal extension of (2.6) if and only if $\pi_{N^*}^*$ *is orthogonal to all lower-degree polynomials with respect to the* (sign-variable) *measure* $d\sigma^*(t) = \pi_{N^\circ}^\circ(t)d\sigma(t)$. Here is how sign-variable measures enter into the process of extending quadrature rules.

We now discuss a number of specific examples.

Example 2.1: Gauss-Kronrod formulae.

This is the case $N^\circ = n$, $\pi_{N^\circ}^\circ(\cdot) = \pi_n(\cdot\,; d\sigma)$, $N^* = n+1$, so that $N = 2n + 1$, $d = 3n + 1$, and (2.8_{N^*}) takes the form

$$\int_{\mathbb{R}} \pi_{n+1}^*(t)p(t)\pi_n(t\,; d\sigma)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_n. \tag{2.9}$$

(We must necessarily have $N^* \geq n + 1$ in this case; cf. Monegato [1980].) In other words, the classical n-point Gauss-Christoffel formula is optimally extended to a $(2n + 1)$-point formula of the form

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f). \tag{2.10}$$

The measure involved in the orthogonality relation (2.9) is $d\sigma^*(t) = \pi_n(t\,; d\sigma)d\sigma(t)$, which for $d\sigma(t) = dt$ is precisely the one considered by Stieltjes. We call π_{n+1}^* in (2.9) the *Stieltjes polynomial* associated with $d\sigma$ and denote it by $\pi_{n+1}^*(\cdot) = \pi_{n+1}^*(\cdot\,; d\sigma)$. It is easily seen that π_{n+1}^* (assumed monic of degree $n+1$) is uniquely determined by (2.9).

For the weights in (2.10) one finds (see, e.g., Monegato [1976])

$$\begin{aligned} \sigma_\nu &= \gamma_\nu + \frac{\|\pi_n\|_{d\sigma}^2}{\pi_{n+1}^*(\tau_\nu)\,\pi_n'(\tau_\nu)}, \qquad \nu = 1, 2, \ldots, n; \\ \sigma_\mu^* &= \frac{\|\pi_n\|_{d\sigma}^2}{\pi_n(\tau_\mu^*)\,\pi_{n+1}^{*\prime}(\tau_\mu^*)}, \qquad \mu = 1, 2, \ldots, n+1, \end{aligned} \tag{2.11}$$

where $\gamma_\nu = \gamma_\nu^{(n)}(d\sigma)$ are the Christoffel numbers, and $\|\cdot\|_{d\sigma}$ the L_2-norm for the measure $d\sigma$.

For symmetric measures, i.e., $d\sigma(-t) = d\sigma(t)$ and the support of $d\sigma$ symmetric with respect to the origin, it follows easily from uniqueness that

$$\pi_n(-t\,; d\sigma) = (-1)^n\,\pi_n(t\,; d\sigma), \quad \pi_{n+1}^*(-t\,; d\sigma) = (-1)^{n+1}\,\pi_{n+1}^*(t\,; d\sigma) \quad (d\sigma \text{ symmetric}), \tag{2.12}$$

so that (2.9) holds trivially for even polynomials p and is therefore valid for all $p \in \mathbb{P}_{n+1}$ if n is odd. Thus, $d = 3n + 1$ if n is even, and $d = 3n + 2$ if n is odd. (In special cases, the degree of exactness can be even higher; see Subsections 3.3 and 3.5 for examples.)

Example 2.2: Kronrod extension of Gauss-Radau formulae.

For definiteness we consider only the Radau formula with fixed node τ_0 at -1. The case $\tau_0 = 1$ is treated similarly.

We assume that $d\sigma$ is supported on $[-1,1]$ and that the measure $(1+t)d\sigma(t)$ allows $(2n + 1)$-point Kronrod extension, i.e., the Stieltjes polynomial $\pi^*_{n+1}(\cdot\,; (1+t)d\sigma)$ has distinct real zeros, all in $(-1,1)$ and all different from the zeros of $\pi_n(\cdot\,; (1+t)d\sigma)$. Then there exists a unique optimal extension of the Gauss-Radau formula for the measure $d\sigma$. It has the form

$$\int_{-1}^{1} f(t)d\sigma(t) = \sigma_0 f(-1) + \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma^*_\mu f(\tau^*_\mu) + R_n(f) \tag{2.13}$$

and corresponds to the case $N^\circ = n+1$, $\pi^\circ_{N^\circ}(t) = (1+t)\pi_n(t\,; (1+t)d\sigma)$, $N^* = n+1$, hence has degree of exactness (at least) $d = 3n + 2$. The orthogonality condition (2.8_{N^*}) assumes the form

$$\int_{-1}^{1} \pi^*_{n+1}(t)p(t)\pi_n(t\,; (1+t)d\sigma)(1+t)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_n. \tag{2.14}$$

Thus, as far as the nodes τ^*_μ are concerned, we can obtain them exactly as if we were to extend the Gauss formula for the measure $(1+t)d\sigma(t)$. Also, the quantities $(1 + \tau_\nu)\sigma_\nu$ and $(1 + \tau^*_\mu)\sigma^*_\mu$ can be obtained by expressions which are identical to the ones on the right-hand sides of (2.11), where the Christoffel numbers and norm refer to the measure $(1+t)d\sigma(t)$. The weight σ_0 then follows from $\sigma_0 + \sum_{\nu=1}^{n} \sigma_\nu + \sum_{\mu=1}^{n+1} \sigma^*_\mu = \mu_0, \quad \mu_0 = \int_{\mathbb{R}} d\sigma(t)$.

Example 2.3: Kronrod extension of Gauss-Lobatto formulae.

We assume, similarly as in Example 2.2, that the measure $(1 - t^2)d\sigma(t)$, supported on $[-1,1]$, allows Kronrod extension. Then the unique optimal extension of the $(n+2)$-point Gauss-Lobatto formula for the measure $d\sigma$ is given by

$$\int_{-1}^{1} f(t)d\sigma(t) = \sigma_0 f(-1) + \sigma_{n+1} f(1) + \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma^*_\mu f(\tau^*_\mu) + R_n(f), \tag{2.15}$$

and is the case $N^\circ = n+2$, $\pi^\circ_{N^\circ}(t) = (1-t^2)\,\pi_n(t\,; (1-t^2)d\sigma)$, $N^* = n+1$ of (2.5), with the degree of exactness now being (at least) $d = 3n + 3$. The orthogonality condition (2.8_{N^*}) becomes

$$\int_{-1}^{1} \pi^*_{n+1}(t)p(t)\pi_n(t\,; (1-t^2)d\sigma)(1-t^2)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_n, \tag{2.16}$$

and is the same as for Kronrod extension of the n-point Gauss formula for the measure $(1-t^2)d\sigma(t)$. Again, the quantities $(1-\tau_\nu^2)\sigma_\nu$ and $(1-\tau_\mu^{*2})\sigma_\mu^*$ have representations identical to those on the right of (2.11), the measure being $(1-t^2)d\sigma(t)$ throughout. The remaining weights σ_0, σ_{n+1} are most easily obtained by solving the system of two linear equations expressing exactness of (2.15) for $f(t)=1$ and $f(t)=t$.

We remark that in the special case of Jacobi measures $d\sigma^{(\alpha,\beta)}(t)=(1-t)^\alpha(1+t)^\beta dt$, $\alpha>-1$, $\beta>-1$, we have

$$\pi_n(\cdot\,;(1-t^2)d\sigma^{(\alpha,\beta)})=\frac{1}{n+1}\,\pi'_{n+1}(\cdot\,;d\sigma^{(\alpha,\beta)}), \tag{2.17}$$

as follows readily from the identity $(d/dt)P_{n+1}^{(\alpha,\beta)}(t)=\tfrac{1}{2}(n+\alpha+\beta+2)\,P_n^{(\alpha+1,\beta+1)}(t)$ for Jacobi polynomials.

Example 2.4: "Kronrod-heavy" extension of Gauss formulae.

The "Kronrod nodes" τ_μ^* and "Gauss nodes" τ_ν in the Gauss-Kronrod formula (2.10) are nicely balanced, in that exactly one Kronrod node fits into the space between two consecutive Gauss nodes and between the extreme Gauss nodes and the respective endpoints (possibly $\pm\infty$) of the support interval of $d\sigma$. There are, however, occasions (for example, in cases of nonexistence; cf. Subsection 3.4) where it might be necessary to forgo this balance in favor of more Kronrod nodes; we call such extensions *Kronrod-heavy*. These also fit into the general scheme (2.5), where $N^\circ=n$, $\pi_{N^\circ}^\circ(\cdot)=\pi_n(\cdot\,;d\sigma)$, $N^*=n+q$ with $q>1$, and give rise to the orthogonality condition

$$\int_{\mathbb{R}}\pi_{n+q}^*(t)p(t)\pi_n(t\,;d\sigma)d\sigma(t)=0,\quad\text{all}\quad p\in\mathbb{P}_{n+q-1}. \tag{2.18}$$

In contrast to Gauss-Kronrod formulae, the unique existence of π_{n+q}^*, let alone the reality of its zeros, is no longer assured. Starting with the unique $\pi_{n+1}^*(\cdot\,;d\sigma)=\pi_{n+1,n}^*(\cdot)$, however, there is an infinite sequence $\{\pi_{n+q_m,n}^*\}_{m=1}^\infty$ of uniquely determined polynomials $\pi_{n+q_m}^*=\pi_{n+q_m,n}^*$ of exact degree $n+q_m$, $1=q_1<q_2<q_3<\cdots$, such that (2.18) holds with $q=q_m$, and such that no polynomial $\pi_{n+q_m}^*$ of degree $<n+q_m$ exists for which (2.18) holds with $q=q_m$ (Monegato [1980]).

One can try, of course, to extend in this manner other quadrature formulae, e.g., the Gauss-Radau or Gauss-Lobatto formulae.

Example 2.5: Repeated Kronrod extension of Gauss formulae.

Given an n-point Gauss formula, one can try to extend it optimally to a $(2n + 1)$-point formula as in Example 2.1, then extend this formula once again to a $(4n + 3)$-point formula (by optimally adding $2n + 2$ new nodes), and so on. The likelihood of such repeated extensions to all exist (i.e., have real distinct nodes) is probably quite small. Remarkably, however, for n=3 and $d\sigma(t) = dt$ on [−1,1], such extensions, even with all weights positive, have been successfully computed by Patterson [1968a], [1973] up to the 255-point formula.

For the second extension, for example, the node polynomial π^*_{2n+2} must be orthogonal to all lower-degree polynomials with respect to the measure $d\sigma^*(t) = \pi_n(t\,; d\sigma)\,\pi^*_{n+1}(t\,; d\sigma)d\sigma(t)$.

Example 2.6: Extension by contraction.

As contradictory as this may sound, the point here is that one starts with a "base formula" containing a sufficiently large number of nodes, then successively removes subsets of nodes to generate a sequence of quadrature rules having fewer and fewer nodes. Looking at this sequence in the opposite direction then turns it into a sequence of (finitely often) extended quadrature rules.

More specifically, following Patterson [1968b], one takes as base formula any $(2^r + 1)$-point formula and then defines r subsets of points by successively deleting alternate points from the preceding subset (keeping the first and the last). For example, if r=3, the successive three subsets of the original points with index set {1,2,3,4,5,6,7,8,9} contain the points with indices {1,3,5,7,9}, {1,5,9} and {1,9}, respectively. A sequence of r+1 quadrature formulae can now be defined by taking the interpolatory formulae for the original node set and all r subsets of nodes. (A slightly different procedure is proposed by Rabinowitz, Kautsky and Elhay; see Rabinowitz, Kautsky, Elhay and Butcher [1987, Appendix A, p.125].)

The reality of the nodes is thereby trivially guaranteed, but not necessarily the positivity of the weights. Patterson [1968b], nevertheless, finds by computation that all weights remain posi-

tive if one starts with the 33-point, or 65-point Gauss-Legendre formula (r=5 and r=6, respectively), or with the 65-point Lobatto formula (r=6) as base formulae.

Another example of a suitable base formula, which in fact (Imhof [1963], Brass [1977, Satz 77]) has positivity of all weights built in, is the Clenshaw-Curtis formula (Clenshaw and Curtis [1960]) based on the initial point set $\tau_\nu = \cos(\nu\pi/2^r)$, $\nu = 0, 1, 2, \ldots, 2^r$.

If one is willing to delete successively one point at a time, then the following result of Rabinowitz, Kautsky, Elhay and Butcher [1987] is of interest: Given any interpolatory quadrature rule with all weights positive, it is possible to delete one of its points such that the interpolatory rule based on the reduced point set has all weights nonnegative.

All sequences of extended quadrature rules in Example 2.6 are examples of nonoptimal, in fact interpolatory, extensions. Other examples of nonoptimal, even subinterpolatory, extensions are those of product rules given by Dagnino [1983] (see also Dagnino [1986]). The severe sacrifice in polynomial degree of exactness is justified in this reference in terms of a simplified convergence and stability theory.

We restricted our discussion here to quadrature rules of the simplest type (2.1). There is little work in the literature on the extension of quadrature rules involving derivatives. Bellen and Guerra [1982], however, extend Turán-type formulae, but work them out only in very simple special cases.

3. Existence, nonexistence and remainder term. We consider here mainly the Gauss-Kronrod formula as defined in Example 2.1, that is,

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \quad R_n(\mathbb{P}_{3n+1}) = 0. \tag{3.1}$$

We say that the nodes τ_ν, τ_μ^* in (3.1) *interlace* if they are all real and, when arranged decreasingly, satisfy

$$-\infty < \tau_{n+1}^* < \tau_n < \tau_n^* < \cdots < \tau_2^* < \tau_1 < \tau_1^* < \infty. \tag{3.2}$$

For any given $n \geq 1$, the following properties are of interest:

(a) The nodes τ_ν, τ_μ^* interlace.

(b) All nodes τ_ν, τ_μ^*, in addition to interlacing, are contained in the interior of the smallest interval containing the support of $d\sigma$.

(c) The nodes interlace and each weight σ_ν is positive. (It is known, cf. Monegato [1976], that the interlacing property is equivalent to $\sigma_\mu^* > 0$, all μ.)

(d) All nodes, without necessarily satisfying (a) and/or (b), are real.

Little has been *proved* with regard to these properties; any new piece of information, from whatever source – computational or otherwise – should therefore be greeted with appreciation. In this section, we give an account of what is known, or what can be conjectured, for some classical and nonclassical measures.

3.1 *Gegenbauer measures* $d\sigma^{(\lambda)}(t) = (1-t^2)^{\lambda-\frac{1}{2}}dt$ *on* $[-1,1]$, $\lambda > -\frac{1}{2}$. Properties (a) and (b), as already mentioned in Section 1, have been proved for all $n \geq 1$ by Szegö [1935], when $0 < \lambda \leq 2$, and property (c) by Monegato [1978a], when $0 \leq \lambda \leq 1$. Properties (a) and (b) also hold for the extension of Lobatto formulae, if $-\frac{1}{2} < \lambda \leq 1$ (cf. Example 2.3), but nothing as yet has been proved concerning property (c). This, then, is the extent of what is known rigorously, for arbitrary n, at this time.

A good deal more, however, can be uncovered for specific values of n, if we let the parameter λ move continuously away from the above intervals and observe the resulting motion of the nodes τ_ν, τ_μ^* and the movement of the weights σ_ν, σ_μ^*. Given n, property (a) will cease to hold at the very moment a node τ_ν collides (for the first time) with a node τ_μ^*. This event is coincident with the vanishing of the resultant of $\pi_n(\cdot\,; d\sigma^{(\lambda)})$ and $\pi_{n+1}^*(\cdot\,; d\sigma^{(\lambda)})$. When λ has moved beyond this critical value, the nodes τ_ν and τ_μ^* involved in the collision have likely crossed each other, so that two Kronrod nodes now lie between consecutive Gauss nodes. Only now is it possible that two Kronrod nodes may collide and split into a pair of complex nodes, an event that is signaled by the vanishing of the discriminant of $\pi_{n+1}^*(\cdot\,; d\sigma^{(\lambda)})$. By using purely algebraic methods, it is thus possible to delineate parameter intervals in which properties (a) and (d) are valid. The

subintervals of the first of these, in which properties (b) and (c) hold, can be determined rather more easily, in an obvious manner.

This has been carried out computationally in Gautschi and Notaris [submitted] for values of n up to 40. Based on these results it is conjectured (and proved for $n \leq 4$) that property (p) holds for $\lambda_n^p < \lambda < \Lambda_n^p$, where the bounds λ_n^p and Λ_n^p for $p = a, b, c, d$ are as shown in Table 3.1.

n	λ_n^a	Λ_n^a	λ_n^b	Λ_n^b	λ_n^c	Λ_n^c	λ_n^d	Λ_n^d
1	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞
2	$-\frac{1}{2}$	∞	0	∞	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞
3	$-\frac{1}{2}$	16	0	16	$-\frac{1}{2}$	6.552...	$-\frac{1}{2}$	16
4	$-\frac{1}{2}$	∞	0	∞	$-\frac{1}{2}$	51.78...	$-\frac{1}{2}$	∞
≥ 5	$-\frac{1}{2}$	Λ_n^a	0	Λ_n^a	$-\frac{1}{2}$	Λ_n^c	$-\frac{1}{2}$	Λ_n^d

Table 3.1. *Property (p) for Gegenbauer measures*

Here, Λ_n^a, Λ_n^c, Λ_n^d are certain constants satisfying $1 < \Lambda_n^a < \infty$, $1 < \Lambda_n^c < \Lambda_n^a$ and $\Lambda_n^d \geq \Lambda_n^a$ with equality precisely when $n = 4r - 1$, $r = 1, 2, 3, \ldots$. Numerical values of these constants, to 10 decimal places, are provided in the cited reference for $n = 5(1)20(4)40$.

The fact that Kronrod extension (satisfying properties (c) and (d)) cannot exist for all $n \geq 1$ when λ is sufficiently large, not even if the degree of exactness is lowered to $[2rn + l]$, $r > 1$, l an integer, is claimed by Monegato [1979]. (The proof given is erroneous, but can be repaired; Monegato [1987].)

3.2 *Jacobi measures* $d\sigma^{(\alpha,\beta)}(t) = (1-t)^\alpha(1+t)^\beta dt$ *on* $[-1,1]$. Since interchanging the parameters α and β has the effect of changing the signs of the nodes τ_ν and τ_μ^*, hence, if the order (3.2) is maintained, of renumbering them in reverse order, and the same renumbering applies to the weights σ_ν and σ_μ^*, the validity of property (p), $p = a, b, c, d$, is unaffected by such an interchange. We will assume, therefore, that $-1 < \alpha \leq \beta$.

Except for the cases $|\alpha| = |\beta| = \frac{1}{2}$ (considered in Subsection 3.3) and the transformations to Gegenbauer measures noted below, the only known proven result is that property (b) is false for $\alpha = -\frac{1}{2}, -\frac{1}{2} < \beta < \frac{1}{2}$ when n is even, and for $\alpha = -\frac{1}{2}, \frac{1}{2} < \beta \leq 3/2$ when n is odd (Rabinowitz [1983, p.75] †).

(†) There is a misprint on p.75 of this reference: The superscript $\mu + \frac{1}{2}$ should be replaced by $\mu - \frac{1}{2}$ twice in Eq. (68), and twice in the discussion immediately following Eq. (69).

Monegato [1982] notes that $\pi_{n+1}^{*(\alpha,-\frac{1}{2})}(2t^2-1) = 2^{n+1}\, t\, \pi_{2n+1}^{*(\alpha,\alpha)}(t) - d_n$, where d_n is an explicitly given constant, and similarly, $\pi_{n+1}^{*(\alpha,\frac{1}{2})}(2t^2-1) = 2^{n+1}\, \pi_{2n+2}^{*(\alpha,\alpha)}(t)$. In the latter case, there are also simple relationships between the weights σ_ν, σ_μ^* of the respective Gauss-Kronrod formulae (3.1); cf. Gautschi and Notaris [submitted, Thm. 5.1]. The cases $\alpha > -1$, $\beta = \pm \frac{1}{2}$ can thus be reduced to the Gegenbauer case, and appeal can be made to the empirical results of Subsection 3.1, at least when $\beta = \frac{1}{2}$. A similar reduction is possible in the case $\alpha > -1$, $\beta = \alpha + 1$ (Monegato [1982, Eq. (36)]), which is of interest in connection with Kronrod extension of Gauss-Radau formulae for Gegenbauer measures (cf. Example 2.2).

The algebraic methods described in Subsection 3.1 have also been applied to general Jacobi measures (Gautschi and Notaris [submitted]) and the results for $2 \le n \le 10$ displayed by means of graphs. There are marked qualitative differences for n even and n odd, as is shown in Figure 3.1 for the cases n=6 and n=7. The region of validity for property (p) is consistently below the curve labeled "p", except for p=b and n even, when it is above and to the right of the curve.

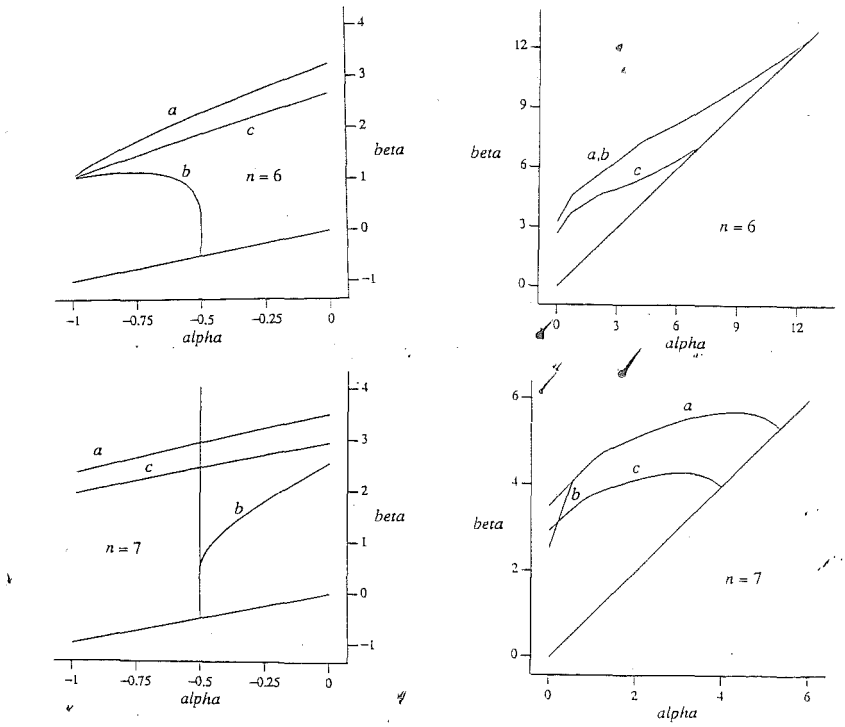


Figure 3.1. *Property (p), p = a, b, c for the Jacobi measure* $d\sigma^{(\alpha,\beta)}$ *when n=6 and n=7*

3.3 *Chebyshev measures of* 1st, 2nd *and* 3rd *kind.* These are the cases $|\alpha| = |\beta| = \frac{1}{2}$ of the Jacobi measure $d\sigma^{(\alpha,\beta)}$. They are the only known cases in which both the Gauss formulae and their Kronrod extensions can be written down explicitly (in terms of trigonometric functions). If $\alpha = \beta = -\frac{1}{2}$, the (optimal) extension of the n-point Gauss-Chebyshev formula of the first kind, when $n \geq 2$, is simply the $(2n + 1)$-point Lobatto formula for the same weight function. (For n=1, it is the 3-point Gauss-Chebyshev rule.) To get the Kronrod extension of the n-point Gauss-Chebyshev formula of the second kind ($\alpha = \beta = \frac{1}{2}$), it suffices to replace n by $2n + 1$ in the same formula. Finally, for $\alpha = -\frac{1}{2}$, $\beta = \frac{1}{2}$, the Kronrod extension is the Radau formula (with fixed node at 1) for the same weight function. All these extended formulae have elevated degrees of exactness, namely $4n - 1$, $4n + 1$ and $4n$, respectively, and enjoy property (p) for all $p = a, b, c$ (hence also d). These elegant relationships have been noted as early as 1964 by Mysovskih [1964]; see also Monegato [1982, p.147]. For the first two cases, Monegato [1976] points out that the formulae can be extended infinitely often in an explicit manner.

3.4 *Laguerre and Hermite measures.* Here is another instance in which a nonexistence result is known (Kahaner and Monegato [1978]): For the generalized Laguerre measure $d\sigma^{(\alpha)}(t) = t^{\alpha}e^{-t}\,dt$ on $[0,\infty]$, $-1 < \alpha \leq 1$, the Kronrod extension of the n-point Gauss-Laguerre formula, with real nodes and positive weights, does not exist when $n \geq 23$, and if α=0 not even for n>1. As a corollary, n-point Gauss-Hermite formulae cannot be so extended, unless $n = 1, 2$ or 4, confirming earlier empirical results of Ramskiĭ [1974]. These negative results led Kahaner, Waldvogel and Fullerton [1982], [1984] to explore the feasibility of Kronrod-heavy extensions for the Laguerre measure. Computational experience is reported for $n = 1(1)10$ and q=8 (11 for n=1 and 9 for n=2), where q is defined as in Example 2.4.

3.5 *Other measures.* At the heart of Geronimus' theory (Geronimus [1930]) is the measure $d\sigma_{\mu}(t) = (1-t^2)^{\frac{1}{2}}dt/(1 - \mu t^2)$ on $[-1,1]$, $-\infty < \mu \leq 1$. The corresponding polynomials $\pi_n(\cdot\,; d\sigma_{\mu})$ and $\pi^*_{n+1}(\cdot\,; d\sigma_{\mu})$ turn out to be linear combinations of Chebyshev polynomials U_n, U_{n-2} and T_{n+1}, T_{n-1}, respectively. This allows explicit construction of the associated Gauss-Kronrod extension and verification of all properties (a) – (c); cf. Gautschi and Rivlin [submitted]. In addition,

the degree of exactness is exceptionally high (Monegato [1982, p.146]). Similar expressions for π_n and π^*_{n+1} result if the denominator of $d\sigma_\mu$ is replaced by a positive, not necessarily even, polynomial of degree 2 (Monegato and Palamara Orsi [1985]).

Gautschi and Notaris [submitted, Thm. 5.2] observe that the problem of Kronrod extension for the measure $^\gamma d\sigma^{(\alpha)}(t) = |t|^\gamma (1-t^2)^\alpha dt$ on $[-1,1]$, $\alpha > -1$, $\gamma > -1$, can be reduced, when n is odd, to the analogous problem for the Jacobi measure $d\sigma^{(\alpha,(\gamma+1)/2)}$.

Very little is known for measures unrelated to classical measures. One that is likely to admit satisfactory Kronrod extension for every $n \geq 1$ (judging from numerical results of Caliò, Gautschi and Marchetti [1986]) is the logarithmic measure $d\sigma(t) = \ln(1/t)dt$ on $[0,1]$ for which properties (a), (b) and (c) appear to be all true. The same is conjectured for measures $d\sigma(t) = t^\alpha \ln(1/t)dt$, $\alpha = \pm \frac{1}{2}$, except for $\alpha = -\frac{1}{2}$ and n odd, in which case property (b), though not (d), fails, the polynomial $\pi^*_{n+1}(\cdot\,; d\sigma)$ having exactly one negative zero.

3.6 *Remainder term.* The Gauss-Kronrod formula (3.1) can be characterized in the manner of Markov [1885] as the unique quadrature formula (if it exists) obtained by integrating the interpolation polynomial $p_{3n+1}(f\,; \tau_\nu, \tau^*_\mu, \tau^*_\mu; \cdot)$ (with simple knots τ_ν and double knots τ^*_μ) of degree $\leq 3n + 1$ and by requiring (if possible) that the coefficients of all derivative terms in the resulting quadrature sum be zero. The elementary Hermite interpolation polynomials g_ν, h_μ, k_μ associated with this interpolation process can be easily expressed in terms of the fundamental Lagrange polynomials l_ν and l^*_μ for the nodes $\tau_1, \tau_2, \ldots, \tau_n$ and $\tau^*_1, \tau^*_2, \ldots, \tau^*_{n+1}$, respectively (see, e.g., Caliò, Gautschi and Marchetti [1986, Eq. (3.13)]). The coefficients $\sigma^{*\prime}_\mu$ required to be zero are then

$$\sigma^{*\prime}_\mu = \int_{\mathbb{R}} k_\mu(t) d\sigma(t), \quad \mu = 1, 2, \ldots, n+1, \tag{3.3}$$

where

$$k_\mu(t) = \frac{\pi_n(t)}{\pi_n(\tau^*_\mu)} [l^*_\mu(t)]^2 (t - \tau^*_\mu), \quad \pi_n(\cdot) = \pi_n(\cdot\,; d\sigma). \tag{3.4}$$

Thus we must have

$$\pi^*_{n+1}{}'(\tau^*_\mu) \int_{\mathbb{R}} \pi_n(t)[l^*_\mu(t)]^2 (t - \tau^*_\mu) d\sigma(t) = \int_{\mathbb{R}} \pi^*_{n+1}(t) l^*_\mu(t) \pi_n(t) d\sigma(t) = 0, \quad \mu = 1, 2, \ldots, n+1, \tag{3.5}$$

which, by the linear independence of the l_μ^*, is equivalent to the orthogonality condition (2.9).

From interpolation theory there follows that

$$R_n(f) = \frac{1}{(3n+2)!} \int_R [\pi_{n+1}^*(t)]^2 f^{(3n+2)}(\tau(t))\pi_n(t)d\sigma(t), \tag{3.6}$$

provided $f \in C^{3n+2}$ on an interval containing $\text{supp}(d\sigma)$. For Gegenbauer measures $d\sigma(t) = (1-t^2)^{\lambda-\frac{1}{2}}dt$ on $[-1,1]$, with $0 < \lambda < 1$, Monegato [1978b], relying heavily on Szegö's theory, shows that $|\pi_{n+1}^*(t\,; d\sigma)| < 2^{-n}$ on $[-1,1]$, which in combination with known bounds for $|\pi_n(\cdot\,; d\sigma)|$ yields an explicit upper bound for $|R_n(f)|$ in terms of $||f^{(3n+2)}||_\infty$. Rabinowitz [1980] improves this bound slightly and extends it to the case $1 < \lambda < 2$, as well as to Kronrod extensions of Gauss-Lobatto rules for $-\frac{1}{2} < \lambda \leq 1$, $\lambda \neq 0$. He also proves that for $0 < \lambda \leq 2$, $\lambda \neq 1$ the degrees $d = 3n+1$ and $d = 3n+2$ for n even and odd, respectively, are indeed the exact degrees of precision. (When $\lambda = 1$, one has exact degree $4n + 1$, and when $\lambda = 0$ exact degree $4n - 1$.) Analogous statements are proved for the Kronrod extension of the Gauss-Lobatto rule. Szegö's work, again, proves invaluable for this analysis, as it does, in combination with a result of Akrivis and Förster [1984, Proposition 1], to show that the remainder term $R_n(f)$ is indefinite if $0 < \lambda < 1$ and $n \geq 2$ (Rabinowitz [1986b]). For $\lambda > 1$, the question of definiteness is still open; it is also open for Kronrod extensions of Gauss-Lobatto rules for any λ (with the obvious exceptions).

Error constants in Davis-Rabinowitz type estimates of the remainder (Davis and Rabinowitz [1954]) for functions analytic on elliptic domains are given by Patterson [1968a] for his repeated extensions of the 3-point Gauss formula. They are compared with the corresponding constants for the Gauss and Clenshaw-Curtis formulae having the same number of points.

4. Computational methods, numerical tables, computer programs and applications.

4.1 *Computational methods.* Kronrod originally computed the Stieltjes polynomial $\pi_{n+1}^*(\cdot\,; dt)$ in power form, requiring it to be orthogonal (in the sense of (1.14)) to all monomials of degree $\leq n$. The zeros of π_{n+1}^* are then obtained by a rootfinding procedure, and the weights

σ_ν, σ_μ^* from a system of linear equations expressing exactness of (1.10) for the first $2n+1$ monomials. (Symmetry, of course, was used throughout.) As he himself observes, the procedure is subject to considerable loss of accuracy and therefore requires elevated precision. Patterson [1968a] achieves better stability by expanding π_{n+1}^* in Legendre polynomials and computing the coefficients recursively. He does so not only for the Kronrod extension of the Gauss formula, but likewise for the extension of the Lobatto formula. Further improvements and simplifications result from expansion in Chebyshev polynomials; cf. Piessens and Branders [1974]. Their procedure, even somewhat simplified and generalized to Gegenbauer measures, actually can be extracted from the work of Szegö [1935], as is pointed out by Monegato [1978b]; see also Monegato [1979], [1982]. For Gegenbauer measures, then, this seems to be the method of choice. Once the nodes have been computed, the weights can be obtained, e.g., by the formulae in (2.11).

Expansion of $\pi_{n+1}^*(\cdot\,; d\sigma)$ in orthogonal polynomials $\pi_k(\cdot\,; d\sigma)$, $k = 0, 1, \ldots, n+1$, however, is possible for arbitrary measures $d\sigma$. Replacing $p(\cdot)$ in (2.9) successively by $\pi_i(\cdot\,; d\sigma)$, $i = 0, 1, \ldots, n$, indeed yields a triangular system of equations which can be readily solved. Its coefficients can be computed, e.g., by Gauss-Christoffel quadrature relative to the measure $d\sigma$, using $[(3n + 3)/2]$ points; cf. Caliò, Gautschi and Marchetti [1986, Sec. 4]. (For another method, see Caliò, Marchetti and Pizzi [1984] and Caliò and Marchetti [1987].)

A rather different approach, resembling (in fact, generalizing) the well-known Golub-Welsch procedure (Golub and Welsch [1969]) for computing Gauss-Christoffel quadrature formulae is developed by Kautsky and Elhay [1984] and Elhay and Kautsky [1984] and relies on eigenvalues of suitably constructed matrices. For the weights, these authors use their own methods and software for generating interpolatory quadrature rules (Kautsky and Elhay [1982], Elhay and Kautsky [1985]).

Instead of computing, as above, the Gauss-Kronrod formula piecemeal – first the Stieltjes polynomial, then its zeros, and finally the weights – it might be preferable to compute these components all at once, for example by applying Newton's method to the system of $3n + 2$ (non-

linear) equations expressing exactness of the quadrature rule (2.10) for some set of basis functions in $\mathbb{P}_{3n+1}$. The feasibility of this idea is demonstrated in Caliò, Gautschi and Marchetti [1986], where the numerical condition of the underlying problem, hence the stability of the procedure, is also analyzed. It appears, though, that this method runs into severe ill-conditioning when one attempts to use it for repeated Kronrod extension (Gautschi and Notaris [in preparation]).

4.2. *Numerical tables.* There are a number of places where Kronrod extensions of n-point Gauss formulae can be found tabulated: Kronrod himself (Kronrod [1964b]) has them (transformed to the interval [0,1]) for $n = 1(1)40$ to 16 decimals (also in binary form!). In addition, he tabulates errors incurred when the formulae are applied to monomials. Patterson [1968a] (on microfiche) gives 20 S values for $n = 65$, and Piessens [1973] 16 S values for $n = 10$. The most accurate are the 33-decimal tables for $n = 7$, 10(5)30 in Piessens et al. [1983, pp. 19–23]. Extensions of $(n+2)$-point Lobatto formulae, $n = 1(1)7$ and $n = 63$, can be found to 20 decimals in Patterson [1968a] (on microfiche), and extensions of the $(n+1)$-point Radau formula, $n = 2(2)16$ (but incomplete), to 15 decimals in Baratella [1979].

Repeated Gauss-Kronrod extensions of the 3-point Gauss formula, as far up as the 127-point formula, are given to 20 significant digits in Patterson [1968a] (on microfiche), and the 255-point formula to the same accuracy in Patterson [1973] (in a Fortran data statement). The repeatedly extended 10-point formula, through the one with 87 points, is given to 33 decimals in Piessens et al. [1983, pp. 19, 26–27]. Extensions in the sense of Example 2.6 are tabulated to 20 decimals in Patterson [1968b] (on microfiche), using the 33-point and 65-point Gauss formula, as well as the 65-point Lobatto formula as "base formulae".

For measures other than the constant weight measure, there are 25 S tables of $(2n+1)$-point Gauss-Kronrod formulae for $d\sigma(t) = t^{\alpha} \ln(1/t)dt$ on [0,1], $\alpha = 0, \pm\frac{1}{2}$, where $n = 5(5)25$ for $\alpha = 0, \frac{1}{2}$, and $n = 4(4)24$ for $\alpha = -\frac{1}{2}$ (Caliò, Gautschi and Marchetti [1986, Suppl. S57–S63]). 15 S tables for the same weight functions, but with $n = 4$ and 12 for $\alpha = 0, \frac{1}{2}$, and $n = 6$ and 12 for $\alpha = -\frac{1}{2}$, are given in Caliò and Marchetti [1987]. Kahaner, Waldvogel and Fullerton [1984]

provide 15–18 S tables of Kronrod-heavy extensions of the Gauss-Laguerre formula ($d\sigma(t) = e^{-t}\,dt$ on $[0, \infty]$) with $n = 1$, $q = 3(1)6$ and $n = 10$, $q = 18$ (in the notation of Example 2.4).

We finally mention the 16 S tables of Piessens [1969] of the complex Gauss-Kronrod formulae, with $n = 2(1)12$, for the Bromwich integral, and the 15 S table of the interpolatory $(n+1)$-point formula based solely on the Kronrod nodes, given by Monegato [1982] for $d\sigma(t) = dt$ and $n = 2(1)9$.

4.3. *Computer programs.* Fortran programs for Kronrod extension of the n-point Gauss formula are provided in Squire [1970, p. 279] for $n = 20$, and in Piessens and Branders [1974] for arbitrary n. Dagnino and Fiorentino [1984] describe a Fortran program (listed in Dagnino and Fiorentino [1983]) generating Gauss-Kronrod formulae for Gegenbauer measures $d\sigma(t) = (1-t^2)^{\lambda - \frac{1}{2}}dt$ on $[-1,1]$, $0 \leq \lambda \leq 2$, $\lambda \neq 1$, using the recursive algorithm of Szegö as resurrected by Monegato (cf. Subsection 4.1). Programs for more general measures are described and listed in Caliò and Marchetti [1987], [1985], respectively.

A number of routines employing Gauss-Kronrod quadrature in the context of automatic integration are discussed and listed in Piessens et al. [1983].

4.4. *Applications.* The original motivation came from a desire to estimate the error of Gaussian, or other quadrature formulae (taking the more accurate Kronrod extension as a substitute for the exact answer). The need for such error estimates has recently been highlighted in connection with the development of automatic integration schemes; see, e.g., Cranley and Patterson [1971], Patterson [1973], Piessens [1973] and Piessens et al. [1983]. For an interesting interpretation of the Kronrod scheme of error estimation, see Laurie [1985]. A rather different estimation procedure is proposed in Berntsen and Espelid [1984].

Patterson's repeated extensions of the 3-point Gauss-Legendre rule (cf. Example 2.5) has been used with some success in certain methods to compute improper integrals arising in weakly singular integral equations. One method employs the ε – algorithm to accelerate a sequence of approximants (Evans, Hyslop and Morgan [1983]), another suitable transformations of variables to attenuate the singularity (Evans, Forbes and Hyslop [1983]).

Kronrod's idea has been applied to other types of integrals, for example, as already mentioned, to the Bromwich integral for the inversion of Laplace transforms (Piessens [1969]), and to Cauchy type singular integrals involving Gegenbauer measures (Rabinowitz [1983]). These applications, especially the latter, are not entirely straightforward, as the occurrence of numerical cancellation, or derivative values, may present difficulties. They can be surmounted, to some extent, by more stable implementations (Rabinowitz [1986a]), using, in part, Kronrod-heavy extensions (with $q = 2$; see Example 2.4). For an application of Kronrod's idea to cubature formulae, see Malik [1980], Genz and Malik [1980], [1983], Laurie [1982], Neumann [1982], Cools and Haegemans [1986], [1987] and Berntsen and Espelid [1987].

An interesting application, first noted by Barrucand [1970], is the use of Gauss-Kronrod formulae for computing Fourier coefficients in orthogonal expansions,

$$c_n(f) = \|\pi_n\|_{d\sigma}^{-1} \int_{\mathbb{R}} \pi_n(t) f(t) d\sigma(t), \quad n = 0, 1, 2, \ldots, \tag{4.1}$$

where $\pi_n(\cdot) = \pi_n(\cdot; d\sigma)$ is the nth degree orthogonal polynomial associated with the measure $d\sigma$. The $(2n+1)$-point Gauss-Kronrod formula (for the coefficient c_n), in this case, reduces to an $(n+1)$-point formula,

$$c_n(f) = \|\pi_n\|_{d\sigma}^{-1} \left[\sum_{\mu=1}^{n+1} \sigma_\mu^* \pi_n(\tau_\mu^*) f(\tau_\mu^*) + R_n(\pi_n f) \right], \tag{4.2}$$

but still has degree of exactness (at least) $2n + 1$. The new weights, $\sigma_\mu^* \pi_n(\tau_\mu^*)$, however, even if all σ_μ^* are positive, alternate in sign, which somewhat detracts from the usefulness of these formulae. For Gegenbauer measures $d\sigma^{(\lambda)} = (1-t^2)^{\lambda - \frac{1}{2}}$, $\lambda \neq 0,1$, Rabinowitz [1980] shows that the degree of exactness $2n + 1$ ($2n + 2$ if n is odd) is best possible. (4.2) is exact for polynomials of degree $3n - 1$, when $\lambda = 0$, and of degree $3n + 1$, when $\lambda = 1$, both of which is again best possible. The highest precision is thus obtained for Fourier-Chebyshev coefficients of the second kind.

Finite element and projection methods frequently rely on numerical integration but so far, Gauss-Kronrod formulae, unlike the Gauss formulae, have been shunned. An exception is Bellen [1980], who uses them in his "extended collocation-least squares" method.

Acknowledgment. The author is indebted to Professor P. Rabinowitz for providing additional references, particularly on multidimensional integration.

References

Akrivis, G. and Förster, K.–J. [1984]: *On the definiteness of quadrature formulae of Clenshaw-Curtis type,* Computing *33,* 363–366.

Atkinson, K.E. [1978]: *An Introduction to Numerical Analysis,* Wiley, New York, 1978.

Baillaud, B. and Bourget, H. [1905]: *Correspondance d'Hermite et de Stieltjes* I, II. Gauthier-Villars, Paris.

Baratella, P. [1979]: *Un' estensione ottimale della formula di quadratura di Radau,* Rend. Sem. Mat. Univ. e Politec. Torino *37,* 147–158.

Barrucand, P. [1970]: *Intégration numérique, abscisse de Kronrod-Patterson et polynomes de Szegö,* C.R. Acad. Sci. Paris *270,* 336–338.

Bellen, A. [1980]: *Metodi di proiezioni estesi,* Boll. Un. Mat. Ital. Suppl., no. *1,* 239–251.

Bellen, A. and Guerra, S. [1982]: *Su alcune possibili estensioni delle formule di quadratura gaussiane,* Calcolo *19,* 87–97.

Berntsen, J. and Espelid, T.O. [1984]: *On the use of Gauss quadratures in adaptive automatic integration schemes,* BIT *24,* 239–242.

_____, _____ [1987]: *On the construction of higher degree three dimensional embedded integration rules* (abstract), in: Numerical Integration – Recent Developments, Software and Applications (P. Keast and G. Fairweather, eds.), NATO Advanced Science Institute Series, Series C: Mathematical and Physical Sciences, Vol. *203,* 173–174. D. Reidel Publ. Co., Dordrecht.

Bognár, J. [1974]: *Indefinite Inner Product Spaces,* Ergebnisse der Mathematik und ihrer Grenzgebiete, Vol. 78, Springer, New York.

Brass, H. [1977]: *Quadraturverfahren,* Vandenhoeck and Ruprecht, Göttingen.

Caliò, F. and Marchetti, E. [1985]: *A program code of an algorithm to evaluate singular integrals,* Internal Report, Dipartimento di Matematica, Politecnico di Milano.

_____, _____ [1987]: *Derivation and implementation of an algorithm for singular integrals,* Computing *38,* 235–245.

_____, Gautschi, W. and Marchetti, E. [1986]: *On computing Gauss-Kronrod quadrature formulae,* Math. Comp. *47,* 639–650.

_____, Marchetti, E. and Pizzi, G. [1984]: *Valutazione numerica di alcuni integrali con singolarità di tipo logaritmico,* Rend. Sem. Fac. Sci. Univ. Cagliari *54,* 31–40.

Clenshaw, C.W. and Curtis, A.R. [1960]: *A method for numerical integration on an automatic computer,* Numer. Math. *2,* 197–205.

Cools, R. and Haegemans, A. [1986]: *Optimal addition of knots to cubature formulae for planar regions,* Numer. Math. *49,* 269–274.

_____, _____ [1987]: *Construction of sequences of embedded cubature formulae for circular symmetric planar regions*, in: Numerical Integration – Recent Developments, Software and Applications (P. Keast and G. Fairweather, eds.), NATO Advanced Science Institute Series, Series C: Mathematical and Physical Sciences, Vol. *203*, 165–172. D. Reidel Publ. Co., Dordrecht.

Cranley, R. and Patterson, T.N.L. [1971]: *On the automatic numerical evaluation of definite integrals*, Comp. J. *14*, 189–198.

Dagnino, C. [1983]: *Extended product integration rules*, BIT *23*, 488–499.

_____ [1986]: *Extensions of some results for interpolatory product integration rules to rules not necessarily of interpolatory type*, SIAM J. Numer. Anal. *23*, 1284–1289.

_____ and Fiorentino, C. [1983]: *A Fortran code for the computation of nodes and weights of extended Gaussian rules*, Internal Report, Dipartimento di Matematica, Politecnico di Torino.

_____, _____ [1984]: *Computation of nodes and weights of extended Gaussian rules*, Computing *23*, 271–278.

Davis, P.J. and Rabinowitz, P. [1954]: *On the estimation of quadrature errors for analytic functions*, Math. Tables Aids Comput. *8*, 193–203.

_____, _____ [1984]: *Methods of Numerical Integration*, 2nd ed., Academic Press, Orlando, Florida.

Elhay, S. and Kautsky, J. [1984]: *A method for computing quadratures of the Kronrod Patterson type*, Austral. Comput. Sci. Comm. *6*, no. *1*, 15.1–15.9. Department of Computer Science, University of Adelaide, Adelaide, South Australia.

_____, _____ [1985]: *IQPACK – Fortran subroutines for the weights of interpolatory quadratures*, School of Mathematical Sciences, The Flinders University of South Australia.

Evans, G.A., Hyslop, J. and Morgan, A.P.G. [1983]: *An extrapolation procedure for the evaluation of singular integrals*, Internat. J. Comput. Math. *12*, 251–265.

_____, Forbes, R.C. and Hyslop, J. [1983]: *Polynomial transformations for singular integrals*, Internat. J. Comput. Math. *14*, 157–170.

Gautschi, W. [1981]: *A survey of Gauss-Christoffel quadrature formulae*, in: E.B. Christoffel (P.L. Butzer and F. Fehér, eds.), 72–147. Birkhäuser, Basel.

_____, Notaris, S. [submitted]: *An algebraic study of Gauss-Kronrod quadrature formulae for Jacobi weight functions*.

_____, _____ [in preparation]: *Newton's method and Gauss-Kronrod quadrature*.

_____, Rivlin, T.J. [submitted]: *A family of Gauss-Kronrod quadrature formulae*.

Genz, A.C. and Malik, A.A. [1980]: *Algorithm 019 – Remarks on algorithm 006: An adaptive algorithm for numerical integration over an N-dimensional rectangular region*, J. Comput. Appl. Math. *6*, 295–302.

_____, _____ [1983]: *An imbedded family of fully symmetric numerical integration rules*, SIAM J. Numer. Anal. *20*, 580–588.

Geronimus, J. [1930]: *On a set of polynomials*, Ann. of Math. *31*, 681–686.

Golub, G.H. and Welsch, J.H. [1969]: *Calculation of Gauss quadrature rules*, Math. Comp. *23*, 221–230.

Imhof, J.P. [1963]: *On the method for numerical integration of Clenshaw and Curtis*, Numer. Math. *5*, 138–141.

Jacobi, C.G.J. [1826]: *Ueber Gauβ neue Methode, die Werthe der Integrale näherungsweise zu finden*, J. Reine Angew. Math. *1*, 301–308.

Kahaner, D.K. [1987]: *Personal communication.*

_____, Monegato, G. [1978]: *Nonexistence of extended Gauss-Laguerre and Gauss-Hermite quadrature rules with positive weights*, Z. Angew. Math. Phys. *29*, 983–986.

_____, Waldvogel, J. and Fullerton, L.W. [1982]: *Addition of points to Gauss-Laguerre quadrature formulas*, IMSL Tech. Rep. Series, 8205. IMSL, Houston.

_____, _____, _____ [1984]: *Addition of points to Gauss-Laguerre quadrature formulas*, SIAM J. Sci. Stat. Comput. *5*, 42–55.

Kautsky, J. and Elhay S. [1982]: *Calculation of the weights of interpolatory quadratures*, Numer. Math. *40*, 407–422.

_____, _____ [1984]: *Gauss quadratures and Jacobi matrices for weight functions not of one sign*, Math. Comp. *43*, 543–550.

Kronrod, A.S. [1964a]: *Integration with control of accuracy* (Russian), Dokl. Akad. Nauk SSSR *154*, 283–286.

_____ [1964b]: *Nodes and Weights for Quadrature Formulae. Sixteen-place Tables* (Russian). Izdat "Nauka", Moscow. [English translation: Consultants Bureau, New York, 1965.]

Laurie, D.P. [1982]: *Algorithm 584 – CUBTRI: Automatic cubature over a triangle*, ACM Trans. Math. Software *8*, 210–218.

_____ [1985]: *Practical error estimation in numerical integration*, J. Comput. Appl. Math. *12 & 13*, 425–431.

Malik, A.A. [1980]: *Some new fully symmetric rules for multiple integrals with a variable order adaptive algorithm*, Ph.D. thesis, University of Kent, Canterbury.

Markov, A. [1885]: *Sur la méthode de Gauss pour le calcul approché des intégrales*, Math. Ann. *25*, 427–432.

Monegato, G. [1976]: *A note on extended Gaussian quadrature rules*, Math. Comp. *30*, 812–817.

_____ [1978a]: *Positivity of the weights of extended Gauss-Legendre quadrature rules*, Math. Comp. *32*, 243–245.

_____ [1978b]: *Some remarks on the construction of extended Gaussian quadrature rules*, Math. Comp. *32*, 247–252.

_____ [1979]: *An overview of results and questions related to Kronrod schemes*, in: Numerische Integration (G. Hämmerlin, ed.), ISNM *45*, 231–240. Birkhäuser, Basel.

_____ [1980]: *On polynomials orthogonal with respect to particular variable-signed weight functions*, Z. Angew. Math. Phys. *31*, 549–555.

_____ [1982]: *Stieltjes polynomials and related quadrature rules*, SIAM Review *24*, 137–158.

_____ [1987]: *Personal communication.*

_____, Palamara Orsi, A. [1985]: *On a set of polynomials of Geronimus*, Boll. Un. Mat. Ital. *B* (6) *4*, 491–501.

Mysovskih, I.P. [1964]: *A special case of quadrature formulae containing preassigned nodes* (Russian), Vesci Akad. Navuk BSSR Ser. Fiz.-Tehn. Navuk, No. *4*, 125–127.

Neumann, G. [1982]: *Boolesche interpolatorische Kubatur*, Ph.D. thesis, Universität GH Siegen.

Patterson, T.N.L. [1968a]: *The optimum addition of points to quadrature formulae*, Math. Comp. *22*, 847–856. Loose microfiche suppl. C1–C11. [Errata, ibid. *23* (1969), 892.]

_____ [1968b]: *On some Gauss and Lobatto based integration formulae*, Math. Comp. *22*, 877–881. Loose microfiche suppl. D1–D5.

_____ [1973]: *Algorithm 468 – Algorithm for automatic numerical integration over a finite interval*, Comm. ACM *16*, 694–699.

Piessens, R. [1969]: *New quadrature formulas for the numerical inversion of the Laplace transform*, BIT *9*, 351–361.

_____ [1973]: *An algorithm for automatic integration*, Angew. Informatik, Heft *9*, 399–401.

_____, Branders, M. [1974]: *A note on the optimal addition of abscissas to quadrature formulas of Gauss and Lobatto type*, Math. Comp. *28*, 135–139. Suppl., ibid., 344–347.

_____, de Doncker-Kapenga, E., Überhuber, C.W. and Kahaner, D.K. [1983]: *QUADPACK: A Subroutine Package for Automatic Integration.* Springer Series in Computational Mathematics *1*. Springer, Berlin.

Rabinowitz, P. [1980]: *The exact degree of precision of generalized Gauss-Kronrod integration rules*, Math. Comp. *35*, 1275–1283. [Corrigendum: ibid. *46* (1986), 226 footnote.]

_____ [1983]: *Gauss-Kronrod integration rules for Cauchy principal value integrals,* Math. Comp. *41,* 63–78. [Corrigenda: ibid. *45* (1985), 277.]

_____ [1986a]: *A stable Gauss-Kronrod algorithm for Cauchy principal-value integrals,* Comput. Math. Appl. *12B,* 1249–1254.

_____ [1986b]: *On the definiteness of Gauss-Kronrod integration rules,* Math. Comp. *46,* 225–227.

_____, Kautsky, J., Elhay, S. and Butcher, J.C. [1987]: *On sequences of imbedded integration rules,* in: Numerical Integration – Recent Developments, Software and Applications (P. Keast and G. Fairweather, eds.), NATO Advanced Science Institute Series, Series C: Mathematical and Physical Sciences, Vol. *203,* 113–139. D. Reidel Publ. Co., Dordrecht.

Ramskiĭ, Ju. S. [1974]: *The improvement of a certain quadrature formula of Gauss type* (Russian), Vyčisl. Prikl. Mat. (Kiev) Vyp. *22,* 143–146.

Squire, W. [1970]: *Integration for Engineers and Scientists,* American Elsevier, New York.

Struble, G.W. [1963]: *Orthogonal polynomials – Variable-signed weight functions,* Numer. Math. *5,* 88–94.

Szegö, G. [1935]: *Über gewisse orthogonale Polynome, die zu einer oszillierenden Belegungsfunktion gehören,* Math. Ann. *110,* 501–513. [Collected Papers (R. Askey, ed.), Vol. 2, 545–557.]

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

THE HOLOGRAPHIC TRANSFORM

WALTER SCHEMPP

ABSTRACT: The basic idea of holography is to record analog signals as complex-valued functions on the (complexified) time-frequency plane. The holographic transform is a sesquilinear integral transformation which performs a planar encoding of the time and the frequency domains of signals simultaneously by means of interference patterns in the holographic plane. The 'frozen' interference patterns are recorded in the holographic plane by the hologram. The phase differences between the reference wave and the signal waves may be decoded by the coherent light of a laser beam in order to reconstruct the three-dimensional object from the planar hologram. - The present paper establishes an analog of the Paley-Wiener theorem for the holographic transform. Moreover, the holographic transform of the Hermite (or oscillator wave) functions is calculated explicitly in terms of Laguerre and Poisson-Charlier polynomials, and a series of holographic identities for digital signals are established. As a result, new identities for theta-null values are popping up. The energy preserving invariants of the holographic identities are classified by the ornamental groups (= dihedral groups D_m under the crystallographic restriction $m \in \{1,2,3,4,6\}$) via the elliptic Möbius transforms of the holographic plane $\mathbf{C}$. The orbits of the plane crystallographic groups D_m ($m \in \{2,3,4,6\}$) in the holographic plane $\mathbf{C}$ admit far-reaching applications to computerized holography, information theory, and neuromathematics.

0. CONTENTS

1. THE PERFECT LOW-PASS FILTER SINC

Recall the Paley-Wiener theorem which is at the basis of the classical sampling theorem.

Theorem 1 (Paley-Wiener). Let ψ denote an entire holomorphic function such that

$$\int_{\mathbf{R}} |\psi(x)|^2 dx < +\infty$$

and the estimate

$$|\psi(z)| \leq C\, e^{2\pi A|z|} \qquad (z \in \mathbf{C})$$

holds for positive constants A and C. Then there exists a function $\Psi \in L^2(-A,+A)$ such that

$$\psi(z) = \int_{-A}^{+A} \Psi(t) e^{2\pi i z t} dt \qquad (z \in \mathbf{C})$$

("finite Fourier cotransform" of Ψ).

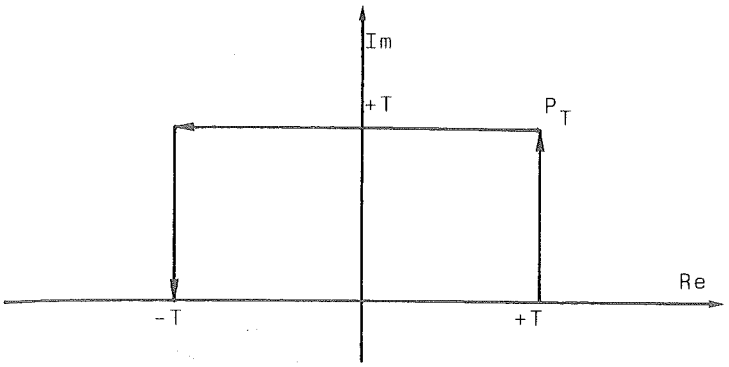
Proof. In order to establish that $\Psi = \mathscr{F}_{\mathbf{R}}\psi$ vanishes almost everywhere outside the compact interval $[-A,+A]$ of the real line $\mathbf{R}$, it will be sufficient by Cauchy's theorem to prove

$$\lim_{T \to +\infty} I_t = 0 \qquad (|t| > A)$$

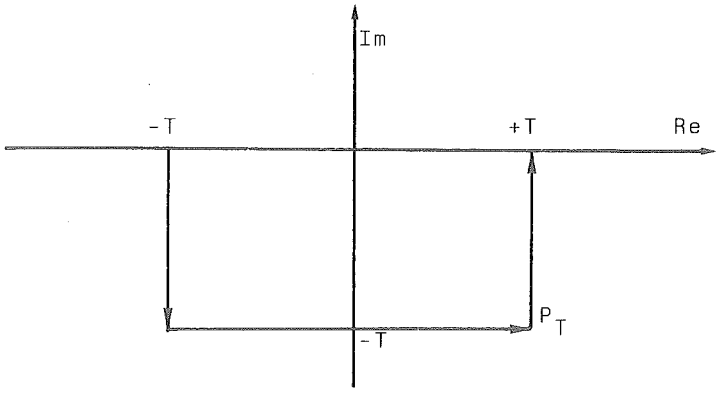
where the compact path P_T of the complex contour integral

$$I_t = \int_{P_T} \psi(z) e^{-2\pi i t z} dz \qquad (T > 0)$$

is defined in the following way:



$t < -A$



$t > +A$

The Phragmén-Lindelöf principle (a far-reaching generalization of the maximum modulus principle) implies that an entire holomorphic function of exponential type that is bounded on a line must be bounded on every parallel line in $\mathbf{C}$. It follows

$$|\psi(x+iT)| \leq Me^{2\pi AT} \qquad (x \in \mathbf{R})$$

where $M > 0$ is an appropriate constant. Without loss of generality, suppose $t < -A$. Then this estimate shows that the part of the complex contour integral I_t that belongs to the horizontal line of P_T vanishes as $T \to +\infty$. The line integrals belonging to the vertical parts of P_T can be

handled in a similar way. Indeed, consider the right vertical line of the path P_T. Then the corresponding line integral admits in absolute value an estimate by

$$\int_0^T e^{2\pi ty} |\psi(T+iy)| \, dy = \int_0^{T'} e^{2\pi ty} |\psi(T+iy)| \, dy + \int_{T'}^T e^{2\pi ty} |\psi(T+iy)| \, dy$$

where $T' \in]0,T[$. Another Phragmén-Lindelöf argument shows that

$$\lim_{T \to +\infty} \psi(T+iy) = 0$$

holds uniformly in $y \in [0,T']$. Consequently, we have

$$\lim_{T \to +\infty} \int_0^{T'} e^{2\pi ty} |\psi(T+iy)| \, dy = 0 \qquad (T' > 0).$$

Again by appealing to the Phragmén-Lindelöf principle, we conclude the estimate

$$\int_{T'}^T e^{2\pi ty} |\psi(T+iy)| \, dy \leq M \int_{T'}^T e^{2\pi(t+A)y} dy = \frac{M}{2\pi(t+A)} \left(e^{2\pi(t+A)T} - e^{2\pi(t+A)T'}\right)$$

Since $t < -A$, the last terms approach zero as $T' \to +\infty$ and $T \to +\infty$.-

The preceding proof can be traced back to lectures given by G.H. Hardy. For details of the simplified version, see the monograph by Boas [1].

The complex vector space $\mathcal{PW}(\mathbf{C})$ of all entire holomorphic functions of exponential type at most $(A \leq \frac{1}{2})$ that are square integrable along the real axis $\mathbf{R}$ forms a complex Hilbert space under the standard scalar product

$$\langle \psi | \phi \rangle = \int_{\mathbf{R}} \psi(x) \bar{\phi}(x) dx .$$

Let $\mathbf{T}$ denote the compact circle group. Then the Fourier transform $\mathcal{F}_{\mathbf{R}}$ is an isometric isomorphism of the Paley-Wiener space $\mathcal{PW}(\mathbf{C})$ onto the complex Hilbert space $L^2(\mathbf{T})$. By taking the Fourier cotransform $\bar{\mathcal{F}}_{\mathbf{R}}$ of the modes $e^{2\pi i \mu t}$ $(\mu \in \mathbf{Z})$ it follows that the sequence of functions

$$\operatorname{sinc}(z-\mu) = \begin{cases} \dfrac{\sin \pi(z-\mu)}{\pi(z-\mu)} & (z \neq \mu) \\ 1 & (z = \mu) \end{cases}$$

forms a Hilbert basis of $\mathcal{PW}(\mathbf{C})$. Accordingly each function $\psi \in \mathcal{PW}(\mathbf{C})$ admits a unique expansion of the form

$$\psi(z) = \sum_{\mu \in \mathbf{Z}} c_\mu \operatorname{sinc}(z-\mu) \quad (z \in \mathbf{C})$$

with $\|\psi\|^2 = \sum_{\mu \in \mathbf{Z}} |c_\mu|^2$. It follows $c_\mu = \psi(\mu)$ for all $\mu \in \mathbf{Z}$ and therefore we established the so-called sampling theorem.

Theorem 2 (Whittaker-Nyquist-Shannon-Kotel'nikov). A function $\psi \in \mathcal{PW}(\mathbf{C})$ can be recaptured from its values at the integers by the cardinal series:

$$\psi(z) = \sum_{\mu \in \mathbf{Z}} \psi(\mu) \operatorname{sinc}(z-\mu) \qquad (z \in \mathbf{C}).$$

The cardinal series is uniformly convergent in each horizontal strip in $\mathbf{C}$.

In terms of electrical engineering, a band-limited function ψ can be recovered from its equidistant samples by passing the data samples $(\psi(\mu))_{\mu \in Z}$ through a perfect low-pass filter. Since voice and video form band-limited signals, the sampling theorem is at the basis of digital signal processing. The scaled sinc-function serves as a perfect low-pass filter.

Example: CD-ROM (=Compact Disc Read Only Memory) for linear sequential digital signal processing. The encoding process is normally based on CIRC (= Cross-Interleaved Reed-Solomon Code).

Corollary 1. For all functions ψ and ϕ in $\mathcal{PW}(\mathbf{C})$ the sesquilinear quadrature formula

$$\sum_{n \in Z} \psi(n)\bar{\phi}(n) = \int_R \psi(x)\bar{\phi}(x)dx$$

holds.

Corollary 2. The complex Hilbert space $\mathcal{PW}(\mathbf{C})$ admits the reproducing kernel

$$(z,w) \longrightarrow \text{sinc } (z-\bar{w}).$$

For all functions $\psi \in \mathcal{PW}(\mathbf{C})$ the integral representation

$$\psi(z) = \int_R \psi(t) \text{ sinc } (t-z)\ dt$$

is valid for all $z \in \mathbf{C}$.

For a survey of the Whittaker-Nyquist-Shannon-Kotel'nikov sampling theorem, the reader is referred to the articles by Butzer [3], and Higgins [7]. Higgins also reviews some of the mathematics connected with the cardinal series and traces the origins of the result to before Whittaker. Also see the paper [21] for a proof of the sampling theorem via harmonic analysis on the compact Heisenberg nilmanifold.

As a final application of the Paley-Wiener theorem, we establish the following result due to S.N. Bernstein (1923).

Theorem 3 (Bernstein's inequality). Let $\psi \in \mathcal{PW}(\mathbf{C})$ - then

$$\|\psi' | \mathbf{R}\|_\infty \leq \pi \|\psi|\mathbf{R}\|_\infty .$$

Proof. Apply Theorem 1 to the entire holomorphic function

$$\phi_\varepsilon : z \longrightarrow \psi(z)\operatorname{sinc} \varepsilon z \qquad (\varepsilon > 0)$$

and observe that $\lim_{\varepsilon \to 0+} \phi_\varepsilon'(x) = \psi(x)$ holds for all $x \in \mathbf{R}$.-

Thus the derivative of a band-limited function on the real line $\mathbf{R}$ cannot get too large compared with the value of the function. This constraint is a fundamental one which has strong impact to vision. See Marr [12].

2. HOLOGRAPHY

The reasoning of the preceding section depends upon the duality of the complex Hilbert spaces

$$L^2(\mathbf{R}) \text{ and } L^2(\mathbf{R})$$

or

$$\mathcal{PW}(\mathbf{C}) \text{ and } L^2(\mathbf{T})$$

performed by the (linear) Fourier transform

$$\psi \longrightarrow \mathcal{F}_{\mathbf{R}}\psi.$$

From the physical point of view, however, the separation of the time and the frequency domains of (band-limited) signals is artificial. Moreover, it leads to serial algorithms which are not very efficient ways of signal processing.

The holography or wave front reconstruction (cf. Gabor [5]) is based on the following main idea: Consider for parallel signal processing the wave functions $\psi \in L^2(\mathbf{R})$ and their Fourier transformed versions $\mathscr{F}_{\mathbf{R}}\psi \in L^2(\mathbf{R})$ simultaneously.

From the mathematical point of view, the simultaneous encoding of time and frequency in the holographic plane can be performed by introducing the quadratic Fourier transform $H(\psi;.,.)$ of $\psi \in L^2(\mathbf{R})$ according to the prescription

$$H(\psi;x,y) = \int_{\mathbf{R}} \psi(t+x)\bar{\psi}(t)e^{2\pi iyt}dt$$

with $(x,y) \in \mathbf{R} \oplus \mathbf{R}$. If $\phi \in L^2(\mathbf{R})$, the sesquilinear analog reads as follows:

$$H(\psi,\phi;x,y) = \int_{\mathbf{R}} \psi(t+x)\bar{\phi}(t)e^{2\pi iyt}dt$$

Definition. The cross-correlator

$$L^2(\mathbf{R}) \times L^2(\mathbf{R}) \ni (\psi,\phi) \rightsquigarrow H(\psi,\phi;.,.)$$

is called the sesquilinear holographic transform. Its restriction to the diagonal, i.e., the corresponding autocorrelator, is called the quadratic holographic transform.

Key observation: Let $A(\mathbf{R})$ denote the three-dimensional real Heisenberg two-step nilpotent Lie group with one-dimensional center Z [23]. The projection $A(\mathbf{R})/Z$ of $A(\mathbf{R})$ along Z induces a symplectic structure on the plane $\mathbf{R} \oplus \mathbf{R}$ and a twisted convolution product on $L^2(\mathbf{R} \oplus \mathbf{R})$. The infinite dimensional, topologically irreducible, unitary, linear representations of $A(\mathbf{R})$ are square integrable mod Z. The sesquilinear holographic transform $H(\psi,\phi;.,.)$ coincides with the projection of the matrix coefficient of the linear Schrödinger re-

presentation U_1 of $A(\mathbf{R})$ defined by $\psi \in L^2(\mathbf{R})$ and $\phi \in L^2(\mathbf{R})$ along Z to the holographic plane [24],[25],[26]. The coadjoint orbit associated with U_1 under the Kirillov correspondence carries the symplectic form $(X,X') \rightarrowtail \det(X,X')$ and is isomorphic to the holographic plane by the exponential mapping.

Obviously the quadratic holographic transform satisfies the "peak property"

$$H(\psi;0,0) = \| \psi \|^2.$$

By virtue of the Cauchy-Schwarz-Bunjakovsky inequality, the sesquilinear holographic transform satisfies the estimate

$$H(\psi;\phi;x,y) \leq \| \psi \|.\| \phi \| \qquad ((x,y) \in \mathbf{R} \oplus \mathbf{R})$$

for all $\psi, \phi \in L^2(\mathbf{R})$. More important is the following result:

Theorem 4. For all functions ψ', ϕ' and ψ, ϕ in $L^2(\mathbf{R})$ the orthogonality relations

$$\int_{\mathbf{R}\oplus\mathbf{R}} \int H(\psi',\phi';x,y)\bar{H}(\psi,\phi;x,y)dxdy = \langle\psi'|\psi\rangle\langle\phi|\phi'\rangle$$

are valid.

As a consequence the following analog of the classical Paley-Wiener theorem (Theorem 1 supra) obtains.

Corollary. The sesquilinear holographic transform

$$\psi \otimes \phi \rightarrowtail H(\psi,\phi;.,.)$$

extends to an isometry of $L^2(\mathbf{R}) \hat{\otimes}_2 L^2(\mathbf{R})$ to the complex Hilbert space of Hilbert-Schmidt operators K on $L^2(\mathbf{R})$ realized as kernel operators

$$K\psi(x) = \int_{\mathbf{R}} k(x,y)\psi(y)dy \qquad (\psi \in L^2(\mathbf{R}))$$

with kernels $k \in L^2(\mathbf{R} \oplus \mathbf{R})$.

It is known (see Segal [27]) that the kernel k takes the form

$$k_f(x,y) = ({}_2\bar{\mathscr{F}}_{\mathbf{R}} f)(x-y,y) \qquad ((x,y) \in \mathbf{R} \oplus \mathbf{R})$$

where $f \in L^2(\mathbf{R} \oplus \mathbf{R})$ and ${}_2\bar{\mathscr{F}}_{\mathbf{R}}$ denotes the partial Fourier co-transform with respect to the second variable of the holographic plane. The bijective linear mapping

$$L^2(\mathbf{R} \oplus \mathbf{R}) \ni f \rightarrowtail k_f \in L^2(\mathbf{R} \oplus \mathbf{R})$$

is the Weyl transform. It gives rise to the natural Hilbert-Schmidt extension of the sesquilinear holographic transform and hence to the following result:

Theorem 5. A hologram generated on the holographic plane $\mathbf{R} \oplus \mathbf{R}$ by the Weyl transform

$$f \rightarrowtail k_f$$

acts by the Hilbert-Schmidt extension of the holographic transform as a linear spatial filter in a coherent optical system.

In Section 6 infra the preceding result will be used to point out an algorithm for generating sampled Fourier transform holograms.

3. RADIALITY

The property of the quadratic holographic transform $H(\psi;.,.)$ to form a radial function on the holographic plane $\mathbf{R} \oplus \mathbf{R}$ implies a serious restriction on the wave function $\psi \in L^2(\mathbf{R})$.

Theorem 6. Let $\psi \in L^2(\mathbf{R})$ be given and suppose that its quadratic holographic transform $H(\psi;.,.)$ is a radial function on the holographic plane $\mathbf{R} \oplus \mathbf{R}$. Then

$$\psi = \xi_n H_n$$

where $\xi_n \in \mathbf{C}$ is a constant and H_n is the Hermite function of degree $n \geq 0$.

4. SOME ORTHOGONAL POLYNOMIALS

a) Recall the definition of the Hermite functions

$$H_n(x) = e^{-\frac{1}{2}x^2} h_n(x) \qquad (x \in \mathbf{R})$$

where h_n denotes the Hermite polynomial of degree $n \geq 0$ satisfying the orthogonality relation

$$\int_{\mathbf{R}} h_n(x) h_m(x) e^{-x^2} dx = \delta_{nm}.$$

b) Let $L_n^{(\alpha)}$ denote the Laguerre function, i.e.,

$$L_n^{(\alpha)}(x) = e^{-\frac{1}{2}x} l_n^{(\alpha)}(x) \qquad (x \in \mathbf{R})$$

where $l_n^{(\alpha)}$ denotes the Laguerre polynomial of degree $n \geq 0$ and order $\alpha > -1$ satisfying the orthogonality relations

$$\int_0^\infty l_n^{(\alpha)}(x) l_m^{(\alpha)}(x) x^\alpha e^{-\frac{1}{2}x} dx = \delta_{nm}.$$

c) Finally, the Charlier-Poisson polynomials $c_n(.;a)$ on $\mathbf{N}$ of degree $n \geq 0$ and parameter value $a > 0$ are needed. The polynomials $c_n(.;a)$ satisfy the discrete orthogonality relations

$$\sum_{x \in \mathbf{N}} c_n(x;a)c_m(x;a)\frac{a^x}{x!} = e^a a^n n!\,\delta_{nm}$$

where $a > 0$.

Using the preceding orthogonality relations, we get the following result:

Theorem 7. The holographic transform of the Hermite functions reads in terms of Laguerre functions and Charlier-Poisson polynomials as follows:

$$H(H_m,H_n;x,y) = \sqrt{\frac{n!}{m!}}\,(\sqrt{\pi}(x+iy))^{m-n} L_n^{(m-n)}(\pi(x^2+y^2))$$

$$= \frac{(-1)^n}{\sqrt{m!n!}}\, z^{m-n}\,|z|^{2n}\, e^{-\frac{1}{2}|z|^2}\, c_n(m;|z|^2)$$

where $m \geq n \geq 0$ and $z = \sqrt{\pi}(x+iy) \in \mathbf{C}$.

5. THE HOLOGRAPHIC IDENTITIES

In the preceding theorem we identified the holographic plane $\mathbf{R} \oplus \mathbf{R}$ with the complex plane $\mathbf{C}$. If we restrict the holographic transform $H(\psi,\phi;.,.)$ to the quadratic lattice $\mathbf{Z} \oplus \mathbf{Z}$ in $\mathbf{R} \oplus \mathbf{R}$, i.e., to the lattice $\mathbf{Z}[i]$ of Gaussian integers in $\mathbf{C}$ we get

Theorem 8. Let ψ and ϕ be elements of $L^2(\mathbf{R})$ then the holographic identity

$$\sum_{(\mu,\nu) \in \mathbf{Z}\oplus\mathbf{Z}} H(\psi;\mu,\nu).\bar{H}(\phi;\mu,\nu) = \sum_{(\mu,\nu) \in \mathbf{Z}\oplus\mathbf{Z}} |H(\psi,\phi;\mu,\nu)|^2$$

is valid.

On the left hand side the signal terms occur whereas the right hand side encompass the interference terms. This explains the name. It can be established that the holographic identity implies the classical sampling theorem as a special case. However, it implies more.

In view of Theorem 7 we get by choosing for ψ and ϕ the Hermite functions:

Theorem 9. Let m,n be integers such that $m \geq n \geq 0$ - then the identity

$$\sum_{(\mu,\nu) \in Z \oplus Z} L_m^{(0)}(\pi(\mu^2+\nu^2)) \cdot L_n^{(0)}(\pi(\mu^2+\nu^2)) =$$

$$\frac{n!}{m!}\pi^{m-n} \sum_{(\mu,\nu) \in Z \oplus Z} (\mu^2+\nu^2)^{m-n} (L_n^{(m-n)}(\pi(\mu^2+\nu^2)))^2$$

holds.

The theta function is defined by means of the Fourier series

$$\vartheta(z,\tau) = \sum_{\mu \in Z} e^{-\pi\mu^2\tau} e^{2\pi i\mu z}$$

which is normally convergent in the domain $\{(z,\tau) \in \mathbb{C}^2 \mid \mathrm{Re}\,\tau > 0\}$. It was C.G.J. Jacobi (1804-1851) who invented the theta-function in the 1820s. Since then it has been used in many investigations by generations of number theorists. It is involved in many fascinating identities of number-theoretical and combinatorial import, and it provides one of the most effective ways to construct automorphic forms. According to D. Newman (Lecture in honour of A. Sharma, Edmonton 1986) the theta-function actually belongs to theology, and not to mathematics. In the early 1960s André Weil, inspired especially by the work of C.L. Siegel, provided a representation-

theoretic foundation for the theory of theta-function. See the classical paper by Weil [30]. He found that the theta-function is intimately connected with the metaplectic (or oscillator) representation, which forms a most singular projective unitary linear representation of the symplectic group. This representation arises by virtue of the existence of an action by automorphisms of the symplectic group on the Heisenberg two-step nilpotent Lie group A(R) mentioned in Section 2 supra. Moreover, André Weil showed the intimate relationship to the law of quadratic reciprocity (cf.[22]). The preceding theorem implies the following identities for the odd powers of π which can be considered as identities for the classical theta-function ("theta-null value")

$$\vartheta(\tau) = \vartheta(0,\tau) = \sum_{\mu \in Z} e^{-\pi\mu^2\tau} \qquad (\mathrm{Re}\,\tau > 0)$$

at the point $\tau = 1$ of the right half-plane.

m = 1, n = 0

$$\pi = \frac{\sum_{\mu \in Z} e^{-\pi\mu^2}}{4\sum_{\mu \in Z} \mu^2 e^{-\pi\mu^2}}$$

See Advanced Problem # 6491, Amer. Math. Monthly 92 (1985), 217.

m = 2, n = 1

$$\pi^3 = \frac{15\sum_{\mu \in Z} (8\pi^2\mu^4 - 1)e^{-\pi\mu^2}}{32\sum_{\mu \in Z} \mu^6 e^{-\pi\mu^2}}$$

$m = 3, \quad n = 2$

$$\pi^5 = \frac{45 \sum_{\mu \in Z} (16\pi^4\mu^8 - 140\pi^2\mu^4 + 21) e^{-\pi\mu^2}}{64 \sum_{\mu \in Z} \mu^{10} e^{-\pi\mu^2}}$$

$m = 4, \quad n = 3$

$$\pi^7 = \frac{91 \sum_{\mu \in Z} (256\pi^6\mu^{12} - 15840\pi^4\mu^8 + 166320\pi^2\mu^4 - 25245) e^{-\pi\mu^2}}{1024 \sum_{\mu \in Z} \mu^{14} e^{-\pi\mu^2}}$$

.

.

.

See Amer. Math. Monthly 93 (1986), 822-823 and Proc. Amer. Math. Soc. 92 (1984), 103-110.

6. HOLOGRAPHIC ENCODING

A mapping of the holographic plane

$$\sigma: \mathbf{R} \oplus \mathbf{R} \longrightarrow \mathbf{R} \oplus \mathbf{R}$$

is said to be an invariant of the quadratic holographic transform $H(\psi;.,.)$, if the identity

$$H(\psi;x,y) = H(\psi_\sigma;\sigma(x,y))$$

holds for all pairs $(x,y) \in \mathbf{R} \oplus \mathbf{R}$ and all functions $\psi \in L^2(\mathbf{R})$

in such a way that the assignment $\psi \longrightarrow \psi_\sigma$ defines a unitary operator in $L^2(\mathbb{R})$.

Theorem 10. A mapping of the holographic plane

$$\sigma : \mathbb{R} \oplus \mathbb{R} \longrightarrow \mathbb{R} \oplus \mathbb{R}$$

is an invariant of H, if and only if

$$\sigma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad \det \sigma = 1$$

with real coefficients a,b,c,d, i.e., $\sigma \in SL(2,\mathbb{R})$.

In the case when σ preserves the lattice $\mathbb{Z}[i]$ and the radiality of H, the choices of σ are drastically reduced.

Theorem 11. Let σ be an invariant of the holographic identity displayed in Theorem 8 supra. Then

$$\sigma = \begin{bmatrix} \cos 2\pi\frac{k}{m} & \sin 2\pi\frac{k}{m} \\ -\sin 2\pi\frac{k}{m} & \cos 2\pi\frac{k}{m} \end{bmatrix} \qquad (0 \leq |k| \leq m-1)$$

and m satisfies the cristallographic restriction

$$m \in \{1,2,3,4,6\}.$$

Proof. Since σ preserves the lattice $\mathbb{Z}[i]$, the coefficients a,b,c,d are integers. If

$$\sigma \notin \{-id_{\mathbb{R}\oplus\mathbb{R}},\ id_{\mathbb{R}\oplus\mathbb{R}}\}$$

preserves the radiality of H, the mapping

$$z \longrightarrow \frac{az+b}{cz+d}$$

defines an elliptic Möbius transformation of the upper complex half-plane preserving $\mathbf{R}$. It follows

$$|\mathrm{tr}\ \sigma| < 2$$

and since $\mathrm{tr}\ \sigma \in \mathbf{Z}$ obviously

$$\mathrm{tr}\ \sigma \in \{-1,0,+1\}.$$

Therefore σ is a turn through $\pm\pi/3$, $\pm\pi/2$ or $\pm 2\pi/3$, and no other turn is allowed.-

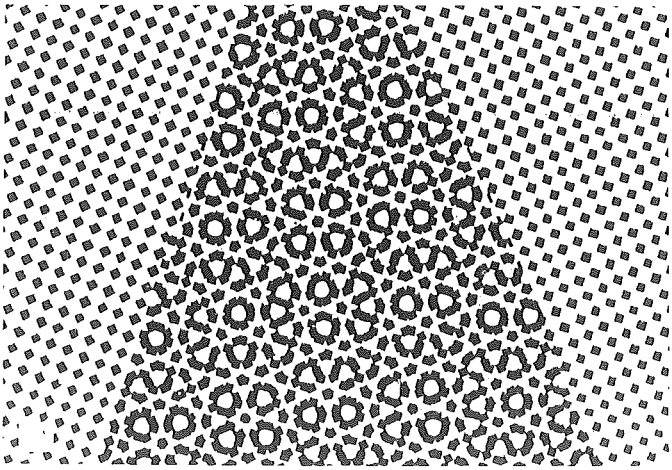
It follows that the holographic identities have the dihedral groups D_m ($m \in \{1,2,3,4,6\}$) as their groups of invariants. Nothing like a turn through $\pm\pi/5$ is possible. Only the classical planar crystal symmetries (or ornamental groups) and none of the forbidden fivefold symmetries, well-known from the theory of quasi-crystals, are allowed. For similar patterns arising in long crested wave models, see the paper by Schachter [20].

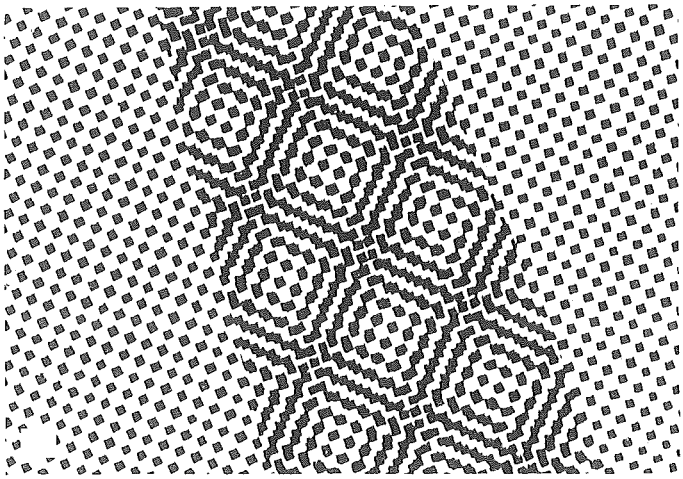
It should be observed that the dihedral groups D_m have order $2m$ and not the order m of the cyclic groups $\mathbf{Z}/m\mathbf{Z}$. Actually this fact reflects that a hologram generates two images, a real pseudoscopic image and a virtual orthoscopic image. It can be shown that the generation of orthoscopic and pseudoscopic images is at the basis of non-linear laser optics and in particular of non-linear optical phase-conjugation [24].

The figures on the next page show two superpositions of patterns formed by squares ($m = 4$).

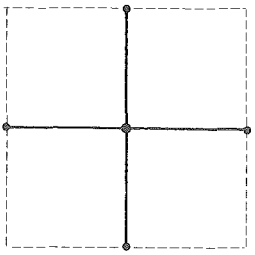
7. COMPUTERIZED HOLOGRAPHY

The periodic tilings of the holographic plane $\mathbf{R} \oplus \mathbf{R}$ enable to implement numerically various discretizations of the kernel

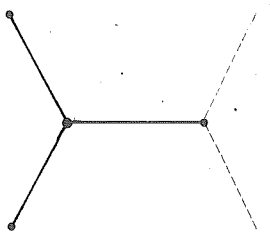




function k_f, the image of $f \in L^2(\mathbb{R} \oplus \mathbb{R})$ under the Weyl transform. In this way an algorithm arises by Theorem 5 supra to generate computer holograms. One way to do this is to compute in a first step by the FFT algorithm the Fourier transform $f = \mathcal{F}_{\mathbb{R} \oplus \mathbb{R}}\, g$ of the "two-dimensional image" g on the lattice with group D_m ($m \in \{2,3,4,6\}$) of invariants and then the second step is to compute the kernel k_f on the grid. In the case m = 4 we get Lee's encoding scheme of generating sampled Fourier transform holograms by decomposing the complex-valued functions to be synthesized into four components [10], [11]. Four times more samples are used along one direction than the other are required by this encoding technique.



In the case m = 6 we get Burckhardt's encoding scheme of generating sampled Fourier transform holograms by decomposing the complex valued functions to be synthesized into three components [2]. Also see Yaroslavskii [29].



For processing hexagonally sampled two-dimensional signals, the reader should consult Mersereau [13]. A similar procedure is possible in the cases m = 3 and m = 2.

3. THE NEURAL HOLOGRAPHIC MODEL

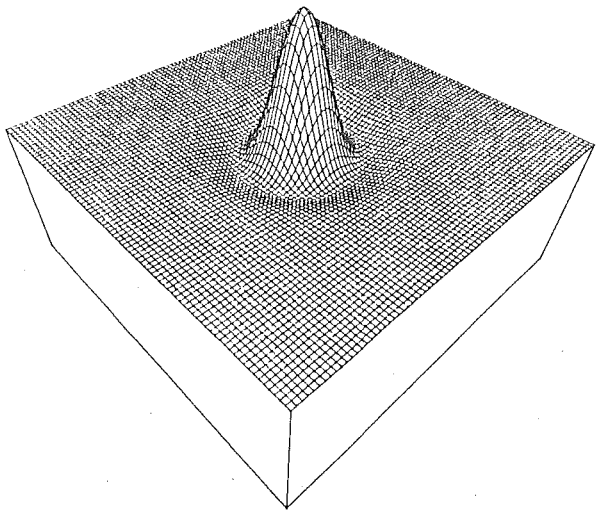
A growing number of theorists in the field of neurophysiology have invoked the principles of holography to explain certain aspects of brain function. One of the best established facts about brain mechanisms and memory is that large destructions within a neural system do not seriously impair its function. Indeed, the pioneering experiments by Lashley [9] showed that 80% or more of the visual cortex of a rat could be damaged without loss of the ability to correctly respond to patterns. Moreover, Robert Galambos (see Galambos, Norton, and Frommer [6]) has surgically removed as much as 98% of the optic tracts of cats with little effect on visual recognition behaviour. These and similar tests on monkeys and even men (performed during neurosurgery) have been interpreted to indicate that the neural elements necessary to the recognition and recall processes must be distributed throughout the brain systems involved. The problem that then confronts neurophysiologists is essential this: how can the relationships between neural activity become distributed and stored (temporarily or permanently) by a neural network. The neural holographic model developed by P.R. Westlake, K.H. Pribram and co-workers (Pribram [15], [16], [17]; also see Pribram, Nuwer, and Baron [18], Ferguson [4]) explains the property of distributed storage. Indeed, what makes the hologram unique as a storage device is that every element of the original image is distributed by the holographic transform and the Weyl transform (cf. Theorem 5 supra) over the entire holographic plane. Aside from this property, holographic memories show large capacities, parallel processing, and content addressability for rapid recognition, associative storage for perceptual completion, and for associative recall. The holographic hypothesis is in agreement with the experimental results of Rodieck [19] who found circularly symmetric excitability profiles of visual receptive fields which are conformal to Theorems 6 and 7 supra and also with the

thematical results by Marr [12]. See the figure on the following page and also Kronauer and Zeevi [8]. Moreover, Theorem 11 supra is in agreement with the results by Welt, Aschoff, Kameda, and Brooks [28] who found that "sensory convergence into the motor (sensory) cortex is superimposed on topographically uniform output organization in radial arrays, the diameter of which is estimated to be 0.1 to 0.4 mm. Thus, neurons with fixed local receptive fields provide a radially oriented framework (a reference system) for common peripheral inputs..." More precisely, Nicolis [14] concludes from his model of thalamocortical pacemaker that "specifically cognition is manifested at the cortex as a result of a matching process between pairs of spatial-temporal patterns, each containing a great number of elemental units (neurons). In each pair, one pattern (the same for all pairs) is the unknown information; it is embodied in incoming triggers, coded either in sequences of pulses from the peripheral nervous system, or, if it comes from other areas of the central nervous system, encoded in strings of macromolecular (neuro-transmitter/hormonal) releases from presynaptic endings. The other pattern of the pair is one of the pattern/attractors created by the processor; it constitutes a prestored spatial-temporal "mosaic" embodied in a set of partly synchronized post-synaptic membrane potentials or a spatial-temporal pattern of post-synaptic membrane receptors. The coupling or cross-correlation between the above two patterns of each pair takes place dynamically via energy exchanges between equal or neighbouring frequency pairs shared by both spectra... The result of the cross-correlation in phase and amplitude determines the "degree of cognition" between the incoming and the preset or the unknown and the expected patterns..."

It follows that the holographic transform provides a rigorous basis of neuromathematics. It includes the transference of phase informations to bijective linear transformations of the

holographic plane by the metaplectic representation of the symplectic group which explains the neural encoding of signal pulses emphasized by D.H. Hubel and T.N. Wiesel as well as the parallel processing of information emphasized by F.W. Campbell and D.A. Pollen.

Finally, let us quote P. Greguss (Lecture presented at the International Conference on Holography Applications, Beijing 1986): "I would like to express my belief that the holographic concept of Gabor is as fundamental as the general relativity theorem of Einstein, and it has to be explored further for a better understanding of nature in which we live."



Acknowledgments. The author is grateful to Professors Pál Greguss (Technical University Budapest) and Yehoshua Y. Zeevi (Harvard and Technion) for stimulating discussions. Moreover, he acknowledges the constant support and constructive criticisms by Miklôs Nyari (Technical University Budapest). Finally, the hospitality of the Mathematical Research Institute at Oberwolfach is gratefully acknowledged, where parts of this work has been done.

REFERENCES

1. R.P. BOAS, JR.: Entire functions. Academic Press, New York, N.Y., 1954.

2. C.B. BURCKHARDT: A simplification of Lee's method of generating holograms by computer. Applied Optics **9** (1970), 1949, 2813.

3. P.L. BUTZER: A survey of the Whittaker-Shannon sampling theorem and some of its extensions. J. Math. Research Exposition **3** (1983), 185-212.

4. M. FERGUSON: Wirklichkeit und Wandel - Karl Pribram als Pionier der Gehirn- und Bewußtseinsforschung. In: Das holographische Weltbild (K. Wilber, Hrsg.), Scherz-Verlag, Bern, München, Wien, 1986, pp. 12-26.

5. D. GABOR: Associative holographic memories. IBM J. of Research and Development **13** (1969), 156-159.

6. R. GALAMBOS, T.T. NORTON and C.P. FROMMER: Optic tract lesions sparing pattern vision in cats. Experimental Neurology **18** (1967), 8-25.

7. J.R. HIGGINS: Five short stories about the cardinal series. Bull. (New Series) Amer. Math. Soc. **12** (1985), 45-89.

8. R.E. KRONAUER and Y.Y. ZEEVI: Reorganization and diversification of signals in vision. IEEE Trans. Syst., Man, Cybern. **13** (1985), 91-101.

9. K.S. LASHLEY: In search of the engram. In: Physiological Mechanisms in Animal Behaviour. Academic Press, New York, N.Y. 1951, pp. 112-146.

10. W.H. LEE: Sampled Fourier transform hologram generated by computer. Applied Optics **9** (1970), 639-643.

11. W.H. LEE: Computer-generated holograms: Techniques and applications. In: Progress in Optics, Vol. XVI (E. Wolf, ed.), North-Holland, Amsterdam, New York, Oxford, 1978, pp. 119-232.

12. D. MARR: Vision. W.H. Freeman, San Francisco, 1982.

13. R.M. MERSEREAU: The processing of hexagonally sampled two-dimensional signals. Proc. IEEE **67** (1979), 930-949.

14. J.S. NICOLIS: Dynamics of hierarchical systems. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1986.

15. K.H. PRIBRAM: Languages of the brain: Experimental paradoxes and principles in neuropsychology. 5th ed. Brandon House, Bronx, N.Y., 1982.

16. K.H. PRIBRAM: Worum geht es beim holographischen Paradigma? In: Das holographische Weltbild (K. Wilber, Hrsg.), Scherz-Verlag Bern, München, Wien, 1986, pp. 27-36.

17. K.H. PRIBRAM: Holography and brain function. In: Encyclopedia of neuroscience (G. Adelman, ed.), Vol. I. Birkhäuser Verlag, Boston, Basel, Stuttgart, 1987, pp. 499-500.

18. K.H. PRIBRAM, M. NUWER and R.J. BARON: The holographic hypothesis of memory structure in brain function and perception. In: Measurement, Psychophysics, and Neural Information Processing (D.H. Krantz, R.D. Luce, R.C. Atkinson, P. Suppes, eds.), W.H. Freeman, San Francisco, 1974, pp. 416-457.

19. R.W. RODIECK: Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vison Res. **5** (1965), 583-601.

20. B. SCHACHTER: Long crested wave models. Computer Graphics and Image Processing **12** (1980), 187-201. Also in: Image Modeling (A. Rosenfeld, ed.), Academic Press, New York, London, Toronto, Sydney, San Francisco, 1981, pp. 327-341.

21. W. SCHEMPP: Gruppentheoretische Aspekte der Signalübertragung und der kardinalen Interpolationssplines I. Math. Methods Appl. Sci. **5** (1983), 195-215.

22. W. SCHEMPP: Group theoretical methods in approximation theory, elementary number theory, and computational signal geometry. In: Approximation Theory V (C.K. Chui, L.L. Schumaker, J.D. Ward, eds.), Academic Press, Boston, Orlando, San Diego, New York, Austin, London, Sydney, Tokyo, Toronto, 1986, pp. 129-171.

23. W. SCHEMPP: Harmonic analysis on the Heisenberg nilpotent Lie group, with applications to signal theory. Pitman Research Notes in Math., Vol. **147**. Longman Scientific and Technical, Harlow, Essex, 1986.

24. W. SCHEMPP: Signal geometry (to appear).

25. W. SCHEMPP: The holographic transformation (to appear).

26. W. SCHEMPP: The holographic plane (to appear).

27. I.E. SEGAL: Transforms for operators and symplectic automorphisms over a locally compact abelian group. Math. Scand. **13** (1963), 31-43.

28. C. WELT, J.C. ASCHOFF, K. KAMEDA and V.B. BROOKS: Intracortical organization of cat's motorsensory neurons. In: Neuropysiological Basis of Normal and Abnormal Motor Activities (M.D. Yahr, D.P. Purpura, eds.), Raven Press, Hewlett, N.Y., 1967, pp. 255-294.

29. L.P. YAROSLAVSKII: Applied problems of digital optics. In: Advances in Electronics and Electron Physics (P.W. Hawkes, ed.), Academic Press, Orlando, San Diego, New York, Austin, London, Montreal, Sydney, Tokyo, Toronto, 1986, pp. 1-140.

30. A. WEIL: Sur certains groupes d'opérateurs unitaires. Acta Math. **111** (1964), 143-211. Also in: Collected papers, Vol. III. Springer-Verlag, New York, Heidelberg, Berlin, 1980, pp. 1-69.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

THE MOVING GRID METHOD FOR BLN PROBLEM

M. ALIĆ and R. MANGER

ABSTRACT. *We consider Godunov method for the Bardos, Leraux and Nedelec initial-boundary problem in the case of nonuniform grids. Computer code and results are also included.*

Let $f \in C^2(R)$, a_0, $a_L \in R$ and let $Q_T =]0,L[\times]0,T[$, $\tilde{Q}_T = [0,L]\times]0,T[$ for $T > 0$, $L > 0$. For $u \in BV(Q_T)$, $c \in R$ and $\varphi \in C_0^1(\tilde{Q}_T)$ we introduce the notation

$$E(u,\varphi,c) = -\int_0^L \int_0^T \{|u-c| \frac{\partial \varphi}{\partial t} + \operatorname{sign}(u-c)[f(u)-f(c)] \frac{\partial \varphi}{\partial x}\} dx\, dt +$$

$$+ \int_0^T \operatorname{sign}(c-a_0)[f(T_0u)-f(c)]\varphi(0,t)dt -$$

$$- \int_0^T \operatorname{sign}(c-a_L)[f(T_Lu) - f(c)]\varphi(L,t)dt ,$$

where $(T_0u)(t) = \lim_{x\to 0+} u(x,t)$ in $L^1(0,T)$ and $(T_Lu)(t) = \lim_{x\to 1-} u(x,t)$ in $L^1(0,T)$.

For a given $u_0 \in BV(]0,L[)$ we consider Bardos, Leroux and Nedelec problem

$$(1) \quad \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \qquad \text{in } Q_T$$

$$(2) \quad u(x,0) = u_0(x) \qquad \text{in }]0,1[$$

$$(3) \quad \min_{c\in J[T_0u(t),a_0]} \operatorname{sign}(a_0-T_0u(t))[f(T_0u(t))-f(c)] = 0$$

$$(4) \quad \min_{c\in J[T_Lu(t),a_L]} \operatorname{sign}(T_Lu(t)-a_L)[f(T_Lu(t))-f(c)] = 0 ,$$

for $t \in]0,T[$ where

$$J[\alpha,\beta] = [\min\{\alpha,\beta\}, \max\{\alpha,\beta\}].$$

DEFINITION. A function $u \in BV(Q_T)$ is a solution of the problem (1)-(4) if it satisfies the initial condition (2) almost everywhere in $]0,T[$ and if

(5) $$E(u,\varphi,c) \leq 0$$

for all $c \in \mathbf{R}$ and all non negative $\varphi \in C_0^1(Q_T)$.

Bardos, Leroux and Nedelec have proven in [1] the existence and uniqueness theorem for the above problem.

The following lemma is a fundamental one for our consideration:

LEMMA. Let $u_0 \in BV(]0,L[)$ be a step function. If $u \in BV(Q_{T'})$ is the solution of (1)-(4) and if $w \in BV(\mathbf{R}\times]0,T'[)$ is the solution of the Cauchy problem

(6) $$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x} f(w) = 0 \quad \text{in} \quad \mathbf{R}\times]0,T'[$$

(7) $$w(x,0) = \begin{cases} a_0, & x \leq 0 \\ u_0(x), & x \in]0,L[\\ a_L, & x \geq L \end{cases}$$

then

$$u = v|_{Q_T},$$

for a short time $T' > 0$.

The proof of this lemma follows from the fact that if w is short time solution of Cauchy problem (with short time T given by some Courant condition, see [6]) then w satisfies boundary conditions (3) and (4) as a solution of a Riemann problem (see [2]).

For $\delta \in]0,\delta_0[$ we consider a set of grids $\{G_\delta\}$ in Q_T where $G_\delta = \{(x_i^j, t^j)\}$ and where

$$0 = t^0 < t < \ldots < t^n = T$$

$$0 = x_0^j < x_1^j < \ldots < x_{m_j}^j = L \ .$$

We suppose that there exist positive constants C_0, C_1, k_0, k_1 such that

$$(8) \qquad k_0 \cdot \delta \leq x_{i+1}^j - x_i^j = \Delta x_i^j \leq k_1 \delta$$

$$(9) \qquad C_0 \leq \frac{\Delta^+ t^j}{\Delta x_i^j} = \frac{t^{j+1} - t^j}{\Delta x_i^j} \leq C_1$$

where

$$(10) \qquad C_1 = \frac{1}{2 \max\limits_{n \in I} |f'(u)|} \ ,$$

$I = [\min\{a_0, a_L, \inf u_0(x)\}, \max\{a_0, a_L, \sup u_0(x)\}$

(this is Courant condition!).

It follows from (8) and (9) that

$$(11) \qquad n\delta \leq C_2$$

where $C_2 = \frac{T}{C_0 k_0}$. We define a regular set of grids as a set of grids with properties (8), (9) and (10).

For the formulation of Godunov method we use the solution operator $S(t)$ for BLN problem (1)-(4) and the averaging operator A_j, $j=0,\ldots,n-1$. The operator A_j is defined for $u \in L^1(0,1)$ by the formula

$$(A_j u)(x) = \frac{1}{\Delta x_i^j} \int_{x_i^j}^{x_{i+1}^j} u(\xi) d\xi$$

for $x \in [x_i^j, x_{i+1}^j[$. We define an approximation v^δ by

$$v^\delta(x,t) = v^j(x,t) \ ,$$

for $(x,t) \in]0,L[\times]t^j, t^{j+1}[$ and $j=0,\ldots,n-1$, where

$$v^0(x,0) = A_0 u_0(x)$$

$$v^j(x,t^j) = A_j v^{j-1}(x,t^j), \quad j=1,\dots,n-1,$$

and

$$v^j(x,t) = S(t-t^j)v^j(x,t^j)$$

for $t \in [t^j,t^{j+1}]$ and $j=0,\dots,n-1$.

THEOREM. If $u \in BV(Q_T)$ is the solution of the problem (1)-(4) and $\{G_\delta\}$ a regular set of grids in Q_T then $u = \lim_{\delta\to 0} v^\delta$ in $L^\infty(0,T:L(0,L))$.

Proof. Let w^j be the solution of Cauchy problem

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \ , \text{ in } R \times [t^j,t^{j+1}] \ , \tag{12}$$

$$w(x,t^j) = \begin{cases} a_0 \ , & x \le 0 \\ v^j(x,t^j), & x \in]0,L[\\ a_L \ , & x \ge L \ . \end{cases} \tag{13}$$

For w^j and $t \in [t^j,t^{j+1}]$ we have fundamental Kružkov estimates:

$$\inf_x w^j(x,t^j) \le w^j(x,t) \le \sup_x w^j(x,t^j) \ , \tag{14}$$

$$Var[w^j(\cdot,t);R] \le Var\,[w^j(\cdot,t^j);R] \tag{15}$$

and

$$\|w^j(\cdot,t+\tau)-w^j(\cdot,t)\|_{L^1(R)} \le |\tau| \cdot L \cdot Var[w^j(\cdot\, t^j) \ ; \ R \ ] \tag{16}$$

for $t+\tau \in [t^j,t^{j+1}]$ where L is the Lipschitz constant of f on I. We define the function $\overline{v}^\delta$ by the formula

$$\overline{v}^\delta = \sum_{j=0}^{n-1} w^j \cdot \chi_{]0,L]\times[t^j,t^{j+1}[}$$

such that

$$v^\delta = \overline{v}^\delta|_{Q_T}.$$

By using the results of Lemma 3.1. and Lemma 3.2. from [5]

we obtain that for some positive constant C_1, C_2 and C_3

$$\|\bar{v}^{\delta}\|_{L^{\infty}(R^2)} \leq C_1 \,, \tag{17}$$

$$\mathrm{Var}[\,\bar{v}^{\delta}(\cdot,t);\ R] \leq C_2 \,, \tag{18}$$

and

$$\|\,\bar{v}^{\delta}(\cdot,t+\tau)-\bar{v}^{\delta}(\cdot,t)\|_{L^{\infty}(R^2)} \leq C_3(|\tau|+\delta). \tag{19}$$

Estimates (17), (18) and (19) imply that every sequence (v^{δ}) with δ tending to zero has a subsequence converging to a limit in $L^{\infty}(0,T;L(0,L))$. This limit is also in $BV(Q_T)$, by some integral criterium (see [7], C.IV.§3). By the inequality (17) there exists a subsequence (v^{δ_r}) such that $v^{\delta_r} \to v$ in $L^1(Q_T)$, $f(T_0 v^{\delta_r}) \to p$ and $f(T_L v^{\delta_r}) \to q$ weak star in $L^{\infty}(0,T)$. Similarly as in [5] it follows that $p = f(T_0 v)$ and $q = f(T_L v)$ if $\lim_{r\to\infty} \bar{t}(v^{\delta_r},\varphi,C) \leq 0$. Indeed, from inequalities $E(v^j,\psi,C) \leq 0$ for

$$\psi \in C_0([0,T]\times]t^j,t^{j+1}[\,,$$

$\psi \geq 0$, $j=0,\ldots,n-1$, $C \in R$

we obtain the inequality

$$E(v^{\delta},\varphi,c) \leq \sum_{j=0}^{n-1} \{\int_0^L |v^j(x,t^j)-c|\varphi(x,t^j)dx\} -$$
$$- \int_0^L |v^j(x,t^{j+1})- c|\varphi(x,t^{j+1})dx\}\,, \tag{20}$$

for $\varphi \in C_0(\tilde{Q}_T)$, $\varphi \geq 0$. The inequality (20) implies, as in [6] the inequality

$$E(v^{\delta},\varphi,c) \leq K\delta$$

for a positive K and we finally have

$$E(v,\varphi,c) \leq 0$$

which, because of uniqueness theorem, completes the proof of Theorem.

Define $v_i^j = v^\delta(x,t^j)$ for $x \in [x_i^j, x_{i+1}^j[$. Using the fact that v^δ is the exact solution on the strip $]0,L[\times]t^j,t^{j+1}[$ and the divergence theorem on the trapezium with vertices (t^{j+1},x_{i+1}^{j+1}), (t^{j+1},x_i^{j+1}), (t^j,x_k^j), (t^j,x_l^j) we obtain the Godunov sheme

$$(21) \qquad v_i^{j+1}\,\Delta x_i^j = \sum_{r=k}^{l-1} v_r^j \Delta x_r^j - \Delta^+ t^j [Y_{l,i+1}^j - Y_{k,i}^j] \ ,$$

where

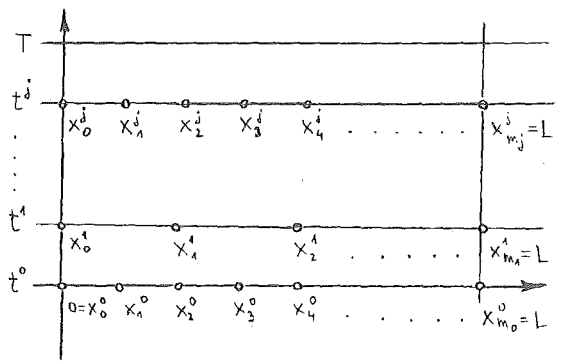
$$Y_{k,i}^j = \begin{cases} \min\limits_{u\in[v_{K-1}^j, v_K^j]} [f(u) - X_{K,i}^j u] \ , & \text{if} \quad v_{K-1}^j < v_K^j \\ \max\limits_{u\in[v_K^j, v_{K-1}^j]} [f(u) - X_{K,i}^j u] \ , & \text{if} \quad v_K^j < v_{k-1}^j \end{cases} ,$$

and where

$$X_{k,i}^j = \frac{\Delta x_i^j}{\Delta^+ t^j} \ .$$

The index k (or l) is chosen such that the line segment $(x_i^{j+1},t^{j+1}),(x_k^j,t^j)$ (or $(x_{i+1}^{j+1},t^{j+1}),(x_l^j,t^j)$) nowhere transversaly crosses any Riemann fan.

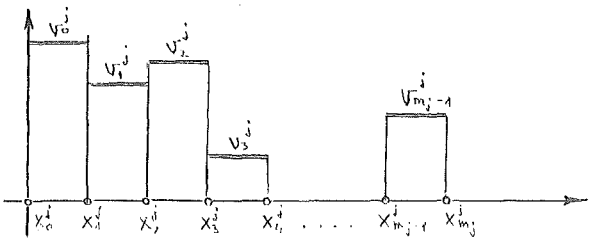
In order to test the method numerically, we have made a computer program which solves the BLN problem. Our program is only one of many possible implementations of the method. A grid of points (x_i^j,t^j), $j=0,\ldots,n$, $i=0,\ldots,m_j$ is automatically constructed: The grid covers the domain of our problem, i.e. $x_j^0=0, x_{m_j}^j=L$ $t^0=0, t^n \geq T$



The parameters n, m_j $(j=1,\ldots,n)$, $t^j (j=1,\ldots,n)$ are chosen by the program during the computation. On one time layer (for fixed j) the grid is uniform, i.e.:

$$x_1^j - x_0^j = x_2^j - x_1^j = \ldots = x_{m_j}^j - x_{m_j-1}^j = L/m_j = :h_j$$

Yet, the whole grid is still nonuniform, since m_j and $\Delta t^j = t^{j+1} - t^j$ depend on j. There is even stronger regularity among the numbers m_j: m_{j+1} can be either equal to m_j or two times greater than m_j or two times less than m_j. For given t^j, the solution $u(x,t^j)$ of the BLN problem is approximated by a step function:

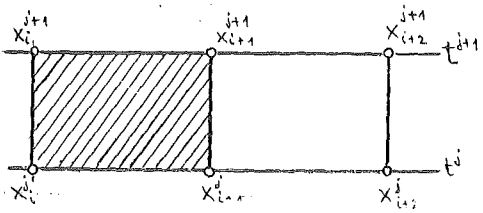


The set of all computed step functions is stored in random-access files. Additional modules of the program use these files to produce printed or plotted reports containing approximate versions of the functions $u(x,t^j)$ (for some user-specified values t^j).

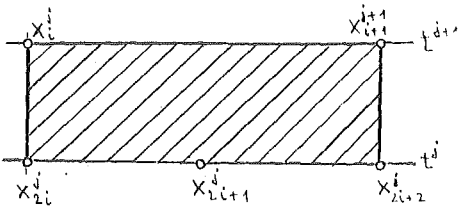
Now, we will describe essential features of the algorithm used in our program:

- The computation is advancing time layer by time layer. The grid is being constructed in the course of computation
- In order to construct the next time layer, the program selects one of three possible patterns:

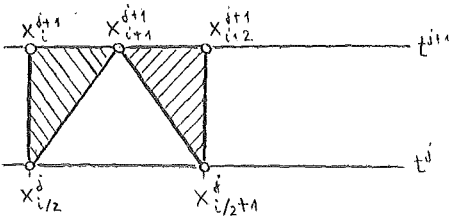
1^o the grid has equal number of x-intervals on j-th and on (j+1)-th time layer



2^o the grid is two times sparser on (j+1)-th time layer



3^o the grid is two times denser on (j+1)-th time layer



- The decision which pattern to choose is based on the following simple heuristics:

"If the function $u(x,t^j)$ is oscillating very much and/or has big discontinuities, then it should be computed more precisely (i.e. with denser grid)"

To measure oscillations and discontinuities of the function $u(x,t^j)$, the following variational norm is introduced:

$$d^j = \sum_{j=0}^{m_j} (v_i^j - v_{i-1}^j)^2, \text{ where } v_{-1}^j = a_0,\ v_{m_j}^j = a_L$$

If d^j (computed using the grid resulting from pattern 1^o) is significantly greater than d^j, then pattern 3^o, is rather used. Else if d^{j+1} is significantly less than d^j, then pattern 2^o is rather used.

- Time step Δt^j is chosen so as to keep the quotient $\Delta t^j/h^j$ constant through the whole computation. Initial value $\Delta t^0/h^0$ is determined as to satisfy the Courant condition.

- To compute the next step-function (on j+1-th time layer) the program uses formulae for v_i^{j+1} which are derived from more general formula (21). The trapesium (used in formula (21)) is substituted by a rectangle (pattern 1^o, pattern 2^o) or by one of the triangles (pattern 3^o). On figures illustrating the patterns, rectangles and triangles used are shaded.
- Since the quotient $\Delta t^j/h^j$ is kept constant, the "Godunov function" (used in formula (21)) can be replaced by three simpler functions:

$$g_0(v,\overline{v}) = \begin{cases} \min\limits_{u\in[v,\overline{v}]} f(u) \ , & \text{for } v \le \overline{v} \\ \max\limits_{u\in[\overline{v},v]} f(u) \ , & \text{for } v > \overline{v} \end{cases}$$

$$g_{-1}(v,\overline{v}) = \begin{cases} \min\limits_{u\in[v,\overline{v}]} \left(f(u) - \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v \le \overline{v} \\ \max\limits_{u\in[\overline{v},v]} \left(f(u) - \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v > \overline{v} \end{cases}$$

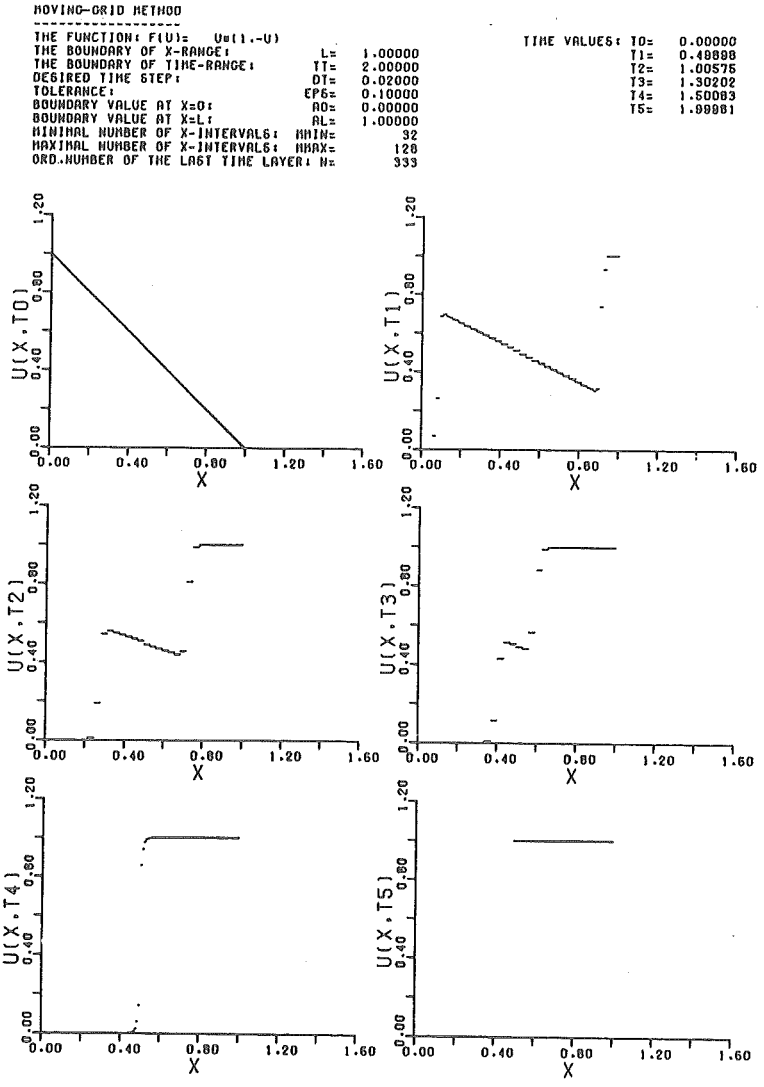
$$g_1(v,\overline{v}) = \begin{cases} \min\limits_{u\in[v,\overline{v}]} \left(f(u) + \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v \le \overline{v} \\ \max\limits_{u\in[\overline{v},v]} \left(f(u) + \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v > \overline{v} \end{cases}$$

- Each evaluation of any of the functions g_0, g_{-1}, g_1 involves a constrained optimization problem. In order to solve these optimization problems efficiently, our program initially finds all local extrema of functions

$$f(u) \ , \ f(u) - \frac{h^o}{2\Delta t^o} u \ , \ f(u) + \frac{h^o}{2\Delta t^o} u \ .$$

The table of local extrema is used whenever one of the functions g_0, g_{-1}, g_1 is being evaluated.

The program was tested on a number of examples involving three different f-functions. The results were compared with known exact solutions (or numerical solutions obtained by a different method) as given in papers [3], [4]. There is a good accordance between the computed and expected values. On the following page a plotted report generated by our program is reproduced. All data describing the corresponding BLN problem are quoted. Approximate versions of the functions



$u(x,t^j)$ for six different values of t^j are plotted. Since the value t^o is equal to 0, the initial condition is also visible.

REMARKS. The work described in this paper was carried out as a part of authors' collaboration with INA Naftaplin petroleum Industry from Zagreb, Yugoslavia. The considered BLN problem has an interpretation which arises in the study of one-dimensional flow of two immiscible fluids (i.e. oil and water) through a porous medium. By solving a series of BLN problems and comparing the solutions with experimental results, one can estimate parameters describing physical properties of a given porous material. This is an important step leading to a reliable petroleum reservoir simulation.

REFERENCES:

1. C.BARDOS, A.Y.LEROUX and J.C.NEDELEC: First order quasilinear equations with boundary conditions, Rapp.Intervue CMA 38(1978)
2. Y.BRENIER and S.OSHER: Approximate Riemann solvers and numerical flux functions, SIAM J.Numer.Anal.2 (1986), 259-273.
3. G.CHAVENT and G.SALZANO: 1-D water flooding problem with gravity, J.Comput.Phys. 45(1982),307-343.
4. P.CONCUS and W.PROSKUROWSKI: Numerical solutions of a nonlinear hyperbolic equation by the random choice method, J.Comput.Phys. 30(1979),153-166.
5. A.Y.Le ROUX: Etude de probleme mixte pour une equation quasi-lineaire du premier ordre, C.R.Acad.Sc.Paris 285(1977), 351-354
6. R.SANDERS: The moving grid method for nonlinear hyperbolic conservation laws, SIAM J.Numer.Anal. 4(1985), 713-728.
7. A.I.VOLJPERT, S.I.HUDJAJEV: Analiz v klassah razryvnyh funkcij i uravnenija matematičeskoj fiziki, Nauka, Moskva, 1975.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

THE SPLINE TRANSFORM AND ITS APPLICATION IN THE PROBLEMS OF SIGNALS' DIGITAL TREATMENT

A.H. ARAKELIAN and M.R. VOSKANIAN

ABSTRACT: In this work the calculating formulas for computing the spectral characteristics are brought, obtained by the application of wide class of splines. The programmes of computing the spectral characteristics were used for investigating the medical-biological curves.

The practical application of splines shows, that for obtaining a considerable degree of closeness of a spline to the interpolating function, it is sufficient that the degree of splines be limited by four.

INTRODUCTION

A great number of papers is devoted to the treatment and prophylaxis of postcholecystectomy syndrome. But the number of research works on usage of differentiated health resort treatment complexes depending upon clinical variations of postcholecystectomy syndrome course is as far extremely insufficient. In a number of papers the significance of sanatoria and health resort treatment using mineral waters at early period after cholecystectomy is especially emphasized as a prophylactic method of serious complications after cholecystectomy.

Relative to the problem we have supposed that it would be timely to make clear the possibility and expedience of the usage at early periods after operation on bilare tract the health resort factors in particular, mineral water "Jermuk" of complex chemical composition, with the purpose of prophylaxis of postcholecystectomy syndrome and most rapid restoration of working capacity. There are no information concerning effect of Armenian mineral waters both under health resort and under common conditions on patient rehabilitation at early periods after cholecystectomy.

The paper presented is devoted to the investigation of the effect of complex treatment method developed for the patients after cholecystectomy operation, on the disease course. The investigation was conducted by means of rheopatography namely, utilization of registration of hepatic blood supply regularities by means of rheohepatograms (RHG). Fig.1 presents a typical RHG-registration.

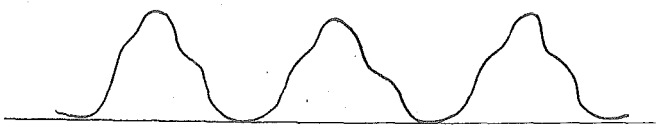


Fig. 1.

1. THE AIM OF INVESTIGATIONS

The basic aim of investigations conducted consists in the development of effective methods of patients early rehabilitation after cholecystectomy by physical factors depending upon the character of pathologic process in hepatobiliary system, in the estimation of efficiency, and in the recommendations for these methods usage.

The pecularities of clinical course of the patients condition after cholecystectomy, the effect of mineral water "Jermuk" combined with pine bath, massage of portal fissure zone muscles, remedial gymnastics on the patient condition after cholecystectomy, laboratory indices characterizing the hepatobiliary system condition, the alteration of liver hemodynamics according to rheohepatography data and modification dynamics of RGH-curve spectral characteristics were investigated.

2. METHODS OF SPECTRAL ANALYSIS

75 patients were observed. 50 of them have been transferred from surgical clinic to gastroenterological department in 2-3 weeks after cholecystectomy 25 patients had a period from 6 months to 6 years after operation. Almost all the patients investigated had a pain in the right hypochondrium, discomfort in epigastric region after eating, and at times nausea, heartburn, bitter taste and xerostomia. Often the patients have mentioned disorders of intestine emptying function.

In the first group of patients the phenomen of asthenic syndrome has been observed. The patients from both groups have received the same treatment complex during 24-26 days of hospital treatment. The analysis of data received has shown that for 95,8% patients in first group and 76,5% patients in second group the pain in the right hypochondrium has disappeared and for the others the pain intensity has essentially decreased. An analogous effect was observed for the pain in epigastric region. But, unlike the pain a number of patients have continued to complain of gastric dyspepsia effects in particular, heartburn, eructations although with decreased frequency of their appearance. For almost all the patients bitter taste, gastric flatulence, asthenic effects (weakness, erethism) have disappeared, defecation has normalized. For the registration of hepatic blood supply character the active electrode was placed across the medioclavicular line to the right, in the region of its intersection with the costal arch, and the passive one across the medioscapular line to the right, in the center between the angle of the scapulae and the crest of the iliac bone.

For calculation of integral Fourier transformation a method based on the approximate representation of integrand function by means of Hermite spline was used [1,3]. The spectral processing of the signals was conducted on a computer in real-time.

Special attention was drawn to functional condition of liver affected mostly by cholelithiasis. The liver condition was investigated by means of rheohepatography. Analysis of used treatment complex results was conducted by means of RHG-curve registration received by 4RG1A apparatus.

The interesting RHG characteristics are splash values m=0,038 Hz before and n=0,07 Hz after treatment. Fig.2 gives the typical shape of RHG spectrum before and after treatment with distinguished peaks. The typical frequency values are about 0,036-0,04 Hz for m and about 0,067-0,071 Hz for n.

3. ALGORITHM OF RHG SPECTRAL ANALYSIS

In the investigations conducted a calculation procedure based on RHG-curve Fourier transform represented by Hermite spline is used. The method permits in contrast with the visual one not only to reduce the investigation time but also to free the investigator from elements of subjectively peculiar to the visual method [2,3].

The calculation procedure of amplitude-frequency characteristic determination proceeds as follows:

Let $f(x) \in C^m[a,b]$ function be analysed where $C^m[a,b]$ is a space of real functions continuous on [a,b] interval and has continuous derivatives of m-degree.

Let

$$\Delta_n : a = x_0 < x_1 < x_2 < \ldots < x_n = b, \quad n \in N$$

is a net given on a finite interval [a,b].

Let us denote $(x-y_i)_+^{2m} = max[0; (x-y_i)]^m$.

Definition [1]. $S_{2m}(x) = S_{2m}(x; f)$, $m \geq 1$ function is called Hermite interpolation spline for a function $f(x) \in C^m[a,b]$, $m \in N$, if

a) $S_{2m}(x; f) = C^m[a,b]$ and

$$S_{2m}(x; f) = \sum_{s=0}^{2m} \frac{a_s^{(i)}}{s!}(x-x_i)^s + \frac{a_{2m+1}^{(i)}}{(2m)!}(x-y_i)_+^{2m}$$

for every $x \in [x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$.

b) $S_{2m}^{(k)}(x_i) = f^k(x_i)$,

$k = 0, 1, \ldots, m$; $i = 0, 1, \ldots, n$.

If $f(x) \in C^m[a,b]$, then [1,4] spline transform of Fourier for Hermite spline representation of $f(x)$ function is equal to $\bar{H}(j\omega) = \frac{1}{2\pi}\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} S_{2m}(x; f) =$

$$\frac{1}{2\pi}\left[\sum_{i=0}^{n-1}\sum_{s=0}^{2m}\frac{a_s^{(i)}}{s!}\int_{x_i}^{x_{i+1}}(x-x_i)^s e^{-j\omega x}dx + \int_{y_i}^{x_{i+1}}(x-y_i)^{2m}e^{-j\omega x}dx\right].$$

If we denote

$$Q^{(i)}(s,\omega) = \int_{x_i}^{x_{i+1}}(x-x_i)^s e^{-j\omega x}dx; \quad L^i(2m,\omega) = \int_{y_i}^{x_{i+1}}(x-y_i)^{2m}e^{-j\omega x}dx,$$

then it is possible to construct the following iteration procedures for their definition

$$Q^{(i)}(0,\omega) = (e^{-j\omega x_i} - e^{-j\omega x_{i+1}})/j\omega; \quad Q^{(i)}(1,\omega) = [Q^{(i)}(0,\omega) - (x_{i+1}-x_i)e^{-j\omega x_{i+1}}]/j\omega;$$

$$Q^{(i)}(s,\omega) = [sQ(s-1,\omega) - (x_{i+1}-x_i)^s e^{-j\omega x_{i+1}}]/j\omega$$

and $L^i(0,\omega) = -(e^{-j\omega x_{i+1}} - e^{-j\omega y_i})/j\omega$; $L^{(i)}(1,\omega) = [L^{(i)}(0,\omega) - (x_{i+1}-y_i)e^{-j\omega x_{i+1}}]/j\omega$;

$$L^{(\kappa)}(\kappa,\omega) = [\kappa Q(\kappa-1,\omega) - (x_{i+1} - y_i)^{\kappa} e^{-j\omega x_{i+1}}]/j\omega$$

respectively.

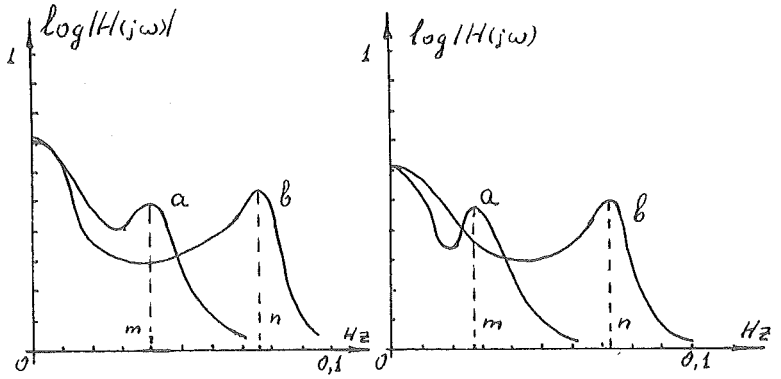
If we denote the Fourier transform of $f(x)$ function by $H(j\omega)$, then $|H(j\omega) - \bar{H}(j\omega)| = \frac{1}{2\pi} \left| \int_a^b (f(x) - S_{2m}(x; f)) e^{-j\omega x} dx \right| \leq$

$$\frac{1}{2\pi} (2m+1) \frac{\|\Delta_n\|^{2m} \tilde{\omega}(f^{(2m)}, \|\Delta_n\|)}{2^{2m}(2m)!},$$

where $\|\Delta_n\| = \max_{0 \leq i \leq n-1} (x_{i+1} - x_i)$, $\tilde{\omega}(f^{(\kappa)}, \|\Delta_n\|) = \max_{0 \leq i \leq n-1} \max_{y \in [x_i, x_{i+1}]} |f^{(\kappa)}(x) - f^{(\kappa)}(y)|$.

4. RHG SIGNAL SPECTRAL ANALYSIS

It is known that RHG represents an electrical signal generated by an non-stationary source, the liver. Fig.2 and 3 represent the results of RHG-curve spectral analysis for the first and second groups of patients, respectively.



The results of investigations conducted have shown that the increase speed of frequency value at which the spectrum peak is observed, is conditioned by intensity increase of hepatic blood flow at the expense of both arterial inflow and venous outflow. Besides, the first group of patients has larger values of spectral characteristics than second one.

Thus, the observations have shown that the used treatment complex including inner dose of carbonate-hydrocarbonate-sulphate-chlorine-natrium-calcium-magnesium mineral water "Jermuk" permits to improve essentially the liver

hemodynamics for both groups of patients. At the same time information obtained testifies more evidently expressed decrease of liver hypoxia, for the first group of patients for which the rehabilitating treatment began at earlier period after operation.

5. CONCLUSIONS

1. The rehabilitating treatment involving the balneologic factors and conducted at early period after cholecystectomy promotes the favourable dynamics of post-operational syndromes, the normalization of lives functional condition, circulation of the blood in liver, the most rapid restoration of working capacity.

2. The use of signal digital processing methods and algorithms and their realizing programs stipulates for possibility of objective quantitative estimation of RHG-curve. The problem of determining when and under what conditions RHG-curve changes was so far not solved. The RHG spectral processing method used in the paper permits to receive the quantitative information about violations of hepatic blood supply character.

References

1. ARAKELIAN A.H.: Optimization methods of dialogue information systems for testing, Acad. Science of Armenian SSR, Yerevan, 1985.
2. ARAKELIAN A.H., AGAIAN S.S.: On an algorithm of spectral analysis, Cyber. and Syst. Res. 2, 1984. R. Trappl. North-Holl., pp.273-275.
3. VOSKANIAN M.R., ARAKELIAN A.H.: The Spectr. Analy. of EGH and its Application. IN: Proceedings of a IX All-Union Conf. on Operational Problems. M.1983, p.394
4. VOSKANIAN M.R., KHONDKARIAN N.S.: Spline transform of Fourier in the Problems of RHG Spectral Analysis. In: Proceedings of a II All-Union Conference. "The Realization of Mathematical Methods in Clinical and Experimental Medicine", Moscow, 1986, pp.92-93.

AN IMPLEMENTATION OF A SEMI-DEFINITE PROGRAMMING METHOD TO CHEBYSHEV APPROXIMATION PROBLEMS

M.D. AŠIĆ and V.V. KOVAČEVIĆ-VUJČIĆ

ABSTRACT: A discretization method for solving linear semi-infinite programming problems arising from Chebyshev approximation is presented. The method is based on selective refinement of the initial coarse grid, which enables an efficient treatment of multidimensional problems. Numerical examples from Chebyshev approximation are also presented.

1. INTRODUCTION

This paper is a natural extension of a sequence of papers on semi-infinite programming methods ([2] , [3] , [4] , [5] , [6]). We consider here the following Chebyshev approximation problem:

Let $C=[p_1,q_1]\times\ldots\times[p_r,q_r]$ and let $g_i:C\to R$, $i=1,\ldots,m$ and $f:C\to R$ be given functions. Find $x_1,\ldots,x_m$ such that

$$\max_{t\in C} |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \tag{1}$$

is minimized.

It is easy to see that this problem can be reformulated as the linear semi-infinite programming problem:

$$\min x_{m+1}$$
$$x_{m+1}\geq |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \quad \text{for all } t\in C. \tag{2}$$

For brevity, let $x=(x_1,\ldots,x_{m+1})$,

$$c_1(x,t)=x_1g_1(t)+\ldots+x_mg_m(t)-x_{m+1}$$
$$c_2(x,t)=-x_1g_1(t)-\ldots-x_mg_m(t)-x_{m+1}.$$

Then (2) becomes

$$\min x_{m+1}$$
$$x\in X,\ X=\{x\in R^{m+1} \mid c_1(x,t)\leq f(t),\ c_2(x,t)\leq -f(t) \quad \text{for all } t\in C\}. \tag{3}$$

In the sequel we shall use the following:

Assumption 1. (i) There exists an $\bar{x}\in X$ such that the set

hemodynamics for both groups of patients. At the same time information obtained testifies more evidently expressed decrease of liver hypoxia, for the first group of patients for which the rehabilitating treatment began at earlier period after operation.

5. CONCLUSIONS

1. The rehabilitating treatment involving the balneologic factors and conducted at early period after cholecystectomy promotes the favourable dynamics of post-operational syndromes, the normalization of lives functional condition, circulation of the blood in liver, the most rapid restoration of working capacity.

2. The use of signal digital processing methods and algorithms and their realizing programs stipulates for possibility of objective quantitative estimation of RHG-curve. The problem of determining when and under what conditions RHG-curve changes was so far not solved. The RHG spectral processing method used in the paper permits to receive the quantitative information about violations of hepatic blood supply character.

References

1. ARAKELIAN A.H.: Optimization methods of dialogue information systems for testing, Acad. Science of Armenian SSR, Yerevan, 1985.
2. ARAKELIAN A.H., AGAIAN S.S.: On an algorithm of spectral analysis, Cyber. and Syst. Res. 2, 1984. R. Trappl. North-Holl., pp.273-275.
3. VOSKANIAN M.R., ARAKELIAN A.H.: The Spectr. Analy. of EGH and its Application. IN: Proceedings of a IX All-Union Conf. on Operational Problems. M.1983, p.394
4. VOSKANIAN M.R., KHONDKARIAN N.S.: Spline transform of Fourier in the Problems of RHG Spectral Analysis. In: Proceedings of a II All-Union Conference. "The Realization of Mathematical Methods in Clinical and Experimental Medicine", Moscow, 1986, pp.92-93.

AN IMPLEMENTATION OF A SEMI-DEFINITE PROGRAMMING METHOD TO CHEBYSHEV APPROXIMATION PROBLEMS

M.D. AŠIĆ and V.V. KOVAČEVIĆ-VUJČIĆ

ABSTRACT: A discretization method for solving linear semi-infinite programming problems arising from Chebyshev approximation is presented. The method is based on selective refinement of the initial coarse grid, which enables an efficient treatment of multidimensional problems. Numerical examples from Chebyshev approximation are also presented.

1. INTRODUCTION

This paper is a natural extension of a sequence of papers on semi-infinite programming methods ([2] , [3] , [4] , [5] , [6]). We consider here the following Chebyshev approximation problem:

Let $C=[p_1,q_1]\times\ldots\times[p_r,q_r]$ and let $g_i:C\to R$, $i=1,\ldots,m$ and $f:C\to R$ be given functions. Find $x_1,\ldots,x_m$ such that

$$\max_{t\in C} |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \tag{1}$$

is minimized.

It is easy to see that this problem can be reformulated as the linear semi-infinite programming problem:

$$\min x_{m+1}$$
$$x_{m+1}\geq |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \quad \text{for all } t\in C. \tag{2}$$

For brevity, let $x=(x_1,\ldots,x_{m+1})$,

$$c_1(x,t)=x_1g_1(t)+\ldots+x_mg_m(t)-x_{m+1}$$
$$c_2(x,t)=-x_1g_1(t)-\ldots-x_mg_m(t)-x_{m+1}.$$

Then (2) becomes

$$\min x_{m+1}$$
$$x\in X,\ X=\{x\in R^{m+1} \mid c_1(x,t)\leq f(t),\ c_2(x,t)\leq -f(t) \quad \text{for all } t\in C\}. \tag{3}$$

In the sequel we shall use the following:

Assumption 1. (i) There exists an $\bar{x}\in X$ such that the set

$$\bar{X} = X \cap \{x \in R^{m+1} \mid x_{m+1} \leq \bar{x}_{m+1}\}$$

is bounded.

(ii) The functions $g_1,\ldots,g_m$ and f satisfy the Lipschitz condition.

It is clear that Assumption 1 implies the existence of a uniform Lipschitz constant L for functions $c_1(x,t)$, $c_2(x,t)$ and $f(t)$, i.e.

$$|c_i(x,t')-c_i(x,t")| \leq L\|t'-t"\| \quad , \quad x \in \bar{X}, \quad i=1,2$$
$$|f(t')-f(t")| \leq L\|t'-t"\| .$$

The main idea of the method which will be described in Section 2 is to use selective discretization of the index set C in order to replace semi-infinite programming problem (3) by a sequence of linear programming problems. The method starts with a uniform grid which depends on the Lipschitz constant L and suc cessive refinements are made in such a way to ensure linear grow of the number of grid points, while retaining the usual converge ce properties.

2. THE METHOD

In order to describe the algorithm of the method we need th following notation:

Let (M_j) denote the sequence of uniform discretizations of the set C defined by

$$M_j = \{(p_1+k_1h_1^j,\ldots,p_r+k_rh_r^j) \mid k_i \in \{0,1,\ldots,2^j m_i\} \ , \ i=1,\ldots,r\} \ ,$$

where $h_i^j=(q_i-p_i)/(2^j m_i)$, and m_i are appropriately chosen posit ve integers. Furthermore, for given $y \in R^{m+1}$, $t \in C$, $h_1>0,\ldots,h_r>0$ let q_1 and q_2 be the functions defined by

$$q_1(s) = c_1(y,t)-f(t)+ \sum_{i=1}^{r} \bar{A}_{i1}(s_i-t_i) + \sum_{i=1}^{r} \bar{B}_{i1}|s_i-t_i|$$

(4)

$$q_2(s) = c_2(y,t)+f(t)+ \sum_{i=1}^{r} \bar{A}_{i2}(s_i-t_i) + \sum_{i=1}^{r} \bar{B}_{i2}|s_i-t_i| \ ,$$

where $\bar{A}_{i1},\bar{A}_{i2},\bar{B}_{i1},\bar{B}_{i2}$ are such that

$$q_1(s) \geq c_1(y,s)-f(s), \quad q_2(s) \geq c_2(y,s)+f(s)$$

for all s satisfying $|s_i-t_i| \leq h_i$, $i=1,\ldots,r$. It is obvious that and q_2 depend also on $y,t,h_1,\ldots,h_r$ and that they are actually piecewise linear majorants of c_1-f and c_2+f, respectively.

lgorithm 1. Input parameters: $\beta_0>0$, Lipschitz constant L, integers $m_1,\ldots,m_r$ satisfying

$$m_i>(q_i-p_i)L\sqrt{r}/(2\beta_0),\quad i=1,\ldots,r.$$

tep 0. Set $\gamma_0=\beta_0/L$, $C_0=M_0$, $k=0$.

tep 1. Solve the linear programming problem:

$$\min x_{m+1}$$

$$x\in Y_k,\quad Y_k=\{x\in R^{m+1} \mid c_1(x,t)\le f(t)-\beta_k,\ c_2(x,t)\le -f(t)-\beta_k,\ t\in C_k\}$$

nd let y be a solution. Set $j=0$, $E_0=C_0$.

tep 2. If $j=k$ go to Step 4. Otherwise, for each $t\in E_j$ find unctions q_1, q_2 such that

$$q_1(s)\ge c_1(y,s)-f(s),\quad q_2(s)\ge c_2(y,s)+f(s),$$

or all s satisfying $|s_i-t_i|\le h_i^j$, $i=1,\ldots,r$.

Let E'_{j+1} be the set of points $t\in E_j$ for which either q_1 or q_2 s greater than or equal to $-\beta_k$ at some extreme point of the set

$$([t_1-h_1^j,t_1+h_1^j]\times\ldots\times[t_r-h_r^j,t_r+h_r^j])\cap C.$$

et E_{j+1} be the set of points in M_{j+1} whose distance from the set , $j+1$ does not exceed γ_j. Replace j by j+1.

tep 3. If for all $t\in E_j$

$$c_1(y,t)\le f(t)-\beta_k,\quad c_2(y,t)\le -f(t)-\beta_k,$$

o to Step 2. Otherwise, let

$$T=\{t\in E_j \mid c_1(y,t)>f(t)-\beta_k \text{ or } c_2(y,t)>-f(t)-\beta_k\},$$

eplace C_k by $C_k\cup T$ and go to Step 1.

tep 4. Set $x^{k+1}=y$, $C_{k+1}=C_k$, $\beta_{k+1}=\beta_k/2$, $\gamma_{k+1}=\gamma_k/2$, replace k by +1 and go to Step 1.

It is easy to see that Algorithm 1 belongs to the class of ethods defined in [4]. The result in [4] implies that the Alorithm is well defined and that each cluster point of the sequnce (x^k) generated by the Algorithm is a solution to (3). Moreover, $x^k\in X$ for all $k=0,1,\ldots$.

The efficiency of the Algorithm depends on the cardinalities of the sets E_j and on the number of inner cycles of the type Step 1 → Step 2 → Step 3 → Step 1 at the k-th iteration, which determines the cardinality of the set C_k. For further analysis we need the following:

Assumption 2. (i) x^* is the unique solution to (3).

(ii) $(c_1(x^*,t)-f(t))(c_2(x^*,t)+f(t))=0$ for finitely many t's. Let T^* be a complete list.

(iii) Functions $f,g_1,\ldots,g_m$ are twice continuously differentiable on C.

(iv) For each $t^* \in T^*$ the following property holds: If t_i^* is an endpoint of $[p_i,q_i]$ and, say, $c_1(x^*,t^*)-f(t^*)=0$ then $\partial(c_1(x^*,t^*)-f(t^*))/\partial t_i \neq 0$. Moreover, the Hessian matrix with respect to the remaining t_i's is negative definite at t^*. Similar property holds if $c_2(x^*,t^*)+f(t^*)=0$.

The following result on the cardinalities of the sets E_j ho

Theorem 1. Let Assumptions 1 and 2 be satisfied and let q_1 a q_2 have the form (4). Moreover, assume that $\bar{B}_{i1} \to 0$, $\bar{B}_{i2} \to 0$ as $h_1 \to 0,\ldots,h_r \to 0$ for all $i=1,\ldots,r$. Then the cardinalities of the sets E_j generated by Algorithm 1 are bounded above by a constant independent on j and k.

The proof is similar to that of Theorem 3.1 in [5] and is t omitted. Let us only point out that under Assumption 2 functions q_1 and q_2 of the type (4) satisfying $\bar{B}_{i1} \to 0$, $\bar{B}_{i2} \to 0$ as $h_1 \to 0,..$ $h_r \to 0$, $i=1,\ldots,r$ can be obtained using the first order Taylor e pansion of c_1-f and c_2+f and rounding up the remainder term.

It remains to analyze the number of inner cycles at the k- iteration, which is done by the following theorem:

Theorem 2. Assume that the functions $g_1,\ldots,g_m$ satisfy the condition on the set C and $f \notin \operatorname{span}\{g_1,\ldots,g_m\}$. Suppose furth more that Assumptions 1 and 2 hold. Then the number of cycles the type Step 1 → Step 2 → Step 3 → Step 1 in Algorithm 1 is bounde a constant independent on k.

The proof follows directly from the following three Lemma

Lemma 1. Suppose that the assumptions of Theorem 2 are sati fied. Let y be one of the points generated in Step 1 of Algori during the (k+1)-th iteration. Then there is a positive consta independent on k such that

$$\bar{y}_{m+1} \leq y_{m+1} \leq \bar{y}_{m+1} + D\beta_{k+1}^2 \quad , \text{ where } \bar{y}=(x_1^k,\ldots,x_m^k,x_{m+1}^k-\beta_{k+1}$$

The proof of Lemma 1 is similar to that of Lemma 3.1 in [6 and is omitted.

Lemma 2. Suppose that the assumptions of Theorem 2 are satisfied Let y be one of the points generated in Step 1 of Algorithm 1 during the iteration k+1. Then there is a positive constant E independent on k such that

$$\|y-\bar{y}\| \leq E\beta_{k+1}^2, \text{ where } \bar{y}=(x_1^k,\ldots,x_m^k,x_{m+1}^k-\beta_{k+1}).$$

Proof : Note first that $\bar{y}$ is a solution to the problem

$$\min x_{m+1}$$

$$c_1(x,t)\leq f(t)-\beta_k,\ c_2(x,t)\leq -f(t)-\beta_k,\ t\in C_k.$$

By duality theorem there are nonpositive numbers $d_1(t)$, $d_2(t)$, $t\in C_1$ such that

$$\sum_{t\in C_k} d_1(t)g_i(t)-\sum_{t\in C_k} d_2(t)g_i(t) = 0,\ i=1,\ldots,m$$

$$\sum_{t\in C_k} d_1(t)+\sum_{t\in C_k} d_2(t) = -1.$$

By Caratheodory's theorem we may assume that

$$\sum_{j=1}^{m'} d_1(t^j)g_i(t^j) - \sum_{j=m'+1}^{m''} d_2(t^j)g_i(t^j) = 0,\ i=1,\ldots,m$$

$$\sum_{j=1}^{m'} d_1(t^j) + \sum_{j=m'+1}^{m''} d_2(t^j) = -1, \tag{5}$$

where $d_1(t^j)<0$, $j=1,\ldots,m'$, $d_2(t^j)<0$, $j=m'+1,\ldots,m''$ and $m''\leq m+1$. It easily follows that for k large enough each t^j is in the neigh borhood of some point in T^*. Without loss of generality we may as sume that these points are $t^{*1},\ldots,t^{*\bar{m}'},t^{*\bar{m}'+1},\ldots,t^{*\bar{m}''}$ and that the corresponding neighborhoods are disjoint. It is clear that $\bar{m}''\leq m''\leq$ We will show that $\bar{m}''=m+1$.

Assume the contrary and let

$$F = \begin{bmatrix} g_1(t^{*1}) & \ldots & g_1(t^{*\bar{m}'}) & -g_1(t^{*\bar{m}'+1}) & \ldots & -g_1(t^{*\bar{m}''}) \\ \vdots & & & & & \\ g_m(t^{*1}) & \ldots & g_m(t^{*\bar{m}'}) & -g_m(t^{*\bar{m}'+1}) & \ldots & -g_m(t^{*\bar{m}''}) \\ 1 & \ldots & 1 & 1 & \ldots & 1 \end{bmatrix}.$$

Due to the Haar condition, the system

$$F^T u = [1\ \ldots\ 1]^T$$

$$u_{m+1} = -1$$

has a solution $\bar{u}_1,\ldots,\bar{u}_m,\bar{u}_{m+1}$. Moreover, for k large enough,

$$g_1(t^j)\bar{u}_1+\ldots+g_m(t^j)\bar{u}_m+\bar{u}_{m+1}>0\ ,$$
$$-g_1(t^j)\bar{u}_1-\ldots-g_m(t^j)\bar{u}_m+\bar{u}_{m+1}>0\ .$$

Multiplying the equalities (5) by $\bar{u}_i$'s and adding we obtain:

$$0>\sum_{j=1}^{m'} d_1(t^j)(g_1(t^j)\bar{u}_1+\ldots+g_m(t^j)\bar{u}_m+\bar{u}_{m+1}) +$$
$$+\sum_{j=m'+1}^{m''} d_2(t^j)(-g_1(t^j)\bar{u}_1-\ldots-g_m(t^j)\bar{u}_m+\bar{u}_{m+1}) = -\bar{u}_{m+1} = 1,$$

which is a contradiction.

Hence, $\bar{m}''=m+1$ and (5) holds with $m''=m+1$. It is easy to see now that $d_1(t^j)$ and $d_2(t^j)$ in (5) are bounded above by a negative constant G. Multiplying the equalities (5) by $\bar{y}_1,\ldots,\bar{y}_m,-\bar{y}_{m+1}$, respectively, and adding, we obtain

$$x^k_{m+1}-\beta_{k+1}=\sum_{j=1}^{m'} d_1(t^j)c_1(\bar{y},t^j)+\sum_{j=m'+1}^{m+1} d_2(t^j)c_2(\bar{y},t^j)\ ,$$

which implies

$$x^k_{m+1}-\beta_{k+1}=\sum_{j=1}^{m'} d_1(t^j)f(t^j)-\sum_{j=m'+1}^{m+1} d_2(t^j)f(t^j)+\beta_{k+1}. \tag{6}$$

Let y be an arbitrary point generated at Step 1 during the iteration k+1. Then

$$\begin{aligned} c_1(y,t^j) &= f(t^j)-\beta_{k+1}+v_j,\ j=1,\ldots,m' \\ c_2(y,t^j) &= -f(t^j)-\beta_{k+1}+v_j,\ j=m'+1,\ldots,m+1, \end{aligned} \tag{7}$$

where $v_j\leq 0$, $j=1,\ldots,m+1$. Multiplying the equalities (5) by $y_1,\ldots,y_m,-y_{m+1}$, respectively, and adding, we obtain

$$y_{m+1}=\sum_{j=1}^{m'} d_1(t^j)c_1(y,t^j)+\sum_{j=m'+1}^{m+1} d_2(t^j)c_2(y,t^j),$$

which by (7) implies

$$y_{m+1}=\sum_{j=1}^{m'} d_1(t^j)(f(t^j)-\beta_{k+1}+v_j)+\sum_{j=m'+1}^{m+1} d_2(t^j)(-f(t^j)-\beta_{k+1}+v_j).$$

Now using (6) we obtain

$$y_{m+1}=x^k_{m+1}-\beta_{k+1}+\sum_{j=1}^{m'} d_1(t^j)v_j+\sum_{j=m'+1}^{m+1} d_2(t^j)v_j\leq x^k_{m+1}-\beta_{k+1}+D\beta^2_{k+1}$$

where the inequality follows from Lemma 1. Hence,

$$\sum_{j=1}^{m'} d_1(t^j)v_j+\sum_{j=m'+1}^{m+1} d_2(t^j)v_j\leq D\beta^2_{k+1}\ ,$$

so that

(8) $$0 \geq v_i \geq D\beta_{k+1}^2/G \ .$$

Note that y and $\bar{y}$ can be thought of as solutions to system of linear equations (7) and the corresponding system when v_i's are replaced by 0. By Cramer's rule we obtain

$$\|y-\bar{y}\| \leq D_1(|v_1|+\ldots+|v_n|) \ ,$$

where D_1 does not depend on k. Finally, (8) yields

$$\|y-\bar{y}\| \leq E\beta_{k+1}^2 .$$

Lemma 3. Suppose that the assumptions of Theorem 2 are fulfille and let $\bar{y}=(x_1^k,\ldots,x_m^k,x_{m+1}^k-\beta_{k+1})$, $w_1(t)=c_1(\bar{y},t)-f(t)+\beta_{k+1}$, $w_2(t)=c_2(\bar{y},t)+f(t)+\beta_{k+1}$. Let $\hat{t}$ be any point added to C_k during the iteration k+1. Then for k large enough either $w_1(t)$ or $w_2(t)$ has a local maximum $\bar{t}$ such that $\|\hat{t}-\bar{t}\| \leq F\beta_{k+1}$, where F does not depend on k

The proof of Lemma 3 and Theorem 2 is analogous to the proof of Lemma 3.3 and Theorem 3.2 in [6].

Let us note that an immediate consequence of Theorem 2 is that the cardinality of the sets C_k generated by Algorithm 1 grows at most linearly with k. Theorems 1 and 2 also imply that the total number of points generated by the algorithm at the k-th iteration is bounded above by a function linear in k, while at the same time the cardinality of the uniform grid M_k depends exponentially on k. Numerical experience seems to indicate that this linear behaviour is retained also when Haar's condition is omitted. It should be pointed out that the existing discretization methods (see e.g. [9], see also [8]) have an exponential upper bound on the number of points generated at the k-th step.

3. NUMERICAL EXPERIENCE

The method described in Section 2 was tested on a number of test problems, mostly taken from [1] and [7] . The obtained result agree very well with the data in the literature. Here we give the details for three examples. In the corresponding tables $N(C_k)$ and $N(E_j)$ stand for the cardinality of C_k and the average cardinality of E_j at the k-th iteration, respectively.

Example 1.[1]. Approximate $(t_1)^2t_2$ by $v_1=1$, $v_2=t_1$, $v_3=(t_1)^2$, $v_4=t_2$, $v_5=(t_2)^2$, $v_6=t_1t_2$ on $(t_1)^2+(t_2)^2 \leq 1$.

Following the authors in [1] the problem is reduced to the approximation problem on $[0,1] \times [0,2\pi]$. Input parameters are : $\beta_0=4.4$, $L=10.5$, $m_1=2$, $m_2=11$.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$N(C_k)$	36	42	46	50	54	58	62	66	70	74	78	82	86	90
$N(E_j)$	86	64	61	60	53	47	47	47	41	45	46	41	42	41

x^{13}=(0.0000,0.0000,0.0000,0.2500,0.0000,0.0000).

Exact solution x^*=(0,0,0,1/4,0,0).

Example 2.[7] . Approximate $\exp(-(t_1)^2-t_2)$ by functions $v_1=1$, $v_2=t_1$, $v_3=t_2$, $v_4=2(t_1)^2-1$, $v_5=t_1t_2$, $v_6=2(t_2)^2-1$ on the set $[0,1] \times [0,1]$.

Input parameters are: $\beta_0=2.5$, $L=24.2$, $m_1=m_2=7$.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$N(C_k)$	64	70	76	79	83	85	88	93	97	100	103	106	108	111
$N(E_j)$	0	46	44	40	40	40	42	39	36	37	35	36	34	36

x^{13}=(0.9858,-0.3480,-0.9027,-0.1446,0.4246,0.1129).

Solution in [7] : (0.9858,-0.3480,-0.9027,-0.1446,0.4246,0.1129).

Example 3. [1] . Approximate t^2 by functions $v_1=t$, $v_2=\exp(t)$ on the interval $[0,2]$.

Input parameters are: $\beta_0=2$, $L=13$, $m_1=8$.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$N(C_k)$	9	11	12	13	14	15	16	17	18	19	20	21	22	23
$N(E_j)$	0	7	7	6	6	6	6	6	6	6	6	6	5	5

x^{13}=(0.1842,0.4186).

Solution in [1] : (0.1842,0.4186).

REFERENCES

1. D.D. ANDREASSEN and G.A. WATSON : Linear Chebyshev approximati without Chebyshev sets. BIT 16 (1976), 349-362.

2. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : An application of semi-infinite programming to approximation theory. In: Proceedings of the XI Yugoslavian Symposium on Operations Research, Herceg Novi, 1984, pp. 55-64.

3. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : A semi-infinite programming method and its application to boundary value problems. ZAMM 66 (1986), 403-405.

4. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : An interior semi-infinite programming method. JOTA 59 (1988) .

5. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : Computational complexity of some semi-infinite programming methods. In: System Modelling and Optimization (A. Prekopa, B. Strazicky, J.Szelezsan, eds.), Springer-Verlag, Berlin, 1986, pp. 34-42.

6. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : Linear semiinfinite programming problem: A discretization method with linearly growing number of points. In: Proceedings of 17. Jahrestagung "Mathematische Optimierung" (K. Lommatzsch ed .), Seminarbericht 85, Humboldt Universitat zu Berlin, 1986, pp. 1-10.

7. K. GLASHOFF and S.A. GUSTAFSON : Linear Optimization and Approximation. Springer-Verlag, Berlin, 1983.

8. R. HETTICH : A review of numerical methods for semi-infinite optimization. In: Semi-Infinite Programming and Applications (A.V. Fiacco, K.O. Kortanek, eds.), Springer-Verlag, Berlin, 1983, pp. 158-178.

9. R. HETTICH: An implementation of a discretization method for semi-infinite programming. Mathematical Programming 34 (1986), 354-361.

ON THE ZEROS OF A POLYNOMIAL

M. BIDKHAM and K.K. DEWAN

ABSTRACT: In this paper we have considered the problem of finding the maximum number of zeros in a prescribed region.

1. INTRODUCTION AND STATEMENT OF RESULTS

The following result is due to Mohammad [4]

THEOREM A. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ *be a polynomial of degree* n *such that*

$$a_n \geq a_{n-1} \geq \cdots \geq a_1 \geq a_0 > 0 ,$$

then the number of zeros of $p(z)$ *in* $|z| \leq \frac{1}{2}$ *does not exceed*

$$1 + \frac{1}{\log 2} \log \frac{a_n}{a_0}$$

As a generalization of Theorem A, Dewan [1] proved

THEOREM B. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ *be a polynomial of degree* n *with complex coefficients such that*

$$|\arg a_k - \beta| \leq \alpha \leq \pi/2 , \quad k = 0, 1, \ldots, n$$

for some real β , *and*

$$|a_n| \geq |a_{n-1}| \geq \cdots \geq |a_1| \geq |a_0| ,$$

then the number of zeros of $p(z)$ *in* $|z| \leq \frac{1}{2}$ *does not exceed*

$$\frac{1}{\log 2} \log \frac{|a_n|(\cos\alpha + \sin\alpha + 1) + 2\sin\alpha \sum_{k=0}^{n-1} |a_k|}{|a_0|}$$

THEOREM C. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ *be a polynomial of degree* n *with complex coefficients. If* $\operatorname{Re} a_k = \alpha_k$, $\operatorname{Im} a_k = \beta_k$, *for*

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

$$\alpha_n \geq \alpha_{n-1} \geq \ldots \geq \alpha_1 \geq \alpha_0 > 0 ,$$

then the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed

$$1 + \frac{1}{\log 2} \log \frac{\alpha_n + \sum_{k=0}^{n} |\beta_k|}{|a_0|} .$$

In this paper, we generalize Theorems A, B and C for different classes of polynomials which in turn also refine upon them. More precisely, we prove the following.

THEOREM 1. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n with complex coefficients. If for some real β, $|\arg a_i - \beta| \leq \alpha \leq \pi/2$, $0 \leq i \leq n$ and for some $0 < t \leq 1$

$$|a_0| \leq t|a_1| \leq \ldots \leq t^k|a_k| \geq t^{k+1}|a_{k+1}| \geq \ldots \geq t^n|a_n|, \quad 0 \leq k \leq n;$$

then the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed.

$$\frac{1}{\log 2} \log \frac{2t^{k+1}|a_k|\cos\alpha + 2t\sin\alpha \sum_{i=0}^{n} t^i|a_i| - (\cos\alpha + \sin\alpha - 1)\, t^{n+1}|a_n|}{t|a_0|} .$$

REMARK 1. For $t = 1$ and $k = n$ the above theorem reduces to Theorem B. If in addition to $t = 1$ and $k = n$, $\alpha = \beta = 0$ then it reduces to Theorem A.

THEOREM 2. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n with compex coefficients. If $\operatorname{Re} a_i = \alpha_i$, $\operatorname{Im} a_i = \beta_i$, for $i = 0, 1, \ldots, n$ and for some $0 < t \leq 1$

$$0 < \alpha_0 \leq t\alpha_1 \leq \ldots \leq t^k \alpha_k \geq t^{k+1} \alpha_{k+1} \geq \ldots \geq t^n \alpha_n$$

then the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed

$$1 + \frac{1}{\log 2} \log \left\{ \frac{t^{k+1}\alpha_k + t \sum_{i=0}^{n} |\beta_i| t^i}{t|\alpha_0|} \right\}$$

REMARK 2. For $k = n$ and $t = 1$, Theorem 2 reduces to Theorem C nd for $k = n$, $t = 1$ and $\beta_i = 0$, $0 \leq i \leq n$, it reduces to Theorem A.

The proof of next theorem follows on combining Theorem B and Lemma 2.

THEOREM 3. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n with complex coefficients such that

$$|\arg a_i - \beta| \leq \alpha \leq \pi/2 , \quad i = 0, 1, \ldots, n$$

for some real β, and

$$|a_n| \geq |a_{n-1}| \geq \cdots \geq |a_1| \geq |a_0|,$$

then the number of zeros of $p(z)$ in $R_2 \leq |z| \leq \frac{1}{2}$ does not exceed

$$\frac{1}{\log 2} \log \frac{|a_n|(\cos \alpha + \sin \alpha + 1) + 2 \sin \alpha \sum_{i=0}^{n} |a_i|}{|a_0|}$$

where R_2 is the same as defined in Lemma 2.

The above Thoerem is a refinement of Theorem B. In particular, if $\alpha = \beta = 0$, then it gives a refinement of Theorem A.

THEOREM 4. Let $p(z) = \sum_{i=0}^{n} a_i z^i$. If $\operatorname{Re} a_i = \alpha_i$, $\operatorname{Im} a_i = \beta_i$, for $i = 0, 1, \ldots, n$, and

$$\alpha_n \geq \alpha_{n-1} \geq \cdots \geq \alpha_1 \geq \alpha_0 > 0 , \quad \alpha_n > 0 ,$$

then the number of zeros of $p(z)$ in $R_4 \leq |z| \leq \frac{1}{2}$ does not exceed

$$1 + \frac{1}{\log 2} \log \frac{\alpha_n + \sum_{i=0}^{n} |\beta_i|}{|a_0|}$$

where R_4 is the same as defined in Lemma 3 .

Theorem 4, follows from Theorem C and Lemma 3. If $\beta_i = 0$ for $i = 0, 1, \ldots, n$ then it gives a refinement of Theorem A otherwise it is a refinement of Theorem C.

2. LEMMAS

LEMMA 1. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n such that $|\arg a_i - \beta| \leq \alpha \leq \pi/2$ for $i = 0, 1, \ldots, n$ and for some real β , then for some $t > 0$

$$|ta_i - a_{i-1}| \leq |t|a_i| - |a_{i-1}||\text{Cos } \alpha + (t|a_i| + |a_{i-1}|)\text{Sin } \alpha .$$

The proof of the above lemma is omitted as it follows immediately from the Lemma in [3] .

LEMMA 2. Let $p(z)$ be the same as defined in Theorem B. Then $p(z)$ has all its zeros in the ring shaped region given by

$$R_2 \leq |z| \leq R_1 .$$

Here

$$R_1 = \frac{c}{2}\left(\frac{1}{|a_n|} - \frac{1}{M_1}\right) + \left\{ \frac{c^2}{4}\left(\frac{1}{|a_n|} - \frac{1}{M_1}\right)^2 + \frac{M_1}{|a_n|} \right\}^{\frac{1}{2}}$$

and

$$R_2 = \frac{1}{2M_2^2}\left[-R_1^2|b|(M_2 - |a_0|) + \left\{4|a_0|R_1^2M_2^3 + R_1^4|b|^2(M_2 - |a_0|)^2\right\}^{\frac{1}{2}}\right]$$

where

$$M_1 = |a_n|(\text{Cos } \alpha + \text{Sin } \alpha) + 2 \text{ Sin } \alpha \sum_{k=0}^{n-1} |a_k| ,$$

$$M_2 = |a_n|R_1^n\left[\frac{2 \text{ Sin } \alpha}{|a_n|}\sum_{k=0}^{n-1} |a_k| + R_1\left(1 - \frac{|a_n|}{|a_n|}\right)(\text{Cos } \alpha + \text{Sin } \alpha)\right] ,$$

$$c = |a_n - a_{n-1}| ,$$

$$b = a_1 - a_0$$

LEMMA 3. Let $p(z)$ be defined as in Theorem C . Then $p(z)$ has all its zeros in the ring shaped region given by

$$R_4 \leq |z| \leq R_3 .$$

Here

$$R_3 = \frac{c}{2}\left(\frac{1}{\alpha_n} - \frac{1}{M_3}\right) + \left\{ \frac{c^2}{4}\left(\frac{1}{\alpha_n} - \frac{1}{M_3}\right)^2 + \frac{M_3}{\alpha_n} \right\}^{\frac{1}{2}}$$

and

$$R_4 = \frac{1}{2M_4^2}\left[-R_3^2|b|(M_4 - |a_0|) + \left\{ 4|a_0|R_3^2 M_4^3 + R_3^4 |b|^2 (M_4 - |a_0|)^2 \right\}^{\frac{1}{2}} \right] ,$$

where

$$M_3 = \alpha_n R ,$$

$$R = 1 + \frac{1}{\alpha_n}\left[2 \sum_{k=0}^{n-1} |\beta_k| + |\beta_n| \right] ,$$

$$M_4 = R_3^n \left[(\alpha_n + |\beta_n|)R_3 + \alpha_n R - (\alpha_0 + |\beta_0|) \right] ,$$

$$c = |a_n - a_{n-1}| ,$$

$$b = a_1 - a_0 .$$

Lemmas 2 and 3 are due to Govil and Jain [2] .

3. PROOFS OF THE THEOREMS

Proof of Theorem 1. Consider

$$F(z) = (t - z)p(z)$$

$$= (t - z)(a_0 + a_1 z + \ldots + a_n z^n)$$

$$= -a_n z^{n+1} + ta_0 + \sum_{i=1}^{n} (ta_i - a_{i-1}) z^i$$

For $|z| \leq t \leq 1$,

$$|F(z)| \leq t^{n+1} |a_n| + t|a_0| + \sum_{i=1}^{n} |t|a_i| - |a_{i-1}||t^i \text{ Cos } \alpha$$

$$+ \sum_{i=1}^{n} (t|a_i| + |a_{i-1}|)\, t^i \text{ Sin } \alpha \text{ , (by Lemma 1)}$$

$$\leq t^{n+1} |a_n| + t|a_0| + \sum_{i=1}^{k} (t|a_i| - |a_{i-1}|)t^i \text{ Cos } \alpha$$

$$+ \sum_{i=k+1}^{n} (|a_{i-1}| - t|a_i|)t^i \text{ Cos } \alpha$$

$$+ \sum_{i=1}^{n} (t|a_i| + |a_{i-1}|)t^i \text{ Sin } \alpha$$

$$= 2t^{k+1} |a_k| \text{ Cos } \alpha + 2t \text{ Sin } \alpha \sum_{i=0}^{n} t^i |a_i|$$

$$- t|a_0|(\text{Cos } \alpha + \text{Sin } \alpha - 1) - t^{n+1} |a_n|(\text{Cos } \alpha + \text{Sin } \alpha - 1)$$

$$\leq 2t^{k+1}|a_k| \text{ Cos } \alpha + 2t \text{ Sin } \alpha \sum_{i=0}^{n} t^i |a_i|$$

$$- t^{n+1} |a_n|(\text{Cos } \alpha + \text{Sin } \alpha - 1).$$

Now it is known (see [5], p. 171) that if $P(z)$ is regular, $P(0) \neq 0$ and $|F(z)| \leq M$ in $|z| \leq 1$ then the number of zeros of $P(z)$ in $|z| \leq \frac{1}{2}$ does not exceed $\frac{1}{\log 2} \{\log \frac{M}{|P(0)|}\}$. Applying this result to $F(z)$, we get that the number of zeros of $F(z)$ in $|z| \leq \frac{1}{2}$ does not exceed

$$\frac{1}{\log 2}\left\{\log \frac{2t^{k+1}|a_k|\cos\alpha + 2t\sin\alpha \sum_{i=0}^{n} t^i|a_i| - (\cos\alpha \quad \sin\alpha - 1)|a_n|t^{n+1}}{t|a_0|}\right\}.$$

As the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed the number of zeros of $F(z)$ in $|z| \leq \frac{1}{2}$, the theorem follows.

Proof of Theorem 2. Consider

$$F(z) = (t - z)\,p(z)$$

$$= -a_n z^{n+1} + ta_0 + \sum_{i=1}^{n}(ta_i - a_{i-1})z^i$$

For $|z| \leq t \leq 1$,

$$|F(z)| \leq t^{n+1}|a_n| + t|a_0| + \sum_{i=1}^{n}|ta_i - a_{i-1}|t^i$$

$$\leq t^{n+1}|a_n| + t|a_0| + \sum_{i=1}^{n}|t\alpha_i - \alpha_{i-1}|t^i$$

$$+ \sum_{i=1}^{n}(|\beta_{i-1}| + t|\beta_i|)t^i$$

$$\leq t^{n+1}|a_n| + t|a_0| + \sum_{i=1}^{k}(t\alpha_i - \alpha_{i-1})t^i$$

$$+ \sum_{i=k+1}^{n}(\alpha_{i-1} - t\alpha_i)t^i + \sum_{i=1}^{n}(|\beta_{i-1}| + t|\beta_i|)t^i$$

$$\leq t^{n+1}|a_n| + t|a_0| + 2t^{k+1}\alpha_k - t\alpha_0 - t^{n+1}\alpha_n + t|\beta_0|$$

$$- t^{n+1}|\beta_n| + 2t\sum_{i=1}^{n} t^i|\beta_i|$$

$$\leq 2\left(t^{k+1}\alpha_k + t\sum_{i=0}^{n} t^i|\beta_i|\right)$$

and following on the lines of the proof of Theorem 1, the proof of Theorem 2 can be completed.

REFERENCES

1. K.K. Dewan: Extremal properties and coefficient estimates for polynomials with restricted zeros and on location of zeros of polynomials. Ph.D. Thesis, I.I.T., Delhi, New Delhi, 1980.

2. N.K. Govil and V.K. Jain: On the Eneström-Kakeya Theorem II. Jour. of Approx. Theory 22 (1978), 1-10.

3. N.K. Govil and Q.I. Rahman: On the Eneström-Kakeya Theorem. Tôhoku Math. J. 20 (1968), 126-136.

4. Q.G. Mohammad: On the zeros of the polynomials. Amer. Math. Monthly, 72 (1965), 631-633.

5. E.C. Titchmarsh: The theory of functions, 2nd ed., Oxford University Press, London, 1939.

A POSTERIORI ERROR BOUNDS FOR EIGENSYSTEMS OF MATRICES

Z. BOHTE

ABSTRACT: In this paper an a posteriori error bound for approximate eigenvectors corresponding to simple eigenvalues of non-defective matrices is obtained. Under some additional assumptions the computable bound for the condition number is derived. Some illustrative numerical examples are given.

1. INTRODUCTION

A posteriori error bounds for computed eigenvalues of non-defective matrices and for computed eigenvectors of normal matrices are well-known (see [4]).

Let us summarize some of these known results.

Throughout this paper let A be a non-defective square complex matrix of order n and denote its eigenpairs by (λ_i, x_i), so that

$$Ax_i = \lambda_i x_i , \quad i = 1, \ldots , n \tag{1}$$

Denote by X the matrix of eigenvectors

$$X = [x_1, \ldots , x_n]$$

which is by assumption non-singular.

Let (λ, x) be an approximate eigenpair, usually computed by some numerical method, and let

$$r = Ax - \lambda x \tag{2}$$

be the corresponding residual vector. Then there exists an eigenvalue of the matrix A such that

$$\min_{1 \leq i \leq n} |\lambda_i - \lambda| \leq k(A) \|r\| / \|x\| \tag{3}$$

where

$$k(A) = \|X\| \, \|X^{-1}\| \tag{4}$$

The bound (3) holds for any of the norms 1, 2 or ∞. The number $k(A)$ is called the condition number of the matrix A with respect to the eigenvalue problem. For normal matrices $k_2(A) = 1$ and (3) gives the most satisfactory and easily computable a posteriori bound.

For normal matrices Wilkinson [4] gives the corresponding a posteriori error bound for the approximate eigenvector x. Let λ be an approximation to λ_1, let x_1 and x be normalized so that

$$\|x_1\|_2 = \|x\|_2 = 1 \tag{5}$$

and suppose that x_1 is multiplied by such a complex factor of modulus 1 that in

$$x = a_1 x_1 + \ldots + a_n x_n \tag{6}$$

the coefficient a_1 is non-negative:

$$a_1 \geq 0 \tag{7}$$

Further, let

$$d = \min_{2 \leq i \leq n} |\lambda_i - \lambda| \neq 0 \tag{8}$$

then

$$\|x - x_1\|_2 \leq (c/d)(1 + (c/d)^2)^{1/2} \tag{9}$$

where

$$c = \|r\|_2 \tag{10}$$

To use (9) in practice we need some information about other eigenvalues so that we can estimate the distance d from bel Unless c is significantly less then d, (9) provides no useful boun

Let us now consider a general non-defective matrix.

2. ERROR BOUND FOR APPROXIMATE EIGENVECTOR

Under the same conditions as above we shall prove that for the general non-defective matrix the bound for the error in the approximate eigenvector x is

$$\|x - x_1\|_2 \leq 2k_2(A)c/d \tag{11}$$

From (1), (2) and (6) it follows

$$r = (\lambda_1 - \lambda)a_1x_1 + \ldots + (\lambda_n - \lambda)a_nx_n$$

If we define

$$D = \mathrm{diag}(0, (\lambda_2 - \lambda)^{-1}, \ldots, (\lambda_n - \lambda)^{-1})$$

we have

$$XDX^{-1}r = a_2x_2 + \ldots + a_nx_n = u \tag{12}$$

and

$$\|u\|_2 \leq \|r\|_2 \|D\|_2 k_2(A) \tag{13}$$

Clearly,

$$\|D\|_2 = 1/d$$

where d is defined by (8). Using the notation (4) and (10) we can write the bound (13) in the form

$$\|u\|_2 \leq k_2(A)c/d \tag{14}$$

Further, under the assumptions (5) - (7) we have

$$\|a_1x_1\|_2 = a_1 \tag{15}$$

Since

$$x - x_1 = (a_1 - 1)x_1 + u$$

where u is defined by (12), we have

$$\|x - x_1\|_2 \leq |a_1 - 1| + \|u\|_2 \tag{16}$$

On the other hand

$$a_1x_1 = x - u$$

and using (15) and (5) we have

$$1 - \|u\|_2 \leq a_1 \leq 1 + \|u\|_2$$

From this two-sided inequality it follows

$$|a_1 - 1| \leq \|u\|_2 \tag{17}$$

The bound (11) follows directly from (16), (17) and (14).

For normal matrices the bound (11) is slightly weaker then the bound (9) where the orthogonality of eigenvectors has been taken into account.

In order to be able to use the bound (11) in practice we need also approximations to all other eigenvectors. The practical difficulty is that we must calculate an upper bound for $k_2(A)$ and a lower bound for d from an approximate eigensystem.

3. THE COMPUTABLE UPPER BOUND FOR THE CONDITION NUMBER

Let us denote by k the spectral condition number

$$k = k_2(A) = \|X\|_2 \|X^{-1}\|_2$$

In order to be able to compute a reliable upper bound for k we shall make a number of additional assumptions.

First, suppose that all the eigenvalues λ_i are simple and that we have calculated an approximate eigensystem (μ_i, y_i), $i = 1, \ldots, n$. Let all eigenvectors x_i and their approximations y_i be normalized

$$\|x_i\|_2 = \|y_i\|_2 = 1 , \quad i = 1, \ldots, n$$

and similarly to (6) and (7) we suppose that x_i are such that in

$$y_i = a_1^{(i)} x_1 + \ldots + a_i^{(i)} x_i + \ldots + a_n^{(i)} x_n$$

all

$$a_i^{(i)} \geq 0$$

Denote the matrix of approximate eigenvectors by

$$Y = [y_1, \ldots, y_n]$$

Then, clearly an approximation to k is the number

$$a = \|Y\|_E \|Z\|_E = \sqrt{n}\, \|Z\|_E \tag{18}$$

but it may not be an upper bound for it. This may happen because Y is only an approximation to X and it may be ill-conditioned and Z may be a poor approximation to Y^{-1}.

We shall have to calculate all the residual vectors

$$r_i = Ay_i - \mu_i y_i , \quad i = 1, \ldots , n$$

Denote

$$r = \max_{1 \leq i \leq n} \|r_i\|_2$$

and

$$m = \min_{i \neq j} |\mu_i - \mu_j|$$

Now, let us make the main assumption, that all the circles

$$C_i = (\mu_i, rk) , \quad i = 1, \ldots , n$$

with the centres μ_i and radii rk in the complex plane are disjoint. This means that in every one of them lies exactly one eigenvalue of the matrix A. We call λ_i the eigenvalue of A lying in C_i and from (3) we have the bounds

$$|\lambda_i - \mu_i| \leq rk , \quad i = 1, \ldots , n \tag{19}$$

To obtain the bounds for the errors in y_i we need a lower bound for

$$d_i = \min_{j \neq i} |\mu_i - \lambda_j| , \quad i = 1, \ldots , n$$

and clearly it follows from (19) that

$$d_i \geq m - rk = e , \quad i = 1, \ldots , n \tag{20}$$

From the assumption that all C_i are disjoint it is obvious that

$$e > 0$$

Therefore it follows from (11)

$$\|x_i - y_i\|_2 \leq 2 \|r_i\|_2 k/e , \quad i = 1, \ldots , n$$

These inequalities may be written in the form

$$\|X - Y\|_E \leq 2 \|R\|_E k/e \tag{21}$$

where R is the residual matrix

$$R = [r_1, \ldots , r_n]$$

Because

(22) $$k \leq \|X\|_E \|X^{-1}\|_E$$

and

(23) $$\|X\|_E = \sqrt{n}$$

we need only a bound for $\|X^{-1}\|_E$.

Let us denote

$$F = YZ - I$$

Then,

(24) $$E = XZ - I = F + (X - Y)Z$$

and using (21) we have the bound

(25) $$\|E\|_E \leq \|F\|_E + 2\|R\|_E \|Z\|_E k/e = g$$

Suppose that

(26) $$g < 1$$

This means that the matrix A should not be too ill-conditioned with respect to other terms in the right-hand side of (25). From (24) - (26) it follows directly

(27) $$\|X^{-1}\|_E \leq \|Z\|_E/(1 - g)$$

and we have the final inequality from (22), (23), (27), and (20)

(28) $$k \leq \sqrt{n}\|Z\|_E/(1 - \|F\|_E - 2\|R\|_E \|Z\|_E k/(m - rk))$$

Under the assumptions (20) and (26) both denominators on the right-hand side of (28) are positive.

Denoting

$$b = 1 - \|F\|_E , \quad c = 2\|R\|_E \|Z\|_E$$

and recalling (18) we can write (28) in the form

$$k \leq a/(b - ck/(m - rk)) = a(m - rk)/(bm - (c + br)k)$$

leading to the quadratic inequality

(29) $$(c + br)k^2 - (ar + bm)k + am \geq 0$$

For the exact eigensystem $r = c = 0$ and we obtain from (29) an

obvious bound

$$k \leq a/b$$

where the only errors are made in the computation of the inverse of the matrix of eigenvectors.

From (29) we obtain the bound

$$k \leq (p - (p^2 - 4amq)^{1/2})/(2q) = K \tag{30}$$

where

$$p = ar + bm, \quad q = c + br$$

This bound can be computed directly from approximate eigensystems. It can be shown that for sufficiently small r the number K is greater then 1 and gives therefore a useful bound for the condition number $k_2(A)$. It may happen, of course, that the number K is complex and then we have no bound for k.

The bound (30) can be used in the bounds (3) and (9) for individual eigenpairs. For the errors in eigenvalues we have from (3)

$$|\lambda_i - \mu_i| \leq K \|r_i\|_2 \tag{31}$$

and for the errors in eigenvectors we have from (9)

$$\|x_i - y_i\|_2 \leq 2K \|r_i\|_2 / g_i \tag{32}$$

where

$$g_i = \min_{j \neq i} (|\mu_i - \mu_j| - K \|r_j\|_2)$$

We must remember that these bounds hold provided all the above assumptions are fulfilled.

4. NUMERICAL EXAMPLES

Let us illustrate the obtained bounds by some simple examples of matrices of order $n = 3$.

(i) The matrix

$$A = \begin{bmatrix} 3 & 5 & -16 \\ 6 & 12 & -36 \\ 2 & 5 & -15 \end{bmatrix}$$

has eigenvalues

$$\lambda_1 = 1, \quad \lambda_2 = 2, \quad \lambda_3 = -3$$

and the spectral condition number

$$k_2(A) = 32{\cdot}14\ldots$$

If we take

$$\mu_1 = 1 + 10^{-5}, \quad \mu_2 = 2 + 2.10^{-5}, \quad \mu_3 = 3 - 3.10^{-5}$$

and for eigenvectors the correct eigenvectors rounded to 5 decimal places and also the correct inverse to 5 places, we obtain the bound

$$K = 52{\cdot}44\ldots$$

By the way, the number

$$a = 33{\cdot}45\ldots$$

is a very good approximation to $k_2(A)$. The bounds (31) and (32) are severe overestimates in this case. For instance,

$$|\lambda_1 - \mu_1| = 10^{-5}$$

but

$$K \|r_1\|_2 = 910.10^{-5}$$

and

$$\|x_1 - y_1\|_2 = 0{\cdot}52.10^{-5}$$

but

$$2K \|r_1\|_2 / g_1 = 1836.10^{-5}$$

(ii) The matrix

$$A = \begin{bmatrix} 19 & 7 & -2 \\ 10 & 16 & -2 \\ 4 & -8 & 19 \end{bmatrix}$$

has eigenvalues

$$\lambda_1 = 9, \quad \lambda_2 = 18, \quad \lambda_3 = 27$$

and

$$k_2(A) = 1 + \sqrt{2} = 2{\cdot}41\ldots$$

If we take

$$\mu_1 = 9 + 3.10^{-5}, \quad \mu_2 = 18 + 5.10^{-5}, \quad \mu_3 = 27 + 6.10^{-5}$$

and similarly round the eigenvectors and the inverse matrix, we obtain the bounds for the condition number

$$K = 3{\cdot}87\ldots$$

for the error in the third eigenvalue

$$K \|r_3\|_2 = 24.10^{-5}$$

and for the error in the third eigenvector

$$2K \|r_3\|_2 / g_3 = 5.10^{-5}$$

which is very satisfactory. The approximate condition number a is almost the same 3·87.

(iii) The upper triangular matrix

$$A = \begin{bmatrix} 1 & 100 & 0 \\ & 2 & 0 \\ & & 100 \end{bmatrix}$$

with eigenvalues

$$\lambda_1 = 1, \quad \lambda_2 = 2, \quad \lambda_3 = 100$$

is very ill-conditioned, namely

$$k_2(A) = 200{\cdot}05\ldots$$

With

$$\mu_1 = 1 + 10^{-5}, \quad \mu_2 = 2 + 2.10^{-5}, \quad \mu_3 = 100 + 10^{-3}$$

and rounded eigenvectors we get a complex number K and cannot use any of the bounds for the errors. The approximate condition number a is equal to 245.

5. CONCLUSIONS

It is rarely justified to use expensive a posteriori bounds which are usually too pessimistic. But compared to a posteriori bounds for the solution of the system of linear algebraic equations where the bound is approximately 6 times more expensive as the solution by the

Gauss elimination, here, even with the most economic methods (e.g. the QR method), the additional number of arithmetic operations $6n^3$ is not worrying. Of course, for practical use, some sort of iterative improvement of the approximate eigensystem is more desirable (see [3]).

Recently Chu [2] generalized the Bauer-Fike theorem [1] to defective matrices. Along these lines it would be worthwhile to attempt finding a posteriori bounds for the computed eigenvalues and eigenvectors using the Schur form.

Acknowledgement. I wish to express my sincere thanks to I. Vidav who proved the bound (11).

REFERENCES:

1. F. L. Bauer and C. T. Fike: Norms and exclusion theorems, Numer. Math. 2 (1960), 42 - 53.

2. K. E. Chu: Generalization of the Bauer-Fike theorem, Numer. Math. 9 (1986), 685 - 691.

3. H. J. Symm and J. H. Wilkinson: Realistic error bounds for a simple eigenvalue and its associated eigenvector, Numer. Math. 35 (1980), 113 - 126.

4. J. H. Wilkinson: The algebraic eigenvalue problem, Clarendon Press, Oxford 1965.

ON THE UNIFORM CONVERGENCE OF MODIFIED GAUSSIAN RULES FOR THE NUMERICAL EVALUATION OF DERIVATIVES OF PRINCIPAL VALUE INTEGRALS

G. CRISCUOLO and G. MASTROIANNI

ABSTRACT: *The authors prove some convergence theorems of a modified gaussian rule for the evaluation of the derivatives of Cauchy principal value integrals.*

1. INTRODUCTION

Let $\Phi(wf;t)$ denote the integral in the Cauchy principal value sense of the function f, associated with the weight w and defined by

$$(1)\quad \Phi(wf;t)=\int_{-1}^{1} \frac{f(x)}{x-t}\, w(x)dx=\lim_{\varepsilon\to 0^+}\int_{|x-t|\geq\varepsilon} \frac{f(x)}{x-t}\, w(x)dx\ ,\ -1<t<1.$$

In order to approximate the integral (1) we may consider the gaussian rule

$$\Phi_m(wf;t)=f(x)\int_{-1}^{1} \frac{w(x)}{x-t}\,dx+\sum_{i=1}^{m}\lambda_{m,i}\frac{f(x_{m,i})-f(t)}{x_{m,i}-t}\ ,\ t\neq x_{m,i},\ i=1,2,\ldots,m\ ,$$

where $x_{m,i}$, i=1,2,...m, are the zeros of the m-th orthogonal polynomial associated with the function w and $\lambda_{m,i}$, i=1,2,...m, are the Christoffel constants.

If the function f is "sufficiently smooth", then the sequence $\{\Phi_m(wf;t)\}$converges to $\Phi(wf;t)$. Furthermore, it is easy to prove that the inequality

$$|\Phi(wf;t)-\Phi_m(wf;t)| \leq \text{const } m^{-k}\ \omega(f^{(k)};m^{-1})\log m\ ,\quad f\in C^k(I),\ k\geq 1,$$

hold on every closed interval $\Delta\subset(-1,1)$.

Unfortunately, in the general rule, if f is an Hölder continuous function, then $\{\Phi_m(wf;t)\}$ does not converge to $\Phi(wf;t)$ almost everywhere in (-1,1), (see[4]).

In order to avoid this problem, the authors introduced in [1] a new formula $\Phi_m^*(wf;t)$; this is defined by

$$(2)\qquad \Phi_m^*(wf;t)=f(t)\int_{-1}^{1}\frac{w(x)}{x-t}\,dx+\sum_{\substack{i=1\\ i\neq c}}^{m}\lambda_{m,i}\,\frac{f(x_{m,i})-f(t)}{x_{m,i}-t}\ ,\quad m\in N\ ,$$

where c denotes the index corresponding to the "closest knot" $x_{c(m)}=x_{m,c}$ to the singularity t, defined by $|t-x_{m,c}|=\min\{t-x_{m,d},x_{m,d+1}-t\}$, $x_{m,d}\leq t\leq x_{m,d+1}$ for some $d\in\{0,1,\ldots,m\}$ with $x_{m,o}=-1$, $x_{m,m+1}=1$.

The "modified gaussian rule" $\Phi_m^*(wf;t)$ has degree of exactness 0; nevertheless the hypothesis $x_{m,i}\neq t$, $i=1,2,\ldots,m$ becomes unnecessary.

Notice that the derivative $\frac{d}{dt}\,\Phi(wf;t)$ appears in some integrodifferential equations concerning several branches of physics and engineering.

Further, the analytic solution of the integral equations with logarithmic singularities in the kernel may be represented by the derivatives of Cauchy principal value integrals.

In this paper we study the uniform convergence of the sequence

$$\left\{\frac{d^p}{dt^p}\,\Phi_m^*(wf;t)\right\}\ \text{to}\ \frac{d^p}{dt^p}\,\Phi(wf;t)\ \text{on}\ (-1,1)\ \text{for}\ p\geq 0.$$

This is of interest in solving singular integral equations with a collocation method too. Indeed, uniform convergence results of a quadrature rule on the whole interval (-1,1) are necessary to study the convergence of the method when, for example, the collocation points are zeros of orthogonal polynomials in [-1,1].

The convergence theorems are stated in the Section 2; they generalize

previous results [2] and are proved in the Section 3.

2. CONVERGENCE THEOREMS AND ESTIMATES OF THE REMAINDER

We start with some notation. Throughout this paper DT denotes the space of the continuous functions in I:=[-1,1] satisfying a "Dini type" condition, and $\mathrm{Lip}_M\,\lambda$ the space of the Hölder continuous functions; i.e.:

$$DT:=\{f\in C(I)/\ \int_0^1 \delta^{-1}\,\omega(f;\delta)d\delta < \infty\}$$

$$\mathrm{Lip}_M\lambda:=\{f\in C(I)/\ \omega(f;\delta) \leq M\delta^{\lambda}\ ,\ M>0,\ 0<\lambda\leq 1\}$$

where $\omega(f;\delta)=\max_{|x-y|<\delta} |f(x)-f(y)|$, $x,y\in I$, $\delta\geq 0$, is the modulus of continuity of the function f. We ought to remark that $DT\supset \mathrm{Lip}_M\lambda$.

In the computation of the integral $\Phi(wf;t)$ defined by (1) we suppose that the weight function w can be written in the form $w(x)=\psi(x)\,u^{\alpha,\beta}(x)$, $x\in I$, with $u^{\alpha,\beta}(x)=(1-x)^{\alpha}(1+x)^{\beta}$, $\alpha,\beta>-1$ and $0<\psi\in DT$.

Let $\{P_m(w)\}$ be the sequence of the orthonormal polynomials on I associated with the weight function w; we denote the zeros of

$$P_m(x)=P_m(w;x)=\alpha_m x^m + \text{lower degree terms}\ ,\quad \alpha_m>0,$$

by $x_{m,i}=x_{m,i}(w)=\cos\,\Theta_{m,i}$, $i=1,2,\ldots,m$, so that

$$0=\Theta_{m,m+1}<\Theta_{m,m}<\ldots<\Theta_{m,2}<\Theta_{m,1}<\Theta_{m,o}=\pi .$$

Furthermore, the numbers $\lambda_{m,i}=\lambda_{m,i}(w)$, $i=1,2,\ldots m$, are the Christoffel constants defined by $\lambda_{m,i}(w)=\lambda_m(w;x_{m,i})$ where $\lambda_m(w;x)=\left[\sum_{k=o}^{m-1} P_k^2(w;x)\right]^{-1}$ is the m-th Christoffel function.

Denoting by $E_m^*(wf)=\Phi(wf)-\Phi_m^*(wf)$ the remainder term of the formula $\Phi_m^*(wf)$ defined by (2), we can state the following

THEOREM 1.

If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $\alpha,\beta\geq 0$, then for any function f such that $f^{(p)}\in DT$ the sequence $\left\{\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)\right\}$ converges uniformly to $\frac{d^p}{dt^p}\ \Phi(wf;t)$ on the whole open interval $(-1,1)$ $p\geq 0$.

Moreover, if $f^{(p)}\in Lip_M\lambda$, $0<\lambda\leq 1$, it is also

$$\left|\frac{d^p}{dt^p}\ E_m^*(wf;t)\right| \leq \text{const } m^{-\lambda}\ \log m, \quad -1<t<1,\ p\geq 0$$

THEOREM 2.

If $w=\psi u^{\alpha\ \beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-1<\alpha,\beta<0$, then for any function f such that $f^{(p)}\in Lip_M\lambda$, $0<\lambda\leq 1$, it results

$$(3)\qquad \left|\frac{d^p}{dt^p}\ E_m^*(wf;t)\right| \leq \text{const } m^{-\lambda} u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\ \log\frac{m}{\sqrt{1-t^2}},\quad -1<t<1,\ p\geq 0$$

In particular, if $\alpha+\lambda/2,\beta+\lambda/2\geq 0$ then by Theorem 2 it follows

Corollary 3. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-\frac{1}{2}<\alpha,\beta<0$, then for any function f such that $f^{(p)}\in Lip_M\lambda$, $-2\ \min(\alpha,\beta)<\lambda\leq 1$, the sequence $\left\{\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)\right\}$ converges uniformly to $\frac{d^p}{dt^p}\ \Phi(wf;t)$ on the whole open interval $(-1,1)$, $p\geq 0$.

Moreover, taking into account (3), it seems that the sequence $\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)$ can not converge uniformly on $(-1,1)$ for $\alpha,\beta\leq -\frac{1}{2}$, generally. Nevertheless, a favourable case of interest in the applications comes true when $t\in\Delta_m:=[-1+\text{const } m^{-2},\ 1-\text{const } m^{-2}]$. In fact, by Theorem 2 we deduce also

Corollary 4. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-1<\alpha,\beta<0$, then for any function f such that $f^{(p)}\in Lip_M\lambda$, $-2\min(\alpha,\beta)\leq\lambda\leq 1$, the sequence $\left\{\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)\right\}$ converges uniformly to $\frac{d^p}{dt^p}\ \Phi(wf;t)$ on Δ_m, $p\geq 0$.

Moreover, it results

$$\left|\frac{d^p}{dt^p}\ E_m^*(wf;t)\right| \leq \text{const } m^{-\lambda}\ u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\ \log m,\quad t\in\Delta_m,\ p\geq 0.$$

orollary 5. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-1<\alpha,\beta<0$, then for any function f uch that $f^{(p)}\in Lip_M\lambda$, $-\min(\alpha,\beta)=\gamma<\lambda<-2\min(\alpha,\beta)$, the sequence $\left\{\frac{d^p}{dt^p}\phi_m^*(wf;t)\right\}$ onverges uniformly to $\frac{d^p}{dt^p}\Phi(wf;t)$ on Δ_m, $p\geq 0$.

loreover, it results

$$\left|\frac{d^p}{dt^p}E_m^*(wf;t)\right| \leq \text{const}\ \frac{\log m}{m^{2(\lambda-\gamma)}}\ , \quad t\in\Delta_m, \quad p\geq 0.$$

3. PROOF SKETCH OF THE MAIN RESULTS

For the convenience of the reader, we collect some properties of the orthonormal polynomials $P_m(w)$ with $w(x)=\psi(x)u^{\alpha,\beta}(x)$, $-1\leq x\leq 1$, $u^{\alpha,\beta}(x)=(1-x)^\alpha(1+x)^\beta$, $\alpha,\beta>-1$, $0<\psi\in DT$, which will be used in the following.

The equivalence

(4) $\quad \Theta_{m,k} - \Theta_{m,k+1} \sim m^{-1}$, uniformly for $0\leq k\leq m$, $m\in N$,

(5) $\quad \lambda_{m,k} \sim m^{-1}\ u^{\alpha+1/2,\beta+1/2}(x_{m,k})$, uniformly for $1\leq k\leq m$, $m\in N$

holds for the zeros of $P_m(w)$ and for the Christoffel constants respectively. One can find the relations (4) and (5) in [3].

Furthermore, it follows from (4) that

(6) $\quad u^{\alpha,\beta}(t) \sim u^{\alpha,\beta}(x_{m,k})\ , \qquad x_{m,k} \leq t \leq x_{m,k+1}$,

for $k=2,3,\ldots,m-1$ (see[3,p.48]).

To derive the proofs of the theorems stated in the previous section, the following lemmas are needed.

Lemma 1. If $f\in C^r(I)$, $r\geq 0$, then for each $m\in N$ there exists a polynomial t_m of degree at most $m\geq 4(r+1)$ such that

$$\left|f^{(k)}(x)-t_m^{(k)}(x)\right|\leq \text{const}\left[m^{-1}\sqrt{1-x^2}\right]^{r-k}\omega(f^{(r)};m^{-1}\sqrt{1-x^2}),0\leq k\leq r,\ -1\leq x\leq 1,$$

$$\left|t_m^{(p)}(x)\right|\leq \text{const}\,[\Delta_m(x)]^{r-p}\omega(f^{(r)};\ \Delta_m(x)),\ p>r,\quad -1\leq x\leq 1,$$

where $\Delta_m(x)=m^{-1}\sqrt{1-x^2}+m^{-2}$.

Lemma 2. If $w=\psi u^{\alpha,\beta}$, $0<\psi\in DT$, $\alpha,\beta>-1$ then for any function $f\in C(I)$ the inequality

$$\lambda_{m,c}\left|\frac{t_m(x_{m,c})-t_m(t)}{x_{m,c}-t}\right|\leq \text{const}\left[\sqrt{1-t}+m^{-1}\right]^{2\alpha}\left[\sqrt{1+t}+m^{-1}\right]^{2\beta}\omega(f;\Delta_m(t)),$$

holds uniformly for $t\in(-1,1)$, where t_m is the polynomial of Lemma 1, $x_{m,c}$ is the closest knot to the point t, and $\Delta_m(t)=m^{-1}\sqrt{1-t^2}+m^{-2}$.

Setting $\sigma_m^*(t)=\sum\limits_{\substack{i=1\\ i\neq c}}^{m}\frac{\lambda_{m,i}}{|x_{m,i}-t|}$, we can state

Lemma 3. If $w=\psi u^{\alpha,\beta}$, $0<\psi\in DT$, then the inequalities

$$\sigma_m^*(t)\leq \text{const}\ \log m\ ,\quad \text{if } \alpha,\beta\geq 0,$$

$$\sigma_m^*(t)\leq \text{const}\ u^{\alpha,\beta}(t)\ \log m\ ,\quad \text{if } -1<\alpha,\beta<0,$$

hold uniformly for $t\in(-1,1)$.

Lemma 4. If $w=\psi u^{\alpha,\beta}$, $0<\psi\in DT$, then the inequalities

$$\sum_{\substack{i=1\\ i\neq c}}^{m}\lambda_{m,i}\frac{|r_m(x_{m,i})-r_m(t)|}{|x_{m,i}-t|}\leq \text{const}\begin{cases}\omega(t;m^{-1})\log m, & \text{if } \alpha,\beta\geq 0,\quad f\in C(I)\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log m, & \\ \text{if } -1<\alpha,\beta<0, & f\in \text{Lip}_M\lambda,\ 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\varepsilon(-1,1)$, where $r_m=f-t_m$, being t_m the polynomial of Lemma 1.

Lemma 5. If $w=\psi u^{\alpha,\beta}$, $0<\psi\varepsilon DT$, then the inequalities

$$\int_{-1}^{1}\left|\frac{r_m(x)-r_m(t)}{x-t}\right|w(x)dx\leq const\begin{cases}\omega(f;m^{-1})\log m\ , & \text{if } \alpha,\beta\geq 0,\quad f\varepsilon C(I),\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log\frac{m}{\sqrt{1-t^2}} & \\ \text{if } -1<\alpha,\beta<0,\quad f\varepsilon Lip_M\lambda,\ 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\varepsilon(-1,1)$, where $r_m=f-t_m$ being t_m the polynomial of Lemma 1.

Lemma 1 can be found in [5]; instead, the other previous lemmas are proved in [2].

Lemma 6. Let $v\varepsilon L_1[-1,1]$, i.e. $\int_{-1}^{1}|v(x)|dx<\infty$, possibly having singularities at $v_1,\ldots,v_s$, and assume that v is continuous on each closed interval enclosed in $I-\{v_1,\ldots,v_1\}$. If g is a function such that $g^{(p)}\varepsilon DT$, $p\geq 1$, then the integral $\int_{-1}^{1}\frac{d^p}{dt^p}\left[\frac{g(x)-g(t)}{x-t}\right]v(x)dx$ exists and the identity

$$\frac{d^p}{dt^p}\int_{-1}^{1}\frac{g(x)-g(t)}{x-t}v(x)dx=\int_{-1}^{1}\frac{d^p}{dt^p}\left[\frac{g(x)-g(t)}{x-t}\right]v(x)dx,$$

holds whenever t is in a closed set enclosed in $I-\{v_1,\ldots,v_s\}$ and $p\geq 1$.

The proof of Lemma 6 is based on known results of classical analysis and elementary inequalities for the modulus of continuity.

Lemma 7. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\varepsilon DT$, $\alpha,\beta>-1$, then for any function $f\varepsilon C^p(I)$ the inequality

$$\lambda_{m,c}\left|\frac{d^p}{dt^p}\,\frac{t_m(x_{m,c})-t_m(t)}{x_{m,c}-t}\right|\leq\text{const}\left[\sqrt{1-t}+m^{-1}\right]^{2\alpha}\left[\sqrt{1+t}+m^{-1}\right]^{2\beta}\omega(f^{(p)};\,\Delta_m(t)),$$

holds uniformly for $t\in(-1,1)$, $p\geq 0$, where t_m is the polynomial of Lemma 1, $x_{m,c}$ is the closest knot to the point t, and $\Delta_m(t)=m^{-1}\sqrt{1-t^2}+m^{-2}$.

The proof of this lemma can be deduced by Lemmas 1,2,6 and applying the inequality (6).

Lemma 8. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, then the inequalities

$$\sum_{\substack{i=1\\ i\neq c}}^{m}\lambda_{m,i}\left|\frac{d^p}{dt^p}\,\frac{r_m(x_{m,i})-r_m(t)}{x_{m,i}-t}\right|\leq\text{const}\begin{cases}\omega(f^{(p)};m^{-1})\log m, & \text{if } \alpha,\beta\geq 0,\ f\in C^p(I)\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log m, & \text{if } -1<\alpha,\beta<0,\\ f^{(p)}\in \mathrm{Lip}_M\lambda, & 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\in(-1,1)$, $p\geq 0$, where $r_m=f-t_m$, being t_m the polynomial of Lemma 1.

Lemma 8 follows from Lemmas 1,3,4,6 and taking into account the relation (6).

Lemma 9. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, then the inequalities

$$\left|\frac{d^p}{dt^p}\int_{-1}^{1}\frac{r_m(x)-r_m(t)}{x-t}\,w(x)dx\right|\leq\text{const}\begin{cases}\omega(f^{(p)};m^{-1})\log m, & \text{if } \alpha,\beta\geq 0,\ f\in C^p(I)\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log\frac{m}{\sqrt{1-t^2}}, & \\ \text{if } -1<\alpha,\beta<0, & f^{(p)}\in \mathrm{Lip}_M\lambda,\quad 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\in(-1,1)$, $p\geq 0$, where $r_m=f-t_m$ being t_m, the polynomial of Lemma

Applying again the inequality (6), the proof of Lemma 9 follows from the results of Lemma 1,5,6.

Now, since rule (2) has degree of exactness 0, we have

$$\left|\frac{d^p}{dt^p} E_m^*(wf;t)\right| \leq \lambda_{m,c}\left|\frac{t_m(x_{m,c})-t_m(t)}{x_{m,c}-t}\right| + \sum_{\substack{i=1 \\ i\neq c}}^{m} \lambda_{m,i}\left|\frac{d^p}{dt^p}\,\frac{r_m(x_{m,i})-r_m(t)}{x_{m,i}-t}\right| +$$

$$+ \left|\frac{d^p}{dt^p}\int_{-1}^{1} \frac{r_m(x)-r_m(t)}{x-t}\, w(x)dx\right| ,$$

where $r_m = f - t_m$, being t_m the polynomial of Lemma 1. Thus, Theorem 1 and Theorem 2 follow from Lemmas 7,8 and 9.

REFERENCES

[1] G.CRISCUOLO and G.MASTROIANNI - On the convergence of the Gauss quadrature rules for the Cauchy Principal Value integrals.- Ricerche di Matematica XXXV (1986), 45-60.

[2] G.CRISCUOLO and G.MASTROIANNI - On the uniform convergence of gaussian quadrature rules for Cauchy Principal Value integrals.- To appear on Numeriscke Mathematik.

[3] P.NEVAI and P.VERTESI - Mean convergence of Hermite-Fejér Interpolation.- J.Math. Anal. and Appl. 105 (1985), 26-58.

[4] P.RABINOWITZ - On the convergence of Hunter's method for Cauchy Principal Value integrals. In: Numerical Solution of Singular Integral Equations (A.Gerasoulis, R.Vichmevetsky, eds.) IMACS, 1984.

[5] P.O.RUNCK - Bemerkungen zu den approximatiossätzen von Jackson und Jackson-Timan.- ISNM 10 (1969), 303-308.

APPROXIMATE EXPANSIONS OF DIFFERENTIABLE FUNCTIONS IN POLYNOMIAL SERIES

M.R. DA SILVA

ABSTRACT : One of the most important tools in applied analysis is the expansion of a given function $y = y(x)$ in a series of polynomials. If these are orthogonal, then there are explicit, well-known formulas for the expansion coefficients, but they involve quadratures which are generally difficult to perform. To avoid the evaluation of those integrals, we use a simple approximation principle which leads naturally to good polynomial approximants of y in the sense of the τ - method and is much more amenable to computer programming than Lanczos' original perturbation idea.

I. INTRODUCTION

It is well-known how important it is in applied analysis to be able to develop a given function in a series of algebraic polynomials. If these are orthogonal, then there are explicit, well-known formulas for the expansion coefficients, but they are often unsuitable for numerical evaluation, as they involve quadratures which are generally difficult to perform. For some important particular orthogonal polynomial systems we may approximate the corresponding expansion coefficients recursively, with and without numerical quadratures.

1.1. A RECURSIVE METHOD FOR THE EXPANSION COEFFICIENTS OF DIFFERENTIABLE FUNCTIONS IN SERIES OF JACOBI POLYNOMIALS IN $[0, 1]$.

Let $P_k^*(x) = P_k^{(\alpha,\beta)}(2x - 1)$, $k = 0, 1\ldots,$ $0 \le x \le 1$, $\alpha,\beta > -1$, be the standard shifted Jacobi orthogonal polynomials and assume, formally, that

$$y(x) = \sum_{k=0}^{\infty} a_k P_k^*(x) . \tag{1.1}$$

Multiplying both sides of (1.1) by $(1-x)^{\alpha} x^{\beta} P_n^*(x)$ and integrating from 0 to 1, we get

$$a_n = \frac{1}{\gamma_n} \int_0^1 (1-x)^{\alpha} x^{\beta} P_n^*(x)\, y(x)\, dx \qquad n = 0, 1, \ldots$$

$$\gamma_n = \int_0^1 (1-x)^{\alpha} x^{\beta} (P_n^*(x))^2\, dx \tag{1.2}$$

It is well-known that the fact that we can calculate a_n, $n = 0, 1, \ldots$, does not guarantee that the series in (1.1) converges, or, if the series converges, that its sum is $y(x)$.

Using Rodrigues' formula for $P_n^*(x)$,

$$P_n^*(x) = \frac{(-1)^n}{n!} \frac{D^n\{(1-x)^{\alpha+n} x^{\beta+n}\}}{(1-x)^{\alpha} x^{\beta}}, \quad D = \frac{d}{dx},$$

we obtain

$$a_n = \frac{(-1)^n}{\gamma_n n!} \int_0^1 D^n\{(1-x)^{\alpha+n} x^{\beta+n}\}\, y(x)\, dx \tag{1.3}$$

and after integrating (1.3) by parts n times,

$$a_n = \frac{(-1)^n}{\gamma_n n!} \int_0^1 (1-x)^{\alpha+n} x^{\beta+n} y^{(n)}(x)\, dx. \tag{1.4}$$

Instead of the integral transforms (1.2) - (1.4) we can solve a linear lower triangular system and obtain the coefficients a_n, $n = 0, 1, \ldots$, recursively.

Inserting Rodrigues' formula for $P_k^*(x)$ in (1.1) gives

$$(1-x)^{\alpha} x^{\beta} y(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} a_k D^k\{(1-x)^{\alpha+k} x^{\beta+k}\}. \tag{1.5}$$

Defining

$$I^{n+1} g(\xi) \equiv \int_0^{\xi} \cdots \int_0^{\xi} g(\xi)\, d\xi = \int_0^{\xi} \frac{(\xi - x)^n}{n!} g(x)\, dx,$$

then

$$I^{n+1}\{(1-\xi)^{\alpha} \xi^{\beta} y(\xi)\} = \int_0^{\xi} \frac{(\xi - x)^n}{n!} (1-x)^{\alpha} x^{\beta} y(x)\, dx. \tag{1.6}$$

Also, from (1.5),

$$I^{n+1}\{(1-\xi)^{\alpha}\,\xi^{\beta}\,y(\xi)\} = \sum_{k=0}^{n}\frac{(-1)^k}{k!}\,a_k\,I^{n-k+1}\{(1-\xi)^{\alpha+k}\,\xi^{\beta+k}\}$$

$$+\sum_{k=n+1}^{\infty}\frac{(-1)^k}{k!}\,a_k\,D^{k-n-1}\{(1-\xi^{\alpha+k}\,\xi^{\beta+k}\}$$

$$(1.7)\qquad = \sum_{k=0}^{n}\frac{(-1)^k}{k!}\,a_k\int_0^{\xi}\frac{(\xi-x)^{n-k}}{(n-k)!}\,(1-x)^{\alpha+k}.\,x^{\beta+k}\,dx+\ldots$$

Taking $\xi = 1$ and comparing (1.6) with (1.7) we get the following seemingly new formulas for the coefficients of $y(x)$ in series of Jacobi polynomials

$$(1.8)\qquad \sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{\Gamma(\alpha+n+1)\,\Gamma(\beta+k+1)}{\Gamma(\alpha+\beta+n+k+2)}\,a_k = \int_0^1(1-x)^{\alpha+n}\,x^{\beta}\,y(x)\,dx\ ,$$

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

In particular, for the shifted Legendre polynomials, $\alpha = \beta = 0$ and (1.8) leads to

$$\sum_{k=0}^{n}(-1)^k\frac{\binom{n}{k}}{\binom{n+k+1}{k}}\,a_k = (n+1)\int_0^1(1-x)^n\,y(x)\,dx\ ,\quad n = 0,\ 1,\ldots$$

1.2. POLYNOMIAL SERIES DEVELOPMENTS AND BASIC IDEA OF THE LANCZOS'S τ - METHOD

To avoid the evaluation of the integrals in (1.2), Lanczos [7,8,9] conceived the following perturbation technique, which has long been known as the τ - method.

Given an equation of the form

$$(1.9)\qquad Dy(x) = f(x)\ ,\quad a \le x \le b\ ,\quad |a|, |b| < \infty\ ,$$

where $f(x)$ is an Nth degree algebraic polynomial and D a νth order linear differential operator with polynomial coefficients,

$$(1.10)\qquad D \equiv \sum_{r=0}^{\nu} p_r(x)\,\frac{d^r}{dx^r}\ ,$$

together with ν supplementary (initial, boundary or mixed) conditions through linear combinations of function and derivative values of y, which we may write as

$$g_j(y) = \sigma_j \quad , \quad j = 1(1)\nu \quad , \tag{1.11}$$

where the g_j's are given linear functionals, the basic idea of the Lanczos' τ-method for the construction of a polynomial approximation y_n to the solution y of the problem in (1.9) - (1.11) in a form suitable for numerical evaluation is to perturb the given equation (1.9) through the addition to its r.h.s. of an algebraic polynomial H_n, usually chosen to be a linear combination of Chebyshev or Legendre polynomials with free coefficients, called the τ-parameters, which are to be determined so that y_n is the unique polynomial solution of the perturbed problem

$$Dy_n(x) = f(x) + H_n(x) \quad , \quad a \leq x \leq b \quad , \tag{1.12}$$

$$g_j(y_n) = \sigma_j \quad , \quad j = 1(1)\nu \quad , \tag{1.13}$$

to be called the τ-problem in the sequel.

The choice of H_n is essentially made on the basis of

i) The given supplementary conditions, so H_n is to contain ν τ-parameters to be determined to ensure satisfaction of conditions (1.13);

ii) Intrinsic properties of D, namely its range R_D, which is not, in general, the whole space $\mathbb{P}$ of algebraic polynomials, and its height h, $h = \sup_{n \in \mathbb{N}_0} \{\partial(Dx^n) - n\}$, where $\mathbb{N}_0$ is the set of nonnegative integers, and ∂ stands for "degree of", so H_n is to have degree $\leq n + h$ and to be in R_D. To be more precise, there generally exists for D an s-dimensional residual subspace R_S complementary to R_D,

$$\mathbb{P} = R_D \oplus R_S \quad , \quad R_D \cap R_S = \{0\} \quad , \tag{1.14}$$

$$R_S = \text{span}\,\{x^k : k \in S\} \quad , \quad S = \{k \in \mathbb{N}_0 : x^k \notin R_D\} \quad ,$$

and so H_n is also to contain s τ-parameters to be determined to ensure compatibility of the τ-problem, i.e., that no component of H_n lies in R_S [22];

iii) Approximation properties that y_n is required to possess. Clearly, the quality of y_n as an approximation of y depends on H_n, as follows from the fact that the τ-error function $\varepsilon_n = y - y_n$ is such that

$$D\varepsilon_n(x) = -H_n(x) \quad , \quad a \leq x \leq b \tag{1.15}$$

$$g_j(\varepsilon_n) = 0 \quad , \quad j = 1(1)\nu \quad ,$$

hence

$$\varepsilon_n(x) = -\int_a^b G(x, t)\, H_n(t)\, dt \quad , \tag{1.16}$$

where $G(x, t)$ is the corresponding Green's function, so H_n should be small, e.g., in the sense of the uniform norm, $\| H_n \| = \max_{a \leq x \leq b} | H_n(x) |$.

In the hope that the smallness of H_n will imply that of ε_n, one usually chooses

$$H_n(x) = \sum_{i=0}^{r} \tau_{m-i}^{(n)}\, v_{m-i}(x) \quad , \quad r = \nu + s - 1 \quad , \quad m = n + h \quad ,$$

where the $\tau_j^{(n)}$'s are the τ-parameters to be determined and the v_j's are Chebyshev or Legendre polynomials according as $y_n(x)$ is required to be a good (nearly uniform) global or endpoint approximation of $y(n)$, respectively on $a \leq x \leq b$ or at $x = b$ (see [10] and [16,17] for details and applications).

The existence, uniqueness, and convergence questions for the Lanczos' τ-approximation problem are reduced to the corresponding questions for the τ-parameters. These are uniquely determined by the supplementary and compatibility conditions referred to above and tend exponentially to zero as $n \to \infty$ [2,3].

As for the questions of existence, uniqueness, and characterization of perturbations leading to τ-approximants endowed with prescribed properties, they are still open, as far as we are aware, and we conclude that, in general, the choice of a suitable perturbation is not a simple matter.

1.3. AN ALTERNATIVE PRINCIPLE FOR τ-METHOD APPROXIMATION

As an alternative to the Lanczos' original perturbation idea, the following approximation principle [23,24] has evolved.

Choose a basis $v = \{v_k\}_{k=0}^{n}$ for $\mathbb{P}_n = \{P \in \mathbb{P} : \partial(P) \leq n\}$,

preferably orthogonal for rapid convergence, express y_n in it,

$$y_n = \sum_{k=0}^{n} \alpha_k v_k \quad ,$$

and determine the α_k's by making y_n satisfy the supplementary conditions in (1.13) and Dy_n agree with Dy as far as possible or desired.

This alternative principle, which emerges from [6] and [18], is shown in [23] to lead naturally to an approximation of the solution y of the problem in (1.9) - (1.11) in the sense of the τ - method, to be much more amenable to computer programming than Lanczos' original idea, and to be applicable to any kind of equation involving a linear (algebraic, differential, or integral) operator mapping $\mathbb{P}$ into itself, such as D in (1.10) or its integrated forms. There are, however, important differences to be considered between this approximation principle and Lanczos' original idea. For instance, in the Lanczos' τ - method the perturbation H_n is chosen in advance, whereas here it is not.

1.4. NUMERICAL SOLUTION OF THE τ - APPROXIMATION PROBLEM

There are essentially two approaches to the numerical solution of the τ - problem (1.12) - (1.13), one in terms of the matrix operator representation of D acting on v [23], to be described next for completeness and the other in terms of the sequence $Q = \{Q_k(x)\}_{k \in \mathbb{N}_0}$ of canonical polynomials associated with D and v, which are given by the functional equation

$$DQ_k(x) = v_k(x) \quad , \quad k = 0, 1, \ldots \quad , \tag{1.17}$$

(cf. [8,9]) obviously inconsistent for $k \in S$, or by the Ortiz' [14,15] redefining equation

$$\begin{aligned} DQ_k(x) &= v_k(x) + r_k(x) \quad , \quad r_k \in R_S \quad , \quad k \notin S \quad , \\ r_k(x) &= -v_k(x) \qquad\qquad\quad , \quad k \in S \quad , \end{aligned} \tag{1.18}$$

based on the fact in (1.14) that every element of $\mathbb{P}$ is uniquely decomposed into the sum of an element in R_D with another in R_S.

Canonical polynomials are, in fact, equivalence classes modulo $K_D = \{P \in \mathbb{P} : DP = 0\}$, the set of exact polynomial solutions of the given

equation, but this is a technical point, the details of which are to be found in [14].

Needless to say, the above definitions (1.17) - (1.18) extend immediately to any linear operator L mapping $\mathbb{P}$ into itself.

2. POLYNOMIAL τ - METHOD APPROXIMATION IN MATRIX OPERATIONAL TERMS

To show that the polynomial y_n in (1.12) - (1.13) is a τ - approximant of the solution y of the problem in (1.9) - (1.11), we review and extend some basic definitions and notation relative to the principle of using matrix operations in the Lanczos' τ - method, which has been developed in [13], [21], and [19,20].

By furnishing zero components, if need be, all vectors in the sequel are infinite - dimensional. Columnvectors are underlined once and row - vectors twice.

Let $\underline{v}$ and $D\underline{v}$ stand for the vectors with components v_k and Dv_k respectively, $k \geq 0$, then

$$D\underline{v} = \Pi_v \underline{v} \quad ,$$

Π_v being the matrix operator representation of D acting on $\mathbb{P}$ when we take for $\mathbb{P}$ the basis v. Π_v may be obtained directly whenever the laws of differentiation and multiplication in v are simple enough, otherwise we may work in the basis $\{x^k\}_{k=0,1\ldots}$ to get

$$D\underline{x} = \Pi_x \underline{x} \quad , \qquad \Pi_x = \sum_{r=0}^{\nu} \eta^r p_r(\mu) \quad ,$$

$$\eta = [\underline{e}_1, 2\underline{e}_2, 3\underline{e}_3, \ldots] \quad , \qquad \mu = [\underline{0}, \underline{e}_0, \underline{e}_1, \ldots] \quad ,$$

$\underline{e}_k$ being the vector with 1 in the kth position and 0 elsewhere, and switch to the basis v to get $\Pi_v = V \Pi_x V^{-1}$, V being such that $\underline{v} = V\underline{x}$. We refer to [18] and [21] for computational details and for structural properties of Π_x and Π_v. Π_x is a band matrix operator and its band width is $\leq \nu + h + 1$. Π_v is no longer banded from below, but is still banded from above.

If we let $y = \underline{\alpha}\,\underline{v}$ be the formal v - series expansion of the solution of the problem in (1.9) - (1.11), express $f(x)$ in the basis v, $f(x) = \underline{\underline{F}}\,\underline{v}$, introduce the vector $\underline{\sigma} = (\sigma_1, \ldots, \sigma_\nu, 0, 0, \ldots)$ and the matrix

$$B_\nu = (b_{ij}) \ , \quad b_{ij} = \begin{cases} g_j(v_i) \ , & j = 1(1)\nu \ ; \ i = 0, 1, \ldots \\ 0 \ , & j > \nu \ , \end{cases}$$

to express the supplementary conditions in (1.11) in the form

$$\underline{\underline{\alpha}} \, B_\nu = \underline{\underline{\sigma}} \ ,$$

and define the matrix

$$\Gamma_\nu = B_\nu + \Pi_\nu \, \mu^\nu$$

and the vector

$$\underline{\underline{\beta}} = \underline{\underline{\sigma}} + \underline{\underline{F}} \, \mu^\nu \ .$$

(postmultiplication of Π_ν and $\underline{\underline{F}}$ by μ^ν shifts their column entries ν places to the right), then $\underline{\underline{\alpha}}$ satisfies the infinite system of linear algebraic equations

$$\underline{\underline{\alpha}} \, \Gamma_\nu = \underline{\underline{\beta}} \ ,$$

whose truncation to its first $n+1$ equations, $n \geq N + \nu$,

$$\underline{\underline{\alpha}}^{(n)} \, \Gamma_\nu = \underline{\underline{\beta}} \ , \tag{2.1}$$

leads to the coefficient vector $\underline{\underline{\alpha}}^{(n)} = (\alpha_0^{(n)}, \ldots, \alpha_n^{(n)}, 0, 0, \ldots)$ of $y_n = \underline{\underline{\alpha}}^{(n)} \, \underline{v}$, which is a polynomial approximation of y in the sense of the τ-method. Indeed, with $\underline{\underline{\alpha}}^{(n)}$ given by

$$\underline{\underline{\alpha}}^{(n)} \, B_\nu = \underline{\underline{\beta}} \ , \quad \underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_j = F_j \ , \quad j = 0(1) \ n - \nu \ ,$$

the first ν equations representing the supplementary conditions in (1.13), then

$$\begin{aligned} Dy_n(x) &\equiv \underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{v} \\ &= \sum_{j=0}^{n-\nu} (\underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_j) \, v_j + \sum_{j=n-\nu+1}^{n+h} (\underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_j) \, v_j \\ &= f(x) + \sum_{j=1}^{\nu+h} (\underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_{n-\nu+i}) \, v_{n-\nu+i} \end{aligned}$$

agrees with $Dy(x)$ as far as possible, i.e., up to $v_{n-\nu}$, and we have

solved the τ - problem in (1.12) - (1.13) with the perturbation

$$(2.2)\quad H_n = \sum_{i=1}^{\nu+h} \tau_i^{(n)} v_{n-\nu+i} \ , \quad \tau_i^{(n)} = \underline{\alpha}^{(n)} \Pi_\nu \underline{e}_{n-\nu+i} \ , \quad i = 1(1)\, \nu+h \ .$$

For rapid convergence, the components of $\underline{\alpha}^{(n)}$ should tend to zero fast. To achieve this, a convenient orthogonal basis has to be chosen. Lanczos would have taken a perturbation like that in (2.2), with $v_k = T_k^*(x)$, $k = 0, 1, \ldots$, the Chebyshev polynomials shifted to $a \le x \le b$, to get a good global polynomial approximation y_n of y . There are, however, important differences to be considered between the above approximation principle and Lanczos' original perturbation idea. For instance, in the Lanczos' τ - method, the perturbation is chosen in advance, whereas here it is not. On the other hand, the evaluation of the Lanczos' τ - parameters is not generally amenable to computer programming, whereas (2.2) gives them immediately, the moment $\underline{\alpha}^{(n)}$ is obtained.

While error analysis for the general polynomial τ - method approximation is undoubtedly difficult, an upper bound for $\| \varepsilon_n \|$ may be easily obtained from (1.16). Also, from (1.15), an efficient estimation of ε_n may be obtained as follows (see [23] and references given there for details, applications, and numerical examples).

Let $\varepsilon_{n,m}$, $m > n$, be an m th order polynomial approximation of ε_n , then $\varepsilon_{n,m}$ satisfies the perturbed ODE

$$D\, \varepsilon_{n,m}(x) = - H_n(x) + H_m(x) \ , \qquad a \le x \le b \ ,$$

is such that

$$\varepsilon_{n,m}(x) = y_m(x) - y_n(x) \ ,$$

and thus, every time two τ - approximants $y_n(x)$ and $y_m(x)$ are computed, an estimation of $\varepsilon_n(x)$ is obtained.

Clearly, all we have said about the differential operator D extends immediately to any linear operator L mapping $\mathbb{P}$ into itself. As pointed out in [6], the only embarrassment of the method is the accidental singularity of the system (2.1). If this happens, we merely change n to $n+1$ and keep doing this until we find a nonsingular system. This must ultimately exist if the solution of the given problem has a convergent ν - series expansion.

3. RECURSIVE CONSTRUCTION OF POLYNOMIAL τ - APPROXIMANTS IN TERMS OF CANONICAL POLYNOMIALS

The canonical and residual polynomials associated with a given linear operator $L : \mathbb{P} \to \mathbb{P}$ and a given basis v may be computed recursively [15].

Writing

$$L v_n = \sum_{j=0}^{m} \Pi_{nj} v_j \quad , \quad m = n + h \quad ,$$

$$= \Pi_{nm} v_m + \sum_{j=0}^{m-1} \Pi_{nj} (L Q_j - r_j) \quad ,$$

$$L \left(v_n - \sum_{j=0}^{m-1} \Pi_{nj} Q_j\right) = \Pi_{nm} v_m - \sum_{j=0}^{m-1} \Pi_{nj} r_j \quad ,$$

assuming that $\Pi_{nm} \neq 0$ and that Q_j and r_j , $j = 0(1)\ m-1$, have already been computed, then

$$Q_m = \left(v_n - \sum_{j=0}^{m-1} \Pi_{nj} Q_j\right) / \Pi_{nm}$$

$$r_m = -\left(\sum_{j=0}^{m-1} \Pi_{nj} r_j\right) / \Pi_{nm} \qquad n = 0,\ 1, \ldots$$

Once the first $M = \max \{m , N\}$ canonical and residual polynomials are known, the τ - solution y_n of the equation

$$Ly = f \quad , \qquad f = \sum_{k=0}^{N} F_k v_k \quad ,$$

which is the exact polynomial solution of the perturbed equation

$$Ly_n = f + H_n \quad , \quad H_n = \sum_{k=0}^{m} h_k v_k \quad ,$$

$$= Z_M \quad , \quad Z_M = \sum_{k=0}^{M} z_k v_k \quad , \quad z_k = F_k + h_k \quad ,$$

is immediately at hand,

$$y_n = \sum_{k=0}^{M} z_k Q_k \quad ,$$

as well as the corresponding compatibility conditions,

$$\sum_{k=0}^{M} z_k \, r_k \equiv 0 \quad ,$$

which are easily seen to be equivalent to the formal use of the undefined canonical polynomials (see (1.17)) and the subsequent cancellation of their coefficients.

In addition to satisfying a self-starting recurrence relation and being an efficient basis for the representation of τ - solutions, canonical polynomials have a number of other useful properties, namely, they are

i) Permanent, and so, if we need y_{n+1} after y_n has been constructed, namely to improve the approximation accuracy, only Q_{m+1} has to be computed ;

ii) Independent of the given supplementary conditions, hence initial and boundary value problems are treated alike ;

iii) Independent of the approximation interval, and so piecewise polynomial and rational τ - approximants are easily constructed (see, e.g., [1], [2], [4,5], [11,12], and [16,17], where they choose a convenient perturbation in advance, and [24], where we just accept the perturbation the given problem leads to).

4. THE LINK BETWEEN THE TWO FOREGOING APPROACHES TO THE NUMERICAL SOLUTION OF THE τ - APPROXIMATION PROBLEM

Let Π_v be the matrix operator representation of a given linear operator $L : \mathbb{P} \to \mathbb{P}$ when we take for $\mathbb{P}$ the basis v, i.e.,

$$L\underline{v} = \Pi_v \underline{v} \quad ,$$

let $Q = \{Q_k\}_{k \in \mathbb{N}_0 - S}$ and $r = \{r_k\}_{k \in \mathbb{N}_0 - S}$ be the sequences of canonical and residual polynomials associated with L and v, and assume, with no loss of generality, that $S = \{0, 1, \ldots, s-1\}$. Following [20], we define the vectors $\underline{r} = (r_s, r_{s+1}, \ldots)$ and $\underline{Q} = (Q_s, Q_{s+1}, \ldots)$ and the matrices $[R;0]$ and C such that

$$\underline{r} = [R;0] \, \underline{v} \quad , \quad \underline{Q} = C \, \underline{v} \quad ,$$

then Ortiz' funcional equations (1.18) may be written as

$$L\underline{Q} = [R;I]\ \underline{v} \quad ,$$

but $L\underline{Q} = C\,\Pi_v\,\underline{v}$, so $C\,\Pi_v = [R;I]$, and if we set $\Pi_v = [\Pi_S;\Pi_Q]$, then

$$C\,\Pi_Q = I \quad , \quad C\,\Pi_S = R \ .$$

The calculation of the canonical polynomials Q_k , $k \notin S$, is, therefore, equivalent to the inversion of the matrix Π_v stripped of its columns of order $k \in S$.

5. EXAMPLES OF APPROXIMATE EXPANSIONS IN SERIES OF ORTHOGONAL POLYNOMIALS

1) To construct approximate expansions of the function

$$y(x) = ((1-x)/2)^{1/2} \quad , \qquad -1 \le x \le 1 \quad ,$$

in the Legendre basis $v = \{P_n(x)\}_{n=0,1,\ldots}$,

$$y(x) = \underline{\underline{\alpha}}\,\underline{v} = \frac{2}{3}P_0(x) - 2\sum_{n=1}^{\infty}\frac{P_n(x)}{(2n-1)(2n+3)} \quad , \tag{5.1}$$

we choose a definition of $y(x)$ in terms of a linear operator $L: \mathbb{P} \to \mathbb{P}$, e.g.,

$$L\,y(x) \equiv (1-x)\,y(x) + \frac{3}{2}\int_{-1}^{x} y(t)\,dt = 2 \quad ,$$

integrated form of the IVP

$$D\,y(x) \equiv 2(1-x)\,y'(x) + y(x) = 0 \quad , \quad y(-1) = 1 \quad ,$$

and construct the matrix operator Π_v such that $L\underline{v} = \Pi_v\,\underline{v}$, i.e.,

$$L\,P_0 = \frac{5}{2}P_0 + \frac{1}{2}P_1$$

$$L\,P_n = -\frac{2n+3}{2(2n+1)}P_{n-1} + P_n - \frac{2n-1}{2(2n+1)}P_{n+1} \quad , \quad n = 1, 2, \ldots$$

Thanks to the structure of Π_v , the polynomial approximation $y_n = \underline{\underline{\alpha}}^{(n)}\,\underline{v}$ of y is such that

$$L\,y_n = 2 + \alpha_n^{(n)}\,P_{n+1} \quad , \quad n = 0, 1, \ldots \tag{5.2}$$

and its coefficient vector $\underline{\alpha}^{(n)}$ may be obtained either by solving the following system of linear algebraic equations

$$\underline{\alpha}^{(n)} \Pi_v \underline{e}_0 = 2 , \quad \underline{\alpha}^{(n)} \Pi_v \underline{e}_j = 0 , \quad j = 1(1) n ,$$

or by using the canonical and residual polynomials associated with L and v ,

$$Q_0 = 0 , \quad r_0 = -1 ; \quad Q_1 = 2 P_0 , \quad r_1 = 5 ;$$

$$Q_{n+1} = -\frac{1}{2n-1} \left[2(2n+1)(P_n - Q_n) + (2n+3) Q_{n-1}\right] ,$$

$$r_{n+1} = \frac{1}{2n-1} \left[2(2n+1) r_n - (2n+3) r_{n-1}\right] , \quad n = 1, 2, \ldots$$

From (5.2) and (1.18) we obtain

$$y_n = \alpha_n^{(n)} Q_{n+1} , \quad \alpha_n^{(n)} = \frac{2}{r_{n+1}} , \quad n = 0, 1, \ldots ,$$

successive approximations of the corresponding partial sums of the Legendre series in (5.1). In particular, the coefficients of y_4 ,

$$y_4 = \frac{1}{45045} (7525 P_0 - 4452 P_1 - 985 P_2 - 378 P_3 - 135 P_4) ,$$

approximate the corresponding coefficients of y with absolute error $\leq 1.4 \times 10^{-2}$.

2) To construct approximate expansions of

$$y(x) = e^{-x} , \quad 0 \leq x \leq 1 ,$$

in the Hermite basis $v = \{H_n^*(x)\}_{n=0,1\ldots}$,

$$y(x) = \underline{\alpha}\, \underline{v} = e^{1/4} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n n!} H_n^*(x) ,$$

let us define $y(x)$ in terms of the following linear operator

$$L y(x) \equiv y(x) + \int_0^x y(t)\, dt = 1 .$$

Recalling that

$$\int_0^x H_n^*(t)\,dt = \frac{1}{2(n+1)}\left[H_{n+1}^*(x) - H_{n+1}^*(0)\right] ,$$

$$H_{2n}^*(0) = (-1)^n \frac{(2n)!}{n!} , \quad H_{2n+1}^*(0) = 0 , \quad n = 0, 1,\ldots ,$$

we get

$$L H_n^*(x) = \frac{1}{2(n+1)}\left[H_{n+1}^*(x) + 2(n+1)\, H_n^*(x) - H_{n+1}^*(0)\right] ,$$

hence

$$Q_0 = 0 , \quad r_0 = -1$$

$$Q_n = 2n(H_{n-1}^* - Q_{n-1}) , \quad r_n = -2n\, r_{n-1} - H_n^*(0) , \quad n = 1, 2,\ldots$$

The polynomial approximation $y_n = \underline{\alpha}^{(n)} \underline{v}$ of $y = \underline{\alpha}\, \underline{v}$ satisfies the perturbed equation

$$L y_n = 1 + \alpha_n^{(n)} H_{n+1}^* , \quad n = 0, 1,\ldots ,$$

and is such that

$$y_n = \frac{Q_{n+1}}{r_{n+1}} , \quad n = 0, 1,\ldots .$$

In particular, the coefficients $\alpha_k^{(n)}$, $n = 0(1)\,5$, $k = 0(1)\,n$, of the first six y_n's ,

$$y_0 = H_0$$

$$y_1 = \frac{2}{3}(2H_0^* - H_1^*)$$

$$y_2 = \frac{1}{6}(8H_0^* - 4H_1^* + H_2^*)$$

$$y_3 = \frac{2}{75}(48H_0^* - 24H_1^* + 6H_2^* - H_3^*)$$

$$y_4 = \frac{1}{300}(384H_0^* - 192H_1^* + 48H_2^* - 8H_3^* + H_4^*)$$

$$y_5 = \frac{1}{2990}(3840H_0^* - 1920H_1^* + 480H_2^* - 80H_3^* + 10H_4^* - H_5^*) ,$$

approximate the corresponding coefficients α_k of y with the errors given below

n \ k	$\alpha_k - \alpha_k^{(n)}$					
0	2.84×10^{-1}					
1	-4.93×10^{-2}	2.47×10^{-2}				
2	-4.93×10^{-2}	2.47×10^{-2}	-6.16×10^{-3}			
3	4.03×10^{-3}	-2.01×10^{-3}	5.03×10^{-4}	-8.39×10^{-5}		
4	4.03×10^{-3}	-2.01×10^{-3}	5.03×10^{-4}	-8.39×10^{-5}	1.05×10^{-5}	
5	-2.56×10^{-4}	1.28×10^{-4}	-3.19×10^{-5}	5.32×10^{-6}	-6.67×10^{-7}	6.60×10^{-8}

REFERENCES

1. J.P. COLEMAN: The Lanczos tau-method. J. Inst. Maths. Applics 17 (1976), 85-97.

2. M.R. CRISCI: The tau method with perturbation term depending on the differential operator. J. Comp. Appl. Maths. 15 (1986), 123-136.

3. M.R. CRISCI and E.L. ORTIZ: Existence and convergence results for the numerical solution of differential equations with the tau method. Imperial College NAS Res. Rep., University of London, London 1981.

4. M.R. CRISCI and E. RUSSO: A stability of a class of methods for the numerical integration of certain linear systems of ordinary differential equations. Math. Comp. 38 (1982), 431-435.

5. M.R. CRISCI and E. RUSSO: An extension of Ortiz' recursive formulation of the tau method to certain linear systems of ordinary differential equations. Math. Comp. 41 (1983), 27-42.

6. L. FOX and I.B. PARKER: Chebyshev polynomials in numerical analysis. Oxford Univ. Press, London 1968.

7. C. LANCZOS: Trigonometric interpolation of empirical and analytical functions. J. Math. Phys. 17 (1938), 123-199.

8. C. LANCZOS: Tables of Chebyshev polynomials $S_n(x)$ and $C_n(x)$, Introduction. Nat. Bur. Standards Appl. Math. Ser. 9, U.S. Govt. Printing Office, Washington 1952.

9. C. LANCZOS: Applied analysis. Pitman, London 1957.

10. C. LANCZOS: Legendre versus Chebyshev polynomials. In: Topics in numerical analysis (J. J. H. Miller, ed.), Academic Press, London 1973, pp. 191-201.

11. Y.L. LUKE: The special functions and their approximations. Academic Press, New York 1969.

12. Y.L. LUKE: Mathematical functions and their approximations. Academic Press, London 1975.

13. P. ONUMANYI, E.L. ORTIZ, and H. SAMARA: Software for a method of finite approximations for the numerical solution of differential equations. Appl. Math. Modelling 5 (1981), 282 - 286.

14. E.L. ORTIZ: The tau method. SIAM J. Numer. Anal. 6 (1969), 480 - 492.

15. E.L. ORTIZ: Canonical polynomials in the Lanczos tau method. In: Studies in numerical analysis (B.K.P. Scaife, ed.), Academic Press, London 1974, pp. 73 - 93.

16. E.L. ORTIZ: Step by step tau method - part I. Piecewise polynomials approximations. Comp. & Maths. with Appls. 1 (1975), 381 - 392.

17. E.L. ORTIZ: Sur quelques nouvelles applications de la methode tau. In: Séminaires IRIA analyse et contrôle de systèmes, IRIA, Paris 1975, pp. 247 - 257.

18. E.L. ORTIZ and H. SAMARA: An operational approach to the tau method for the numerical solution of nonlinear differential equations. Computing 27 (1981), 15 - 25.

19. E.L. ORTIZ and H. SAMARA: Matrix displacement mappings in the numerical solution of functional and nonlinear differential equations with the tau method. Num. Funct. Anal. and Optimiz. 6 (1983), 379 - 398.

20. E.L. ORTIZ and H. SAMARA: Numerical solution of differential eigenvalue problems with an operational approach to the tau method. Computing 31 (1983), 95 - 103.

21. M.R. da SILVA: LACALGEBRA versions of Lanczos' tau method for the numerical solution of differential equations. Port. Math. 41 (1982), 295 - 316.

22. M.R. da SILVA: A quick survey of recent developments and applications of the τ - method. In: Numerical approximation of partial differential equations (E.L. Ortiz, ed.), North-Holland, Amsterdam 1987, pp. 297-308.

23. M.R. da SILVA: Numerical treatment of differential equations with the τ - method. J. Comp. Appl. Math. 20 (1987), 1 - 7.

24. M.R. da SILVA and M.J. RODRIGUES: A simple alternative principle for rational τ - method approximation. To appear in the proceedings of the conference on Nonlinear Numerical Methods and Rational Approximation, Antwerp 1987.

ON MONOTONICITY OF SOME LINEAR POSITIVE OPERATORS

B. DELLA VECCHIA

ABSTRACT: In this paper we study the monotonicity of the sequences of some linear positive operators, to which we apply the iterated Nörlund operator. As particular cases, we find the results established by D.D. Stancu for the sequence $\{(B_n f)^{(m)}(x)\}_n$ and by the author for the sequences $\{(M_n f)^{(m)}(x)\}_n$ and $\{(P_n f)^{(m)}(x)\}_n$ where M_n and P_n are the Favard-Szasz-Mirakyan and Baskakov operator respectively.

1. INTRODUCTION

It is well known that the Bernstein Polynomials corresponding to functions convex in $[0,1]$ verify the following monotonicity relationship [1,18,28]:

$$(1)\quad (B_{n+1}f)(x) = B_{n+1}f(x) \leqq B_n f(x), \quad 0 \leqq x \leqq 1$$

This property has been extended to other linear positive operators [3,8,9,10,11,12,19].

Later Stancu in [24] studied the derivatives of the sequence of Bernstein polynomials and obtained interesting monotonicity properties for this sequence.

Then Horová in [7] established a relation of type (1) for the first derivatives of Favard-Szasz-Mirakyan operator [5,15,16].

Recently in [4] we have proved monotonicity properties for the derivatives of order s, $s > 1$, of the sequence of Favard-Szasz-Mirakyan operator and for the derivatives of order s, $s \geqq 1$, of the sequence of Baskakov operator [2,6,27].

On the other hand some authors introduced separately discrete type operators generalizing Bernstein, Favard-Szasz-Mirakyan and Baskakov operators.

The main purpose of this paper is to extend the procedure given in [24] to the sequence of these operators, to which we apply the iterated Nörlund difference operator, instead of the differentiation operator.

As a particular case, we find the results established in [4].

2. PRELIMINARY RESULTS

Let f be a function defined on an interval I of the real axis. As usual, we denote by $[t_0,t_1,\dots,t_n;f]$ the divided difference of order n, of the function f, with respect to the distinct nodes $t_0,t_1,\dots,t_n \in I$.

We recall also that f is called convex, non-concave, polynomial, non-convex respectively concave of n-order on an interval $I=[a,b]$, if all its divided differences of order n+1, on n+2 distinct nodes from I, are >0, $\geqq 0$, $=0$, $\leqq 0$, resp. <0.

We use in the sequel also the formula

$$(2.1)\quad D_\alpha^m[f(x)g(x)] = \sum_{i=0}^{m} \binom{m}{i} D_\alpha^i f(x+(m-i)\alpha) D_\alpha^{m-i} g(x)$$

with $m \in N$, f and g defined on I, $x \in I$, $\alpha \in R^+$ and

$$D_\alpha g(x) = [g(x+\alpha)-g(x)]\alpha^{-1}, \quad D_\alpha^m = D_\alpha(D_\alpha^{m-1}), \quad D_\alpha^0 g(x) = g(x).$$

Then let r and n be two integers, with $0 \leqq r \leqq n$, and consider the following points of the interval I: $a_i=a+ih$, $i=0,1,\dots,n$ and $b_j=a+j\ell$, $j=1,2,\dots,n$, where $0<h\leqq\frac{b-a}{n}$, $0<\ell<\frac{b-a}{n}$.

Now we denote by $T_k^{(\nu)}$, $0\leqq k\leqq n$, $1<\nu\leqq r+1$, the linear functionals defined recursively as follows

$$T_k^{(2)} f = [a_k, a_{k+1}, b_{k+1}; f], \quad 0\leqq k\leqq n-1$$

$$(2.2)\quad T_k^{(\nu+1)} f = T_{k+1}^{(\nu)} f - T_k^{(\nu)} f, \quad 1<\nu\leqq r, \quad 0\leqq k\leqq n-r$$

where f is a function defined on [a,b].

This functional has been introduced in [24] by Stancu, who proved there that

$$\text{2.3)}\quad \text{f convex (non-concave) of order } r+1 \Longrightarrow T_k^{(r+2)} f > 0 \quad (T_k^{(r+2)} f \geqq 0), \quad 0 \leqq k \leqq n-r$$

We consider now the class of linear and positive operators V_n^α defined by

$$\text{2.4)}\quad V_n^\alpha f(x) = \sum_{k=0}^{\infty} (-1)^k D_\alpha^k \phi_n^\alpha(x) \frac{x^{(k,-\alpha)}}{k!} f\left(\frac{k}{n}\right)$$

where: α is a non-negative parameter that can depend only on $n \in N$; $x^{(k,-\alpha)} = x(x+\alpha)\ldots(x+(k-1)\alpha)$; the functions ϕ_n^α $(n \in N)$ are defined on R and verify the following conditions:

i) $\phi_n^\alpha(0) = 1$;

ii) $\forall k \in N$ and $\forall x \in R\ (-1)^k D_\alpha^k \phi_n^\alpha(x) \geqq 0$;

iii) $\sum_{k=0}^{\infty} (-1)^k D_\alpha^k \phi_n^\alpha(x) \frac{x^{(k,-\alpha)}}{k!} = 1$;

$f \in C^*$, where C^* denotes the set of functions defined on $[0,+\infty[$ and such that (2.4) has meaning.

This operator has been introduced and studied in [11,12,23].

Setting

$$\gamma_{n,i}^\alpha = \frac{(-1)^i D_{-\alpha}^i \phi_n^\alpha(0)}{n^i}, \quad \gamma_{n,0}^\alpha = 1$$

it is known [11] that

Theorem 2.1. If

$$\lim_n \gamma_{n,r+i}^\alpha = 1, \quad i=0,1,2, \quad r \in N$$

with $0<\alpha=\alpha(n)\to 0$, when $n\to\infty$, then we have

$$\lim_n \left\| f^{(r)} - D_\alpha^r V_n^\alpha f \right\| = 0,$$

$\forall f^{(r)} \in \bar{C}^\circ$, where $\bar{C}^\circ$ denotes the set of functions defined on $[0,+\infty[$, there bounded and uniformly continuous, and $\forall r \in N$.

We notice that, by choosing suitably ϕ_n^α functions, V_n^α beco-

mes well-known linear positive operators, studied separately by some authors in [13,14,17,19,20-22,25].

Here we want to consider the following three particular cases.

1) If

$$\phi_n^{\alpha}(x) = \frac{(1-x)^{(n,-\alpha)}}{1^{(n,-\alpha)}}, \qquad 0 \leqq x \leqq 1$$

V_n^{α} coincides with the Stancu operator S_n^{α} defined by

$$(2.5)\quad S_n^{\alpha} f(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{(1-x)^{(n-k,-\alpha)}}{1^{(n,-\alpha)}} x^{(k,-\alpha)} f(\frac{k}{n}), \qquad \begin{matrix} 0 \leqq x \leqq 1 \\ f \in C([0,1]). \end{matrix}$$

Stancu introduced in [19] this operator, which later has been studied in [11,13,14,20,22,23].

We notice that, for $\alpha=0$, S_n^{α} becomes equal to Bernstein operator B_n.

2) If

$$\phi_n^{\alpha}(x) = (1+n\alpha)^{-x/\alpha}, \qquad x \geqq 0, \qquad 0 \leqq n\alpha \leqq 1$$

V_n^{α} coincides with the M_n^{α} operator defined by

$$(2.6)\quad M_n^{\alpha} f(x) = (1+n\alpha)^{-x/\alpha} \sum_{k=0}^{\infty} (\alpha+\frac{1}{n})^{-k} \frac{x^{(k,-\alpha)}}{k!} f(\frac{k}{n})$$

$f \in E_{\beta,B}$, where $E_{\beta,B}$ denotes the set of functions defined on $[0,+\infty[$, continuous in $[0,B]$, $(B>0)$ and such that $f(x)=O(2^{\beta x})$ $(x\to\infty)$, with β a positive fixed number.

This operator was proposed by different approaches in [12,17,23,25]. We notice that, for $\alpha=0$, (2.6) becomes

$$(2.7)\quad M_n f(x) = e^{-nx} \sum_{k=0}^{\infty} \frac{(nx)^k}{k!} f(\frac{k}{n})$$

where M_n is Favard-Szasz-Mirakyan operator [3,5,7,9,15,19,26

3) If

$$\phi_n^{\alpha}(x) = (1+x)^{(-n,-\alpha)}\, 1^{(n,-\alpha)} = \frac{1^{(n,-\alpha)}}{(1+x)^{(n,-\alpha)}} \qquad \begin{matrix} x \geqq 0 \\ 0 \leqq \alpha < 1/2 \end{matrix}$$

V_n^{α} coincides with Baskakov-Stancu operator P_n^{α} defined by

$$(2.8)\quad P_n^{\alpha} f(x) = 1^{(n,-\alpha)} \sum_{k=0}^{\infty} \binom{n+k-1}{k} \frac{x^{(k,-\alpha)}}{(1+x)^{(n+k,-\alpha)}} f(\frac{k}{n}), \quad f \in C^*$$

This operator was introduced by Stancu in [21,23] and later studied by Mastroianni in [11]. As a particular case, for $\alpha=0$, P_n^{α} becomes equal to Baskakov operator [2,6,9,27]

$$2.9)\quad P_n f(x) = \sum_{k=0}^{\infty} \binom{n+k-1}{k} \frac{x^k}{(1+x)^{n+k}} f\left(\frac{k}{n}\right)$$

. ON THE MONOTONICITY OF THE SEQUENCE $\{D_{\alpha}^{m} S_n^{\alpha} f(x)\}_n$

Let $S_n^{\alpha} f$ be the Stancu operator defined by (2.5). It is ell known [19] that two consecutive terms of the sequence $S_n^{\alpha} f(x)\}_n$ verify the following relationship:

$$(3.1)\quad \begin{aligned} (S_{n+1}^{\alpha} - S_n^{\alpha}) f(x) &= -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot \\ &\cdot \sum_{\nu=0}^{n-1} \frac{(-1)^{\nu} D_{\alpha}^{\nu} \phi_{n-1}^{\alpha}(x-\alpha)(x+\alpha)^{(\nu,-\alpha)}}{\nu!} \left[\frac{\nu}{n}, \frac{\nu+1}{n+1}, \frac{\nu+1}{n}; f\right] \end{aligned} \qquad \begin{matrix} f\in C([0,1]) \\ x\in[0,1] \end{matrix}$$

If we choose the following points

$$a_{\nu+i} = \frac{\nu+i}{n+i}, \qquad b_n = \frac{\nu}{n}, \qquad \nu=0,1,\ldots,n-1, \qquad i=0,1$$

using the functional $T_{\nu}^{(2)}$ defined by (2.2), (3.1) becomes

$$(3.2)\quad \begin{aligned} (S_{n+1}^{\alpha} - S_n^{\alpha}) f(x) &= -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot \\ &\cdot \sum_{\nu=0}^{n-1} (-1)^{\nu} D_{\alpha}^{\nu} \phi_{n-1}^{\alpha}(x-\alpha)(x+\alpha)^{(\nu,-\alpha)} T_{\nu}^{(2)} f \end{aligned}$$

Now we introduce, $\forall r\in N$ and $\forall f\in C([0,1])$, the linear positive operator $S_n^{\alpha,r}$

$$S_n^{\alpha,r} f(x) = \sum_{\nu=0}^{n-r} (-1)^{\nu+r-1} \frac{D_{\alpha}^{\nu+r-1} \phi_{n-1}^{\alpha}(x-\alpha)(x+r\alpha)^{(\nu,-\alpha)}}{\nu!} f\left(\frac{\nu}{n}\right)$$

From (3.2) it follows that

$$(3.3)\quad (S_{n+1}^{\alpha}-S_n^{\alpha})f(x) = -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} S_n^{\alpha,1} T_{\nu}^{(2)} f(x)$$

Now, in order to study the monotonicity of the sequence $\{D_{\alpha}^{m} S_n^{\alpha} f(x)\}_n$, we prove

Theorem 3.1. The following relationship holds:

$$D_{\alpha}^{m}(S_{n+1}^{\alpha}-S_n^{\alpha})f(x) = -\frac{1}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot$$

$$(3.4)\quad \cdot\Big[(x+m\alpha)(1-x-m\alpha)S_n^{\alpha,m+1} T_{\nu}^{m+2} f(x) +$$

$$+ m(1-2x-(2m-1)\alpha)S_n^{\alpha,m} T_{\nu}^{(m+1)} f(x) - m(m-1)S_n^{\alpha,m-1} T_{\nu}^{(m)} f(x)\Big]$$

with $m \leqq n$ and $\alpha \geqq 0$

Proof.

In fact, using (2.1) in (3.3), with

$$f(x) = -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \text{ and } g(x) = S_n^{\alpha,1} T_{\nu}^{(2)} f(x),$$

we have

$$D_{\alpha}^{m}(S_{n+1}^{\alpha}-S_n^{\alpha})f(x) = -\frac{1}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot$$

$$\cdot \Big\{(x+m\alpha)(1-x-m\alpha)D_{\alpha}^{m} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) +$$

$$+ mD_{\alpha}^{1}[(x+(m-1)\alpha)(1-x-(m-1)\alpha)]\, D_{\alpha}^{m-1} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) +$$

(3.5)

$$+\frac{m(m-1)}{2} D_{\alpha}^{2}[(x+(m-2)\alpha)(1-x-(m-2)\alpha)]\, D_{\alpha}^{m-2} S_n^{\alpha,1} T_{\nu}^{(2)} f(x)\Big\} =$$

$$= -\frac{1}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)}\Big\{(x+m\alpha)(1-x-m\alpha)D_{\alpha}^{m} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) +$$

$$+m(1-2x-(2m-1)\alpha)D_{\alpha}^{m-1} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) - m(m-1)D_{\alpha}^{m-2} S_n^{\alpha,1} T_{\nu}^{(2)} f(x)\Big\}$$

On the other hand, one can easily verify by induction that

$$D_{\alpha}^{m} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) = S_n^{\alpha,m+1} T_{\nu}^{(m+2)} f(x)$$

and making use of this last relation in (3.5), the theorem follows.

From Theorem 3.1, by (2.3), we obtain

Corollary 3.2. The sequence

$$(3.6)\quad \{D_{\alpha}^{m}S_{n}^{\alpha}f(x)\}_{n}, \qquad 0\leq m\alpha\leq 1, \qquad m\leq n$$

verifies the following monotonicity properties:

i) for m=1

a) If on the interval $\left[0,\frac{1-\alpha}{2}\right]$ the function f is convex (concave) of first and second order, then the sequence (3.6) is decreasing (increasing) on $\left[0,\frac{1-\alpha}{2}\right]$;

b) If on the interval $\left[\frac{1-\alpha}{2},1-\alpha\right]$ the function f is concave (convex) of first order and convex (concave) of second order, then the sequence (3.6) is decreasing (increasing) on the interval $\left[\frac{1-\alpha}{2},1-\alpha\right]$;

c) If on the interval $[1-\alpha,1]$ the function f is concave (convex) of first and second order, then the sequence (3.6) is decreasing (increasing) on $[1-\alpha,1]$.

ii) for m≥2

a) If on the interval $\left[0,\frac{1-(2m-1)\alpha}{2}\right]$ the function f is concave (convex) of order m-1 and convex (concave) of order m and m+1, then the sequence (3.6) is decreasing (increasing) on $\left[0,\frac{1-(2m-1)\alpha}{2}\right]$;

b) If on the interval $\left[\frac{1-(2m-1)\alpha}{2},1-m\alpha\right]$ the function f is concave (convex)of order m-1 and m and convex (concave) of order m+1, then the sequence (3.6) is decreasing (increasing) on $\left[\frac{1-(2m-1)\alpha}{2},1-m\alpha\right]$;

c) If on the interval $[1-m\alpha,1]$ the function f is concave (convex) of order m-1, m and m+1, then the sequence (3.6) is decreasing (increasing) on $[1-m\alpha,1]$.

We notice that, for α=0, Corollary 3.2 gives us a monotonicity result for the derivatives of the sequence of Bernstein polynomials, obtained by Stancu in [24].

4. ON THE MONOTONICITY OF THE SEQUENCE $\{D_\alpha^m M_n^\alpha f(x)\}_n$

Let $M_n^\alpha f$ be the operator introduced in (2.6).
It is known [12] that the following relationship holds for two consecutive terms of the sequence $\{M_n^\alpha f(x)\}_n$:

$$(4.1)\qquad (M_{n+1}^\alpha - M_n^\alpha) f(x) = -\frac{x}{n(n+1)} \cdot \sum_{k=0}^{\infty} (\alpha - \frac{1}{n})^{k+1} D_\alpha^{k+1} \phi_n^\alpha(x) \frac{(x+\alpha)^{(k,-\alpha)}}{k!} \left[\frac{k}{n}, \frac{k+1}{n+1}, \frac{k+1}{n}; f\right] \quad \begin{matrix} f \in E_{\beta,B} \\ x \geq 0 \end{matrix}$$

We introduce now, $\forall r \in N$ and $\forall f \in E_{\beta,B}$, the operator $M_n^{\alpha,r}$ defined as follows

$$M_n^{\alpha,r} f(x) = \frac{1}{n} \sum_{k=0}^{\infty} (-1)^{k+r} D_\alpha^{k+r} \phi_n^\alpha(x) \frac{(x+r\alpha)^{(k,-\alpha)}}{k!} f(\frac{k}{n})$$

One can easily verify that this operator is linear and positive.
Letting then

$$a_{k+i} = \frac{k+i}{n+i} \quad \text{and} \quad b_k = \frac{k}{n}, \qquad k = 0,1,\ldots, \qquad i = 0,1$$

and recalling the definition of the functional $T_k^{(2)}$ introduced in (2.2), (4.1) becomes

$$(4.2)\qquad (M_{n+1}^\alpha - M_n^\alpha) f(x) = -\frac{x}{n(n+1)} M_n^{\alpha,1} T_k^{(2)} f(x)$$

Now, in order to study the monotonicity of the sequence $\{D_\alpha^m M_n^\alpha f(x)\}_n$, we prove

Theorem 4.1. The following relationship holds

$$(4.3)\qquad D_\alpha^m [M_{n+1}^\alpha - M_n^\alpha] f(x) = -\frac{x+m\alpha}{n(n+1)} M_n^{\alpha,m+1} T_k^{(m+2)} f(x) + \\ - \frac{m}{n(n+1)} M_n^{\alpha,m} T_k^{(m+1)} f(x), \qquad m \in N \quad \text{and} \quad 0 \leq n\alpha \leq 1$$

Proof.
Indeed, by applying (2.1) to (4.2), with

$$f(x) = -\frac{x}{n(n+1)} \quad \text{and} \quad g(x) = M_n^{\alpha,1} T_k^{(2)} f(x),$$

we have

$$D_\alpha^m [M_{n+1}^\alpha - M_n^\alpha] f(x) = -\frac{1}{n(n+1)} \sum_{i=0}^{1} \binom{m}{i} D_\alpha^i [x+(m-i)\alpha] \cdot$$

$$(4.4) \quad \cdot D_\alpha^{m-i} M_n^{\alpha,1} T_k^{(2)} f(x) =$$

$$= -\frac{1}{n(n+1)} \left[(x+m\alpha) D_\alpha^m M_n^{\alpha,1} T_k^{(2)} f(x) + m D_\alpha^{m-1} M_n^{\alpha,1} T_k^{(2)} f(x) \right]$$

Moreover, one can easily prove by induction that

$$D_\alpha^m M_n^{\alpha,1} T_k^{(2)} f(x) = M_n^{\alpha,m+1} T_k^{(m+2)} f(x)$$

and, by using this last relationship in (4.4), the theorem follows.

From Theorem 4.1, by (2.3), we have

Corollary 4.2. For the sequence

$$(4.5) \quad \{D_\alpha^m M_n^\alpha f(x)\}_n, \qquad m \in N \quad 0 \leqq n\alpha \leqq 1$$

the following monotonicity property holds:
if on the interval $[0,+\infty[$ the function f is convex (concave) of order m and m+1, then the sequence (4.5) is decreasing (increasing) on $[0,+\infty[$.

We recall that, for $\alpha=0$, M_n^α operator coincides with M_n operator defined by (2.7); so, Corollary 4.2 represents an extension of a result previously established in [4] for the sequence $\{(M_n f)^{(m)}(x)\}_n$.

5. ON THE MONOTONICITY OF THE SEQUENCE $\{D_\alpha^m P_n^\alpha f(x)\}_n$

Let $P_n^\alpha f$ be the Baskakov-Stancu operator defined by (2.8). It is well-known [11] that the difference between two consecutive terms of the sequence $\{P_n^\alpha f(x)\}_n$ can be expressed as follows:

$$(P^{\alpha}_{n+1}-P^{\alpha}_{n})\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}x}{n(n+1)}\;\cdot$$

(5.1)
$$\cdot\sum_{k=0}^{\infty}\frac{(n+k+1)\ldots(n+2)}{(1+x)^{(n+k+1,-\alpha)}}\,\frac{(x+\alpha)^{(k,-\alpha)}}{k!}\left[\frac{k}{n+1},\frac{k+1}{n+1},\frac{k+1}{n};f\right],\quad x\geqq 0$$

Letting then

$$a_k=\frac{k}{n+1}\qquad\text{and}\qquad b_k=\frac{k}{n},\qquad k=0,1,\ldots$$

taking (2.2) into account, we have

$$(P^{\alpha}_{n+1}-P^{\alpha}_{n})\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}x}{n(n+1)}\;\cdot$$

(5.2)
$$\cdot\sum_{k=0}^{\infty}\frac{(n+k+1)\ldots(n+2)}{k!}\;\frac{(x+\alpha)^{(k,-\alpha)}}{(1+x)^{(n+k+1,-\alpha)}}\;T^{(2)}_{k}f$$

We introduce now, $\forall r\in N$, the linear positive operator $P^{\alpha,r}_{n}$

$$P^{\alpha,r}_{n}\,f(x) \;=\; \sum_{k=0}^{\infty}\frac{(n+k+1)\ldots(n+2)(x+r\alpha)^{(k,-\alpha)}}{(1+x)^{(n+k+r,-\alpha)}k!}\;f\left(\frac{k}{n}\right)$$

So (5.2) can be written as follows

$$(5.3)\quad (P^{\alpha}_{n+1}-P^{\alpha}_{n})\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}x}{n(n+1)}\;P^{\alpha,1}_{n}\;T^{(2)}_{k}f\,(x)$$

Now, in order to study the monotonicity of the sequence $\{D^{m}_{\alpha}P^{\alpha}_{n}\,f(x)\}_{n}$, we prove

Theorem 5.1. The following relationship holds

$$D^{m}_{\alpha}[P^{\alpha}_{n+1}-P^{\alpha}_{n}]\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}}{n(n+1)}\;m!\;\Big\{(x+m\alpha)\cdot$$

$$(5.4)\quad \cdot\Big[P^{\alpha,\,m+1}_{n}\binom{n+k+m}{m}T^{(m+2)}_{k}f(x)+P^{\alpha,m+1}_{n}\binom{n+k+m}{m-1}T^{(m+1)}_{k+1}f(x)\Big]+$$

$$+\Big[P^{\alpha,m}_{n}\binom{n+k+m-1}{m-1}T^{(m+1)}_{k}f(x)+P^{\alpha,m}_{n}\binom{n+k+m-1}{m-2}T^{(m)}_{k+1}f(x)\Big]\Big\}$$

$\forall m\in N$ and $0\leqq\alpha<\frac{1}{2}$

:oof.

, using (2.1) in (5.3), with

$$f(x) = -\frac{1^{(n,-\alpha)}x}{n(n+1)} \quad \text{and} \quad g(x) = P_n^{\alpha}, {}_1T_k^{(2)} f(x)$$

e obtain

$$D_\alpha^m[P_{n+1}^\alpha - P_n^\alpha] f(x) = -\frac{1^{(n,-\alpha)}}{n(n+1)} \cdot$$

$$\cdot \sum_{i=0}^{1} \binom{m}{i} D_\alpha^i (x+(m-i)\alpha) D_\alpha^{m-i} P_n^\alpha, {}_1T_k^{(2)} f(x) =$$

$$= -\frac{1^{(n,-\alpha)}}{n(n+1)} \left[(x+m\alpha) D_\alpha^m P_n^\alpha, {}_1T_k^{(2)} f(x) + m D_\alpha^{m-1} P_n^\alpha, {}_1T_k^{(2)} f(x) \right]$$

n the other hand, we can easily verify by induction that

$$D_\alpha^m P_n^\alpha, {}_1T_k^{(2)} f(x) = m! \left[P_n^{\alpha,m+1} \binom{n+k+m}{m} T_k^{(m+2)} f(x) + P_n^{\alpha,m+1} \binom{n+k+m}{m-1} T_{k+1}^{(m+1)} f(x) \right]$$

nd, by taking this last relation into account in (5.5), the heorem follows.

rom Theorem 5.1, by (2.3), we have

orollary 5.2. The sequence

$$5.6) \quad \{ D_\alpha^m P_n^\alpha f(x) \}_n , \quad m \in \mathbb{N}, \quad 0 \leq \alpha < \frac{1}{2}$$

erifies the following monotonicity properties

) for m=1

if on the interval $[0,+\infty[$ the function f is convex (concave) of first and second order, then the sequence (5.6) is decreasing (increasing) on $[0,+\infty[$;

i) for $m \geq 2$

if on the interval $[0,+\infty[$ the function f is convex (concave) of order m-1, m and m+1, then the sequence (5.6) is decreasing (increasing) on $[0,+\infty[$.

We notice that for $\alpha=0$, P_n^α operator becomes equal to P_n operator defined by (2.9); so, Corollary 5.2 generalizes a re-

sult established in [4] for the sequence $\{(P_n f)^{(m)}(x)\}_n$.

REFERENCES

1. O. ARAMĂ: Proprietăţi privind monotonia şirului polinoamelor de interpolare ale lui S.N. Bernstein şi aplicarea lor la studiul aproximării funcţiilor. Stud. Cerc. Mat. (Cluj) 8 (1957), 195-210.
2. V.A. BASKAKOV: An instance of a sequence of linear positive operators in the space of continuous functions. Dokl. Akad. Nauk. SSSR 113 (1957), 249-251.
3. E.W. CHENEY and A. SHARMA: Bernstein power series. Canad. J. Math. XVI, 2 (1964), 241-252.
4. B. Della Vecchia: On the monotonicity of the derivatives of the sequences of Favard and Baskakov operators. Submitted to Ricerche di Matematica (1987).
5. J. FAVARD: Sur les multiplicateurs d'interpolation. J. Math. Pures Appl. 23 (1944), 219-247.
6. T. HERMANN: On Baskakov-type operators. Acta Math. Sci. Hungar. 31, (3-4) (1978), 307-316.
7. I. HOROVÁ: A note on the sequence formed by the first order derivatives of the Szász-Mirakyan operators. Mathematica 24 (47) (1982), 49-52.
8. I. HOROVÁ: Linear positive operators and their applications to differnetial equations. Arch. Math. (Brno) 20 (1984), 1-8.
9. A. LUPAŞ: On Bernstein power series. Mathematica 8 (31) (1966), 287-296.
10. A. LUPAŞ and M.W. MÜLLER: Approximation properties of the M_n-Operators. Aeq. Math. 5 (1970), 19-37.
11. G. MASTROIANNI: Su una classe di operatori lineari e positivi. Rend. Accad. Scien. M.F.N. Serie IV XLVIII (1980).
12. G. MASTROIANNI: Una generalizzazione dell'operatore di

Mirakyan. Rend. Accad. Sci. M.F.N. Serie IV XLVIII (1980).

3. G. MASTROIANNI and M.R. OCCORSIO: Sulle derivate dei polinomi di Stancu. Rend. Accad. Sci. M.F.N. Serie IV XLV (1978).

4. G. MASTROIANNI and M.R. OCCORSIO: Una generalizzazione dell'operatore di Stancu. Rend. Accad. Sci. M.F.N. Serie IV XLV (1978).

5. G. MIRAKYAN: Approximation des fonctions continues au moyen de polynomes de la forme ... Dokl. Akad. Nauk. SSSR 31 (1941), 201-205.

6. O. SZASZ: Generalization of Bernstein's polynomials to the infinite interval. J. Res. Nat. Bur. Standards 45 (1950), 239-245.

7. S.P. PETHE and G.C. JAIN: Approximation of functions by a Bernstein-type operator. Canad. Math. Bull. 15 (4) (1972), 551-557.

8. D.D. STANCU: On the monotonicity of the sequence formed by the first order derivatives of the Bernstein polynomials. Math. Z. 98 (1967), 46-51.

9. D.D. STANCU: Approximation of functions by a new class of linear polynomial operators. Rev. Roumanie Math. Pures Appl. 13 (1968), 1173-1194.

10. D.D. STANCU: Approximation properties of a class of linear positive operators. Studia Univ. Babes-Bolyai, Cluj, Ser. Mat. 2 (1970), 33-38.

11. D.D. STANCU: Two classes of positive linear operators. Analele Univ. Timişoara, Ser. Mat. 8 (1970), 213-220.

12. D.D. STANCU: On the remainder of approximation of functions by means of parameter-dependent linear polynomial operator. Studia Univ. Babes-Bolyai, Cluj 16 (1971), 59-66.

23. D.D. STANCU: Approximation of functions by means of some new classes of positive linear operators. In: Proc. Conf.

Math. Res. Inst. Oberwolfach (L. Collatz, G. Meinardus eds.), 1972.

24. D.D. STANCU: Application of divided differences to the study of monotonicity of the derivatives of the sequences of Bernstein polynomials. Calcolo 16 (1979), 431-445.
25. D.D. STANCU: A study of the remainder in an approximation formula using a Favard-Szasz type operator. Studia Univ. Babes-Bolyai, Mathematica XXV (1980), 70-76.
26. F. STANCU: Asupra restului în formulele de aproximare prin operatorii lui Mirakian de una şi două variabile. Analele Şt. Univ. "Al. I. Cuza", Iaşi 14 (1968), 415-422.
27. F. STANCU: Asupra aproximării funcţiilor de una şi două variabile cu ajutorul operatorilor lui Baskakov. St. Cerc. Mat. 22 (1970), 531-542.
28. W.B. TEMPLE: Stieltjes integral representation of convex functions. Duke Math. J. 21 (1954), 527-531.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

OPTIMAL PERIODIC INTERPOLATION IN THE MEAN

F.-J. DELVOS

ABSTRACT: *The concept of periodic Hilbert spaces was introduced by Babuska in connection with universally optimal quadrature formulas. It was shown by Prager, Locher, Knauff - Kress, and the author that periodic Hilbert spacesform an appropriate tool for constructing periodic interpolation splines and some of its extensions such as rational trigonometric interpolation. It was pointed out by Subbotin that it is natural to approximate functions from L^1 via interpolation in the mean splines. In this paper we will develop the method of optimal interpolation in the mean in periodic Hilbert spaces. Applications to periodic splines are presented.*

1. TRIGONOMETRIC INTERPOLATION IN THE MEAN

We denote by $\tau_{0,n-1}$ the n-dimensional space of trigonometric polynomials spanned by the functions

$$e_k(t) = \exp(ikt) \quad (0 \le k < n) .$$

Recall that $\tau_{0,n-1}$ is the appropriate space for discussing the discrete Fourier transform method. Assume that there are n real numbers $t_0,\ldots,t_{n-1}$ and a positive real number h satisfying

$$0 = t_0 < t_1 < t_2 < \ldots < t_{n-1} < 2\pi$$

and

$$0 < h \le 1/n , \quad h \le t_{j+1}-t_j \quad (0 \le j < n)$$

with $t_n = 2\pi$. The *interpolation in the mean functionals* are given by

$$L_{j,h}(f) = \frac{1}{h} \int_{t_j}^{t_j+h} f(t)\, dt \quad (0 \le j < n) .$$

Note that the interpolation functionals

$$L_j(f) = f(t_j) \quad (0 \le j < n)$$

are obtained formally by setting $h = 0$.

Proposition 1.1

For any $f \in L^1_{2\pi}$ there is a unique trigonometric polynomial $H_n(f) \in \tau_{0,n-1}$ satisfying the interpolation conditions

$$L_{j,h}(H_n(f)) = L_{j,h}(f) \quad (0 \le j < n) \ .$$

Proof. It is sufficient to show that

$$A = (a_{j,k}) := (L_{j,h}(e_k))_{0\le j,k<n} \ .$$

is a regular matrix. It is easily seen that

$$L_{j,h}(e_0) = 1 \quad (0 \le j < n) \ ,$$
$$L_{j,h}(e_k) = \exp(ikt_j)(-1+\exp(ikh))/(ikh) \quad (0\le j<n \ , \ 0<k<n) \ .$$

This implies $A = VD$ with

$$V = (\exp(ikt_j))_{0\le j,k<n}$$

and

$$D = \mathrm{diag}(d_0,\ldots,d_{n-1}) \ ,$$

$$d_0 = 1 \ , \ d_k = (\exp(ikh) - 1)/(ikh) \quad (0 < k < n) \ .$$

V is a Vandermonde matrix and D is a regular diagonal matrixin view of $\exp(ikh) \neq 1$. This completes the proof of Proposition 1.1

2. PERIODIC HILBERT SPACES

In this section we recall the properties of periodic Hilbert spaces as described in Prager [10] . Let

$$d_k \quad (k \in \mathbb{Z})$$

be a biinfinite sequence of real numbers from l^1 satisfying

(2.1) $$d_k = d_{-k} > 0 \quad (k \in \mathbb{Z}) \ , \ d_0 = 1 \ .$$

Then there exists a unique function ψ from the Wiener algebra $A_{2\pi}$ satisfying

$$d_k = (\psi, e_k) = \frac{1}{2\pi}\int_0^{2\pi} \psi(t)\exp(-ikt)\,dt \quad (k \in \mathbb{Z}) \ .$$

It is obvious that ψ is real-valued and even :

$$\psi(t) \quad = \quad 1 + 2 \sum_{k=1}^{\infty} d_k \cos(kt) \quad .$$

The periodic Hilbert space H_d related to $d = (d_k)$, respectively ψ, is defined by

$$H_d \quad = \quad \{ f \in L^2_{2\pi} : \sum_{k=-\infty}^{\infty} (f,e_k)(e_k,f)/d_k \; < \; \infty \} \quad .$$

The inner product of H_d is given by

$$(f,g)_d \quad = \quad \sum_{k=-\infty}^{\infty} (f,e_k)(e_k,g)/d_k \quad .$$

Obviously, H_d contains the algebra τ of trigonometric polynomials. Moreover, Prager showed that

(2.2) $\qquad H_d \quad \subseteq \quad A_{2\pi} \quad .$

It is easily seen that for $f \in H_d$ and $a \in \mathbb{R}$ we have

$$f(.-a) \in H_d \quad ,$$

i. e., H_d is closed with respect to translation. Moreover, we have

(2.3) $\qquad (f(.-a),g(.-a))_d \quad = \quad (f,g)_d$

for all functions $f,g \in H_d$.

Proposition 2.2

Let $f \in H_d$ and $x \in \mathbb{R}$. Then we have

$$f(x) \quad = \quad (f,\psi(.-x))_d \quad .$$

Thus, H_d is a reproducing kernel Hilbert space of periodic functions with kernel $K(y,x) = \psi(y-x)$.

The function $\psi(.-x)$ is the representer of the Dirac measure δ_x which is a bounded linear functional on H_d . Let L be a bounded linear functional on H_d and $u \in H_d$ be its representer. Then we have

(2.4) $\qquad L(f) \quad = \quad (f,u)_d$

for all functions $f \in H_d$. It was shown by Prager [10] that the Fourier series of u is given by the formula

$$(2.5) \qquad u(t) = \sum_{k=-\infty}^{\infty} d_k \overline{L(e_k)}\, e_k(t) \quad .$$

As an example we consider the construction of the periodic Sobolev space $W_{2\pi}^r$ with $r \in \mathbb{N}$. The defining sequence d is given by

$$d_0 = 1 \quad , \quad d_k = k^{-2r} \quad (k \neq 0)$$

and we have $H_d = W_{2\pi}^r$. The reproducing kernel of $W_{2\pi}^r$ satisfies the relation

$$(2.6) \qquad K_r(y,x) = 1 + (-1)^r B_{2r}(y-x) = \psi(y-x) \quad ,$$

$$B_q(x) = \sum_{k\neq 0} (ik)^{-q} e_k(x) \quad .$$

$B_q(x)$ is the periodic Bernoulli function of order q which is defined uniquely by the relations

$$(2.7) \qquad B_1(x) = \pi - x \quad , \quad B'_{q+1}(x) = B_q(x) \quad , \quad (B_{q+1}, e_0) = 0$$

$$(0 < x < 2\pi) \quad .$$

3. OPTIMAL INTERPOLATION IN THE MEAN

In this section we will study interpolation in the mean as a minimum norm interpolation problem in the reproducing kernel Hilbert space H_d [8].

Proposition 3.1

The linear functionals $L_{0,h}, \ldots\ldots, L_{n-1,h}$ are bounded and linearly independent over H_d.

Proof. It follows from Proposition 2.2 that the following estimate

$$(3.1) \qquad \|f\|_\infty \leq \sqrt{\psi(0)}\, \|f\|_d$$

holds for all functions $f \in H_d$. Thus, the interpolation in the mean functionals $L_{j,h}$, $0 \leq j < n$, are bounded. Since the trigonometric polynomials are contained in H_d it follows from Proposition 1.1 that $L_{0,h}, \ldots\ldots, L_{n-1,h}$ are linearly independent over H_d.

we will determine the representers u_j of $L_{j,h}$ for $j =$ $,\ldots,n-1$. For this construction we need the *periodic integral* Ψ of . Recall that Ψ is the unique function from $C^1_{2\pi}$ such that

3.2) $\Psi'(t) = \psi(t) - (\psi,e_0)$, $(\Psi,e_0) = 0$.

he Fourier series of Ψ is given by

3.3) $\Psi(t) = \sum_{k\neq 0} d_k (ik)^{-1} e_k(t) = \sum_{k=1}^{\infty} 2d_k \sin(kt)/k$.

roposition 3.2

The representer u_j of $L_{j,h}$ is given by the formula

3.4) $u_j(t) = 1 + (\Psi(t-t_j) - \Psi(t-t_j-h))/h$

ith $0 \leq j < n$.

roof. Recall that

$L_{j,h}(e_0) = 1$, $L_{j,h}(e_k) = e_k(t_j)(e_k(h)-1)/(ikh)$ ($k \neq 0$).

Jsing relation (2.5) and relation (3.3) we can conclude

$$u_j(t) = \overline{L_{j,h}(e_0)} + \sum_{k\neq 0} d_k \overline{L_{j,h}(e_k)}\, e_k(t)$$

$$= 1 - \sum_{k\neq 0} (d_k/(ikh))\, e_k(-t_j-h)\, e_k(t)$$

$$+ \sum_{k\neq 0} (d_k/(ikh))\, e_k(-t_j)\, e_k(t)$$

$$= 1 - (\Psi(t-t_j) - \Psi(t-t_j-h))/h \quad .$$

This completes the proof of Proposition 3.2 .

_et U_n be the linear span of the representers $u_0,\ldots,u_{n-1}$ and let S_n be the orthogonal projector in H_d with

$$\mathfrak{R}(S_n) = U_n = \langle u_0,\ldots,u_{n-1} \rangle \quad .$$

The following result is a consequence of the method of minimum norm interpolation in Hilbert spaces (Prager [10]).

Proposition 3.3

Let $f \in H_d$ be given. Then $S_n(f) \in U_n$ is the unique function in H_d having the characteristic properties

$$(i) \quad \frac{1}{h}\int_{t_j}^{t_j+h} S_n(f)(t)\,dt = \frac{1}{h}\int_{t_j}^{t_j+h} f(t)\,dt \quad (0 \le j < n) ,$$

$$(ii) \quad \|S_n(f)\|_d \le \|g\|_d \text{ if } L_{j,h}(g) = L_{j,h}(f) \quad (0 \le j < n) .$$

As an example we determine the *periodic interpolation in the mean splines* which are obtained by choosing $H_d = W^r_{2\pi}$ [2,6,8,11] . It follows from Proposition 3.2 and relations (2.6) and (2.7) that the representers $u_0,\dots,u_{n-1}$ are given by

$$(3.5) \quad u_j(t) = 1 + (-1)^r(B_{2r+1}(t-t_j) - B_{2r+1}(t-t_j-h))/h \quad (0 \le j < n) .$$

The properties of the Bernoulli functions imply that u_j is a periodic spline of degree $2r$ with spline knots $t_j+2\pi k$, $t_j+h+2\pi k$ $(k \in \mathbb{Z})$. As a consequence U_n is an n-dimensional space of periodic splines of degree $2r$ with splines knots

$$t_j + 2\pi k \ , \ t_j + h + 2\pi k \quad (0 \le j < n \ , \ k \in \mathbb{Z}) .$$

Let us consider the case of a uniform mesh, i. e. ,

$$t_j = 2\pi j/n \quad (j \in \mathbb{Z}) .$$

Then the space of interpolation in the mean splines is generated by translation from the *generating function*

$$(3.6) \quad u(t) = 1 + (B_{2r+1}(t) - B_{2r+1}(t-h))/h ,$$

i. e., we have

$$(3.7) \quad U_n = \langle u(.-t_0),\dots,u(.-t_{n-1})\rangle =: V_n(u) .$$

It should be noted that for the special case $h = t_1$ the space $V_n(u)$ is just the space of periodic splines of degree $2r$ with knots t_j. This follows from the fact that

$$1 = (u(t-t_0) + u(t-t_1) + \dots + u(t-t_{n-1}))/n .$$

Let $v \in U_n = V_n(u)$ be the unique function satisfying

3.8) $L_{j,h}(v) = \delta_{0,j} \quad (0 \le j < n)$.

Since U_n is translation invariant with respect to t_1 it follows from the relation

3.9) $$L_{j,h}(f) = \frac{1}{h} \int_0^h f(t+t_j)\, dt = L_{0,h}(f(.+t_j))$$
$(0 \le j < n)$

that the interpolation in the mean spline $S_n(f)$ is given by the formula

3.10) $$S_n(f)(t) = \sum_{j=0}^{n-1} L_{j,h}(f)\, v(t-t_j) \quad .$$

(*Interpolation in the mean by translation*).

4. THE CONSTRUCTION OF THE FUNDAMENTAL FUNCTION

For the case of a uniform mesh we will apply the method of the discrete Fourier transform to derive an explicit formula for *interpolation in the mean fundamental function* v . Recall that

(4.1) $$u(t) = 1 + (\Psi(t) - \Psi(t-h))/h \quad ,$$

$$u_k(t) = u(t-t_k) \quad (0 \le k < n) .$$

Furthermore, let ϕ be the periodic integral of Ψ , i. e.,

(4.2) $$\phi'(t) = \Psi(t) \quad , \quad (\phi, e_0) = 0 \quad .$$

Since $\Psi(-t) = -\Psi(t)$ it follows

(4.3) $$\phi(-t) = \phi(t) \quad .$$

Then we can conclude

$$L_{j,h}(u_k)$$

$$= 1 + h^{-2}\left(\int_{t_j}^{t_j+h} \Psi(t-t_k)\, dt - \int_{t_j}^{t_j+h} \Psi(t-t_k-h)\, dt \right)$$

$$= 1 + h^{-2} \left(\phi(t_{j-k}+h) + \phi(t_{j-k}-h) - 2\phi(t_{j-k}) \right) \quad ,$$

i. e., we have

$$(4.4)\quad L_{j,h}(u_k) = 1 + h^{-2}(\phi(t_{j-k}+h)+\phi(t_{j-k}-h)-2\phi(t_{j-k})) =: w(j-k)$$
$$(\ 0 \le j,k < n\ )\quad .$$

It follows from the definition of u_k that the matrix

$$(4.5)\quad T = (L_{j,h}(u_k))_{0\le j,k<n}$$

is a circulant Toeplitz matrix which is also positive definite. The interpolation in the mean fundamental function v is given by

$$(4.6)\quad v(t) = \sum_{k=0}^{n-1} c_k\, u(t-t_k)$$

with

$$(4.7)\quad \sum_{k=0}^{n-1} w(j-k)\, c_k = \delta_{0,k} \quad (\ 0 \le j < n\ )\quad .$$

Using discrete Fourier transform methods [3] we obtain the following explicit formulas

$$(4.8)\quad c_k = \frac{1}{n}\sum_{j=0}^{n-1} e_k(t_j)/a_j \quad (\ 0 \le k < n\ )\quad ,$$

$$a_j = \sum_{k=0}^{n-1} w(k)\, e_j(-t_k) \quad (\ 0 \le j < n\ )\quad .$$

For the practically important case $n = 2m$ these formulas reduce to

$$(4.9)\quad a_j = w(0) + w(m)\cos(\pi j) + 2\sum_{k=1}^{m-1} w(k)\cos(jt_k)\quad ,$$

$$c_k = \frac{1}{n}(a_0^{-1} + \cos(\pi k)a_m^{-1} + 2\sum_{j=1}^{m-1}\cos(kt_j)a_j^{-1}) \quad (\ 0 \le j,k < n\ )\ .$$

5. EXAMPLES

In this section we determine for two special choices of ψ the related functions Ψ and ϕ . The first example is concerned with the function

$$(5.1)\quad \psi(t) = 1 + (-1)^r B_{2r}(t)\quad ,\ r \in \mathbb{N}\ .$$

The related periodic Hilbert space H_d is the periodic Sobolev space

$w_{2\pi}$. Using the properties of the Bernoulli functions we obtain the following formulas:

$$(5.1') \quad \Psi(t) = (-1)^r B_{2r+1}(t) \quad , \quad \phi(t) = (-1)^r B_{2r+2}(t) \quad .$$

The space of optimal periodic interpolants in the mean $\mathfrak{K}(S_n)$ consists of periodic splines of even degree.

The second example is characterized by the function

$$(5.2) \quad \psi(t) = 1 - (\cosh(b) - 1)/(\cosh(b) - \cos(t))^2 \quad , \; b > 0.$$

In this case H_d is a space of periodic functions having holomorphic extensions in the strip $|\mathrm{Im}(z)| < b$:

$$H_d = \{ f \in L^2_{2\pi} : \sum_{k\neq 0} |(f,e_k)|^2 |k|^{-1} \exp(b|k|) < \infty \} \quad .$$

It is easily seen that the following relations hold :

$$(5.2') \quad \Psi(t) = -\sin(t)/(\cosh(b) - \cos(t)) \quad ,$$

$$\phi(t) = -\ln(\cosh(b) - \cos(t)) \quad .$$

In this case $\mathfrak{K}(S_n)$ is a linear space of rational trigonometric functions.

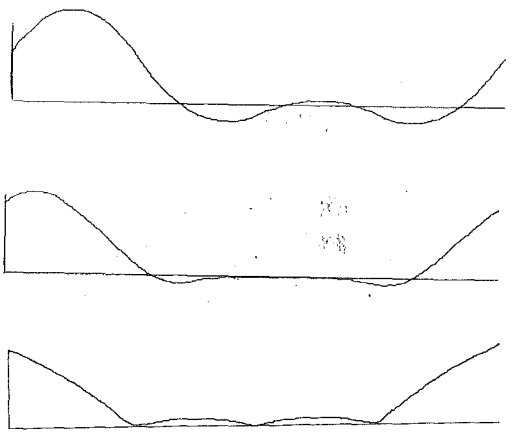


Fig. 1 Fundamental interpolation in the mean function v for $\psi(t) = 1 - B_2(t)$ with $n = 4$ and $h = t_1, t_1/2, t_1/128$

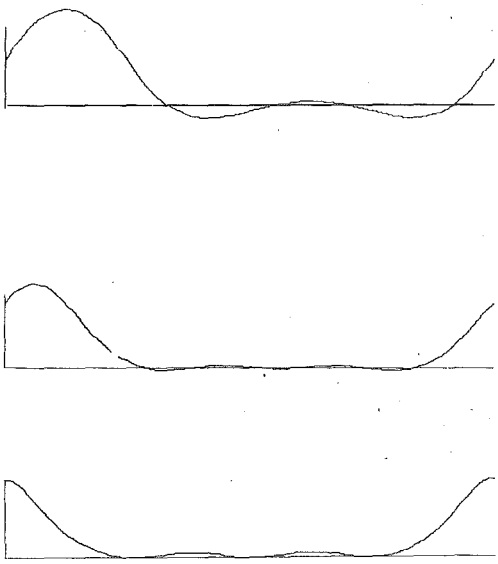


Fig. 2 Fundamental interpolation in the mean function v for $\psi(t) = 1-(\cosh(1)-1)(\cosh(1)-\cos(t))^{-2}$, $n = 4$ and $h = t_1,\ t_1/2,\ t_1/128$.

References

1. I. BABUSKA : Ueber universal optimale Quadraturformeln, Teil 1, Apl. mat. 13(1968), 304-338, Teil 2, Apl. mat. 13(1968), 388-404 .

2. G. BASZENSKI and L.L. SCHUMAKER : On a method for fitting an unknown function based on mean value measurements . SIAM J. Numer. Anal. 24(1987), 725-736 .

3. D.R. BRILLINGER: "Time series: data analysis and theory". Holt, Rinehardt and Winston, Inc., New York, 1975.

4. F.-J. DELVOS: Periodic interpolation on uniform meshes . J. Approx. Theory 49(1987) (to appear) .

5. F.-J. DELVOS: Convergence on interpolation by translation. Colloquia Mathematica János Bolyai 49, In: Alfred Haar Memorial Conference , Budapest (Hungary), 1985, pp. 273-287.

6. F.-J. DELVOS: Periodic area matching interpolation.
In: Numerical Methods of Approximation Theory (L. Collatz, G. Meinardus, G. Nürnberger , eds.), ISNM 81, Birkhäuser Verlag, Basel 1987, pp. 54-66.

7. W.KNAUFF and R. KRESS: Optimale Approximation linearer Funktionale auf periodischen Funktionen. Numer. Math. 22(1974), 187-205.

8. P.-J. LAURENT: Approximation et optimisation. Herrmann, Paris, 1972.

9. F. LOCHER: Interpolation on uniform meshes by translates of one function and related attentuation factors. Mathematics of Computation 37(1981), 403-416 .

10. M. PRAGER: Universally optimal approximation of functionals. Apl. mat. 24(1979), 406-420 .

11. Yu. SUBBOTIN: Extremal problems of functional interpolation and splines for interpolation in the mean. Proc Steklov Inst. Math. 138(1975), 127-185.

Numerical Methods and Approximation Theory III (Niš, August 18-21, 1987)

ACCURATE EXPLICIT FINITE DIFFERENCE SOLUTION OF THE SHOCK TUBE PROBLEM

S.K. DEY and CHARLIE DEY

Abstract

A simple predictor-corrector algorithm has been developed in [1] for numerical solution of initial-value problems. In this article we will discuss computer experimentation of this method for solution of the shock tube problem.

Introduction

One dimensional motion of compressible flow is described by:

$$U_t + F_x = 0 \tag{1}$$

where

$$U = (\rho,\ \rho u,\ e)^T$$

$$F = (a,\ b,\ c)^T$$

$$a = \rho u, \quad b = (\gamma-1)e + ((3-\gamma)/2)\ \rho u^2$$

$$c = \gamma e u - ((\gamma-1)2)\ \rho u^3. \tag{2}$$

ρ = density, u = velocity, p = pressure, e = total energy per unit volume, given by $e = \rho\varepsilon + u^2/2$, ε = internal energy per unit mass. For a perfect gas pressure p is defined by

$$p = (\gamma-1)(e - \rho u^2/2), \quad \gamma = 1.4$$

subject to various initial/boundary conditions, this set of equations will describe various compressible flow models.

Now let us consider the shock tube problem. Let two gases separated by a diaphragm be in equilibrium in a tube. Let the densities of them be unequal. If the diaphragm is suddenly broken, the gas molecules start mixing. This mixing phenomenon is often referred to as the shock tube problem. Here we will use the following initial conditions:

At $t = 0$, $\rho = 1$, $u = 0$, $e = 1/(\gamma-1)$ for $0 \leq x \leq 1.9$ and $\rho = 0.1$, $u = 0$, $e = 0.1/(\gamma-1)$ for $1.9 < x \leq 5$.

As the gas molecules start mixing, sharp changes of density, pressure, velocity and energy take place at several points along the x-axis. To describe this phenomenon appropriately by a numerical algorithm is often a challenge for researchers in computational fluid dynamics.

This challenge has been undertaken by many researchers in the past [2, 3, 4], and in some cases excellent results have been found. In this work an explicit finite difference scheme, whose algorithm is much simpler than all the above methods, has been successfully applied to obtain quite accurate numerical solutions of the shock tube problem. The algorithm has been developed by the second author and applied extensively by him to solve several linear and nonlinear models in Engineering and Applied Mathematics [1]. Let us briefly describe the algorithm and some of its properties.

The Algorithm

Let us consider an initial-value problem

$$du/dt = f(u, t), \quad u(t_o) = u_o \tag{3}$$

A predictor-corrector algorithm to solve (3) may be expressed as:

$$\hat{U} = U_n + \Delta t\, f(U_n, t_n) \qquad \text{predictor} \tag{4a}$$

$$U_{n+1} = (1-\gamma)\,\hat{U} + \gamma\,\{U_n + \Delta t\, f(\hat{U}, t_{n+1})\} \qquad \text{corrector} \tag{4b}$$

where $U_n = U(t_n)$, Δt = step size, $\hat{U}$ is the predicted value of U at t_{n+1}, U_{n+1}, is the corrected value of U at t_{n+1}.

γ is called a filtering parameter, and it is assumed that $0 < \gamma < 1$. The predictor is Euler's forward difference approximation. If $\gamma = 0.5$, then (4a) and (4b) are reduced to a second-order Runge-Kutta scheme. It is expected that the corrector should filter most errors generated by the predictor. But one must choose a value of γ before the algorithm may be used. Such a choice for the value of γ may be obtained if we do the stability analysis of this numerical method [1].

Linearized Stability Analysis

If we linearize (3), write $du/dt = \lambda u$ and use (4a) and (4b), we get the combined form of (4a) and (4b) as,

$$U_{n+1} = \sigma U_n, \text{ where } \sigma = 1+z+\gamma z^2, \; z = \lambda \Delta t.$$

For stability, $|\sigma| \leq 1$. Let us consider the following example to look into an interesting property of this method. $du/dt = -80u$, $u(0) = 1$. The analytical solution is $u(t) = e^{-80t}$. Here $\sigma = 1-80h + 6400\gamma h^2$. With $\gamma = 0.1$, if $h = 0.01$, $|\sigma| < 1$ (stable), if $h = 0.07$, $|\sigma| > 1$ (unstable) and if $h = 0.1$, $|\sigma| < 1$ (stable). When the algorithm is stable, it gives quite accurate steady-state solutions. But often time accurate solutions given by this scheme are not up to expectations. This is true for one equation or a system of equations.

Since λ may be complex, z may be taken to be a complex variable, and hence $\sigma(z)$ is a complex function. Lomax [5] developed a computer code such that for a given γ, $\sigma(z)$ may be plotted in a complex plane and the region of stability may be found. Some of these contours have been described in [1].

Difference Approximation of (1)

The equation (1) may be approximated as follows:

$$U_j^{n+1} = U_j^n + \Delta t \, (F_{j-1}^n - F_{j+1}^n)/(2\Delta x) \qquad (5)$$

or

$$U_j^{n+1} = U_j^n + \Delta t\ (F_{j-1}^n - F_j^n)/\Delta x \qquad (6)$$

where $U_j^n = U(x_j, t_n)$. To stabilize the numerical process, an artificial viscosity term was introduced. This second-order derivative was approximated by central differences. Using (5), the predictor-corrector algorithm is:

$$\hat{U}_j = U_j^n + (\Delta t/2\Delta x)(F_{j-1}^n - F_{j+1}^n)$$

$$U_j^{n+1} = (1-\gamma)\ \hat{U}_j + \gamma\cdot\{U_j^n + (\Delta t/2\Delta x)(\hat{F}_{j-1}^{n+1} - \hat{F}_{j+1})\} \qquad (7)$$

Since there are three components of U, for each component a unique value of γ may be chosen.

Discussions

Figures 1, 2, and 3 which describe distributions of density, pressure and velocity were obtained by using the predictor-corrector with (5) as predictor. Figures 4, 5, and 6 describe distributions of the same and were obtained using the same algorithm with (6) as predictor. If the filtering parameters are not selected properly the results may not be reasonably correct. This often poses a problem, since (1) is a nonlinear model. For nonlinear Burgers' equation the model was linearized and γ was computed using the contours of stability [1]. This has not yet been done for the Euler's equation (1). We hope that such a stability analysis may resolve the problem in the future.

Acknowledgement

This research was conducted at the University of Siedlce, Poland, where both authors received research fellowships to undertake this work.

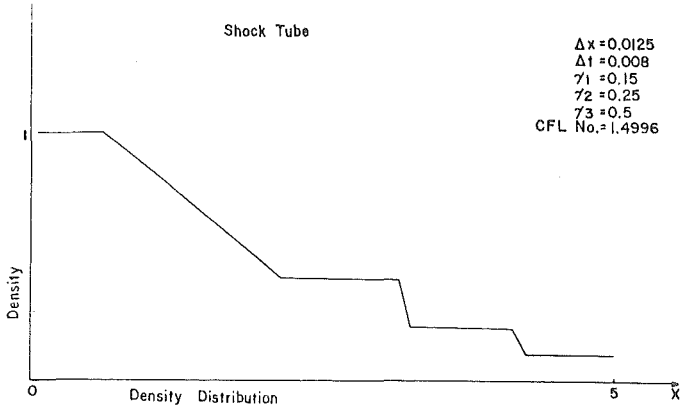


FIGURE 1

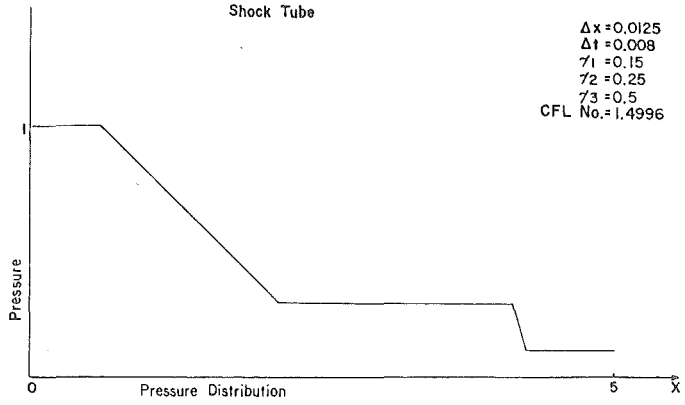


FIGURE 2

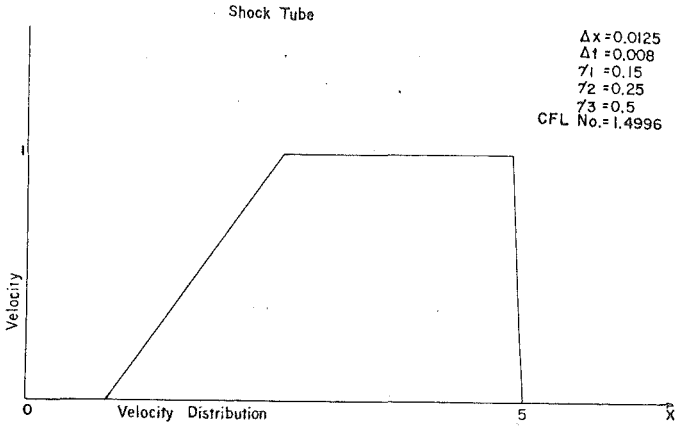


FIGURE 3

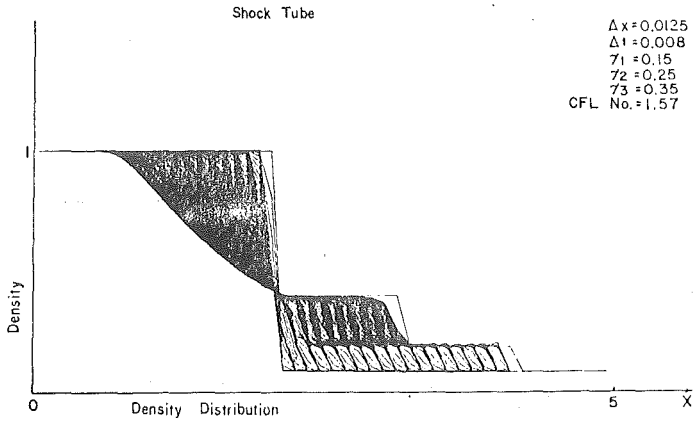


FIGURE 4

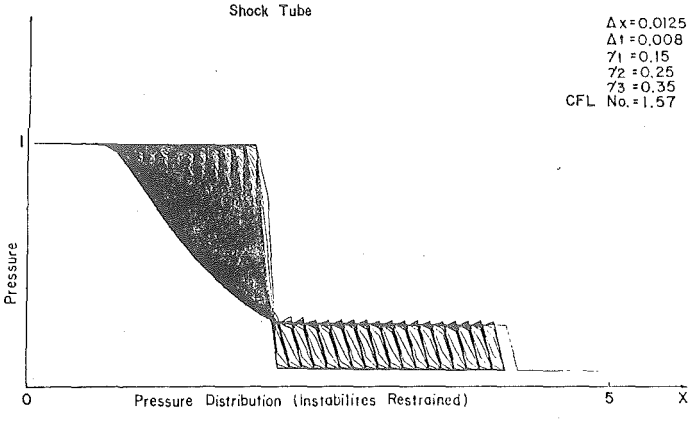


FIGURE 5

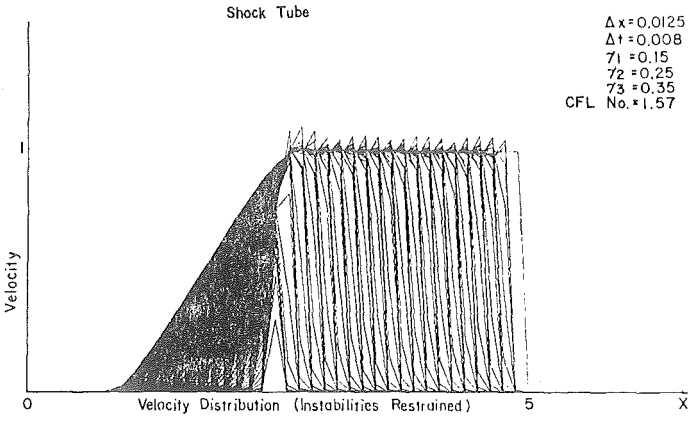


FIGURE 6

References

1. S. K. Dey and Charlie Dey, "An Explicit Predictor-Corrector Solver with Applications to Burgers' Equation." NASA Technical Memo 84402, September 1983. Ames Research Center, Moffett Field, CA 94035.

2. P. L. Roe, "Approximate Riemann Solvers, Parameter Vectors and Difference Schemes." J. Comp. Physics, Vol. 43, 1981.

3. E. J. Kansa, "Highly Accurate Shock Flow Calculations with Moving Grids and Mesh Refinement." Pro. Int. Congress on Sci. Comp. IMAC, Oslo, Norway, 1985.

4. S. K. Dey, "Numerical Solution of Euler's Equation by Perturbed Functionals." Lectures in Applied Math, AMS, Vol. 22, 1985.

5. H. Lomax, NASA Ames Research Center, Private Communication.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

SOME ASPECTS OF AUTOMATIC DIFFERENTIATION

HERBERT FISCHER

Abstract: Gradient and Hessian matrix of an explicitely given function can be computed automatically and straightforward by way of "automatic differentiation". This method is applicable to a broad class of functions. No quotients of differences are used. And no symbolic manipulation of symbols is involved. Complexity considerations show that "automatic differentiation" is competitive and efficient.

1. INTRODUCTION

Gradient and Hessian matrix of a real function of several variables play an important role in many numerical methods, especially in Nonlinear Optimization. But little effort has been devoted to the computation of these entities so far. "The Hessian matrix is not available." This statement used to be an axiom in the optimization folklore for decades. It led to the construction of well-known algorithms for nonlinear optimization problems, where the Hessian matrix respectively its inverse is approximated. Nevertheless, we will show how to obtain gradient and Hessian matrix "automatically" in an easy and straightforward manner. No manipulation of symbols is involved, we deal with numbers, not with formulas. This complies with the fact that, within the implementation of a relevant numerical method, gradient and Hessian matrix themselves are of interest rather than formulas for them.

Let us revisit the Hessian situation. Assume f is a twice differentiable function of several variables and $\bar{x}$ is a point in the domain of f. Assume further, we need the Hessian matrix $H(\bar{x})$ of f at $\bar{x}$. There are various approaches to compute $H(\bar{x})$, for instance

)

(1) derive a formula for H(x) and evaluate this formula for the specific $\bar{x}$,

(2) approximate $H(\bar{x})$ by a matrix of quotients of differences, using the gradient of f or an approximation thereof,

(3) "update" somehow a previously obtained approximation to $H(\bar{y})$, where $\bar{y}$ is "near" $\bar{x}$,

(4) use Automatic Differentiation.

The approach (1) is cumbersome, time consuming and prone to error, even if an outside computer-program for manipulation of symbols is used.
The numerical differentiation mentioned in (2) inevitably leads into the well-known predicament: a large stepsize yields inaccurate values and a small stepsize makes the computational process instable.
The way (3) may be considered an emergency measure.
The approach (4) seems to be the easiest one. It is astonishing that for a long time the idea of Automatic Differentiation has been overseen or ignored, despite quite a number of publications in this direction. We should mention that most of the protagonists' papers were too intermingled with programming languages or had to establish their own differentiation programming system. This may have hindered general recognition and delayed use. The breakthrough in Automatic Differentiation came with the work of L.B. Rall.
The automatic generation of gradient and Hessian matrix for a broad class of functions $\mathbb{R}^n \to \mathbb{R}$ may well restrict the above "axiom" to very expensive functions and to functions which are defined implicitely.

2. THE IDEA

In this section we sketch the basic idea of Automatic Differentiation as far as gradient and Hessian matrix are concerned.
Assume we have a rational function

$$r: D \subseteq \mathbb{R}^n \to \mathbb{R}$$

in explicit form. This means, for r(x) we have a formula which only contains the components $x_1, x_2, \ldots, x_n$ of x, some real numbers, the four arithmetic operations addition, subtraction, multiplication and division, and parentheses at proper places.

Let $r_G(x)$ denote the gradient of r at $x \in D$ and

let $r_H(x)$ denote the Hessian matrix of r at $x \in D$.

case: r is primitive

$r(x) = x_i$ = i-th component of x, for some $i \in \{1,2,\ldots,n\}$.

For any $x \in D$ we have $r_G(x)$ = i-th unit-vector, $r_H(x)$ = zero-matrix.

case: r is constant

$r(x) = \text{const} = c$, for some $c \in \mathbb{R}$.

For any $x \in D$ we have $r_G(x)$ = zero-vector , $r_H(x)$ = zero-matrix.

case: r is neither primitive nor constant

We employ Cesar's rule *divide et impera*, which of course in our situation reads

split and differentiate!

In splitting the formula for r(x), we obtain one of the four cases

(A) $r(x) = a(x) + b(x)$

(S) $r(x) = a(x) - b(x)$

(M) $r(x) = a(x) \cdot b(x)$

(D) $r(x) = a(x) / b(x)$

where a and b are rational functions $D \subseteq \mathbb{R}^n \to \mathbb{R}$. Furthermore, the functions a and b are available in explicit form and the formulas for a(x) and b(x) are shorter than the formula for r(x).

To follow the rule, we differentiate the function r. This yields

(A') $r_G = a_G + b_G$

(S') $r_G = a_G - b_G$

(M') $r_G = b \cdot a_G + a \cdot b_G$

(D') $r_G = (a_G - r \cdot b_G) / b$

where a_G = gradient of the function a and b_G = gradient of the function b.

From the formulas A, S, M, D and A', S', M', D' we conclude

> For any $x \in D$, the pair $r(x), r_G(x)$ can be computed from the pairs $a(x), a_G(x)$ and $b(x), b_G(x)$.

Notice that for a given $x \in D$, the pair $r(x), r_G(x)$ is not a pair of formulas, nor is it a pair of functions, it is an element of $\mathbb{R} \times \mathbb{R}^n$.

We differentiate the function r_G: $D \subseteq \mathbb{R}^n \to \mathbb{R}^n$. This yields

(A") $r_H = a_H + b_H$

(S") $r_H = a_H - b_H$

(M") $r_H = b \cdot a_H + b_G \cdot a_G^t + a \cdot b_H + a_G \cdot b_G^t$

(D") $r_H = (a_H - b_G \cdot r_G^t - r \cdot b_H - r_G \cdot b_G^t) / b$

where a_H = Hessian of the function a and b_H = Hessian of the function b.

Now we already know what to conclude

> For any $x \in D$, the triple $r(x), r_G(x), r_H(x)$ can be computed from the triples $a(x), a_G(x), a_H(x)$ and $b(x), b_G(x), b_H(x)$.

3. EXAMPLE

We consider the rational function

$$f: D \subseteq \mathbb{R}^3 \to \mathbb{R} \quad \text{with} \quad f(x) = x_1 - x_2 \cdot x_3 + x_1/(x_2 \cdot x_3 \cdot x_3).$$

First we split the formula for $f(x)$, then we split the parts, and so on. We obtain the tree shown in figure 1. Now we identify equivalent parts of the tree and get the graph shown in figure 2. This graph is a guide-line to compute $f(x)$.

code-list for $f(x)$

$f_1(x) = x_1$ = given
$f_2(x) = x_2$ = given
$f_3(x) = x_3$ = given
$f_4(x) = f_2(x) \cdot f_3(x)$
$f_5(x) = f_4(x) \cdot f_3(x)$
$f_6(x) = f_1(x) - f_4(x)$
$f_7(x) = f_1(x) / f_5(x)$
$f(x) = f_6(x) + f_7(x)$

For convenience we define

$$\overline{f}_i(x) = (f_i(x), f_{iG}(x), f_{iH}(x)) \quad \text{for } i = 1,2,\ldots,7$$
$$\overline{f}(x) = (f(x), f_G(x), f_H(x)) .$$

Now we know from section 2, that we can compute

$\overline{f}_4(x)$ from $\overline{f}_2(x)$ and $\overline{f}_3(x)$
$\overline{f}_5(x)$ from $\overline{f}_4(x)$ and $\overline{f}_3(x)$
$\overline{f}_6(x)$ from $\overline{f}_1(x)$ and $\overline{f}_4(x)$
$\overline{f}_7(x)$ from $\overline{f}_1(x)$ and $\overline{f}_5(x)$
$\overline{f}(x)$ from $\overline{f}_6(x)$ and $\overline{f}_7(x)$

This information allows to draw a graph to compute $\overline{f}(x)$, see figure 3. Notice that there is little difference between the graph to compute $f(x)$ and the graph to compute $\overline{f}(x)$.

The computational activities to get the value $f(x)$, the gradient $f_G(x)$ and the Hessian $f_H(x)$ for some $x \in D$ are obvious:

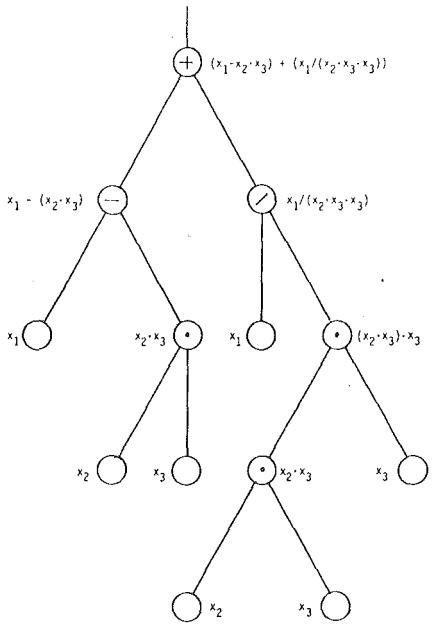


Figure 1: Tree for $f(x) = x_1 - x_2 \cdot x_3 + x_1/(x_2 \cdot x_3 \cdot x_3)$

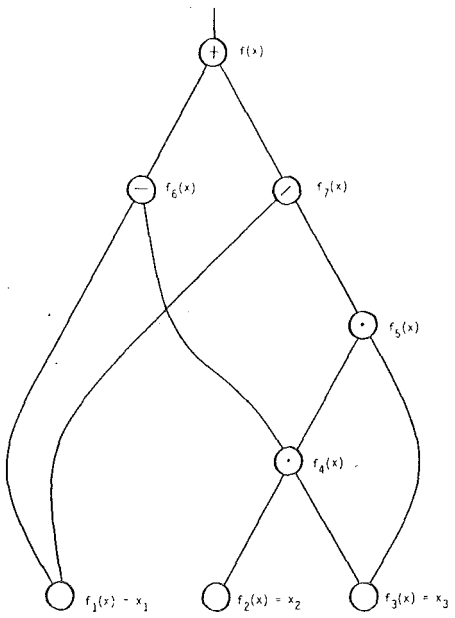


Figure 2: Graph to compute $f(x)$

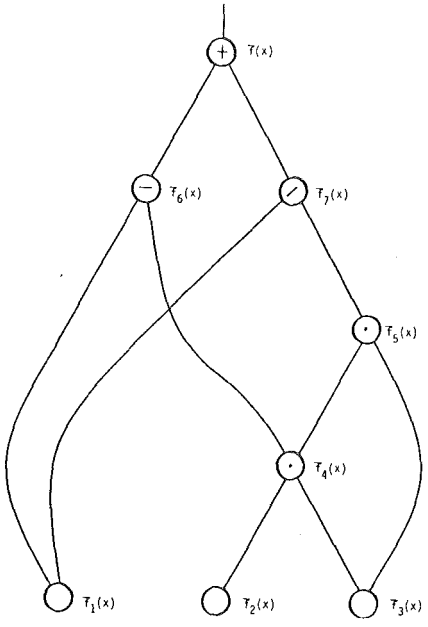


Figure 3: Graph to compute $\bar{f}(x) = (f(x), f_6(x), f_4(x))$

Set $\overline{f}_1(x)$, $\overline{f}_2(x)$, $\overline{f}_3(x)$ to their known values.

Then compute $\overline{f}_4(x)$, $\overline{f}_5(x)$, $\overline{f}_6(x)$, $\overline{f}_7(x)$, $\overline{f}(x)$ in this order.

The final triple $\overline{f}(x)$ provides the desired entities.

4. COMPLEXITY

Assume that $f: D \subseteq \mathbb{R}^n \to \mathbb{R}$ is an explicitely given rational function. This means, we can evaluate $f(x)$ using only addition, subtraction, multiplication and division of real numbers. Hence, we can set up a *characterizing sequence* $f_1, f_2, \ldots, f_s$ of functions $f_i: D \to \mathbb{R}$ such that

(1) for $i \in \{1,2,\ldots,n\}$

$f_i(x) = x_i$ = i-th component of x

(2) for $i \in \{n+1, n+2, \ldots, n+d\}$ with some $d \in \{0,1,\ldots\}$

$f_i(x) = c_i$ = real constant

(3) for $i \in \{n+d+1, n+d+2, \ldots, s\}$

$f_i(x) = a_i(x) * b_i(x)$ with $* \in \{+,-,\cdot,/\}$, and $a_i, b_i \in \{f_1, f_2, \ldots, f_{i-1}\}$

(4) $f_s(x) = f(x)$

In order to avoid superfluous operations, we assume that for $i \in \{n+d+1, \ldots, s-1\}$ the function f_i shows up as operand in at least one of the subsequent functions $f_{i+1}, \ldots, f_s$.

Prior to complexity investigations we have to specify the methods considered. Let us indicate gradient and Hessian matrix of a function by the subscript G resp. H.

Gradient-Method

(0) choose an $x \in D$

(1) for $i = 1, \ldots, n$ set $f_i(x), f_{iG}(x)$ to their values

(2) for $i = n+1, \ldots, n+d$ set $f_i(x), f_{iG}(x)$ to their values

(3) for $i = n+d+1, \ldots, s$ compute $f_i(x), f_{iG}(x)$ according to section 2

(4) then $f_s(x) = f(x)$, $f_{sG}(x) = f_G(x)$

Hessian-Method

(0) choose an $x \in D$

(1) for $i = 1,\ldots,n$ set $f_i(x), f_{iG}(x), f_{iH}(x)$ to their values

(2) for $i = n+1,\ldots,n+d$ set $f_i(x), f_{iG}(x), f_{iH}(x)$ to their values

(3) for $i = n+d+1,\ldots,s$ compute $f_i(x), f_{iG}(x), f_{iH}(x)$
according to section 2

(4) then $f_s(x) = f(x)$, $f_{sG}(x) = f_G(x)$, $f_{sH}(x) = f_H(x)$

What does it cost to compute the gradient $f_G(x)$ and the Hessian $f_H(x)$ of f at some $x \in D$? We answer this question in terms of arithmetic operations. Let us define

$\#(f)$:= number of arithmetic operations to compute $f(x)$ using the characterizing sequence $f_1, f_2, \ldots, f_s$

$\#(f, f_G)$:= number of arithmetic operations to compute $f(x)$ and $f_G(x)$

$\#(f, f_G, f_H)$:= number of arithmetic operations to compute $f(x)$, $f_G(x)$ and $f_H(x)$

Of course, $\#(f, f_G)$ and $\#(f, f_G, f_H)$ depend on the method used.

For the Gradient-Method we obtain

$$\#(f, f_G) \leq (3n + 1)\cdot\#(f) ,$$

and for the Hessian-Method we get

$$\#(f, f_G, f_H) \leq (\frac{7}{2}n^2 + \frac{13}{2}n + 1)\cdot\#(f) .$$

These bounds show that Automatic Differentiation is competitive if compared with numerical methods which approximate components of gradient and Hessian matrix by quotients of differences. Furthermore, it should be mentioned that, by some sophisticated organization, we are able to establish Automatic Differentiation methods with

$$\#(f, f_G) \leq 4\cdot\#(f) \quad \text{and} \quad \#(f, f_G, f_H) \leq (12n+8)\cdot\#(f) .$$

5. REMARKS

a) In section 2 we used a rational function r to point out the basic idea of Automatic Differentiation as far as gradient and Hessian matrix are concerned. But the formulas A', S', M', D' and A", S", M", D" mentioned there are not restricted to rational functions. The crucial point is the *rational composition* of r, rather than the rational character of the parts of r.

b) In section 4 we assumed that f is a *rational* function. This restriction was set only for didactic reasons. In the more general case where the formula for f(x) also involves some library functions like sin, cos, ..., the key is: Assume that the functions

$$a\colon D \subseteq \mathbb{R}^n \to \mathbb{R} \quad \text{and} \quad \lambda\colon E \subseteq \mathbb{R} \to \mathbb{R}$$

are twice differentiable. Under the provision $a(D) \subseteq E$ define the function

$$r\colon D \subseteq \mathbb{R}^n \to \mathbb{R} \quad \text{with} \quad r(x) := \lambda(a(x)) \ .$$

Then r is twice differentiable and

$$r_G(x) = \lambda'(a(x)) \cdot a_G(x)$$
$$r_H(x) = \lambda''(a(x)) \cdot a_G(x) \cdot a_G^t(x) + \lambda'(a(x)) \cdot a_H(x)$$

These formulas reveal the following fact:

> For any $x \in D$, the triple $r(x), r_G(x), r_H(x)$ can be computed from the triple $a(x), a_G(x), a_H(x)$ using λ, λ', λ''.

c) An efficient implementation of the Gradient- and the Hessian-Method has to take care of system-zeros (zeros in gradient and Hessian of primitive and constant functions).

REFERENCES

1. W. BAUR and V. STRASSEN: The complexity of partial derivatives. Theoretical Computer Science 22 (1983), 317-330.

2. H. KAGIWADA et al.: Numerical Derivatives and Nonlinear Analysis. Plenum Press, New York and London 1986.

3. L.B. RALL: Automatic Differentiation: Techniques and Applications. Lecture Notes in Computer Science No. 120, Springer-Verlag, Berlin-Heidelberg-New York 1981.

4. L.B. RALL: Differentiation in Pascal-SC: Type GRADIENT. ACM Trans. Math. Software 10 (1984), 161-184.

5. L.B. RALL: Global Optimization Using Automatic Differentiation And Interval Iteration. MRC Technical Summary Report #2832, University of Wisconsin, Madison 1985.

Numerical Methods and Approximation Theory III (Niš, August 18-21, 1987)

ON TWO SIDED APPROXIMATION FOR SOME SECOND ORDER VALUE BOUNDARY PROBLEMS*

P. GHELARDONI, G. GHERI and P. MARZULLI

ABSTRACT - This paper is concerned with boundary value problems for second order systems of the form $y'' = f(x,y)$. From a theorem proving under suitable conditions the existence of a solution, using Picard's iterations, a numerical procedure is derived to find actually two sided approximations of the solution. To this purpose a class of linear two-step methods is shown to be efficient, when two formulas of the class, with error constants of opposite sign, are alternatively used.
As a numerical application three test problems are developped.

1. INTRODUCTION.

Let the two point boundary value problem

(1) $y'' = f(x,y)$, $y(0) = \alpha$, $y(1) = \beta$,

be given with $f, y, \alpha, \beta \in R^m$, $x \in [0,1]$ of R^1 and let a function $y_1(x)$ be chosen so that $y_1(0) = \alpha$, $y_1(1) = \beta$; it is well known that under suitable hypotheses, Picard's iterations defined by

$$y_n'' = f(x,y_{n-1}), \quad y_n(0) = \alpha, \quad y_n(1) = \beta, \quad n>1,$$

can generate monotone sequences converging to a solution of (1). In a different way monotone sequences bounding the solution of

(1) Work supported by the Italian Ministero della Pubblica Istruzione.

(1) can be obtained by quasi-linearization ([8], ch. 5). Both those methods require to solve a linear problem at each step, bu the numerical solutions of these linear problems can bound the solution of (1) in the same fashion as the theoretical solution only if certain additional assumptions are verified ([2], p. p. 98-100 and corresponding references). Abiding by this frame, this paper we present a numerical method based on Picard's iter which give a theoretical two sided approximation to the solutio (1). The numerical method is suitable to solve the linear probl arising at each step, under assumptions sufficient to preserve property of a two sided approximation.

2. A PICARD'S SOLUTION OF THE PROBLEM.

Among several theorems assuring the existence of a solutio the problem (1), we are interested in the following formulation in [3], concerned with a more general problem than (1). Such formulation, which is given for a scalar equation, is a particu case of that reported in [1] (theor. 3.3., p. 34).

THEOREM 1. Given the boundary value problem

(1') $y'' = f(x,y,y')$, $y(0) = \alpha$, $y(1) = \beta$,

f, y, α, $\beta \in R$, let $f(x,y,y')$ be continous on

$$\{(x,y,y') \mid x \in [0,1],\ |y|<\infty,\ |y'|<\infty\}$$

and satisfy the Lipschitz conditions

$$|f(x,y,y')-f(x,y^*,y')| \leq L_1|y-y^*|,$$

$$|f(x,y,y')-f(x,y,y'^*)| \leq L_2|y'-y'^*|.$$

Then if

$$L_1 + 4L_2 < 8$$

there exists at least one solution of the problem (1').

The proof is given by constructing the sequence $\{F_n(x)\}$, $x \in [0,1]$,

$$F_1(x) = (\beta-\alpha)x + \alpha,$$
$$F_n''(x) = f(x,F_{n-1}(x),F_{n-1}'(x)), \quad n>1,$$
$$F_n(0) = \alpha, \quad F_n(1) = \beta;$$

this sequence is shown to converge uniformly on $[0,1]$ to a solution $y(x)$ of the problem (1').

This theorem can be extended with some slight modification to the problem (1).

We denote, for simplicity, $y_B(x) = (\beta-\alpha)x + \alpha$ the linear function satisfying the boundary conditions and

$$S = \{z(x) \mid z(x) \in C^0[0,1],\ z(0) = \alpha,\ z(1) = \beta\},$$

where the vector norm $\| \cdot \|$ is given by

$$\| z(x) \| = \sum_{i=1}^{m} \max_{0\leq x\leq 1} |z_i(x)|.$$

Then the following result holds.

THEOREM 2. Consider the problem (1) where $y(x) \in S$. Assuming that

(2) $\quad f_i(x,y) \in C^0([0,1]\times S),\ i = 1,2,\ldots,m,$

and, for any $y, y^* \in S$,

(3) $\quad |f_i(x,y)-f_i(x,y^*)| \leq L_i \sum_{j=1}^{m} |y_j-y_j^*|,$

(4) $\quad L_i < 8/m,\ i = 1,2,\ldots,m,$

then the Picard's iterations of S

$$
\begin{aligned}
&F_1(x) = y_B(x),\\
(5)\qquad &F_n''(x) = f(x,F_{n-1}(x)),\ n>1,\\
&F_n(0) = \alpha,\ F_n(1) = \beta,
\end{aligned}
$$

converge uniformly on $[\,0,1\,]$ to a solution of (1).

PROOF. The main feature of the proof is to prove that the sum $F_1(x) + (F_2(x)-F_1(x)) + \ldots + (F_n(x)-F_{n-1}(x))$ for $n\to\infty$ converges uniformly to a function which is shown to be a solution of (1). Since $F_n(x)$ is a solution of (5) we can write ([2] p. 42-43, [7])

$$
(5')\qquad F_n(x) = y_B(x) + \int_0^1 g(x,t)f(t,F_{n-1}(t))\,dt\ ,
$$

where

$$
g(x,t) = \begin{cases} t(x-1) & \text{for } t\le x\\ x(t-1) & \text{for } t>x \end{cases}
$$

is the Green's function. Thus we have

$$
|F_{2i}(x)-F_{1i}(x)| \le \int_0^1 |g(x,t)|\,|f_i(t,y_B(t))|\,dt\ ,\ i = 1,2,\ldots
$$

Now we observe that, for $i= 1,2,\ldots,m$, it results $\alpha_i\le y_{Bi}(x)\le\beta_i$ or $\alpha_i\ge y_{Bi}(x)\ge\beta_i$ on $[\,0,1\,]$, so that the functions $f_i(x,y_B)$ are defined on a closed and bounded domain D. Thus, from (2), we can set $l_i = \max_D |f_i(x,y_B)|$ and it results, for any $x\in[\,0,1\,]$,

$$
|F_{2i}(x)-F_{1i}(x)| \le \frac{1}{8}l_i,\ i = 1,2,\ldots,m.
$$

Then it follows $\|F_2-F_1\| \le l/8$, where $l = \sum_{i=1}^{m} l_i$.
Analogously, by (3), we obtain

$$
|F_{ni}(x)-F_{n-1,i}(x)| \le \frac{1}{8}L_i\,\|F_{n-1}-F_{n-2}\|\ ,\ i = 1,2,\ldots,m
$$

for any $x\in[\,0,1\,]$. Adding all these inequalities and denoting

$L = \sum_{i=1}^{m} L_i$, we have $\| F_n - F_{n-1} \| \leq \frac{1}{8} L \| F_{n-1} - F_{n-2} \|$, or, by recurrence

$$\| F_n - F_{n-1} \| \leq \frac{1}{8} 1 (\frac{L}{8})^{n-2}.$$

Owing to (4) the uniform convergence of $\{F_n(x)\}$ on $[\,0,1\,]$ is proved.

Let $F(x) = \lim_{n\to\infty} F_n(x)$ and

$$s(x) = F(x) - \int_0^1 g(x,t) f(t,F(t))\, dt - y_B(x).$$

To prove that $F(x)$ is a solution of (1) it is sufficient to verify that $s(x) = 0$ identically.

In fact from (5') we have

$$s(x) = F(x) - F_n(x) - \int_0^1 g(x,t)(f(t,F(t)) - f(t,F_{n-1}(t)))\, dt,$$

and from (3) we can write, for any $x \in [\,0,1\,]$,

$$|f_i(x,F(x)) - f_i(x,F_n(x))| \leq L_i \| F - F_n \|, \quad i = 1,2,\ldots,m.$$

Then, because of $\lim_{n\to\infty} \| F - F_n \| = 0$, it can be seen that for any arbitrary positive δ we can choose n so large that $|s_i(x)| < \delta$, $i = 1,2,\ldots,m$, for any $x \in [\,0,1\,]$. Then $s(x) = 0$ identically, which completes the proof.

We introduce now a partial ordering in R^m defining that $v \geq w$ means $v_i \geq w_i$ for $i = 1,2,\ldots,m$; so we can prove the following theorem giving sufficient conditions for the two sided approximat: to the solution $y(x)$ of (1) by means of the sequence $\{F_n(x)\}$.

THEOREM 3. Let the problem (1) satisfy the hypotheses of the theorem 2 and let the inequalities

(6) $f(x,y(x)) \geq 0$,

(7) $J(x,y(x)) = \partial f/\partial y \geq 0$,

hold in $[0,1] \times S$; then the Picard's sequence $\{F_n(x)\}$, as defined in (5) and converging to $y(x)$, satisfies, for $n \geq 1$, the inequaliti

(8) $F_{2n}(x) \leq y(x) \leq F_{2n-1}(x)$.

This statement is also true if inequalities (6), (7) and (8) are reversed.

PROOF. Since the limit vector $F(x) = y(x)$ is a solution of (1) we have

(9) $$y(x) = y_B(x) + \int_0^1 g(x,t) f(t,y(t))\, dt .$$

As previously observed, we can write

(10) $$F_{n+1}(x) = y_B(x) + \int_0^1 g(x,t) f(t,F_n(t))\, dt .$$

Writing (9) in the equivalent form

$$y(x) - y_B(x) = (x-1)\int_0^x tf(t,y(t))\, dt + x\int_x^1 (t-1)f(t,y(t))\, d$$

we have

(11) $y(x) \leq y_B(x) = F_1(x)$.

Subtracting (10) from (9) and using the mean value theorem we obtain

(12) $$y(x) - F_{n+1}(x) = (x-1)\int_0^x t\hat{J}(y(t)-F_n(t))\, dt + x\int_x^1 (t-1)\hat{\hat{J}}(y(t)-F_n(t))\, dt$$

where $\hat{J}$ and $\hat{\hat{J}}$ are two suitable evaluations of the jacobian matri

From (7) and (12) it follows that

$$y(x) \gtreqless F_n(x) \quad \text{implies} \quad y(x) \lesseqgtr F_{n+1}(x) \ ,$$

and taking into account (11), the inequalities (8) are proved.

3. METHODS FOR TWO SIDED APPROXIMATION.

Consider the well known class of linear q-step methods of the form

$$\sum_{i=0}^{q} \alpha_i y_{k+i-1} = h^2 \sum_{i=0}^{q} \gamma_i f_{k+i-1}, \quad q \geq 2$$

(see [6] p. 27-28 and [5] p. 252-256) and limit our attention to the family of methods

$$(13) \qquad y_{k-1} - 2y_k + y_{k+1} = h^2(\gamma_0 f_{k-1} + \gamma_1 f_k + \gamma_2 f_{k+1}).$$

Moreover we consider for γ_0, γ_1, γ_2 only non negative values guaranteeing among other things to avoid operations where exact significant figures may be lost.

It is easy to verify that: if $\gamma_0 + \gamma_1 + \gamma_2 = 1$, from (13) we have a class of first order formulas at least; if in addition we impose $\gamma_1 + 2\gamma_2 = 1$ we have a class of second order formulas; by adding the further condition $\gamma_1 + 4\gamma_2 = 7/6$ we obtain a unique fourth order formula; finally in the family of formulas (13) with non negative γ_i there are two sub-classes of formulas having error constants of opposite sign.

Denoting M_1 and M_2 a couple of formulas with non negative γ_i and error constants of opposite sign, we can set respectively for the truncation errors

$$\tau^{(1)} = C_1 h^{p_1 - 2} y^{(p_1)}(\zeta),$$

$$\tau^{(2)} = -C_2 h^{p_2 - 2} y^{(p_2)}(\eta),$$

where p_1, $p_2 \geq 3$, ζ and η are suitable values on $(0,1)$ even depending on k, C_1 and C_2 are positive constants.

We want to apply M_1 and M_2 to the problem (5). Discretizing the interval $[0,1]$ with the mesh points $x_k = kh$, $k = 0,1,\ldots,K+1$, $h = 1/(K+1)$, and using M_1 or M_2 to approximate the solution of t problem (5), we have to solve at each step an algebraic linear s In fact, denoting by $g_k^{(n)}$ the m-vector giving the approximate di solution at the current step and at the point x_k, supposing to u an exact arithmetic, from (13) we have

$$g_{k-1}^{(n)} - 2g_k^{(n)} + g_{k+1}^{(n)} = h^2(\gamma_0 f(x_{k-1}, g_{k-1}^{(n-1)}) + \gamma_1 f(x_k, g_k^{(n-1)}) + \gamma_2 f(x_{k+1}, g_{k+1}^{(n-1)})).$$

Collecting all these equations for k varying from 1 to K we have

$$Hg_n = c^{(i)}(g_{n-1}) \quad , \quad n = 2,3,\ldots,$$

where g_n is the mK-vector whose components are $(g_k^{(n)})_j$ $((j=1,2,.$ $k=1,2,\ldots,K)$; (i) means we have used the method M_i $(i=1,2)$; $c^{(i)}$ is a non decreasing mK-vector function of its argument $z \in R^{mK}$ a consequence of the condition (7) and of the non-negativity of t selected γ_0, γ_1, γ_2 in (13); H is an mK-order T-matrix ([9], c of the tridiagonal block form

$$H = \begin{bmatrix} -2I & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & -2I \end{bmatrix}$$

where I is the m-order identity matrix. Moreover we observe that is a M-matrix ([9], p. 42-45) and consequently a monotone matri for which $-Hu \leq -Hv$ implies $u \leq v$ or $Hu \geq Hv$ implies $u \leq v$ as we being u, v, real mK-vectors.

Now let $Y_k^{(n)}$ be the m-vector formed with the exact values of the solution $F_n(x)$ of the current problem (5) at the mesh-point x_k and let $G_k^{(n)}$ be the m-vector approximating $Y_k^{(n)}$ and formed with the corresponding values obtained using M_1 or M_2 in presence of round-of errors.

If $\rho_k^{(n)}$ and $\tau_k^{(n)}$ denote the m-vectors giving respectively the local round-off error and the local truncation error, we have, by definiti for each x_k, $k = 1,2,\ldots,K$,

$$(14)\quad G_{k-1}^{(n)} - 2G_k^{(n)} + G_{k+1}^{(n)} = h^2(\gamma_0 f(x_{k-1},G_{k-1}^{(n-1)}) + \gamma_1 f(x_k,G_k^{(n-1)}) + \gamma_2 f(x_{k+1},G_{k+1}^{(n-1)})) + \rho_k^{(n)}$$

$$(14')\quad Y_{k-1}^{(n)} - 2Y_k^{(n)} + Y_{k+1}^{(n)} = h^2(\gamma_0 f(x_{k-1},Y_{k-1}^{(n-1)}) + \gamma_1 f(x_k,Y_k^{(n-1)}) + \gamma_2 f(x_{k+1},Y_{k+1}^{(n-1)})) + h^2\tau_k$$

Subtracting (14') from (14) and setting $e_k^{(n)} = G_k^{(n)} - Y_k^{(n)}$, we obta

$$e_{k-1}^{(n)} - 2e_k^{(n)} + e_{k+1}^{(n)} = h^2(\gamma_0 J_{k-1} e_{k-1}^{(n-1)} + \gamma_1 J_k e_k^{(n-1)} + \gamma_2 J_{k+1} e_{k+1}^{(n-} + \rho_k^{(n)} - h^2\tau_k^{(n)},$$

where J_{k-1}, J_k, J_{k+1} are suitable determinations of the jacobian matrix of the function f.

Defining now the mK-vectors G_n, Y_n, e_n, ρ_n, τ_n having the componen respectively given by $(G_k^{(n)})_j$, $(Y_k^{(n)})_j$, $(e_k^{(n)})_j$, $(\rho_k^{(n)})_j$, $(\tau_k^{(n)})_j$ $((j=1,2,\ldots,m)$, $k=1,2,\ldots,K)$, the last equation can be written as

$$(15)\quad -He_n = -Qe_{n-1} + h^2\tau_n - \rho_n$$

where Q is a mK-order non-negative matrix. Note that $e_1 = 0$ becaus we assume $G_1 = Y_1$.

Denoting with $|v|$ the vector whose components are the absolute val of the corresponding components of v, the following theorem holds.

THEOREM 4. Let the solution of the problem (5) have both p_1-th and p_2-th derivatives ≥ 0 and let the methods M_1 and M_2 be applied to (5) with

$$|\rho_n| < h^2|\tau_n|. \tag{16}$$

Then the method M_1 gives $G_n > Y_n$ if $G_{n-1} \leq Y_{n-1}$ and the method M_2 gives $G_n < Y_n$ if $G_{n-1} \geq Y_{n-1}$. This statement is also true if all inequalities, but (16), are reversed.

PROOF. Since $-H$ is an M-matrix, $(-H)^{-1}$ is a non-negative matri so that from (15) and (16) and taking into account the present s of τ_n the proof is obtained.

Theorems 3 and 4 enable to carry out a procedure for a two sided approximation of the solution of a problem of the type (1) satisf the condition (2), (3), (4), (6), (7) of the section 2 and having solution $y(x)$ with non-negative p_1-th and p_2-th derivatives (the with reversed inequalities is analogously obtained).

Denoting with Y the mK-vector whose components are the exact val of the solution $y(x)$ of the problem (1) at the mesh-points x_1, ..,x_K, the procedure can be described as follows.

From (8) we have

$$Y_{2n} \leq Y \leq Y_{2n-1}, \quad n \geq 1;$$

on the other hand, according to theorem 4, starting with $G_1 = Y_1$ applying the method M_2 to the problem (5), with $n = 2$, we obtain $G_2 < Y_2$ and this implies that the further application of the met M_1 with $n = 3$ will give $G_3 > Y_3$. The entire process can be repea using alternatively M_2 and M_1 to obtain, in general

$$G_{2n} < Y_{2n} \leq Y \leq Y_{2n-1} < G_{2n-1}.$$

We observe that this result can be also obtained using the monot of $-H$ and because $c^{(i)}(z)$ is a not decreasing function.

Because of the convergence of the sequence $\{Y_n\}$ to Y, it can be expected the couples G_{2n}, G_{2n-1} give, for suitable n, good two sided approximations to Y, according to the accuracy of the methods M_1 and M_2.

4. NUMERICAL TESTS.

We have considered the following formulas of the kind (13) and the corresponding truncation errors:

$$(M_1) \qquad \begin{aligned} & y_{k-1} - 2y_k + y_{k+1} = h^2 f_k, \\ & \tau^{(1)} = \frac{1}{12} h^2 y^{(4)}(\zeta); \end{aligned}$$

$$(M_2) \qquad \begin{aligned} & y_{k-1} - 2y_k + y_{k+1} = h^2 \left(\frac{1}{2} f_{k-1} + \frac{1}{2} f_{k+1}\right), \\ & \tau^{(2)} = -\frac{1}{24} h^2 y^{(4)}(\theta), \end{aligned}$$

$$(M_{2'}) \qquad \begin{aligned} & y_{k-1} - 2y_k + y_{k+1} = h^2 \left(\frac{1}{12} f_{k-1} + \frac{5}{6} f_k + \frac{1}{12} f_{k+1}\right), \\ & \tau^{(2')} = -\frac{1}{240} h^4 y^{(6)}(\eta). \end{aligned}$$

As the results are concerned with three test problems, we have quoted in the tables 1, 2, 3, the m-vectors $e_k^{(n)} = e^{(n)}(x_k)$ with n=8 and n=9.

Coupling the preceding formulas in the two fashions (M_1, M_2) and $(M_1, M_{2'})$, an application has been made, in the test 1 and 2, to a class of linear problems like (1) considered in [4]:

$$(17) \qquad y'' = A(x)y(x),$$

where the matrix $A(x) \equiv A \in R^{m\times m}$ is given by

$$A = W''W + (W')^2 + 2W'DW + WDW' + (W'W)^2 + WDWW'W + WD^2W,$$

$W=W(x) \in R^{m\times m}$ is such that $W^2=I$, and $D=\mathrm{diag}(\lambda_i)$ is a diagonal

matrix of order m with λ_i, i = 1,2,...,m, real parameters not depending on x. For this problem we have the exact solution $y(x) = W(x)z(x)$, where

$$z(x) = (e^{\lambda_1 x}, e^{\lambda_2 x}, \ldots\ldots, e^{\lambda_m x})^T.$$

Test 1. We have selected for W the constant binomial matrix of order m

$$W = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \ldots\ldots \\ 1 & -1 & 0 & 0 & 0 & \ldots\ldots \\ 1 & -2 & 1 & 0 & 0 & \ldots\ldots \\ 1 & -3 & 3 & -1 & 0 & \ldots\ldots \\ 1 & -4 & 6 & -4 & 1 & \ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots & & & & & \end{bmatrix},$$

and chosen $\lambda_i = a^{ib/2}$, i = 1,2,...,m, with a>0 and b real numbers Then we have $A = WD^2W$ whose elements are given by

$$(A)_{ij} = \binom{i-1}{j-1} a^{jb}(1-a^b)^{i-j}, \quad 1 \leq j \leq i \leq m.$$

Choosing a and b so that $0 \leq a^b \leq 1$ it is not difficult to verify that the following properties hold:

i) $A \geq 0$;

ii) $\|A\|_\infty = a^b$, so that the condition (4) of the theorem (2) is equivalent to $m \leq 8/a^b$;

iii) $y^{(p)}(x) \geq 0$ if $m \leq 1+1/a^{\frac{1}{2}pb}$.

Then the test problem is

$$y'' = Ay,$$

$$y(0) = \alpha = (1, 0, \ldots, 0)^T,$$

$$y_i(1) = \beta_i = \sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^k e^{\lambda_k}, \quad i = 1,2,\ldots,m.$$

ssuming m=4, b=2, a=0.5, the conditions of the theorems 2 and 3 re satisfied and we have $y^{(4)}(x)\geq 0$, $y^{(6)}(x)\geq 0$, $x \in [0,1]$, so nat the theorem 4, related to the couples (M_1,M_2) and $(M_1,M_2,)$ olds. In the table 1 we have listed the values of $e^{(8)}(x)$ and $^{(9)}(x)$ corresponding to the odd mesh-points with a step-lenght =0.1.

Table 1

	(M_1,M_2)			$(M_1,M_2,)$	
	$e^{(8)}(x)$	$e^{(9)}(x)$	$(\times 10^{-5})$	$e^{(8)}(x)$	$e^{(9)}(x)$
1	-0.63	0.69		-0.13	0.17
	-0.58	0.67		-0.11	0.16
	-0.57	0.66		-0.11	0.15
	-0.55	0.63		-0.10	0.14
.3	-1.04	1.17		-0.21	0.24
	-0.96	1.09		-0.20	0.23
	-0.87	0.99		-0.19	0.21
	-0.81	0.92		-0.18	0.20
.5	-1.22	1.45		-0.32	0.35
	-1.18	1.33		-0.30	0.33
	-1.05	1.20		-0.26	0.31
	-0.98	1.11		-0.24	0.27
.7	-1.18	1.35		-0.30	0.35
	-1.11	1.27		-0.27	0.33
	-1.03	1.17		-0.26	0.30
	-0.98	1.11		-0.24	0.27
.9	-0.74	0.84		-0.15	0.20
	-0.96	0.79		-0.13	0.19
	-0.66	0.75		-0.12	0.17
	-0.63	0.72		-0.11	0.15

est 2. We consider again the problem (17) with m=2 and

$$W = \begin{bmatrix} a_{11}x+b_{11} & a_{12}x+b_{12} \\ a_{21}x+b_{21} & a_{22}x+b_{22} \end{bmatrix}$$

where a_{ij}, b_{ij}, $1\leq i,j\leq 2$ are real numbers chosen according to the condition $W^2 = I$.

It is easy to verify that $(W')^2 = 0$, $W'W=-WW'$ and, obviously, $W''= 0$; so we have $A = 2W'DW+WD^2W$.

Furthermore the class of the matrices like W for which $W'\neq 0$ is defined by the following conditions

$$a_{11}^2 + a_{12}a_{21} = 0 ,$$
$$a_{12}b_{21} + b_{12}a_{21} + 2a_{11}b_{11} = 0 ,$$
$$b_{11}^2 + b_{12}b_{21} = \pm 1 .$$

Selecting, for example, $a_{11}= a_{12}= b_{12}= a_{22}= 0$, $b_{11}= 1$ and $b_{21}= b_{22} = -1$, we obtain the test problem

$$y'' = (2W'DW+WD^2W)y ,$$
$$y(0) = (1,-2)^T, \quad y(1) = (e^{\lambda_1},-e^{\lambda_2})^T .$$

If $0<\lambda_2\leqq\lambda_1\leqq 2$ and $0\leqq\lambda_1-\lambda_2\leqq 1$, the conditions of the theorems 2 and 3 hold, and $y^{(s)}(x)\geqq 0$ for $s\geqq 2$ so that the theorem 4 is applicable as well.

In the table 2 are displayed the values of $e^{(8)}(x)$ and $e^{(9)}(x)$ relatively to the odd mesh points with step-lenght h=0.1 and $\lambda_1=2$, $\lambda_2=1$.

Table 2

	(M_1,M_2)			$(M_1,M_2,)$	
x	$e^{(8)}(x)$	$e^{(9)}(x)$	$(\times 10^{-4})$	$e^{(8)}(x)$	$e^{(9)}(x)$
0.1	-0.74	1.05		-0.39	0.59
	-0.96	1.03		-0.38	0.58
0.3	-1.56	2.21		-0.82	1.24
	-1.07	1.51		-0.56	0.87
0.5	-3.31	4.80		-1.80	2.73
	-1.57	2.20		-0.80	1.22
0.7	-7.76	10.92		-4.15	6.28
	-2.15	3.03		-1.19	1.70
0.9	-6.49	9.10		-3.37	5.12
	-1.40	1.94		-0.74	1.13

Test 3. As a case different from the type (17), we consider the nonlinear problem

$$y_i''(x) = e^{y_{m-i+1}} , \quad i = 1,2,\ldots,m.$$

n this equation the conditions (6) and (7) are verified, and $^{(4)}(x)>0$.

urthermore, choosing suitably the boundary values of y(x) such hat $0 \geq \alpha > \beta$, we find $y(x) \leq 0$ and $y'(x) \leq 0$ on $[0,1]$.
hen even the conditions $y^{(6)}(x)>0$ and $\|\partial f/\partial y\|_\infty \leq 1$ hold: thus it s possible to apply the two sided approximation process for ≤ 8.

n the table 3 results for $e^{(8)}(x)$ and $e^{(9)}(x)$ at some odd mesh oints with h=0.1 are displayed for a problem with m=4 and boundary onditions given by

$$y(0) = 0,$$
$$y_1(1) = -0.4, \qquad y_3(1) = -0.8,$$
$$y_2(1) = -0.6, \qquad y_4(1) = -1.0.$$

Table 3

	(M_1,M_2)			$(M_1,M_2,)$	
x	$e^{(8)}(x)$	$e^{(9)}(x)$	$(\times 10^{-4})$	$e^{(8)}(x)$	$e^{(9)}(x)$
0.1	-2.49	3.41		-0.35	0.51
	-1.85	2.57		-0.28	0.41
	-1.25	1.71		-0.21	0.30
	-0.83	1.13		-0.12	0.18
0.3	-4.44	6.05		-0.76	1.13
	-3.66	4.98		-0.56	0.82
	-2.58	3.50		-0.30	0.45
	-1.72	2.33		-0.24	0.34
0.5	-4.39	6.00		-0.68	1.00
	-3.57	4.91		-0.56	0.81
	-2.92	3.99		-0.37	0.55
	-1.78	2.44		-0.27	0.39
0.7	-3.20	4.35		-0.49	0.71
	-2.66	3.61		-0.37	0.54
	-1.93	2.62		-0.29	0.43
	-1.28	1.75		-0.20	0.29
0.9	-1.16	1.61		-0.17	0.25
	-0.97	1.33		-0.14	0.21
	-0.80	1.10		-0.09	0.16
	-0.52	0.70		-0.07	0.12

All calculations have been performed on the IBM 370 at the CNUCE of Pisa using a double precision arithmetic (8bytes).

REFERENCES

[1] P.B. BAILEY - L.F. SHAMPINE - P.E. WALTMAN, Nonlinear two point boundary value problems, Academic Press, New York, 1968.

[2] J.W. DANIEL - R.E. MOORE, Computation and theory in ordinary differential equations, W.H. Freeman and Company, San Francisco, 1970.

[3] G. GHELARDONI, Sul problema di valori al contorno per l'equazione differenziale y''=f(x,y,y'), Accademia Nazionale dei Lincei, Serie VIII, vol. XXXIII, fasc. 5, 1962, p. 237-243.

[4] G. GHERI - P. MARZULLI, Collocation for initial value problems based on Hermite interpolation, CALCOLO, vol. XXIII, 1986, p. 115-130.

[5] J.D. LAMBERT, Computational methods in ordinary differen equations, Jhon Wiley, London, 1972.

[6] L. LAPIDUS - J.H. SEINFELD, Numerical solution of ordina differential equations, Academic Press, New York, 197

[7] G.F. ROACH, Green's functions, Cambridge University Pres Cambridge, 1982.

[8] S.M. ROBERTS - J.S. SHIPMAN, Two-point boundary value problems: shooting methods, American Elsevier, New Yor 1972.

[9] D.M. YOUNG, Iterative solution of large linear systems, Academic Press, New York, 1971.

ON THE APPROXIMATE CALCULATION OF INTEGRALS ON A POLYGON IN R^2

ALLAL GUESSAB

Abstract : We will consider the problem of approximating a double integral on a polygon in $\mathbb{R}^2$ as a linear combination of integrals on the real line. Cubature formulas are obtained in such a way as to minimize the exact error bounds of the formulas for a given class of functions.

1. INTRODUCTION : NOTATIONS AND DEFINITIONS

The theory of numeric cubature formulas for functions of one variable is well developed. We refer to Davis-Rabinowitz [1], Stroud-Secrest [23] and Krylov [11] .

In this work, we will consider the problem of approximating a double integral on a polygon K in $\mathbb{R}^2$ as a linear combination of integrals on the real lir

Let us first fix a few notations. For $\alpha = (\alpha_1,\alpha_2)$ in $\mathbb{N}^2$, we denote by X^α the monomial defined by :

$$X^\alpha = x^{\alpha_1} y^{\alpha_2} .$$

For $k = (k_1,k_2)$ in $\mathbb{N}^*.\mathbb{N}^*$, and $m = (m_1,m_2)$ in $\mathbb{N}^*.\mathbb{N}^*$ (such that $k_1 < m_1$ and $k_2 < m_2$) . We have the following definitions :

$$(1.1) \qquad L_k = \left\{ \alpha = (\alpha_1,\alpha_2) \in \mathbb{N}^2 , \quad \alpha_i \le k_i , \quad i = 1,2 \right\} ,$$

$$(1.2) \qquad R_k = \left\{ R \equiv \sum_{\alpha \in L_k} a_\alpha X^\alpha , \quad a_\alpha \in \mathbb{R} \right\} ,$$

$$(1.3) \qquad V_k = \left\{ R \equiv X^k + \sum_{\alpha \in L_{(k_1-1,k_2-1)}} a_\alpha X^\alpha , \ a_\alpha \in \mathbb{R} \right\} .$$

Finally, let c be a real number, which is assumed to be fixed. Let us introduce the following notation :

$M^{k}_{0,K}(c)$ is the set of all functions $f(x,y)$ which have piecewise continuous derivatives

$$\frac{\partial^{i+j}}{\partial x^{i}\partial y^{j}} f(x,y) \ , \quad i = 0,1,\ldots,k_1 \ ; \quad j = 0,1,\ldots,k_2 \tag{1.4}$$

on K and satisfy the conditions

$$\left\| \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f \right\|_{0,K} \leq c \tag{1.5}$$

where :

$$\left\| \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f \right\|_{0,K} = \left(\int_K \left(\frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f(x,y) \right)^2 dxdy \right)^{1/2} \tag{1.6}$$

Let K be a polygon in $\mathbb{R}^2$ and g_k an arbitrary polynomial in V_k. Consequently the following equality is true :

$$I_K(f) = \int_K f(x,y)dxdy = \frac{1}{k_1!k_2!} \int_K f(x,y) \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} g_k(x,y)dxdy \ .$$

Then if $f \in M^{k}_{0,K}(c)$, applying Green's formula to the right-hand side, we obtain the following cubature formula :

$$I_K(f) = Q_K(f,g_k) + E_K(f,g_k) \ , \tag{1.7}$$

where

$$Q_K(f,g_k) = \frac{1}{k_1!k_2!} \left(\sum_{j=0}^{k_1-1} (-1)^j \int_\Gamma \frac{\partial^j}{\partial x^j} f(x,y) \frac{\partial^{k_1-1-j+k_2}}{\partial x^{k_1-1-j}\partial y^{k_2}} g_k(x,y)\vartheta_1 \right.$$
$$\left. + \sum_{j=0}^{k_2-1} (-1)^{k_1+j} \int_\Gamma \frac{\partial^{k_1+j}}{\partial x^{k_1}\partial y^{j}} f(x,y) \frac{\partial^{k_2-1-j}}{\partial y^{k_2-1-j}} g_k(x,y) \ \vartheta_2 d\sigma \right) ,$$

where ϑ_i , is the i-th component of the outer normal vector along K , and

$$E_K(f,g_k) = \frac{(-1)^{k_1+k_2}}{k_1!k_2!} \int_K g_k(x,y) \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f(x,y)dxdy \ . \tag{1.8}$$

APPROACH TO THE PROBLEM

Our goal in this work is to derive optimal cubature formulas of the type (1.7) in the space $M^k_{0,K}(c)$. The formula (1.7) will be optimal in the space $M^k_{0,K}(c)$, if the polynomial g_k is chosen so that the quantity

$$E(M^k_{0,K}(c)) = \sup_{f\in M^k_{0,K}(c)} |E(f,g_k)| .$$

has the minimal value.

. PRELIMINARY RESULTS

Remark 2.1. It is clear that the cubature formula (1.7) is exact on F, i.e.

$$E_K(f,g_k) = 0 , \quad \text{for } f \text{ in } F ,$$

where F is the vector space generated by the family of monomials X^α, $\alpha = (\alpha_1,\alpha_2) \in A$, with

$$A = (\{0,1,2,\ldots,k_1-1\} \cdot \mathbb{N}) \cup (\mathbb{N} \cdot \{0,1,2,\ldots,k_2-1\})$$

Remark 2.2. Assuming $f \in M^k_{0,K}(c)$. By Hölder's inequality, we obtain from (1.8) that :

$$(2.1) \qquad \sup_{f\in M^k_{0,K}(c)} |E_K(f,g_k)| \le \frac{c}{k_1!k_2!} \|g_k\|_{0,K}$$

Proposition 2.1. Assume $Q_K(f,g_k)$ is a formula of type (1.7). Then :

$$(2.2) \qquad \sup_{f\in M^k_{0,K}(c)} |E_K(f,g_k)| = \frac{c}{k_1!k_2!} \|g_k\|_{0,K}$$

Proof : For the function

$$R(x,y) = \frac{(-1)^{k_1+k_2}}{(k_1-1)!(k_2-1)!} \frac{c}{\|g_k\|_{0,K}} \int_0^x \int_0^y (x-u)^{k_1-1} (y-v)^{k_2-1} g_k(u,v)$$

belonging to $M^k_{0,K}(c)$, it follows from (1.8) that

$$|E_K(R,g_k)| = \frac{c}{k_1!k_2!} \|g_k\|_{0,K} .$$

Then we have from (2.1) the equality (2.2).

In the sequel, we will use the following terminology :
K is a polygon in $\mathbb{R}^2$. We establish a triangulation $\mathcal{C}_h$ ((cf. Fig.2.1)) over K, i.e. K is expressed as a finite union

$$K = \bigcup_{i=1}^{n} K_{h,i} \quad , \tag{2.3}$$

of triangles $K_{h,i}$ in such a way that these triangles are non-overlapping, and are all interior to K .

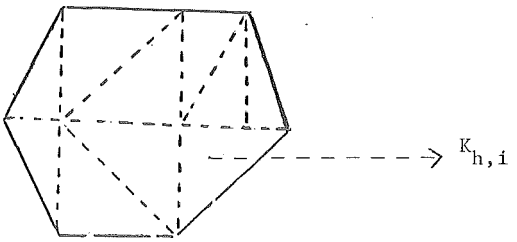


Fig.2.1 Subdivision of K into triangles.

APPROACH TO THE PROBLEM IF K IS A TRIANGLE.

Let K be a triangle with vertices (a,b), $(a+h,b)$ and $(a,b+h')$, where $h, h' \in \mathbb{R}_+^*$. We denote by Γ_i , $i = 1,2,3$ the three sides of K , and

$$\Gamma = \bigcup_{i=1}^{3} \Gamma_i \quad , \tag{3.1}$$

where Γ_1 (resp. Γ_2) is a piece of a line parallel to the y-axis (resp. x-axis) :

$$\Gamma_1 = \{(x,y) \in K \ , \ d_1(x,y) = x - a = 0 \}$$

$$\Gamma_2 = \{(x,y) \in K \ , \ d_2(x,y) = y - b = 0 \} \quad .$$

Definition 3.1. type (1.7) cubature formula $Q_K(f,g_k)$ is said to be optimal with respect to K of order $\tilde{r} = (r_1,r_2) \in \mathbb{N}^* \mathbb{N}^*$ such that $r_1-1 \le k_1$ and $r_2-1 \le k_2$, if and only if the following properties hold :

i) g_k lies in $V_k^{(1)}$, where $V_k^{(1)}$ is the set of polynomials p_k in V_k such that :

$$p_k = p^*_{k-\tilde{r}} d_1^{r_1} d_2^{r_2} \quad , \quad p^*_{k-\tilde{r}} \in V_{k-\tilde{r}} \ .$$

ii) $$E_K^*(f,g_k) = \sup_{f \in M^k_{0,K}(c)} |E_K(f,g_k)| = \inf_{p_k \in V_k^{(1)}} \sup_{f \in M^k_{0,K}(c)} |E_K(f,p_k)| \ .$$

Theorem 3.1. Let K be a triangle of the type (3.1). Then, there exists an optimal cubature formula of the type (1.7) with respect to K, of order $\tilde{r} = (r_1,r_2)$, such that :

$$g_k = g^*_{k-\tilde{r}} d_1^{r_1} d_2^{r_2} \quad ,$$

where $g^*_{k-\tilde{r}}$ is in $V_{k-\tilde{r}}$, and is orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(K)$, when considering the inner product associated to integration on K and weight function $d_1^{2r_1} d_2^{2r_2}$.

Proof : From proposition 2.1, we have :

$$\sup_{f \in M^k_{0,K}(c)} |E_K(f,p_k)| = \frac{c}{k_1!k_2!} \|p_k\|_{0,K} \quad , \text{ for all } p_k \in V_k \ .$$

Then

$$E_K^*(f,g_k) = \inf_{p_k \in V_k^{(1)}} \sup_{f \in M^k_{0,K}(c)} |E_K(f,p_k)| = \frac{c}{k_1!k_2!} \inf_{p_k \in V_k^{(1)}} \|p_k\|_{0,K} \ .$$

So :

$$E_K^*(f,g_k) = \frac{c}{k_1!k_2!} \inf_{p \in V_{k-\tilde{r}}} \left[I_K (d_1^{2r_1} d_2^{2r_2} p^2) \right]^{1/2} \ .$$

It is shown in [12] that this problem has one and only one solution $g^*_{k-\tilde{r}} \in V_{k-\tilde{r}}$, also characterized by

$$I_K\left(d_1^{2r_1} d_2^{2r_2} g^*_{k-\tilde{r}} R\right) = 0 \ , \ \forall\, R \in R_{(k_1-r_1-1,k_2-r_2-1)}(K) \ .$$

Example 3.1. $K = \{(x,y) \in \mathbb{R}^2 \ , \ x + y \le 1 \ , \ x \ge 0 \ , \ y \ge 0\}$, $k_1 = k_2 = 3$, and : $r_1 = r_2 = 2$. In this case : $d_1(x,y) = x$, $d_2(x,y) = y$, and

$g_{(3,3)}(x,y) = x^2y^2 \left(xy - \frac{25}{132}\right)$.

We then get :

$$E^*_K(f,g_{(3,3)}) = \frac{c}{36} \|g_{(3,3)}\|_{0,K} \simeq 1.4\ 10^{-5} c \ .$$

Finally, the optimal cubature formula reads as :

$$\begin{aligned} Q_K(f,g_{(3,3)}) = \frac{1}{36} \Bigg(& 36 \int_0^1 x f(x,1-x)dx - 18 \int_0^1 x^2 \frac{\partial}{\partial x} f(x,1-x)dx \\ & + 6 \int_0^1 x^3 \frac{\partial^2}{\partial x^2} f(x,1-x)dx - \frac{25}{66} \int_0^1 x^2 \frac{\partial^3}{\partial x^3} f(x,0)dx \\ & - \int_0^1 x^2 \left(6x(1-x) - \frac{25}{66}\right) \frac{\partial^3}{\partial x^3} f(x,1-x)dx \\ & + \int_0^1 x^2(1-x)\left(3x(1-x) - \frac{25}{66}\right) \frac{\partial^4}{\partial x^3 \partial y} f(x,1-x)dx \\ & - \int_0^1 x^2(1-x)^2 \left(x(1-x) - \frac{25}{132}\right) \frac{\partial^5}{\partial x^3 \partial y^2} f(x,1-x)dx \Bigg) . \end{aligned} \tag{3.2}$$

which is exact on F , where F is the vector space generated by the family of monomials X^α , $\alpha = (\alpha_1, \alpha_2) \in A$, with

$$A = (\{0,1,2\} \times \mathbb{N}) \cup (\mathbb{N} \times \{0,1,2\}) . \tag{3.3}$$

Example 3.2. $K = \{(x,y) \in \mathbb{R}^2 \ , \ x + y \le 1 \ , \ x \ge 0 \ , \ y \ge 0\}$, $k_1 = k_2 = 5$, and : $r_1 = r_2 = 4$. In this case : $d_1(x,y) = x$, $d_2(x,y) = y$, and

$$g_{(5,5)}(x,y) = x^4y^4 \left(xy - \frac{81}{380}\right) .$$

We then get :

$$E_K^*(f,g_{(5,5)}) = \frac{c}{5!5!}\,\|g_{(5,5)}\|_{0,K} \simeq 10^{-9}c \quad .$$

Finally, the optimal cubature formula reads as :

$$\begin{aligned} Q_K(f,g_{(5,5)}) = \frac{1}{14400}\Bigg(& 14400 \int_0^1 xf(x,1-x)dx - 7200 \int_0^1 x^2 \frac{\partial}{\partial x} f(x,1-x)dx \\ & + 2400 \int_0^1 x^3 \frac{\partial^2}{\partial x^2} f(x,1-x)dx - 600 \int_0^1 x^4 \frac{\partial^3}{\partial x^3} f(x,1-x)dx \\ & + 120 \int_0^1 x^5 \frac{\partial^4}{\partial x^4} f(x,1-x)dx - \frac{486}{95} \int_0^1 x^4 \frac{\partial^5}{\partial x^5} f(x,0)dx \\ & - \int_0^1 x^4(120x(1-x) - \frac{486}{95}) \frac{\partial^5}{\partial x^5} f(x,1-x)dx \\ & + \int_0^1 x^4(1-x)(60x(1-x) - \frac{486}{95}) \frac{\partial^6}{\partial x^5 \partial y} f(x,1-x)dx \\ & - \int_0^1 x^4(1-x)^2(20x(1-x) - \frac{243}{95}) \frac{\partial^7}{\partial X^5 \partial y^2} f(x,1-x)dx \\ & + \int_0^1 x^4(1-x)^3(5x(1-x) - \frac{81}{95}) \frac{\partial^8}{\partial x^5 \partial y^3} f(x,1-x)dx \\ & - \int_0^1 x^4(1-x)^4(x(1-x) - \frac{81}{380}) \frac{\partial^9}{\partial x^5 \partial y^4} f(x,1-x)dx \Bigg) \end{aligned} \tag{3.4}$$

which is exact on F , where F is the vector space generated by the family of monomials X^α , $\alpha = (\alpha_1,\alpha_2) \in A$, with

$$A = (\{0,1,2,3,4\} \,.\, \mathbb{N}) \cup (\mathbb{N} \,.\, \{0,1,2,3,4\}) \quad .$$

Example 3.3 : $K = \{(x,y) \in \mathbb{R}^2 \,,\, x + y \le 1 \,,\, x \ge 0 \,,\, y \ge 0\}$, $k_1 \in \mathbb{N}^*$, $k_2 \in \mathbb{N}^*$ and : $r_1 = k_1 - 1$, $r_2 = k_2 - 1$. In this case : $d_1(x,y)=x$, $d_2(x,y)=y$, and $g_{(k_1,k_2)}(x,y) = x^{k_1-1}y^{k_2-1}(xy - c_k)$, where

$$c_k = \frac{(2k_1-1)(2k_2-1)}{(2k_1+2k_2-1)(2k_1+2k_2)} \quad ,$$

We then get :

$$E_K^*(f,g_{(k_1,k_2)}) = \frac{c}{k_1!k_2!} \left\| g_{(k_1,k_2)} \right\|_{0,K} =$$

$$\frac{1}{k_1!k_2!}\left(\frac{(2k_1)!(2k_2)!}{(2k_1+2k_2+2)!} - 2c_k \frac{(2k_1-1)!(2k_2-1)!}{(2k_1+2k_2)!} + c_k^2 \frac{(2k_1-2)!(2k_2-2)!}{(2k_1+2k_2-2)!} \right)$$

Finally, the optimal cubature formula reads as :

$$Q_k(f,g_k) = \frac{1}{k_1!k_2!} \left(\sum_{j=0}^{k_1-1} (-1)^j A_{j,k} \int_0^1 x^{j+1} \frac{\partial^j}{\partial x^j} f(x,1-x)dx \right.$$

$$\left. + (-1)^{k_2} B_k \int_0^1 x^{k_1-1} \frac{\partial^{k_1}}{\partial x^{k_1}} f(x,0)dx \right.$$

$$\left. + \sum_{j=0}^{k_2-1} (-1)^{k_1+j} \int_0^1 x^{k_1-1} (1-x)^j (C_{j,k}\, x(1-x) - D_{j,k}) \frac{\partial^{k_1+j}}{\partial x^{k_1} \partial y^j} f(x,1-x) \right.$$

where

$$A_{j,k} = k_2!(k_1-j-1)! \binom{j+1}{k_1}$$

$$B_k = (k_2-1)!C_k$$

$$C_{j,k} = (k_2-j-1)! \binom{j+1}{k_2}$$

$$D_{j,k} = (k_2-j-1)! \binom{j}{k_2-1} C_k$$

Remark 3.1. If K is a triangle with vertices (e,f), $(e+h_1,f)$, $(e,f+h_2)$ then using the change of variables $u = \frac{x-e}{h_1}$, $v = \frac{y-f}{h_2}$, we are led back to the situation $\hat{D} = \{(x,y) \in \mathbb{R}^2,\ x \geq 0,\ y \geq 0,\ x+y \leq 1\}$, with :

$$E_k^*(f,g_k) = h_1^{k_1+\frac{1}{2}} h_2^{k_2+\frac{1}{2}} E_{\hat{D}}^*(f,p_k). \quad ,$$

where

$$p_k = x^{r_1} y^{r_2} p_{k-\tilde{r}},$$

and $p_{k-\tilde{r}} \in V_{k-\tilde{r}}$ and is orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(\hat{D})$, with respect to $\hat{D}$ and the weight function $x^{2r_1} y^{2r_2}$.

Remark 3.2. If K is a triangle with vertices (x_1,y_1), (x_2,y_2), (x_3,y_3), then, using the change of variables

$$u(x,y) = (x_2-x_1)x + (x_3-x_1)y + x_1$$
$$v(x,y) = (y_2-y_1)x + (y_3-y_1)y + y_1 \quad ,$$

we have

$$\int_K f(u,v)dudv = |J| \int_{\hat{D}} f(u(x,y), v(x,y))\,dxdy \quad , \tag{3.5}$$

where :

$$\hat{D} = \{ (x,y) \in \mathbb{R}^2 \ , \ x \geq 0 \ , \ y \geq 0 \ , \ x + y \leq 1 \} \quad ,$$

and

$$J = \ell_2 \ell_3 - \ell_1 \ell_4 \quad ,$$

with

$$\ell_1 = x_3-x_1 \ , \ \ell_2 = x_2-x_1 \ , \ \ell_3 = y_3-y_1 \ , \ \ell_4 = y_2-y_1$$

From (3.2) and (3.5), we have the cubature formula :

(3.6)

$$Q_K(f,g_{(3,3)}) = \frac{|J|}{36}\Big(36 \int_0^1 xf(\ell_1 x+\ell_2(1-x),\ell_3 x+\ell_4(1-x)+y_1)dx$$
$$- 18(\ell_1+\ell_3) \int_0^1 x^2 \frac{\partial}{\partial x} f(\ell_1 x+\ell_2(1-x),\ell_3 x+\ell_4(1-x)+y_1)dx$$

$$+ 6(\ell_1+\ell_3)^2 \int_0^1 x^3 \frac{\partial^2}{\partial x^2} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1)\,dx$$

$$+ \frac{25}{66} (\ell_1+\ell_3)^3 \int_0^1 x^2 \frac{\partial^3}{\partial x^3} f(\ell_1 x+x_1 , \ell_3 x+y_1)\,dx$$

$$-(\ell_1+\ell_3)^3 \int_0^1 x^2(6x(1-x) - \frac{25}{66}) \frac{\partial^3}{\partial x^3} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1)\,dx$$

$$+(\ell_1+\ell_3)^3(\ell_2+\ell_4) \int_0^1 x^2(1-x)(3x(1-x)- \frac{25}{66}) \frac{\partial^4}{\partial x^3 \partial y} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1$$

$$- (\ell_1+\ell_3)^3(\ell_2+\ell_4)^2 \int_0^1 x^2(1-x)^2(x(1-x)- \frac{24}{132}) \frac{\partial^5}{\partial x^3 \partial y^2} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1)$$

4. GENERAL PROBLEM

Let K be a polygon in $\mathbb{R}^2$, from (2.3) we can write

$$I_K(f) = \sum_{i=1}^{n} I_{i,h}(f) \quad , \tag{4.1}$$

where

$$I_{i,h}(f) = \int_{K_{i,h}} f(x,y)\,dxdy \quad ,$$

where $K_{i,h}$, are defined in (2.3).

For each $K_{h,i}$, we assume that the boundary of $K_{i,h}$

$\Gamma_{h,i,j} = \bigcup_{i=1}^{3} \Gamma_{h,i,j}$

where $\Gamma_{h,i,j}$, $j = 1,2$ are defined by :

$$\Gamma_{h,i,1} = \{(x,y) \in K_{h,i} , d_{h,i,1}(x,y) = x-e_{h,i} = 0 \}$$

$$\Gamma_{h,i,2} = \{(x,y) \in K_{h,i} , d_{h,i,2}(x,y) = y-f_{h,i} = 0 \}$$

with $e_{h,i} \in \mathbb{R}$, $f_{h,i} \in \mathbb{R}$.

Hence, we obtain the following equality ,

$$I_K(f) = \frac{1}{k_1!k_2!} \left(\sum_{i=1}^{n} I_{i,h} \left(f \frac{\partial^{k_1+k_2}}{\partial x^{k_1} \partial y^{k_2}} g_{k,i} \right) \right)$$

for each $g_{k,i}$ in $V_k(K_{h,i})$.

Now, we may apply (1.7) for each $K_{h,i}$, and it follows :

$$I_K(f) = \sum_{i=1}^{n} Q_{h,i}(f,g_{k,i}) + \sum_{i=1}^{n} E_{h,i}(f,g_{k,i}) \tag{4.2}$$

$$= Q_K(f) + E_K(f) \quad .$$

Assuming $f \in M^k_{0,K}(K)$, from (4.2), we have :

$$\sup_{f \in M^k_{0,K}(c)} |E_K(f)| \leq \frac{c}{k_1!k_2!} \left(\sum_{i=1}^{n} \|g_{k,i}\|_{0,K_{h,i}} \right) .$$

If $g_{k,i}$ is of the form :

$$g_{k,i} = g^{(1)}_{h,i} d^{r_1}_{h,i,1} \cdot d^{r_2}_{h,i,2} \quad ,$$

with $g^{(1)}_{k,i} \in V_{k-\tilde{r}}$, where $\tilde{r} = (r_1,r_2)$.

By Hölder's inequality we obtain from (4.2) that

$$\sup_{f \in M^k_{0,K}(c)} |E_K(f)| \leq \frac{c}{k_1!k_2!} \left(\sum_{i=1}^{n} \|g^*_{k,i}\|_{0,K_{h,i}} \right) \tag{4.3}$$

$$= \inf_{g_{k,i} \in V^{(1)}_k} \left(\sum_{i=1}^{n} \|g_{k,i}\|_{0,K_{h,i}} \right) ,$$

where

$$g^*_{k,i} = g^*_{h,i} d^{r_1}_{h,i,1} d^{r_2}_{h,i,2} \quad , \tag{4.4}$$

with $g^{*}_{h,i} \in V_{k-\tilde{r}}(K_{h,i})$, and orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(K_{h,i})$ when considering the weight function $d^{2r_1}_{h,i,1} d^{2r_2}_{h,i,2}$. We summarize the results of this section by :

Theorem 4.1. Let K be a polygon of $\mathbb{R}^2$. Then, there exists an optimal cubature formula of the type (4.2), where

$$g_{k,i} = g^{*}_{h,i} d^{r_1}_{h,i,1} d^{r_2}_{h,i,2} ,$$

and

$g^{*}_{h,i} \in V_{k-\tilde{r}}(K_{h,i})$, and orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(K_{h,i})$ when considering the weight function $d^{2r_1}_{h,i,1} d^{2r_2}_{h,i,2}$.

And the remainding term satisfies the relation (4.3).

5. NUMERICAL EXAMPLES.

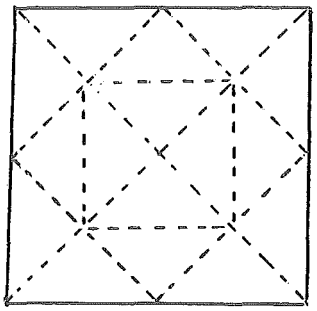
In this section we present some numerical examples in order to demonstrate the performance of the optimal formula (3.6) for various choices of K .

We compare the evaluation of the integral

$$I(f) = \int_K \frac{1}{4+x+y} \, dx \, dy$$

by the optimal formula (3.6). For each of the simple integrals of formula (3.6), we use an 5-point Gauss formula.

Example 5.1 : If $K = [-1.1] \cdot [-1.1]$, and $f(x,y) = \frac{1}{4+x+y}$,

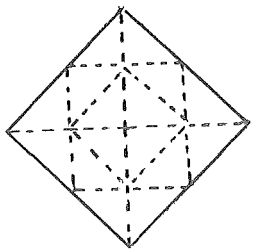


Subdividion of K into triangles.

In this case, we have :

Number of triangles	Approximate value of I(f)
16	1.04659549549661
64	1.04649884004569
256	1.04649633382633
1024	1.04649628828193
4096	1.04649628754098
Exacte value	1.0464962 8752910

Example 5.2 : If K is the Lozenge $(\pm 1,0)$, $(0,\pm 1)$, and $f(x,y) = \frac{1}{4+x+y}$



Subdivision of K into triangles.

In this case, we have :

Number of triangles	Approximate value of I(f)
16	0.509837941482940
64	0.510825848299233
256	0.510825627424528
1024	0.510825623823773
4096	0.510825623762376
Exact value	0.510825623765991

The above calculation were carried out in turbo-pascal (about sixteen significant digits) on the IBM PC AT.

References

[1] P.J. DAVIS & P. RABINOWITZ : "Methods of numerical integration". Academic press, New York, 1975.

[2] I. GANSCA : "Best quadrature formulas with relatively few terms". Revue roumaine de Math. Pures et Appliquées XXI, Vol.2, 1977, pp.143-151.

[3] J.L. GOUT & A. GUESSAB : "Sur les formules de quadrature numérique à nombre minimal de noeuds d'intégration". Numer. Math. Vol.49,1986, pp.439-455.

[4] J.L. GOUT & A. GUESSAB : "Exemples de formules de quadrature numérique à nombre minimal de noeuds sur des domaines à double symétrie axial R.A.I.R.O., vol.20, 1986, pp.287-314.

[5] A. GUESSAB : "Formules de quadrature numérique dans un compact de $\mathbb{R}^n$" . Thèse de 3ème cycle, 1983.

[6] A. GUESSAB : "Cubature formulae which are exact on space P , intermediat between P_k and Q_k". Numer. Math., Vol.49, 1986, pp.561-576.

[7] A. GUESSAB : "Numerical cubature formulas with preassigned knots". to appear. Numer. Math. 1987.

[8] A. GUESSAB : "Numerical cubature with multiple knots" to appear IMA journal or Numer. Anal. 1987.

[9] A. GUESSAB : "Sur les formules de quadrature numérique dans $\mathbb{R}^n$ avec certains noeuds ayant une compoante connue" to appear Applicable Analysis 1987.

[10] C.B. HUELSMAN:"Quadrature formulas over fully symmetric planar regions". Numer. Math. Vol.10, pp.539-552, 1973.

[11] H.I. KRYLOV : "Approximate calculation of integral". Mac. Millan New Yor London 1962.

[12] P.J. LAURENT : "Approximation et Optimation" Hermann, Paris, 1972.

[13] M. LEVIN : "On the approximate calculation of double integrals". Math. Comp. Vol.40, 1983, pp.273-282.

[14] M. LEVIN & J. GIRSHOVICH "Extremal problems for cubature formulas" Soviet Math. dokl., Vol.18, 1977, pp.1355-1358.

[15] H.M. MÖLLER : "Kubaturformeln mit minimaler knotenzhl" Num. Math. Vol.25, 1976, pp.185-200.

[16] H.M. MÖLLER : "Lower Bounds for the number of nodes in cubature formula Birlchöuser Verlag ISNM Vol.45, 1978, pp.221-230.

[17] F.W.J. OLVER : "Asymptotics and special functions" Academic Press, New York San Francisco London 1974.

18] J. PIESSENS & HAEGEMANS : "Cubature formulas of degree seven for symmetric planar regions" Journal of Comp. and applied Math. Vol.1, 1975, pp.79-83.

19] P. RABINOWITZ & N. RICHTER : "Perfectly symmetric two-dimensional integration formulas with minimal numbers of points" Math. Comp., Vol.23, 1969, pp.767-779.

20] P. RABINOWITZ & N. RICHTER : "Asymptotic properties of minimal integration rules" Math. Comp. Vol.24, 1970, pp.593-609.

21] H.J. SCHMIDT : "On cubature formulae with a minimal number of knots". Numer. Math., vol.31, 1978, pp.281-297.

22] A.H. STROUD : "Approximate calculation of multiple integrals". Prentice Hall, Englewood cliffs, N.J. 1971.

23] A.H. STROUD & D. SECREST : "Gaussian quadrature formulas". Prentice Hall, Englewood cliffs, N.J. 1966.

24] G. SZEGÖ : "Orthogonal polynomials" 3rd ed. Amer. Math. Soc. Colloq. Publ., Vol.VVIII. Amer. Math. Soc., New York (1960).

A COMBINATION OF RELAXATION METHODS AND METHOD OF AVERAGING FUNCTIONAL CORRECTIONS

DRAGOSLAV HERCEG and LJILJANA CVETKOVIĆ

ABSTRACT: *We consider a combination of the Accelerated Overrelaxation method for solving linear systems (basic method), introduced by A. Hadjidimos, with the method of Averaging Functional Corrections in order to form the composite method, which is in some cases faster than the basic method. Sufficient conditions for the convergence of this method are obtained. Several numerical examples demonstrate the efficiency of our method.*

1. INTRODUCTION

If we want to solve a system of linear equations

$$x = Mx + d,\ M = [m_{ij}] \in \mathbb{R}^{n,n},\ d = [d_1, \dots, d_n]^T \in \mathbb{R}^n, \tag{1}$$

instead of the basic iterative method

$$x^{k+1} = Mx^k + d, \qquad k = 0, 1, \dots,$$

in order to accelerate the convergence, we can use AFC (method of averaging functional corrections). This method was introduced by Sirenko [5], where it was given in the following form:

Algorithm:

Step 0: Calculate $m = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}$;

Step 1: If $n \leq m$ stop, otherwise go to step 2;

Step 2: Choose $x^0 \in \mathbb{R}^n$;

Step 3: Calculate $s_0 = \frac{1}{n-m}\sum_{i=1}^{n} d_i$; $k = 0$;

Step 4: Calculate $x^{k+1} = M(x^k + s_k\delta) + d$, $\delta = [1, \ldots, 1]^T \in \mathbb{R}^n$;

Step 5: Calculate $s_{k+1} = \frac{1}{n-m}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\left(x_j^{k+1} - x_j^k - s_k\right)$;

Step 6: Take $k = k + 1$ and return to step 4.

Numerical examples show that, very often, AFC method converges faster than the basic method. Because of that, it was the subject of our investigations, [2] the results of which we shall give in section 2.

In this paper, as the basic iterative method, we shall use AOR (Accelerated Overrelaxation) method introduced by A. Hadjidimos [4]. It means that we consider a system of linear equations

$$Ax = b, \quad A = [a_{ij}] \in \mathbb{R}^{n,n}, \quad b \in \mathbb{R}, \tag{2}$$

which we solve by using AOR method

$$x^{k+1} = M_{\sigma,\omega}x^k + d, \quad k = 0,1, \ldots,$$

and, after that, by AFC method. Here we denote by $M_{\sigma,\omega} = (D - \sigma T)^{-1}((1 - \omega)D + (\omega - \sigma)T + \omega S)$, $d = \omega(D - \sigma T)^{-1}b$, where $A = D - T - S$ is the standard splitting of the matrix A into diagonal (D), strictly lower (T) and strictly upper (S) triangular parts, σ and ω are real parameters, $\omega \neq 0$.

2. CONVERGENCE OF THE AFC METHOD

In [2] we proved that the AFC method for solving system (1) can be written in the following form

3) $$x^{k+1} = Bx^k + d', \quad k = 0, 1, \ldots$$

where

$$B = \left(E + \frac{1}{n-m} MP\right) M\left(E - \frac{1}{n}P\right), \quad d' = \left(E + \frac{1}{n-m}MP\right)d,$$

$M = [m_{i,j}] \in \mathbb{R}^{n,n}$ is the iterative matrix of the basic method, $m = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}$, P is the nxn matrix all entries of which are equal to 1 and E is identity matrix. Also, we showed that

4) $$(B)_{ij} = m_{ij} - \frac{1}{n-m} m_i(1 - m_j^*), \quad i,j = 1, 2, \ldots, n,$$

where

$$m_i = \sum_{j=1}^{n} m_{ij}, \qquad m_i^* = \sum_{j=1}^{n} m_{ji}, \qquad i = 1, 2, \ldots, n.$$

Now, it is easy to see that AFC method converges if $\rho(B) < 1$.

3. AOR + AFC METHOD

AOR + AFC method has the matrix form (3), where $M = M_{\sigma,\omega}$. Some sufficient conditions for the convergence of this method we can obtain by analysing the condition $\rho(B) < 1$. So, we obtain the following theorem.

THEOREM 1. *Let* $M_{\sigma,\omega} \geq 0$, $L_1 = \|D^{-1}T\|_1$, $\|D^{-1}(T + S)\|_\infty < 1$, $U_1 = \|D^{-1}S\|_1$, $0 < L_1 < 1$, $L_1 + U_1 \leq 1$,

5) $$\begin{cases} 0 < \omega \leq 1, & -\dfrac{\omega(1 - L_1 - U_1)}{2L_1} \leq \sigma \leq \dfrac{\omega(1 + L_1 - U_1)}{2L_1} \quad or \\ 1 \leq \omega \leq \dfrac{2}{1 + L_1 + U_1}, & \dfrac{-2 + \omega(1 + L_1 + U_1)}{2L_1} \leq \sigma \leq \dfrac{2 - \omega(1 - L_1 + U_1)}{2L_1}, \end{cases}$$

and

(6.1) $$0 < \omega \le 1, \quad -\min_{1\le i\le n} \frac{1 - l_i - u_i}{2l_i} \le \sigma \le \min_{1\le i\le n} \frac{1 + l_i - u_i}{2l_i}$$

or

(6.2) $$1 < \omega < \frac{2}{1 + \max\limits_{1\le i\le n}(l_i + u_i)}, \quad \max_{1\le i\le n} \frac{\omega(1 + l_i + u_i) - 2}{2l_i} < \sigma < 0,$$

where $$l_i = \sum_{j=1}^{n} \left|(D^{-1}T)_{ij}\right|, \quad u_i = \sum_{j=1}^{n} \left|(D^{-1}S)_{ij}\right|, \quad i = 1,2, \dots, n.$$

Then AOR + AFC method converges for any start vector.

P r o o f. If ω,σ are choosen as in (6), then $\|M_{\sigma,\omega}\|_\infty < 1$, (see [3]). From (5) we shall prove that $\|M_{\sigma,\omega}\|_1 \le 1$. Obviously,

$$\|M_{\sigma,\omega}\|_1 \le \|(D - \sigma T)^{-1}\|_1 \|(1 - \omega)D + (\omega - \sigma)T + \omega S\|_1,$$

$$\|M_{\sigma,\omega}\|_1 \le \|(E - \sigma L)^{-1}\|_1 \|(1 - \omega)E + (\omega - \sigma)L + \omega U\|_1.$$

If $|\sigma|L_1 < 1$, we have

$$\|M_{\sigma,\omega}\|_1 \le \frac{1}{1 - |\sigma|L_1}\left(|1 - \omega| + |\omega - \sigma|L_1 + |\omega|U_1\right).$$

If σ and ω are choosen as in (5), then it can be verified that the two following conditions

$$|\sigma|L_1 < 1 \quad \text{and} \quad \frac{1}{1 - |\sigma|L_1}\left(|1 - \omega| + |\omega - \sigma|L_1 + |\omega|U_1\right) \le 1$$

are satisfied. Hence $\|M_{\sigma,\omega}\|_1 \le 1$. Because of that, since $M_{\sigma,\omega} \ge 0$, we have $m_i < 1$, $m_i^* \le 1$, $i = 1, 2, \dots, m$. Now, for the AOR + AFC matrix B it holds:

$$\sum_{s=1}^{n} \left|(B)_{is} - (B)_{js}\right| = \sum_{s=1}^{n} \left| m_{is} - \frac{1}{n-m} m_i(1 - m_s^*) - m_{js} + \frac{1}{n-m} m_j (1 - m_s^*)\right| \le$$

$$\leq \sum_{s=1}^{n} |m_{is} - m_{js}| + \frac{1}{n-m} |m_i - m_j| \sum_{s=1}^{n} (1 - m_s^*) =$$

$$= \sum_{s=1}^{n} |m_{is} - m_{js}| + |m_i - m_j| \leq m_i + m_j + |m_i - m_j| =$$

$$= 2 \max (m_i, m_j) < 2.$$

Matrix B has constant row sums equal to 0. Now, we construct the matrix $C = [c_{ij}] \in \mathbb{R}^{n,n}$ as follows:

$$c_{ij} = (B)_{ij} - b_{j,min}, \quad i, j = 1, 2, \ldots, n,$$

where

$$b_{j,min} = \min_{1 \leq i \leq n} (B)_{i,j}.$$

It follows that $C \geq 0$ and

$$\gamma = \sum_{j=1}^{n} c_{ij} = - \sum_{j=1}^{n} b_{j,min} \geq 0.$$

It holds $\gamma > 0$, except in the trivial case $B = 0$. Now, the matrix $\frac{1}{\gamma} C$ is a stochastic matrix, for which (see [6]) we know that h

$$\rho\left(\frac{1}{\gamma} C\right) \leq \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} \frac{1}{\gamma} |c_{is} - c_{js}|.$$

Now, it follows

$$\rho(C) \leq \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} |c_{is} - c_{js}| = \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} |(B)_{is} - b_{s,min} - (B)_{js} + b_{s,min}|$$

$$\rho(C) \leq \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} |(B)_{is} - (B)_{js}| < 1.$$

It remains to show that $\rho(B) \leq \rho(C)$. Let λ denote an eigenvalue of the matrix B^T, $\lambda \neq 0$, and let y be a corresponding eigenvector. For i = 1, 2, ...,n,

we have

$$(B^T y)_i = \sum_{j=1}^{n} (B)_{ji} y_j = \lambda y_i$$

and

$$\lambda \sum_{i=1}^{n} y_i = \sum_{j=1}^{n} y_j \sum_{i=1}^{n} (B)_{ji} = 0.$$

Since $\lambda \neq 0$, we obtain $\sum_{i=1}^{n} y_i = 0$. Using this, we have for $i = 1, 2, \ldots, n$,

$$(B^T y)_i = \sum_{j=1}^{n} (B)_{ji} y_j = \sum_{j=1}^{n} (B)_{ji} y_j - b_{i,min} \sum_{j=1}^{n} y_j = \sum_{j=1}^{n} \left((B)_{ji} - b_{i,min} \right) y_j = (C^T y)_i.$$

So, λ is an eigenvalue of the matrix C^T, too. Since B and B^T, as well as C and C^T have the same eigenvalues, we can conclude that all eigenvalues of the matrix B (except, might be, $\lambda = 0$) are eigenvalues of the matrix C. Hence, $\rho(B) \leq \rho(C) < 1$, which completes the proof. □

Some of the conditions from Theorem 1 are very restrictive. For example, the absolute and maximum norm of the Jacobi matrix ($M_{0,1}$) have to be less than 1. The condition $M_{\sigma,\omega} \geq 0$ is satisfied, for example, when A is an L-matrix and $0 \leq \sigma \leq \omega \leq 1$, while intervals for σ and ω, given by (5) and (6) are always nonempty (always it is possible to choose $\sigma = 0$, $\omega = 1$). The convergence intervals given in this theorem are always wider than the ones from [1].

From the above discussion we can not conclude when the combined AOR+AFC method converges faster than the basic AOR method. But, numerical examples show that in some cases the AOR+AFC method has better convergence than the AOR method. So, for the simple example 3 the graphs of Figure 2 give the behaviour of the actual error as a function of the iteration k for both AOR and AOR+AFC methods.

4. NUMERICAL EXAMPLES

Example 1. We consider a system of linear equations with the matrix

$$A = \begin{bmatrix} -0.875 & 0.05 & 0.0125 & 0.0125 & 0.0625 & 0.0624 \\ 0.024 & -0.75 & 0.0125 & 0.125 & 0.00624 & 0.01325 \\ 0.12 & 0.0125 & -0.875 & 0.0625 & 0.05 & 0.0625 \\ 0.0125 & 0.12 & 0.0125 & -0.5 & 0.0625 & 0.0624 \\ 0.00625 & 0 & 0.0025 & 0.0625 & -0.9375 & 0.0049 \\ 0.00327 & 0 & 0.024 & 0.0124 & 0.025 & -0.837 \end{bmatrix}$$

The convergence area for the parameters of AOR + AFC method is given in the figure 1.

Example 2. In the Table 1 we can see that AOR + AFC method converges even if the basic method diverges, as well as the convergence is very fast, where the matrix of linear system is:

$$A = \begin{bmatrix} -0.875 & 0.5 & 0.125 & 0.125 & 0.0625 & 0.0624 \\ 0.24 & -0.75 & 0.125 & 0.125 & 0.0624 & 0.1325 \\ 0.12 & 0.125 & -0.875 & 0.0625 & 0.5 & 0.0625 \\ 0.125 & 0.12 & 0.125 & -0.5 & 0.0625 & 0.0624 \\ 0.625 & 0 & 0.25 & 0.0625 & -0.9375 & 0.49 \\ 0.327 & -0.001 & 0.24 & 0.0124 & 0.25 & -0.837 \end{bmatrix}$$

Table 1.

(σ,ω)	$\rho(M_{\sigma,\omega})$	$\rho(B)$
(0,1)	1.070	0.531
(1,1)	1.128	0.650
(0.255,0.882)	1.070	0.252

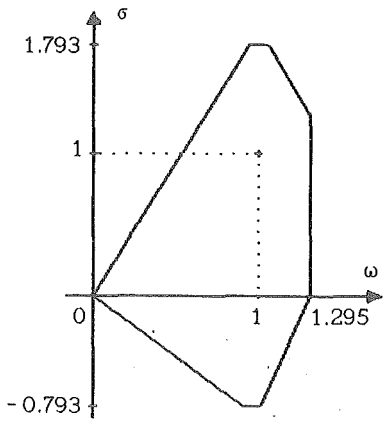


Figure 1.

Example 3. Let

$$A = \begin{bmatrix} -0.8 & 0.22 & 0.26 & 0.24 \\ 1 & -1.02 & -0.06 & -0.04 \\ 0.5 & 0.4 & -1.03 & -0.02 \\ 0.5 & -0.5 & 1 & -0.98 \end{bmatrix} \qquad b = \begin{bmatrix} 0.42 \\ -0.22 \\ -1.20 \\ 0.40 \end{bmatrix}$$

In the following Figure 2 we present the value -logE as a function of iteration k, where

$$E := \frac{\|x - x^k\|_\infty}{\|x\|_\infty},$$

and $x = [1,1,2,2]^T$ is the exact solution of the system $Ax = b$, and x^k is the k-th iteration obrtained by AOR or AOR + AFC method with $x^0 = [100,0,0,-100]$.

The graphs are denoted as follows:

1 - Jacobi method;

2 - Gauss-Seidel method;

3 - AOR method with $\sigma = 0.9875$ and $\omega = 1.27$;

4 - AOR method with $\sigma = 0.972$ and $\omega = 0.965$;

5 - Jacobi + AFC method;

6 - Gauss-Seidel + AFC method;

7 - AOR + AFC method with $\sigma = 0.9875$ and $\omega = 1.27$;

8 - AOR + AFC method with $\sigma = 0.972$ and $\omega = 0.965$.

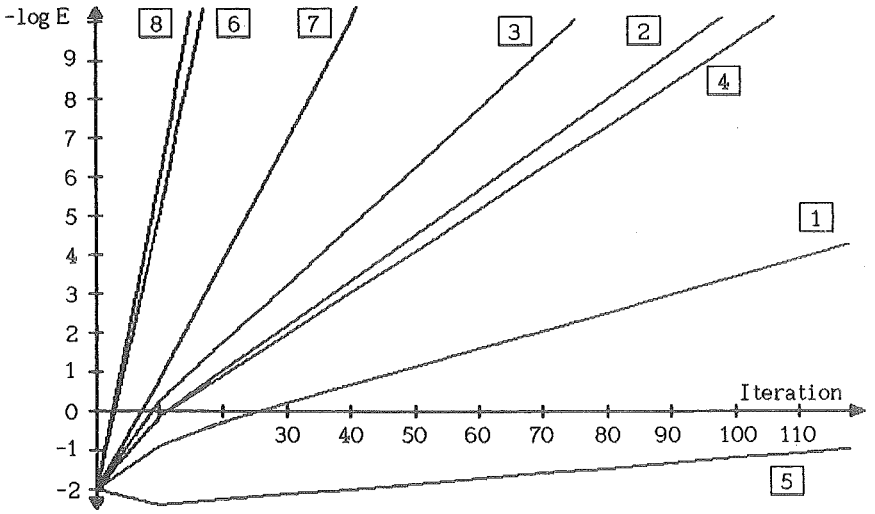


Figure 2.

Acknowledgments. We would like to thank the referee for having improved the presentation of the paper.

REFERENCES

[1] Lj. Cvetković, D. Herceg: Eine Modifikation des AOR-Verfahrens, Z. Angew. Math. Mech. Bd. 67,N.5(1987), 913-914.

[2] Lj. Cvetković, D. Herceg: On the method of averaging functional corrections applied to linear systems, Zb. Rad. Prir.Mat.Fak. Ser.Mat. 17,2(1987) (in print).

[3] Lj. Cvetković, D. Herceg: An improvement for the area of convergence of the AOR method, Anal. Numer. Theor. Approx. 16(1987), 109-115.

[4] A. Hadjidimos: Accelerated overrelaxation method, Math. Comp., 32(1978), 149-157.

[5] В.Х. Сиренко: О численной реализации метода осреднения функциональных поправок, УМЖ 13(1961), 51-66.

[6] Ch. Zenger, A comparation of some bounds for the nontrivial eigenvalues of stochastic matrices, Numer. Math. 19(1972), 209-211.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

ON THE EFFICIENCY OF ITERATIVE METHODS FOR BOUNDING THE INVERSE MATRIX

J. HERZBERGER

ABSTRACT: *In this note we are considering the higher-order interval Schulz methods for improving bounds for the inverse matrix. First we give a different computation scheme for the iteration formula which is more efficient especially for the higher-order formulas. Next, we derive a modification of the methods which has the same properties as the original ones but compares favourably for the higher-order cases. For both versions presented here some efficiency indices are listed and compared with those of the original formulas.*

0. INTRODUCTION

Let $\underset{\cdot}{A}$ be an m×m nonsingular real matrix and $X^{(0)}$ be an m×m interval matrix with $\underset{\cdot}{A}^{-1} \in X^{(0)}$. In [1], Chapter 18 there are described iteration methods which improve $X^{(0)}$ iteratively. These formulas are the following ones:

$$(1) \qquad X^{(n+1)} = m(X^{(n)}) \sum_{i=0}^{r} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^i + X^{(n)} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^{r+1} ,$$

$$(2) \qquad X^{(n+1)} = \{m(X^{(n)}) \sum_{i=0}^{r} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^i + X^{(n)} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^{r+1}\} \cap X^{(n)} .$$

where

$$\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)})$$

and

$$m(X) = m(([x_{ij}^1, x_{ij}^2])) = ((x_{ij}^1 + x_{ij}^2)/2)$$

$(r \geq 0)$.

In Theorem 1 and Theorem 2 of Chapter 18 in [1] it is shown that for the methods (1) and (2)

$$\underset{\cdot}{A}^{-1} \in \underset{\cdot}{X}^{(k)} \quad , \quad k \geq 0$$

holds true. Furthermore, for the R-order of convergence (see [1]) we have the estimations

$$O_R((1),\underset{\cdot}{A}^{-1}) \geq r+2 \quad \text{and} \quad O_R((2),\underset{\cdot}{A}^{-1}) \geq r+2 \ .$$

Since the convergence criterion of (1) is weaker than that of (2), we usually start with (1) and after some contracting iterations switch to method (2) as soon as its convergence criterion is fulfilled. Then method (2) produces a nested sequence of inclusions for $\underset{\cdot}{A}^{-1}$ and thus allows to establish a quite natural stopping rule. For more details see [1], Chapter 18. Formulas (1) and (2) are computed by means of the Horner scheme and require $r+2$ matrix multiplications each of them. Now, the efficiency index (see [4] Appendix C) E_H can be estimated by

$$E_H \geq (r+2)^{\frac{1}{r+2}} \ .$$

1. MODIFIED SCHEMES

We consider the iteration formulas

$$(3) \qquad \underset{\cdot}{X}^{(n+1)} = m(\underset{\cdot}{X}^{(n)}) \prod_{j=0}^{k-1} \Big(\sum_{i=0}^{r} (\underset{\cdot}{T}^{(n)})^{i(r+1)^j} \Big) + \underset{\cdot}{X}^{(n)} (\underset{\cdot}{T}^{(n)})^{(r+1)^k}$$

and

$$(4) \qquad \underset{\cdot}{X}^{(n+1)} = \Big\{ m(\underset{\cdot}{X}^{(n)}) \prod_{j=0}^{k-1} \Big(\sum_{i=0}^{r} (\underset{\cdot}{T}^{(n)})^{i(r+1)^j} \Big) + \underset{\cdot}{X}^{(n)} (\underset{\cdot}{T}^{(n)})^{(r+1)^k} \Big\} \cap \underset{\cdot}{X}^{(n)}$$

($k \geq 1$, $r \geq 0$). Setting $k=1$ in (3) and (4) we formally get the methods (1) and (2) as special cases. On the other hand, by virtue of the equality for real matrices $\underset{\cdot}{C}$

$$\prod_{j=0}^{k-1} \Big(\sum_{i=0}^{r} \underset{\cdot}{C}^{i(r+1)^j} \Big) = \sum_{i=0}^{(r+1)^k-1} \underset{\cdot}{C}^{i}$$

which can be proved by complete induction using a proper rearrangement of the summation terms, we get formulas equivalent to (3) and (4) by

$$(3)' \qquad X^{(n+1)} = m(X^{(n)}) \sum_{i=0}^{(r+1)^k-1} (\dot{T}^{(n)})^i + X^{(n)} (\dot{T}^{(n)})^{(r+1)^k} ,$$

$$(4)' \qquad X^{(n+1)} = \{m(X^{(n)}) \sum_{i=0}^{(r+1)^k-1} (\dot{T}^{(n)})^i + X^{(n)} (\dot{T}^{(n)})^{(r+1)^k}\} \cap X^{(n)} .$$

This shows that (3) and (4) are also just methods of the kind of (1) and (2) and therefore all have the same properties. In particular the R-order of convergence is

$$O_R((3), \dot{A}^{-1}) \geq (r+1)^k + 1 \quad \text{and} \quad O_R((4), \dot{A}^{-1}) \geq (r+1)^k + 1 .$$

Again we measure the amount of work by the necessary matrix multiplications which count exactly $k(r + (1 - \delta_{1k})\varphi(r+1)) + 2$ when using the Horner scheme for the occuring matrix polynomial factors. Here $\varphi(u)$ denotes the number of multiplications required for computing the u-th power. So, we get for the efficiency index E_M the estimation

$$E_M \geq ((r+1)^k+1)^{\frac{1}{k(r+(1-\delta_{1k})\varphi(r+1))+2}} .$$

In comparison to this the efficiency index for the original formulas (1) and (2) of the corresponding order was

$$E_H \geq ((r+1)^k+1)^{\frac{1}{(r+1)^k+1}} .$$

The following tables show for some selected values of the parameters r and k bounds for the efficiency indices.

r=1	k=1	2	3	4	5	6
E_H	1.442	1.379	1.277	1.181	1.112	1.066
E_M	1.442	1.308	1.316	1.328	1.338	1.347

r=2	k=1	2	3
E_H	1.414	1.259	1.126
E_M	1.414	1.259	1.269

Remarks: As the tables show, the modifications (3) and (4) are considerably more efficient for greater values of parameters or - with other words - for higher orders.
The bounds for the efficiency indices E_H and E_M achieve its maximum $\sqrt[3]{3}$ for the formulas of order three as an easy analysis shows.

2. MODIFIED METHODS

Now, we consider the iteration methods

$$y^{(n+1,0)} = m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n)} (I - Am(X^{(n)}))^r ,$$

$$y^{(n+1,i+1)} = m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n+1,i)} (I - Am(X^{(n)}))^r , \tag{5}$$

$$X^{(n+1)} = y^{(n+1,s)} , \qquad 0 \leq i < s ,$$

and

$$y^{(n+1,0)} = \{m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n)} (I - Am(X^{(n)}))^r\} \cap X^{(n)} ,$$

$$y^{(n+1,i+1)} = \{m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n+1,i)} (I - Am(X^{(n)}))^r\} \cap y^{(n+1,i)} , \tag{6}$$

$$X^{(n+1)} = y^{(n+1,s)} , \qquad 0 \leq i < s ,$$

where $r \geq 1$ and $s \geq 0$. (In case $s = 0$ the second statements of the iteration formulas are to be empty.) Setting $s = 0$ we again get the methods (1) and (2) as special cases.

A simple backward substitution of the quantities $y^{(n,i)}$ in (5) leads to the equivalent formula

$$(5)' \qquad X^{(n+1)} = m(X^{(n)}) \sum_{i=0}^{(s+1)r-1} (I - A\,m(X^{(n)}))^i + \\ + (\ldots((X^{(n)}(I - A\,m(X^{(n)}))^r)(I - A\,m(X^{(n)}))^r)\ldots$$

which is again a method like (1). Such a transformation is, however, not possible for (6).

From the equality

$$m(X^{(n)}) \sum_{i=0}^{r-1} (I - A\,m(X^{(n)}))^i = A^{-1} - A^{-1}(I - A\,m(X^{(n)}))^r$$

together with the inclusion monotonicity of the interval operations we get for (5) and (6) by complete induction the property

$$A^{-1} \in X^{(n+1)}, \quad y^{(n+1,i)}, \quad (0 \leq i \leq s), \quad n \geq 0 .$$

Similarly like in [1] Chapter 18 we can prove by a straight forward analysis the same convergence criteria for (5) and (6) as for (1) and (2). As for the R-order of convergence, we immediately get from the representation (5)' of (5)

$$O_R((5), A^{-1}) \geq (s+1)r + 1 .$$

By a similar analysis for the sequences $\{d(X^{(n)})\}$, where d is the width operator, we get in addition to this estimation

$$O_R((6), A^{-1}) \geq (s+1)r + 1 .$$

The amount of work in terms of matrix multiplications is in case of (5) or (6) $r + s + (1 - \delta_{0,s})\varphi(r) + 1$. Thus we get for the efficiency index E_{MM} the estimation

$$E_{MM} \geq ((s+1)r + 1)^{\frac{1}{r+s+(1-\delta_{0,s})\varphi(r)+1}} = \alpha(r,s)$$

in contrast to the corresponding efficiency index E_H

$$E_H \geq ((s+1)r+1)^{\frac{1}{(s+1)r+1}} = \beta(r,s) .$$

It is easy to see that the inequality

$$\beta(r,s) \leq \alpha(r,s)$$

holds true. This means that the unmodified methods are not so efficient as the modified methods.
The following tables give the bounds for E_{MM} and E_H for some selected values of parameters r and s.

r=2	s=1	2	3	4	5
E_H	1.380	1.320	1.277	1.244	1.218
E_{MM}	1.380	1.383	1.367	1.350	1.330

r=3	s=1	2	3	4	5
E_H	1.320	1.259	1.218	1.189	1.168
E_{MM}	1.320	1.344	1.330	1.320	1.307

Remarks: The bounds for the efficiency index E_{MM} achieve its maximum $\sqrt[3]{3}$ for s=0, r=2 or r=s=1 with methods of order three. A direct comparison between the given bounds for E_M and E_{MM} is not possible because the orders of convergence of the produced methods for different values of parameters do not coincide.

REFERENCES

1. G. ALEFELD and J. HERZBERGER: Introduction to Interval Computations. Academic Press, New York 1983.

2. J. HERZBERGER: Some aspects of iterative methods for bounding the inverse matrix. Colloquia Mathematica Societas János Bolyai, 50. Numerical Methods, 1987, 185-199.

3. J. HERZBERGER: Zur Effizienz von intervallmäßigen Schulz-Verfahren höherer Ordnung. Z. angew. Math. Mech. (to appear).

4. J.F. TRAUB: Iterative methods for the solution of equations. Prentice-Hall, Englewood Cliffs N.J. 1964

PROCESS IDENTIFICATION USING B-SPLINES

LJ. M. KOCIĆ and B. DANKOVIĆ

Abstract: An application of B-splines in computation of Laplace transform of the unit step response (and then the transfer function) of a given system of automatic control is considered. The algorithm suggested is based upon de Boor-Cox algorithm for numerically stable calculation of B-splines. Error estimation is given, and a numerical experiment is performed.

1. INTRODUCTION

Suppose that we are given the system (process) of automatic control in a "black box" form (Fig. 1). To identify this process means to find its transfer function $W(s)=Y(s)/X(s)$, where $s \mapsto X(s)$ $(s \mapsto Y(s))$ is the Laplace transform of input (output) signal as function of time: $t \mapsto x(t)$ $(t \mapsto y(t))$. Namely, we shall write $X(s)=L(x(t))=\int_0^\infty e^{-st}x(t)dt$, where we supposed $x, y \in L_2[0, \infty)$, and $x(t)=y(t)=0$ for $t<0$, which is satisfied in a great number of practical situations. When $x(t)$ is the unit step function, we have $W(s)=sY(s)=sL(y(t))$, where $y(t)$ will be regarded as the unit step response of the "black box", and we can gather the information on $y(t)$ only by measuring it in some discrete set of points $\tau=\{\tau_0, \tau_1, \ldots, \tau_N\}$ $(\tau_i<\tau_{i+1}$, $i=0, \ldots, N-1)$. As the result, we should get the set of data $d=\{y_i=y(t_i)\}_{i=1}^N$. Based on τ and d we can calculate the function $y^*(t)$ which approximate $y(t)$ on $[\tau_0, \tau_N]$ so that $\|y-y^*\|<E_N$, where E_N is the prescribed error of appro-

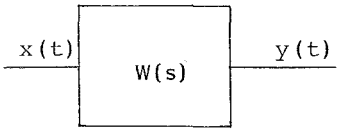


Fig. 1.

ximation and $\|\cdot\|$ is one of the usual norms, taken over the interval $[\tau_0, \tau_N]$. Now, we can find $Y^*(s) = L(y^*(t))$ and then $W^*(s) = sY^*(s)$. But, we must pay attention to an important detail. The unit step response $y(t)$, always (for real systems) approaches to a fixed value, say, y^∞, after a long enough interval of time. It is convenient to suppose that $y^*(t)$ approximates $y(t)$ on $[\tau_0, \tau_N]$ and $y^*(t) = 0$ outside, so with $y^*(t) + y^\infty$ we have the approximation to $y(t)$ completed. Then, we can put $Y^*(s) = L[y^*(t) + y^\infty] = L[y^*(t)] + s^{-1}y^\infty$, and therefore $W^*(s) = sY^*(s) = s\,L[y^*(t)] + y^\infty$. Since $W(s)$ has the similar form, namely $W(s) = sL[y(t)] + y^\infty$, the error (see section 3) will not contain y^∞.

So, the problem is to find the approximation y^* for the unit step response y which has to be "good" in the following sense:

$$\|y^* - y\| \to \min \; ; \quad \left\| \frac{d}{dt} y^* - \frac{d}{dt} y \right\| \to \min .$$

This two requests arises in natural way in the theory of adaptive processes (see for example, [10]).

In this paper we investigate the most convenient way to use polynomial splines in order to compute $L[y^*(t)]$ and to estimate the error $|W(s) - W^*(s)|$.

Let $S_{k,\xi}$ be the space of polynomial splines of order k, with the knot sequence $\xi = \{\xi_1, \xi_2, \ldots, \xi_{n+1}\}$ (which is strictly increasing one). Then, if, for example, $y^* \in S_{k,\xi}$, we have representation via truncated power basis

$$y^*(t) = \sum_{j=0}^{k-1} a_j (t - \xi_1)^j + \sum_{i=2}^{n} b_i (t - \xi_i)_+^{k-1} ,$$

where

$$a_j = \frac{y^*(\xi_1 + 0)^{(j)}}{j!} , \quad b_j = \frac{D^{k-1}y^*(\xi_i + 0) - D^{k-1}y^*(\xi_i - 0)}{(k-1)!} .$$

This representation is very convenient for applying L-transform, but, unfortunately, very unstable for numerical calculations (see [1]). So, we will turn to B-splines which provide very stable numerical process.

Some of known methods (for example the Aizerman's method) use piecewise constant approximation of y. In terms of splines, this means that $y^* \in S_{1,\xi}$ (see [7]). In [3], the function y is approximated by parabolic segments performed to fit the set of data $\mathbf{d}$. Of course, such interpolant, y^* sufferes from low smootheness.

2. B-SPLINE AND ITS L-TRANSFORM

As it is already known, the space $S_{k,\xi}$ has so called B-spline base $\{ B_{i,k} \}_{i=1}^{n}$, where $B_{i,k}$ is defined for the set of knots $t = \{ t_1, \ldots, t_{n+k} \}$ which can be derived from ξ by adding $2k$ new knots $t_1 = \ldots = t_k = \xi_1$ and $t_{n+1} = \ldots = t_{n+k} = \xi_{n+1}$ and so that $\xi_i = t_i$ $(i=k+1,\ldots,n)$. Then, the i-th B-spline of order k is given by

$$(1) \quad B_{i,k}(t) = (t_{i+k} - t_i)[t_i, \ldots, t_{i+k}](\cdot - t)_+^{k-1} ,$$

for $i = 1, \ldots, n$. Now, for $y^* \in S_{k,t}$ we have

$$(2) \qquad y^*(t) = \sum_{i=1}^{n} c_i B_{i,k}(t).$$

Computation of the coefficients c_i depands on approximation scheme we want to use. For example, we can interpolate the data $\mathbf{d}$ in the nodes τ. If we have the freedom of choosing the nodes τ_i it is advisable to take

$$\tau_i = \frac{1}{k-1} (t_{i+1} + \ldots + t_{i+k-1}).$$

We also can calculate c_i in order to smooth the data $\mathbf{d}$. According to [1] we can do that by minimizing the quantity

$$p \sum_{i=1}^{N} \left(\frac{y_i - y^*(\tau_i)}{\delta y_i} \right)^2 + (1-p) \int_{\tau_1}^{\tau_N} (D^m y^*(t))^2 \, dt$$

where δy_i is an estimate of the variance in y_i, and $p \in [0,1]$ is a given parameter. The role of p is to emphasize the closeness to data (when $p \to 1_-$) or smootheness of y^* (when $p \to 0_+$). In this way, we get so called Whittaker spline, and the corresponding package SMOOTH is given in [1]. By the another procedure from [1], named L2APPR, we can calculate c_i from (2), and the spline y^* is then the approximation of data in the sense of least squares.

Another interesting procedures for smoothong data via splines from $S_{k,t}$ can be found in [4] and [5].

Suppose that we have all c_i $(i=1,\ldots,n)$ calculated, so we have y^* completely defined. Now, if we apply the L operator on both sides of (2), we shall get

$$(3) \qquad L(y^*(t)) = \sum_{i=1}^{n} c_i \, L(B_{i,k}(t)) = \sum_{i=1}^{n} c_i \, L_{i,k}(s).$$

So, we must calculate $L_{i,k}(s) = L(B_{i,k}(t))$ $(i=1,\ldots,n)$, and we have to do that in the most efficient way. For example, we do not recommend using the explicit formula

$$B_{i,k}(t) = (t_{i+k} - t_i) \sum_{j=i}^{i+k} \frac{(t_j - t)_+^{k-1}}{\pi'_{k,i}(t_j)} \quad , \qquad \pi_{k,i}(t) = \prod_{j=i}^{i+k} (t - t_j)$$

from the reason of its low accuracy. Instead of that, we shall start from de Boor-Cox algorithm for stable calculation of B-splines ([1] , [2]):

$$(4) \qquad B_{i,1}(t) = \begin{cases} 1 \, , & t_i \leq t < t_{i+1} \, , \\ 0 \, , & \text{otherwise} \, , \end{cases}$$

$$(5) \qquad B_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t).$$

Can we find the similar reccurence formula for $L_{i,k}(s)$? The answer is affirmative. Namely, from (4) we can compute directly

$$(6) \qquad L_{i,1}(s) = \frac{1}{s}\left(e^{-t_i s} - e^{-t_{i+1} s} \right), \quad i=1,\dots,n.$$

Now, we can use the definition relation (1). Owing to the obvious identity

$$(t - x)_+^{k-1} = (t - x)^{k-1} + (-1)^k (x - t)_+^{k-1} ,$$

and the fact that $[t_i,\dots,t_{i+k}]\,(t - \cdot)^{k-1} = 0$, we have

$$(7) \qquad B_{i,k}(t) = (-1)^k (t_{i+k} - t_i)\,[t_i,\dots,t_{i+k}]\,(t - \cdot)_+^{k-1} .$$

The point, named "placeholder", states instead of the variable which the divided difference is applying on. If we apply the L-operator on both sides of (7), we get

$$(8) \qquad L_{i,k}(s) = (-1)^k (t_{i+k} - t_i)\,[t_i,\dots,t_{i+k}]\{L(t - \cdot)_+^{k-1}\}$$

$$= \frac{(k-1)!}{s^k} (-1)^k (t_{i+k} - t_i)[t_i,\dots,t_{i+k}] e^{-s(\cdot)} .$$

The authors of [8] have derived the same formula beginning from Schoenberg's identity.

Now, due to the recurrence relation for divided differences we have

$$L_{i,k}(s) = \frac{(k-1)!}{s^k}(-1)^k (t_{i+k} - t_i)\, x$$

$$x \frac{[t_{i+1},\dots,t_i] e^{-s(\cdot)} - [t_i,\dots,t_{i+k-1}] e^{-s(\cdot)}}{t_{i+k} - t_i}$$

$$= \frac{k-1}{s} \left\{ \frac{1}{t_{i+k-1} - t_i} \frac{(k-2)!}{s^{k-1}} (-1)^k (t_{i+k-1} - t_i)[t_i,\dots,t_{i+k-1}] e^{-s(\cdot)} \right.$$

$$\left. - \frac{1}{t_{i+k} - t_{i+1}} \frac{(k-2)!}{s^{k-1}} (-1)^{k-1} (t_{i+k} - t_{i+1})[t_{i+1},\dots,t_{i+k}] e^{-s(\cdot)} \right\}$$

and thus

$$(9)\quad L_{i,k}(s) = \frac{k-1}{s}\left(\frac{L_{i,k-1}(s)}{t_{i+k-1} - t_i} - \frac{L_{i+1,k-1}(s)}{t_{i+k} - t_{i+1}}\right) \quad i=1,\ldots,n,\ k \geq 2$$

where we have taken into account (8). So, the set $\{L_{i,k}(s)\}_{i=1}^{n}$ is completely defined by (6) and (9). This commpletes the procedure of finding $Y^*(s)$ (see (3)) and then $W^*(s)$ as well.

We must underline that the computation of (9) can also become unstable for small $|s|$, and then we recommand technique given in [6]. For further study of L-transform of B-splines, see [9].

3. ERROR ESTIMATION

The distance between $W(s)$ and $W^*(s)$ in the complex plane $s = \sigma + j\omega$ is given by

$$|W(s) - W^*(s)| = |s|\,|Y(s)-Y^*(s)| = |s|\left|\int_0^{+\infty}\{y(t)-y^*(t)\}e^{-st}dt\right|$$

$$\leq |s| \int_{\tau_0}^{\tau_N} |y(t)-y^*(t)|\, e^{-\sigma t}\,dt \leq |s|\, E_N \int_{\tau_0}^{\tau_N} e^{-\sigma t}\,dt\ ,$$

so we have

$$|W(s) - W^*(s)| \leq \begin{cases} \left[1 + \left(\frac{\omega}{\sigma}\right)^2\right]^{1/2}\left(e^{-\sigma\tau_0} - e^{-\sigma\tau_N}\right)\cdot E_N\ , & \sigma \neq 0, \\ \left[\sigma^2 + \omega^2\right]^{1/2}(\tau_N - \tau_0)\cdot E_N\ , & \sigma = 0. \end{cases}$$

The case $\sigma = 0$ is of especially importance in analysis of stability of automatic control systems. The E_N is the C-norm error of spline approximation. For example, if y^* is a cubic spline interpolant for the data d, we have

$$E_N \leq \frac{5}{385}|\tau|^4 \max_{[\tau_0,\tau_N]}\left|y^{(4)}(t)\right|\ ,\ |\tau| = \max_i \Delta\tau_i$$

and so on.

4. EXAMPLE

For a test-example we can take an ideal system with the unit step response $y(t) = 1 - e^{-t}$ (Fig.2). The graph of its tran-

sfer function, in the case $\sigma = 0$ is a half-circle. It is represented in (P, Q)-plane, where $P + jQ = W(j\omega)$ (Fig. 3). The cubic spline function $y^*(t)$ that interpolates $y(t)$ in the nodes (0., 2., 4., 6.) is constructed and its graph is shown in Fig. 4. The equidistant set of nodes we used is the worst choice we can made. This results in oscilations of y^*. However, the resulting graph of $W^*(j\omega) = P^*(\omega) + jQ^*(\omega)$, shown in Fig.4, has very satisfied form.

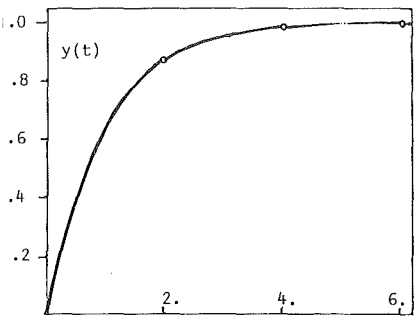


Fig. 2. Ideal system step response $y(t)$, $0. < t < 6.$;

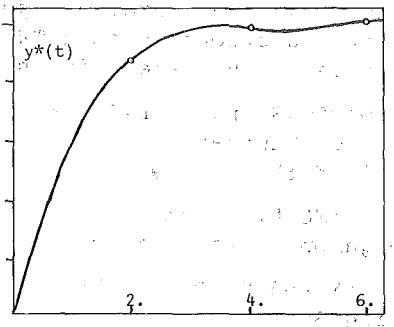


Fig. 3. Cubic spline $y^*(t)$ with the nodes 0.,2.,4.,6.;

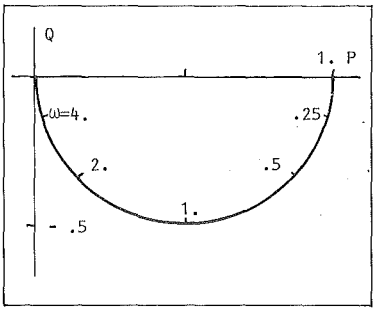


Fig. 4. Transfer function $W(j\omega)$ for the ideal system;

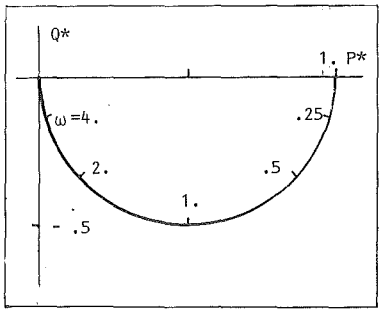


Fig. 5. Approximation of the transfer functiom based on spline approximation y^*.

The authors are grateful to the referees and professor G.V. Milovanović for a number of valuable suggestions which led to improvement of this paper.

REFERENCES

1. C. de BOOR: A Practical Guide to Splines. Springer-Verlag, New York, 1978.
2. M. G. Cox: The Numerical Evaluation of B-Splines. J. Inst. applics., 10 (1972), 132-149.
3. B. Danković, D. Ignjatović: Korišćenje računara pri identifikaciji tehnoloških procesa. In: Informacijski sistemi zasnovani na primeni računara, Zbornik radova, Niš 1985,69-76.
4. P. DIERCXS: An algorithm for smoothing, differentiation and integration of experimental data using spline functions. J. Comput. Appl. Math., 1 (1975), no. 3, 165-184.
5. P. DIERCXS: An Improved Algorithm for Curve Fitting with Spline Functions. Report TW54, July 1981, Dept. of Computer Science, Katholieke Universitet, Leuven (Belgium).
6. P. DIERCXS and R. PIESSENS: Calculation of Fourier Coefficients of Discrete Functions Using Cubic Splines. J. Comp. Appl. Maths. 3 (1977), 207-209.
7. P. EYKNOFF: Trends and progress in system identification. Pergamon Press, Oxford 1981.
8. M. LAX and G. P. AGRAWAL: Evaluation of Fourier Integrals Using B-Splines. Maths. Computation 39 (1982),no.160,535-548
9. W. SCHEMPP: Complex Contour Integral Representation of Cardinal Spline Functions. Contemporary Mathematics, Vol. 7, Amer. Math. Soc., Providence 1982.
10. V. TARAN, S. BRUDNIK, J. KOFANOV: Matematicheskie voprosi avtomatizacii priozvodstvenyh processov. Vyshaya shkola, Moskva 1978.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

ON CALCULATING QUADRATIC B-SPLINES IN TWO VARIABLES*

J. KOZAK and M. LOKAR

ABSTRACT: *One of the encountered problems in practical use of multivariate splines is a stable and efficient evaluation of a spline given as a linear combination of B-splines. No generalization has been found for the well-known univariate recursion scheme. Thus the only way to compute the value of a spline is to compute the values of all B-splines incident at a given point. In the paper we propose a special scheme for calculating all quadratic B-splines (in two dimensions) incident at a point in a certain subregion of the original domain. Our discusion can be viewed as a refinement of the work done by Meyling ([7,8]). We show that our scheme requires minimal constant B-splines evaluations.*

1. Introduction

Multivariate (simplex) splines have attracted quite a lot of attention in the past ten years. However, the theoretical work was not so widely followed by practical applications as one might expect for such a powerfull and flexible tool. There are several reasons for this fact. Perhaps one of the main obstacles is algorithmic and computational complexity of the computer procedures. The purpose of this paper is to tackle one practical aspect in dealing with bivariate quadratic splines, i.e. an evaluation of a spline. Though this is the simplest nontrivial case, several computational problems will be revealed.

---

*Supported by Research Council of Slovenija.

In order to construct a bivariate spline we recall its basis function first. There are several ways to define a B-spline. Perhaps the most apparent is the geometric one ([1]). A bivariate B-spline of degree n-2 is given by

$$M(\underline{x}|\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^n) := \frac{\mathrm{vol}_{n-2}(\{\underline{v}\in\sigma:\underline{v}|_{\mathbb{R}^2}=\underline{x}\})}{vol_n(\sigma)},$$

independently of σ, where $\sigma := [\underline{v}^0,\underline{v}^1,\ldots,\underline{v}^n]$ is n-simplex in $\mathbb{R}^n$ such that

i) $\mathrm{vol}_n(\sigma) > 0$,

ii) $\underline{v}^i|_{\mathbb{R}^2} = \underline{x}^i$, *for* $i = 0,1,\ldots,n$.

In other words, the set $\{\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^n\}$ *contains the orthogonal projections of the vertices* $\underline{v}^i$ *onto* $\mathbb{R}^2$. *Quite clearly M is a piecewise polynomial function of total degree n-2. If* $\underline{x}^i$ *are in general position (no triple lies on a line) then*

$$M(\underline{x}|\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^n) \in C^{n-3}(\mathbb{R}^2).$$

Consider now a given domain $\Omega \subset \mathbb{R}^2$. *Let* Δ *be its triangulation,* $T := |\Delta|$ *and* $V := \{\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^N\}$ *the set of knots. A spline space is derived by the following procedure ([4],[6]): Each knot* $\underline{x}^j \in V$ *is pulled apart in*

$$\underline{x}^{j,0} := \underline{x}^j,\ \underline{x}^{j,1},\ldots,\underline{x}^{j,n-2}.$$

Further, for any triangle

$$\rho_j := [\underline{x}^{j_0},\underline{x}^{j_1},\underline{x}^{j_2}] \in \Delta,\ j_0 < j_1 < j_2,$$

a set C_{ρ_j} of $\binom{n}{2}$ B-splines supports $K_{j,r}$ is constructed. If we associate with every point $\underline{x}^{j}{}_{m,q}$ *an element* (m,q) *of the lattice* $(0,1,2) \times (0,1,\ldots,n)$, *then the knot sets* $K_{j,r}$ *can be identified with nondescending paths along grid lines from* $(0,0)$ *to* $(2,n-2)$.

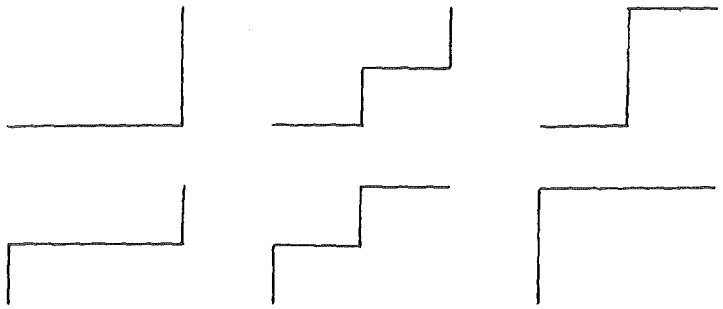


Fig. 1

In the quadratic case this would read

$$C_{\rho_j} = \{K_{j,r},\ r = 1,2,\ldots,6\}$$

where

$$\begin{aligned}
K_{j,1} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_1,0},\ \underline{x}^{j_2,0},\ \underline{x}^{j_2,1},\ \underline{x}^{j_2,2}\},\\
K_{j,2} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_1,0},\ \underline{x}^{j_1,1},\ \underline{x}^{j_2,1},\ \underline{x}^{j_2,2}\},\\
K_{j,3} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_1,0},\ \underline{x}^{j_1,1},\ \underline{x}^{j_1,2},\ \underline{x}^{j_2,2}\},\\
K_{j,4} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_0,1},\ \underline{x}^{j_1,1},\ \underline{x}^{j_2,1},\ \underline{x}^{j_2,2}\},\\
K_{j,5} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_0,1},\ \underline{x}^{j_1,1},\ \underline{x}^{j_1,2},\ \underline{x}^{j_2,2}\},\\
K_{j,6} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_0,1},\ \underline{x}^{j_0,2},\ \underline{x}^{j_1,2},\ \underline{x}^{j_2,2}\}.
\end{aligned}$$

Let $C := \bigcup_{j=1}^{T} C_{\rho_j}$. *The B-splines*

$$M(\underline{x}|K_{j,r}),\ j = 1,2,\ldots,T;\quad r = 1,2,\ldots,\binom{n}{2}$$

are the basis functions of the spline space

$$S(C) := span\{M(\underline{x}|K_{j,r})\}$$

over C with

$$dim\ S(C) = \binom{n}{2} T$$

Thus any $s \in S(C)$ can be expressed as

$$s(\underline{x}) = \sum_{K \in C} c_k M(\underline{x} | K).$$

2. Evaluation of a spline

There is no known analog of the univariate algorithm that computes a value of a spline by repeatedly forming convex combinations of its B-spline coefficients. Thus the value

$$s(\underline{x}) = \sum_{K \in C} c_K M(\underline{x} | K).$$

can be computed only by computing all nonzero B-splines $M(\underline{x} | \underline{x}^0, \underline{x}^1, \ldots, \underline{x}^n)$ incident at a given $\underline{x}$. A far reaching application of Stokes theorem reveals that ([4],[9])

$$M(\underline{x} | \underline{x}^0, \underline{x}^1, \ldots, \underline{x}^n) = \frac{n}{n-2} \sum_{m=0}^{n} \lambda_m M(\underline{x} | \underline{x}^0, \underline{x}^1, \ldots, \underline{x}^{m-1}, \underline{x}^{m+1}, \ldots, \underline{x}^n), \qquad n > 2$$

where $\underline{x}$ is expressed as any affine combination of $\underline{x}^i$,

$$\sum_{m=0}^{n} \lambda_m \underline{x}^m = \underline{x}, \qquad \sum_{m=0}^{n} \lambda_m = 1.$$

Put

$$M(\underline{x} | \underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2) = \begin{cases} \dfrac{1}{\mathrm{vol}_2([\underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2])}, & \underline{x} \in int[\underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2], \\ 0 & \underline{x} \notin [\underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2] \end{cases}$$

and general degree B-spline can be computed by the previous recurrence relation, at least for the points that do not lie

on any of the mesh lines. A constant B-spline on its boundary must be defined on a slightly different way. A simple remedy is as follows: let $\underline{\gamma}$ be a direction that is not parallel to any of mesh lines. A constant B-spline at a boundary point $\underline{x}$ is defined by

$$M(\underline{x}|\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2) := M(\underline{x}+O(\underline{\gamma})|\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2)$$

where $O(\underline{\gamma})$ is a small perturbation in the direction $\underline{\gamma}$. In a special case when all the mesh points lie in general position, one can avoid this difficulty by stopping recurrence when $n=3$ since linear B-splines are in this case continuous ([8]).

Quite clear, λ_i have to be nonnegative in order to assure numerical stability. Further, computational complexity implies that as many λ_i as possible should be zero. Thus in practice the recurrence step for a B-spline with support $[K]$ reads

$$M(\underline{x}|K) = (n-2)\sum_{m=0}^{n}\lambda_m(\underline{x}|\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2)M(\underline{x}|K\setminus\{\underline{x}^{i}m\})$$

where λ_m are nonnegative barycentric coordinates of $\underline{x}$ in a triangle $[\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2] \subset K$. The choice of i_0, i_1, i_2 is in general not unique since $\underline{x}$ may belong to several triangle parts of $[K]$.

3. The quadratic case

Let us start by defining subregions of the region supported by C_{ρ_j} of a particular interest ([5],[7]), i.e.

$$B_{\rho_j} := \bigcup_{(q_0,q_1,q_2)\in Q}[\underline{x}^{j}0,q_0,\underline{x}^{j}1,q_1,\underline{x}^{j}2,q_2]$$

and

$$Q := \{(q_0,q_1,q_2)\in \mathbb{Z}_+^{3}: 0\leq q_0\leq q_1\leq q_2\leq k\}.$$

Here k denotes degree of a B-spline. One can show that only $\binom{k+2}{2}$ B-splines $M(\underline{x}|K)$, $K \in C$ are incident at a point $\underline{x} \in B_{\rho_j}$, all j. Here B_{ρ_j} is a subregion of ρ_j, a and all of these $\binom{k+2}{2}$ B-splines are constructed over ρ_j. On the other hand, at points $\underline{x} \notin B_{\rho_j}$, all j, the number of B-splines $I_{\underline{x}}(C)$ incident at $\underline{x}$ can be quite large ([7]). This is a consequence of the fact that B-splines from adjecent knot set configurations may overlap the triangle ρ_j. $I_{\underline{x}}(C)$ depends on the original triangulation as well as on the pulling-apart procedure. For k = 2 (and general knot position) the following statement can be proved

$$I_{\underline{x}}(C) = \binom{2+2}{2} = 6, \qquad \underline{x} \in B_{\rho_j} \text{ for some } j,$$

$$I(C) := \max_{\underline{x} \in \Omega} I_{\underline{x}}(C) > 6.$$

There is very little hope that a general strategy which minimizes the computational complexity can be found when $\underline{x} \notin B_{\rho_j}$, all j. We shall therefore restrict ourselves to the case $\underline{x} \in B_{\rho_j}$, for some j. If $\underline{x} \notin B_{\rho_j}$, all j, we shall assume that the recurrence is applied in a straightforward way, and no attempt is made to reduce the number of operations. If the steps in pulling-appart procedure are small, then

$$C \setminus \cup B_{\rho_j}$$

is small compared to

$$\cup B_{\rho_j}$$

and care for evaluation in B_{ρ_j} justified.

Let $\rho_j = [\underline{x}^{j_0}, \underline{x}^{j_1}, \underline{x}^{j_2}]$, with $j_0 < j_1 < j_2$, be a triangle in Δ. For calculating all 6 quadratic B-splines incident in B_{ρ_j}, we propose the recursion scheme, which involves common lower order B-splines. A complete recursion

forest is shown in Fig.2.

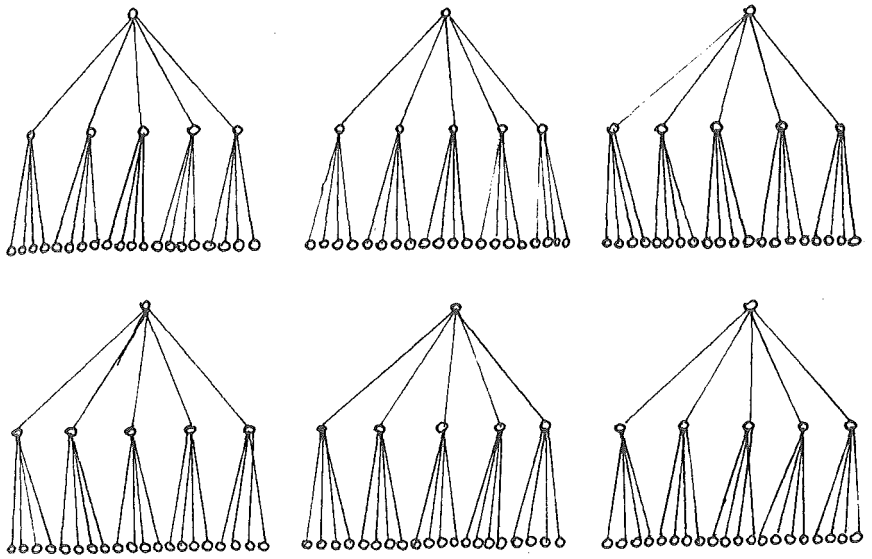


Fig. 2

As already pointed out there is no need to evaluate all of the tree. Even for general k there is enough to compute three tree knots at each tree level (except the root one).

Each quadratic spline is calculated from at most three linear splines that are chosen among five. Also at calculating linear splines we have four possible selections of three constant splines. Thus we have $(\binom{5}{3}\ 4^3)^6$ possible evaluation schemes.

In order to count necessary constant B-spline evaluations let us define a reduced knot set. Let $K \in C_{\rho_j}$ for some j, and denote by K^r the reduced knot set as any of its subsets of cardinality 5 - r. Quite clear, the supports of lower order B-splines in figure 2 are obtained as reduced sets of the root ones. Note also that the reduced set may belong to different tree knots at the same level.

All linear B-splines involved when evaluating the value of a quadratic spline are of type $M(\underline{x}|K^1)$, where K^1 is a reduced knot set consisting of four points. Further, all

constant splines that appear in the scheme are of type $M(\underline{x}|K^2)$ with K^2 a reduced knot set of cardinality 3. A straightforward calculation reveals that there are 24 different reduced sets of cardinality 4 and 37 of cardinality 3. As the full evaluation forest involves 30 sets with 4 points and 120 with 3 points, some reduced knot sets have to appear several times. In fact, an exact upper bound for necessary constant B-spline evaluations can be stated.

Theorem In order to evaluate all $\binom{k+2}{2}$ B-splines of degree k at the point $\underline{x} \in B_{\rho_j}$ it is necessary to compute at most

$$\varphi(k) := \frac{k+1}{3}(5k^2 + 7k + 3)$$

constant B-splines.

The proof is based upon carefull counting of the number of reduced knot sets of order 3. Note that

$$\varphi(0) = 1, \quad \varphi(1) = 10, \quad \varphi(2) = 37, \quad \varphi(3) = 92,$$

etc.

Number of appearence of sets with three points varies from 2 to 8. It is not obvious which reduced knot sets are to be choosen to minimise in general the overall scheme. A heuristic approach that uses more frequently appearing reduced sets can reduce the computational effort significiantly. However, by posing additional requirement on the pulling-apart procedure an optimal algorithm can be found.

If the knots $\underline{x}^{i,q}$, q = 1,2 are chosen in the polygon R_i (the convex polygon which contains $\underline{x}^i$ and is bounded, but not intersected, by the lines passing through the midpoints of any edges belonging to the same triangle with vertex $\underline{x}^i$) 9 linear B-splines vanish in B_{ρ_j}. Using this fact, in [8] an algorithm was presented that computes all 6 C^1 quadratic B-splines by evaluating 6 linear B-splines. The following scheme improves the result to five linear B-splines

evaluations, and all of these five can be computed from 10 constant B-splines. Let K_j^r denotes j-th reduced set at level r of the evaluation forest, counted from the left.

$$M(\underline{x}|K_{j,1}) = 2\lambda_2(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_2,0})M(\underline{x}|K^1_{15})$$

$$M(\underline{x}|K_{j,2}) = 2\,[\lambda_1(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_1,1})M(\underline{x}|K^1_{17}) + \lambda_2(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_1,1})M(\underline{x}|K^1_{15})$$

$$M(\underline{x}|K_{j,3}) = 2\lambda_1(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_2,2})M(\underline{x}|K^1_{16})$$

$$M(\underline{x}|K_{j,4}) = 2\,[\lambda_0(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,1},\underline{x}^{j_2,1})M(\underline{x}|K^1_{17}) + \lambda_2(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,1},\underline{x}^{j_2,1})M(\underline{x}|K^1_{5})\,]$$

$$M(\underline{x}|K_{j,5}) = 2\,[\lambda_0(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^1_{6}) + \lambda_1(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^1_{5})\,]$$

$$M(\underline{x}|K_{j,6}) = 2\lambda_0(\underline{x}|\underline{x}^{j_0,2},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^1_{6})$$

Furtner, linear splines are computed as

$$M(\underline{x}|K^1_6) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{16}) + \lambda_1(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{5}) + \lambda_2(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{3})$$

$$M(\underline{x}|K^1_5) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_0,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{5}) + \lambda_1(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_0,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{15}) + \lambda_2(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_0,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{24})\,]$$

$$M(\underline{x}|K^1_{17}) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_2,1},\underline{x}^{j_1,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{15}) + \lambda_1(\underline{x}|\underline{x}^{j_2,1},\underline{x}^{j_1,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{19}) + \lambda_2(\underline{x}|\underline{x}^{j_2,1},\underline{x}^{j_1,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{36})\,]$$

$$M(\underline{x}|K^1_{15}) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_1,0},\underline{x}^{j_2,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{19}) + \lambda_1(\underline{x}|\underline{x}^{j_1,0},\underline{x}^{j_2,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{12}) + \lambda_2(\underline{x}|\underline{x}^{j_1,0},\underline{x}^{j_2,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{34})\,]$$

$$M(\underline{x}|K^1_{16}) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{16}) + \lambda_1(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{15}) + \lambda_2(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{13})\,]$$

For simplicty let us abreviate $\underline{x}^{j_i,q}$ by iq. The sets K_i then read as follows

$K^1_5=\{00,01,11,22\}$ $K^1_6=\{00,01,12,22\}$ $K^1_{15}=\{00,10,21,22\}$
$K^1_{16}=\{00,11,12,22\}$ $K^1_{17}=\{00,11,21,22\}$

and

$K^2_3 = \{00,01,12\}$ $K^2_5 = \{00,01,22\}$ $K^2_{12} = \{00,10,22\}$
$K^2_{13} = \{00,11,12\}$ $K^2_{15} = \{00,11,22\}$ $K^2_{16} = \{00,12,22\}$
$K^2_{19} = \{00,21,22\}$ $K^2_{24} = \{01,11,22\}$ $K^2_{34} = \{10,21,22\}$
$K^2_{36} = \{11,21,22\}$

Observe that the B-spline evaluation is numericaly stable, since at each point $\underline{x}$ in B_{ρ_j} all barycentric coordinates $\lambda_m(\underline{x}|\underline{x}^{j_0,q_0},\underline{x}^{j_1,q_1},\underline{x}^{j_2,q_2})$, $m = 0,1,2$; $0 \leq q_0 \leq q_1 \leq q_2 \leq 2$ are nonnegative.

We now proceed by showing that this evaluation scheme is the best as far as constant B-spline evaluations are concerned.

Theorem Evaluation of all 6 quadratic B-splines, incident at $\underline{x} \in B_{\rho_j}$, requires at least 10 constant B-spline evaluations.

roof.

Supose that only 9 constant B-spline evaluations are eeded.

In the Fig. 3 linear splines that vanish are denoted y □ .

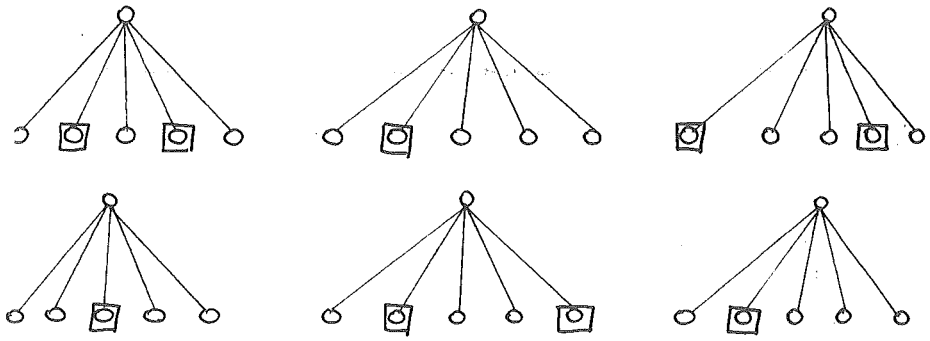


Fig.3

Thus with our 9 constant splines we have to determine :he value of at least one linear B-spline in the first, :hird and fifth tree and at least two in the other trees. Vith a straightforward computer program we can check all :ombinations. It shows up that no combination satisfies requirement presented.■

Since evaluation of all linear B-splines incident at $\underline{x} \in B_{\rho_j}$ requires four constant B-spline evaluations we are :empted to conjucture

$$\binom{k+3}{3}$$

as the lower bound in general case.

4. References

[1] de BOOR, C., Splines as linear combinations of B-splines, in Approximation Theory II, G.G. Lorenz, C.K.Chui & L.L.Schumaker eds., Academic Press, 1976, 1-47.

[2] DAHMEN, W., Polynomials as linear combinations of multivariate B-splines, Math. Z., 169 (1979), 93 - 98.

[3] DAHMEN, W., On multivariate B-splines, SIAM J. Num. Anal., 17 (1980), 179 - 191.

[4] DAHMEN, W., MICCHELLI, C.A., On the linear independence of multivariate B-splines, I:Triangulations of simploids, SIAM J. Num. Anal., 19 (1982), 993 - 1012.

[5] DAHMEN, W., MICCHELLI, C.A., Multivariate splines - a new constructive approach, In Surfaces in Computer Aided Geometric Design, ed. R.E. Barnhill in W. Böhm, North Holland, Amsterdam (1983), 191 - 215.

[6] GMELIG MEYLING, R.H.J., An algorithm for constructing configurations of knots for bivariate B-splines, SIAM J. Num. Anal., xx, xx-xx.

[7] GMELIG MEYLING, R.H.J., On algoritmes and applications for bivariate B-splines, Proc. Conf. Algorithms for the Approximation of Functions and Data, ed. J.C. Mason and M.G.Cox, Shrivenham (1985), xx-xx.

[8] GMELIG MEYLING, R.H.J., Least squares approximation by linear combinations of bivariate B-splines, In Ph.D. Thesis, University of Amsterdam, 1986.

[9] HØLLIG, K., Multivariate splines, SIAM J. Num. Anal., 19 (1982), 1013 - 1031.

ON BOUNDED TENSION INTERPOLATION *

J. KOZAK and M. LOKAR

1. Introduction

The spline in tension goes back to [9], and it was followed by [3], [10], [4], [6], and many others. It was introduced as a (single additional parameter) tool in the shape preserving interpolation, i.e. interpolation that preserves convexity of data. However, it was applied also in other problems (singular perturbation differential equation problems etc.). Thus splines in tension have attracted quite a lot of attention, but they are not very popular in practical computations. Two main reasons for this fact are:

(1) Their use is more time consuming compared to the use of polynomial or rational splines.

(2) The choice od tension parameters is not always apparent.

We shall not bother about (1) since computational complexity for all these spline classes is quite clearly of

ABSTRACT: *Splines in tension are not very popular in practical computations. One of the reasons is obviously the choice of tension parameters that is not always apparent. In this talk, we tackle this problem and we consider the spline in tension as a tool in several approximation problems. In particular, we describe a complete interpolatory tension spline that can be bounded independently of the partition.*

*Supported by Research Council of Slovenija.

the same order, but we shall discuss a remedy for (2): in some applications of splines in tension a natural (and simply computable) choice of tension parameters can be found. However, we shall keep in mind that this talk is far from being meant to show that splines in tension are more useful than for example polynomial ones. They can compete with them only in special circumstances.

Let us recall the definition of a spline in tension. Let $\underline{\tau} := (\tau_i)$ be a strictly increasing partition of $[0,1]$,

$$0 =: \tau_1 < \tau_2 < \ldots < \tau_{n+1} := 1.$$

A spline in tension, with tension parameters $\underline{p} := (p_i)$, $p_i \geq 0$, and breakpoint sequence $\underline{\tau}$ is a function that belongs to

$$S_{4,\underline{\tau},\underline{p}} := C^{(2)}(0,1) \cap N(L)$$

where $L := L_{\underline{p}}$ (and its tension part $M := M_{\underline{p}}$) is piecewise defined by

$$L_{\underline{p}} := \frac{d^4}{dx^4} - \left(\frac{p}{\Delta\tau_i}\right)^2 \frac{d^2}{dx^2} = \frac{d^2}{dx^2}\left(\frac{d^2}{dx^2} - \left(\frac{p}{\Delta\tau_i}\right)^2\right) =:$$

$$=: \frac{d^2}{dx^2}(M_{\underline{p}}) =: \frac{d^2}{dx^2}(M), \qquad \text{on } [\tau_i, \tau_{i+1}).$$

Quite clearly the choice $\underline{p} = (0)$ reproduces the cubic spline space $S_{4,\underline{\tau}}$ as well as $\underline{p} = (\infty)$ the piecewise linear functions $S_{2,\underline{\tau}}$. It is a special case of exponential (and hyperbolic) spline, or more generally L-spline. Some of the ideas presented here could be carried over to more general case. Let us now turn to the examples of practical applications of splines in tension.

2. Shape preserving interpolation

Suppose that a given function g is known at points $\underline{\tau}$. Assume that the partition is extended by

$$\tau_0 := \tau_1, \quad \tau_{n+1} =: \tau_{n+2}$$

in order to simplify the discussion. A shape preserving interpolant is a function that agrees with g at $\underline{\tau}$ and (locally) preserves convexity of the function. However, since g is known only at certain points, an approximation of the second derivative

$$d_j := [\tau_{j-1}, \tau_j, \tau_{j+1}]g,$$

takes over the role. Thus if $d_i d_{i+1} \leq 0$ the second derivative of the interpolant should not change sign in $[\tau_i, \tau_{i+1}]$.

A complete interpolatory spline in tension was the first tool to deal with this (generally nonlinear) problem ([9]). It is easy to see a complete tension interpolatory spline $I_{4,\underline{p}}g$, with interpolatory projector $I_{4,\underline{p}}$ defined by

$$I_{4,\underline{p}} : C(0,1) \longrightarrow S_{4,\underline{\tau},\underline{p}} : f \longrightarrow I_{4,\underline{p}}f := (I_{4,\underline{p}}f|_{\underline{\tau}} = f|_{\underline{\tau}}),$$

is uniquely defined for any tension parameters $\underline{p}$. Thus these additional parameters can be used for smoothing out extraneous inflection points of the interpolant. Quite clearly, a choice $\underline{p} = (\infty)$ would smooth out all inflection points, but produce at most second order approximation. Thus a natural choice ([6]) suggests to choose tension parameters as small as possible in order to preserve the approximation power of the cubic polynomial spline. As it turns out, on each of the subintervals $[\tau_{i-1}, \tau_i]$ it is enough to consider approximate modified derivatives

$$s_0 := \frac{\Delta\tau_i}{\Delta\tau_{i-1}+\Delta\tau_i}([\tau_{i-1}, \tau_i]g - [\tau_i, \tau_{i+1}]g),$$

$$s_1 := \frac{\Delta\tau_i}{\Delta\tau_i+\Delta\tau_{i+1}}([\tau_{i+1}, \tau_{i+2}]g - [\tau_i, \tau_{i+1}]g).$$

If they are of the opposite sign, data indicate that there is no inflexion point, and the interpolant should preserve convexity on the given interval. In this case a quantity $\omega := s_1/(s_1-s_0)$ is studied. A short analysis shows that a spline in tension will not have an inflection point if ω is

trapped in a certain interval. This leads to a simple nonlinear equation for p_i.

3. Tension spline collocation

To start more generaly, consider m-th order linear (to make the discussion simpler) ordinary differential equation for unknown u,

$$Au = f, \text{ on } [0,1]$$

with boundary or initial conditions

$$\beta_i u = c_i, \quad i = 1,2,\ldots,m.$$

Here

$$A := \sum_{i=0}^{m} a_i D^i, \quad a_m \neq 0.$$

We shall assume that the equation has a unique solution u for any f, i.e. there exists a unique Green's function G.

In a simple outfit a collocation approximation to u is constructed as follows:

1) The interval $[0,1]$ is partitioned by a strictly increasing breakpoint sequence $\underline{\tau}$.

2) An approximate solution $u_{\underline{\tau}}$ is looked for as

$$u_{\underline{\tau}} \in C^{(m-1)}(0,1) \cap B^{-1}(P_{k,\underline{\tau}}).$$

where $P_{k,\underline{\tau}}$ denotes as usually the space of piecewise polynomial functions of order k, and B playes the role of an approximation of A. It is usually taken as

$$B = D^m,$$

i.e. the leading part of A. $u_{\underline{\tau}}$ has to satisfy differential equation at collocation points

$$\xi_{ij} \in [\tau_i, \tau_{i+1}), \quad j = 1,2,\ldots,k; \; i = 1,2,\ldots n,$$

$$\xi_{ij} < \xi_{i,j+1}, \text{ all } j$$

and additional conditions

$$\beta_i u_{\underline{\tau}} = c_i, \quad i = 1,2,\ldots,m.$$

Error analysis reveals pointwise error e as

$$e(x) := u(x) - u_{\underline{\tau}}(x) = \int_0^1 G(x,\cdot)(f - Au_{\underline{\tau}})$$

where the Green's function satisifies zero additional conditions,

$$\beta_i G(\cdot,y) = 0, \quad i = 1,2,\ldots,m.$$

The factor $r := f - Au_{\underline{\tau}}$ vanishes at the collocation points ξ_{ij} in $[\tau_i,\tau_{i+1})$. As a consequence, this contributes a factor

$$|\underline{\tau}|^k := \max_i \Delta\tau_i$$

to the L_∞ error bound if $r^{(k)}$ can be properly bounded. The choice of collocation points, based upon orthogonality relations, can further raise the order of approximation up to

$$O(|\underline{\tau}|^{k+m}),$$

and at the breakpoints τ_i even up to

$$O(|\underline{\tau}|^{2k}).$$

For a smooth f, behaviour of r depends on $Au_{\underline{\tau}}$, thus more or less on the quality of approximation of A by $\bar{B}$. Improper choice of $\bar{B}$ could introduce a large error by the method of solution (the choice of collocation functions), regardless of the nature of G that is inherent to the problem, and cannot be avoided. In such a case we can conclude that the nullspaces

$$N(A), \quad N(\bar{B})$$

differ significantly. The usuall choice $B = D^m$ quite clearly fails if the behaviour of solution depends heavily on all of A, not only on its leading term. Consider an example, a second order singular perturbation problem of the form

$$A = -\varepsilon D^2 + a_0 I, \quad 0 < \varepsilon << 1, \; a_0 \neq 0.$$

A solution depends heavily on the sign of a_0, and an approximation B has to take this into account. The best choice would be $B = A$, since then Au_τ reduces to a polynomial. However, such a B cannot be always practically computed. Assume now $a_0 > 0$. A piecewise constant $O(|\tau|)$ approximation

$$B = -\varepsilon D^2 + a_0(\frac{\tau_i + \tau_{i+1}}{2}), \text{ on } [\tau_i, \tau_{i+1}),$$

depends on the sign of a_0 too. Further, $k = 2$ brings us back to the splines in tension as collocation functions, with natural choice of tension parameters

$$p_i = \Delta\tau_i \sqrt{a_0(\frac{\tau_i + \tau_{i+1}}{2})/\varepsilon}, \quad \text{all } i.$$

4. Bounded tension interpolation

It is customary to study approximation power of linear interpolation problems by analysing its bound, expressed as a product of two factors. The first depends on the interpolation scheme, the second on the best approximation of the given function in the space concerned. For a familiar complete spline projector I_4 this inequality, called Lebesgue, would read

$$\| I_4 f - f \| \leq (1 + \| I_4 \|) dist(f, S_{4,\underline{\tau}}).$$

Here $\| \cdot \| := \| \cdot \|_\infty$, and *dist* defined correspondingly. A properly bounded I_4 would quite clearly produce an optimal order approximation. On the other hand, the interpolation error can be bounded also from below by $\| I_4 f \| - \| f \|$. This shows that interpolation error for some

functions f has to be large if $\| I_4 \|$ is large though $\| f \| = 1$. But then one could expect that large norm of interpolation projector would significantly amplify errors in the measured, not accurate data. Thus we can conclude that the bounding of an interpolatory projector has its theoretical as well as practical importance.

It is well known that the projector I_4 can not be bounded independently of $\underline{\tau}$ ([2]), and various restrictions have been imposed on $\underline{\tau}$ in order to produce a bounded projector. One of the approaches (which is also of practical importance for partitions that are close to the geometric one) is to bound I_4 by considering local mesh ratio

$$m_i := \frac{\Delta\tau_i}{\Delta\tau_{i-1}}$$

and its bound

$$m_\Delta := \sup_{|i-j|=1} \frac{\Delta\tau_i}{\Delta\tau_j}.$$

The result that was quite a while looked for can be found in [1]: the complete cubic spline interpolation is bounded (independently of n) if

$$m_\Delta \leq m^* < m_4^* := \frac{3+\sqrt{5}}{2} \quad (m^* \text{ constant}).$$

If the partition $\underline{\tau}$ is too nonuniform one can shift to splines in tension ([7]). The idea is to choose large tension parameters p_i where the partition is changing too rapidly, but to stick to $p_i = 0$ if it is locally uniform. To be precise, a natural choice of $\underline{p}$ is as follows: $I_{4,\underline{p}}$ should be as close to I_4 as possible, but bounded independently of n by a given constant.

The tension parameters are obtained by looking at tension nullsplines. A nullspline $s \in S_{4,\underline{\tau},\underline{p}}$ satisfies

$$s(\tau_i) = 0, \text{ all } i.$$

A tension nullspline is described on each of the intervals by two values which are continuously carried over the interval boundary. A short computation yields,

$$\underline{s}_{i+1} = - A_i \underline{s}_i$$

where

$$A_i := A(m_i, p_i),$$

$$A(p,m) := \begin{pmatrix} \alpha m & 2\frac{\alpha^2 - 1}{\beta} m^2 \\ \frac{\beta}{2} m & \alpha m^2 \end{pmatrix},$$

$$\alpha := \alpha(p) := \frac{p\,\mathrm{ch}(p) - \mathrm{sh}(p)}{\mathrm{sh}(p) - p},$$

$$\beta := \beta(p) := \frac{p^2 \mathrm{sh}(p)}{\mathrm{sh}(p) - p},$$

and

$$\underline{s}_i := \begin{pmatrix} \Delta\tau_{i-1} s'(\tau_i) \\ \frac{\Delta\tau_{i-1}^2}{2} s''(\tau_i) \end{pmatrix}.$$

The choice of $\underline{p}$ has to guarantee that a nullspline increases exponentially in at least one direction. An argument in [5] reduces this to the inequalities (by elements)

$$|A(\frac{1}{m_\Delta}, p_\Delta)| \leq |A(m_i, p_i)| \leq |A(m_\Delta, p_\Delta)|,$$

$$|A^{-1}(m_\Delta, p_\Delta)| \leq |A^{-1}(m_i, p_i)| \leq |A^{-1}(\frac{1}{m_\Delta}, p_\Delta)|.$$

Here, p_Δ is chosen in advance in such a way that the largest eigenvalue λ_2 of the matrix A,

$$\lambda_1 := \lambda_1(m,p) < \lambda_2 := \lambda_2(m,p) < 0$$

satisfies

$$\omega := |\lambda_2(m_\Delta, p_\Delta)| < 1.$$

This assures that the fundamental splines decay by at least a factor ω, and as consequence produces a bounded interpolation. The matrix inequalities are further by a somewhat tedious argument reduced to a single nonlinear equation that determines p_i.

Let us conclude with a brief mention on the approximation power of splines in tension, with the emphasis on its dependence on tension parameters. A general result, with tension parameters hidden in a constant, can be found in [8]. Let us state a refined conclusion ([7]:

Let $f \in C^{(4)}(0,1)$. Then

$$dist_i(f, S_{4,\underline{\tau},\underline{p}}) \leq \| f - I_{4,\underline{p}} f \|_i \leq \frac{\Delta\tau_i^4}{2} C(p_i) \| Mf'' \|_i$$

with

$$C(p) := \frac{1}{p^2}\left(1 - \frac{1}{\mathrm{ch}(p/2)}\right).$$

Here $\| \cdot \|_i$ denotes the sup norm on $[\tau_i, \tau_{i+1}]$, and $dist_i$ is defined correspondingly. Note that $p_i = 0$, all i, reduces the bound to

$$\frac{\Delta\tau_i^4}{16} \| f^{(4)} \|_i,$$

as well as $p_i \longrightarrow \infty$, all i to

$$\frac{\Delta\tau_i^2}{2} \| f'' \|_i.$$

This is (up to the constant) expected.

5. References

1. C. de BOOR: On cubic spline functions which vanish at all knots. Advances in Mathematics 20(1976), 1-17.

2. C. de BOOR: A Practical Guide to Splines. Springer Verlag, New York, 1978.

3. A. CLINE: Scalar- and planar-valued curve fitting in one and two dimensional spaces using splines under tension. Comm. ACM 17(1974), 218-223.

4. J. E. FLAHERTY, W. MATHON: Collocation with polynomial and tension splines for singularly-perturbed boundary value problems. SIAM J. Sci. Stat. Comp. 1(1980), 260-289.

5. S. FRIEDLAND, C.A. MICCHELLI: Bounds of the solutions of difference equations and spline interpolation at knots. Lin. Alg. and its Appl., 20(1978), 219-251.

6. J. KOZAK: Shape preserving approximation. Computers in Industry 7 (1986), 435-440.

7. Y.Y.FENG, J. KOZAK: An approach to the interpolation of nonuniformly spaced data. to appear.

8. L.L. SCHUMAKER: Spline Functions: Basic Theory. John Wiley & Sons, New York, 1981.

9. D.G. SCHWEIKERT: An interpolating curve using a spline in tension. J.Math. Physics 45 (1966), 312-317.

10. H. SPÄTH: Spline-Algorithmen zur Konstruktion glatter Kurven und Flächen. R. Oldenbourg Verlag, München, 1973.

NUMERICAL METHODS IN SEMICONDUCTOR DEVICE SIMULATION

P.A. MARKOWICH, C. SCHMEISER* and S. SELBERHERR

Abstract: The simulation of the electrical behavior of semiconductor devices involves the solution of initial-boundary value problems for a nonlinear elliptic-parabolic system. Two major difficulties in the numerical solution of these problems are discussed:

a) The construction of discretisations is not obvious as the equations are singularly perturbed.
b) The discretised problems are very large systems of nonlinear algebraic equations which have to be solved iteratively.

1. INTRODUCTION

The electrical behavior of a semiconductor device is determined by the flow of two types of free charge carriers, the electrons in the conduction band (density $n(x,t)$) and the defect electrons or holes in the valence band (density $p(x,t)$). Well accepted models for the flow of electrons and holes are the Boltzmann transport equations, but their complexity is prohibitive for the numerical simulation of complicated devices. Perturbation arguments lead to the simplified drift-diffusion approximation of the current densities:

$$J_n = \mu_n(\nabla n + nE) \,, \qquad J_p = -\mu_p(\nabla p - pE) \,. \tag{1.1a}$$

(All the appearing variables and parameters are already in scaled dimensionless form.)

*The work of the second author was supported by "Österreichischer Fonds zur Förderung der wissenschaftlichen Forschung".

In (1.1a) the parameters μ_n, μ_p denote mobilities and E is the electric field which is related to the electrostatic potential ψ by

(1.1b) $\quad E = -\nabla\psi \ .$

Common models for the mobilities depend on n,p,E and the position x.

Maxwell's equations imply the continuity equations

$$\text{(1.1c)} \qquad \begin{aligned} \operatorname{div} J_n - n_t &= R \ , \\ \operatorname{div} J_p + p_t &= -R \end{aligned}$$

and Poisson's equation

(1.1d) $\quad \lambda^2 \Delta\psi = n - p - C(x) \ ,$

where the source term R, called the recombination-generation rate, is the number of electron-hole pairs which are generated $(R<0)$ or disappear $(R>0)$ per unit time. It is usually modelled as a given function of n,p,E and position. The function $C(x)$, the so called doping profile, denotes the concentration of impurity ions. The dimensionless parameter λ is the scaled minimal Debye length and takes small values for realistic semiconductor devices.

The unscaled equations (1.1) are due to Van Roosbroeck [21]. For a derivation from Maxwell's equations and the Boltzmann transport equation see Selberherr [18]. The scaling which leads to (1.1) can be found in Markowich [8].

Mathematically a semiconductor device is given by the doping profile $C(x)$ defined in a bounded domain $\Omega \subseteq \mathbb{R}^3$ which represents the semiconductor part of the device. For the purpose of simulation it often makes sense to reduce the dimension of Ω. Thus, we take $\Omega \subset \mathbb{R}^k$, $k = 1,2$ or 3. The boundary $\partial\Omega$ spli into the union of contact segments $\partial\Omega_D$ where Dirichlet boundary conditions for n,p and ψ are given

(1.2a) $\quad n\big|_{\partial\Omega_D} = n_D \ , \quad p\big|_{\partial\Omega_D} = p_D \ , \quad \psi\big|_{\partial\Omega_D} = \psi_D \ ,$

and the insulating part $\partial\Omega_N$ where the homogeneous Neumann conditions

(1.2b) $\quad (J_n,\nu)\big|_{\partial\Omega_N} = (J_p,\nu)\big|_{\partial\Omega_N} = (E,\nu)\big|_{\partial\Omega_N} = 0$

hold. In (1.2b) ν denotes the outward normal vector of $\partial\Omega$.

Substituting (1.1a) into (1.1c) shows that (1.1) is a system of two parabolic equations for n and p coupled to an elliptic equation for ψ. In order to complete the formulation of an initial-boundary value problem initial conditions for the densities

$$(1.3) \qquad n(x,0) = n_I(x) \ , \quad p(x,0) = p_I(x) \ , \quad x \in \Omega$$

have to be prescribed. The potential at $t = 0$ can be determined by solving Poisson's equation. Several existence and uniqueness results for (1.1)-(1.3) can be found in the literature (see e.g. Mock [12]). Existence results for the corresponding stationary problem are contained in [8] and [12]. Uniqueness cannot be expected in general (see Steinrück [19]).

For the construction and analysis of numerical methods some a priori knowledge of the solution structure is extremely important. This can be gained from a singular perturbation analysis by exploiting the smallness of the parameter λ^2 in (1.1d). In the stationary case such an analysis shows that the solution can be approximated by setting $\lambda = 0$ except in thin layer regions where it varies rapidly (see [8]). For the time dependent problem additionally an initial layer appears (see Ringhofer [14], Szmolyan [20], Markowich [9]). In this paper we will be concerned with the stationary problem. Its analysis is facilitated by the transformation

$$(1.4) \qquad n = e^{\psi}u \ , \quad p = e^{-\psi}v$$

which takes the stationary differential equations to the form

$$\begin{aligned} \lambda^2 \Delta\psi &= e^{\psi}u - e^{-\psi}v - C(x) \\ (1.5) \qquad \mathrm{div}(\mu_n e^{\psi} \nabla u) &= R \\ \mathrm{div}(\mu_p e^{-\psi} \nabla v) &= R \end{aligned}$$

The continuity equations are in self-adjoint form now. Besides u and v are so called slow variables which means that they do not exhibit layer behavior. As opposed to (1.1d) the potential can be determined from the reduced (λ=0) Poisson's equation. Subject to the appropriate boundary conditions each of the equations in (1.5) represents a well posed problem for the variable which appears with the highest differential order, when the other two variables are considered as known.

These properties make it much easier to design numerical methods which are well suited for (1.5) than for the original system. Unfortunately the potential becomes rather large in many applications such that u and v are so out of range that they are impossible to compute with (for different choices of variables and related conditioning questions see Bank et al. [3], Schmeiser et al. [17], Ascher et al. [1]). These facts led to the following approach: Methods are designed and analysed for (1.5). In computations the transformation (1.4) is applied on the discrete level to be able to compute with the original variables ψ,n and p.

2.DISCRETISATIONS

In this section we shall present discretisations for the steady state semiconductor equations which take into account the singular perturbation nature of the problem. The properties of system (1.5) allow for a decoupled approach, where each equation is treated separately.

2.1. Poisson's equation is a semilinear elliptic equation for the potential when u and v are considered to be known. The solution is approximated by a solution of the reduced equation except close to regions of rapid variation of the doping profile and possibly close to the boundaries where the solution varies rapidly. When trying to solve the problem numerically one would expect to be forced to use grids which are fine enough in the regions of rapid variation to resolve the solution structure. For the simulation of complex devices the cost of using such a grid is prohibitive. In order to get around this difficulty, discretisations are used which mimic the above described properties of the continuous problem by the use of lumping for the evaluation of the right hand side. A finite element of finite difference discretisation at node x_i then takes the form

$$\lambda^2(\Delta_h\psi_h)_i = e^{\psi_i}u_i - e^{-\psi_i}v_i - C(x_i) \tag{2.1}$$

where Δ_h is a discretised version of the Laplace-operator (see Markowich [8], Selberherr [18]). The effect of lumping is that the reduced equations in the continuous and the discrete

case are the same. For any discretisation which inherits the stability properties of the continuous operator (maximum principle) the solution structure is similar for the discrete and continuous problems even if a coarse mesh is used. The main difference is that layers in the discrete case may be wider $(O(h))$ than in the continuous case $(O(\lambda))$. This fact will be demonstrated in the following section. It has two effects of major importance. First, even when starting on a very coarse grid adaptive grid refinement will be able to detect the correct solution structure. Second, as the solution is approximated well away from the thin layer regions the approximation error will be small if measured in integral norms although large pointwise errors may occur. The importance of this effect will also be demonstrated in section 3.

2.2. The continuity equations. We shall only deal with the electron continuity equation as the necessary modifications for the hole continuity equation are obvious. Let us first consider the one-dimensional situation. As the variables u and J_n are slow variables - in the language of singular perturbation theory - in this case, the discretisation of

$$J_n' = R \ , \quad J_n = \mu_n e^{\psi} u' \tag{2.2}$$

is not very critical. For simplicity we assume an equidistant grid and replace the first equation at the gridpoint x_i by

$$J_{n,i+1/2} - J_{n,i-1/2} = h R_i \tag{2.3a}$$

where R_i denotes an approximation of the recombination-generation rate at x_i. The second equation is approximated between gridpoints by

$$J_{n,i+1/2} = \mu_{n,i+1/2} (e^{\psi})_{i+1/2} \frac{u_{i+1}-u_i}{h} \ , \tag{2.3b}$$

where the approximation $\mu_{n,i+1/2}$ for μ_n at $\frac{x_i+x_{i+1}}{2}$ depends on the model which is used. For the approximation $(e^{\psi})_{i+1/2}$ two obvious choices are

$$\frac{1}{2}(e^{\psi_i}+e^{\psi_{i+1}}), \quad \exp(\frac{\psi_i+\psi_{i+1}}{2}) \ .$$

A third possibility is obtained by replacing μ_n and J_n by constants and ψ by a linear function in $[x_i, x_{i+1}]$ and solving the second equation in (2.2) explicitely. This results in the approximation

$$(2.3c) \qquad (e^{\psi})_{i+1/2} = \frac{\psi_{i+1} - \psi_i}{e^{-\psi_i} - e^{-\psi_{i+1}}} .$$

This procedure could have also been applied to the equation (1.1a) in the original variable n. The so obtained discretisation which is equivalent to (2.3) is an example of an exponentially fitted method (see Doolan et al. [5]) and bears the names of the engineers Scharfetter and Gummel [16] in the semiconductor device simulation literature.

The difference between the above mentioned discretisations is an unsettled issue from the theoretical point of view, but in practically all of the existing device simulation software the Scharfetter-Gummel scheme is used.

Extensions to finite difference methods in the two- and three-dimensional cases are straightforward (see Selberherrr [18]). Finite element methods which are generalisations of the Scharfett Gummel method to the two-dimensional situation can be found in Buturla and Cottrell [4] and Markowich and Zlamal [10]. It can be shown that the errors only depend on the variation of the current density J_n (see [10], Mock [13]). The drawback in the multidimensional situation is that J_n is not a slow variable in general (see Markowich [8]) which makes it necessary to use fine grids in regions of rapid variation of J_n. However, in most practical situations J_n varies much less than ψ,n and p and the computational effort remains reasonable.

The above error considerations dealt with each equation separately. In order to prove convergence results for the full system one has to assume wellposedness of the problem. Then the error estimates for the single equations can be combined (see [8]).

3.A UNIFORM CONVERGENCE RESULT

When talking about numerical methods for singular perturbation problems uniform convergence means roughly that errors can be estimated independently of the singular perturbation parameter. In particular, errors are even small if the grid ignores layers. Results of this kind can be proven for pointwise errors when using exponentially fitted methods (see Doolan et al. [5]). Such a result cannot be expected for the discretisations of the semiconductor device equations discussed in the preceding section, but this is of minor importance when the goals of device modeling are considered. These goals are basically twofold. One aim is to reveal the solution structure inside the device, the second is to obtain the relation between applied voltages - which enter the Dirichlet boundary conditions - and outflow currents - which are computed by integrals of the current densities along contact segments. Only for the latter part the accuracy of the method is of decisive importance. In this section we prove for a model problem that both aims can be met with reasonable computational effort.

We consider a one-dimensional problem with constant mobilities and vanishing recombination-generation rate. System (1.5) reads

$$
\begin{aligned}
&\lambda^2\psi'' = e^{\psi}u - e^{-\psi}v - C(x) \ , \\
&(e^{\psi}u')' = 0 \ , \\
&(e^{-\psi}v')' = 0
\end{aligned}
\tag{3.1}
$$

in this case. The simulation domain is $\Omega = (0,1)$. System (3.1) is subject to Dirichlet boundary conditions at $x = 0$ and $x = 1$. We consider an equidistant grid on $[0,1]$. Poisson's equation is discretised by using the common three point formula for the approximation of ψ''. The approximate solution ψ_h is obtained by linear interpolation between the gridpoints.

The Scharfetter-Gummel method amounts to replacing ψ by ψ_h in the continuity equations and solving them explicitely because of the assumptions on μ_n, μ_p and R. Problem (3.1) can be written as a fixed point problem by denoting the solutions of the continuity equations for given ψ

$$
\begin{aligned}
u(x) &= u(0) + (u(1)-u(0)) \int_0^x e^{-\psi} \Big/ \int_0^1 e^{-\psi} \,, \\
v(x) &= v(0) + (v(1)-v(0)) \int_0^x e^{\psi} \Big/ \int_0^1 e^{\psi}
\end{aligned} \tag{3.2}
$$

by $u(\psi), v(\psi)$ and the solution of

$$\lambda^2 \phi'' = e^{\phi} u(\psi) - e^{-\phi} v(\psi) - C(x)$$

plus boundary conditions by $\phi = T(\psi)$. A fixed point of the operator T corresponds to a solution.

The discretised problem can be written as

$$
\begin{aligned}
&\frac{\lambda^2}{h^2} (\psi_{i+1} - 2\psi_i + \psi_{i-1}) = e^{\psi_i} u_i - e^{-\psi_i} v_i - C(x_i) \,, \\
&u_h = u(\psi_h) \,, \quad v_h = v(\psi_h) \,.
\end{aligned} \tag{3.3}
$$

Our convergence analysis will be based on the

Lemma 3.1: Let the Frechet derivative of the operator $(I-T)$ at ψ_h be invertible and the inverse be bounded as operator from $L^1(\Omega)$ to $L^1(\Omega)$ independently of λ and h.
Let $\| \psi_h - T(\psi_h) \|_1$ be sufficiently small, where $\|.\|_p$ denotes the L^p-Norm on $(0,1)$.
Then (3.1) has a locally unique solution ψ^* and

$$\| \psi^* - \psi_h \|_1 \leq K_1 \| \psi_h - T(\psi_h) \|_1$$

with K_1 independent of λ and h holds.

The proof is a straightforward application of the implicit function theorem (For similar results see [8],[12]).

Because of Lemma 3.1 we only have to estimate the L^1-Norm of the error in solving Poisson's equation. This is contained in

Lemma 3.2: Let $C(x)$ have a finite number of jump discontinuities in $[0,1]$ and Lipschitz-continuous first derivatives between those points. Let $u(0), u(1), v(0), v(1) > 0$ hold. Then

$$\| \psi_h - T(\psi_h) \|_1 \leq K_2 (\lambda + h)$$

holds with K_2 independent of λ and h.

Outline of a proof:A priori estimates (see [8],[12]) show that

$$e^{\psi_i} u_i + e^{-\psi_i} v_i \geq K > 0$$

holds for the derivative with respect to ψ_i of the right hand side of (3.3). Thus the discrete operator in (3.3) is of inverse monotone type (see Meis-Markowitz [11]). This allows the use of comparison functions for estimates of the solution. Comparison functions can be constructed which are roughly the sum of a solution of the reduced equation and of terms which decay exponentially away from the boundaries and the discontinuities of the doping profile. The L^1-Norm of the decaying terms can be computed and shown to be of the order $O(\lambda+h)$. The argument that the layer terms in the continuous solution are $O(\lambda)$ with respect to the L^1-Norm completes the proof.

A combination of the above lemmata yields the main result of this section

Theorem 3.3: Let the assumptions of the Lemmata 3.1 and 3.2 hold. If the total current density is denoted by $J = J_n + J_p$, the estimate

$$\|\psi^*-\psi_h\|_1 + \|u^*-u_h\|_\infty + \|v^*-v_h\|_\infty + |J-J_h| \leq K_3(\lambda+h)$$

holds with K_3 independent of λ and h.

Proof: The estimate for the error in the potential follows direct from the preceding lemmata. Considering the representation (3.2) for u and v and

$$J_n = (u(1)-u(0)) \Big/ \int_0^1 e^{-\psi}, \quad J_p = (v(0)-v(1)) \Big/ \int_0^1 e^{\psi}$$

for the current densities the proof of the remaining estimates is also immediate.

Supposedly the above result can be extended to one-dimensional problems with less stringent assumptions one the mobilities and the recombination-generation rate. In the multidimensional situation a similar result cannot be expected to hold because layers in the current densities have to be resolved which requires grid-spacings of the order $O(\lambda)$.

4. NONLINEAR ITERATION METHODS

By discretising (1.5) we obtain a large system of nonlinear algebraic equations. Their solution requires the use of apprc priate iteration methods. Although these methods are applied to the discrete problem we discuss them for the continuous eq tions for notational convenience. Assuming again constant mob lities and vanishing recombination-generation we have to solv

$$
\begin{aligned}
&\lambda^2\Delta\psi - e^{\psi}u + e^{-\psi}v + C(x) = b_1 = 0 \ , \\
&\operatorname{div}(e^{\psi}\nabla u) = b_2 = 0 \ , \\
&\operatorname{div}(e^{-\psi}\nabla v) = b_3 = 0 \ .
\end{aligned}
\tag{4.1}
$$

Newton's method for (4.1) reads

$$
\begin{aligned}
&\lambda^2\Delta d\psi - (e^{\psi}u+e^{-\psi}v)d\psi - e^{\psi}du + e^{-\psi}dv = - b_1 \ , \\
&\operatorname{div}(J_n d\psi + e^{\psi}\nabla du) = -b_2 \ , \\
&\operatorname{div}(J_p d\psi + e^{-\psi}\nabla dv) = -b_3 \ .
\end{aligned}
\tag{4.2}
$$

Its application requires the solution of a large linear syste in each iteration step. The computational cost can be reduce by "freezing" the Frechet-derivative for several iteration s For efficient strategies of this kind and their analysis see and Rose [2]. Their concept of approximate Newton methods al for perturbations in the Frechet-derivative. A worthwhile go to find perturbations which decouple the linear system (4.2) certain extent. One method of this kind relies on the assump that the current densities are comparatively small. Obviousl (4.2) is decoupled if J_n and J_p are replaced by zero. Th resulting method amounts to solving the continuity equations given ψ and then the linearized Poisson's equation with the dated u and v in each step. This method was first proposed b Gummel [6]. An alternative which also carries his name is to solve the nonlinear Poisson's equation in each step which ca also be seen as the Picard iteration for the fixed point pro $\psi = T(\psi)$ formulated in the preceding section. Convergence an ses of Gummel's method for small current densities are conta in Markowich [8] and Kerkhoven [7].

When the current densities take values of significant size the convergence of Gummel's method often deteriorates. In view of this situation a different kind of decoupling by approximating the Frechet derivative was introduced in Ringhofer and Schmeiser [15]. Here the singular perturbation character of the linearised problem (4.2) is used. As du and dv are slow variables they are approximated well by the solution of the reduced problem. Thus, we substitute

$$\overline{d\psi} = (-e^{\psi}\overline{du} + e^{-\psi}\overline{dv} + b_1)(e^{\psi}u + e^{-\psi}v)^{-1}$$

into the linearized continuity equations

$$\text{(4.3a)}\qquad \begin{aligned} &\operatorname{div}\Big(\frac{J_n}{e^{\psi}u+e^{-\psi}v}\,(-e^{\psi}\overline{du}+e^{-\psi}\overline{dv}+b_1) + e^{\psi}\nabla\overline{du}\Big) = -b_2\,,\\ &\operatorname{div}\Big(\frac{J_p}{e^{\psi}u+e^{-\psi}v}\,(-e^{\psi}\overline{du}+e^{-\psi}\overline{dv}+b_1) + e^{-\psi}\nabla\overline{dv}\Big) = -b_3\,. \end{aligned}$$

As $d\psi$ is a fast variable, $\overline{d\psi}$ is a good approximation only away from layers. In order to improve on that the full linearised Poisson's equation has to be solved:

$$\text{(4.3b)}\qquad \lambda^2\Delta\widetilde{d\psi} - (e^{\psi}u+e^{-\psi}v)\widetilde{d\psi} - e^{\psi}\overline{du} + e^{-\psi}\overline{dv} = -b_1$$

Instead of the Newton corrections $d\psi, du, dv$ we now use $\widetilde{d\psi}, \overline{du}, \overline{dv}$ In the perturbed problem (4.3) Poisson's equation is decoupled from the continuity equations which are coupled to each other by the terms multiplied by J_n and J_p.

Some of the most important semiconductor devices (e.g.MOSFETs) are so called unipolar devices. They are characterised by the property that only one type of charge carriers (i.e. electrons or holes) contributes significantly to the current flow. This means that one current density (for example J_p) is very small compared to the other. This motivates a further decoupling by replacing J_p by zero in (4.3). The resulting method was proven to converge linearly in [15] with a convergence rate of the form

$$\text{(4.4)}\qquad \text{const}\,(c(\lambda)\,\|J_n\| + \|J_p\|)$$

if the problem is well-posed. In (4.4) $c(\lambda)$ tends to zero as $\lambda \to 0$ and $\|.\|$ denotes a suitable norm. The value of (4.4) is so

small in many applications that the convergence behavior is dominated by the quadratic terms throughout the computations which suggests the use of the term "almost quadratic convergence". The performance of this method was examined in [15] by numerical tests which showed that - compared to Gummel's method - a significant improvement can be achieved.

References:

[1] U.Ascher, P.A.Markowich, C.Schmeiser, H.Steinrück, R.Weiss, Conditioning of the Steady State Semiconductor Device Problem, Techn.Rep.86-18, Comp.Sc., UBC, 1986.

[2] R.E.Banks, D.J.Rose, Global Approximate Newton Methods, Numer.Math.37, 279-295, 1981.

[3] R.E.Bank, D.J.Rose, W.Fichtner, Numerical Methods for Semiconductor Device Simulation, SIAM JSSC 4, No.3, 416-435, 1983.

[4] E.M.Buturla, P.E.Cottrell, Two-Dimensional Finite Element Analysis of Semiconductor Steady Transport Equations, Proc Int.Conf. "Computer Methods in Nonlinear Mechanics", Aust Texas, 512-530, 1974.

[5] E.P.Doolan, J.H.H.Miller, W.H.A.Schilders, Uniform Numeric Methods for Problems with Initial and Boundary Layers, Boc Press, Dublin, 1980.

[6] H.K.Gummel, A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations, IEEE Trans.El.Devices, ED-11, 455-465, 1964.

[7] T.Kerkhoven, Convergence of Gummel's Algorithm for Realist Device Geometries, Proc.NASECODE IV Conf., Boole Press, Dub 1985.

[8] P.A.Markowich, The Stationary Semiconductor Device Equatio Springer-Verlag, Wien, 1986.

[9] P.A.Markowich, Spatial-Temporal Structure of Solutions of the Semiconductor Device Problem, to appear in "Lecture Notes on Appl.Math.", 1987.

[10] P.A.Markowich, M.Zlamal, Inverse-Average-Type Finite Element Discretisations of Self-Adjoint Second Order Elliptic Problems, submitted to Math.Comp., 1987.

[11] T.Meis, U.Marcowitz, Numerische Behandlung Partieller Differentialgleichungen, Springer-Verlag, Berlin, 1978.

[12] M.S.Mock, Analysis of Mathematical Models of Semiconductor Devices, Boole Press, Dublin, 1983.

[13] M.S.Mock, Analysis of a Discretisation Algorithm for Stationary Continuity Equations in Semiconductor Device Models I, COMPEL 2, No.3, 117-139, 1983.

[14] C.A.Ringhofer, The Shape of Solutions to the Basic Semiconductor Equations, to appear in "Lecture Notes on Appl.Math.", 1987.

[15] C.A.Ringhofer, C.Schmeiser, An Approximate Newton Method for the Solution of the Basic Semiconductor Device Equations, submitted to SIAM J.Numer.Anal., 1987.

[16] D.L.Scharfetter, H.K.Gummel, Large Signal Analysis of a Silicon Read Diode Oscillator, IEEE Trans.El. Devices, ED-16, 64-77, 1969.

[17] C.Schmeiser, S.Selberherr, R.Weiss, On Scaling and Norms for Semiconductor Device Simulation, Proc. NASECODE IV Conf., Boole Press, Dublin, 1985.

[18] S.Selberherr, Analysis and Simulation of Semiconductor Devices, Springer-Verlag, Wien, 1984.

[19] H.Steinrück, A Bifurcation Analysis of the One Dimensional Steady State Semiconductor Device Equations, submitted to SIAM J.Appl.Math., 1987.

[20] P.Szmolyan, Ein hyperbolisches System aus der Halbleiterphysik, Thesis, Techn.Univ.Wien, 1987.

[21] W.V.Van Roosbroeck, Theory of Flow of Electrons and Holes in Germanium and Other Semiconductors, Bell Syst.Techn.J. 29, 560-607, 1950.

SOLUTION OF THE DIFFUSION EQUATION IN VLSI PROCESS MODELING BY A NONLINEAR MULTIGRID ALGORITHM

S. MIJALKOVIĆ and N. STOJADINOVIĆ

ABSTRACT: *An application of the nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm for solution of the diffusion equation in VLSI process modeling has been investigated. It is demonstrated that this approach shows high efficiency, which is essentialy independent of physical and numerical parameters of the problem.*

1. INTRODUCTION

For the present underlying physical models of processes used in VLSI process simulation programs, solution of the two-dimensional diffusion equation places heavy demands on computer resources. Moreover, further improvements in kinetic models of point defects, because of their important role in coupling oxidation and diffusion processes, will require at least a tenfold increase in computational throughput for the next generation of VLSI process simulation programs [8]. Therefore, it is clear that more emphasis should be put on numerical approaches that are more efficient than those currently used for diffusion process simulation.

In this view, multigrid methods, well known as the fastest solvers of discretized partial differential equations, seem to be a good choice for this application. Besides their computational efficiency, multigrid methods are fully parallelizable on multiprocessor computers. However, it should be noted that highly efficient and extendable multigrid solvers for more complex problems could be obtained only with a proper choice of a multigrid algorithm and various additional multigrid components [1].

In this paper an application of a nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm for solution of the diffusion equation in VLSI process modeling has been investigated. In the following section mathematical description of diffusion equation and discretization procedure are briefly outlined. The third section describes multigrid algorithm and related multigrid components used. Finally, the two last sections contain analysis of the empirical solution efficiency obtained and some practical examples of actual simulation.

2. PROBLEM DEFINITION

Simulation of the redistribution of impurities in semiconductors under practical processing conditions involves solution of the nonlinear diffusion equation in a domain where one of the boundaries (the silicon--oxide interface) is continually and nonuniformly moving in space as a function of time.

By ignoring diffusion in the oxide (which is justified in many cases), one can consider the diffusion equation [5]

$$\frac{\partial C}{\partial t} - \frac{\partial}{\partial x}(D\frac{\partial C}{\partial x}) - \frac{\partial}{\partial \eta}(D\frac{\partial C}{\partial \eta}) = 0 \tag{1}$$

as two-dimensional initial boundary-value problem in the bounded domain Ω with C being the impurity concentration and D=D(C) the concentration dependent diffusion coefficient.

The boundary conditions are

1) deep in the silicon substrate i.e. at the top of simulated region ($\eta=\eta_1+mU$): $C=C_{min}=10^{13}$ cm^{-3},

2) along the lines of symmetry ($x=0$ and $x=x_1$): $\partial C/\partial n=0$ and

3) on the silicon-oxide interface i.e. at the bottom of simulated region ($\eta=mU$):

$$D\frac{\partial C}{\partial n} - (k-m)\cdot\dot{U}\cdot C\cdot n = 0 \tag{2}$$

where x_1 and η_1 determine the domain extent, $U=U(x,t)$ is the local oxide thickness, k is the segregation coefficient, m is the ratio of silicon thickness consumed to oxide thickness produced and n is the unit vector normal to each boundary.

Since numerical treatment of the problem has been the main goal of the paper, electric field induced flux of impurities as well as coupled impurity diffusion have been excluded for simplicity. Most of physical parameters in (1) and (2) have been modeled according to the program SUPREM [2].

To avoid the problems with discretization at the moving silicon--oxide interface $\eta=B(x,t)=m\cdot U(x,t)$, coordinate transformation [5]

$$y = \eta - B(x,t)$$

which transforms physical domain into the time independent rectangular domain has been used. This yields the following transformed diffusion equation:

$$\frac{\partial C}{\partial t} - \frac{\partial}{\partial x}(D\frac{\partial C}{\partial x}) - (1+B'^2)\frac{\partial}{\partial y}(D\frac{\partial C}{\partial y}) -$$

3)

$$- (B-D\cdot B'')\frac{\partial C}{\partial y} + B'\{\frac{\partial}{\partial x}(D\frac{\partial C}{\partial y}) + \frac{\partial}{\partial y}(D\frac{\partial C}{\partial x})\} = 0$$

ıe boundary condition (2) is converted into

$$D\{(1+B'^2)\frac{\partial C}{\partial y} - B'\frac{\partial C}{\partial x}\} - (k-m)\cdot B\cdot C = 0$$

ıile the other boundary conditions remain unchanged for symmetry
ɜasons.

The multigrid solution of the diffusion equation (1) could be also ɜrformed in physical domain without coordinate transformation [3]. This ɔproach, because of need for special boundary discretization control t the moving oxide-silicon interface and additional modifications in ɹltigrid components used still requires advanced multigrid techniques ɹs the local coordinate transformation [1] near the moving boundary.

The time discretization of the diffusion equation (3) is performed ıy the implicit backward Euler scheme. An automatic time step selection ɔased on Milne´s device [8] has been used. That implies three integration ɜteps in each time step: a crude step and two finer steps with integra-ɜion time half that of the crude step.

The spatial derivatives of (3) are discretized by 9-point central ɟifferences for the second order terms and by upwind differences for the ʃirst order convection terms. For discretization of the Neuman boundary ɔonditions so-called "mirror imaging" [7] has been used. This is the same ɟiscretization as inside the domain substitutuing the missing quantities ɔutside the domain using Neuman boundary conditions and linear interpo-lation.

Two-dimensional ion-implanted concentration profiles based on LSS ʇheory [7] have been used as initial solution for the first time step, ẇhile the following time steps use the results of previous ones as their initial solution.

3. MULTIGRID APPROACH

The discretization of the diffusion equation leads to a nonlinear algebraic system of equations. To solve this system we use a nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm. This approach is in many respects advantageous to linear multigrid method with

Correction Scheme already used for this application [6]. Inherently non-linear, FAS algorithm does not require global linearization of equations. Hence, no extra storage for coefficients of linearized equations is needed and the programing is very convenient. FAS algorithm also gives a natural way to estimate a local truncation error which could be useful for making an efficient stopping criterion for iteration. Finally, in future developments one can benefit from various advanced multigrid techniques as local refinements for grid adaption and local coordinate transformation at the silicon-oxide interface.

The FAS algorithm used employs sequence $\{G_k,H_k\}_{1\leq k\leq M}$ of uniform, non-staggered, rectangulars grids with corresponding meshsizes ($H_{k-1}=2H_k$) where k is the grid level. Regarding the discretized diffusion equation as a discrete eliptic problem $L_MC_M=F_M$ on the finest grid (G_M), it can be solved for the unknown grid function C_M starting with k=M the following recursive procedure

```
procedure FAS (k,ν1,ν2: integer; Ck,Fk: array);
var  j: integer; τk-1,Ck-1,Fk-1: array
begin
  if k=1 then Ck:=L^-1(Fk) else
    begin
      for j:=1 to ν1 do Ck:=Sk(Ck,Fk);
      Ck-1:=R(Ck);
      τk-1:=Lk-1(Ck-1)-R(LkCk);
      Fk-1:=R(Fk)+τk-1;
      FAS(k-1,ν1,ν2,Ck-1,Fk-1);
      Ck:=Ck-1+P(Ck-1-R(Ck));
      for j:=1 to ν2 do Ck:=Sk(Ck,Fk);
    end;
  if (k=M) and (||Fk-Lk(Ck)||>1/3||τk-1||) then
    FAS(k,ν1,ν2,Ck,Fk);
end;
```

For the purpose of smoothing (S_k), succesive-displacement Gauss-Siedel relaxation with lexicographical (LEX) and red-black (RB) ordering of points were tested. RB ordering of points is in some way advantageous because it could readily be fully parallelized. The only linearization required in FAS algorithm is that in smoothing process local to the corresponding grid point

For the purpose of this linearization, so-called "principal linearization" [1], which confines the original form of differential operator has been used.

Normal full weighting (9-point symmetric) formula [1] has been used as the restriction operator (R) which is natural for highly varying grid functions like impurity doping concentration. The prolongation (P) has been performed with a bilinear interpolation. The use of cubic rather than bilinear interpolation has not significantly ameliorated the situation.

On the coarsest grid (G_1), the solution (L^{-1}) is obtained with 5 iterations of S_1 type. The fine-to-coarse defect correction (τ_k) is used to estimate the local truncation error as $\tau_{M-1}/(2^p-1)$ [1] where p is the local approximation order of differential operator. This feature gives an efficient stopping criterion for terminating FAS algorithm.

4. SOLUTION EFFICIENCY

Very important question when the application of a multigrid algorithm is considered is the efficiency of the obtained solution. As a measure of solution efficiency we have considered an average residual error reduction per work unit (WU) i.e. per work equivalent to one relaxation over the finest grid level, which is onwards referred to as convergence rate (ρ_{WU}).

Regarding the generallity and robustness of the algorithm as one of its most significant attributes, our main concerns were the empirical prediction of behaviour of the convergence rate over a wide range of problem parameters and choice of smoothing technique. As a reference for comparation of the FAS algorithm convergence rate, equivalent single-grid (SG) iteration solver on the finest grid level has been used. One time step simulation of highly nonlinear neutral diffusion process from initially implanted arsenic layer has been chosen as an exemplary problem of numerical testing of convergence rates.

The behaviour of residual error L_2 norm for various time steps (Ts1-Ts4) and the different finest grid levels (M_1 and M_2) during solution procedure is shown in Figs.1 and 2, respectively. The dashed lines represent levels of estimated local truncation error norms for the given finest grid level. Relaxation has been performed with RB ordering of points. Table I gives calculated convergence rates for corresponding time steps and grid levels from Figs.1 and 2 with comparation of RB and LEX ordering of points in relaxation.

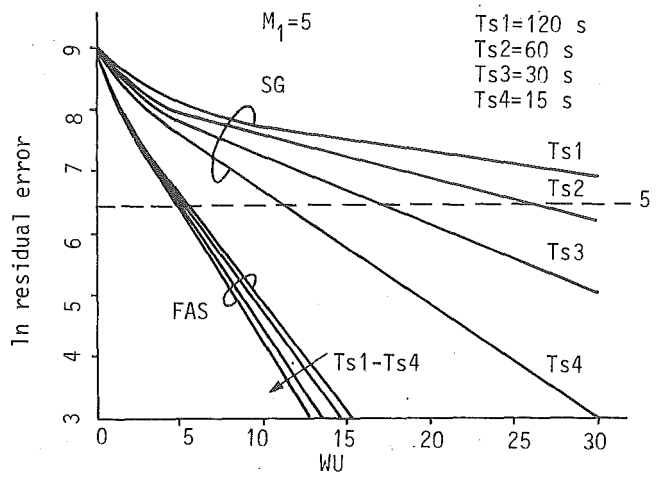


Fig.1 Single-grid and FAS residual restriction for different time step sizes

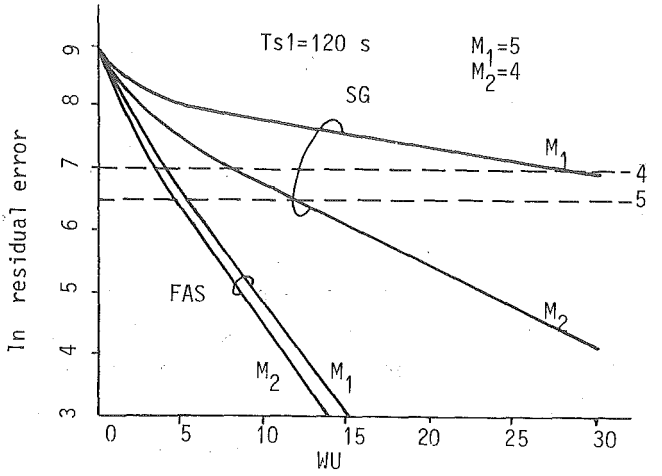


Fig.2 Single-grid and FAS residual restriction for the different finest grid levels

TABLE I Empirical convergence rates (ρ_{WU})

	M_1				Ts1	
	Ts_1	Ts_2	Ts_3	Ts_4	M_1	M_2
SG-RB	0.9019	0.8547	0.7711	0.6512	0.9019	0.6840
SG-LEX	0.9009	0.8574	0.7753	0.6623	0.9009	0.7185
FAS-RB	0.4364	0.4188	0.3987	0.3615	0.4364	0.3768
FAS-LEX	0.4827	0.4554	0.4204	0.3930	0.4827	0.4327

As an empirical observation, it is obvious that FAS algorithm solves a problem to the level of truncation error level in just a few work units which is commonly regarded as "normal" multigrid efficiency [4]. More important fact is that the FAS algorithm convergence rate is almost independent of time step size and choice of the finest grid level which means that for all practical simulation examples the computational cost of FAS algorithm is essentialy problem independent.

5. SAMPLE SIMULATION

Two diffusion process steps typical for fabrication of VLSI NMOS transistor structure are considered as practical examples of simulation. The first process step is the diffusion of boron channel-stop implant during local oxidation (LOCOS). The second process is a high-temperature anneal following arsenic implant for the source/drain region formation.

The boron channel-stop implant through a predefined field-oxide region, with the dose $5 \cdot 10^{12}$ cm^{-2} at 100 keV is followed by the 240 min field oxidation at $1000^{o}C$ in H_2O. On the other hand, arsenic was implanted with the dose 10^{16} cm^{-2} at 150 keV and driven-in for 15 min in neutral ambient at $1000^{o}C$.

The contour plots of the impurity concentrations at various stages of processing are shown in Figs.3 through 6. The boron channel-stop implant and source/drain arsenic implant distributions after the completion of the ion-implantation process step are shown in Figs.3 and 5, respectively. The silicon-nitride mask for boron channel-stop implantation extends from x=0 to x=1μm and for arsenic implantation from x=0 to x=0.25μm. Fig.4 shows the final boron profile after the local oxidation step. Initial oxide thickness for this process step was 0.05μm. Fig.5 shows the final arsenic distribution after the high-temperature anneal.

6. CONCLUSION

The main objective of the next generation VLSI process simulation programs is to reduce the design duration while simultaneously increasing the efficiency in achieving well-designed processes. Special attention should be paid on simulation of diffusion processes which are the most crucial and time consuming simulation steps.

In this paper we have presented an application of a nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm for solution

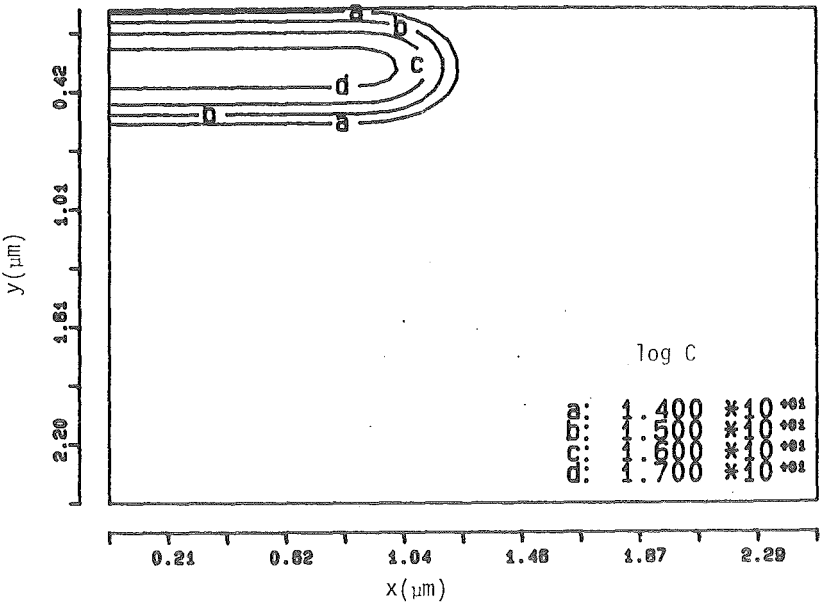


Fig.3 Boron channel-stop implant distribution

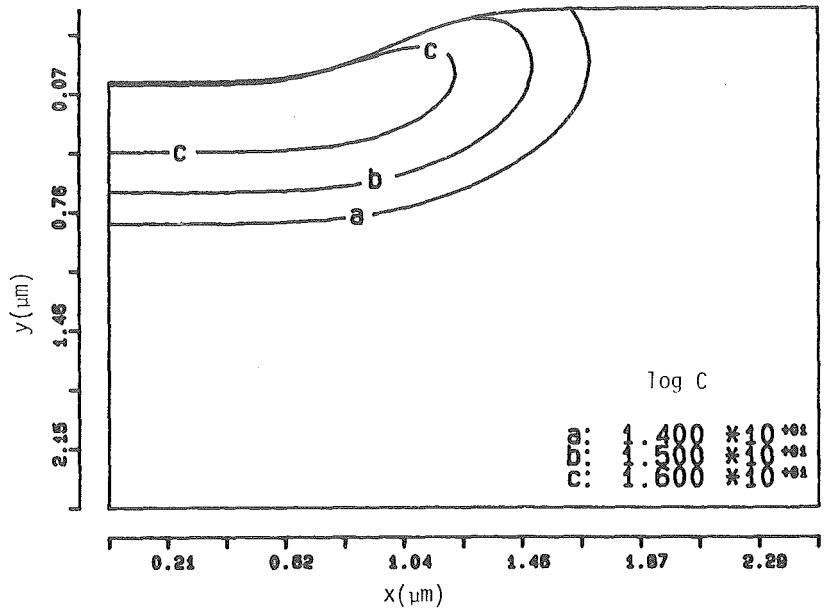


Fig.4 Boron distribution after local oxidation process step

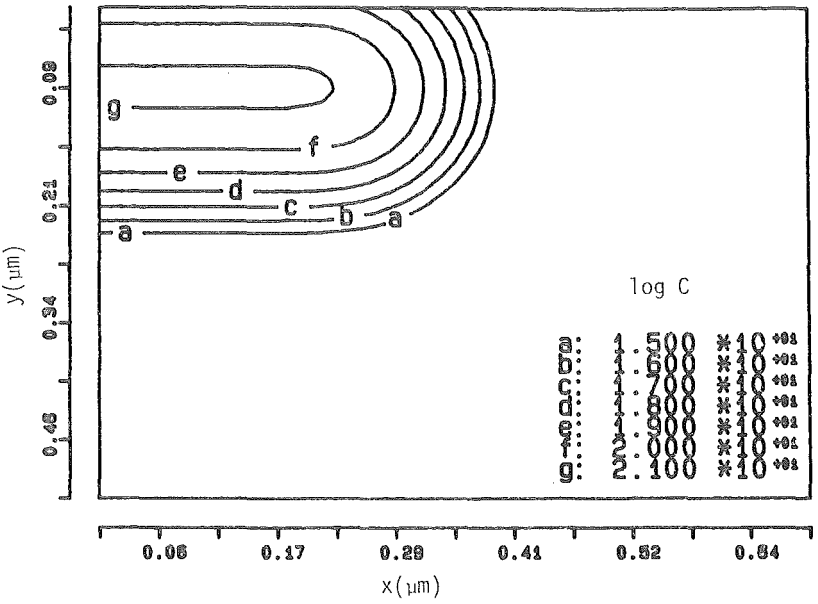


Fig.5 Arsenic source/drain implant distribution

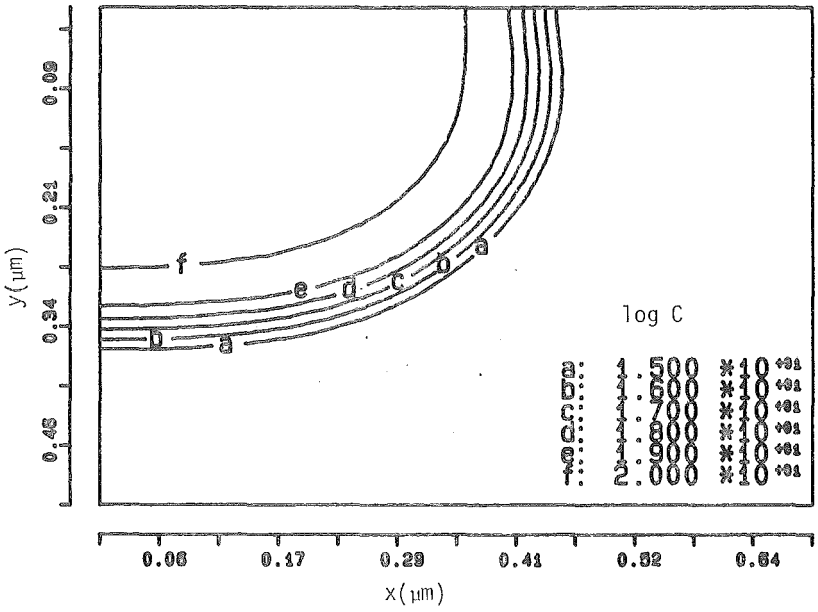


Fig.6 Arsenic distribution after high temperature anneal process step

of diffusion equation in VLSI process modeling. It has been demonstrated that nonlinear FAS algorithm shows high computational efficiency of solution which is almost independent of problem parameters in most practical applications.

Having in mind these features and possible extensions of FAS algorithm for more complex problems this numerical approach could be very effective for the next generation of process simulation programs.

7. REFERENCES

1. A. BRANDT: Multigrid Technique: 1984 Guide with Application to Fluid Dynamics. GMD-Studien Nr.85. Bonn, W. Germany.

2. C. HO, D. PLUMMER, S. HANSEN and R. DUTTON: VLSI Process Modeling-- SUPREM III. IEEE Trans. Electron Device ED-30 (1983), 1438-1453.

3. W. JOPPICH: A Multigrid Method for Solving the Nonlinear Diffusion Equation on a Time-Dependent Domain Using Rectangular Grids in Cartesian Coordinates. In: Proceedings of NASECODE V Conference. Dublin: Boole Press 1987.

4. S. MIJALKOVIĆ and N. STOJADINOVIĆ: Multigrid Method: An Efficient Numerical Tool in VLSI Process Modeling. In: Proceedings of First International Conference on Computer Technology, Systems and Application. Hamburg, 1987, 508-509.

5. B. PENUMALLI: A Comprehensive Two-Dimensional VLSI Process Simulation Program, BICEPS. IEEE Trans. Electron Device ED-30 (1983), 986-992.

6. A. SEIDL: A Multigrid Method for Solution of the Diffusion Equation in VLSI Process Modeling. IEEE Trans. Electron Device ED-30 (1983), 999-1004.

7. S. SELBERHERR: Analysis and Simulation of Semiconductor Devices. Springer-Verlag, Wien 1984.

8. H. YEAGER and R. DUTTON: An Approach to Solving Multiparticle Diffusion Exhibiting Nonlinear Stiff Coupling. IEEE Trans. Electron Device ED-32 (1985) 1964-1975.

CONSTRUCTION OF s - ORTHOGONAL POLYNOMIALS AND TURÁN QUADRATURE FORMULAE

GRADIMIR V. MILOVANOVIĆ

ABSTRACT: *A connection between Turán quadratures and s-orthogonal polynomials with respect to a nonnegative measure on the real line $\mathbb{R}$ is given. Using a discretized Stieltjes procedure and the Newton-Kantorowič method, an iterative method with quadratic convergence for the construction of s-orthogonal polynomials is formulated. Some numerical examples are included. Finally, some considerations about Turán quadrature formulae with Chebyshev measure are given.*

1. INTRODUCTION

In 1950 P. Turán investigated numerical quadratures of the type

$$\int_{-1}^{1} f(t)\,dt = \sum_{\nu=1}^{n} \sum_{i=0}^{k-1} A_{i,\nu} f^{(i)}(\tau_\nu) + R_{n,k}(f), \tag{1.1}$$

where

$$A_{i,\nu} = \int_{-1}^{1} \ell_{\nu,i}(t)\,dt \qquad (\nu=1,\ldots,n;\ i=0,1,\ldots,k-1)$$

and $\ell_{\nu,i}(t)$ are the fundamental functions of Hermite interpolation. The $A_{i,\nu}$ are Cotes numbers of higher order. The formula (1.1) is exact if f is a polynomial of degree at most kn-1 and the points $-1 \leq \tau_1 < \tau_2 < \ldots < \tau_n \leq 1$ are arbitrary.

For k=1 the formula (1.1), i.e.,

$$\int_{-1}^{1} f(t)\,dt = \sum_{\nu=1}^{n} A_{0,\nu} f(\tau_\nu) + R_{n,1}(f),$$

can be exact for all polynomials of degree $\leq 2n-1$ if the nodes τ_ν are the zeros of the Legendre polynomial P_n. That is the well-known Gauss-Legendre quadrature.

Because of the theorem of Gauss it is natural to ask whether knots τ_ν can be chosen so that the quadrature formula (1.1) will be exact for polynomials of degree not exceeding $(k+1)n-1$. P. Turán [17] showed that the answer is negative for $k=2$, and for $k=3$ it is positive. He proved that the knots τ_ν should be chosen as the zeros of the monic polynomial $\pi_n^*(t) = t^n + \dots$ which minimizes the following integral

$$\int_{-1}^{1} \pi_n(t)^4 dt,$$

where $\pi_n(t) = t^n + a_{n-1}t^{n-1} + \dots + a_1 t + a_0$.

More generally, the answer is negative for even, and positive for odd k, and then τ_ν are the zeros of the polynomial minimizing

$$\int_{-1}^{1} \pi_n(t)^{k+1} dt. \tag{1.2}$$

For $k=1$, π_n is the monic Legendre polynomial P_n.

Because of the above, we put $k=2s+1$. Instead of (1.1), it is also interesting to investigate the analogous formula with a weight function $t \to p(t)$,

$$\int_{-1}^{1} f(t)p(t)dt = \sum_{i=0}^{2s} \sum_{\nu=1}^{n} A_{i,\nu} f^{(i)}(\tau_\nu) + R(f),$$

or more generally, with some nonnegative measure $d\lambda(t)$ on the real line $\mathbb{R}$,

$$\int_{\mathbb{R}} f(t)d\lambda(t) = \sum_{i=0}^{2s} \sum_{\nu=1}^{n} A_{i,\nu} f^{(i)}(\tau_\nu) + R(f). \tag{1.3}$$

This paper is organized as follows. In Section 2 we give a connection between Turán quadratures and s-orthogonal polynomials, which were studied extensively by several Italian mathematicians [12], [7], [13], [14]. Also, in this section we mention a recent method of Vincenti [20] for the computation of the coefficients of s-orthogonal polynomials with respect to an even function. In Section 3 we develop a new method for the numerical construction of s-orthogonal polynomials with respect to an arbitrary weight function. Numerical examples are given in Section 4. Section 5 deals with Turán quadratures with Chebyshev measure.

2. TURAN QUADRATURES AND s-ORTHOGONAL POLYNOMIALS

We consider the Turán quadrature formula (1.3), where $d\lambda(t)$ is a nonnegative measure on the real line $\mathbb{R}$, with compact or infinite support, for which all moments $\mu_k = \int_{\mathbb{R}} t^k d\lambda(t)$, $k=0,1,\ldots$, exist and are finite, and $\mu_0 > 0$. The formula (1.3) must be exact for all polynomials of degree at most $2(s+1)n-1$. The role of the integral (1.2) is taken over by

$$F = \int_{\mathbb{R}} \pi_n(t)^{2s+2} d\lambda(t),$$

where $F \equiv F(a_0,\ldots,a_{n-1})$, $\pi_n(t) = \sum_{k=0}^{n} a_k t^k$, $a_n = 1$. In order to minimize F we must have

$$\int_{\mathbb{R}} \pi_n(t)^{2s+1} t^k d\lambda(t) = 0, \quad k=0,1,\ldots,n-1. \tag{2.1}$$

Usually, instead of $\pi_n(t)$ we write $P_{s,n}(t)$.

The case $d\lambda(t) = p(t)dt$ on $[a,b]$ has been considered by the Italian mathematicians A.Ossicini [12], A.Ghizzetti and A.Ossicini [7], S.Guerra [8], [9]. It is known that there exists a unique

$P_{s,n}(t) = \prod_{\nu=1}^{n} (t-\tau_\nu)$, whose zeros τ_ν are real, distinct and located in the interior of the interval [a,b]. These polynomials are known as s-orthogonal (or s-self associated) polynomials in the interval [a,b] with respect to the weight function p.

For s=0 we have the standard case of orthogonal polynomials. The case when s>0 is very difficult. It requires the use of a method with special numerical treatment.

Recently G.Vincenti [20] has considered an iterative process to compute the coefficients of s-orthogonal polynomials in a special case when the interval [a,b] is symmetric with respect to origin, say, [-b,b], and the weight function p is an even function p(-t)=p(t). Then $P_{s,n}(-t) = (-1)^n P_{s,n}(t)$. He considered two cases: when n is even and when n is odd.

In the first case n=2m, $P_{s,n}(t) = \sum_{i=0}^{m} a_i t^{2m-2i}$, $a_0 = 1$. From (2.1) Vincenti obtained a nonlinear system of equations of the form

$$\sum_{i=0}^{m} C_{m+r-i} a_i = -C_{m+r} \qquad (r=0,1,\ldots,m-1),$$

where

$$C_j^{(0)} = \int_0^b p(t) t^{2j} dt, \qquad C_j^{(h)} = \sum_{p,q=0}^{m} C_{j+2m-p-q}^{(h-1)} a_p a_q,$$

and $C_j^{(s)} \equiv C_j$. Then he has solved this system by some iterative method like Newton's method. For n=2m+1, a similar system of equations was obtained.

Vincenti applied his process to the Legendre case. When n and s increase, the process becomes ill-conditioned. So, the author gave numerical results in the following cases: n=2,3, $1 \leq s \leq 10$; n=4,5, $1 \leq s \leq 5$; n=6,7, $1 \leq s \leq 3$; n=8,9, $1 \leq s \leq 2$; n=10,11, s=1. The results were obtained with 18 correct decimal digits,

but using an arithmetic with 36 decimal digits.

From (2.1) we can see that this procedure needs the first $2(s+1)n$ moments of the weight function: $\mu_0, \mu_1, \cdots, \mu_{2(s+1)n-1}$. We see that $c_j^{(0)} = \mu_{2j}/2$. Of course, in this special case, the moments of odd order are zero. Here, we have a nonlinear map $V_{n,s}: \mathbb{R}^{2(s+1)n} \to \mathbb{R}^n$, given by $[\mu_0, \mu_1, \ldots, \mu_{2(s+1)n-1}]^T \to [a_0, a_1, \ldots, a_{n-1}]^T$. The problem itself is highly sensitive to small perturbations in the moments, so that any algorithm which theoretically solves the problem using the moments will be subject to severe growth of errors when executed in an arithmetic of finite precision ([4],[5]). It would be useful to find a numerical condition number of the map $V_{n,s}$, but that will not be our aim here.

3. CONSTRUCTION OF s-ORTHOGONAL POLYNOMIALS

In this section we will give a stable procedure for the numerical construction of s-orthogonal polynomials with respect to $d\lambda(t)$ on $\mathbb{R}$. Namely, we will reduce our problem to the standard theory of orthogonal polynomials, and then we will use the Stieltjes procedure ([3],[5]). The main idea is an interpretation of the "orthogonality conditions" (2.1), i.e.,

$$\int_{\mathbb{R}} \pi_n(t)\, t^k\, \pi_n(t)^{2s}\, d\lambda(t) = 0, \qquad k=0,1,\ldots,n-1.$$

For given n and s, we put $d\mu(t) = d\mu^{s,n}(t) = (\pi_n(t))^{2s} d\lambda(t)$. These conditions can be interpreted as

$$\int_{\mathbb{R}} \pi_k^{s,n}(t)\, t^\nu\, d\mu(t) = 0, \qquad \nu=0,1,\ldots,k-1,$$

where $(\pi_k^{s,n})$ is a sequence of monic orthogonal polynomials with respect to the new measure $d\mu(t)$. Of course, $P_{s,n}(.) = \pi_n^{s,n}(.)$.

As we can see, the polynomials $\pi_k^{s,n}$, $k=0,1,\ldots$, are implicitly defined, because the measure $d\mu(t)$ depends of $\pi_n^{s,n}(t)$. The general class of such polynomials was introduced by H.Engels (see [2, pp. 214-226]).

We will write only $\pi_k(.)$ instead of $\pi_k^{s,n}(.)$. These polynomials satisfy a three-term recurrence relation

$$\pi_{k+1} = (t-\alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \quad k=0,1,\ldots, \tag{3.1}$$

$$\pi_{-1}(t) = 0, \quad \pi_0(t) = 1,$$

where, because of orthogonality,

$$\alpha_k = \alpha_k(s,n) = \frac{\langle t\pi_k, \pi_k\rangle}{\langle \pi_k, \pi_k\rangle} = \frac{\int_{\mathbb{R}} t\pi_k^2(t)\,d\mu(t)}{\int_{\mathbb{R}} \pi_k^2(t)\,d\mu(t)}, \tag{3.2}$$

$$\beta_k = \beta_k(s,n) = \frac{\langle \pi_k, \pi_k\rangle}{\langle \pi_{k-1}, \pi_{k-1}\rangle} = \frac{\int_{\mathbb{R}} \pi_k^2(t)\,d\mu(t)}{\int_{\mathbb{R}} \pi_{k-1}^2(t)\,d\mu(t)},$$

and, for example, $\beta_0 = \int_{\mathbb{R}} d\mu(t)$.

The coefficients α_k and β_k are the fundamental quantities in the constructive theory of orthogonal polynomials. They provide a compact way of representing orthogonal polynomials, requiring only a linear array of parameters. The coefficients of orthogonal polynomials, or their zeros, in contrast need two-dimensional arrays.

Finding the coefficients α_k, β_k $(k=0,1,\ldots,n-1)$ gives us access to the first n+1 orthogonal polynomials $\pi_0, \pi_1, \ldots, \pi_n$. Of course, for a given n, we are interested only in the last of them i.e., $\pi_n (\equiv \pi_n^{s,n})$. So, for $n=0,1,\ldots$, the diagonal (boxed) elements

In the following table are our s-orthogonal polynomials $\pi_n^{s,n}$.

TABLE 3.1

n	$d\mu^{s,n}(t)$	Orthogonal Polynomials
0	$(\pi_0^{s,0}(t))^{2s} d\lambda(t)$	$\boxed{\pi_0^{s,0}}$
1	$(\pi_1^{s,1}(t))^{2s} d\lambda(t)$	$\pi_0^{s,1}$ $\boxed{\pi_1^{s,1}}$
2	$(\pi_2^{s,2}(t))^{2s} d\lambda(t)$	$\pi_0^{s,2}$ $\pi_1^{s,2}$ $\boxed{\pi_2^{s,2}}$
3	$(\pi_3^{s,3}(t))^{2s} d\lambda(t)$	$\pi_0^{s,3}$ $\pi_1^{s,3}$ $\pi_2^{s,3}$ $\boxed{\pi_3^{s,3}}$
⋮		

A stable procedure for finding the coefficients α_k, β_k is the discretized Stieltjes procedure, especially for infinite intervals of orthogonality (see Gautschi [5], and Gautschi, Milovanović [6]). Unfortunately, in our case this procedure cannot be used directly, because the measure $d\mu(t)$ involves an unknown polynomial $\pi_n^{s,n}$. Consequently, we consider the system of nonlinear equations

$$f_0 \equiv \beta_0 - \int_{\mathbf{R}} \pi_n^{2s}(t)\, d\lambda(t) = 0,$$

$$f_{2k+1} \equiv \int_{\mathbf{R}} (\alpha_k - t)\pi_k^2(t)\pi_n^{2s}(t)\, d\lambda(t) = 0, \quad k=0,1,\ldots,n-1, \tag{3.3}$$

$$f_{2k} \equiv \int_{\mathbf{R}} (\beta_k \pi_{k-1}^2(t) - \pi_k^2(t))\pi_n^{2s}(t)\, d\lambda(t) = 0, \quad k=1,\ldots,n-1,$$

which follows from (3.2).

Let x be a $(2n)$-dimensional column vector with components $\alpha_0, \beta_0, \ldots, \alpha_{n-1}, \beta_{n-1}$ and $f(x)$ a $(2n)$-dimensional vector with components $f_0, f_1, \ldots, f_{2n-1}$, given by (3.3). If $W = W(x)$ is the corresponding Jacobi matrix of $f(x)$, then we can apply Newton-Kantorovič's method

$$x^{[\nu+1]} = x^{[\nu]} - W^{-1}(x^{[\nu]})\, f(x^{[\nu]}), \qquad \nu = 0,1,\ldots, \tag{3.4}$$

for determining the coefficients of the recurrence relation (3.1) Starting with a reasonable good approximation $x^{[0]}$, the convergence of the method (3.4) is quadratic.

It is interesting that the elements of Jacobi matrix can be easily computed in the following way:

First, we have to determine the partial derivatives $a_{k,i} = \frac{\partial \pi_k}{\partial \alpha_i}$ and $b_{k,i} = \frac{\partial \pi_k}{\partial \beta_i}$. Differentiating the recurrence relation (3.1) with respect to α_i and β_i we obtain

$$a_{k+1,i} = (t - \alpha_k) a_{k,i} - \beta_k a_{k-1,i} ,$$

and

$$b_{k+1,i} = (t - \alpha_k) b_{k,i} - \beta_k b_{k-1,i} ,$$

where

$$a_{k,i} = 0, \quad b_{k,i} = 0, \quad k \le i,$$

$$a_{i+1,i} = -\pi_i(t), \quad b_{i+1,i} = -\pi_{i-1}(t).$$

These relations are the same as those for π_k, but with other initial values. The elements of the Jacobi matrix are

$$\frac{\partial f_{2k+1}}{\partial \alpha_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t) \left[(\alpha_k - t) p_{k,i}(t) + \frac{1}{2} \delta_{ki} \pi_k^2(t) \pi_n(t) \right] d\lambda(t),$$

$$\frac{\partial f_{2k+1}}{\partial \beta_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t)(\alpha_k - t) q_{k,i}(t) d\lambda(t),$$

3.5)
$$\frac{\partial f_{2k}}{\partial \alpha_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t)(\beta_k p_{k-1,i}(t) - p_{k,i}(t)) d\lambda(t),$$

$$\frac{\partial f_{2k}}{\partial \beta_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t)\{(\beta_k q_{k-1,i}(t) - q_{k,i}(t)) + \frac{1}{2}\delta_{ki}\pi_{k-1}^2(t)\pi_n(t)\} d\lambda(t),$$

here $p_{k,i}(t) = \pi_k(t)(a_{k,i}\pi_n(t) + sa_{n,i}\pi_k(t))$ and $q_{k,i}(t) = k(t)(b_{k,i}\pi_n(t) + sb_{n,i}\pi_k(t))$, and δ_{ki} is Kronecker's delta.

All of the above integrals in (3.3) and (3.5) can be found xactly, except for rounding errors, by using a Gauss-Christoffel uadrature formula with respect to the measure $d\lambda(t)$,

3.6)
$$\int_{\mathbb{R}} g(t) d\lambda(t) = \sum_{k=1}^{N} A_k^{(N)} g(\tau_k^{(N)}) + R_N(g),$$

aking $N=(s+1)n$ knots. This formula is exact for all polynomials f degree at most $2N-1 = 2(s+1)n - 1 = 2(n-1) + 2ns + 1$.

Thus, for all calculations we use only the fundamental three-erm recurrence relation and the Gauss-Christoffel quadrature 3.6). As initial values $\alpha_k^{[0]} = \alpha_k^{[0]}(s,n)$ and $\beta_k^{[0]} = \beta_k^{[0]}(s,n)$ we ta-e the values obtained for n-1, i.e. $\alpha_k^{[0]} = \alpha_k(s,n-1)$, $\beta_k^{[0]} = \beta_k(s,n-1)$, ≦ n-2. For α_{n-1} and β_{n-1} we use the corresponding extrapolated alues.

In the case n=1 we solve the equation

$$\Phi(\alpha_0) = \Phi(\alpha_0(s,1)) = \int_{\mathbb{R}} (t-\alpha_0)^{2s+1} d\lambda(t) = 0,$$

and then determine

$$\beta_0 = \beta_0(s,1) = \int_{\mathbb{R}} (t-\alpha_0)^{2s} d\lambda(t).$$

4. NUMERICAL EXAMPLES

We will consider two examples, involving Laguerre and Legendre measures.

Example 4.1. $d\lambda(t) = e^{-t}dt$ on $(0,\infty)$.

Using the presented method, we determined the recursion coefficients $\alpha_k(s,n)$ and $\beta_k(s,n)$, $k=0,1,\ldots,n-1$, for $s = 1(1)5$ and $n = 1(1)10$. These coefficients and zeros of $\pi_n^{s,n}$, $\tau_k(s,n)$, $k=1,\ldots,n$, for some selected values of s and n, are given in Table 4.1. Numbers in parentheses denote decimal exponents.The zeros $\tau_k(s,n)$, $k=1,\ldots,n$, were obtained as eigenvalues of the (symmetric tridiagonal) Jacobi matrix

$$J_n = \begin{bmatrix} \alpha_0(s,n) & \sqrt{\beta_1(s,n)} & & & O \\ \sqrt{\beta_1(s,n)} & \alpha_1(s,n) & \sqrt{\beta_2(s,n)} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \sqrt{\beta_{n-1}(s,n)} \\ O & & & \sqrt{\beta_{n-1}(s,n)} & \alpha_{n-1}(s,n) \end{bmatrix},$$

using the QR algorithm.

Example 4.2. $d\lambda(t) = dt$ on $(-1,1)$. In this (Legendre) cas the coefficients $\alpha_k(s,n)$ are equal to zero, so the computatio can be simplified. The system of equations (3.3) becomes

$$g_0 = f_0 = \beta_0 - \int_{-1}^{1} \pi_n^{2s}(t)dt = 0,$$

$$g_k = f_{2k} = \int_{-1}^{1} (\beta_k \pi_{k-1}^2(t) - \pi_k^2(t))\pi_n^{2s}(t)dt = 0, \quad k=1,\ldots,n$$

TABLE 4.1

s,n)	k	$\alpha_k(s,n)$	$\beta_k(s,n)$	$\tau_{k+1}(s,n)$
(1,5)	0	1.53297437454020(0)	1.95429735674308(6)	3.8619211523014(-1)
	1	5.58879530809235(0)	3.09769990936949(0)	2.5326808971664(0)
	2	9.67825960904726(0)	1.44873867444755(1)	6.8055964648137(0)
	3	1.38195768909663(1)	3.44094720328124(1)	1.3770543148954(1)
	4	1.79144743230187(1)	6.31554867162230(1)	2.5039067879500(1)
1,10)	0	1.51947559720794(0)	1.15245141095965(18)	1.9845989648554(-1)
	1	5.54285984605682(0)	3.04910058102535(0)	1.2852724641604(0)
	2	9.57433648078956(0)	1.42156585447179(1)	3.3633782573586(0)
	3	1.36134600304330(1)	3.35373242373804(1)	6.4866460030154(0)
	4	1.76626196755034(1)	6.10680235778704(1)	1.0743607524688(1)
	5	2.17262963633088(1)	9.68925818155147(1)	1.6274303555431(1)
	6	2.58125737578751(1)	1.41154607302825(2)	2.3303521691882(1)
	7	2.99352863173826(1)	1.94114444641607(2)	3.2216061440735(1)
	8	3.40937486293287(1)	2.56171126328495(2)	4.3764898673766(1)
	9	3.80755964787433(1)	3.26547484073315(2)	5.9920103669108(1)
(2,5)	0	3.06241261660323(0)	1.11900724691562(16)	5.1108081782716(-1)
	1	8.17357215072018(0)	6.27220780166492(0)	3.6504048515689(0)
	2	1.43542025111386(1)	3.14187808183856(1)	1.0011553444478(1)
	3	2.06411614818251(1)	7.61775799352481(1)	2.0452776123775(1)
	4	2.68361238086797(1)	1.41467716850165(2)	3.7441657331318(1)
(3,5)	0	2.58905931144849(0)	5.71776101144993(27)	6.3593164870754(-1)
	1	1.07564139072170(1)	1.05185172722828(1)	4.7669589415140(0)
	2	1.90289971948242(1)	5.47855138833478(1)	1.3215882166030(1)
	3	2.74628973371076(1)	1.34292199752058(2)	2.7133552841620(1)
	4	3.57594914086306(1)	2.50922121763235(2)	4.9844533561357(1)
(4,5)	0	3.11368201971988(0)	6.65045548992180(40)	7.6048752765420(-1)
	1	1.33381242208130(1)	1.58333974393260(1)	5.8827138815968(0)
	2	2.37031258589862(1)	8.45825858503624(1)	1.6419218171525(1)
	3	3.42845650702239(1)	2.08746777684076(2)	3.3813401673707(1)
	4	4.46834996793661(1)	3.91510787488863(2)	6.2247175594627(1)
(5,5)	0	3.63680292296229(0)	8.46508537128994(54)	8.8474548516636(-1)
	1	1.59190911806156(1)	2.22147113900203(1)	6.9978980073417(0)
	2	2.83768214565758(1)	1.20806800183997(2)	1.9621882995226(1)
	3	4.11061379266411(1)	2.99537220959448(2)	4.0492610101210(1)
	4	5.36078046528124(1)	5.63228814211952(2)	7.4649521550663(1)

Table 4.2 shows the numerical results for s = 1,3,5 and n = 3,5,10. The corresponding zeros $\tau_\nu(s,n)$, k=1,...,n, are given in Table 4.3.

TABLE 4.2

n	ν	$\beta_\nu(1,n)$	$\beta_\nu(3,n)$	$\beta_\nu(5,n)$
3	0	0.483864899809040(-1)	0.999799077102820(-4)	0.284169237312933(-6)
	1	0.396390615424778	0.438361519822241	0.455125737914133
	2	0.266920571579793	0.262372968797798	0.259637334393080
5	0	0.313354730979678(-2)	0.264465724288258(-7)	0.301618113315945(-13)
	1	0.397514379556632	0.440125755974452	0.456936553362545
	2	0.266421480435867	0.261489083023563	0.258693332791772
	3	0.256509353896241	0.254475851257394	0.253414689828449
	4	0.253674592138278	0.252629769731300	0.252061944536419
10	0	0.314536690060498(-5)	0.261903853328827(-16)	0.290667534992279(-27)
	1	0.398771414276302	0.442152192689833	0.459032427879297
	2	0.266409589288295	0.261261065487811	0.258382986818575
	3	0.256307280251967	0.254101849534999	0.253013674028616
	4	0.253361155621508	0.252167886595534	0.251600165348871
	5	0.252110174900276	0.251373736923891	0.251025244228691
	6	0.251467087710631	0.250973891152692	0.250737300372080
	7	0.251096334167793	0.250747641830448	0.250575234680807
	8	0.250866757894766	0.250611009696396	0.250478257846294
	9	0.250718964459874	0.250526857803099	0.250419693077896

TABLE 4.3

n	ν	$\tau_\nu(1,n)$	$\tau_\nu(3,n)$	$\tau_\nu(5,n)$
3	1,3	±0.81443918557776	±0.83709885235857	±0.84543661637477
	2	0.	0.	0.
5	1,5	±0.92711786960989	±0.93810619284349	±0.94197468869998
	2,4	±0.56086741916164	±0.57330378590709	±0.57774579736053
	3	0.	0.	0.
10	1,10	±0.98066259593659	±0.98398991804138	±0.98512298236202
	2,9	±0.87750022098482	±0.88396182054293	±0.88618806147381
	3,8	±0.69262442514005	±0.69957700233546	±0.70197668437523
	4,7	±0.44320099195064	±0.44838741280314	±0.45017897460267
	5,6	±0.15247058767942	±0.15437687188524	±0.15503560566469

5. TURÁN QUADRATURES WITH CHEBYSHEV WEIGHT

Now, we will consider again the quadrature formula of Turán (1.3). If we define ω_ν, by

$$\omega_\nu(t) = \left(\frac{\pi_n(t)}{t-\tau_\nu}\right)^{2s+1}, \quad \nu=1,\dots,n,$$

where $\pi_n(t)=\pi_n^{s,n}(t)$ and $\tau_\nu=\tau_\nu(s,n)$, then the coefficients $A_{i,\nu}$ in Turán quadrature (1.3) can be expressed in the form [16]

$$A_{i,\nu} = \frac{1}{i!(2s-i)!}\left[D^{2s-i}\,\frac{1}{\omega_\nu(t)}\int_{\mathbb{R}}\frac{\pi_n(x)^{2s+1}-\pi_n(t)^{2s+1}}{x-t}\,d\lambda(x)\right]_{t=\tau_\nu},$$

where D is the standard differentiation operator. Especially, for i=2s, we have

$$A_{2s,\nu} = \frac{1}{(2s)!\,(\pi_n'(\tau_\nu))^{2s+1}}\int_{\mathbb{R}}\frac{\pi_n(x)^{2s+1}}{t-\tau_\nu}\,d\lambda(x),$$

i.e.,

$$A_{2s,\nu} = \frac{B_\nu^{(s)}}{(2s)!\,(\pi_n'(\tau_\nu))^{2s}}, \quad \nu=1,\dots,n, \tag{5.1}$$

where $B_\nu^{(s)}$ are the Christoffel numbers of the following quadrature (with respect to the measure $d\mu(t)=\pi_n^{2s}(t)d\lambda(t)$)

$$\int_{\mathbb{R}} g(t)\,d\mu(t) = \sum_{\nu=1}^{n} B_\nu^{(s)} g(\tau_\nu) + R_n(g), \quad R_n(\mathbb{P}_{2n-1})=0, \tag{5.2}$$

So we have $A_{2s,\nu}>0$.

The expressions for the other coefficients (i<2s) become very complicated.

For the numerical calculation we can use a triangular system of linear equations obtained from the formula (1.3) by replacing f

with the Newton polynomials: $1,\ t-\tau_1,\ \ldots,\ (t-\tau_1)^{2s+1}$,

$(t-\tau_1)^{2s+1}(t-\tau_2),\ldots,\ (t-\tau_1)^{2s+1}(t-\tau_2)^{2s+1}\ldots(t-\tau_n)^{2s}$.

Particularly interesting is the case of the Chebyshev weight

$$p(t) = (1-t^2)^{-1/2}.$$

In 1930, S. Bernstein [1] showed that $2^{1-n}T_n(t)$ minimizes all integrals of the form

$$\int_{-1}^{1} \frac{|\pi_n(t)|^{k+1}}{\sqrt{1-t^2}}\,dt, \qquad k \geq 0.$$

So the Turán-Chebyshev formula

$$(5.3)\qquad \int_{-1}^{1} (1-t^2)^{-1/2} f(t)\,dt = \sum_{i=0}^{2s}\sum_{\nu=1}^{n} A_{i,\nu} f^{(i)}(\tau_\nu) + R(f),$$

with $\tau_\nu = \cos\frac{(2\nu-1)\pi}{2n}$, $\nu=1,\ldots,n$, is exact for polynomials of degree not exceeding $2(s+1)n-1$. Turán has stated a problem of explicit determination of $A_{i,\nu}$ and its asymptotic behavior as $n\to\infty$ (Problem XXVI in [18]). In this regard, Micchelli and Rivlin ([11]) have proved the following characterization: If $f\in\mathbb{P}_{2(s+1)n-1}$ then

$$\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}\,dt = \frac{\pi}{n}\left\{\sum_{\nu=1}^{n} f(\tau_\nu) + \sum_{j=1}^{s} \alpha_j f'[\tau_1^{2j},\ldots,\tau_n^{2j}]\right\},$$

where

$$\alpha_j = (-1)^j \frac{\binom{-1/2}{j}}{2j\,4^{(n-1)j}}, \qquad j=1,2,\ldots,$$

and $g[y_1^r,\ldots,y_m^r]$ designate the devided difference of the function g, where each y_j is repeated r times.

For s=1, the solution of the Turán problem XXVI is given by

$$A_{0,\nu} = \frac{\pi}{n}\ ,\ A_{1,\nu} = -\frac{\pi\tau_\nu}{4n^3}\ ,\quad A_{2,\nu} = \frac{\pi}{4n^3}(1-\tau_\nu^2)\,.$$

In 1975 R.D. Riess [15],and in 1984 A.K. Varma[19], using very different methods, obtained the explicit solution of the Turán problem for s=2:

$$A_{0,\nu} = \frac{\pi}{n}\ ,\quad A_{1,\nu} = -\frac{\pi\tau_\nu}{64n^5}(20n^2-1)\,,\ A_{2,\nu} = \frac{\pi}{64n^5}[3+(20n^2-7)(1-\tau_\nu^2)],$$

$$A_{3,\nu} = -\frac{6\pi\tau_\nu}{64n^5}(1-\tau_\nu^2)\,,\quad A_{4,\nu} = \frac{\pi}{64n^5}(1-\tau_\nu^2)^2\,.$$

Notice that (5.1), for the Chebyshev weight, reduces to

$$A_{2s,\nu} = \frac{\pi}{4^s n^{2s+1}(s!)^2}(1-\tau_\nu^2)^s\,,\qquad \nu=1,\dots,n.$$

One simple answer to Turán question was given by O. Kis [10]. His result can be stated in the following form: If g is an even trigonometric polynomial of degree at most 2(s+1)n-1, then

$$\int_0^\pi g(\theta)\,d\theta = \frac{\pi}{n(s!)^2}\sum_{j=0}^{s}\frac{S_j}{4^j n^{2j}}\sum_{\nu=1}^{n} g^{(2j)}\left(\frac{2\nu-1}{2n}\pi\right),$$

where S_{s-j} $(j=0,1,\dots,s)$ denotes the symmetric elementary polynomials with respect to the numbers $1^2, 2^2, \dots, s^2$, i.e.,

$$S_s = 1,\quad S_{s-1} = 1^2+2^2+\dots+s^2,\ \dots,\ S_0 = 1^2\cdot 2^2\cdots s^2.$$

Consequently,

$$\int_{-1}^{1}(1-t^2)^{-1/2}f(t)\,dt = \frac{\pi}{n(s!)^2}\sum_{j=0}^{s}\frac{S_j}{4^j n^{2j}}\sum_{\nu=1}^{n}\left[D^{2j}f(\cos\theta)\right]_{\theta=\frac{2\nu-1}{2n}\pi}.$$

Using the expansion

$$D^{2k} f(\cos\theta) = \sum_{i=1}^{2k} a_{k,i}(t) f^{(i)}(t), \qquad \cos\theta = t, \quad k>0,$$

where the functions $a_{i,j} \equiv a_{i,j}(t)$ are given recursively by

$$a_{k+1,1} = (1-t^2)a'_{k,1} - ta_{k,1},$$

$$a_{k+1,2} = (1-t^2)a'_{k,2} - ta_{k,2} + 2(1-t^2)a_{k,1} - ta_{k,1},$$

$$a_{k+1,i} = (1-t^2)a'_{k,i} - ta_{k,i} + 2(1-t^2)a_{k,i-1} - ta_{k,i-1} + (1-t^2)a_{k,i-2}$$

$$(k = 3, \ldots, 2k),$$

$$a_{k+1,2k+1} = 2(1-t^2)a_{k,2k} - ta_{k,2k} + (1-t^2)a_{k,2k-1},$$

$$a_{k+1,2k+2} = (1-t^2)a_{k,2k},$$

with $a_{1,1} = -t$ and $a_{1,2} = 1-t^2$, we obtain the formula (5.3). For example, when s=3, we have

$$A_{0,\nu} = \frac{\pi}{n}, \quad A_{1,\nu} = \frac{\pi\tau_\nu}{2304n^7}(784n^4+56n^2-1),$$

$$A_{2,\nu} = \frac{\pi}{2304n^7}\left\{(784n^4-392n^2+31)(1-\tau_\nu^2) + 168n^2-15\right\},$$

$$A_{3,\nu} = -\frac{\pi\tau_\nu}{2304n^7}\left\{(336n^2-89)(1-\tau_\nu^2) + 15\right\},$$

$$A_{4,\nu} = \frac{\pi}{2304n^7}\left\{(56n^2-65)(1-\tau_\nu^2)^2 + 45(1-\tau_\nu^2)\right\},$$

$$A_{5,\nu} = \frac{\pi\tau_\nu}{2304n^7}\left\{674(1-\tau_\nu^2)^2 - 240(1-\tau_\nu^2)\right\}, \quad A_{6,\nu} = \frac{\pi}{2304n^7}(1-\tau_\nu^2)^3.$$

To conclude, we mention the corresponding formula (5.2) for the Chebyshev weight,

$$(5.4) \qquad \int_{-1}^{1} g(t)\frac{\hat{T}_n^{2s}(t)}{\sqrt{1-t^2}}dt = \frac{\pi}{4^s n}\binom{2s}{s}\sum_{\nu=1}^{n} g(\tau_\nu) + R_n(g),$$

where $\tau_\nu = \cos(2\nu-1)\frac{\pi}{2n}$, $\nu=1,\dots,n$. Note that all weights are equal, that is, the formula (5.4) is one of Chebyshev type.

The last formula can be reduced to a "cosinus" formula

$$\int_0^\pi f(\cos x)\cos^{2s}(nx)\,dx = \frac{\pi}{n4^s}\binom{2s}{s}\sum_{\nu=1}^{n} f(\cos(2\nu-1)\tfrac{\pi}{2n}) + R_n(f),$$

where $R_n(f) \equiv 0$ if $f\in\mathbb{P}_{2(s+1)n-1}$.

Acknowledgment. The author is grateful to Professor Walter Gautschi for his careful reading of the paper and useful suggestions for better formulations of the material.

REFERENCES

1. S. BERNSTEIN: Sur les polynomes orthogonaux relatifs à un segment fini. J. Math. Pures Appl. (9)9(1930), 127-177.

2. H. ENGELS: Numerical quadrature and cubature. Academic Press, London, 1980.

3. W. GAUTSCHI: Construction of Gauss-Christoffel quadrature formulas. Math. Comp. 22(1968), 251-270.

4. W. GAUTSCHI: A survey of Gauss-Christoffel quadrature formulae. In: E.B. Christoffel - The Influence of his Work on Mathematics and the Physical Sciences (P.L. Butzer and F. Fehèr, eds.), Birkhäuser Verlag, Basel, 1981, pp. 72-147.

5. W. GAUTSCHI: On generating orthogonal polynomials. SIAM J. Sci. Statist. Comput. 3(1982), 289-317.

6. W. GAUTSCHI and G.V. MILOVANOVIĆ: Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series. Math. Comp. 44(1985), 177-190.

7. A. GHIZZETTI and A. OSSICINI: Su un nuovo tipo di sviluppo di una funzione in serie di polinomi. Rend. Accad. Naz. Lincei (8)43(1967), 21-29.

8. S. GUERRA: Polinomi generati da successioni peso e teoremi di rappresentazione di una funzione in serie di tali polinomi. Rend. Ist. Mat. Univ. Trieste 8(1976), 172-194.

9. S. GUERRA: Su un determinante collegato ad un sistema di polinomi ortogonali. Rend. Ist. Mat. Univ. Trieste 10(1978), 66-79.

10. O. KIS: Remark on mechanical quadrature *(Russian)*. Acta Math. Acad. Sci. Hungar. 8(1957), 473-476.

11. C.A. MICCHELLI and T.J. RIVLIN: Turán formulae highest precision quadrature rules for Chebyshev coefficients. IBM J. Res. Develop. 16(1972), 372-379.

12. A. OSSICINI: Costruzione di formule di quadratura di tipo Gaussiano. Ann. Mat. Pura Appl. (4)72(1966), 213-238.

13. A. OSSICINI and F. ROSATI: Funzioni caratteristiche nelle formule di quadratura gaussiane con nodi multipli. Bull. Un. Mat. Ital. (4)11(1975), 224-237.

14. A. OSSICINI and F. ROSATI: Sulla convergenza dei funzionali ipergaussiani. Rend. Mat. (6)11(1978), 97-108.

15. R.D. RIESS: Gauss-Turán quadratures of Chebyshev type and error formulae. Computing 15(1975), 173-179.

16. D.D. STANCU: Asupra unor formule generale de integrare numerica. Acad. R. P. Romîne. Stud. Cerc. Mat. 9(1958), 209-216.

17. P. TURÁN: On the theory of the mechanical quadrature. Acta Sci. Math. Szeged. 12(1950), 30-37.

18. P. TURÁN: On some open problems of approximation theory. J. Approx. Theory 29(1980), 23-85.

19. A.K. VARMA: On optimal quadrature formulae. Studia Sci. Math. Hungar. 19(1984), 437-446.

20. G. VINCENTI: On the computation of the coefficients of s-orthogonal polynomials. SIAM J. Numer. Anal. 23(1986), 1290-1294.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

ON SOME PARALLEL HIGHER-ORDER METHODS OF HALLEY'S TYPE FOR FINDING MULTIPLE POLYNOMIAL ZEROS

M.S. PETKOVIĆ and L.V. STEFANOVIĆ

ABSTRACT: *Using Newton's and Halley's corrections, some modifications of the iterative method for the simultaneous finding multiple complex zeros of a polynomial, based on the Halley-like algorithm, are obtained. The convergence order of the proposed methods is five and six, respectively. Further improvements of these methods are performed by applying the Gauss-Seidel approach. The lower bounds of the R-order of convergence for the accelerated (single-step) methods are also given. Faster convergence is attained without additional calculations which makes the proposed methods be very efficient. The considered iterative procedures are illustrated numerically in the example of an algebraic equation.*

1. ITERATION SCHEMES

Consider a monic polynomial of degre $n \geq 3$

$$P(z) = \prod_{i=1}^{\nu} (z - r_i)^{\mu_i}$$

with real or complex zeros $r_1,\dots,r_\nu$ having the order of multiplicity $\mu_1,\dots,\mu_\nu$ respectively, where $\mu_1 + \cdots + \mu_\nu = n$ $(\nu > 1)$. Let $z \in \mathbb{C}$ and

$$f_k(z) = \frac{P^{(k)}(z)}{P(z)} \qquad (k=1,2) ,$$

$$g(z) = \frac{f_1(z)}{2}\left(1 + \frac{1}{\mu}\right) - \frac{f_2(z)}{2 f_1(z)} ,$$

$$S_i(a,b) = \frac{1}{\mu_i}\left[\sum_{j=1}^{i-1} \mu_j (z-a_j)^{-1} + \sum_{j=i+1}^{\nu} \mu_j (z-b_j)^{-1}\right]^2 + \sum_{j=1}^{i-1} \mu_j (z-b_j)^{-2} + \sum_{j=i+1}^{\nu} \mu_j (z-b_j)^{-2} ,$$

where $a = (a_1,\dots,a_\nu)^T$ and $b = (b_1,\dots,b_\nu)^T$ are some vectors. In particular, according to the above, we have, for example,

$$S_i(a,a) = \frac{1}{\mu_i}\left[\sum_{\substack{j=1\\ j\neq i}}^{\nu} \mu_j (z-a_j)^{-1}\right]^2 + \sum_{\substack{j=1\\ j\neq i}}^{\nu} \mu_j (z-a_j)^{-2} .$$

Using the Bell's polynomials, X. Wang and S. Zheng have derived in [10] the following relations

$$(1) \qquad r_i = z - \left[g(z) - \frac{1}{2f_1(z)} S_i(r,r)\right]^{-1} \qquad (i=1,\ldots,\nu) ,$$

where $r = (r_1,\ldots,r_\nu)^T$ and $\mu=\mu_i$.

Assume that reasonably good approximations $z_1,\ldots,z_\nu$ of the zeros $r_1,\ldots,r_\nu$ were found. Letting $z=z_i$ and $r_i:=\hat{z}_i$ in (1), where $\hat{z}_i$ is the new approximation of the zero r_i, and taking certain approximations of r_j in S_i on the right-hand side of the relation (1), we obtain some iterative methods for simultaneous finding all zeros of the polynomial P.

We shall first define

$$N(z) = \mu/f_1(z) \qquad \text{(the Newton's correction)} ,$$

$$H(z) = 1/g(z) = 2\left[\frac{f_2(z)}{f_1(z)} - (1+\frac{1}{\mu})f_1(z)\right]^{-1}$$

(the Halley's correction)

and introduce the vectors

$$z = (z_1,\ldots,z_\nu)^T \qquad \text{(the former approximations)} ,$$

$$z_N = (z_{N,1},\ldots,z_{N,\nu})^T , \quad z_{N,i} = z_i - N(z_i)$$

(the Newton's approximations),

$$z_H = (z_{H,1},\ldots,z_{H,\nu})^T , \quad z_{H,i} = z_i - H(z_i)$$

(the Halley's approximations),

$$\hat{z} = (\hat{z}_1,\ldots,\hat{z}_\nu)^T \qquad \text{(the new approximations)} .$$

In calculating the approximations $z_{N,i}$ and $z_{H,i}$, as well in all formulas where the function $g(z)$ appears, one has to take $\mu=\mu_i$.

(TS) For $r_j:=z_j$ $(j\neq i)$ we obtain the total-step iteration (TS)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(z,z) \right]^{-1} \qquad (i=1,\ldots,\nu). \tag{2}$$

This method has been discussed in [10] (see, also [11]) as a special case obtained from the family of iterative methods. The formula (2) can be rewritten in the form

$$\hat{z}_i = z_i - H(z_i) \left[1 - \frac{H(z_i)}{2f_1(z_i)} S_i(z,z) \right]^{-1},$$

wherefrom we observe the similarity of the iterative method (2) (which has the convergence order equal to *four*) with the Halley's method (with *cubic* convergence)

$$\hat{z}_i = z_i - H(z_i)$$

for improvement of multiple zero r_i (see [2]). Thus, the correction term in the form of summs provides (i) the increase of convergence order and (ii) the determination of all zeros of a polynomial. Furthermore, we note that the formula (2) is more complicated to the square root iteration (which also has the convergence order equal to *four*, see, e.g. [7]), but (2) does not require the extraction of a root and the selection of appropriate value (of two values) of the square root.

(SS) Let $r_j:=\hat{z}_j$ $(j<i)$ and $r_j:=z_j$ $(j>i)$ (the Gauss-Seidel approach), then we obtain from (1) the single-step iteration (SS)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(\hat{z},z) \right]^{-1} \qquad (i=1,\ldots,\nu). \tag{3}$$

(TSN) Letting $r_j := z_{N,j} = z_j - N(z_j)$ $(j \neq i)$ in (1), one obtains the total-step method with Newton's correction (TSN)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(z_N, z_N) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{4}$$

(SSN) Substituting $r_j := \hat{z}_j$ $(j < i)$, $r_j := z_{N,j} = z_j - N(z_j)$ $(j > i)$ in (1), we obtain the single-step method with Newton's correction (SSN)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(\hat{z}, z_N) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{5}$$

(TSH) Putting $r_j := z_{H,j} = z_j - H(z_j)$ $(j \neq i)$ in (1), similar as for TSN method, we obtain the total-step method with Halley's correction (TSH)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(z_H, z_H) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{6}$$

(SSH) TSH method can be accelerated using the Gauss-Seidel approach: setting $r_j := \hat{z}_j$ $(j < i)$, $r_j := z_{H,j} = z_j - H(z_j)$ $(j > i)$ in (1), we get the single-step method with Halley's correction

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(\hat{z}, z_H) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{7}$$

2. CONVERGENCE ORDER

In this section we shall consider the convergence order of the iterative schemes (2)-(7). For the single-step methods, where the new approximations are used immediately they become available, we shall apply the definition of the R-order of convergence (see [4]). The R-order of convergence of the iterative process IP with the limit point $r = (r_1,\ldots,r_\nu)^T$, where $r_1,\ldots,r_\nu$ are the polynomial zeros, will be denoted by $O_R(IP,r)$.

Let $u_i^{(m)}$ be a multiple of $|z_i^{(m)} - r_i|$ $(i=1,\ldots,\nu)$, where $m=0,1,\ldots$ is the iteration index. Using the technique applied in [1], [6] or that presented in [7], it can be shown that the iterative methods (2)-(7) belong to a class of iterative simul-

taneous methods for which the following relations can be derived under suitable initial conditions

$$(8)\qquad u_i^{(m+1)} = \frac{1}{\nu - 1}\, u_i^{(m)p}\left(\alpha \sum_{j<i} u_j^{(m+1)} + \sum_{j>i} u_j^{(m)q}\right)$$

$$(i=1,\ldots,\nu;\ p,q\in\mathbb{N},\ \alpha\in\{0,1\}).$$

As in [5], we introduce the order triplet $U(IP) = (p,\alpha,q)$ as a characteristic of the relations (8) for the iterative process IP. The integers p and q are the exponents of $u_i^{(m)}$, while $\alpha = 0$ for a TS method and $\alpha = 1$ in the case of an SS method.

In order to determine the convergence order of the algorithms (2)-(7), we shall use the following assertion proved in [5]. Assume that the starting approximations $z_1^{(0)},\ldots,z_\nu^{(0)}$ are chosen sufficiently close to the zeros $r_1,\ldots,r_\nu$ so that $u_i^{(0)} < 1$ $(i=1,\ldots,\nu)$. Then, for the iterative process IP with $U(IP) = (p,\alpha,q)$ we have

$$(9)\qquad \begin{aligned} O_R(IP,\mathbf{r}) &= p + q \qquad \text{if } \alpha = 0 \quad \text{(total-step method)},\\ O_R(IP,\mathbf{r}) &= p + t_\nu \qquad \text{if } \alpha = 1 \quad \text{(single-step method)}, \end{aligned}$$

where t_ν is the unique positive root of the equation

$$(10)\qquad t^\nu - tq^{\nu-1} - pq^{\nu-1} = 0 \quad .$$

Using the results presented in [5] and [8] we can find the following bounds for t_ν:

$$(11)\qquad q + \frac{pq}{(\nu-1)(p+q)} < t_\nu \leq q + \frac{2p}{1+\sqrt{1+\frac{4p}{q}}} \quad .$$

An extensive but elementary analysis, similar as in [1] or [5]-[7], shows that the iterative schemes (2)-(7) have the following characteristics:

$$U(TS) = (2,0,2)\ ,\ U(TSN) = (3,0,2)\ ,\ U(TSH) = (3,0,3)\ ,$$

$$U(SS) = (3,1,1)\ ,\ U(SSN) = (3,1,2)\ ,\ U(SSH) = (3,1,3)\ .$$

According to this and (9) we have the assertions:

THEOREM 1. *The convergence order of the total-step methods TS(2), TSN(4) and TSH(6) is four, five and six, respectively.*

THEOREM 2. *The R-order of convergence of the single-step methods SS(3), SSN(5) and SSH(7) is given by*

$$O_R(SS, r) \geq 3 + \tau_\nu \ ,$$

$$O_R(SSN, r) \geq 3 + x_\nu$$

and

$$O_R(SSH, r) \geq 3 + y_\nu \ ,$$

where τ_ν, x_ν *and* y_ν *are the unique positive roots of the equations*

$$\tau^\nu - \tau - 3 = 0 \ ,$$

$$x^\nu - x \cdot 2^{\nu-1} - 3 \cdot 2^{\nu-1} = 0$$

and

$$y^\nu - y \cdot 3^{\nu-1} - 3^\nu = 0 \ ,$$

respectively.

The values of the lower bounds of the R-order of convergence in the case of the single-step methods can be easily obtained solving the algebraic equation (10) starting from the interval given by (11). These values are displyed in Table 1 and coincide with that concerning the corresponding modifications of square-root iterations (the single-step versions, without or with the Newton's and Halley's corrections) (see [7]).

method \ ν	2	3	4	5	6	7	8	9	10
SS(3)	5.303	4.672	4.453	4.341	4.274	4.229	4.196	4.172	4.153
SSN(5)	6.646	5.862	5.585	5.443	5.357	5.299	5.257	5.225	5.200
SSH(7)	7.854	6.974	6.662	6.502	6.404	6.338	6.291	6.255	6.227

Table 1

3. NUMERICAL RESULTS

In order to test the presented iterative schemes, a FORTRAN routine was realised on a HONEYWELL 66 system in double-precision arithmetic (about 18 significant decimal digits). In realising the TSN, SSN, TSH and SSH methods with Newton's and Halley's corrections, before calculating new approximations $z_i^{(m+1)}$, the values $f_k(z_i^{(m)})$ $(k=1,2\ ;\ m=0,1,\ldots)$ were calculated. The same values are used for calculating the function

$$g(z_i^{(m)}) = \frac{1}{2}\ (1+\frac{1}{\mu_i})\ f_1(z_i^{(m)}) - \frac{1}{2}\cdot\frac{f_2(z_i^{(m)})}{f_1(z_i^{(m)})}\quad ,$$

the Newton's correction

$$N(z_i^{(m)}) = \frac{\mu_i}{f_1(z_i^{(m)})}$$

and Halley's correction

$$H(z_i^{(m)}) = \frac{1}{g(z_i^{(m)})} = 2\left[\frac{f_2(z_i^{(m)})}{f_1(z_i^{(m)})} - (1+\frac{1}{\mu_i})\ f_1(z_i^{(m)})\right]^{-1}.$$

Thus, the proposed iterative methods with Newton's and Halley's correction terms require slightly more numerical opperations in relation to the basic methods (algorithms (2) and (3)). Taking into account the significantly increased order of convergence, it is obvious that the proposed methods have a greater efficiency.

In order to illustrate numerically the efficiency of the modified methods, the algorithms TS(2), SS(3), TSN(4), SSN(5), TSH(6) and SSH(7) were applied for the improvement of zeros of the polynomial

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

The exact zeros of this polynomial are $r_1 = 1+2i$, $r_2 = 1-2i$, $r_3 = -1$ and $r_4 = 3$, with multiplicities $\mu_1 = 2$, $\mu_2 = 2$, $\mu_3 = 3$ and $\mu_4 = 2$. As initial approximations to these zeros the

following complex numbers were taken:

$$z_1^{(0)} = 1.7 + 2.7i, \quad z_2^{(0)} = 1.7 - 2.7i,$$
$$z_3^{(0)} = -0.3 - 0.7i, \quad z_4^{(0)} = 2.4 - 0.6i.$$

In spite of crude initial approximations ($\min_i |z_i^{(0)} - r_i| \cong 1$) the modified methods demonstrate very fast convergence. Numerical results, obtained in the second iteration, are given in Table 2.

method	i	Re $\{z_i^{(2)}\}$	Im $\{z_i^{(2)}\}$
TS(2)	1	0.999999703872727	1.999999577023530
	2	1.000004966234449	-1.999858354626263
	3	-1.000001724263487	1.28×10^{-6}
	4	3.000175153200852	4.58×10^{-5}
SS(3)	1	0.999999603833368	2.000000538829041
	2	0.999997513035036	-2.000168291520113
	3	-1.000001434643141	8.31×10^{-7}
	4	3.000000000400398	4.03×10^{-9}
TSN(4)	1	1.000005463270708	1.999990357789566
	2	1.000000009930465	-2.000000025656453
	3	-0.999999370541218	-2.44×10^{-7}
	4	2.999980969476169	5.24×10^{-6}
SSN(5)	1	0.999998904155992	1.999998927299469
	2	0.999999988521851	-1.999999982758255
	3	-1.000000001576391	5.07×10^{-9}
	4	3.000000000001254	-3.26×10^{-12}
TSH(6)	1	1.000000002444691	2.000000000565806
	2	1.000000002639924	-2.000000001014728
	3	-0.999999999964674	-2.81×10^{-12}
	4	3.000000003876174	-3.25×10^{-10}
SSH(7)	1	1.000000000020514	2.000000000101261
	2	1.000000000012086	-1.999999999988034
	3	-1.000000000000157	-2.86×10^{-13}
	4	3.000000000000029	-2.88×10^{-14}

Table 2

REFERENCES

. G. ALEFELD and J. HERZBERGER: *On the convergence speed of some algorithms for the simultaneous approximation of polynomial roots*. SIAM J. Numer. Anal. 11 (1974), 237-243.

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

. G. V. MILOVANOVIĆ and M. S. PETKOVIĆ: *On the convergence order of a modified method for simultaneous finding polynomial zeros*. Computing 30 (1983), 171-178.

. J. M. ORTEGA and W. C. RHEINBOLDT: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York 1970.

. M. S. PETKOVIĆ, G. V. MILOVANOVIĆ and L. V. STEFANOVIĆ: *Some higher-order methods for the simultaneous approximation of multiple polynomial zeros*. Comput. Math. Appls. 9 (1986), 951-962.

. M. S. PETKOVIĆ and L. V. STEFANOVIĆ: *On the convergence order of accelerated root iterations*. Numer. Math. 44 (1984), 463-476.

'. M. S. PETKOVIĆ and L. V. STEFANOVIĆ: *On some improvements of square root iteration for polynomial complex zeros*. J. Comput. Appl. Math. 15 (1986), 13-25.

3. L. V. STEFANOVIĆ: *Some iterative methods for the simultaneos finding of polynomial zeros* (in Serbo-Croatian). Ph. D. Thesis, University of Niš, Niš 1986.

). D. WANG and Y. WU: *Some modifications of the parallel Halley iteration method and their convergence*. Computing 38 (1987), 75-87.

). X. WANG and S. ZHENG: *A family of parallel and interval iteration for finding simultaneously all roots of a polynomial with rapid convergence (I)*. J. Comput. Math. 2 (1984), 70-76.

L. X. WANG and S. ZHENG: *A family of parallel and interval iteration for finding simultaneously all roots of a polynomial with rapid convergence (II)*. (in Chinese). J. Comput. Math. 4 (1985), 433-444.

Numerical Methods and Approximation Theory III (Niš, August 18-21, 1987)

PADÉ-APPROXIMATION AND BAND-LIMITED PROCESSES

TIBOR K. POGÁNY

ABSTRACT: In the paper we apply the Padé-approximation method to the approximation of spectral densities which are analytical at the origin.The observed densities are positive on a finite interval $I = [-w,w] \subset \mathbb{R}$ and vanish otherwise. Some results are given on the lower and upper bound of Padé-approximants on I and the convergence for some approximant sequences of the observed density was investigated.Related convergence results are given for the sequences of Padé-processes.

1. INTRODUCTION

The estimation theory of wide-sense stationary stochastic processes use the so called Wiener-Hopf equation,which Yaglom has solved explicitly for the class of processes with rational spectral densities.The Wiener-Hopf equation can be solved only in this class.

The concept of a band-limited process is an important one in practice.Many processes in applied sciences have a spectrum f(u) which is concetrated on a finite interval I.Practically these processes are band-limited: The harmonic oscillations $f(u)e^{iut}$ with frequencies u outside I have very small energy.

Because some Padé-approximants of the spectral density of a band-limited process have identical properties as the spectral densities,with the help of the convergence results for approximant sequences and Padé-process sequences,we may

map the rational approximation problem into a stochastic process class.

The final step is: solving the estimative problems for a rational, Padé-density class.

2. PRELIMINARIES AND SOME DEFINITIONS

Let $f(u)$ be a real function analytic at the origin. The Padé-approximant (in further PA) of order (L,M) of the function $f(u)$ is the rational expression $(L/M)_f(u) = P_L(u)/Q_M(u)$, $Q_M(o) = 1$, which has $L+M$-order contact with $f(u)$ at the origin. We can write:

$$Q_M(u)f(u) - P_L(u) = O(u^{L+M+1}) \tag{1}$$

or equivalently

$$Q_M(u)f(u) - P_L(u) = u^{L+M+1}\, h_{L,M}(u) \tag{2}$$

where $h_{L,M}(o) \neq o$. The coefficients of the polynomials $P_L(u) = \sum_{o}^{L} p_k u^k$, $Q_M(u) = \sum_{o}^{M} q_k u^k$ can be computed from (1), see for example [1].

The formal power series of $f(u)$ is

$$f(u) = \sum_{k=o}^{\infty} f_k u^k = \sum_{k=o}^{\infty} \frac{f^{(k)}(o)}{k!}\, u^k \quad , \tag{3}$$

where the series on the right side of (3) converges uniformly on the interval $(-r,r)$, $r^{-1} = \lim_{k\to\infty} \sup |f_k|^{1/k}$. When $r = +\infty$, $f(u)$ is an entire function; for $r = o$ the power series converges only at the origin and the power series (3) is formal.

The real function $f(u)$ is analytic on the interval $I =$

$= [a,b]$ if in some neighborhood $u_o - \delta < u < u_o + \delta$ of all points $u_o \in I$ there exists the expansion

$$f(u) = \sum_{k=o}^{\infty} f_k (u - u_o)^k \tag{4}$$

where the coefficients f_k are real.

Let $f(u)$ be a real function, analytic on I, and $\hat{f}(z)$ an analytic function on some region D which contains I, and $f(u) = \hat{f}(u)$ on I. Then $\hat{f}(z)$ said be the analytical continuation of $f(u)$ from I into the region D.

The series

$$\sum_{k=o}^{\infty} f_k (z - u_o)^k \tag{5}$$

we obtain from (4) with a complex value $z = u + iv$. It converges on the disk $|z - u_o| < \delta$ and its sum is $\hat{f}(z)$. Of course the sums (4) and (5) are identical on I. Finally, $f(u)$ can be analytical continued from I to some region D, which is symmetric with respect to the real axis: this fact follows from the Riemann-Schwartz principle of symmetry.

The function $h_{L,M}(z)$ has an integral representation:

$$\frac{1}{2\pi i} \oint_{\Gamma} \frac{f(t) Q_M(t)}{t^{L+M+1}(t - z)} \, dt \tag{6}$$

where Γ is a positively oriented contour in $\mathbb{C}$ which satisfies the following conditions:

(i) the origin and the point $t=z$ are inside Γ,

(ii) $f(u)$ is analytic on and within Γ.

Naturally we choose in the integral representation (6) the analytical continuation of the functions $f(u), Q_M(u)$ to the whole complex region G ($\partial G = \Gamma$). We shall lightly recognize, based on the context, the nature of the investigated functions. Instead of $\hat{f}, \hat{h}, \hat{Q}_M$ and $\hat{P}_L$ we shall write f, h, Q_M and P_L .

The existence of PAs was discussed for example in [1] .

A wide-sense stationary stochastic process has the spectral representation in the form:

$$X(t) = \int_{\mathbb{R}} e^{itu} \, dZ_X(u),$$

where $Z_X(u)$ is the so called spectral process of $X(t)$. The connection between the process and its spectral density $f(u)$ is given with the correlation function $K_X(t)$ and the so called Bochner-Khintchine's theorem:

$$EX(t)\overline{X(o)} = K_X(t) = \int_{\mathbb{R}} e^{itu} f(u) du \ .$$

From $K_X(t) = EX(t)\overline{X(o)} = E\,\overline{X(o)\overline{X(t)}} = \overline{K_X(-t)}$ and from

$$f(u) = \frac{1}{2} \int_{\mathbb{R}} e^{-itu} K_X(t) \, dt$$

it follows that a spectral density is nonnegative, selfconjugate and $L_1(\mathbb{R})$ - integrable. The quantity $K_X(o) = E|X(t)|^2$ is the variance of the process $X(t)$, we note $K_X(o) = DX(t)$.

A wide-sense stationary stochastic process is said to be band-limited if there exists a positive real number w, such that

$$K(t) = \int_{-W}^{W} e^{itu} f(u) \, du \ .$$

In other words, the spectral density vanishes outside of $I = [-w,w]$.

In our investigations we consider only the centered processes, i.e. $EX(t) = 0$.

3. PADÉ-APPROXIMATION OF SPECTRAL DENSITIES

Let $f(u)$ be a real function, analytic at the origin and

$$f(u) \begin{cases} >0 & u \in [-w,w] \\ =0 & \text{otherwise} \end{cases} . \tag{7}$$

Because $f(u)$ is a real function, the poles of $f(u)$ are complex when $f(u)$ is bounded or L_1-integrable. All functions which satisfy the condition (7), form the class $\mathbb{F}$. $\mathbb{F}$ is a subclass of the basic function class.

Theorem 1: Let $f \in \mathbb{F}$. If $P_L(u) > 0$ on $I = [-w,w]$, it follows that $(L/M)_f(u) \in L_1(\mathbb{R})$.

Proof: From relation (2) follows

$$Q_M(z)f(z) - P_L(z) = \frac{z^{L+M+1}}{2\pi i} \oint_{\Gamma} \frac{f(t)Q_M(t)}{t^{L+M+1}(t-z)}\, dt, \tag{8}$$

where Γ is a closed, positively oriented contour which contains I. We can now evaluate the quantity $|h_{L,M}(z)|$ through

$$|h_{L,M}(z)| \leq \max|f(z)| \left| \frac{1}{2\pi i} \oint_{\Gamma} \frac{Q_M(t)}{t^{L+M+1}(t-z)}\, dt \right|$$

$$\leq f^{+}\left(\sum_{k=1}^{L+M+1} \left| \frac{1}{2\pi i} \oint_{\Gamma} \frac{A_k}{t^k}\, dt \right| + \left| \frac{1}{2\pi i} \oint_{\Gamma} \frac{B}{t-z}\, dt \right| \right) = H,$$

where $f^+ = \max_G |f(z)|$, and

$$A_{L+M-k+1} = \begin{cases} -z^{-k-1} \sum_{j=0}^{k} q_j z^j & k = \overline{0,M-1} \\ -z^{-k-1} Q_M(z) & k = \overline{M,L+M} \end{cases} ,$$

$B = -A_1 = Q_M(z)/z^{L+M+1}$. From the Cauchy's integral formula it follows that

$$H = f^+(|A_1| + |B|) = 2f^+ |Q_M(z)| / |z|^{L+M+1} .$$

Finally

$$|h_{L,M}(z)| \leq 2f^+ |Q_M(z)| / |z|^{L+M+1} . \tag{9}$$

Now, from (2), (9) and $||a| - |b|| \leq |a - b|$ follows

$$\left| |Q_M(z)| |f(z)| - |P_L(z)| \right| \leq |Q_M(z)f(z) - P_L(z)| \leq 2f^+ |Q_M(z)|$$

for all $z \in G$. Further, we choose the restrictions of the investigated functions to the real axis. The last evaluation gives

$$|Q_M(u)| \geq |P_L(u)| / 3f^+ . \tag{10}$$

Let $P_L^- = \min_I |P_L(u)|$. From the positivity of P_L^- and (10) we have

$$0 < P_L^- / 3f^+ \leq \left| \frac{P_L(u)}{Q_M(u)} \right| = |(L/M)_f(u)| . \tag{11}$$

The upper bound of $(L/M)_f(u)$ there exists: (10) guarantees that

$(L/M)_f(u)$ has no poles on I.So $(L/M)_f(u)$ is a bounded,positive function on I.The positivity of $(L/M)_f(u)$ is a simple consequence of $(L/M)_f(o) = f(o) > o$.

The proof is complete. □

Consequence: Let f $\mathbb{F}$ .Then $(o/2m)_f(u)$ satisfies the inequality

$$(12) \qquad o < (Q_M^+)^{-1} \leq \left|(o/2m)_f(u)\right| \leq \frac{3f^+}{f(o)}$$

where Q_M^+ is the restriction to I of $\max_G |Q_M(z)|$.

Proof: Based on the maximum modulus principle for the closed region G is $|Q_M(z)| \leq \max_G |Q_M(z)| = Q_M^+ \cdot P_L^- = f(o)$ and from (1o) and (11) follows the statement of the consequence.□

Remark: From foregoing considerations it is clear that we observe the function $(L/M)_f(u)$ only on I.Exactly,we think that $(L/M)_f(u)$ vanishes outside of I.

A special type of Padé-approximants,the (o,2m) order PAs have a very interesting property: it can be written as

$$(13) \qquad (o/2m)_f(u) = f(o)/|A_m(u)|^2$$

for some complex coefficient polynomial $A_m(u)$ retaining the properties given in (12),if f(u) is an <u>even</u> function.

From (1) we have

$$(14) \qquad \sum_o^k q_j f_{k-j} = p_k \quad (k = \overline{o,L}); \quad \sum_o^k q_j f_{k-j} = o \quad (k = \overline{L+1,M+L}).$$

If f(u) is an even function the formal power series (3) contains only the eventh order elements: $f_{2k-1} = o$, $k \in \mathbb{N}$.

For the (o,2m) order PA $L = o, M = 2m$ and the system (14) reduces to

$$(15) \qquad q_o = 1 \ , \qquad \sum_{j=o}^{k} q_j f_{k-j} = o \qquad (k = \overline{1,2m})$$

and $q_1 = q_3 = \dots = q_{2m-1} = o$. The matrix form of (15) is

$$\begin{bmatrix} 1 & o & o & \cdot & \cdot & \cdot & o \\ f_2 & f(o) & o & \cdot & \cdot & \cdot & o \\ f_4 & f_2 & f(o) & \cdot & \cdot & \cdot & o \\ \cdot & \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & & \cdot & \cdot \\ f_{2m} & f_{2m-2} & f_{2m-4} & \cdot & \cdot & \cdot & f(o) \end{bmatrix} \cdot \begin{bmatrix} q_o \\ q_2 \\ q_4 \\ \cdot \\ \cdot \\ \cdot \\ q_{2m} \end{bmatrix} = \begin{bmatrix} 1 \\ o \\ o \\ \cdot \\ \cdot \\ \cdot \\ o \end{bmatrix}$$

The solution of the previous system is unique and nontrivial:

$$q_{2j} = \frac{(-1)^j}{f(o)^j} \begin{vmatrix} f_2 & f(o) & \cdot & \cdot & \cdot & o \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & f(o) \\ & & & & \cdot & \\ f_{2j} & f_{2j-2} & \cdot & \cdot & \cdot & f_2 \end{vmatrix}$$

where $j = \overline{o,m}$.

The connection between the coefficients of $Q_{2m}(u)$ and $A_m(u)$ must be from (13):

$$(16) \qquad q_k = \sum_{\substack{i+j=k \\ i,j \in \{o,\dots,m\}}} a_i \overline{a_j} \ ,$$

where $A_m(u) = \sum_{o}^{m} a_k u^k$, $a_k \in \mathbb{C}$. We can now solve (16) with

respect to a_k ,but this solution is not unique.For example:

(i) $|a_o|^2 = 1$ i.e. $a_o = e^{i\varphi_o}$.

(ii) Let us take $a_j = r_j e^{i\varphi_j}$, $j = \overline{1,m}$
$a_o\overline{a_1} + \overline{a_o}a_1 = 2\mathrm{Re}(\overline{a_o}a_1) = o$ give us $a_1 = a_o r_1 i(-1)^{k_1}$
for some integer k_1 and arbitrary $r_1 > o$.

(iii) $a_o\overline{a_2} + |a_1|^2 + \overline{a_o}a_2 = 2\mathrm{Re}(\overline{a_o}a_2) + r_1^2 = -f_2/f(o)$,
i.e. $r_2\cos(\varphi_2-\varphi_o) = -1/2(r_1^2 + f_2/f(o))$.The last equation has a solution but it is not unique etc.

In this way we prove the existence of the coefficients a_k. Hence,the relation (13) is valid,and $(o/2m)_f(u)$ is positive on I,therefore the function

$$(o/2m)_f(u) = \begin{cases} f(o)/|A_m(u)|^2 & u \in I = [-w,w] \\ o & \text{elsewhere} \end{cases}$$

is the rational spectral density of a band-limited process $(o/2m)_X(t)$ if the band-limited process X(t) has the spectral density f(u).

4. CONVERGENCE OF PA SEQUENCES

We consider a sequence $\{(o/2m)_f(u)\}$ of PAs of a band-limited density f(u).The uniform convergence of such sequences was investigated by many authors: De Montessus,Beardon, Pommerenke etc. in the following cases: $n/m \to \infty$; $n=am$, $a \in (o,1)$ etc..Now,we investigate the pointwise convergence on the possible largest interval on (-r,r).We prove first a result for

the sequence of $(L/M)_f(u)$ PAs.

Theorem 3: The truncation error for PA approximation of $f \in \mathbb{F}$ is

$$\left| f(u) - (L/M)_f(u) \right| = O\left(\left(\frac{|u|}{r} \right)^{L+M+1} \right) \text{ on } I \cap (-r,r).$$

Proof: From (2) follows

$$\left| f(z) - (L/M)_f(z) \right| \leq \frac{|z|^{L+M+1}}{2\pi |Q_M(z)|} \left| \oint_{\Gamma} \frac{f(t) Q_M(t)}{t^{L+M+1}(t - z)} \, dt \right| .$$

We choose a new integration contour $C_r = \left\{ re^{is} : 0 \leq s \leq 2\pi \right\}$, that contains the point z .It follows

$$\left| f(z) - (L/M)_f(z) \right| \leq \frac{|z|^{L+M+1}}{2\pi |Q_M(z)|} \int_0^{2\pi} \frac{\left| f(re^{is}) Q_M(re^{is}) \right|}{r^{L+M} \left| re^{is} - z \right|} \, ds \quad .$$

From the maximum modulus principle $|Q_M(re^{is})|$ has its maximum on the integration contour C_r , this value is $Q^+_{M,r}$ .Theorem 1 give us the estimates

$$\left| f(z) - (L/M)_f(z) \right| \leq \frac{|z|^{L+M+1}}{2\pi r^{L+M} P_L^-} 3 Q^+_{M,r} (f^+)^2 \int_0^{2\pi} \frac{ds}{\left| re^{is} - z \right|}$$

$$\leq \frac{|z|^{L+M+1} \, 3 (f^+)^2}{r^{L+M}(r - |z|) P_L^-} Q^+_{M,r} \quad .$$

Naturally,we retain the solution on the positivity of $P_L(u)$. Finally,we get the inequality:

$$\left| f(z) - (L/M)_f(z) \right| \leq 3 Q^+_{M,r} \frac{(f^+)^2}{P_L^-} (|z|/r)^{L+M+1} \left(1 - \frac{|z|}{r}\right)^{-1} .$$

ıoosing the restrictions of the functions in the last in-
ıuality to I $=[-w,w]\cap(-r,r)$ it is not hard to show that its
ɔal-valued variant is equivalent to the assertion of the
ıeorem.□

As a consequence of the previous theorem we can formu-
ıte the

Theorem 4: The sequence of rational spectral densities $\{(o/2m)_f(u)\}$ tends to an even band-limited density f(u) point-
ise on $I\cap(-r,r)$, when m tends to infinity.□

The elements of the sequence $\{(o/2m)_f(u)\}$ are spectral
.ensities.We can so approximate poitwise the even spectral
.ensity of a band-limited process with rational spectral den-
ities.This result has very interesting consequences.

3. PADÉ-PROCESSES

The PA of the spectral density of a band-limited process
.s a spectral density when a) f(u) is an even function, b)
;he PA is of the order (o,2m).The first condition is a sim-
)le consequence of the reality of X(t).The connected process
)f the density $(o/2m)_f(u)$ we note $(o/2m)_X(t)$ and it is the
so called Padé-process.What can be said about the mean square
convergence of the sequence $(o/2m)_X(t)$ to X(t) if m tends to
.nfinity?Before we give an answer to this question,we discuss
;he connection between w and r.

1. $w \leqq r$.The interval of the convergence of $(o/2m)_f(u)$ to f(u) is (-r,r).Viewing in the light of the m.s.

convergence this case is interesting : we cannot lose any information on the nature of $f(u)$ and $(o/2m)_f(u)$, moreover on the processes $X(t)$ and $(o/2m)_X(t)$ too.

2. $w > r$. Outside $(-r,r)$ we cannot consider the pointwise convergence of $(o/2m)_f(u)$, therefore the convergence in the mean of $(o/2m)_X(t)$ is senseless.

For example the class of the basic functions D_∞ (which are infinitely differentiable and vanish outside of a finite interval) of L. Schwartz satisfy the property 1.

Thus in the following considerations we suppose that $w \leq r$.

The linear transformation (or filter) of the process $X(t)$ is a transformation $A: X(t) \to Y(t)$ where:

$$(17) \qquad Y(t) = \int_{\mathbb{R}} e^{itu} h_Y(u)\, dZ_X(u) \quad .$$

$Z_X(u)$ is the spectral process of $X(t)$ (section 2.) and the $L_2(f_X(u)du)$-integrable function $h_Y(t)$ is the spectral characteristic function of the filter A. Some classical examples are: the differential operator $\mathbb{D}$ with the spectral characteristic $h_{\mathbb{D}}(u) = iu$, the integration operator $\mathbb{I}$ with $h_{\mathbb{I}}(u) = 1/iu$. Let now $\hat{h}_Y(s)$ be the inverse Fourier-transform of $h_Y(u)$. Another representation of $Y(t)$ is (equivalently to (17)):

$$(18) \qquad Y(t) = \int_{\mathbb{R}} \hat{h}_Y(s) X(t - s)\, ds \quad .$$

Consequently $Y(t)$ is the response of the process $X(t)$ on the input A.

A Rozanov theorem gives us the connection:

$$f_X(u)\left|h_Y(u)\right|^2 = f_Y(u) \quad ,$$

where $f_X(u)$ and $f_Y(u)$ are the spectral densities of X(t) and Y(t). In our case we consider the process $(o/2m)_X(t)$ as the response on the input $(o/2m)_{(\cdot)}$ to the band-limited process X(t). It has a spectral density like f(u) in formula (7). The characteristic function is

$$h_{(o/2m)}(u) = \frac{f_X(o)^{1/2}}{A_m(u)}\, f_X^{-1/2}(u) \tag{19}$$

from Rozanov's theorem, where $f_X^{1/2}$ is the positive root of the equation $(f_X^{1/2})^2 = f_X$. Naturally, $h_{(o/2m)}(u)$ is L_2-integrable on the measure $f_X(u)du$:

$$\int_{\mathbb{R}} \left|h_{(o/2m)}(u)\right|^2 f_X(u)\, du = \int_{\mathbb{R}} (o/2m)_f(u)\, du =$$

$$= E\left|(o/2m)_X(t)\right|^2 = D(o/2m)_X(t) < \infty \, ,$$

and the process $(o/2m)_X(t)$ has bounded second moment.

Theorem 5: $\left|h_{(o/2m)}(u)\right|^2 \longrightarrow 1$ pointwise on the interval (-r,r) when m tends to infinity.

Proof: the statement follows from Theorem 4 and (19). □

The cross-correlation function $K_{X,P}(t)$ of the process X(t) and $(o/2m)_X(t)$ was defined with

$$K_{X,P}(t) = EX(t)\overline{(o/2m)_X(o)} = \int_{\mathbb{R}} e^{itu} f_{X,P}(u)\, du$$

where $f_{X,P}(u)$ is the so called cross-spectral density. Another result by Rozanov states that $f_{X,P}(u) = f_X(u)\, h_{(o/2m)}(u)$, for $f_X(u) \in \mathbb{F}$. It is clear that

$$(2o) \qquad E|X(t) - (o/2m)_X(t)|^2 = \int_{\mathbb{R}} |1 - h_{(o/2m)}(u)|^2 f_X(u)du.$$

It is not hard to show that there exists a positive real number C' and a positive integer m_o for which is

$$|1 - h_{(o/2m)}(u)|^2 \leq C'(1 - |h_{(o/2m)}(u)|^2)^2$$

if $m > m_o$. Based on theorem 5 we give

Theorem 6: $E|X(t) - (o/2m)_X(t)|^2 \xrightarrow[m\to\infty]{} o$. □

Of course, we can state that the theorems 5 and 6 are valid on the whole of $\mathbb{R}$. Naturally, we choose only the positive-r densities from $\mathbb{F}$ for which $w \leq r$. The formal spectral densities have no practical importance.

REFERENCES

1. G.A.BAKER Jr. and P.R.GRAVES-MORRIS: Padé-Approximants, Addison-Wesley Publishing co., Reading, Massachusets, 1981.
2. D.ELLIOTT: Truncation error in Padé-approximants to certain functions: an alternative approach, Math.Comp.21(1967), 3o8-32
3. T.POGÁNY: Singuläre zufällige Prozesse und mittelquadratisch Konvergenz, Publ.Math.Debrecen 34(1987), 197-2o5.
4. YU.A.ROZANOV: Stationary Random Prozesses, Fizmatgiz, Moscow, 1963.
5. A.M.YAGLOM: An Intriduction to the Theory of Stationary Random Functions, Dover Publications, New York, 1973.

AN APPLICATION OF VARIATIONAL CALCULUS IN MECHANICS AND SOME PROPERTIES OF THE EIGENVALUES OF THE LAPLACIAN

THEMISTOCLES M. RASSIAS

\BSTRACT. In this survey paper we present:
I. The stability and oscillations or small motions of a soap film suspended between parallel coaxial rings. The solution to the problem relates the radius of the film r to the displacement z along the axis of symmetry by the equation of

$$r=a\cosh\frac{z-b}{a}.$$

The constants a and b are to be determined by requiring that r be equal to the fixed radii of the rings for z=0 and h, where h is the separation of the rings.
We study this equilibrium problem using eigenfunction methods and prove that the dynamical stability of the film is determined by the sign of the lowest eigenvalue λ_1 of an associated Sturm-Liouville problem, with the film stable for $\lambda_1>0$ and unstable for $\lambda_1<0$. This follows Durand [6].

II. Some of the most important properties of the eigenvalues of the Laplacian with some remarks on the smoothness of eigenfunctions and a generalization of Courant's nodal domain theorem (see [19],[2

I. An Application of Variational Calculus in Mechanics

In this section we consider the stability and oscillations or small motions of a soap film suspended between parallel coaxial rings, as this has been analyzed in Durand [6]. It is a standard problem used to introduce variational calculus in mechanics to determine the equilibrium shape of a soap film suspended between two parallel coaxial circular rings. The solution to the problem relates the radius of the film r to the displacement z along the axis of symmetry by the equation of the catenary

$$r=a\cosh\frac{z-b}{a}. \tag{1}$$

The constants a and b are to be determined by requiring that r be equal to the fixed radii of the rings for z=0 and h, where h is the separation of the rings. If the rings are of equal radius r_0, the surface is symmetrical about $z=\frac{h}{2}$, b is equal to $\frac{h}{2}$,and a,the minimum radius of the film, is to be found by solving the equation

$$(2) \qquad r_0 = a \cosh \frac{h}{2a} .$$

There are two solutions for $\frac{h}{2r_0} < 0.66274...$, *only one of which is stable, and no solutions at all for* $\frac{h}{2r_0} < 0.66274...$ In the second case, *the tubular configuration of the soap film is unstable*. From the *experimental point of view* this can be demostrated as follows (see [6]): We start with a stable tubular film with $\frac{h}{2r_0} < 0.66274$. and gradually increasing the separation between the rings until $\frac{h}{2r_0}$ approaches and then exceeds the critical value. For $\frac{h}{2r_0}$ greater than the critical value, the film collapses in the center and splits into two planar films, one on each ring. As $\frac{h}{2r_0}$ approaches the critical value, any perturbation results in a characteristic low-frequency oscillation of the film.

We shall give a mathematical analysis of this equilibrium problem (following [6]) using eigenfunction methods, and show that the *dynamical stability* of the film is determined by the sign of the lowest eigenvalue λ_1 of an associated *Sturm-Liouville problem* with the film *stable for* $\lambda_1 > 0$ *and unstable for* $\lambda_1 < 0$.

The energy of an ideal static soap film with surface area S and surface tension σ is given, neglecting gravity, by

$$(3) \qquad V[S] = 2\sigma S.$$

The possible equilibrium shapes of the film are determined by finding those surfaces for which V[S] has a local minimum. We require that the film be attached to two plane parallel coaxial rings with radii r_1 and r_2 separated by a distance h, and have no other boundaries. The equilibrium surfaces are axially symmetric, with a surface energy given by

$$(4) \qquad V[S] = 2\sigma \int_S dS = 4\pi\sigma \int_S r\sqrt{dr^2+dz^2}$$

If V[S] is to be an extremal for a surface S, there must be no first-order change in V[S] when S is varied slightly subject

to the fixed boundary conditions, i.e. $\delta V[S]=0$. We will specify the shape of the surface by giving its radius r as a function of z. Thus we get that $V[S]$ vanishes if r(z) satisfies the Euler equation

$$(5)\quad \frac{d}{dz}\left(\frac{rr_z}{\sqrt{1+r_z^2}}\right) - \sqrt{1+r_z^2} = 0,\quad r_z \equiv \frac{d}{dz}\, r(z),$$

or equivalently, if

$$(6)\quad \frac{1}{r_z}\,\frac{d}{dz}\left(\frac{r}{\sqrt{1+r_z^2}}\right) = 0.$$

Equation (6) is satisfied if either

$$(7)\quad \frac{r}{\sqrt{1+r_z^2}} = a,$$

where a is a positive constant, or r_z is infinite. Solving (7) we obtain the equation of a hollow tube,

$$(8)\quad r(z)=a\cosh\left(\frac{z-b}{a}\right),$$

where b is a constant of integration. In the second case, $z_r=\frac{1}{r_z}$ vanishes, thus z does not vary with r, and the surface S consists for our boundary conditions of two disconnected plane disks which fill the rings. Any variation of the disks about the plane configuration clearly increases their surface area. As a result $V[S]$ has at least a local minimum, and the double soap film is stable against small perturbations. From the topological point of view the double soap film is distinct from thehollow tube. The constants of integration a and b in (8) must be specified in such a way that $r(0)=r_1$ and $r(h)=r_2$. We will consider the case of rings of equal radius r_0. Similar methods can be applied for the asymmetrical case We obtain $b=\frac{h}{2}$,

$$(9)\quad r(z)=a\cosh\left(\frac{z}{a}-\frac{h}{2a}\right),$$

and the boundary value problem reduces to that of determining a from the equation

$$(10)\quad r_0=a\cosh\frac{h}{2a},\quad a>0,$$

which again can be rewriten as

$$(11)\qquad \frac{2r_0}{h}=\frac{2a}{h}\cosh\frac{h}{2a}=u_0^{-1}\cosh u_0,\ u_0=\frac{h}{2a}\ .$$

The function $u^{-1}\cosh u$ is positive, diverges for $u\to 0$ and $u\to\infty$ (as $a\to\infty,0$), and has a finite minimum value 1.5089...for utanhu=1, $u=u_c=1.1997...$It follows that there exist two solutions to the boundary value problem for $\frac{2r_0}{h}<1.509$, a single solution for $\frac{2r_0}{h}=1.509$, and no solutions at all for $\frac{2r}{h}<1.509$ ($\frac{h}{2r_0}>$ $>0.66274...$).

We can solve the boundary value problem (11) by iteration starting with $a=r_0$, and find that

$$(12)\quad a=\frac{r_0}{\cosh\frac{h}{2a}}\approx r_0\left(1-\frac{h^2}{8a^2}+...\right)$$

$$\approx r_0\left(1-\frac{h^2}{8r_0^2}+...\right),\ \frac{h}{2r_0}<<1.$$

The shape of the film is given in the same approximation

$$(13)\qquad r=r_0\left(1-\frac{1}{2r_0^2}\,z(h-z)+...\right),\ \frac{h}{2r_0}<<1,$$

and is cylindrical up to terms of order $\frac{h^2}{4r_0^2}$.

The area of the film is

$$(14)\qquad S=\pi a^2\left(\sinh\frac{h}{a}+\frac{h}{a}\right)=2\pi\left(r_0(\sqrt{r_0^2-a^2})+\frac{1}{2}ha\right).$$

For the nearly cylindrical film (12),

$$(15)\qquad S=2\pi r_0 h\left[1+0\left(\frac{h^2}{4r_0^2}\right)\right],$$

where the correction terms are negative. It follows that this co figuration is stable, therefore that S has at least a local min mum. The second limiting solution for closely spaced rings $\frac{2r_0}{h}>>1$ is that for which $u_0=\frac{h}{2a}$ is large, and a is small,a<<h<< In fact, if we rewrite (11) as

$$(16)\qquad \frac{r_0}{a}=\frac{2r_0}{h}\cosh^{-1}\frac{r_0}{a}=\frac{2r_0}{h}\left[\ln\frac{2r_0}{a}+0\left(\frac{a^2}{r_0^2}\right)\right],\ \frac{a}{r_0}\ll 1,$$

We get

$$(17)\qquad a \approx \frac{h}{2}\left\{\ln\left[\frac{4r_0}{h}\left(\ln\frac{4r_0}{h}(\ldots)\right)\right]\right\}^{-1} < \frac{h}{2}\,.$$

For very closely spaced rings the extremal surface consists of two nearly planar surfaces connected by a narrow neck with radius $a<<h<<r_0$.

The area of the surface is then

$$(18)\qquad 2\pi r_0^2\left[1+0\left(\frac{h^2}{4r_0^2}\right)\right],\quad \frac{a}{r_0}<<1.$$

The correction terms are positive. The *nearby* configuration of two separate plane disks has a smaller area $2\pi r_0^2$, and can be approached arbitrarily closely by letting $\frac{h}{2r_0}\to 0$ $(\frac{a}{r_0}\to 0)$.

We get that the narrow-necked surface is probably unstable, therefore that S probably has a local maximum for this configuration. Continuity arguments imply that the entire branch of the solution curve with $\frac{h}{2a} > 1.2$ *is unstable*, while that with $\frac{h}{2a} < 1.2$ *is stable*.

Stability of the soap film. Suppose $r(z)=f(z)$ describe an initial surface S_0 (not necessarily an extremal surface) and consider a perturbed surface described the equation

$$(19)\qquad r(z)=f(z)+g(z),$$

where $g(z)$ is an infinitesimal twice-differentiable function with $g(0)=g(h)=0$. Assume also that g_z is infinitesimal for $0\leq z\leq h$. Then $V[S]$ can be written as a power series in g, g_z as follows:

$$(20)\qquad V[S]=4\pi\sigma\int_0^h \sqrt{1+r_z^2}\; r\,dz$$

$$=4\pi\sigma\int_0^h \left(f\sqrt{1+f_z^2}+g\sqrt{1+f_z^2}+\frac{ff_zg_z}{\sqrt{1+f_z^2}}\right.$$

$$(21)\qquad \left.+\frac{gg_zf_z}{\sqrt{1+f_z^2}}+\frac{fg_z^2}{2(1+f_z)^{3/2}}+0(z^3)\right)dz$$

$$=V[S_0]+4\pi\sigma\int_0^h g\left(\sqrt{1+f_z^2}-\frac{d}{dz}\,\frac{ff_z}{\sqrt{1+f_z^2}}\right)dz\,.$$

$$+2\pi\sigma \int_0^h (fg_z^2 - f_{zz}g^2) \frac{dz}{(1+f_z^2)^{3/2}} + O(z^3)$$

(23) $$= V[S_0] + \delta V[S_0] + \delta^2 V[S_0] + \ldots$$

For S_0 an extremal surface, f(z) satisfies the Euler equation (5) and $\delta V[S_0]=0$.

Let us consider now the case of symmetrical rings.

We obtain

(24) $$f = a \cosh\left(\frac{z}{a} - \frac{h}{2a}\right) = a\cosh u, \quad u = \frac{z}{a} - \frac{h}{2a}$$

and

(25) $$\delta^2 V[S_0] = 2\pi\sigma a \int_0^h \left(g_z^2 - \frac{1}{a^2} g^2\right) \cosh^{-2}\left(\frac{z}{a} - \frac{h}{2a}\right) dz$$

$$= 2\pi\sigma \int_{-u_0}^{u_0} (g_u^2 - g^2) \frac{du}{\cosh^2 u}, \quad u_0 = \frac{h}{2a}.$$

It is known from the work of Legendre, Jacobi, and Weierstrass that an extremal curve will give a minimum of $V[S]$ if (i) the second derivative of the integrand in (20) with respect to r_z is positive for all z and r in a neighborhood of the curve and all finite r_z; and (ii) there is no point *conjugate to z=0* on the interval $0 \leq z \leq h$. It is easy to see that both these conditions are satisfied in our case.

Another way to be used in order to verify the conditions for minimum is to convert the weak stability problem into one of the determining *the sign of the lowest eigenvalue of an appropriate Sturm-Liouville operator*. The weak form of condition (i) will enter when we define the Sturm-Liouville operator. The conjugate points are just the nodes of the lowest eigenfunction of this problem, and the condition (ii) is replaced by the requirement that the lowest eigenvalue be positive. We will change at this point from the radial displacements g(z) used above to equivalent infinitesimal displacements $\xi(z)$ perpedicular to the initial surface of the film. The ξ's are the natural coordinates for the study of the oscillations. The vector displacement of a point $\bar{r}=(r,z)$ associated with a perpedicular displacement $\bar{\xi}(z)$ is

$$\bar{r}'-\bar{r}=\bar{\xi}(z)=\hat{n}(z)\xi(z), \tag{27}$$

where $\hat{n}(z)$ is the normal to the surface at $(r,z)=(f(z),z)$,

$$\hat{n}=(\hat{r}-f_z\,\hat{z})\sqrt{1+f_z^2} \tag{28}$$

Therefore

$$r'(z')=r+\xi_r\ f(z)\ +\frac{\xi(z)}{\sqrt{1+f_z^2}} \tag{29}$$

$$z'=z+\xi_z=z-\ \frac{f_z}{\sqrt{1+f_z^2}}\ \ \xi(z) \tag{30}$$

If we substitute for z' in r'(z') and expand, we obtain

$$r'(z)=f+\xi\ \sqrt{1+f_z^2}\ +0(\xi^2) \tag{31}$$

$$=f+\xi\cosh\ u\ +0(\xi^2), \tag{32}$$

thus g(z), the first order change in r at fixed z, is given by

$$g=\xi\cosh\ u. \tag{33}$$

After some standard computation we derive

$$\delta^2V[S_0]=2\pi\sigma\int_{-u_0}^{u_0}(\xi_u^2-\ \frac{2}{\cosh^2u}\ \ \xi^2)\ du \tag{34}$$

$$=2\pi\sigma\int_{-u_0}^{u_0}\xi(-\xi_{uu}-\ \frac{2}{\cosh^2u}\ \xi)\ du. \tag{35}$$

The problem now of whether or not the extremal configuration of a soap film specified by a given value of $u_0=\frac{h}{2a}$ is stable can be restated at this point in terms of the operator

$$L=-\ \frac{d^2}{du^2}\ -\ \frac{2}{\cosh^2u}\ . \tag{36}$$

We have

$$\delta^2V[S_0]=2\pi\sigma\ (\xi,L\xi), \tag{37}$$

where the inner product is defined by the integral in (35). If L is a positive operator, that is, if $(\xi,L\xi)$ is positive for any ξ, then δ^2V is positive for any variation of the extremal configuration, and the soap film is stable. Now L is a positive operator if and only if its lowest eigenvalue is positive. Consider the

Sturm-Liouville eigenvalue problem defined by the differential equation

$$(38) \qquad L\Psi_n=\lambda_n w(u)\Psi_n$$

with the boundary conditions $\Psi_n(u_0)=\Psi_n(-u_0)=0$. The weight functi w(u) must be strictly positive, but arbitrary.

Set $w(u)=\cosh^2 u$. The equation (38) becomes

$$(39) \qquad \frac{d^2\Psi_n(u)}{du^2}+(\lambda_n\cosh^2 u+\frac{2}{\cosh^2 u})\Psi_n(u)=0.$$

The eigenfunctions Ψ_n can be chosen to be real, and we will assume also that they have been normalized. The orthogonality relation for the Ψ's is then

$$(40) \qquad \int_{-u_0}^{u_0}\Psi_n(u)\Psi_m(u)\cosh^2 u\,du=\delta_{nm}.$$

The eigenvalues λ_n, n=1,2,... are real and discrete, and will be assumed to be

$$\lambda_1<\lambda_2<\lambda_3<\cdots$$

We expand $\xi(u)$ in (25) as a series in the complete set o eigenfunctions Ψ_n,

$$(41) \qquad \xi(u)=\sum_{n=1}^{\infty}c_n\Psi_n(u),\ \{c_n\}\ \text{real},$$

and we find that

$$(42) \qquad \delta^2V[S_0]=2\pi\sigma\sum_{n=1}^{\infty}\lambda_n c_n^2 .$$

We now see that a given extremal configuration of the soa film is stable (resp. unstable) if the lowest eigenvalue λ_1 is positive (resp. negative). If λ_1 is positive, all the eigenvalues are positive, and δ^2V is positive for any choice of the c_n. Thus the area of the film increases for any variation $\xi(u)$ which satisfies the boundary conditions. This is the condition for stability. If λ_1 is negative, the choice $c_1\neq 0$, $c_n=0$ for n>1 gives a va riation ξ which decreases the area of the film, $\delta^2V<0$, and the configuration is unstable. If $\lambda_1=0$, the configuration is in neutr equilibrium since an infinitesimal displacement

$$\xi(u)=c\ \Psi_1(u)$$

gives

$$\delta V=\delta^2 V=0.$$

It can be shown that λ_1 is exactly zero for the critical value of $u_0, u_0=u_c=1.20$, $\frac{h}{2r_0}=0.663$. It can also be shown that the soap film is stable for $u_0<u_c$ and unstable for $u_0>u_c$.

II. Eigenvalues of the Laplacian

II.1 Let $D\subset\mathbb{R}^m, m\geq 2$ be a domain with a smooth boundary. We consider solutions of

$$(1)\qquad \begin{cases} \Delta u+\lambda u=0 & \text{in } D \\ u=0 & \text{on } \partial D \end{cases}$$

where D is a region (and ∂D is the boundary of D) such that the spectrum is discrete. For $m=2, \Delta u+\lambda u=0$ in D is also known as the *Helmholtz equation* [10]. Someone reduces to it from separating the time variable out of the wave equation. The eigenvalue problem (1) for m=2 may represent the vibration of a *fixed membrame*, with the eigenvalue $\lambda=k^2$, where k is proportional to a *principal frequency of vibration*, and the eigenfunction u represents the shape of *a mode of vibration*. These are also the frequencies and modes of the *simply supported plate* of the same plate (see [12]).

Suppose that the spectrum i.e., those values of λ for which a non-trivial solution exists, is discrete. We order the eigenvalues

$$0<\lambda_1<\lambda_2\leq\lambda_3\leq\ldots\leq\lambda_n\leq\ldots$$

and we normalize the corresponding eigenfunctions $u_1, u_2, \ldots, u_n, \ldots$ such that

$$(2)\qquad \int_D u_i u_j=\delta_{ij},\quad i,j=1,2,\ldots$$

Theorem 1 ([9]). *Let* σ *be the unique solution on* (λ_n,∞) *of the equation*

$$(3)\qquad \sum_{i=1}^{n}\frac{\lambda_i}{\sigma-\lambda_i}=\frac{mn}{4}.$$

Then

(4) $\lambda_{n+1} \leq \sigma$.

To prove Theorem 1, Hile and Protter have first established the following proposition.

Proposition 2. I. *For each integer* l *with* $1 \leq l \leq n$, *the following inequality holds:*

$$(5) \qquad \lambda_{n+1} \leq \frac{1}{2}(\lambda_n+\lambda_1) + \frac{2}{mn} \sum_{i=1}^{n} \lambda_i + \frac{2}{mn}\left\{\left[\frac{mn}{4}(\lambda_n-\lambda_1) + \sum_{i=1}^{n}\lambda_i\right]^2 - mn(\lambda_n-\lambda_1) \sum_{i=1}^{l}\lambda_i\right\}^{1/2}$$

II. *The first* $(n+1)$ *eigenvalues of the Laplacian satisfy the inequality*

$$(6) \qquad \sum_{i=1}^{n} \frac{\lambda_i}{\lambda_{n+1}-\lambda_i} \geq \frac{1}{4} mn$$

Remark. Inequality (6) is of interest only when λ_{n+1} is strictly greater than λ_n.

Proof of Theorem 1 ([9]). Consider the n trial functions

$$(7) \qquad \phi_i = x_1 u_i - \sum_{j=1}^{n} a_{ij} u_j, \quad i=1,2,\ldots,n,$$

such that

$$(8) \qquad a_{ij} = \int_D x_1 u_i u_j, \quad i,j=1,2,\ldots,n.$$

It follows that each ϕ_i is orthogonal to $u_1, u_2, \ldots, u_n$ and because of the fact $\phi_i = 0$ on ∂D, we obtain

$$(9) \qquad \lambda_{n+1} \leq \frac{-\int \phi_i \Delta \phi_i}{\int \phi_i^2} , \quad i=1,2,\ldots,n$$

It follows easily that

$$(10) \qquad \lambda_{n+1} \int \phi_i^2 \leq \lambda_i \int \phi_i^2 - 2 \int u_{i,x_1} \phi_i, \quad i=1,2,\ldots,n$$

Thus

$$(11) \qquad \lambda_{n+1} \sum_{i=1}^{n} \int \phi_i^2 \leq \sum_{i=1}^{n} \lambda_i \int \phi_i^2 - 2 \sum_{i=1}^{n} \int u_{i,x_1} \phi_i .$$

Because of the fact a_{ij} are symmetric,

$$-2 \sum_{i=1}^{n} \int u_{i,x_1} \phi_i = -2 \sum_{i=1}^{n} \int x_1 u_i u_{i,x_1} + 2 \sum_{i,j=1}^{n}$$

$$a_{ij}\int u_j u_{i,x_1} = \sum_{i=1}^{n}\int u_i^2 = n.$$

Therefore (11) becomes

$$(12)\qquad \lambda_{n+1}\sum_{i=1}^{n}\int\phi_i^2 \leq \sum_{i=1}^{n}\lambda_i\int\phi_i^2 - 2(1+\beta)\sum_{i=1}^{n}\int u_{i,x_1}\phi_i - n\beta$$

where the real parameter β will be chosen later

Let $\tau_1,\tau_2,\ldots,\tau_n$ be positive constants and apply Cauchy's inequality, then (12) reduces to

$$(13)\ \lambda_{n+1}\sum_{i=1}^{n}\int\phi_i^2 \leq \sum_{i=1}^{n}(\lambda_i+\tau_i)\int\phi_i^2 + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int u_{i,x_1}^2 - n\beta\ .$$

Set $\tau_n=\tau$ and choose the τ_i, $i=1,2,\ldots,n-1$ so that

$$\tau_i=\tau+\lambda_n-\lambda_i,\quad i=1,2,\ldots,n-1$$

Define

$$S_1=\sum_{i=1}^{n}\int\phi_i^2$$

Then (13) can be written as follows

$$(14)\ \lambda_{n+1}\ S_1 \leq (\lambda_n+\tau)S_1 + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int u_{i,x_1}^2 - n\beta\ .$$

In place of the trial functions (7) we may choose the functions

$$(15)\ \phi_{ik}=x_k u_i - \sum_{j=1}^{n} a_{ijk}u_j, i=1,2,\ldots,n: k=1,2,\ldots,m$$

Peforming an analysis as above for $k=2,3,\ldots,m$ and,denotin

$$S_k=\sum_{i=1}^{n}\int\phi_{ik}^2\ ,\quad k=1,2,\ldots,m,$$

we obtain the m inequalities

$$(16)\ \lambda_{n+1}S_k \leq (\lambda_n+\tau)S_k + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int u_{i,x_k}^2 - n\beta, k=1,2,\ldots,m$$

Setting

$$S=\sum_{k=1}^{m}S_k,\ \text{we obtain}$$

$$\lambda_{n+1}S \leq (\lambda_n+\tau)S + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int|\nabla u_i|^2 - mn\beta,$$

or

$$(17)\quad \lambda_{n+1}S \leq (\lambda_n+\tau)S+(1+\beta)^2 \sum_{i=1}^{n} \lambda_i\tau_i^{-1}-mn\beta$$

The selection of τ such that

$$(18)\quad \sum_{i=1}^{n} \lambda_i\tau_i^{-1} \leq (1+\beta)^{-2}\, mn\beta$$

implies an inequality for the τ_i as a function of β. Therefore the condition on τ becomes

$$\sum_{i=1}^{n} \lambda_i(\tau+\lambda_n-\lambda_i)^{-1} \leq \frac{mn}{4}$$

We note that

$$f(\tau)=\sum_{i=1}^{n} \frac{\lambda_i}{\tau+\lambda_n-\lambda_i}$$

is a decreasing function of τ on (o,∞) and $\lim_{\tau\to 0^+} f(\tau)=+\infty, \lim_{\tau\to\infty} f(\tau)=0$.

Thus setting $\sigma=\tau+\lambda_n$ we observe that there is a unique solution of (3) on (λ_n,∞) and (17),(18) imply (4).

The equation (3) can be written also in the form

$$(19)\quad \prod_{i=1}^{n}(\sigma-\lambda_i)-\frac{4}{mn}\sum_{i=1}^{n}\lambda_i \prod_{\substack{j=1\\ j\neq i}}^{n}(\sigma-\lambda_i)=0.$$

We denote

$$P(\sigma)\equiv \prod_{i=1}^{n}(\sigma-\lambda_i)=\sigma^n+\sum_{k=1}^{n}(-1)^k a_k\sigma^{n-k},$$

where a_i is the i-th elementary symmetric function of $\lambda_1,\lambda_2,\ldots,$ and also denote

$$R(\sigma)\equiv \sum_{i=1}^{n}\lambda_i \prod_{\substack{j=1\\ j\neq i}}^{n}(\sigma-\lambda_j)=\sum_{k=1}^{n}(-1)^{k+1}\, k a_k\sigma^{n-k}$$

Then (19) reduces to the following form

$$(20)\quad \sigma^n+\sum_{k=1}^{n}(-1)^k a_k\left(1+\frac{4k}{mn}\right)\sigma^{n-k}.$$

Hence (4) is given by the unique root of (20) on the interval (λ_n,∞).

Q.E.D.

Remark. The above result of Hile-Protter generalizes the one given by Payne, Polya and Weinberger [14], which states that:

For domains in $\mathbb{R}^2$, *the inequality*

$$\lambda_{n+1} \leq \lambda_n + \frac{2}{n} \sum_{i=1}^{n} \lambda_i, \quad n=1,2,\ldots$$

holds if the spectrum is discrete.

In the following we outline a new method of Hile-Protter [9] which can be used to improve the upper bound estimates for λ_2. For this, let u denote the first normalized eigenfunction for

$$\begin{cases} \Delta u + \lambda_1 u = 0 & \text{in } D, \\ u = 0 & \text{on } \partial D, \end{cases}$$

and let f be any C^1 function in $D \cup \partial D$,

Theorem 3 [9]. *Let*

$$(21) \qquad c = \frac{1}{\lambda_1}\left(\frac{1}{2\lambda_1}\right)^{n-1} \frac{1}{(2n+1)^2}$$

and suppose $n \geq \lambda_1$. *Then for any domain* D *in* $\mathbb{R}^2$ *contained in the unit disk,*

$$(22) \qquad \lambda_2 \leq k\lambda_1,$$

with

$$k = \frac{(5-2c)+\sqrt{(5-2c)^2+8}}{4}$$

The proof of Theorem 3 has been based upon the following series of Lemmas [9]

Lemma 1. *Suppose*

$$\int_D f u^2 = 0$$

Then

$$(23) \qquad \lambda_2 \leq \lambda_1 + \frac{\int u^2 |\nabla f|^2}{\int f^2 u^2}.$$

Let

$$(24) \qquad A(\alpha) = \frac{\int u^{2\alpha}}{(\int u^{\alpha+1})^2}.$$

Lemma 2. *The following inequality holds:*

$$(25)\quad \frac{1}{\int x^2u^2}+\frac{1}{\int y^2u^2}\leq\frac{(\alpha+1)^2}{2\alpha-1}\lambda_1 A(\alpha),\ a\geq 1$$

Lemma 3. *Define* $\nu=\frac{\lambda_2}{\lambda_1}$; *then the inequality*

$$(26)\quad A(\alpha)\leq\frac{(2\alpha-1)(\nu-1)}{(2\alpha-1)\nu-\alpha^2}$$

holds for $1\leq\alpha<\nu+\sqrt{\nu^2-\nu}$

Lemma 4. *The following inequalities hold:*

$$(27)\quad \frac{1}{\int x^2u^2}+\frac{1}{\int y^2u^2}\leq\lambda_1\frac{3\nu+1}{\nu}$$

and, provided the axes are rotated properly,

$$\frac{1}{\int x^2u^2}\leq\frac{1}{2}\lambda_1\cdot\frac{3\nu+1}{\nu}$$

Lemma 5. *For* $\alpha\geq 1$ *define*

$$B_\alpha=\int x^\alpha u^2$$

and choose coordinate axes so that $B_1=\int xu^2=0$.

Let

$$J\equiv\frac{\left[(2n+1)B_{2n}-\lambda_1(\nu-1)B_{2n+2}\right]^2}{(2n+1)^2B_2B_{4n}}$$

with n *a positive integer. Then*

$$(28)\quad \lambda_2\leq\lambda_1+\frac{1}{B_2}-J$$

Lemma 6. *For* $n\geq\lambda_1$, *the following inequality holds:*

$$(29)\quad J\geq\left(\frac{1}{2\lambda_1}\right)^{n-1}\cdot\frac{1}{(2n+1)^2}$$

Remark. Applying Theorem 3, Hile and Protter were able to derive the following inequality of J.J.A.M. Brands [4]

$$(30)\quad \lambda_2\leq\frac{5+\sqrt{33}}{4}\lambda_1,$$

which is essentially the inequality (22) for c=0. The inequality of Brands dor $\mathbb{R}^m$ becomes

$$(31)\quad \frac{\lambda_2}{\lambda_1}\leq\frac{m+3+\sqrt{m^2+10m+9}}{2m}.$$

II.2. Smoothness of eigenfunctions. The eigenfunctions are chara-

cterized with the *unique continuation property*, that is, a function cannot satisfy $\Delta u+\lambda u=0$ in D and vanish on an open subset of D without vanishing identically in D. Each eigenfunction u_n is infinitely differentiable (i.e. $u_n \in C^\infty$) at the interior points of D(cf. [3]). At a straight line segment of the boundary, u_n can be reflected as an odd function accross the boundary. The resulting function satisfies $\Delta u+\lambda u=0$ in D in a whole neighborhood of that portion of the boundary and thus is C^∞ accross the boundary on straight line segments.

II.3. Nodal lines. The set of points in D where $u_n=0$ is the *nodal set* of u_n. Applying the unique continuation property, the nodal set consists of *curves* that are C^∞ in the interior of D. It is a very interesting property to be noted that where nodal lines cross, they form equal angles (cf [5]).

Courant's nodal line theorem [5] states that the nodal lines of the nth eigenfunction divide D into no more than (n-1) subregions which are called *nodal domains*. We note that u_1 has no interior nodes and thus λ_1 is an eigenvalue of multiplicity one. In the special case where D is a convex region, then u_1 has convex level curves (a fact which is not hard to be seen geometrically). Pleijel [16] has given an elegant proof of the nodal line theorem by applying the minimax property and unique continuation. It is an interesting fact to be noted that equality cannot hold for more than a finite number of n. This follows from the *Faber-Krahn inequality* ([7], [11]) for each nodal domain and *Weyl's law*, which is the asymptotic relation for the nth eigenvalue.

$$(32) \qquad \lambda_n \sim \frac{4\pi n}{A} \quad \text{as } n\to\infty$$

where A is the area of D.

It is a standard fact that the nth eigenvalue λ_n of D is the first eigenvalue for each of its nodal domains and a higher eigenvalue for a union of nodal domains.

A generalization of Courant's nodal domain Theorem.

In the following we outline J.Peetre's approach [15] for an extension of A.Pleijel's nodal domain theorem [16] to Riemannian manifolds.

Assume M is a 2-dimensional Riemannian manifold. The *Beltrami-Laplace operator* in M is

$$(33) \qquad \Delta = -g^{-\frac{1}{2}} \frac{\partial}{\partial x^j} \left(g^{\frac{1}{2}} g^{jk} \frac{\partial}{\partial x^k}\right),$$

where g_{kj} and g^{jk} are the covariant and contravariant components of the metric tensor in a local coordinate system and $g=\det g_{jk}$.

Assume now that D is a relatively compact connected domain in M. Consider the eigenvalue problem.

$$(34) \qquad \begin{aligned} \Delta u - \lambda u &= 0 \quad \text{in } D \\ u &= 0 \quad \text{on } \partial D \text{(boundary of D)} \end{aligned}$$

Our program is to compute the number of nodal domains N of the n-th eigenfunction of (34). We suppose that M is homeomorphic to a disk in the Euclidean plane.

Theorem 4.([15]). *Let* D_0 *be the least simply connected domain containing* D. *Suppose that*

$$(35) \qquad V_0 \sup_{D_0} K^+ \leq \pi,$$

where K *is the Caussian curvature,*

$K^+=\max(K,0)$, *and* V_0 *is the area of* D_0.

Then

$$(36) \qquad S^2 \geq 4\pi V \left(1 - \frac{1}{2\pi} \int_D K^+ dV\right),$$

where S *is the length of* ∂D *and* V *the area of* D. *Equality holds if and only if* $K=0$ *and* Ω *is a circle*

Proof ([15]). If D is simply connected ($D=D_0$) then (36) is a theorem of A.Huber (1954). If D is multiply connected then applying Huber's theorem to D_0 we obtain

$$(37) \qquad S_0^2 \geq 4\pi V_0 \left(1 - \frac{1}{2\pi} \int_{D_0} K^+ dV\right),$$

where S_0 measures the length of ∂D_0.

Suppose now that Σ is the interior of D_0-D and set $U=V_0-V$. Then we get

$$V_0 \int_{D_0} K^+ dV = V \int_D K^+ dV + V \int_\Sigma K^+ dV + U \int_{D_0} K^+ dV$$

$$\leq V \int_D K^+ dV + 2UV_0 \sup_{D_0} K^+$$

Therefore

$$(38)\qquad V(2\pi - \int_D K^+ dV) \leqq V_0(2\pi - \int_{D_0} K^+ dV).$$

Now (36) follows as a result of (37),(38) and $S \geqq S_0$. If equality holds in (36), then D must be simply connected and the last assertion of the theorem follows from Huber's theorem.

Q.E.D.

Theorem 5. ([15]). *Let* (36) *be satisfied and let* λ_1 *be the first eigenvalue of* (34).
Then

$$(39)\qquad \lambda_1 V \geqq \pi j^2 (1 - \frac{1}{2\pi} \int_D K^+ dV),$$

where j *is the first positive zero of the Bessel function* J_0. *Equality holds if and only if* K=0 *and* D *is a circle.*

Proof ([15]). Following the method of Faber [7] and Krahn [11] we can write:Let $u=u_1$ be the first eigenfunction. Set

$$(40)\qquad \begin{cases} D(\rho)=\{x \mid u(x)>\rho\}, \quad 0<\rho<\max u \\ \Delta(\rho)= \int_{D(\rho)} |\text{grad } u|^2 dV, \\ V(\rho)= \int_{D(\rho)} dV, \\ S(\rho)= \int_{\partial D(\rho)} dS, \\ H(\rho)= \int_{D(\rho)} u^2 dV. \end{cases}$$

Then

$$|\Delta'(\rho)| = -\Delta'(\rho) = \int_{\partial D(\rho)} |\text{grad } u| \, dS$$

and

$$|V'(\rho)| = -V'(\rho) = \int_{\partial D(\rho)} |\text{grad } u|^{-1} \, dS.$$

From Schwarz's inequality

$$(S(\rho))^2 \leqq |\Delta'(\rho)| \, |V'(\rho)|,$$

and from Theorem 4

$$(41)\qquad 4\pi(1 - \frac{1}{2\pi} \int_D K^+ dV) \; \frac{V(\rho)}{|V'(\rho)|} \leqq |\Delta'(\rho)|$$

If we apply a symmetrization process we get

$$(42) \qquad 4\pi\left(1-\frac{1}{2\pi}\int_D K^+ dV\right)\frac{\tilde{V}(\rho)}{|\tilde{V}'(\rho)|} \leq |\Delta'(\rho)|$$

Here we have replaced the domains $D(\rho)$ by concentric circles $\tilde{D}(\rho)$ with the same areas in the Euclidean plane, and the function u by a function $\tilde{u}$ which equals ρ on $\partial D(\rho)$.

It is true that $\tilde{V}(\rho)=V(\rho)$ and $\tilde{V}'(\rho)=V'(\rho)$. We also get

$$4\pi\frac{\tilde{V}(\rho)}{|\tilde{V}'(\rho)|} = |\tilde{\Delta}'(\rho)| \ ;$$

for $(\tilde{S}(\rho))^2=|\tilde{\Delta}'(\rho)||\tilde{V}'(\rho)|$ and $4\pi\tilde{V}(\rho)=|\tilde{S}(\rho)|^2$.

Therefore

$$\left(1-\frac{1}{2\pi}\int K^+ dV\right)|\tilde{\Delta}'(\rho)| \leq |\Delta'(\rho)|$$

Integrating over the interval $0<\rho<\max u$ we obtain

$$\left(1-\frac{1}{2\pi}\int_D K^+ dV\right)\tilde{\Delta} \leq \Delta$$

Also $\tilde{H}(\rho)=H(\rho)$ and $\tilde{H}=H$. It follows from Rayleigh's inequality that

$$\lambda_1 = \frac{D}{H} \ , \ \tilde{\lambda}_1 \leq \frac{\tilde{D}}{\tilde{H}};$$

and therefore (39) follows. If equality holds in (39), then the last assertion of the theorem follows from Theorem 4.

Q.E.D.

Theorem 6 ([15]). *There is a number $\alpha<1$ such that*

$$(43) \qquad \lim_{n\to\infty} \sup \frac{N}{n} \leq \alpha$$

Proof ([15]). Suppose now λ_n is the n-th eigenvalue and u_n is the n-th eigenfunction. Suppose also that $D_1, D_2, \ldots, D_N$ are the nodal domains of u_n. For each $D_l (l=1,2,\ldots,N)$ the value λ_n is the lowest eigenvalue. If we apply Theorem 5 to each D_l, we get

$$(44) \qquad \lambda_n V_l \geq \pi j^2 \left(1-\frac{1}{2\pi}\int_{D_l} K^+ dV\right)$$

If we take the sum of all inequalities (43) for $l=1,2,\ldots,N$ we obtain

$$\lambda_n V \geq \pi j^2 \left(N-\frac{1}{2\pi}\int_D K^+ dV\right)$$

But $\lim_{n\to\infty} n^{-1}\lambda_n V=4\pi$, therefore

$$\lim_{n\to\infty} \sup \frac{N}{n} \leq \left(\frac{2}{j}\right)^2<1. \tag{45}$$

Q.E.D.

Remark. It is easy to see that (43) remains true if (34) is replaced by an eigenvalue problem of the form

$$\Delta u+a(x)u=\lambda u,$$

where a(x) is a smooth bounded function.

Applying a similar argument, as in Theorem 5, we get [15]

$$(\lambda_1-\inf a(x))\, V \geq \pi j^2\left(1-\frac{1}{2\pi}\int_D K^+ dV\right), \tag{46}$$

and therefore (45) still follows.
It is now not difficult to extend the previous results to the case of a k-dimensional Riemannian manifold of constant curvature.

II.4. An orthogonal projection theorem for mappings and the Rayleigh quotient.

The Rayleigh quotient for the Jacobi operator $L[f]$ is defined by

$$R[f]=\frac{<L[f],f>}{<f,f>}$$

where L is defined in a Hilbert space H, with discrete point spectrum tending to infinity.

The eigenvalue λ_k of L, according to the classical principles of the Calculus of Variations (see for example [5]) of R.Courant and E.Fischer, can be written in te following form.

$$\lambda_k=\max_{W}\min_{f\perp W} R[f]=\min_{V}\max_{f\in V} R[f],$$

where W is any (k-1)-dimensional linear subspace of H and V is any k-dimensional linear subspace of H.Consider Σ_k to be a set (not a linear subspace), such that given any (k-1)-dimensional subspace W of H, there is a non-zero element of Σ_k that is orthogonal to W. The symbol k in this context is in order to know that Σ_k corresponds to the eigenvalue λ_k.

For any chosen $g\in\Sigma_k$ such that $g\perp W$, it follows that

$\min_{f \perp W} R[f] \leq R[g]$ and $R[g] \leq \max_{f \in \Sigma_k} R[f]$. Then $\min_{f \perp W} R[f] \leq \max_{f \in \Sigma_k} R[f]$

Then

$$\lambda_k = \max_{W} \min_{f \perp W} R[f] \leq \sup_{f \in \Sigma_k} R[f].$$

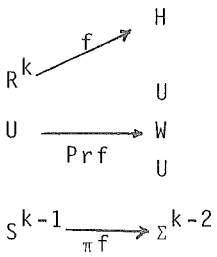
Therefore

$$\lambda_k \leq \sup_{f \in \Sigma_k} R[f]$$

This upper bound for the k^{th} eigenvalue λ_k, namely $\sup_{f \in \Sigma_k} R[f]$ may be a finite or infinite number and someone must be careful for a suitable choise of Σ_k in order for this upper bound to be a finite real number, and even more an accurate approximation of λ_k.

Proposition ([20]) *Let* H *be a Hilbert space and* $f:R^K \to H$ *a continuous mapping, homogeneous of odd degree (i.e.* $f(\lambda x)=\lambda^m f(x)$ *for some odd positive integer m) and satisfying* $f(x) \neq 0$ *for* $x \neq 0$. *Let* W *be a* (k-1)*-dimensional subspace of* H. *Then a vector* $x \neq 0$ *exists such that* $f(x) \perp W$

Proof. Assume that this is not the case and thus the mapping $f:R^K \to H$ has the property that for any W, a(k-1)-dimensional subspace of H, there is no vector $x \neq 0$ such that $f(x) \perp W$. Consider the orthogonal projection mapping Prf, $Prf:R^K \to W$, of $f:R^K \to H$, onto W.



Then $Prf(x) \neq 0$ for any $x \neq 0$, $x \in R^k$, and the mapping

$$\pi f = \frac{Prf}{||Prf||} : S^{K-1} \to \Sigma^{K-2}$$

is well defined, where

$$S^{K-1} = \{x \in R^K : ||x|| = 1\} \quad \text{and}$$

$$\Sigma^{K-2} = \{w \in W : ||w|| = 1\}.$$

Because of the fact $f:R^K \to H$ is a continuous mapping, homegeneous of

odd degree, i.e., $f(\lambda x)=\lambda^m f(x)$ for some odd positive integer $m, \lambda \in R$ and $x \in R^K$, it follows for $\lambda=-1$ that $f(-x)=-f(x)$ for $x \in R^K$. Therefore f is an odd mapping. However by the Borsuk-Ulam antipodal point theorem (see for example [21,p.266]) there is no such a mapping f, and thus we have proved that there exists a vector $x \neq 0$ such that $f(x) \perp W$.

Q.E.D.

Applications. Applying geometrical inequalities some very nice estimates can be deduced in function theory and in mathematical physics [1], [2].

We describe below a few results which are direct consequences of inequalities on two dimensional surfaces.

Suppose D is a simply-connected domain in the complex z- plane, $z_0 \in D$ an arbitrary point and

$$f(z)=(z-z_0)+a_2(z-z_0)^2+\ldots$$

a complex one-to one function mapping D conformally onto the circle $\{w: |w|<R_{z_0}\}$.
It follows from the Riemann mapping theorem that such a function exists and that R_{z_0} is uniquely defined. R_{z_0} *is called the conformal radius of* D *with respect to* z_0 *and*

$$\dot{R}=\sup \{R_{z_0} : z_0 \in D\}$$

is called the maximal conformal radius of D.

Pólya and Szegö [18] have discovered a fundamental inequality which relates the area A of D and the conformal radius:

$$\pi \dot{R}^2 \leq A$$

The equality sign being attained if and only if D is a circle. Consider in D a Riemannian metric $d\sigma^2=p ds^2$ of bounded Gaussian curvature K_0 and denote by A_σ the total area of D with respect to this metric. Then the following inequality (cf. [20]) holds:

$$R_z^2 \leq \frac{4A_\sigma}{p(z)(4\pi-K_0A_\sigma)} \quad \text{if } K_0A_\sigma<4\pi.$$

The above estimate holds for the maximal conformal radius if z is the point such that $R_z=\dot{R}$. Equality holds for the circle centered at the origin with the metric of constant Gaussian curvature K that is

$$d\sigma^2 = \frac{b}{\left(1+\frac{bK_o}{4}|z|^2\right)^2}\, ds^2 = e^{\hat{u}(r;b,K_o)}\, ds^2.$$

Because of the variational characterization of the eigenvalues upper bounds are relatively easier to construct and there are several isoperimetric inequalities providing such bounds (cf. [1], [2]). We would also like to mention the Pólya-Schiffers's inequality [17] concerning the connection of the maximal conformal radius with the sum of the reciprocal first n eigenvalues. Thi can be stated in the following way:

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ *be the first* n *eigenvalues of the fixed membrane equation in a simply connected domain* D *and let* $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n$, *be the corresponding eigenvalues of the circle of radius* 1. *Then*

$$\sum_{i=1}^{n} \lambda_i^{-1} \geq \dot{R}^2 \sum_{i=1}^{n} \hat{\lambda}_i^{-1},$$

where $\dot{R}$ *denotes the maximal conformal radius of* D.

This inequality has a natural extension to non-homogeneous membran [1], which can be stated as follows:

Theorem 7. ([1]). *Let* D *be a simply connected domain,* $z_0 \in D$ *an arbitrary point and* p *a mass density satisfying*

$$\Delta \log p + 2K_o p \geq 0, \text{ and}$$

$$K_o \int_D p\, dx \leq 2\pi$$

Set

$$\beta = p(z_0) R_{z_0}^2, \quad \text{and}$$

$$e^{\hat{u}(r,\beta;K_o)} = \frac{\beta}{\left(1+\frac{\beta K_o r^2}{4}\right)^2}.$$

Note that β *is a conformal invariant. If* $\hat{\lambda}_i$ *is the ith eigenvalu of*

$$\begin{cases} \Delta\hat{\phi} + \hat{\lambda}\, e^{\hat{u}(r,\beta;K_o)} \hat{\phi} = 0 & \text{in } \{x: |x|<1\} \\ \hat{\phi} = 0 & \text{on } \{x: |x|=1\} \end{cases}$$

then

$$\sum_{i=1}^{n} \lambda_i^{-1} \geq \sum_{i=1}^{n} \hat{\lambda}_i^{-1}.$$

Some sharper versions of Pólya and Schiffer's result for symmetric regions and an extension to multiply connected domains can be found in the very nice book of C.Bandle [1].

Very little is known for the *free membrane* described by the eigenvalue problem

$$\begin{cases} \Delta\psi+\nu\psi=0 & \text{in } D\subset\mathbb{R}^2, \\ \frac{\partial\psi}{\partial n}=0 & \text{on } \partial D, \end{cases}$$

where $\frac{\partial}{\partial n}$ denotes the outer normal derivative. By standard results there exists a countable number of eigenvalues $0=\nu_1<\nu_2\leq\ldots$ The following extremal property holds for a circle:
Among all domains of given area the circle yields the highest second eigenvalue ν_2.

This result can take the form of an inequality in the following way:

$$\nu_2\leq\frac{\pi p_1^2}{A},$$

where $p_1=1.841\ldots$ zero of the Bessel function J_1.
This result can easily be extended to the problem

$$\begin{cases} \Delta_S\psi+\nu\psi=0 & \text{in } D\subset\mathbb{R}^2 \\ \frac{\partial\psi}{\partial n}=0 & \text{on } \partial D \end{cases}$$

Theorem 8 ([1]). *Let* D *be a simply connected domain on* S *whose Gaussian curvature is bounded from above by* K_0. *If the total area* A_σ *of* D *satisfies* $K_0A_\sigma\leq 2\pi$, *then the value of*

$$\frac{1}{\mu_{\nu_2}}+\frac{1}{\mu_{\nu_3}}$$

takes its minimum for a geodesic circle on a surface of constant curvature K_0.

Nehari [13] considered membranes with *mixed boundary conditions*

$$\begin{cases} \Delta\phi+\mu\phi=0 & \text{in } D\subset\mathbb{R}^2 \\ \phi=0 & \text{on } \Gamma \\ \frac{\partial\phi}{\partial n}=0 & \text{on } \gamma \end{cases}$$

where $\Gamma\cup\gamma=\partial D$ and $\Gamma\cap\gamma=\phi$. Nehari proved the following theorem

Theorem. *Let γ be a concave arc. Then* $\mu_1 \geq \frac{\pi j_0^2}{2A}$ (μ_1 = *the lowest eigenvalue). Equality holds for semi-circles with* Γ *as circular arc and* γ *as the straight segment.*

Bandle [1] has generalized Nehari's theorem in various ways. In fact the concavity of γ has been dropped and extensions to inhomogeneous membranes have been considered. Then terms involving the curvature of γ enter into the inequalities.

From the topological and geometrical point of view the spectrum of the Laplacian on a Riemannian manifold has been studied, and some very useful estimates for the first non-trivial eigenvalue μ_1 have been investigated. Lichnerowicz and Obata have proved the interesting result that *for compact 2-dimensional manifolds of positive Gaussian curvature* $K(x) \geq \kappa_0 > 0$ *the following holds:*

$$\mu_1 \geq 2\kappa_0$$

The equality holds only for surface isometric to the sphere of radius $\frac{1}{\sqrt{\kappa_0}}$.

Hersch [8] has obtained some upper bounds for μ_1 in the case of a surface homeomorphic to the sphere. He has obtained among other results that

$$\left(\frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} \right) \frac{1}{A} \geq \frac{3}{8\pi} ,$$

where A denotes the area of the surface, with equality holding for the sphere.

In [20] we have investigated main topological and stability properties of some of the most important examples of complete minimal surfaces in R^3, by making use of the Morse-Smale index theorem (see also [19]) which we have formulated in terms of eigenvalues. This way we have completed a global analysis of the index for the stability of a complete minimal surface in R^3.

REFERENCES

1. C. Bandle, *Isoperimetric Inequalities and Applications*, Pitman Publ. London (1980).

2. C. Bandle, *Isoperimetric inequalities*, Convexity and its Applications (eds:P.M. Gruber and J.M. Wills), Birkhäser Verlag, Basel, 1983, pp 30-48.

3. D.L. Bernstein *Existence Theorems in Partial Differential Equations*, Annals of Math. Studies 23, Princeton Univ.Press, Princeton, NJ,1950.

4. J.J.A.M. Brands, Bounds for the ratios of the first three membrane eigenvalues, Arch. Rational Mech.Anal.16(1964), 265-268.

5. R.Courant and D.Hilbert, *Methods of Mathematical Physics*, Vol.I, Interscience, New York,1953.

6. L.Durand, *Stability and oscillations of a soap film:An analytic treatment*, Amer. J.Phys. 49 (1981),334-343.

7. G.Faber, *Beweis dass unter allen homogenen Membrane von gleicher Fläche und gleicher Spannung die kreisformige den tiefsten Grundton gibt*, Sitz. bayer. Akad.,Wiss.(1923), 169-172.

8. J.Hersch, *Quatre propriétés isopérimetriques de membranes spériques homogenes*, C.R. Acad. Sci. Paris A 270(1970), 1645-1648.

9. G.N.Hile and M.H. Protter, *Inequalities for eigenvalues of the Laplacian*, Indiana University Mathematics Journal 29 (1980), 523-538.

10. H.Von Helmholtz, *Die Lehre von den Tonempfindungen*, 1862.

11. E.Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaf des Kreises*, Math. Ann. 94(1924), 97-100.

12. J.R.Kuttler and V.G. Sigillito, *Eigenvalues of the Laplacian in two dimensions*, SIAM Review, 26(1984), 163-193.

13. L.Nehari, *On the principal frequency of a membrane*, Pac.J. Math. 8(1958),285-293.

14. L.E. Payne, G.Polya and H.F.Weinberger, *On the ratio of cons*

cutive eigenvalues, Journal of Math. and Physics 35(1956), 289-298.

15. J.Peetre, *A generalization of Courant's nodal domain theorem*, Math. Scand. 5(1957), 15-20.

16. A.Pleijel, *Remarks on Courant's nodal line theorem*, Comm. Pure Appl. Math. 9(1956), 543-550.

17. G.Pólya and M.Schiffer, *Convexity of functionals by transplantation*, J.d'Anal.Math. 3(1954), 245-345.

18. G.Pólya and G.Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Princeton University Press (1951).

19. Th.M.Rassias, *Sur la multiplicité du premier bord conjugué d'une hypersurface minimale de* R^n, $n \geq 3$, C.R. Acad.Sciences Paris 284(1977), 497-499.

20. Th. M.Rassias, *Foundations of Global Nonlinear Analysis*, Teubner-Texte zur Mathematik, Band 86, Leipzig,1986.

21. E.Spanier, *Algebraic Topology*, McGraw-Hill, New York,1966.

CLOSED FORM EXPRESSIONS FOR SOME SERIES INVOLVING BESSEL FUNCTIONS OF THE FIRST KIND

M.S. STANKOVIĆ, D.M. PETKOVIĆ and M.V. DJURIĆ

ABSTRACT: *During a few last years a large number of papers have been written on the summation of series of Bessel functions. Most of these works dealt with some particular cases of series (1) and (2). There are only two notable exceptions to these; works by M.L. Glasser,* [7] *and B. C. Berndt,* [2], [4], *that serves as excellent background for the advanced material discussed here. In this paper we evaluate and represent the series (1), (2) as the series over Riemann zeta and related functions, which degenerate in closed form formulas in certain cases.*

1. INTRODUCTION

In mathematical physics, particularly in certain problems of telecommunication theory, electrostatics, etc., one often requires numerical values of sums involving Bessel functions, (1) and product of Bessel functions, (2). So, it is useful to have closed form expressions of as many of these as possible.

$$(1) \qquad S_{\nu,\alpha} = \sum_{n=1}^{\infty} \frac{(s)^{n-1} J_{\nu}((an-b)x)}{(an-b)^{\alpha}} \qquad s=1 \text{ or } -1, \quad \mu,\nu,\alpha \in R, \quad \alpha > 0$$

$$(2) \qquad S_{\mu,\nu,\alpha} = \sum_{n=1}^{\infty} \frac{(s)^{n-1} J_{\mu}((an-b)x) J_{\nu}((an-b)x)}{(an-b)^{\alpha}}$$

$J_{\nu}(x)$ are Bessel functions of the first kind and of order ν.

Various special cases can be derived from the general forms (1),(2) and have been treated in [3], [17], [19], [22], [25] and [5], [6], [7], [21], [23] respectively. It seems unli-

kely that these series can be expressed in closed form when the only restrictions are those which are essential to secure the convergence.

Motivated by impossibility to obtain closed form formulas in the general cases, we find them, under some restrictions, for the most frequently occuring class of series, i.e. for $a=1$, $b=0$ and $a=2$, $b=1$, in terms of Riemann zeta functions and other known sums of reciprocal powers.

Inspired by closed form expressions of trigonometric series, the general terms of which are reciprocal powers of integral variable, [15], [20], [21], [22], we expanded an analytical procedure in order to obtain the formulas of interest.

2. PRELIMINARIES

This section deals with some results connected with trigonometric series (3), [20]

$$(3)\qquad \sum_{n=1}^{\infty} \frac{(s)^{n-1} f((an-b)x)}{(an-b)^{\alpha}} = c\,\frac{\pi}{2\Gamma(\alpha) f(\frac{\pi\alpha}{2})}\, x^{\alpha-1} + \sum_{i=o}^{\infty} \frac{(-)^{i} F(\alpha-2i-\delta)}{(2i+\delta)!}\, x^{2i+\delta},$$

$$\alpha \in R$$

$$\alpha > 0$$

where $f = \{\begin{smallmatrix}\sin\\ \cos\end{smallmatrix}\}$, $\delta = \{\begin{smallmatrix}1\\ 0\end{smallmatrix}\}$ and where all relevant parameters are given in the table I. $\zeta(\alpha)$, $\eta(\alpha)$, $\lambda(\alpha)$ and $\beta(\alpha)$ are Riemann zeta functions and other sums of reciprocal powers, [1], [8].

Note that when $f(x)=\sin x$ and $\alpha \to 2m$ or $f(x)=\cos x$ and $\alpha \to 2m+1$, $m \in N_o$, the limiting value of the right-hand side of (3) should be taken into account, [14], [20].

Another important occurence of (3) is when the right-hand side series truncate due to the vanishing of F functions, so in the completely different way one can get closed form

Table I: corresponding F and c

a	b	s	c	F	for
1	0	1	1	ζ	$0 < x < 2\pi$
		-1	0	η	$-\pi < x < \pi$
2	1	1	$\frac{1}{2}$	λ	$0 < x < \pi$
		-1	0	β	$-\frac{\pi}{2} < x < \frac{\pi}{2}$

Table II: closed form cases

F	f	α
ζ, η, λ	sin	$2m+1$
	cos	$2m$
β	sin	$2m$
	cos	$2m+1$

formulas as in [20], [21]. These cases, which are of great importance for our further discussion, are pointed out in the table II. Two formulas of that type are known from Cesaro´s work, 1936., see e.g. [20] and, as always, some really particular cases can be found in [1], [10], [18], [25]. If the paper [15] is not the compilation, then the author rediscovered the results from [20], [21].

It should be mentioned that when (3) has the closed form, it is a simple matter to obtain the following recursion formulas, [16]:

$$F(2m+\delta) = c\,\frac{(-)^{m+1}\pi^{2m}}{2(2m)!} + \sum_{i=1}^{m} \frac{(-)^{i+1}F(2m-2i+\delta)\pi^{2i}}{2^{2i\delta}(2i+1-\delta)!}, \quad m \geq 1,$$

Table III:
Corresponding c and δ

F	ζ	η	λ	β
c	1	0	$\frac{1}{2}$	0
δ	0	0	0	1

Namely, $\zeta(2m)$, $\eta(2m)$, $\lambda(2m)$ are in proportion to π^{2m} and $\beta(2m+1)$ to π^{2m+1}. This fact is very useful in all closed form formulas we discuss here. In [11] one can find corresponding formula for $\zeta(2m)$.

3. OUTLINE OF THE BASIC PROCEDURE

The procedure we shall use is based on undoubtedly well known integral representation of Bessel functions:

$$(4) \qquad J_\nu(z) = 2\,\frac{(\frac{z}{2})^\nu}{\Gamma(\frac{1}{2})\Gamma(\nu+\frac{1}{2})}\int_0^{\frac{\pi}{2}} \sin^{2\nu}\theta\cos(z\cos\theta)\,d\theta, \quad \mathrm{Re}\,\nu > -\frac{1}{2}\,.$$

We shall substitute (4) in (1). It also states that it is possible to interchange the order of summation and integration. When we use this fact, the series (1) can be presented as follows:

$$S_{\nu,\alpha} = \frac{2(\frac{x}{2})^\nu}{\Gamma(\frac{1}{2})\Gamma(\nu+\frac{1}{2})}\int_0^{\frac{\pi}{2}} \sin^{2\nu}\theta \sum_{n=1}^{\infty} \frac{(s)^{n-1}\cos((an-b)x\cos\theta)}{(an-b)^{\alpha-\nu}}\,d\theta, \quad \alpha-\nu>0$$

Obviously, the part of the integrand is the series of the type (3). Further, we use (3) and this procedure leads to the integral the type of which is:

$$\int_0^{\frac{\pi}{2}} \sin^{\mu-1}x\cos^{\nu-1}x\;dx = \frac{1}{2}B(\frac{\mu}{2},\frac{\nu}{2}), \qquad \mathrm{Re}\,\mu>0, \quad \mathrm{Re}\,\nu>0.$$

We shall not go into details and instead merely state the final result (6). The condition $\alpha-\nu>0$ restricts this result to be of the most general character. That's why we recall the integral representation of Bessel functions, but of integral order this time:

$$J_n(z) = \frac{1}{\pi}\int_0^{\pi}\cos(z\sin\theta-n\theta)\,d\theta, \qquad n\in\mathbb{N}_0\,.$$

The same procedure as above leads to the integrals of the type

$$\int_0^{\pi} \sin^{\mu}x\, f(\nu x)\,dx = \frac{\pi}{2^{\mu}} f\left(\frac{\nu\pi}{2}\right) \frac{\Gamma(\mu+1)}{\Gamma\left(\frac{\mu+\nu}{2}+1\right)\Gamma\left(\frac{\mu-\nu}{2}+1\right)}, \quad f=\begin{Bmatrix}\sin\\ \cos\end{Bmatrix}, \quad \mathrm{Re}\,\mu>-1$$

and finally to the result (7).

The treatment of the series over product of Bessel functions, (2), demands the different integral representation, [24]

$$(5) \qquad J_{\mu}(z)J_{\nu}(z) = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} J_{\mu+\nu}(2z\cos\theta)\cos(\mu-\nu)\theta\,d\theta, \qquad \mu,\nu \in R,\ \mu+\nu>-1$$

As it was done previosly, we insert the integral representation into the series under consideration. Changing the order of summation and integration gives

$$S_{\mu,\nu,\alpha} = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \cos(\mu-\nu)\theta \sum_{n=1}^{\infty} \frac{(s)^{n-1} J_{\mu+\nu}(2(an-b)x\cos\theta)}{(an-b)^{\alpha}}\,d\theta, \qquad \alpha>0\ .$$

The series in the integrand is of the type (1) and for $\alpha>\mu+\nu>-\frac{1}{2}$ has the sum (6) and for $\mu+\nu\in N_0$ has the sum (7). In this way we easily obtain the final results (8) and (9), where the integral of the type

$$\int_0^{\frac{\pi}{2}} \cos^{\mu}x\,\cos\nu x\,dx = \frac{\pi}{2^{\mu+1}} \frac{\Gamma(\mu+1)}{\Gamma\left(\frac{\mu+\nu}{2}+1\right)\Gamma\left(\frac{\mu-\nu}{2}+1\right)}, \qquad \mathrm{Re}\,\mu>-1$$

is tacitly used.

4. RESULTS AND DISCUSSION

We are now in position to give the sum of the series (1):

$$(6)\qquad S_{\nu,\alpha} = c\frac{\Gamma\left(\frac{\nu-\alpha+1}{2}\right)}{2\Gamma\left(\frac{\alpha+\nu+1}{2}\right)}\left(\frac{x}{2}\right)^{\alpha-1} + \sum_{i=0}^{\infty}\frac{(-)^i F(\alpha-\nu-2i)}{i!\,\Gamma(\nu+i+1)}\left(\frac{x}{2}\right)^{2i+\nu},$$

$$\alpha,\nu \in R,\qquad \alpha>0,$$

$$\alpha>\nu>-\frac{1}{2},$$

$$(7)\qquad S_{m,\alpha} = c\frac{(-)^{\frac{m-\delta}{2}}\pi}{2\Gamma\left(\frac{\alpha+m+1}{2}\right)\Gamma\left(\frac{\alpha-m+1}{2}\right)f\left(\frac{\pi\alpha}{2}\right)}\left(\frac{x}{2}\right)^{\alpha-1} + \sum_{i=0}^{\infty}\frac{(-)^i F(\alpha-m-2i)}{i!\,(m+i)!}\left(\frac{x}{2}\right)^{2i+m},$$

$$\alpha\in R,$$

$$\alpha>0,$$

where $m=\{{}^{2k+1}_{2k}\}$, $f=\{{}^{\sin}_{\cos}\}$, $\delta=\{{}^{1}_{0}\}$, $k\in N_o$ and where c and F are readable from the table I.

In the case $\alpha-\nu=2k+1$ in (6) and $m-\alpha=2k+1$ in (7), $k\in N_o$, one should work either with limiting values or with principal values of gamma functions.

The chief disadvantage of the formula (6) is unvalidity for $\nu>\alpha$ and therefore (7) is derived, but only for $m\in N_o$. Even in the case $\alpha=\nu=1$ formula (6) holds true, although it does not seem possible, and gives the same result as (7).

Based on Mellin transform, one can find in [17] slightly different and less general ($a=1$, $b=0$, $s=1$, $\alpha-\nu\neq 2k+1$, $\max\{1-\alpha,-\nu\}>\frac{3}{2}$) result than (6).

In spite of the simplicity of the applied procedure it seems that (6) and (7) are the best published results and have not been noticed until now, as far as the authors are informed.

Special, but very useful cases of (6) and (7), [17], [19], [22], [25], we get for $\alpha-\nu-\delta$ even, where δ is given in the table III. Then the right-hand series terminate due to the vanishing of F functions, as it is already pointed out. Particularly, for $s=-1$ and $m>\alpha\in N$ the sum (7) is equal to zero and

it is very useful in accelerating the convergence of certain class of Bessel series.

It is almost obvious that for $\nu=k+\frac{1}{2}$, $k\in N_o$, (6) reduces to:

$$\sum_{n=1}^{\infty}\frac{(s)^{n-1}j_k((an-b)x)}{(an-b)^{\alpha}}=c\,\frac{\sqrt{\pi}}{4}\,\frac{\Gamma(\frac{k-\alpha+1}{2})}{\Gamma(\frac{k+\alpha}{2}+1)}(\frac{x}{2})^{\alpha-1}+\sum_{i=o}^{\infty}\frac{(-)^i F(\alpha-k-2i)}{(2i)!!(2k+2i+1)!!}x^{2i+k},$$

$$\alpha\in R$$

$$\alpha>k$$

where $j_k(x)$ are the spherical Bessel functions of the first kind.

Here we note in passing that the closed form expressions exist for the desired sums for $\alpha-k-\delta$ even.

Let us represent now the sum of series (2):

$$(8)\qquad S_{\mu,\nu,\alpha}=c\,\frac{\Gamma(\alpha)\Gamma(\frac{\mu+\nu-\alpha+1}{2})}{2\Gamma(\frac{\alpha+\mu+\nu+1}{2})\Gamma(\frac{\alpha+\mu-\nu+1}{2})\Gamma(\frac{\alpha-\mu+\nu+1}{2})}(\frac{x}{2})^{\alpha-1}+$$

$$+\sum_{i=o}^{\infty}\frac{(-)^i\Gamma(2i+\mu+\nu+1)F(\alpha-\mu-\nu-2i)}{i!\,\Gamma(i+\mu+1)\Gamma(i+\nu+1)\Gamma(i+\mu+\nu+1)}(\frac{x}{2})^{2i+\mu+\nu},$$

$$\alpha,\mu,\nu\in R,\qquad \alpha>0$$

$$\alpha>\mu+\nu>-\frac{1}{2},$$

$$(9)\qquad S_{\mu,\nu,\alpha}=c\,\frac{(-)^{\frac{\mu+\nu-\delta}{2}}\pi\Gamma(\alpha)}{2\Gamma(\frac{\alpha+\mu+\nu+1}{2})\Gamma(\frac{\alpha-\mu-\nu+1}{2})\Gamma(\frac{\alpha+\mu-\nu+1}{2})\Gamma(\frac{\alpha-\mu+\nu+1}{2})f(\frac{\pi\alpha}{2})}(\frac{x}{2})^{\alpha-1}+$$

$$+\sum_{i=o}^{\infty}\frac{(-)^i\binom{2i+\mu+\nu}{i}F(\alpha-\mu-\nu-2i)}{\Gamma(i+\mu+1)\Gamma(i+\nu+1)}(\frac{x}{2})^{2i+\mu+\nu},$$

$$\alpha,\mu,\nu\in R,\qquad \alpha>0,$$

$$\mu+\nu\in N_o,$$

where $\mu+\nu=\binom{2k+1}{2k}$, $f=\binom{\sin}{\cos}$, $\delta=\binom{1}{0}$, $k\in N_o$. F and c are given in the table I, where $2x$ should be taken in instead of x.

Analogously to (6) and (7), limiting or principal values of gamma functions are necessary for $\alpha-\mu-\nu=2k+1$ in (8) and for $\mu+\nu-\alpha=2k+1$ in (9), $k\in N_o$.

The shortcoming of (8), $\alpha>\mu+\nu>-\frac{1}{2}$, we due to the condition in (6). To overcome this, we additionaly give (9), but only for $\mu+\nu\in N_o$.

The reader will observe that the results just established have more general character than those discovered in [6]; besides, one of them is wrong ($\mu=1$, $\nu=0$, $\alpha=1$, $s=1$). This note one can find in [23], which is partialy incorrect, too.

For $\alpha-\mu-\nu$ even and for $s=1$, $a=1$, $b=0$ we obtain results from [7].

According to the concept of this paper, we wish to have closed form expressions and we get them from (8) and (9) for $\alpha-\mu-\nu+1+\delta$ even, where δ is given in table III. In that way the problem stated in [5] is more generaly solved. Some of these results are also given in [21].

The strange opinion of some colleagues is that some special cases of (6), (7) and (8), (9) should be pointed out. Thus, from the rich variety of closed form formulas we consider in particular (6) or (7) which for $a=2$, $b=1$, $s=1$ and $\nu=0$, $\alpha=2$ degenerate in:

$$\sum_{n=1}^{\infty} \frac{J_o((2n-1)x)}{(2n-1)^2} = \frac{\pi^2}{8} - \frac{x}{2}, \qquad 0 \le x \le \pi .$$

The formula (7) for $a=1$, $b=0$, $s=-1$ and $\alpha=m$ gives:

$$\sum_{n=1}^{\infty} \frac{(-)^{n-1} J_m(nx)}{n^m} = \frac{x^m}{2^{m+1}\Gamma(m+1)} .$$

The both formulas are in full agreement with (8), page 634. and (2), page 635. in [24].

Formulas (8) and (9) for $a=1$, $b=0$, $s=1$ lead to the known results: (13) in [7] and (10) in [23].

The fulfillments of the rest of whishes, as above, has no practical sense at present.

5. CONCLUSION

The series over Bessel functions are extremely useful for both analysis of Bessel functions and various applications. One important class of problems is obtaining closed form formulas. Although most of these formulas have been known for a long time, it seems that this important problem nevertheless has not been solved entirely.

In this paper we give some closed form expressions for two classes of series and we belive that these formulas might be useful for reducing further series, the sums of which are not known, to simpler cases or to the series the sums of which are known now. Also, we belive that this paper does contain some simple but fresh ideas.

Acknowledgments: *The authors express their sincere thanks to professors L. Gatteschi, University of Turin and B. Danković, University of Nish for fruitful discussions.*

REFERENCES

[1] *Abramowitz, M.; Stegun, A.: Handbook of Mathematical Functions, with Formulas, Graphs and Mathematical Tables, Dover Publications, N.Y., 1972.*

[2] *Askey, R.A.: Theory and Application of Special Functions, Academic Press, Inc., N.Y., 1975.*

[3] Berkesh, B.: Einige Formeln uber unendlichen Reihen Besselscher Funktionen, *Glasnik mat.fiz.astr.*,10 (1955), 161-170.

[4] Berndt, B.C.: The evaluation of character series by contour integration, *Univ.Beograd.Publ.Elektrotehn.Fak.Ser.Mat.Fiz.*, 386 (1972), 25-29.

[5] De Doelder, P.J.: On a series of product of Bessel functions of integral order, *Simon Stevin*, oct., (1960), 54-57.

[6] De Doelder, P.J.: Two infinite sums, problem 79-12, *SIAM Rev.*, 21 (1979), 395-396.

[7] Glasser, M.L.: A class of Bessel summations, *Math. comp.*, 37 (1981), 54-57.

[8] Glasser, M.L.: The evaluation of lattice sums. I. Analytic procedure, *J.Math.Phys.*, 14 (1973), 409-413.

[9] Glasser, M.L.: Private communications, Clarkson University, 1987.

[10] Gradshtein, I.S.; Ryzhik, I.M.: Tablicy Integralov, Summ, Ryadov i. Proizvedenii, Nauka, Moskva, 1971.

[11] Janković, Z.: Two recurrence formulas for the sums S_{2k}, *Glas.Mat. Ser.II*, 8 (1953), 27-29.

[12] Korenev, B.G.: Vvedenie v Teoriyu Besselevykh Funktsii, Nauka, Moskva, 1971.

[13] Lossers, O.P.: Private communications, Eindhoven University

[14] Mitrinović, D.S.; Adamović, D.D.: Nizovi i Redovi, Naučna knjiga, Beograd, 1980.

[15] Moiseev, A.I.: O razlozhenii summ $\sum n^{-\alpha} \cos 2\pi n\theta$ i $\sum n^{-\alpha} \sin 2\pi n\theta$ po stepenyam θ, *Izv.Vyssh.Uchebn.Zaved.Mat.*, 4 (1986), 75-77, RZhMat, 9 Б 1795, 1986.

[16] Petković, D.M.: Problem H-381, *Fibonacci Quart.*, feb., (1985), 89.

[17] Petković, D.M.: Infinite sums of Bessel functions, problem 85-14, solution by Glasser, M.L., *SIAM Rev.*, 28 (1986), 402-403.

[18] Prudnikov, A.P.; Brychov, Yu.A.; Marichev, O.I.: Integraly i Ryady. Elementarnye Funktsii, Nauka, Moskva, 1981.

[19] Prudnikov, A.P.; Brychov, Yu.A.; Marichev, O.I.: Integraly i Ryady. Spetsialnye Funktsii, Nauka, Moskva, 1983.

[20] Slavić, D.V.: On summation of trigonometric series, Univ.Beograd.Publ. Elektrotehn.Fak.Ser.Mat.Fiz., 263 (1969), 103-114.

[21] Stanković, M.S.; Petković, D.M.: O sumiranju nekih redova pomoću Riemannovih zeta funkcija, Informatica'81, Ljubljana, okt., (1981), 3_106.

[22] Stanković, M.S.; Petković, D.M.; Djurić, M.V.: Short table of summable series of Bessel functions, Conf.Appl.Math.5., Ljubljana, sept.,(1986), 147-152, see RZhMat. 4 Б 16, 1987.

[23] Toshić, D.Dj.: Some series of product of Bessel functions, Univ.Beograd.Publ.Elektrotehn.Fak.Ser.Mat.Fiz., 678-715 (1980), 105-110.

[24] Watson, G.N.: A Treatise on the Theory of Bessel Functions, Cambridge Univ. Press, Cambridge, 1958.

[25] Zaezdnyi, A.M.: Garmonycheskii Sintez v Radiotekhnike i Elektrosvyazi, Energiya, Leningrad, 1972.

ASYMPTOTIC BEHAVIOUR OF THE OSCILLATION OF THE SEQUENCES OF THE LINEAR TRANSFORMATIONS OF THE FOURIER SERIES

VLADIMIR N. SAVIĆ

A B S T R A C T

In this paper we consider the asymptotic behaviour of

$$(I) \qquad \varepsilon_{mn}(W^r H^\omega;\ U_n) = \sup_{f \in W^r H^\omega} \|U_m(f) - U_n(f)\|_C$$

$$(m, n, r = 1, 2, \ldots;\ m > n)$$

where $U_n(f,x)$ is the sum of Fejér, Cesaro, Rogosinski, ... of the Fourier series of the function $f \in W^r H^\omega$.

ASIMPTOTSKO PONAŠANJE OSCILACIJE NIZA LINEARNIH TRANSFORMACIJA FOURIER-OVOG REDA FUNKCIJE f. U ovom radu razmatramo asimptotsko ponašanje izraza (I), gde je $U_n(f,x)$ suma Fejér-a, Cesaro-a, Rogosinskog, ... Fourier-ovog reda funkcije $f \in W^r H^\omega$.

If n is fixed, and m sufficiently large than ε_{mn} is approximatively equal the distance between U_n and f for each $f \in W^r H^\omega$.

Definition. Let $W^r H^\omega$ $(r \in N)$ be a set 2π-periodic continuous functions f, such that $f^{(r)} \in H^\omega$, or equivalent

$$(\forall x_1, x_2 \in R)\ |f^{(r)}(x_1) - f^{(r)}(x_2)| \leq \omega(|x_1 - x_2|)$$

where ω is the modulus of continuity.

For $f\in W^r$ and $f\in W^rH^\alpha$ $(0<\alpha\leqslant 1)$ we have [1] and [2] with the corresponding results.

The fundamental results follow from the lemma 1 (see [5]) and the lemma 2 (see [4])

Lemma 1. Let $\psi\in L[a,b]$, and suppose that

(i) $\quad \Psi(x)=\int_a^x \psi(t)dt$

(ii) $\quad \Psi\uparrow(\downarrow)$ on $]a,c[$ $\quad (a<c\leqslant b)$, and

$\quad \Psi\downarrow(\uparrow)$ on $]c,b[$

(iii) $\quad \Psi(b)=0.$

Then

$$\text{(II)}\quad \sup_{f\in H^\omega[a,b]}\left|\int_a^b \psi(t)f(t)dt\right| \leqslant \int_a^c |\psi(t)|\,\omega\big(\rho(t)-t\big)dt = \int_c^b |\psi(t)|\,\omega\big(t-\rho^{-1}(t)\big)dt,$$

where the function ρ is defined with

$$\Psi(x)=\Psi\big(\rho(x)\big)\quad \big(a\leqslant x\leqslant c\leqslant \rho(x)\leqslant b\big)$$

and ρ^{-1} is the inverse function of the function ρ.

If ω is a convex modulus of continuity, then, for the function $F(x)+C$ ($C\in R$ is arbitrary constant) we have $=$ in the formula (II), and

$$F(x)=\begin{cases} -\int_x^c \omega'\big(\rho(t)-t\big)dt, & a\leqslant x\leqslant c \\ \int_c^x \omega'\big(t-\rho^{-1}(t)\big)dt, & c\leqslant x\leqslant b \end{cases}$$

Lemma 2. Let (λ_{nk}) $(n,k\in N)$ be a matrix of real numbers such that $\lambda_{nk}=0$ for $k>n$, and the sequence

$$\left\{\frac{\lambda_{n+1,k}-\lambda_{nk}}{k^2}\right\}_{k=1,+\infty} \qquad (\forall n\in N)$$

is non - increasing. If

$$U_n(f,x) = \frac{1}{\pi}\int_0^{2\pi} f(x+t)\left[\frac{1}{2}+\sum_{k=1}^{n}\lambda_{nk}\cos kt\right]dt$$

then, for a convex modulus of continuity ω, for all $r\in N$ $(r \geq 3)$ and for all $m, n\in N$ $(m > n)$

$$\varepsilon_{m,n}(W^r H^\omega;\ U_n) =$$

$$= \begin{cases} \dfrac{2}{\pi}\displaystyle\sum_{k=0}^{\left[\frac{m-1}{2}\right]} \frac{\lambda_{m,2k+1}-\lambda_{n,2k+1}}{(2k+1)^r}\int_0^{\frac{\pi}{2}}\omega(2t)\sin(2k+1)t\,dt\ ; \quad r = 2i-1 \\ \qquad\qquad (i = 2, 3, \ldots) \\ \dfrac{1}{\pi}\displaystyle\int_0^{2\pi}\psi_{r,\lambda}(t)F_{r,\lambda}(t)dt; \quad r = 2i \quad (i = 2,3,\ldots) \end{cases}$$

where

$$F_{r,\lambda}(t) = \begin{cases} F^1_{r,\lambda}(t), & 0 \leq t \leq \pi \\ -F^1_{r,\lambda}(t), & \pi \leq t \leq 2\pi \end{cases},$$

$$F^1_{r,\lambda}(t) = \begin{cases} -\displaystyle\int_x^{t_0}\omega'\left(\rho(t)-t\right)dt, & 0 \leq x \leq t_0 \\ \displaystyle\int_{t_0}^{x}\omega'\left(t-\rho^{-1}(t)\right)dt, & t_0 \leq x \leq \pi. \end{cases},$$

t_0 is a zero of the function

$$\psi_{r,\lambda}(t) = \sum_{k=1}^{m}\frac{\lambda_{m,k}-\lambda_{n,k}}{k^r}$$

on $[0,\pi]$, and the function ρ is defined with

$$\int_0^x \psi_{r,\lambda}(t)dt = \int_0^{\rho(x)}\psi_{r,\lambda}(t)dt \quad (0 \leq x \leq t_0 \leq \rho(x) \leq \pi),$$

and ρ^{-1} is the inverse function of the function ρ.

Let, now, $(\sigma_n(f,x))$ be a sequence of the sums of Fejér of the Fourier series of the function f, i.e.

$$\sigma_n(f,x) = \frac{1}{\pi}\int_0^{2\pi} f(x+t)\left[\frac{1}{2} + \sum_{k=1}^{n}\left(1 - \frac{k}{n+1}\right)\cos kt\right]dt$$

where

$$\lambda_{nk} = \begin{cases} 1 - \dfrac{k}{n+1}, & k \leq n \\ 0, & k > n \end{cases}$$

Now, we prove

Theorem 1. For all $m, n \in N$ $(m > n)$ and for a convex modulus of continuity ω we have the asymptotic equality

$$\varepsilon_{m,n}(W^r H^\omega;\ U_n) = \begin{cases} C_1 \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\omega\left(\frac{1}{n}\right)\right)\right), & r = 1 \\ C_2 \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\frac{1}{n}\right)\right), & r = 2 \\ C_r \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\dfrac{1}{n^{r-1}}\right)\omega\left(\frac{1}{n}\right)\right), & r = 2i-1 \ (i = 2,3,\ldots) \\ C_r^1 \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\dfrac{1}{n^{r-1}}\omega\left(\frac{1}{n}\right)\right)\right), & r = 2i \ (i = 2,3,\ldots) \end{cases}$$

where

$$C_1 = \frac{2}{\pi}\sum_{k=0}^{\infty}\int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt$$

$$C_2 = -\frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\int_0^{2\pi} F_{2,\lambda}(t)\cos kt\,dt$$

$$C_r = \frac{2}{\pi}\sum_{k=0}^{\infty}\frac{1}{(2k+1)^{r-1}}\int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt$$

$$C_r^1 = -\frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k^{r-1}}\int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt$$

Proof. For $r = 2i-1$ and $r = 2i$ from the theorem 1 (see [4]) we have

$$(1)\quad \varepsilon_{mn}(W^r H^\omega;\ U_n) = \frac{2}{\pi}\left[\frac{m-n}{(m+1)(n+1)}\sum_{k=0}^{\infty}\frac{1}{(2k+1)^{r-1}}\int_0^{\frac{\pi}{2}}\omega(2t)\sin(2k+1)t\,dt -\right.$$

$$- \frac{m-n}{(m+1)(n+1)} \sum_{k=\left[\frac{n-1}{2}\right]+1}^{\infty} \frac{1}{(2k+1)^{r-1}} \int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt +$$

$$+ \frac{1}{m+1} \sum_{k=\left[\frac{n-1}{2}\right]+1}^{\left[\frac{m-1}{2}\right]} \frac{m-2k}{(2k+1)^{r}} \int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt \Bigg], \quad r = 2i-1, \quad (i = 2,3,\ldots)$$

$$(2)\quad \varepsilon_{mn}(W^r H^{\omega};\ U_n) = \frac{1}{\pi}\left[- \frac{m-n}{(m+1)(n+1)} \sum_{k=1}^{\infty} \frac{1}{k^{r-1}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt + \right.$$

$$+ \frac{m-n}{(m+1)(n+1)} \sum_{k=n+1}^{\infty} \frac{1}{k^{r-1}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt +$$

$$\left. + \sum_{k=n+1}^{m} \frac{k-(m+1)}{(m+1)k^{r}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt \right] = \sum^{0} + \sum^{1} + \sum^{2}, \quad r = 2i, \quad (i = 2,3,\ldots)$$

Since we have

$$\int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt = 0\left(\frac{1}{2k+1}\,\omega\left(\frac{1}{2k+1}\right)\right), \qquad (k \to +\infty)$$

from (1), we obtain the theorem 1, for $r = 2i-1$ $(i = 2,3,\ldots)$.

Let, now, $f_1(t) = F_{r,\lambda}(t) - F_{r,\lambda}(0)$,

$$D_n^{(r)}(t) = \sum_{k=n+1}^{\infty} \frac{\cos kt}{k^{r-1}}, \qquad D_{mn}^{(r)}(t) = \sum_{k=n+1}^{m} \left(1 - \frac{k}{m+1}\right)\cos kt,$$

then, by [3], we get

$$(3)\quad |\sum^{1}| = \left|\frac{1}{\pi}\int_0^{2\pi} f_1(t) D_n^{(r)}(t)\,dt\right| = 0\left(\frac{m-n}{mn^{r}}\,\omega\left(\frac{1}{n}\right)\right)$$

$$(4)\quad |\sum^{2}| = \left|\frac{1}{\pi}\int_0^{2\pi} f_1(t) D_{mn}^{(r)}(t)\,dt\right| = 0\left(\frac{m-n}{mn^{r}}\,\omega\left(\frac{1}{n}\right)\right)$$

From (3), (4) and (2) it follows the theorem 1 for $r = 2i$ $(i = 2,3,\ldots)$.

If $(\tilde{\sigma}_n(f; x)$ is a sequence of the conjugate sums of Fejér of the Fourier series of the function f, then, we have, by the theorem 2 from [4]

Theorem 2. For all m, n$\in$N (m > n) and for a convex modulus of continuity ω we have the asymptotic equality

$$\varepsilon_{mn}(W^r H^\omega;\ \tilde{\sigma}_n) =$$

$$= \begin{cases} \frac{2}{\pi}\,\frac{m-n}{(m+1)(n+1)}\left\{\bar{C}_r + O\left(\frac{1}{n^{r-1}}\,\omega\left(\frac{1}{n}\right)\right)\right\} & r = 2i \\ (i = 1,2,\ldots) \\ \frac{1}{\pi}\,\frac{m-n}{(m+1)(n+1)}\left\{\bar{\bar{C}}_r + O\left(\frac{1}{n^{r-1}}\,\omega\left(\frac{1}{n}\right)\right)\right\} & r = 2i+1 \\ (i = 1,2,\ldots) \end{cases}$$

where

$$\bar{C}_r = \sum_{k=0}^{\infty} \frac{1}{(2k+1)^{r-1}} \int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt$$

$$\bar{\bar{C}}_r = -\sum_{k=1}^{\infty} \frac{1}{k^{r-1}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt.$$

R E F E R E N C E S

1. Рыжанкова Г. И. О колебаниях последовательностей некотор линейных преобразований рядов Фурье, автореферат кандидатской диссертации, Киев, 1972.
2. Филипповский В. Г. О колебании последовательности полиномов, порождаемых линейными метдами суммирования рядов Фурье на классах функций Гельдера, Сборник статей: "Теория приближения функций и ее приложения", Издание Института математики АН УССР, Киев, 1974, 158-181.
3. Ефимов А. В. Приближение непрерывных периодических функци суммами Фурье, Известия АН СССР, 24 (1960), 243-296.

. Savić V.N. The oscillation of the sequences of the linear transformations of the Fourier series of the function f. Collection of scientific papers of the Faculty of Science Kragujevac, 8(1987).

5. Savić V.N. O jednom ekstremalnom problemu u prostoru neprekidnih funkcija od n promenljivih. Mat.vesnik 6 (19) (34), 1982. 165-172.

lumerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

UNIFORMLY CONVERGENT SPLINE COLLOCATION METHOD FOR A DIFFERENTIAL EQUATION WITH A SMALL PARAMETAR

K. SURLA

ABSTRACT: *For the problem:* $\varepsilon y'' + p(x)y' = f(x)$, $-\alpha y(0) + y'(0) = \alpha_0$, $y(1) = \alpha_1$, $p(x) \geq \bar{p} > 0$, *the cubic spline collocation method is derived. The uniform convergence of the first order on locally bounded mesch is achived. The method has the second order of the convergence for fixed* ε.

1. INTRODUCTION

Consider the singularly perturbed two-point boundary value problem:

$$(1) \qquad \begin{cases} Ly = \varepsilon y'' + p(x)y' = f(x), & 0 \leq x \leq 1, \quad 0 < \varepsilon \ll 1, \\ y'(0) - \alpha y(0) = \alpha_0, & y(1) = \alpha_1; \quad \alpha_0, \alpha_1 \in \mathbb{R}, \quad \alpha \geq 0, \end{cases}$$

where the functions $p, f \in C^2[0,1]$, $p(x) \geq \bar{p} > 0$. Under these assumptions problem (1) has a unique solution $y = y(x)$, which exibits a boundary layer at $x = 0$ for small ε, [2].

The ordinary cubic spline collocation methods when applied to (1) have an inherent formal cell Reynolds number limitation, i.e. $h_j p(x_j)/2\varepsilon$ must be less than or equal to 1, [3]. For "small" ε this leads to the spurious oscillations or large inaccuracies in the approximate solution, (see [1],[2]), $h_j = x_{j+1} - x_j$, $j = 0(1)n$, x_j are the points of the grid Δ.

$$\Delta: \ 0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1.$$

In order to avoid these difficulties in case of Diriclet's boundary conditions in [5] the exponential features

of the exact solution are transfered to the spline coefficients by introducing the relaxation parameter affecting the highest derivative. This parameter is determined in such a way that the truncation error of the corresponding difference scheme for the boundary layer function in case $p(x) = p = \text{const}$, vanishes. This procedure is known as the exponential fitting or the introduction of "artificial viscosity". The spline difference schemes have the same order of accuracy on a uniform and non-uniform mesh [6]. It might be expected that the exponentially fitted spline difference schemes preserve this property. However, in the case of Dirichlet's conditions the uniform convergence in [5] is obtained by putting some special conditions on the grid. The dependence of the exact solution of ε in the case of mixed boundary conditions of the type (1) is smoller then in the case of Dirichlet's one. Because of that the first order of the uniform convergence can be obtained with properly bounded local mesh ration: $h_i/h_{i\pm1} \le M$, M is a constant independent of ε and h_i. Thus, in this case the exponentially fitted spline difference scheme has the same order of the accuracy on the equidistant grid and non-equidistant one (as in [6]).

2. DERIVATION OF THE SCHEME

We seek the solution of the problem (1) in the form of the cubic spline $v(x) \in C^2[0,1]$ on the grid Δ. On each interval $[x_i, x_{i+1}]$, spline $v(x)$ has the form:

$$(2) \qquad v(x) = v_j(x) = v_j^{(0)} + (x-x_j)v_j^{(1)} + (x-x_j)^2\frac{v_j^{(2)}}{2} + \frac{(x-x_j)^3}{6}v_j^{(3)}.$$

The constants $v_j^{(k)}$ are determined from the equations (see [3]):

$$(3) \qquad \varepsilon\bar{\sigma}_j v_j^{(2)} + p_j v_j^{(1)} = f_j, \quad j = 0(1)n+1,$$

$$\bar{\sigma}_j = \rho_j \operatorname{cth}\rho_j, \quad \rho_j = h_j p_j / (2\varepsilon),$$

$$(4) \qquad v_j^{(k)}(x_j) = v_{j-1}^{(k)}(x_j), \quad k = 0,1,2; \quad j = 1(1)n,$$

$$5) \qquad v_0^{(1)} - \alpha v_0^{(0)} = \alpha_0, \quad v_{n+1}^{(0)} = \alpha_1.$$

The system (3)-(5) has 4n+4 unknowns and 4n+4 equations. The first equation presents the collocation relaxed by introducing the parameter $\bar{\sigma}_i$ (fitting factor). Equations (4) are the consequence of the continuity conditions, $v(x) \in C^2[0,1]$. By eliminating the unknowns $v_j^{(k)}$, $k = 1,2,3$; from the above equations we obtain the scheme:

$$(6) \qquad \begin{aligned} & R_h v_j = Q_h f_j, \quad j = 0(1)n \\ & v_{n+1} = \alpha_1, \quad \text{where} \\ & R_h v_j = r_j^- v_{j-1} + r_j^c v_j + r_j^+ v_{j+1}, \quad \text{for } j = 0(1)n, \\ & Q_h f_j = q_j^- f_{j-1} + q_j^c f_j + q_j^+ f_{j+1}, \quad \text{for } j=1(1)n \quad \text{and} \end{aligned}$$

$$r_j^- = \frac{3(w_{j-1}-1)}{h_{j-1}A_{j-1}}, \quad r_j^+ = \frac{3(w_{j+1}+H_j)}{h_j A_j}, \quad r_j^c = -r_j^- - r_j^+ ,$$

$$A_j = 3w_j w_{j+1} + 2H_j w_j - 2w_{j+1} - H_j, \quad w_j = c\,\mathrm{th}\,\rho_j,$$

$$q_j^+ = H_j/(p_{j+1}A_j), \quad q_j^- = 1/(p_{j-1}A_{j-1}),$$

$$q_j^c = \frac{H_{j-1}(2w_{j-1}-1)}{p_j A_{j-1} w_j} + \frac{2w_{j+1}+H_j}{p_j A_j w_j} \quad , \text{ for } j = 1(1)n.$$

Further,

$$r_0^- = 0, \quad r_0^+ = \gamma_1^{-1}, \quad r_0^c = -(1+\gamma_1\alpha)\gamma_1^{-1}$$

$$Q_h f_0 = -\alpha_0 - s_1\gamma_1^{-1}, \quad \gamma_1 = h_0\left(1 - \frac{h_0 p_0}{3\sigma_0} - h_0\,\frac{b_1 p_1}{6\sigma_1 a_1}\right),$$

$$s_1 = \frac{h_0^2}{6}\left(2\,\frac{f_0}{\sigma_0} - \frac{p_1 R_1}{\sigma_1 a_1} + \frac{f_1}{\sigma_1}\right), \quad a_j = 1 + \frac{h_{j-1}p_j}{2\sigma_j}$$

$$b_j = 1 - \frac{h_{j-1}p_{j-1}}{2\sigma_{j-1}}, \quad R_1 = \frac{h_0}{2}\left(\frac{f_0}{\sigma_0} + \frac{f_1}{\sigma_1}\right), \quad H_j = \frac{h_j}{h_{j+1}}, \quad p_j = p(x_j),$$

$$f_j = f(x_j), \qquad \sigma_j = \varepsilon\bar{\sigma}_j.$$

3. THE PROOF OF THE UNIFORM CONVERGENCE

The proof is based on the comparison function method which requires the following lemmas, [1].

LEMMA 1. ([2]). Let $f,p \in C^2[0,1]$. Then the solution of (1) satisfies the inequalities

$$|y^{(i)}(x)| \le M(1+\varepsilon^{-i+1}\exp(-2\delta x/\varepsilon)), \quad i = 0(1)4.$$

M and δ are constants independent of ε.

LEMMA 2. (maximum principle)

Let $\{v_j\}$ be a set of values at the grid points x_j satisfying $R_h v_j \ge 0$, $j = 0(1)n$. Then, $v_j \le 0$, $j = 0(1)n$.

Throughout the paper M denotes the different constants independent of ε and h_i.

LEMMA 3. There exist constants M and β independent of h_j and ε such that for $j = 1(1)n$.

a) $R_h\phi_j \ge M\min(\frac{h_j^2}{\varepsilon^2}, 1)$,

b) $R_h\psi_j \ge M\mu_j(\beta)h_j^{-1}\min(h_j^3/\varepsilon^3, 1)$,

c) $R_h\phi_o \ge M$,

d) $R_h\psi_o \ge M\mu_o(\beta)h_o^{-1}\min(h_o/\varepsilon, 1)$,

$\phi_j = -2+x_j$, $\psi_j = -\exp(-\beta t_j)$, $\mu_j(\beta) = \exp(-\beta t_j)$,

$t_j = x_j/\varepsilon$.

Functions $\phi(x)$ and $\psi(x)$ are comparison functions and we use them in order to determine how the operator R_h affects the characteristic parts of the solution $y(x)$ (β is the smallest of various positive constants appearing in the proof). From Lemma 2 and Lemma 3 we can see that

$$|v_j - y_j| \le k_1|\phi_j| + k_2|\psi_j| \tag{7}$$

if

$k_1(h_j,\varepsilon) \geq 0$ and $k_2(h_j,\varepsilon) \geq 0$ are such functions that

$$R_h(k_1\phi_j + k_2\psi_j) \geq R_h(\pm z_j) = \pm\tau_j(y) \tag{8}$$

$z_j = y_j - v_j$, $\tau_j(y)$ is a truncation error of the scheme (6) for the function y. For an arbitrary smooth function g, $\tau_j(g)$ is given by

$$\tau_j(g) = R_h g_j - Q_h(Lg)_j \ .$$

LEMMA 4. The truncation error $\tau_j(y)$ can be written in the form

$$\tau_j(y) = R_h z_j = \left(\frac{a_j\phi_{j+1,2}}{\gamma_{j+1}} - b_j\frac{\phi_{j,2}}{\gamma_j} + \phi_{j,1}\right)/(w_j+H_{j-1})$$

$j = 1(1)n$, where

$$\phi_{j,1} = \psi_{j,1} - \frac{h_{j-1}}{2}\psi_{j,2} + \frac{h_{j-1}}{2}\left(\frac{\eta_{j-1}}{\sigma_{j-1}} + \frac{\eta_j}{\sigma_j}\right)$$

$$\phi_{j,2} = \psi_{j,0} + h_{j-1}^2\left(\frac{\eta_{j-1}}{3\sigma_{j-1}} + \frac{\eta_j}{6\sigma_j} - \frac{\psi_{j,2}}{6} + p_j\frac{\phi_{j,1}}{6a_j\sigma_j}\right),$$

$\eta_j = y_j''(\sigma_j - \varepsilon)$, $\sigma_j = \varepsilon\rho_j w_j$, $\psi_{j,k} = \dfrac{h_{j-1}^{4-k}}{(4-k)!}y^{IV}(\xi_j)$, ξ_j is a fixed point belongs to $[x_{j-1},x_j]$.

For the proof see [3] or [5].

THEOREM 1. Let $f,p \in C^2[0,1]$, $p(x) \geq \bar{p} > 0$. Let v_j be defined by (6) on the grid Δ, where $h_j/h_{j\pm1} \leq M$. Then

$$|y(x_j)-v_j| \leq Mh_j^2/(\varepsilon+h_j). \tag{9}$$

Proof. From Lemma 4 and Lemma 1 we have

$$|\tau_j(y)| \leq M\frac{h_j^2}{h_j+\varepsilon}\, h_j^2\varepsilon^{-2} + \exp(-\delta x_j/\varepsilon)h_j^4\varepsilon^{-4}, \quad h_j \leq \varepsilon, \ j=1(1)n. \tag{10}$$

From Taylor's development about x_j we also have

(11) $$|\tau_0(y)| \le Mh_0^2/(h_0+\varepsilon) + h_0^2\varepsilon^{-2}\exp(-\delta x_0/\varepsilon).$$

Further, for $\varepsilon \le h_j$, after several Taylor's expansions we obtain

$$\tau_j(y) = r_j^- \frac{h_{j-1}^2}{2} y''(\xi_{1j}) + r_j^+ \frac{h_j^2}{2} y''(\xi_{2j}) + q_j^- p_{j-1} h_{j-1} y''(\xi_{3j}) -$$

$$-q_j + p_j h_j y''(\xi_{4j}) - \varepsilon(q_j^- y''_{j-1} + q_j^c y''_j + q_j^+ y''_{j+1}) ,$$

$$x_{j-1} \le \xi_{1,j},\ \xi_{3,j} \le x_j \le \xi_{2,j},\ \xi_{4j} \le x_{j+1}.$$

Since

$$|w_j| \le M,\ h_j/h_{j\pm1} \le M,\ h^k/\varepsilon^k \exp(-\delta x_j/\varepsilon) \le M\exp(-\delta x_j/2\varepsilon),$$

we have $|r_j^{\pm}| \le Mh_j^{-1}$, $|q_j^{\pm c}| \le M$ and

(12) $$|\tau_j(y)| \le M(h_j + \exp(-\delta x_{j-1}/\varepsilon)),\ j=1(1)n.$$

In a similar way, we obtain, for j=0

(13) $$|\tau_0(y)| \le M(h_0 + \exp(-\delta x_0/\varepsilon)).$$

If we take $k_1(h_j,\varepsilon) = h_j^2/(h_j+\varepsilon)$ and $k_2(h_j,\varepsilon) = h_j^2/\varepsilon$ for $h_j \le \varepsilon$, from (10), (11) and Lemma 2 we can see that (8) holds.

This leads to the estimates (7) and (9).

If $\varepsilon \le h_j$ we can take $k_1(h_j,\varepsilon) = 1$, $k_2(h_j,\varepsilon) = h_j$ and from (12),(13) and Lemma 3 we have that (8) holds, and so does Theorem 1.

THEOREM 2. Let the conditions of Theorem 1 be satisfied. Then

(14) $$|y(x)-v(x)| \le Mh^2/(\varepsilon+h), \quad h = \max_i h_i,$$

M is a constant independent of ε and h.

Proof. Since $z(x)=y(x)-v(x) \in C^4[x_j,x_{j+1}]$ we have that:

$$(15)\qquad z(x) = z_j^{(0)} + z_j^{(1)}(x-x_j) + z_j^{(2)}\frac{(x-x_j)^2}{2} + z_j^{(3)}\frac{(x-x_j)^3}{3!} +$$

$$+ y^{IV}(\xi_j)\frac{(x-x_j)^4}{4!}\,,\quad x_j \le \xi_j \le x_{j+1},\quad x \in [x_j, x_{j+1}]$$

$$z_o^{(1)} - \alpha z_o^{(0)} = 0 \quad \text{and} \quad |z_o^{(1)}| \le Mh_o^2/(h_o+\varepsilon).$$

Further,

$$a_j z_j^{(1)} = b_j z_{j-1}^{(1)} + \phi_{j,1},\ \text{and}\ |z_j^{(1)}| \le Mh_j/(h_j+\varepsilon),\ j=1(1)n.$$

$$|z_j^{(2)}| \le |(\eta_j - p_j z_j^{(1)})/\sigma_j| \le M\varepsilon^{-1}h_j/(\varepsilon+h_j),$$

$$|z_{j-1}^{(3)}| \le (z_j^{(2)} - z_{j-1}^{(2)} - \psi_{j,2})/h_{j-1} \le Mh_j^{-1}/(h_j+\varepsilon).$$

After replacing these estimates in (15) we obtain estimate (14) for $h_j \le \varepsilon$.

In the case $\varepsilon \le h_j$ we can take the form

$$(16)\qquad |z(x)| = |z_j^{(0)}| + |(x-x_j)z'(\xi_j)|,\ x \in [x_j, x_{j+1}],$$

$$x_j \le \xi_j \le x_{j+1}\ .$$

From (3) and (4) we obtain

$$(17)\qquad a_j v_j^{(1)} = b_j v_{j-1}^{(1)}\ ,\quad j = 1(1)n,$$

$$(18)\qquad v_j^{(2)} = (f_j - p_j v_j^{(1)})/\sigma_j\ ,\ j = 0(1)n+1,$$

$$(19)\qquad v_j^{(3)} = (v_{j+1}^{(2)} - v_j^{(2)})/h_j,\ j = 0(1)n.$$

Since $|z_o^{(1)}| \le Mh_o^2/(h_o+\varepsilon)$ from Lemma 1 we have

$|v_o^{(1)}| \le M$. From (17) and $|a_j| \le M$, $|b_j| \le M$ we have $|v_j^{(1)}| \le M$, $j=0(1)n$. Because of that from (18) and (19) we obtain $|v_j^{(2)}| \le M/h_j$, $|v_j^{(3)}| \le M/h_j^2$. Since $|y^{(1)}(x)| \le M$ and

$$v_j^{(1)}(x) = v_{j-1}^{(1)} + h_{j-1}v_{j-1}^{(2)} + \frac{h_{j-1}^{(3)}}{2}v_{j-1}^{(3)}\ ,\ j=1(1)n,\ \text{we have}$$

$$|v_j^{(1)}(x)| \leq M \quad \text{and} \quad |z_j^{(1)}(x)| \leq M.$$

Thus, according to (16), Theorem 2 holds.

REFERENCES

[1] A.E.Berger, J.M.Solomon, M.Ciment: An Analysis of a Uniformly Accurate Difference Method for a Singular Perturbation Problem, Math.Comput. 37(1981), 79-94.

[2] U.K.Emeljanov: O raznostnom metode rešenija tretej krajevoj zadači dlja differencijal´nogo uravnenija s malym parametrom pri staršej proizvodnoj. Žurn. vyčislit. mat. i mat.fiz. (1975),15 No. 6. 1457-1465.

[3] V.P.Il´in: O splajnovih rešenijah obyknovenyh differencijal´nyh uravnenij. žur. vyčislit, mat. i mat. fiz. (1978), 3, 621-627.

[4] K.Surla: Singularly perturbed spline collocation method for boundary value problems with mixed boundary conditions, Zb.Radova Prir.-Mat.Fak. u Novom Sadu, Ser. za Mat., 16,2(1980), 132-143.

[5] K.Surla and M.Stojanović: Singularly perturbed spline difference schemes on non-equdistant grid. Z. angew. Math. Mech.68 (1988) 3, 171-180..

[6] Ju.S.Zavijalov, B.I.Kvasov, Z.L.Mirošničenko: Metody splajn funkcii, Moskva 1980.

THE MEASURE OF APPROXIMATION FOR THE PARTICULAR SOLUTION

DJ. TAKAČI

ABSTRACT. *We observe the linear partial differential equation in the field of Mikusinski operators, F, with homogeneous conditions. For the approximate particular solution constructed in [4] we construct and estimate new measures of approximation both in a subspace F_0 of the field, F, as well as in the space L of local-integrable functions.*

1. INTRODUCTION

The nonhomogeneous differential equation with constant coefficients

$$(1)\qquad \sum_{\mu=0}^{m}\sum_{k=0}^{n} \alpha_{\mu,k} \frac{\partial^{\mu+k} x(\lambda,t)}{\partial\lambda^{\mu}\partial t^{k}} = f_1(\lambda,t); \qquad \begin{matrix} 0 \le \lambda \le \lambda_1 \\ 0 \le t \le \infty \end{matrix}$$

with conditions

$$(2)\qquad \frac{\partial^{\mu+k} x(\lambda,0)}{\partial\lambda^{\mu}\partial t^{k}} = 0 \quad \text{for} \quad \begin{matrix} \mu = 0,\dots m \\ k = 0,\dots n-1 \end{matrix}$$

$$(3)\qquad \frac{\partial^{\mu} x(0,t)}{\partial\lambda^{\mu}} = 0 \quad \text{for} \quad \mu = 0,\dots m-2 \text{ and}$$

$$\frac{\partial^{m-1} x(0,t)}{\partial\lambda^{m-1}} = \frac{t^{r-1}}{\Gamma(r)} \quad \text{for} \quad r > 0$$

($f_1(\lambda,t)$ is a continuous function) corresponds in the field F to the equation

$$(4)\qquad \sum_{\mu=0}^{m}\sum_{k=0}^{n} \alpha_{\mu,k} s^{k} x^{(\mu)}(\lambda) = f(\lambda)$$

where s is the differential operator, ℓ is the integral operator, $s = \ell^{-1}$, and $f(\lambda) = \{f_1(\lambda,t)\}$ with the conditions

$$(5)\qquad x^{(\mu)}(0) = 0 \quad \text{for } \mu = 0,\dots m-2 \text{ and } x^{(m-1)}(0) = \ell^{r}.$$

The particular solution of equation (4) can be written in the form (see [1])

$$(6)\qquad x_p(\lambda) = \frac{1}{\ell^r a_m} \int_0^\lambda f(\kappa)x_h(\lambda-\kappa)d\kappa \ ,$$

where

$$a_m = \sum_{k=0}^{n} \alpha_{m,k}s^k \quad \text{and}$$

$$x_h(\lambda) = \sum_{j=o}^{m} b_j \exp(\lambda\omega_j), \quad b_j \text{ are operators, and}$$

$$\omega_j = \sum_{i=0}^{\infty} c_{i,j}\ell^{i\alpha_j-\beta_j} \ , \quad \alpha_j > 0, \ \beta_j \le 1.$$

The approximate particular solution of equation (4) can be treated in the form (see [4])

$$(7)\qquad x_{p,n} = \frac{1}{\ell^r a_m} \int_0^\lambda f(\kappa)x_{h,n}(\lambda-\kappa)d\kappa,$$

where

$$(8)\qquad x_{h,n}(\lambda) = \sum_{j=0}^{m} b_j \exp(\lambda\omega_{j,n}) \quad \text{and}$$

$$(9)\qquad \omega_{j,n} = \sum_{i=0}^{n} c_{i,j}\ell^{i\alpha_j-\beta_j} \ ,$$

The convergence in the space of locally integrable functions L, is the convergence in all seminorms

$$(10)\qquad \|f\|_T = \int_0^T |f(t)|dt \ .$$

L_o is the subspace of L consisting of all functions f, such that $\|f\|_T > 0$, for every $T > 0$, and F_o is the algebra of all operators of the form f/g where $f \in L$ and $g \in L_o$.

The convergence type I′ in F_o is equivalent to the convergence defined by the functional $A(\cdot)$ (see [3])

$$(11)\qquad A(x) = \sum_{i=0}^{\infty} \frac{\beta_{i,1/i}(x)}{e^{ie^{i^2}(1+\beta_{i,1/i}(x))}} \ , \ x \in F_o \ ,$$

where

$$(12)\qquad \beta_{T,\varepsilon}(x) = \inf\{\|f\|_T : x = f/g, \ \|g\|_T < 1, \ \|\ell-\ell g\|_T < \varepsilon\}$$

was introduced by Burzyk ([1]).

Also, we need the following definitions.

DEFINITION 1 ([3]). Operator $\tilde{x} \in F_o$ is the approximation of the operator $x \in F_o$, according to the functional $A(x)$ with the measure of approximation $\delta > 0$ if $A(x-\tilde{x}) < \delta$.

DEFINITION 2 ([3]). The function $\tilde{f}$ from L is the approximation of the function f from L according to the functional

$$F(f) = \sum_{i=1}^{\infty} \frac{\|f\|_T}{e^{ie^{i^2}}(1+\|f\|_T)} \tag{13}$$

with the measure of approximation $\delta_L > 0$ if $F(f-\tilde{f}) < \delta_L$.

2. THE ESTIMATIONS IN THE SPACE F_o

Supposing that

$$\frac{I}{\ell^{r_{a_m}}} f(\lambda) = \{f_2(\lambda,t)\}, \quad gx_{h,n}(\lambda) \quad \text{and} \quad gx_h(\lambda)$$

for $g \in F_o$ represent functions from L, then the operators

$$z_n(\lambda) = \frac{\frac{I}{\ell^{r_{a_m}}} \int_o^{\lambda} f(\lambda) gx_{h,n}(\lambda-\kappa)d\kappa}{g} \tag{14}$$

$$z(\lambda) = \frac{\frac{I}{\ell^{r_{a_m}}} \int_o^{\lambda} f(\kappa) gx_h(\lambda-\kappa)d\kappa}{g} \tag{15}$$

belong to F_o.

Denoting by

$$\frac{\{J_g(\lambda,t)\}}{g} := \frac{\frac{I}{\ell^{r_{a_m}}} \int_o^{\lambda} f(\kappa) g(x_{h,n}(\lambda-\kappa)-x_h(\lambda-\kappa))d\kappa}{g}$$

and using relation (12), we can write

$$\beta_{T,\varepsilon}(y_n(\lambda) - y(\lambda) \le \|J_{g_1}(\lambda,t)\|_T$$

where $g_1 = \frac{\ell}{I+k} \cdot k$ satisfy for, $k > 0$, the inequalities

$$\|g_1\|_T < 1, \quad \|\ell-\ell g_1\|_T < \frac{1}{k} \quad .$$

Now, using paper [3] we obtain

$$\|Jg_1(\lambda,t)\|_T \leq \lambda M(\lambda,t) \cdot \sum_{j=1}^{m} k_{g_1}^{M}(\lambda,T,\alpha_j,\beta_j) \cdot \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j-\beta_j-1}{2}+1\right)}$$

where

$$M(\lambda,T) = \max_{0\leq\kappa\leq\lambda} \int_0^T |f_2(\kappa,t)|dt \quad \text{and} \tag{16}$$

$$k_{g_1}^{M}(\lambda,T,\alpha_j,\beta_j) = \max_{0\leq\kappa\leq\lambda} k_{g_1}((\lambda-\kappa),T,\alpha_j,\beta_j)$$

(the constants $k_{g_1}((\lambda-\kappa),T,\alpha_j,\beta_j)$ may be obtained analogous as in [3])

So, we can prove easily

LEMMA 1. The function $A(z_n(\lambda) - z(\lambda))$, where $z_n(\lambda)$ and $z(\lambda)$ are given by relations (14) and (15), can be estimated as:

$$A(z_n(\lambda)-z(\lambda)) \leq \sum_{j=1}^{m} \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j-\beta_j-1}{2}+1\right)} Q_{g_1}^{p}(\lambda,\alpha_j,\beta_j) = \delta \tag{17}$$

where

$$Q_{g_1}^{p}(\lambda,\alpha_j,\beta_j) \geq \sum_{i=1}^{\infty} \frac{\lambda M(\lambda,i) k_{g_1}^{M}(\lambda,i\alpha_j,\beta_j)}{e^{ie^{i^2}}}$$

So, we have

THEOREM 1. The sequence of approximate solutions $(x_{p,n}(\lambda))_n$ converges to the exact solution $x_p(\lambda)$ in the I' type convergence.

On using definition 1. we can say that the measure of approximation in F_o is given by (17).

3. THE ESTIMATION IN THE SPACE L

Let us suppose that the exact and the approximate particular solutions are the functions from L. Then analogously as in paper [3] we can obtain the estimation:

$$\|x_{p,n}(\lambda) - x_p(\lambda)\|_T \leq \lambda M(\lambda,T) \sum_{j=1}^{m} k^M(\lambda,T,\alpha_j,\beta_j) \cdot \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j - \beta_j - 1}{2} + 1\right)}$$

where $M(\lambda,T)$ is given by (16) and

$$k^M(\lambda,T,\alpha_j,\beta_j) = \max_{0 \leq \kappa \leq \lambda} k(\lambda-\kappa,T,\alpha_j,\beta_j) \quad \text{see [3]}$$

Now, we can prove

LEMMA 2. If $x_{p,n}(\lambda)$ and $x_p(\lambda)$ are given by relations (7) and (6) and represent functions from L, then we have:

$$(18)\quad F(x_{p,n}(\lambda) - x_p(\lambda)) \leq \sum_{j=1}^{m} \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j - \beta_j - 1}{2} + 1\right)} Q^p(\lambda,\alpha_j,\beta_j) = \delta_L$$

where

$$Q^p(\lambda,\alpha_j,\beta_j) \geq \sum_{i=1}^{\infty} \frac{M(i,\lambda) k^M(\lambda,i,\alpha_j,\beta_j)}{e^{ie^{i^2}}}$$

From previous Lemma follows

THEOREM 2. If $x_{p,n}(\lambda)$ and $x_p(\lambda)$ represent functions from L, then the sequence $(x_{p,n}(\lambda))_n$ converges to $x(\lambda)$ in L.

On using definition 2 the measure of approximation in L is given by (18).

It can be remarked that the measures of approximation obtained in this paper does not depend of the lenth of the interval.

EXAMPLE. The differential equation

$$\frac{\partial^2 x(\lambda,t)}{\partial\lambda\partial t} - \frac{\partial x(\lambda,t)}{\partial t} = x(\lambda,t) + f(\lambda), \quad 0 \leq \lambda \leq \lambda_o, \quad t \geq 0$$

with conditions

$$\frac{\partial x(\lambda,0)}{\partial\lambda} = 0, \ \lambda > 0, \quad x(0,t) = 1, \ t > 0$$

corresponds to the equation

$$(s-1)x'(\lambda) - x(\lambda) = f(\lambda)\ell \ ; \quad x(0) = \ell$$

in the field F

The exact particular solution is

$$x(\lambda) = \frac{1}{s-1} \int_0^{\lambda} f(\kappa)\exp((\lambda-\kappa)w)d\kappa$$

where

$$w = \sum_{i=0}^{\infty} \ell^{i+1}$$

and the approximate one is

$$x_n(\lambda) = \frac{1}{s-1} \int_0^{\lambda} f(\kappa)\exp((\lambda-\kappa)w_n)d\kappa$$

where

$$w_n = \sum_{i=0}^{n} \ell^{i+1}$$

However $x_n(\lambda)$ and $x(\lambda)$ belong to L we can find (for $\lambda = 1$ and $f(\lambda) = e^{\lambda}$)

$$\delta_L = \frac{1}{\Gamma(\frac{n+1}{2}+1)} \cdot e\left(\frac{e^{2e^2}}{e} + \frac{e}{e-1}\right) .$$

REFERENCES

1. BURZYK J.: On convergence in the Mikusinski operational calculus, Stud. Math. 75(1983) 313-333.
2. MIKUSINSKI J.: Operational calculus, Pergamon Press, Warsawa (1959).
3. PAP E., TAKAČI Dj.: "Estimations for the solutions of operator linear differential equations, Proc.GFCA-87.(in print).
4. TAKAČI Dj.: The approximate solution of a differential equation in many steps, Zb.radova PMF u Novom Sadu knjiga (1983) 51-61.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

EXPONENTIALLY FITTED QUADRATIC SPLINE DIFFERENCE SCHEMES

Z. UZELAC and K. SURLA

ABSTRACT: *For the problem (1.1) a family of difference schemes is derived using quadratic splines* $v(x) \in C^1[0,1]$. *The schemes are uniformly convergent with first order accuracy. Numerical examples are presented.*

1. DERIVATION OF THE SCHEMES

We consider collocation spline difference schemes for singularly perturbed two point boundary value problems

$$\begin{aligned} &Ly(x) \equiv \varepsilon y''(x) + p(x)y'(x) = f(x), \qquad 0<x<1, \\ &y(0) = \alpha, \quad y(1) = \beta, \end{aligned} \tag{1.1}$$

where ε is a small parameter in $(0,1]$, $p(x)$ and $f(x)$ are sufficiently smooth functions and $p(x) \geq p > 0$. Under these assumptions (1.1) has a unique solution $y(x)$ which in general displays a boundary layer at $x=0$ for small ε. The following lemma discribes some properties of the exact solution $y(x)$.

LEMMA 1.1 ([1]) *Let* $p(x)$, $f(x) \in C^3[0,1]$. *Then the solution of (1.1) can be written in the form* $y(x) = u(x) + W(x)$ *where*

$$u(x) = \varepsilon y'(0) \exp(-p(0)x/\varepsilon)/p(0) \tag{1.2}$$

$$|W^{(i)}(x)| \leq M(1+\varepsilon^{i-1}\exp(-\delta x/\varepsilon)), \qquad i=0,1,\ldots,4,$$

and M *and* δ *are constants independent of* ε.

Let us define the uniform mesh $\{x_j\}$, $j=0(1)n+1$ by $x_j=jh$ where n is an positive integer and the mesh length $h=1/(n+1)$. We will find an approximation to the solution $y(x)$ of (1.1) in the form of a quadratic spline $v(x) \in C^1[0,1]$ wich satisfies on the each interval $I_j=[x_j,x_{j+1}]$, $j=0(1)n$:

$$v_j(x) = v_j^{(0)} + (x-x_j)v_j^{(1)} + (x-x_j)^2 v_j^{(2)}/2.$$

The approximation to $y_j = y(x_j)$ is denoted by $v_j^{(0)} = v_j(x_j)$.

Let define the fitting "comparison" problem associated with (1.1) by:

$$
\begin{aligned}
&\tilde{L}\tilde{y}(x) \equiv \tilde{\sigma}(x,\varepsilon)\tilde{y}''(x) + \tilde{p}(x)\tilde{y}'(x) = \tilde{f}(x), \qquad 0<x<1 \\
&\tilde{y}(0) = \alpha, \quad \tilde{y}(1) = \beta,
\end{aligned} \tag{1.3}
$$

where $\tilde{\sigma}(x,\varepsilon)$, $\tilde{p}(x)$ and $\tilde{f}(x)$ are piecewise polynomial approximations to $\sigma(x,\varepsilon)$, $p(x)$ and $f(x)$ respectively (the fitting factor $\sigma(x,\varepsilon)$ will be determined). It is well-known that the solution $\tilde{y}(x)$ of the "comparison" problem ($\tilde{\sigma}(x,\varepsilon)=\varepsilon$) is a good approximation to the solution $y(x)$ of (1.1) (see Berger et al. [1]).

The unknown coefficients $v_j^{(k)}$, $k=0,1,2$, $j=0(1)n$ are determined from the conditions:

- $v(x)$ satisfies equation (1.3) at the points $x_{j+1/2} \equiv (x_j+x_{j+1})/2$, $j=0(1)n$ and the boundary conditions,
- $v(x) \in C^1[0,1]$.

The above conditions give the system of $3n+3$ unknowns with the same number of equations:

$$
\begin{aligned}
&\tilde{L}v_j(x)_{x=x_{j+1/2}} = \tilde{f}_j(x)_{x=x_{j+1/2}}, \qquad j=0(1)n, \\
&v_j(x)_{x=x_{j+1}} = v_{j+1}(x)_{x=x_{j+1}}, \qquad j=0(1)n-1, \\
&v_j'(x)_{x=x_{j+1}} = v_{j+1}'(x)_{x=x_{j+1}}, \qquad j=0(1)n-1, \\
&v_0(0) = \alpha, \quad v_n(1) = \beta.
\end{aligned} \tag{1.4}
$$

When $\tilde{\sigma}(x,\varepsilon)$, $\tilde{p}(x)$ and $\tilde{f}(x)$ are piecewise constant approximations of $\sigma(x,\varepsilon)$, $p(x)$ and $f(x)$ ($\tilde{p}_j(x)=\tilde{p}_j$, $x \in I_j$, etc.), the systeme (1.4) has the following form on the interval I_{j-1}:

$$
\begin{aligned}
&\tilde{\sigma}_{j-1} v_{j-1}^{(2)} + \tilde{p}_{j-1}\left(v_{j-1}^{(1)} + \frac{h}{2} v_{j-1}^{(2)}\right) = \tilde{f}_{j-1} \\
&v_{j-1}^{(0)} + h\, v_{j-1}^{(1)} + \frac{h^2}{2} v_{j-1}^{(2)} = v_j^{(0)} \\
&v_{j-1}^{(1)} + h\, v_{j-1}^{(2)} = v_j^{(1)}.
\end{aligned} \tag{1.5}
$$

By expessing $v_{j-1}^{(2)}$ from the first equation the system (1.5) has the following form:

$$
v_j^{(0)} = v_{j-1}^{(0)} + h\, v_{j-1}^{(1)} \tilde{\gamma}_{j-1} + \tilde{f}_{j-1} \tilde{S}_{j-1} \tag{1.6}
$$

$$
v_j^{(1)} = v_{j-1}^{(1)} \tilde{A}_{j-1} + h\, \tilde{f}_{j-1} / \tilde{s}_{j-1}, \tag{1.7}
$$

where

$$\tilde{s}_{j-1} = \tilde{\sigma}_{j-1} + h\,\tilde{p}_{j-1}/2, \quad , \quad \tilde{S}_{j-1} = h^2/(2\tilde{s}_{j-1})$$

$$\tilde{\gamma}_{j-1} = h(1-h\,\tilde{p}_{j-1}/(2\tilde{s}_{j-1})), \quad \tilde{A}_{j-1} = (1-h\,\tilde{p}_{j-1}/\tilde{s}_{j-1})$$

Similarly, we have that for $x \in I_j$:

$$(1.8) \qquad v_{j+1}^{(0)} = v_j^{(0)} + \tilde{\gamma}_j\, v_j^{(1)} + \tilde{S}_j \tilde{f}_j$$

$$(1.9) \qquad v_{j+1}^{(1)} = \tilde{A}_j\, v_j^{(1)} + h\tilde{f}_j/\tilde{s}_j.$$

From (1.6) we get:

$$v_{j-1}^{(1)} = (v_j^{(0)} - v_{j-1}^{(0)} - \tilde{S}_{j-1}\,\tilde{f}_{j-1})/\tilde{\gamma}_{j-1}$$

and from (1.8):

$$v_j^{(1)} = (v_{j+1}^{(0)} - v_j^{(0)} - \tilde{S}_j\tilde{f}_j)/\tilde{\gamma}_j.$$

By substituing the above expression for $v_{j-1}^{(1)}$ and $v_j^{(1)}$ into (1.7) we get a spline difference scheme which is a member of family of implicit schemes:

$$(1.10) \quad \frac{\tilde{A}_{j-1}}{\tilde{\gamma}_{j-1}} v_{j-1}^{(0)} - \left(\frac{\tilde{A}_{j-1}}{\tilde{\gamma}_{j-1}} + \frac{1}{\tilde{\gamma}_j}\right) v_j^{(0)} + \frac{1}{\tilde{\gamma}_j} v_{j+1}^{(0)} = \frac{\tilde{S}_j}{\tilde{\gamma}_j}\tilde{f}_j + \left(\frac{h}{\tilde{s}_{j-1}} - \frac{\tilde{A}_{j-1}\tilde{S}_{j-1}}{\tilde{\gamma}_{j-1}}\right)\tilde{f}_{j-1}$$

We introduce the following notation:

$$\tilde{r}_j^- = \tilde{A}_{j-1}/\tilde{\gamma}_{j-1}, \quad \tilde{r}_j^+ = 1/\tilde{\gamma}_j, \quad \tilde{r}_j^c = -\tilde{r}_j^- - \tilde{r}_j^+,$$

$$\tilde{q}_j^- = \frac{h}{\tilde{s}_{j-1}} - \frac{\tilde{A}_{j-1}\tilde{S}_{j-1}}{\tilde{\gamma}_{j-1}}, \quad \tilde{q}_j^c = \frac{\tilde{S}_j}{\tilde{\gamma}_j}, \quad \tilde{q}_j^+ = 0.$$

Then scheme (1.10) has the abbreviated form:

$$\tilde{R}v_j^{(0)} = \tilde{Q}\tilde{f}_j$$

where $\tilde{R}v_j^{(0)} = \tilde{r}_j^-\, v_{j-1}^{(0)} + \tilde{r}_j^c\, v_j^{(0)} + \tilde{r}_j^+\, v_{j+1}^{(0)}$

$$\tilde{Q}\tilde{f}_j = \tilde{q}_j^-\tilde{f}_{j-1} + \tilde{q}_j^c\tilde{f}_j + \tilde{q}_j^+\tilde{f}_{j+1}.$$

The truncation error for the boundary layer function u(x) (1.2) for

$p(x)=p=\text{const}$ is equal to zero when

$$(1.11)\qquad \tilde{r}_j^-/\tilde{r}_j^+ = \exp(-ph/\varepsilon).$$

If $\tilde{\sigma}(x,\varepsilon)=\sigma(\varepsilon)$ when $p(x)=p=\text{const}$ then condition (1.11) gives $\sigma(\varepsilon) = \frac{hp}{2}\,\mathrm{cth}(hp/(2\varepsilon))$. When $p(x)\neq\text{const}$ we define

$$\tilde{\sigma}_j(x,\varepsilon) = \frac{h\tilde{p}_j}{2}\,\tilde{\omega}_j, \quad x \in I_j \quad \text{where} \quad \tilde{\omega}_j = \mathrm{cth}(h\tilde{p}_j/(2\varepsilon)).$$

The coefficients of the family of the spline difference schemes defined by (1.10) have the following form

$$(1.12)\qquad \begin{aligned} &\tilde{r}_j^- = (1-1/\tilde{\omega}_{j-1})/h, \quad \tilde{r}_j^+ = (1+1/\tilde{\omega}_j)/h, \quad \tilde{r}_j^c = -\tilde{r}_j^+-\tilde{r}_j^-,\\ &\tilde{q}_j^- = 1/(\tilde{p}_{j-1}\tilde{\omega}_{j-1}), \quad \tilde{q}_j^c = 1/(\tilde{p}_j\tilde{\omega}_j), \quad \tilde{q}_j^+ = 0. \end{aligned}$$

The choice of approximation to $p(x)$ and $f(x)$ determines the particular scheme.

Let $\tilde{p}_{j-1} = \tilde{p}_j = p_j$, $\tilde{f}_{j-1} = \tilde{f}_j = f_j$

then the scheme (1.12) becomes $Rv_j^{(0)} = Qf_j$ where

$$r_j^- = (\omega_j-1)p_j/(2h), \quad r_j^+ = (\omega_j+1)p_j/(2h), \quad r_j^c = -\omega_j p_j/h,$$

$$q_j^- = q_j^+ = 0, \quad q_j^c = 1, \quad \omega_j = \mathrm{cth}(hp_j/(2\varepsilon)).$$

This scheme is precisely the Allen-Southwell-Il´in scheme for which first order uniform convergence at the nodes was proved in [3] and [4]. So, the quadratic spline difference scheme has the same property.

Choosing $\tilde{p}_{j-1} = (p_{j-1}+p_j)/2$, $\tilde{f}_{j-1} = (f_{j-1}+f_j)/2$,

$$\tilde{p}_j = (p_j+p_{j+1})/2, \qquad \tilde{f}_j = (f_j+f_{j+1})/2,$$

the corresponding implicit difference scheme has the coefficients:

$$(1.13)\qquad \begin{aligned} &r_j^- = (1-1/\tilde{\omega}_{j-1})/h, \quad r_j^+ = (1+1/\tilde{\omega}_j)/h, \quad -r_j^c = r_j^+ + r_j^-,\\ &q_j^- = 1/(2\tilde{p}_{j-1}\tilde{\omega}_{j-1}), \quad q_j^+ = 1/(2\tilde{p}_j\tilde{\omega}_j), \quad q_j^c = q_j^- + q_j^+. \end{aligned}$$

This scheme will be analysed in Section 2.

2. PROOF OF THE UNIFORM CONVERGENCE

The proof is based on the comparison functions method developed by Kellogg & Tsan [4] and Berger et al. [1] .

LEMMA 2.1 *Let* $\{V_j\}$ *be a set of values at the grid points* $\{x_j\}$, $j=0(1)n+1$ *satisfying* $V_0 \leq 0$, $V_{n+1} \leq 0$ *and* $RV_j \geq 0$, $j=1(2)n$. *Then* $V_j \leq 0$ *for* $j=0(1)n+1$.

We use two comparison functions $\phi_j=-2+x_j$ and $\psi_j=-\exp(-\beta x_j/\varepsilon)=-(\mu(\beta))^j$ where $\mu(\beta)=\exp(-\beta h/\varepsilon)$, $\beta>0$ will be chosen appropriately. Lemma 2.1 implies

LEMMA 2.2 *If* $K_1(h,\varepsilon) \geq 0$ *and* $K_2(h,\varepsilon) \geq 0$ *are functions that satisfy:*

$$R(K_1(h,\varepsilon)\phi_j + K_2(h,\varepsilon)\psi_j) \geq R(\pm Z_j) = \pm\tau_j(y)$$

where $Z_j = y_j - v_j^{(0)}$, *then*

$$|Z_j| \leq K_1(h,\varepsilon)|\phi_j| + K_2(h,\varepsilon)|\psi_j|.$$

Throughout the paper $\delta, M, M_1, \ldots$ will be used to denote generic constants independent of x, h and ε.

LEMMA 2.3 *There are constants* M_1 *and* M_2 *such that for* $h \leq M_1$, $0<\beta<M_2$ *and* $j=1(2)n$ *the following holds:*

(2.1) $\quad R\phi_j \geq Mh/\varepsilon , \qquad h \leq \varepsilon,$

(2.2) $\quad R\phi_j \geq M , \qquad \varepsilon \leq h,$

(2.3) $\quad R\psi_j \geq M\mu^j(\beta)h/\varepsilon^2, \qquad h \leq \varepsilon,$

(2.4) $\quad R\psi_j \geq M\mu^j(\beta)/h, \qquad \varepsilon \leq h.$

Proof. $R\phi_j=1/\tilde{\omega}_j+1/\tilde{\omega}_{j-1}$. Hence (2.1) and (2.2) holds. Now,

(2.5) $\quad R\psi_j = \mu^{j-1}(\beta)\, r_j^+ \,(1-\mu(\beta))(\mu\beta) - r_j^-/r_j^+).$

Let $h \leq C\varepsilon$ where C is a constant independent of h and ε, then $r_j^+ \geq M/h$, $r_j^-/r_j^+ \leq \exp(-\tilde{p}_j h/\varepsilon)+Mh$ and $\mu(\beta)-r_j^-/r_j^+ \geq M\mu(\beta)h/\varepsilon$, $1-\mu(\beta)=\beta h\exp(-Q\beta h/\varepsilon)/\varepsilon$, $0<Q<1$.

From (2.5) and the above estimates we see that (2.3) holds for $h \le C_1$.

Let $\varepsilon \le C^{-1}h$ for C sufficiently large. Then $r_j^+ \ge M/h$, $r_j^-/r_j^+ \le M\exp(-\tilde{p}_{j-1}h/\varepsilon)$, $\mu(\beta) - r_j^-/r_j^+ \ge M\mu(\beta)$, for appropriately chosen C and β. Moreover (2.4) holds for $\varepsilon \le C^{-1}h$. Since (2.3) holds for $h \le C\varepsilon$ we have $R\psi_j \ge M\mu^j(\beta)/\varepsilon \ge M\mu^j(\beta)/h$.

LEMMA 2.4 *The following estimates for the truncation error of the scheme (1.13) holds:*

$$|\tau_j(y)| \le M\left(\frac{h^2}{h+\varepsilon} \cdot \frac{h}{\varepsilon} + \frac{h^3}{\varepsilon^3} \cdot \exp\left(-\delta x_j/\varepsilon\right)\right), \quad j=1(2)n, \ h \le \varepsilon \tag{2.6}$$

$$|\tau_j(y)| \le M\left(h + \exp\left(-\delta x_{j-1}/\varepsilon\right)\right), \quad j=1(2)n, \ \varepsilon \le h. \tag{2.7}$$

Proof. Let $h \le \varepsilon$, then we take

$$\begin{aligned}\tau_j(y) = {} & T_0 y_j + T_1 y_j' + T_2 y_j'' + T_3 y_j''' + r_j^- R_3(x_j, x_j - h, y) + \\ & + r_j^+ R_3(x_j, x_j + h, y) - q_j^- \varepsilon R_1(x_j, x_j - h, y'') - \\ & - q_j^- p_{j-1} R_2(x_j, x_j - h, y') - \varepsilon q_j^+ R_1(x_j, x_j + h, y'') - \\ & - q_j^+ p_{j+1} R_2(x_j, x_j + h, y')\end{aligned}$$

where

$$Rn(a,b,g) = g^{(n+1)}(\xi)(b-a)^{(n+1)}/(n+1)! = \frac{1}{n!}\int_a^b (b-s)^n g^{(n+1)}(s)ds, \quad \xi \in (a,b).$$

Since $T_0 = T_1 = 0$ we will estimate T_2, T_3 and the remainder terms.

$$T_2 = h^2(r_j^+ + r_j^-)/2 + (p_{j-1}h - 2\varepsilon)q_j^- - (p_{j+1}h + 2\varepsilon)q_j^+.$$

Since $|hp\omega - 2\varepsilon| \le Mh^2/(\varepsilon+h)$ (see [4]) we have for

$p(x) = p = \text{const}$: $|T_2^c| \le M\frac{h^2}{h+\varepsilon} \cdot \frac{h}{\varepsilon}$.

Let $\rho_{j-1} = \tilde{p}_{j-1}h/(2\varepsilon)$, $r_j^- = r^-(\rho_{j-1})$, $r_j^+ = r^+(\rho_j)$, $q_j^- = q^-(\rho_{j-1})$, $q_j^+ = 1^+(\rho_j)$.

When $p(x) \ne \text{const}$ we expand T_2 at ρ_{j-1} and using the estimation for T_2^c we get $|T_2| \le M\frac{h^3}{\varepsilon(\varepsilon+h)}$.

Consider now

$$T_3 = h^3(r_j^+ - r_j^-) + (\varepsilon h - p_{j-1}h^2/2)q_j^- - (p_{j+1}h^2/2 + \varepsilon h)q_j^+.$$

By Taylor's expansion about ρ_{j-1} we get $|T_3| \le Mh^3/\varepsilon$.

The remainder terms are bounded by

$$Mh^3(1 + \exp(-2\delta x_j/\varepsilon)/\varepsilon^3).$$

Using Lemma 1.1 we have

$$|\tau_j(W)| \le M \frac{h^3}{(\varepsilon+h)\varepsilon}(1+\exp(-2\delta x_j/\varepsilon)/\varepsilon), \quad \text{for} \quad h \le \varepsilon. \tag{2.8}$$

Since $\tau_j(y)=\tau_j(u)+\tau_j(W)$ it remains to estimate $\tau_j(u)$. Let us denote by $\tilde{\tau}_j(u)$ the truncation error for $p(x)=p(0)=p$. As $\tilde{\tau}_j \equiv 0$ after some algebra we get

$$|T_2-\tilde{T}_2|u_j'' \le M(h^3/\varepsilon+h^3x_j/\varepsilon^2)u_j/\varepsilon^2,$$

and

$$|T_3-\tilde{T}_3|u_j''' \le M\, h^3x_j u_j/\varepsilon^4.$$

The remainder terms are bounded by $Mh^3\exp(-\delta x_j/\varepsilon)/\varepsilon^3$, thus

$$|\tau_j(u)| \le Mh^3x_j \exp(-\delta x_j/\varepsilon)/\varepsilon^3 \quad \text{for} \quad h \le \varepsilon. \tag{2.9}$$

From (2.8) and (2.9) we get (2.6).

Let $\varepsilon \le h$, then we consider the truncation error in the following form

$$\begin{aligned}\tau_j(y) = {} & T_2y_j'' + r_j^- R_2(x_j,x_{j-1},y) + r_j^+ R_2(x_j,x_{j+1},y) - \\ & - q_j^- \varepsilon\, R_0(x_j,x_{j-1},y'') - q_j^- p_{j-1} R_1(x_j,x_{j-1},y') - \\ & - q_j^+ \varepsilon\, R_0(x_j,x_{j+1},y'') - q_j^+ p_{j+1}\, R_1(x_k,x_{j+1},y').\end{aligned}$$

As before, we estimate $\tau_j(u)$ and $\tau_j(W)$ separately:

$$|\tau_j(W)| \le M(h+\exp(-\delta x_{j-1}/\varepsilon)), \quad \varepsilon \le h, \tag{2.10}$$

$$|\tau_j(u)| \le M \exp(-\delta x_{j-1}/\varepsilon), \quad \varepsilon \le h. \tag{2.11}$$

From (2.10), (2.11) we get that (2.7) holds.

THEOREM 2.1 *Let $p(x)$, $f(x) \in C^3[0,1]$ in (1.1). Let $\{v_j^{(0)}\}$, $j=0(1)n+1$ be the approximation to the solution $y(x)$ of (1.1) obtained using (1.13). Then, there exist constants M and δ independent of h and ε such that for $j=0(1)n+1$*

$$|v_j^{(0)}-y_j| \le M\left(\frac{h^2}{\varepsilon+h} + \frac{h^2}{\varepsilon}\exp(-\delta x_j/\varepsilon)\right), \quad h \le \varepsilon, \tag{2.12}$$

(2.13) $\quad |v_j^{(0)} - y_j| \le Mh(1+\exp(-\delta x_{j-1}/\varepsilon)), \quad \varepsilon \le h.$

Proof. From (2.8) and (2.9) we can see that the funcions $K_1(h,\varepsilon) = h^2/(h+\varepsilon)$ and $K_2(h,\varepsilon) = h^2/\varepsilon$ satisfy Lemma 2.2, and (2.12) hold. For $\varepsilon \le h$ we use $K_1(h,\varepsilon) = h$ and $K_2(h,\varepsilon) = h\exp(h\delta/\varepsilon)$. Using Lemma 2.2, (2.10) and (2.11) we get (2.13).

3. NUMERICAL RESULTS

We present the numerical results obtained by the scheme (1.13). We consider the following simple problems:

(3.1) $\quad \varepsilon y'' + y' = x, \quad y(0) = y(1) = 0$

which the solution is

$$y(x) = (\varepsilon - 1/2)(1-\exp(-x/\varepsilon))/(1-\exp(-1/\varepsilon)) - \varepsilon x + x^2/2,$$

and

(3.2) $\quad y'' + (1+x^2)y' = -(e^x + x^2), \quad y(0) = -1, \quad y(1) = 0.$

For each problem the mesh length $h = 1/J$ was succesively halved starting with $j = 16$ and ending with $J = 1024$. The maximum error et all the mesh points $E^\infty = \max_j |y_j - v_j^{(0)}|$ is listed in Table 2. under E^∞. The numerical rate of convergence is determined as in [2]:

$$\text{rate} \equiv (\ln Z_{K,\varepsilon} - \ln Z_{K+1,\varepsilon})/\ln 2,$$

where $\quad Z_{K,\varepsilon} = \max_j \left| v_j^{h/2^K} - v_j^{h/2^{K+1}} \right|, \quad K = 0(1)4$

and $v_j^{h/2^K}$ denotes the value of $v_j^{(0)}$ et the mesh point x_j for the mesh length $h/2^K$.

Table 1: Numerical rate of convergence for (1.13) applied to (3.2)

ε / K	1/2 rate	1/4 rate	1/8 rate	1/16 rate	1/32 rate	1/64 rate	1/128 rate	1/156 rate	1/512 rate
0	2.00	1.98	1.98	1.92	1.81	1.58	1.20	.98	.95
1	2.00	2.00	2.00	1.98	1.95	1.85	1.59	1.23	1.00
2	2.00	2.00	2.00	2.00	1.99	1.96	1.85	1.59	1.23
3	2.00	2.00	2.00	2.00	2.00	1.99	1.96	1.86	1.60
4	2.00	2.00	2.00	2.00	2.00	2.00	1.99	1.96	1.86

Table 2: Numerical results for (1.13) applied to (3.1)

K	ε / J	1/2 E∞ rate	1/4 E∞ rate	1/8 E∞ rate	1/16 E∞ rate	1/32 E∞ rate	1/64 E∞ rate	1/128 E∞ rate	1/156 E∞ rate	1/512 E∞ rate
	16	1.5 E-4	5.4 E-4	1.5 E-3	3.9 E-3	8.3 E-3	1.5 E-2	2.1 E-2	2.5 E-2	2.7 E-2
0		2.00	2.00	1.99	1.98	1.92	1.72	1.32	1.00	.95
	32	3.8 E-5	1.3 E-4	3.9 E-4	9.8 E-4	2.1 E-3	4.4 E-3	7.9 E-3	1.1 E-2	1.3 E-2
1		2.00	2.00	2.00	1.99	1.97	1.92	1.72	1.34	1.03
	64	9.6 E-6	3.3 E-5	1.0 E-4	2.4 E-4	5.5 E-4	1.1 E-3	2.3 E-3	4.0 E-3	5.7 E-3
2		2.00	2.00	2.00	2.00	1.99	1.98	1.91	1.71	1.34
	128	2.4 E-6	8.4 E-6	2.5 E-5	6.2 E-5	1.3 E-4	2.9 E-4	6.1 E-4	1.1 E-3	2.0 E-3
3		2.00	2.00	2.00	2.00	2.00	1.99	1.98	1.92	1.72
	256	6.0 E-7	2.1 E-6	6.2 E-6	1.5 E-5	3.5 E-5	7.4 E-5	1.5 E-4	3.1 E-4	6.0 E-4
4		2.00	2.00	2.00	2.00	2.00	2.00	1.99	1.98	1.92
	512	1.5 E-7	5.3 E-7	1.5 E-6	3.8 E-6	8.7 E-6	1.8 E-5	3.8 E-5	7.8 E-5	1.5 E-4
	1024	3.7 E-8	1.3 E-7	3.9 E-7	9.7 E-7	2.1 E-6	4.6 E-6	9.7 E-6	1.9 E-5	3.9 E-5

REFERENCES

1. A. E. Berger & J. M. Solomon & M. Ciment, *An Analysis of a Uniformly Accurate Difference Method for a Singular Perturbation Problem.* Math. Comput. 37 (1981), 79-94.

2. E. P. Doolan & J. J. Miller & W. H. A. Schilders, *Uniform Numerical Methods for Problems with Initial and Boundary Layers*. Dublin, Boole Press (1980)

3. A. M. Il´in, *Differencing scheme for a differential equation with a small parameter affecting the highest derivative*. Mat.Zametki 6 (1969) 237-248.

4. R. B. Kellogg & A. Tsan, *Analysis of some difference approximations for a singular perturbation problem without turning points*. Math.Comp. 32 (1978) 1025-1039.

ON A NUMERICAL SOLUTION OF A POWER LAYER PROBLEM

RELJA VULANOVIĆ

ABSTRACT: *A singularly perturbed boundary value problem, whose solution has a power boundary layer, is considered. A first order numerical method, uniform in the perturbation parameter, is constructed.*

1. INTRODUCTION

The following boundary value problem is considered :

$$Lu := -(\varepsilon+x)^2 u'' + c(x)u = f(x),\ x \in I := [0, 1], \tag{1a}$$

$$Bu := (u(0), u(1)) = (U_0, U_1), \tag{1b}$$

where ε is a perturbation parameter : $0 < \varepsilon \le 1$ (usually $\varepsilon << 1$). The functions c, f and numbers U_0, U_1 are given. We assume that

$$c, f \in C^1(I), \tag{2a}$$

$$c(x) \ge 0,\ x \in I, \tag{2b}$$

$$c(0) > 0. \tag{2c}$$

Under these assumptions we shall show that the unique solution u to the problem (1) has the form :

$$u(x) = s\,v(x) + z(x), \tag{3a}$$

where

$$s \in \mathbb{R},\ |s| \le M,\ v(x) = (1+x/\varepsilon)^{-r},\ r = (\sqrt{1+4c(0)} - 1)/2,$$

$$|z^{(i)}(x)| \le M(\varepsilon+x)^{1-i},\ i=1, 2, 3,\ x \in I. \tag{3b}$$

Here and troughout the paper M denotes any positive constant independent of ε. The function v is a power boundary layer function.

The asymptotic behaviour of the solutions to problems of the power layer type was considered in many papers by S. A. Lomov, see [3]. The numerical solution of power layer problems has not been investigated to the same extent as exponential layer problems. A numerical method for another power layer problem was given in [2].

Here we shall solve (1) numerically by using standard difference schemes on a special non-equidistant mesh which is dense in the layer. The mesh is generated by a suitable function. This approach has been applied successfully to various problems of exponential layer type, cf. [4], [5], for instance. Our main result is linear convergence uniform in ε.

2. ANALYSIS OF THE CONTINUOUS PROBLEM

After giving some lemmas we shall prove (3). For the technique cf. [1], [6].

LEMMA 1. *Let* (2) *hold. Then there exists a unique solution* $u \in C^3$ *to the problem* (1) *and it satisfies*

(4a) $$|u(x)| \le M, \quad x \in I,$$

(4b) $$|u'(0)| \le M/\varepsilon, \quad |u'(1)| \le M.$$

P r o o f : The operator (L, B) is inverse monotone and the uniqueness is guaranteed. The existence and uniform boundedness follow because there exist upper and lower solutions to (1), which are bounded uniformly in ε. Indeed, let $g(x) = M(2-x^2)$, where M is a constant (independent of ε) such that

$$g(t) \ge |U_t|, \quad t = 0, 1,$$

and

$$Lg(x) = 2M(\varepsilon+x)^2 + Mc(x)(2-x^2) \ge |f(x)|.$$

Such an M exists since

$$Lg(x) \ge \gamma M(2-\delta^2) \text{ if } 0 \le x \le \delta,$$

and

$$Lg(x) \ge 2M\delta^2 \text{ if } \delta \le x \le 1,$$

where δ is a number from $(0,1]$, such that

$$c(x) \geq \gamma > 0, \quad x \in [0,\delta],$$

(δ and γ are independent of ε). Then $g(x)$ is the upper solution and $-g(x)$ is the lower solution to the problem (1). Thus, (4a) is proved.

To prove (4b) use

$$u'(b) = u'(a) + \int_a^b (\varepsilon+x)^{-2}(cu-f)(x)\,dx$$

for some $a, b \in I$. Now using (4a) and the choice $b=0$ and $a \in (0,\varepsilon)$, such that $u'(a) = (u(\varepsilon)-u(0))/\varepsilon$, we get the first inequality in (4b). Similarly, with $b=1$ and $a \in (1/2,1)$, such that $u'(a) = 2(u(1)-u(1/2))$, the second inequality follows. □

We shall need estimates for the solution to the following auxiliary problem :

(5a) $$L_1 y := -((\varepsilon+x)^2 y')' + c(x)y = f(x), \quad x \in I,$$

(5b) $$By = (U_0, U_1),$$

LEMMA 2. *Let* (2) *hold and let* $y \in C^3(I)$ *be the solution to the problem* (5). *Then :*

(6) $$|y^i(x)| \leq M(\varepsilon+x)^{-i}, \quad i=0,1,2, \quad x \in I.$$

P r o o f : The case $i=0$ can be easily proved analogously to the proof of (4a). Furthermore, analogously to the proof of the first inequality in (4b) we can get $|y'(0)| \leq M/\varepsilon$. Then we have :

$$|y'(x)| = |\varepsilon^2(\varepsilon+x)^{-2}y'(0) + (\varepsilon+x)^{-2}\int_0^x (cy-f)(t)\,dt| \leq$$

$$\leq M(\varepsilon(\varepsilon+x)^{-2} + x(\varepsilon+x)^{-2}) \leq M/(\varepsilon+x),$$

hence (6) is proved for $i=1$. Then the estimate for $i=2$ follows directly from (5a). □

Now we can prove (3):

THEOREM 1. *Let* (2) *hold. Then the unique solution to the problem* (1) *satisfies* (3).

P r o o f : Let $s = u'(0)/v'(0) = -\varepsilon u'(0)/r$. Then because of (4) we have $|s| \le M$ and

$$z'(0) = 0, \ |z'(1)| \le M,$$
$$L_1 z'(x) = F(x), \ |F(x)| \le M, \ x \in I .$$

Indeed, $F = f' - c'u - s L_1 v'$, and

$$|L_1 v'(x)| = |r((r+1)(r+2) - 2(r+1) - c(x))(\varepsilon+x)^{-1}(1+x/\varepsilon)^{-r}| =$$
$$= |r(c(0) - c(x))(\varepsilon+x)^{-1}(1+x/\varepsilon)^{-r}| \le Mx(\varepsilon+x)^{-1} \le M.$$

Thus, z' satisfies a problem of type (5) and from Lemma 2 it follows

$$|z^{(i+1)}(x)| \le M(\varepsilon+x)^{-i}, \ i = 0, 1, 2. \ \square$$

3. THE DISCRETIZATION

The discretization mesh I_h has points:

(7a) $$x_i = \lambda(t_i), \ t_i = ih, \ i = 0, 1, \ldots, n, \ h = \frac{1}{n}, \ n \in \mathbb{N},$$

where λ is a mesh generating function, cf. [4], [5], of the form:

(7b) $$\lambda(t) = \begin{cases} \omega(t) := a\varepsilon((q/(q-t))^p - 1), & t \in [0, \alpha], \\ \pi(t) := \omega'(\alpha)(t-\alpha) + \omega(\alpha), & t \in [\alpha, 1]. \end{cases}$$

Here $\alpha \in (0,1)$ is given, $q = \alpha + \varepsilon^{1/(p+1)}$, $p \ge 1/r$ (r is given in (3)) and a is determined from the condition $\pi(1) = 1$. Hence, $\lambda \in C^1(I)$.

The properties of function λ are:

(8a) $$\lambda^{(i)}(t) \geq 0, \quad i=0, 1, 2, \quad t \in I,$$

(8b) $$\lambda'(t) \leq M, \quad t \in I,$$

(8c) $$\lambda(t) \geq M\varepsilon^{1/(p+1)}, \quad t \geq \alpha,$$

(8d) $$\lambda(t) \geq M\varepsilon n, \quad t \geq q - Mh > 0.$$

In this Section constants M will be independent of h as well.

Let $w_h = [w_0, w_1, \ldots, w_n]^T \in \mathbb{R}^{n+1}$ be a mesh function on I_h. Then the discrete problem corresponding to (1) reads:

$$w_0 = U_0,$$

(9) $$L^h w_i := -(\varepsilon + x_{i-1})^2 D_h^{\cdot\cdot} w_i + c(x_i) w_i = f(x_i), \quad i=1, 2, \ldots, n-1,$$

$$w_n = U_1,$$

where

$$D_h^{\cdot\cdot} w_i = 2\left(h_{i+1} w_{i-1} - (h_i + h_{i+1}) w_i + h_i w_{i+1}\right) / \left(h_i h_{i+1} (h_i + h_{i+1})\right),$$

$$h_i = x_i - x_{i-1}, \quad i=1, 2, \ldots, n.$$

Note the shift $\varepsilon + x_{i-1}$ (instead of $\varepsilon + x_i$) which is introduced for technical reasons, (see the proof of Theorem 2 below).

Rewrite (9) in the matrix form:

$$A_h w_h = d_h,$$

where $d_h = [U_0, f(x_1), f(x_2), \ldots, f(x_{n-1}), U_1]^T \in \mathbb{R}^{n+1}$ and $A_h \in \mathbb{R}^{n+1, n+1}$ is the corresponding tridiagonal matrix. Let $\|\cdot\|$ denote the maximum norm both in $\mathbb{R}^{n+1}$ and $\mathbb{R}^{n+1, n+1}$. Then we have

(10) $$\|A_h^{-1}\| \leq M,$$

provided that h is sufficiently small, but independent of ε. Indeed, A_h is an L-matrix and for $y_h = [\,2-x_0^2,\ 2-x_1^2,\ \ldots,\ 2-x_n^2\,]^T \in \mathbb{R}$ we have:

$$A_h y_h \geq M,$$

since

$$L^h(2-x_i^2) = 2(\varepsilon+x_{i-1})^2 + c(x_i)(2-x_i^2) \geq M$$

if h is sufficiently small (compare with the proof of (4a) in Lemma 1). Thus $A_h^{-1} \geq 0$ (to be understood componentwise) and the stability (10), uniform in ε, follows.

THEOREM 2. *Let* (2) *hold and let* u *be the solution to the problem* (1). *Let* w_h *be the solution to the discrete problem* (9) *on the mesh* (7) *with sufficiently small* h *independent of* ε. *Then:*

$$\|u_h - w_h\| \leq Mh,$$

where $u_h = [\,u(x_0),\ u(x_1),\ \ldots,\ u(x_n)\,]^T \in \mathbb{R}^{n+1}$.

P r o o f : Because of (10) it is sufficient to prove

(11a) $$|R_i(v)| \leq Mh\ ,\ i = 1,\ 2,\ \ldots\ ,\ n-1,$$

and

(11b) $$|R_i(z)| \leq Mh\ ,\ i = 1,\ 2,\ \ldots\ ,\ n-1,$$

where

$$R_i(g) := L^h g(x_i) - (Lg)(x_i) = -(\varepsilon+x_{i-1})^2 D_h^{\cdot\cdot} g(x_i) + (\varepsilon+x_i)^2 g''(x_i)$$

for any $g \in C^2(I)$. Let

$$R_i(g) = R_i^1(g) + R_i^2(g)\ ,$$

$$R_i^1(g) = \left((\varepsilon+x_i)^2 - (\varepsilon+x_{i-1})^2\right) g''(x_i),$$

$$R_i^2(g) = (\varepsilon+x_{i-1})^2 \left(g''(x_i) - D_h^{\cdot\cdot} g(x_i)\right).$$

In the next steps of the proof we shall use the Taylor expansion of R_i^2 , (3b) and (8).

First it is obvious that

$$|R_i^1(z)| \le Mh_i(x_i+x_{i-1}+2\varepsilon)/(\varepsilon+x_i) \le Mh$$

because (8b) implies $h_i \le Mh$. On the other hand

$$|R_i^2(z)| \le Mh(\varepsilon+x_{i-1})^2 \max_{x_{i-1}\le x\le x_{i+1}} |z^{(3)}(x)| \le Mh$$

and (11b) is proved.

Let us now prove (11a).

1^0 Let $t_{i-1} \ge \alpha$. By using (8 a, b, c) we get for k=1, 2

$$|R_i^k(v)| \le M h (\varepsilon+x_{i-1})^{-1}(\varepsilon/(\varepsilon+x_{i-1}))^r \le M h \varepsilon^r x_{i-1}^{-(r+1)} \le M h \varepsilon^{r-(r+1)/(p+1)} \le M h .$$

2^0 Let $t_{i-1} < \alpha$ and $t_{i-1} \le q-3h$. Then $t_{i+1} < q$ and $q-t_{i+1} \ge (q-t_i)/3$. Now for k=1,2:

$$|R_i^k(v)| \le M h \lambda'(t_{i+1})\varepsilon^r(\varepsilon+x_{i-1})^{-(r+1)} \le M h (q-t_{i+1})^{-(p+1)}(\varepsilon/(\varepsilon+\lambda(t_{i-1})))^{r+1} \le$$

$$\le M h (q-t_{i-1})^{p(r+1)-(p+1)} \le M h.$$

3^0 The remaining case is: $q-3h < t_{i-1} < \alpha$. Suppose that $q-3h > 0$. Then from (8d) it follows:

$$|R_i^k(v)| \le M (\varepsilon/(\varepsilon+x_{i-1}))^r \le M h^{pr} \le M h \ , \ k=1,2.$$

Hence (11a) is proved and so is the theorem. □

4. NUMERICAL RESULTS

We shall consider the following test problem :

$$-(\varepsilon+x)^2 u'' + u = x \ , \ u(0) = 1 \ , \ u(1) = 1 + (\varepsilon/(\varepsilon+1))^r, \ r = (\sqrt{5}-1)/2.$$

Its solution is given by

$$u(x) = (1 + x/\varepsilon)^{-r} + x.$$

Let $E = \| w_h - u_h \|$, using the notation of Theorem 2. We have the following tables :

Table 1. $p = 1/r$, $\alpha = 0.8$

E	ε				
	1. -3	**1. -6**	**1. -9 , 1. -12 , 1. -18**		
n = 20	.119	.149	.153		
n = 50	.0490	.0606	.0618		

Table 2. $p = 2/r$, $\alpha = 0.8$

E	ε				
	1. -3	**1. -6**	**1. -9**	**1. -12**	**1. -18**
n = 20	6.95-3	6.14-3	7.69-3	8.66-3	8.90-3
n = 50	1.13-3	1.09-3	1.45-3	1.65-3	1.69-3

The usual notation $1.-3 = 10^{-3}$ etc. is used.

Table 1 contains the results of the method described above. The linear convergence uniform in ε is obvious. However, for all problems whose solutions have form : $u(x) = Mv(x) + b(x)$, where $|b^{(i)}(x)| \leq M$, $i = 0, 1, 2, 3$, $x \in I$, and hence for this test problem, we can prove quadratic convergence uniform in ε if we take $p \geq 2/r$ and $(\varepsilon + x_i)^2$ instead of $(\varepsilon + x_{i-1})^2$ in (9). This is illustrated by the results in Table 2.

REFERENCES

[1] R.B.Kellogg, A.Tsan : Analysis of some difference approximations for a singular perturbation problem without turning points, Math.Comput. 32 (1978), 1025-1039.

[2] В.Д.Лисейкин : О численном решении уравнений со степенным погранслоем, Ж. вычисл. матем. и матем. физ. 26 (1986), 1813-1820.

[3] В. А. Треногин : Развитие и приложения асимптотического метода Люстерника-Вышика, Успехи матем. наук 25 (1970), 123-156.

[4] R.Vulanović : On a numerical solution of a type of singularly perturbed boundary value problem by using a special discretization mesh, Zb. Rad.Prir.-Mat.Fak.Univ.Novom Sadu, Ser.Mat. 13 (1983), 187-201.

[5] R.Vulanović : Mesh construstion for discretization of singularly perturbed boundary value problems, Ph. D. Thesis, University of Novi Sad, 1986.

[6] R.Vulanović : An exponentially fitted scheme on a non-uniform mesh, Zb. Rad.Prir.-Mat.Fak.Univ.Novom Sadu, Ser. Mat. 12 (1982), 205-215.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

A PROBLEM ON SIMULTANEOUS APPROXIMATION AND A CONJUCTURE OF HASSON

S. ZHOU

Let $C^N_{[-1,\ 1]}$ be the class of functions, which have N continuous derivatives, $P_n(f;\ x)$ be the polynomial of best approximation of degree $\leq n$ to $f \in C_{[-1,\ 1]}$ and $\Delta_n(x)=(1-x^2)^{1/2}/n+1/$

$E_n(f)= \|f - P_n(f)\| = \max\limits_{-1\leq x\leq 1} |f(x) - P_n(f,\ x)|$.

Both important and interesting question of approximation theory is: Do the derivatives of polynomials of best approximati achieve the best approximation to derivatives of function? In p riodic case, this problem had been solved long before. In algeb ic case, a classical result is that, if $f(x) \in C^N_{[-1,\ 1]}$, then there exists a $P(x) \in \Pi_n$ for $x \in (-1,\ 1)$ such that .

(1) $$|f^{(k)}(x) - P^{(k)}(x)| \leq C(N)\Delta_n^{N-k}(x)\,\omega(f^{(N)},\ \Delta_n(x)),$$

for $0\leq k\leq N$, $n\geq N$, where $\omega(f,\delta)$ is the modulus of continuity of Π_n is the set of algebraic polynomials of degree $\leq n$, $C(N)$ is a constant only depending upon N.

Considering the inequality (1), we notice that $P(x)$ is nc necessary to be the polynomial of best approximation to $f(x)$. T refore, it is natural to ask: What can one say about $P_n(f,\ x)$? M.Hasson [1] and D.Leviatan [2] have studied this problem rece tly. The result of Leviatan is that, if $f(x) \in C^N_{[-1,\ 1]}$, then

(2) $$|f_n^{(k)}(x) - P_n^{(k)}(f,x)| \leq \frac{C(N)}{n^k}\,(\Delta_n(x))^{-k}\,E_{n-k}(f^{(k)}),$$

where $x \in (-1,\ 1)$, $0\leq k\leq N$, $n\geq N$. The new question is: Can ineq lity (2) be improved? About this, M. Hasson [1] raised a conje

e as follows:

Let $N \geq 1$, then there exsists a function $f_0(x) \in C^{2N}_{[-1,\ 1]}$ for $+1 \leq k \leq 2N$ such that

$$P_n^{(k)}(f_0,\ 1) \nrightarrow f_0(1), \quad n \to \infty .$$

In this paper, we shall give a positive result to this conecture. This means, that the inequality (2) can not be improved.

Theorem. Let $N \geq 1$, then there exists a function $f_0(x) \in C^{2N}_{[-1,1]}$ or $N+1 \leq k \leq 2N$ such that

$$\overline{\lim_{n \to \infty}} \left| f_0^{(k)}(1) - P_n^{(k)}(f_0,\ 1) \right| > 0 .$$

Proof. Let n be an odd number,

$$T_n(x) = \cos(n \arccos x) = \frac{n}{2} \sum_{k=0}^{[\frac{n}{2}]} a_k\, x^{n-2k} \text{ , where } a_0 = 2^n/n ,$$

$$a_k = (-1)^k \frac{(n-k-1)!}{k!\,(n-2k)!} 2^{n-2k} \ ; \ S_{n+1}(x) = x^{2N} T_{n+1}^{(2N)}(\sqrt{1-x^2}) = \sum_{k=0}^{\frac{n+1}{2}} b_k\, x^{n-2k+1},$$

here $b_0 = (-1)^{\frac{n+1}{2}-N} 2^{n-1}(n+1)n \ldots (n-2N+2)$,

$$H_{n+2N+1}(x) = \int_0^x dx_1 \int_0^{x_1} dx_2 \ \ldots \int_0^{x_{2N-1}} S_{n+1}(x_{2N})\, dx_{2N} = \sum_{k=0}^{\frac{n+2N+1}{2}} C_k\, x^{n+2N-2k+1} ,$$

here $C_o = C_o(n) = (-1)^{\frac{n+1}{2}-N} 2^{n-1} \dfrac{(n+1)n \ldots (n-2N+2)}{(n+2)(n+3)\ldots(n+2N+1)}$.

bviously,

$$H_{n+2N+1}^{(2N)}(x) = S_{n+1}(x),$$

sing the known inequality of Bernstein type

$$\left| T_{n+1}^{(2N)}(\sqrt{1-x^2}) \right| \leq M_N\, n^{2N}\, \Delta_n^{-2N}(\sqrt{1-x^2}) \leq M_N\, x^{-2N}\, n^{2N} ,$$

hence

$$\| H_{n+2N+1}^{(2N)} \| \leq C_1(N)\, n^{2N} . \tag{3}$$

Notice that on $n+2N+2$ points $t_k = \cos(k\pi/(n+2N+1))$, $k=0,\ldots,n+2N+1$,

we have

$$T_{n+2N+1}(t_k) = (-1)^k \|T_{n+2N+1}\| ,$$

so that

$$(4) \quad H_{n+2N+1}(x) - P_{n+2N}(H_{n+2N+1}, x) = 2^{-2N-n} C_0(n) T_{n+2N+1}(x).$$

In view of the extreme properties of Chebyshev polynomials ([1])

$$(5) \quad \left| T_m^{(k)}(\pm 1) \right| = \|T_m^{(k)}(x)\| = \sup\{ \|f^{(k)}\| : f \in \Pi_m, \ \|f\| \le 1 \} = C_2(k) m^{2k}.$$

Take $\{n_j\}$ to be a sequence of odd numbers with

$$n_j^{2N+2}/n_{j+1} \longrightarrow 0, \quad j \longrightarrow \infty ,$$

for example, $n_{j+1} = (2j+1) n_j^{2N+1}$, and define

$$\frac{H_{n_\ell+2N+1}(x)}{n_\ell^{2N+2}} = h_\ell(x), \quad \sum_{\ell=j}^{\infty} h_\ell(x) = f_j(x), \quad j=0,1,\ldots .$$

Let $1 \le i \le N$. Due to (3), $f_0(x) \in C^{2N}_{[-1,1]}$ and for $0 \le k \le 2N$ from

$$\| H_{n+2N+1}^{(k)} \| \le \| S_{n+1} \| \le M_N n^{2N} ,$$

we got

$$(6) \qquad \|f_j^{(k)}(x)\| \le C_3(N) n_j^{-1} , \quad 0 \le k \le 2N.$$

From (4) ve get

$$h_j(x) - P_{n_j+2N}(h_j, x) = 2^{-2N-n_j} C_0(n_j) n_j^{-2N-2} T_{n_j+2N+1}(x) ,$$

and if combine it with (5) we obtain

$$(7) \qquad \left| h_j^{(N+i)}(1) - P_{n_j+2N}^{(N+i)}(h_j, 1) \right| \ge C(N) n_j^{2i-2} .$$

Further, from $f_0(x) - P_{n_j+2N}(f_0, x) = f_j(x) - P_{n_j+2N}(f_j, x)$ we can write

UNIVERSITY OF NIŠ
FACULTY OF ELECTRONIC ENGINEERING

NUMERICAL METHODS AND APPROXIMATION THEORY III

Niš, August, 18 - 21, 1987

Edited by G. V. Milovanović

N i š , 1988

Numerical Methods and Approximation Theory III

This publication was in part supported by The Regional Science Fondations of Niš.

Published by. Faculty of Electronic Engineering, University of Niš, P. O. Box 73, 18000 Niš, Yugoslavia

Technical support: Lj. M. Kocić and S. Zinovijev

Printed by: Prosveta, Niš

Numbers of copies: 500

PREFACE

The third conference on Numerical Methods and Approximation Theory was held in Niš at the Faculty of Electronic Engineering, University of Niš, August 18–21, 1987. It was attended by 140 participants from 20 countries. There were 85 papers presented in three sections.

Previous conferences were held in Niš (1984) and Novi Sad (1985) with 55 and 68 participants, respectively.

Two types of selected and refereed papers appear in this Proceedings: four long survey papers, based on 45–minute invited lectures, and 31 shorter research papers, presented at the thirty– and fifteen–minute talks. The papers were submitted in the prescribed form ready for copying. In both parts, Invited papers and Contributed papers, they are published in the alphabetic order of the surnames of the first authors.

I wish to thank the members of the Organizing Committee and all the referees for their voluntary work.

G. V. Milovanović

C O N T E N T S

LIST OF AUTHORS

ALIĆ, MLADEN
Dept. of Mathematics, Univ. of Zagreb, p.o.box 173, 41001 Zagreb

AŠIĆ, MIROSLAV D.
Dept. of Mathematics, Faculty of Natural Sciences and Mathematic
Studentski Trg 16, 11000 Belgrade, YU

ARAKELIAN, ARAM H.
Academy of Sciences of Aramenian SSR, P.Sevaka 1, 375044Yerevan,

BIDKHAM, MOHAMMAD
Dept. of Mathematics, Faculty of Natural Science and Technology,
mia Millia Islamia, 110025 New Delhi, INDIA

BOHTE, ZVONIMIR
Institute of Mathematics, Physics and Mechanics, Jadranska 19,
61000 Ljubljana, YU

BUTZER, PAUL L.
Lehrstuhl A für Mathematik, R.W.T.H., Templergraben 55, 5100 Aac
FRG

CRISCUOLO, GIULIANA
Istituto per Applicazioni della Matematica - C.N.R., via P. Cas-
llino 111, 80131 Napoli, ITALY

CVETKOVIĆ, LJILJANA
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad,

DANKOVIĆ, BRATISLAV
Dept. of Automatics, Faculty of Electronic Engineering, p.o.box
18000 Niš, YU

DA SILVA, MANUEL
Grupo de Matemática Aplicada, Faculdade de Ciências, Universida
do Porto, 4000 Porto, PORTUGAL

DELLA VECCHIA, BIANCAMARIA
Istituto per Applicazioni della Matematica - C.N.R., via P. Cas
llino 111, 80131 Napoli, ITALY

DELVOS, FRANZ-JÜRGEN
Lehrstuhl für Mathematik I, Univ. of Siegen, Hölderlin Str. 3,
5900 Siegen, FRG

DEWAN, KUM KUM
Dept. of Mathematics, Faculty of Natural Science and Technology
mia Millia Islamia, 110025 New Delhi, INDIA

DEY, CHARLIE
Charleston High School, Charleston, IL 61920, USA

, SUHRIT KUMMAR
Dept. of Mathematics, Eastern Illinois Univ.,Charleston,IL 61920,USA

DJURIĆ, MIRJANA
Fakultet zaštite na radu, Čarnojevićeva 10a, 18000 Niš, YU

FISCHER, HERBERT
Institut für Angewandte Mathematik und Statistik, Technische Universität München, Arcisstrasse 21, 8000 München 2, FRG

GATTESCHI, LUIGI
Dipartimento di Matematica dell'Univesità,Via Carlo Alberto 10,10123 Torino, ITALY

GAUTSCHI, WALTER
Dept. of Computer Sci.,Purdue Univ.,West Lafayette,IN 47907, USA

GHELARDONI, PAOLO
Istituto di Matematiche Applicate "U. Dini", Facoltà di Ingegneria, Università di Pisa, Via Bonanno 25B, 56100 Pisa, ITALY

GHERI, GIOVANNI
Istituto di Matematiche Applicate "U. DINI", Facoltà di Ingegneria, Università di Pisa, Via Bonanno 25B, 56100 Pisa, ITALY

GUESSAB, ALLAL
Département de Mathématiques,Ave. de l'Université,64000 Pau, FRANCE

HERCEG, DRAGOSLAV
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad, YU

HERZBERGER, JÜRGEN
Fachbereich Mathematik, Universität Oldenburg, 2900 Oldenburg, FRG

KOCIĆ, LJUBIŠA
Dept. of Mathematics, Faculty of Elactronic Engineering, p.o.box 73 18000 Niš, YU

KOVAČEVIĆ-VUJČIĆ, VERA V.
Faculty of Organizational Sciences, Jove Ilića 154,11040 Belgrade,

KOZAK, JERNEJ
Dept. of Mathematics and Mechanics, E.K. University of Ljubljana,Jaranska 19, 61111 Ljubljana, YU

LOKAR, MATIJA
Dept. of Mathematics and Mechanics, E.K. University of Ljubljana,Jaranska 19, 61111 Ljubljana, YU

MANGER, ROBERT
Rade Končar Institute, Baštijanova bb, 41000 Zagreb, YU

MARKOWICH, P. A.
Institut für Angewandte und Numerische Mathematik, Wiedner Hauptstr 8-10/115, 1040 Wien, AUSTRIA

MARZULLI, PIETRO
Istituto di Matematiche Applicate "U. DINI", Facoltà di Ingegneria, Università di Pisa, Via Bonanno 25B, 56100 Pisa, ITALY

MASTROIANNI, GIUSEPPE
Università degli Studi della Basilicata, Via N.Sauro, Potenza, ITA

MIJALKOVIĆ, SLOBODAN
Dept. of Microelectronics, Faculty of Electronic Engineering, p.o box 73, 18000 Niš, YU

MILOVANOVIĆ, GRADIMIR
Dept. of Mathematics, Faculty of Electronic Engineering, p.o.box 18000 Niš, YU

PETKOVIĆ, DEJAN
Fakultet zaštite na radu, Čarnojevićeva 10a, 18000 Niš, YU

PETKOVIĆ, MIODRAG
Dept. of Mathematics, Faculty of Electronic Engineering, p.o.box 18000 Niš, YU

POGÁNY, TIBOR
Tehnički fakultet Bor, JNA 12, 19210 Bor, YU

RASSIAS, THEMISTOCLES
Dept. of Mathematics, Univ. of LaVerne, p.o.box 51105, Kifissia, Athens, GRECE 145 10

SAVIĆ, VLADIMIR
PMF Kragujevac, R. Domanovića 12, 34000 Kragujevac, YU

SCHEMPP, WALTER
Lehrstuhl für Mathematik 1, Univ. of Siegen, 5900 Siegen, FRG

SCHMEISER, CHRISTIAN
Institut für Angewandte und Numerische Mathematik, Wiedner Haupt 8-10/115, 1040 Wien, AUSTRIA

SELBERHERR, S.
Institut für allgemeine Elektrotechnik, TU Wien, AUSTRIA

STANKOVIĆ, MIOMIR
Fakultet zaštite na radu, Čarnojevićeva 10a, 18000 Niš, YU

STEFANOVIĆ, LIDIJA
Dept. of Mathematics, Faculty of Electronic Engineering, p.o.bo: 18000 Niš, YU

STENS, R. L.
Lehrs. A für Mathematik, R.W.T.H.,Templergraben 55, 5100 Aachen

STOJADINOVIĆ, NINOSLAV
Dept. of Microelectronics, Faculty of Electronic Engineering, p box 73, 18000 Niš, YU

SURLA, KATARINA
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad,

TAKAČI, DJURDJICA
Institute of Mathematics, dr Ilije Djuričića 4, 21000 Novi Sad,

›resent instant. Therefore the question: is it possible to re-
:onstruct a bandlimited function (to begin with) from its
;amples taken exclusively from the past, i.e., taking into
ıccount only those f(t) for which $t < t_o$?

One answer to this question is the following: can one find
:oefficients $a_{kn} \in \mathbb{R}$ such that f can be reconstructed from its
;amples taken at the points t_o-T/W, t_o-2T/W, t_o-3T/W, ... from
:he past, in terms of

$$f(t_o) = \lim_{n\to\infty} \sum_{k=1}^{n} a_{kn}\, f(t_o - \tfrac{kT}{W}) \qquad (1.2)$$

:or each $t_o \in \mathbb{R}$? This would determine the value of f at the
›resent time instance $t = t_o$. It is the question of predicting
from its past samples.

There are two problems in this respect: (i) the role of T, naturally $T \in (0,1]$ - the closer T is to 1 the wider apart can the sampling points t_o-kT/W, $k \in \mathbb{N}$ be - and whether for each $T \in (0,1]$ the existence of the predictor coefficients is guaranteed, (ii) the evaluation of these coefficients, i.e., the construction of prediction formulae (1.2) in dependence on T - the closer T is to 1 the nearer is the sampling rate to that of the classical sampling theorem, namely the Nyquist rate 1/W.

Regarding the first problem, by applying a general result due to G. Szegö (1920) or a more general one due to N. Levinson (1940) one can show that for each T with $0 < T < 1$ there exist predictor coefficients a_{kn} such that (1.2) holds uniformly in $t_o \in \mathbb{R}$.

Regarding the second, Wainstein and Zubakow [25] (1962) showed that (1.2) is valid with $a_{kn} := (-1)^{k+1}\binom{n}{k}$ provided $0 < T < 1/3$; J.L. Brown Jr. [2] (1972) extended T to $T < 1/2$ for the coefficient choice $a_{kn} := (-1)^{k+1}\binom{n}{k}(\cos \pi T)^k$. This result was extended even further by W. Splettstoesser [22,23] (1981/82) who showed that (1.2) holds uniformly in $t_o \in \mathbb{R}$ for $a_{kn} := (-1)^{k+1}\binom{n+k-1}{k} 4^{-k}$ with $0 < T < \pi^{-1}$ arc cos $(-8^{-1}) \approx 0.5399$. Thus a sampling rate (even) larger than half the Nyquist rate

is possible in predicting bandlimited functions with coefficients a_{kn} that are even independent of T. Generally, the closer T is to 1, the more complicated will the coefficients a_{kn} (dependent on T) be.

The coefficients that are best, in the sense that the mean square error is minimized, are the solutions of the linear system

$$\sum_{k=1}^{n} a_{kn}\, si(\pi(k-j)TW) = si(\pi jTW) \qquad (1 \leq j \leq n) \tag{1.3}$$

where $si(x) = \sin x/x$. Since these are difficult to determine, and because they depend on n, the foregoing sub-optimal coefficients are more efficient.

Now it is known that a function being bandlimited is a rather restrictive condition. Such a function cannot be simultaneously duration limited, and it is the latter class of functions which actually occurs in practice. Further, beginning with bandlimited functions $f \in L^2(\mathbb{R})$, then f can be extended to the complex plane as an entire function (so one that is extremely smooth) that is of exponential type πW. The next question therefore is whether prediction can be carried out for functions that are <u>not</u> necessarily bandlimited. In this respect W. Splettstoesser [24] showed that if the (r+1)th derivative $f^{(r+1)} \in C(\mathbb{R})$ (= space of all uniformly continuous and bounded functions on $\mathbb{R}$), then

$$\sup_{t\in\mathbb{R}} \left| f(t) - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} (\cos \pi T)^k\, f\left(t - \frac{kT}{W}\right) \right| = O\left[(1+\cos \pi T)^n\, W^{-r-1} + (\sin \pi T)^n \sqrt{W}\right] \qquad (n, W \to \infty) \tag{1.4}$$

for each $0 < T < 1/2$. Since both terms on the right of (1.4) contain a factor tending to zero and one to infinity for $n, W \to \infty$, one has to choose n in dependence on W (or vice versa) such that both terms still tend to zero. It turns out that all the sample instants accumulate at t for $n, W \to \infty$. The details are to be found in [24].

The disadvantages in the prediction procedure described so far are (i) the sampling rates are just T/W with $0 < T << 1$ instead of the Nyquist rate $1/W$; (ii) the sample points in (1.2) depend on t, thus all the sample values have to be computed or measured anew when the series are to be evaluated for another t; (iii) in the case of prediction of not necessarily bandlimited functions generally the number of samples plus the distance between the sample points has to be regulated appropriately (recall (1.4)); (iv) to improve the approximation of f by the series in (1.2) or (1.4) the number n of samples has to be increased; (v) the sampling series (1.2) does not have the (classical) convolution structure for sums as given by the Shannon series (1.1).

To avoid these disadvantages, let us try to reconstruct functions from its past samples by the convolution series

$$(1.5) \qquad (S_W^{\varphi} f)(t) := \sum_{k=-\infty}^{\infty} f\left(\frac{k}{W}\right) \varphi(W(t - \frac{k}{W}))$$

for $W \to \infty$, where the kernel φ will be assumed to be continuous and have compact support contained in $[T_0, T_1]$ for some $0 < T_0 < T_1$. This means that $\varphi(Wt-k) \neq 0$ only for those $k \in \mathbb{Z}$ for which $k/W \in (t - T_1/W, t - T_0/W)$, so that only a finite number of samples taken from the past will be needed to evaluate (1.5), and this number will be fixed for all f, W and t. Increasing W in the series (1.5) will only mean that the distance between the sample points will decrease. Further, f need not necessarily be bandlimited. Of course, the coefficients $\varphi(Wt - k)$ depend on t, but the evaluation of φ should be simpler than that of the signal f to be sought.

It will be seen that our results enable one to predict or extrapolate the value of a signal even arbitrarily far ahead of the sample values.

The aim of this paper is to present a well-motivated overview of recent results obtained at Aachen in the matter. Most of the details, including the proofs of results stated, are to be

found in [7]. See also Chapter 5 of [6] which deals with prediction theory. Regarding the specific examples of Sections 2.3 and 4, they are treated here for the first time in actual detail.

For a continuation of the above approach of Splett-stoesser in the matter, see especially [23], [18], [19].

Connections of the present study with the basic work of A.N. Kolmogorov [12] (1941), N. Wiener [26] (1949) as well as of M.G. Krein [14] (1954) in the subject will be sketched in Section 6.

Concerning possible applications, one of the main ones is to speech processing, see e.g. [17], including differential pulse-code modulation [10]. Further applications are to economic prediction and forecasting, see e.g. [1], to geophysics and medicine, see e.g. [16].

2. PREDICTION OF DETERMINISTIC SIGNALS

2.1. GENERAL RESULTS

Let us now study sampling series of the form $(S_W^{\varphi} f)(t)$, defined in (1.5), where the si -function has been replaced by a kernel $\varphi \in C_{oo}(\mathbb{R})$ (=those $f \in C(\mathbb{R})$ that have compact support). Firstly, $S_W^{\varphi} f$ defines a family of bounded, linear operators from $C(\mathbb{R})$ into itself, with the operator norm

$$\| S_W^{\varphi} \|_{[C,C]} = m_o(\varphi) \qquad (W > 0),$$

$m_r(\varphi)$ denoting the absolute (sum) moment of φ of order $r \in \mathbb{N}_o$, namely

$$m_r(\varphi) := \sup_{t \in \mathbb{R}} \sum_{k=-\infty}^{\infty} |t-k|^r |\varphi(t-k)| .$$

Denote the Fourier transform of $g \in L^1(\mathbb{R})$ by

$$g^{\wedge}(v) = (1/\sqrt{2\pi}) \int_{-\infty}^{\infty} g(t)\ e^{-ivt} dt \qquad (v \in \mathbb{R}).$$

Proposition 1. Let $\varphi \in C_{OO}(\mathbb{R})$. The following three assertions are equivalent:

(i) $\lim_{W\to\infty} (S_W^{\varphi} f)(t) = f(t)$

for each $f \in C(\mathbb{R})$ and each $t \in \mathbb{R}$;

(ii) $\sum_{k=-\infty}^{\infty} \varphi(t-k) = 1 \qquad$ (each $t \in \mathbb{R}$);

(iii) $\varphi^{\wedge}(2k\pi) = \begin{cases} 1/\sqrt{2\pi}\ , & k = 0 \\ 0\ , & k \in \mathbb{Z} \setminus \{0\}\ . \end{cases}$

Proposition 2. Let $\varphi \in C_{OO}(\mathbb{R})$, $r \in \mathbb{N}$. If, in addition to the properties (i), (ii) or (iii) of Proposition 1, there holds

(ii)* $\sum_{k=-\infty}^{\infty} (t-k)^j \varphi(t-k) = 0 \qquad (j = 1,2,\ldots,r-1;\ t \in \mathbb{R})$

or, equivalently,

(iii)* $\varphi^{\wedge(j)}(2k\pi) = 0 \qquad (j = 1,2,\ldots,r-1;\ k \in \mathbb{Z})$

(for r=1 only one condition of Prop. 1 need hold), then there hold the estimates

(2.1) $$\|S_W^{\varphi} g - g\|_C \leq \frac{m_r(\varphi)}{r!} \|g^{(r)}\|_C\ W^{-r} \qquad (g \in C^r(\mathbb{R});\ W > 0)$$

$$\|S_W^{\varphi} f - f\|_C \leq K\omega_r(W^{-1}; f; C(\mathbb{R})) \qquad (f \in C(\mathbb{R});\ W > 0),$$

the constant K depending only on φ. In particular, if $f^{(r-1)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$, then

$$\|S_W^{\varphi} f - f\|_C = O(W^{-r+1-\alpha})\ , \quad W \to \infty\ .$$

Above, $\omega_r(\delta; f; C(\mathbb{R}))$ stands for the rth modulus of continuity of $f \in C(\mathbb{R})$, and $\mathrm{Lip}(\alpha; C(\mathbb{R}))$ for the Lipschitz class of order α. Regarding the foregoing propositions, see

e.g. Ries and Stens [21], [5]. Conditions of the type (ii)*, (iii)* were already used in connection with finite element approximation in Fix and Strang [9].

2.2. CONSTRUCTION OF KERNELS

Fejér's kernel F, defined by

$$F(t) := \frac{1}{2\pi}\left[\frac{\sin t/2}{t/2}\right]^2 , \quad F^{\wedge}(v) = \frac{1}{\sqrt{2\pi}} \begin{cases} 1-|v|, & |v| \leq 1 \\ 0, & |v| > 1 \end{cases}$$

satisfies property (ii)* for r=1. Likewise does de la Vallée Poussin's kernel. However, these kernels have unbounded support. The best examples of φ having compact support are the so-called central B-splines of order $r \geq 2$, defined by

$$M_r(t) := \frac{1}{(r-1)!} \sum_{k=0}^{r} (-1)^k \binom{r}{k} (t+\frac{r}{2}-k)_+^{r-1}$$

where $t_+^r = \max(t^r, 0)$, their Fourier transforms being simply

$$M_r^{\wedge}(v) = \frac{1}{\sqrt{2\pi}} \left(\frac{\sin v/2}{v/2}\right)^r \qquad (v \in \mathbb{R}).$$

The M_r are piecewise polynomials of degree r-1 having support [-r/2,r/2]. It is compact, but not contained in $(0,\infty)$, as required.

Let us now construct kernels without the latter deficiency for which Proposition 2 holds by taking appropriate linear combinations of translations of the M_r.

Proposition 3. For $\varepsilon_0 \in \mathbb{R}$ and $r \in \mathbb{N}$, $r \geq 2$, let $a_{\mu r}$, $\mu = 0,1,\ldots,r-1$ be the unique solutions of the linear system

$$(2.2) \qquad \sum_{\mu=0}^{r-1} a_{\mu r} \, (-i(\varepsilon_0+\mu))^j = (1/\sqrt{2\pi}\, M_r^{\wedge})^{(j)}(0) \quad (j=0,1,\ldots,r-1)$$

where $i = \sqrt{-1}$. Then

$$\varphi_r(t) := \sum_{\mu=0}^{r-1} a_{\mu r}\, M_r(t-\varepsilon_0-\mu) \qquad (t \in \mathbb{R})$$

is a polynomial spline of order r satisfying conditions (ii) and (ii)*, having support contained in $[T_0, T_1]$ with $T_0 = \varepsilon_0 - r/2$, $T_1 = \varepsilon_0 + 3r/2 - 1$.

Since $M_r^{\wedge}$ is even, the right side of (2.2) vanishes for j odd. So the solutions $a_{\mu r}$ are all real.

Corollary. In regard to $\varphi_r(t)$ there holds for $f^{(r-1)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$,

$$\|S_W^{\varphi_r} f - f\|_C = O(W^{-r+1-\alpha}) \,. \tag{2.3}$$

For a proof of Proposition 3 see Butzer and Stens [7]. In order to solve equation (2.2), one needs to know the derivatives $(1/M_r^{\wedge})^{(j)}(0)$, at least for small values of r. This can be achieved with the aid of the expansion

$$\left(\frac{v/2}{\sin v/2}\right)^r = \sum_{k=0}^{\infty} b_{kr}\, v^{2k} \qquad (|v| < 2\pi),$$

$$b_{kr} := (-1)^k \frac{(2k+r)!}{r!} \sum_{l=0}^{2k} (-1)^l \frac{r}{r+l} \cdot \frac{T(2k+l,l)}{(2k-k)!\,(2k+l)!},$$

where T(k,l) are the central factorial numbers of the second kind.

These derivatives can be taken from the following table which could readily be enlarged.

Table 1: $(1/\sqrt{2\pi}\, M_r^{\wedge})^{(j)}(0)$: r = 2,3,4,5; j = 0,1,2,3,4.

r \ j	0	1	2	3	4
2	1	0	-	-	-
3	1	0	1/4	-	-
4	1	0	1/3	0	-
5	1	0	5/12	0	9/16

2.3. SPECIFIC EXAMPLES

1. Take $r=2$, $\varepsilon_o=2$, so that $\varepsilon_o > r/2$ and $[T_o,T_1] = [1,4]$. The system (2.2) then reads, noting Table 1, $a_{02} + a_{12} = 1$, $a_{02}(-2i) + a_{12}(-3i) = 0$ for which $a_{02} = 3$, $a_{12} = -2$. Hence

$$\varphi_2(t) = 3M_2(t-2) - 2M_2(t-3) \ ;$$

the associated sampling series (1.5) involves only those samples at $k \in \mathbb{Z}$ for which $k/W \in (t-4/W, t-1/W)$. For example, if t would lie in the interval $(1/W, 2/W)$, the series consists of three terms only, namely for $k = 0,-1,-2$ for which $k/W < t-1/W < t$. If $f' \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, then by (2.3),

$$\| f(t) - \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W})\, \varphi_2(Wt-k) \|_C = \mathcal{O}(W^{-1-\alpha}) \ ,$$

enabling one to predict at least $1/W$ units ahead with error $\mathcal{O}(W^{-1-\alpha})$. If $f'' \in C(\mathbb{R})$ with $\|f''\|_C \leq M$, so that $\alpha=1$, then, according to (2.1), the large-$\mathcal{O}$ constant in (2.3) is $M\, m_2(\varphi_2)/2!$, which is bounded by $15M$ (a fact which cannot be derived theoretically but by employing a computer).

If one would take $r=2$ as above, but $\varepsilon_o > r/2$ arbitrary, then $[T_o,T_1] = [\varepsilon_o - 1, \varepsilon_o + 2]$, and

$$\varphi_{2,\varepsilon_o}(t) = (1+\varepsilon_o)\, M_2(t-\varepsilon_o) - \varepsilon_o M_2(t-\varepsilon_o-1) \ .$$

Here the samples are taken at $k \in \mathbb{Z}$: $k/W \in (t-(\varepsilon_o+2)/W, t-(\varepsilon_o-1)/W)$. In particular, if $\varepsilon_o=8$ and $t \in (2/W, 3/W)$, the series consists of three terms at $k = -5,-6,-7$, for which $k/W < t-7/W < t$. Whereas this is at least $7/W$ units to the left of t, the prediction instant, it was only $1/W$ units in the case of the kernel φ_2. Thus the kernel $\varphi_{2,\varepsilon_o}$ allows one to predict much further ahead with the same number of sampled values (the constant $m_2(\varphi_{2,\varepsilon_o})$ will, however, be much larger than 2·15). In fact, this procedure even enables one to predict or extrapolate a signal arbitrarily far ahead.

2. Now take $r=3$, $\varepsilon_o=2$, so that $[T_o,T_1]=[1/2,\ 11/2]$. The system (2.2) now reads

$$a_{03} + a_{13} + a_{23} = 1$$

$$-2i\,a_{03} - 3i\,a_{13} - 4i\,a_{23} = 0$$

$$4\,a_{03} + 9\,a_{13} + 16\,a_{23} = 1/4$$

which has as solutions $a_{03}=47/8$, $a_{13}=-62/8$, $a_{23}=23/8$. Whence

$$\varphi_3(t) = \frac{1}{8}\,[47M_3(t-2) - 62M_3(t-3) + 23M_3(t-4)]\ , \tag{2.4}$$

the sampling series now consisting of those $k\in\mathbb{Z}$ for which $k/W\in(t-11/2W, t-1/2W)$, thus of five terms for which $k/W<t-1/2W<t$.

In particular, if $\|f'''\|_C\leqslant M$, then $\|f-S_W^{\varphi_3}f\|_C\leqslant M\cdot 54W^{-3}$, noting that $m_3(\varphi_3)/3!\leqslant 54$.

3. Let us finally take $r=4$, $\varepsilon_o=3$, so that $[T_o,T_1]=[1,8]$. By solving a system of four equations in four unknowns one can readily show that

$$\varphi_4(t) = \frac{1}{6}\,[115M_4(t-3) - 256M_4(t-4) + 203M_4(t-5) - 56M_4(t-6)]\ .$$

This time the series consists of seven terms (at most), namely those $k\in\mathbb{Z}$ for which $t-10/W<k/W<t-1/W<t$. In particular, if $\|f^{(4)}\|_C\leqslant M$, then the corresponding rate of approximation can, in comparison with Example 1, be improved to $970\cdot M\,W^{-4}$. By enlarging the $\varepsilon_o\,(\geqslant 4)$ one could again achieve that, instead of being able to predict just (at least) $1/W$ units ahead (from k/W $(<t-1/W)$ to t), one could even predict $(\varepsilon_o-2)/W$ units ahead. Then of course the kernel $\varphi_4(t)$ would take on a different form.

In case $r=5$, $\varepsilon_o=3$ so that $[T_o,T_1]=[4/3,19/2]$, then

$$\varphi_5(t) = \frac{1}{1152}\{36767M_5(t-3) - 108188M_5(t-4)$$
$$+ 127914M_5(t-5) + 14927M_5(t-6)\} .$$

Here seven samples will be needed, the order of approximation being $O(W^{-5})$ provided $\|f^{(5)}\|_C \leq M$. The constant in the order is however large; in fact $m_5(\varphi_5)/5! \leq 3400$.

More generally, if $f^{(r)} \in C(\mathbb{R})$ with $\|f^{(r)}\|_C \leq M$, it is possible to construct a kernel $\varphi_r(t)$ such that the number of samples needed in the convolution sum is just 2r-1 and the associated order is $O(W^{-r})$. However, the constant will be correspondingly large. By this method one cannot increase the approximation order by taking more samples without increasing the order r of $\varphi_r(t)$.

Observe that it is an open question whether there exists a closed form of the solutions $a_{\mu r}$, $\mu = 0,1,\ldots,r-1$ of (2.2). So far the construction can be used in actual practice only for smaller values of r. However, as already the simplest Example 1 shows, even the case r=2 gives the pretty good rate $15\,MW^{-2}$, $W \to \infty$, if $\|f''\|_C \leq M$.

3. TIME-JITTER AND AMPLITUDE ERRORS

It is especially easy to treat time-jitter errors in this frame. These arise when the sample instants are not correctly met but might differ from the exact k/W by δ_k, so that the sampled values are now $f(k/W + \delta_k)$. Here one is interested in estimating the error occurring when f(t) is approximated by the series $S^{\varphi}_{W,\delta}f(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{W} + \delta_k)\varphi(Wt-k)$. This error can be split up as

$$|f(t) - S^{\varphi}_{W,\delta}f(t)| \leq |f(t) - S^{\varphi}_{W}f(t)| + (J_\delta f)(t) ,$$

$$(J_\delta f)(t) := \Big|\sum_{k=-\infty}^{\infty} [f(\tfrac{k}{W}) - f(\tfrac{k}{W} + \delta_k)]\,\varphi(Wt-k)\Big|$$

being the so-called total time-jitter error. It can be esti-

mated in terms of the modulus of continuity, assuming $|\delta_k| \leq \delta$, $k \in \mathbb{Z}$, by

$$|(J_\delta f)(t)| \leq \{\sup_{k\in\mathbb{Z}} \|f(\cdot) - f(\cdot+\delta_k)\|_C\} \{\sup_{t\in\mathbb{R}} \sum_{k=-\infty}^{\infty} |\varphi(t-k)|\}$$

$$\leq m_o(\varphi)\cdot\omega_1(\delta;f;C(\mathbb{R})) \qquad (t \in \mathbb{R}).$$

As a consequence we have

Proposition 4. There hold

a) $$\|f(\cdot) - \sum_{k=-\infty}^{\infty} f(\frac{k}{W} + \delta_k)\varphi(W\cdot - k)\|_C$$

$$\leq \|f - S_W^\varphi f\|_C + m_o(\varphi)\cdot\omega_1(\delta;f;C(\mathbb{R})) \qquad (f \in C(\mathbb{R})).$$

b) If $f \in \mathrm{Lip}(\alpha;C(\mathbb{R}))$, $0 < \alpha \leq 1$, then, provided $\delta \leq 1/W$, $W \geq 1$, the order in part a) is given by $O(W^{-\alpha})$.

Note that this order cannot be improved even if f possesses derivatives of arbitrary order. On the other hand, if $W^{-1} \leq \delta$, then the order in part a) is $O(\delta^\alpha)$.

Thus the prediction series $S_W^\varphi f(t)$ exemplifies stability with respect to the sample points, a small error in each of the sample points produces a correspondingly small error in the prediction series.

There is also the amplitude error $(A_\varepsilon f)(t)$, arising if the exact sample values $f(k/W)$ are not at one's disposal but only falsified values $\overline{f}(k/W)$, differing by $\varepsilon_k := f(k/W) - \overline{f}(k/W)$ with $|\varepsilon_k| \leq \varepsilon$, $k \in \mathbb{Z}$, for some $\varepsilon > 0$. This falsification may be due to rounding-off, quantization or noise. The total amplitude error

$$|(A_\varepsilon f)(t)| := |(S_W^\varphi f)(t) - (S_W^\varphi \overline{f})(t)| \leq \varepsilon\, m_o(\varphi) ,$$

so that the error occurring when $f(t)$ is approximated by $S_W^\varphi \overline{f}(t)$ can be estimated by

Proposition 5. There hold

a) $$\| \sum_{k=-\infty}^{\infty} \overline{f}(\tfrac{k}{W})\, \varphi(W\cdot - k) - f(\cdot) \|_C \leq \| S_W^{\varphi} f - f \|_C + \varepsilon\, m_o(\varphi) \ .$$

b) If $f \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$, then the order in part a) is $O(W^{-\alpha})$ provided $\varepsilon \leq W^{-1}$, $W \geq 1$.

Thus the prediction series also illustrates stability with respect to the function values, a uniformly small change in the function values at all of the sample points produces a correspondingly small change in the prediction series.

4. PREDICTION OF DERIVATIVES $f^{(s)}$ BY SAMPLES OF f

Let us now consider the prediction of derivatives $f^{(s)}$ of a signal f by samples of f only, in terms of derivatives of $S_W^{\varphi} f$, i.e., of

$$(S_W^{\varphi})^{(s)} f(t) = (\tfrac{d}{dt})^s (S_W^{\varphi} f)(t) = W^s \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W}) \varphi^{(s)}(Wt-k)$$

$$(s \in \mathbb{N}_o) \ .$$

Proposition 6. Let $\varphi \in C_{OO}^{(s)}(\mathbb{R})$ satisfy (ii), (ii)* for some $r \geq s+1$ with $s \in \mathbb{N}_o$, $r \in \mathbb{N}$. Then $(S_W^{\varphi})^{(s)} f$ defines a family of bounded, linear operators mapping $C^{(s)}(\mathbb{R})$ into $C(\mathbb{R})$, with norm

$$\| S_W^{\varphi(s)} \|_{[C^{(s)}, C]} \leq \frac{m_s(\varphi^{(s)}_{\cdot})}{s!} \qquad (W > 0) \ .$$

Further,

$$\| (S_W^{\varphi})^{(s)} g - g^{(s)} \|_C \leq \frac{m_r(\varphi^{(s)})}{r!} \| g^{(r)} \|_C \ W^{-r+s}$$

$$(g \in C^{(r)}(\mathbb{R});\ W > 0) \ ,$$

$$\| (S_W^{\varphi})^{(s)} f - f^{(s)} \|_C \leq K \omega_{r-s}(W^{-1}; f^{(s)}; C(\mathbb{R}))$$

$$(f \in C^{(s)}(\mathbb{R});\ W > 0).$$

In particular, one has for $f \in C^{(s)}(\mathbb{R})$,

$$\lim_{W \to \infty} \left(\frac{d}{dt}\right)^s (S_W f)(t) = \left(\frac{d}{dt}\right)^s f(t)$$

uniformly in $t \in \mathbb{R}$; if $f^{(r-1)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, then $\| (S_W^{\varphi})^{(s)} f - f^{(s)} \|_C = O(W^{-r+1+s-\alpha})$, $W \to \infty$.

These results would enable one to predict the speed or acceleration of flying objects.

Let us consider an example. For this purpose we begin with example 2 of Section 2.3 where $r=3$, $\varepsilon_o = 2$, $[T_0, T_1] = [1/2, 5\ 1/2]$ and $\varphi_3(t)$ is given by (2.4). Let us apply Proposition 6 to $\varphi_3(t)$ in the case $s=1$. Noting that

$$M_r'(t) = M_{r-1}(t + 1/2) - M_{r-1}(t - 1/2) \qquad (t \in \mathbb{R}),$$

$$\varphi_3'(t) = \frac{1}{8}[47\, M_2(t - 3/2) - 109\, M_2(t - 5/2) + 85\, M_2(t - 7/2)$$

$$- 23\, M_2(t - 9/2)].$$

Here $\varphi_3' \in C_{OO}(\mathbb{R})$. In particular, if $f^{(2)} \in \mathrm{Lip}(\alpha; C(\mathbb{R}))$, $0 < \alpha \leq 1$, then

$$\| W \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W}) \varphi_3'(Wt - k) - f'(t) \|_C = O(W^{-1-\alpha}) .$$

This result enables one to predict the derivative $f'(t)$ in terms of a series which involves just five samples of f which all lie to the left of $t - 1/2W < t$. If $\| f^{(3)} \|_C \leq M$, then the large-O constant is given by $M\, m_3(\varphi_3')/3!$.

5. PREDICTION OF RANDOM SIGNALS

Signal functions are often of random character, random signals play an important role in signal processing and sampling prediction. For this purpose one often uses stochastic processes which are stationary in the weak sense as a model. Given a probability space (Ω,A,P), a real-valued stochastic (random) process, namely an A-measurable function $X = X(t) = X(t,\omega)$ of $\omega \in \Omega$ for each $t \in \mathbb{R}$, is said to be weak sense stationary (w.s.s.), if its autocorrelation function

$$R_X(t,t+\tau) := \int_\Omega X(t,\omega)X(t+\tau,\omega)dP(\omega)$$

is independent of $t \in \mathbb{R}$, i.e., $R_X(t,t+\tau) = R_X(\tau)$. Here X is assumed to belong to $L^2(\Omega)$, i.e., the norm

$$\|X(t,\cdot)\|_2 := \{\int_\Omega |X(t,\omega)|^2 dP(\omega)\}^{1/2} := \{E[|X(t)|^2]\}^{1/2} \tag{5.1}$$

is finite for all $t \in \mathbb{R}$. Note that $R_X(\tau)$ is even in τ, $\|R_X\|_C = R_X(0)$, and the norm (5.1) is independent of t, equalling $\|R_X\|_C^{1/2}$.

For the prediction of such a process $X \in L^2(\Omega)$ let us consider the prediction series

$$(S_W^\varphi X)(t,\omega) := \sum_{k=-\infty}^{\infty} X(\frac{k}{W},\omega)\varphi(Wt-k) \qquad (t \in \mathbb{R}).$$

It defines a family of bounded, linear operators from $L^2(\Omega)$ into itself, with

$$\|S_W^\varphi X(t,\cdot)\|_2 = \{\sum_{k,\mu=-\infty}^{\infty} R_X(\frac{k-\mu}{W})\varphi(Wt-k)\varphi(Wt-\mu)\}^{1/2}$$

$$\leq R_X(0)^{1/2} m_0(\varphi) = m_0(\varphi)\|X\|_2 .$$

Proposition 7. Let $\varphi \in C_{OO}(\mathbb{R})$ satisfy (ii), (ii)* with r-1 replaced by 2(r-1) for some $r \in \mathbb{N}$. If X is a w.s.s. process with $X^{(r)} \in L^2(\Omega)$, then

$$\{E[|S_W X - X|^2]\}^{1/2} \leq \left\{\frac{(m_0(\varphi)+3)m_{2r}(\varphi)}{2r!}\right\}^{1/2} \cdot$$

$$\cdot \frac{\{E[|X^{(r)}|^2]\}^{1/2}}{W^r} \quad (t \in \mathbb{R};\ W > 0).$$

There exists a constant $K > 0$ such that for any w.s.s. process $X \in L^2(\Omega)$, continuous in the mean,

$$\{E[|S_W X - X|^2]\}^{1/2} \leq K\omega_r(W^{-1}; X; L^2(\Omega)) \quad (t \in \mathbb{R};\ W > 0).$$

Regarding proofs in the case of random processes, one reduces the matter to the deterministic case, namely from assertions dealing with the random process X to those concerned with the deterministic function R_X, by the following basic connections:

i) the rth derivative (in mean) $X^{(r)}$ exists at $t_0 \in \mathbb{R}$ if $R_X \in C^{(2r)}(\mathbb{R})$;

ii) $$\omega_s(\delta; X; L^2(\Omega)) = \{\omega_{2s}(\delta; R_X; C(\mathbb{R}))\}^{1/2};$$

iii) $$E[|S_W^{\varphi} X - X|^2] = R_X(0) - 2\sum_{k=-\infty}^{\infty} R_X(\frac{k}{W} - t)\cdot\varphi(Wt-k) +$$

$$+ \frac{1}{2\pi}\sum_{k,\mu=-\infty}^{\infty} R_X(\frac{k-\mu}{W})\varphi(Wt-k)\cdot\varphi(Wt-\mu)$$

$$\leq (m_0(\varphi)+3)\sup_{u\in\mathbb{R}} |(S_W^{\varphi}\tau_u R_X)(t) - (\tau_u R_X)(t)$$

where $(\tau_u f)(t) = f(t-u)$.

The following table gives the best possible order of approximation according to Proposition 7 for the kernels φ_r of Section 2.3.

Table 2

Kernels	φ_2	φ_3	φ_4
Orders	$O(W^{-1})$	$O(W^{-1})$	$O(W^{-2})$

6. THE APPROACHES OF WIENER AND KREIN IN COMPARISON

Let us finally roughly compare the present approach with the work of Wiener [26] and M.G. Krein [14] (1954) in the matter. For this purpose let us express our convolution sum (1.5), thinking of the commutativity of convolution products, as

$$(6.1) \qquad \sum_{k=-\infty}^{\infty} f(\tfrac{k}{W})\varphi(W(t-\tfrac{k}{W})) \cong \sum_{k=-\infty}^{\infty} f(t-\tfrac{k}{W})\varphi(k) \ .$$

Although the two sums are generally not equal (except under special conditions, see [6]), it is nevertheless also possible to set up our approach to prediction for the right hand one (using parallel arguments, see [7]). If φ has compact support in $[T_0,T_1]$, then the right sum only runs over all k with $T_0 < k < T_1$ so that one can see from it right off where the prediction points lie, namely to the left of t at $\ldots\ t-k/W,\ \ldots,\ t-1/W$.

Now Wiener's aim was to predict the future at time t from the <u>whole</u> past $f(u)$: $-\infty \leqslant u \leqslant t-\varepsilon_0$, $\varepsilon_0 > 0$ prescribed, in a non-discrete setting (where our sum is replaced by an integral). In fact, his aim was to minimize as a function of the kernel Φ the mean-square error

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \left| f(t) - \int_0^{\infty} f(t-\varepsilon_0-u)\Phi(u)du \right|^2 dt \ .$$

He showed that his problem amounts to solving the integral equation

$$R(t) = \int_0^\infty R(t-\varepsilon_o-u)\Phi(u)du \qquad (t \geq \varepsilon_o), \tag{6.2}$$

where R is the auto-correlation function,

$$R(t) := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t+u)f(u)du \ .$$

Now it is to be emphasized that the equation (6.2) only holds for $t \geq \varepsilon_o$, and not for all $t \in \mathbb{R}$. So it is not solvable by routine Fourier methods. The so-called Wiener-Hopf technique (of 1931) has to be employed. In this respect Wiener notes [26; p.65] that there are "limitations and precautions which must be observed" in solving (6.2) and illustrated his method by several examples. In fact, Dym and McKean add [8 , p.92] "it is not clear how to proceed much further in the present direction save by examples". In any case, for a formal derivation as well as exellent coverage of the matter see the treatment in [8] pp. ix, 2-5, 82-96. For good information concerning effective computation see Lee [15], pp. 354-439, Kailath [11], also Noble [20]. For further literature see the extensive reference lists in the commentaries on the work of Wiener by P. Masani, H. Salehi, T. Kailath, P.S. Muhly and G. Kallianpur in [27].

Now the problem treated in this paper is actually that of predicting the future from only a <u>part</u> $f(u) : -T \leq u \leq t-\varepsilon_o$ of the past, in the case of discrete u. Especially in the non-discrete case was this problem solved by Krein [14]; it required even much heavier machinery than that of the (Kolmogorov)-Wiener problem, namely a so-called "method of strings" in the context of operator theory, complex function theory and Hardy functions, wave and spectral functions,all combined with the theory of spaces of entire functions (in the sense of de Branges [3]). This theory was carried out in expert fashion by Dym-McKean [8] pp. 146-278, applied to the actual prediction problem on pp. 279-91; there is an overview on pp. 5-9. However, as these authors write (p. X): "it is hoped that electrical engineers and other people dealing with the practical aspects of prediction will find in it

[our volume] something to interest them too, though it has to be confessed that the computations to which the theory leads are usually difficult to perform and that their statistical content is often obscure; in fact, much remains to be done to clarify the statistical content of the whole subject."

The methods needed to prove the results of this overview, presented in [7] are, in comparison, elementary indeed. Thus Proposition 1 is based upon a simple application of the Poisson summation formula of Fourier analysis, Proposition 2 upon Taylor's formula and elementary approximation theory, while Proposition 3 uses elementary results on B-splines (together with some new results on central factorial numbers). Proposition 7 shows that the treatment of random prediction theory can essentially be reduced to that of the deterministic situation so that no separate approach is necessary.

Most of the results discussed in this overview arose from questions posed by electrical and communication engineers in the course of some seven years of cooperative work. It is to be expected that they can also follow the proofs. The fact that the matter is indeed easy to apply has been demonstrated with the various examples.

LITERATURE

1. G.E.T. BOX and G.M. JENKINS: Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, CA, 1976 (rev. edition).

2. J.L. BROWN: Uniform prediction of bandlimited processes from past samples. IEEE Trans. Inform. Theory IT-18 (1972), 662-664.

3. L. DE BRANGES: Hilbert Spaces of Entire Functions. Prentice Hall, Englewood Cliffs, N.J., 1968.

4. P.L. BUTZER, W. ENGELS, S. RIES and R.L. STENS: The Shannon sampling series and the reconstruction of signals in terms of linear, quadratic and cubic splines. SIAM J. Appl. Math. 46 (1986), 299-323.

5. P.L. BUTZER, S. RIES and R.L. STENS: Approximation of continuous and discontinuous functions by generalized sampling series. J. Approx. Theory 50 (1987), 25-39.

6. P.L. BUTZER, W. SPLETTSTOESSER and R.L. STENS: The sampling theorem and linear prediction in signal analysis. Jahresber. Deutsch. Math.-Verein. 90 (1988), (in print).

7. P.L. BUTZER and R.L. STENS: Prediction of non-band-limited signals from past samples in terms of splines of low degree. Math. Nachr. (in print).

8. H. DYM and H.P. McKEAN: Gaussian Processes, Function Theory, and the Inverse Spectral Problem. Academic Press, New York, 1976, xii +333 pp.

9. G. FIX and G. STRANG: Fourier analysis of the finite element method in Ritz-Galerkin theory. Studies Appl. Math. 48 (1969), 268-273.

10. S. HAYKIN: Introduction to Adaptive Filters. MacMillan, New York and London, 1984, xii +217 pp.

11. T. KAILATH: Lectures on Linear Least-Square Estimation. CISM Courses and Lectures No. 140. Springer, Wien / New York 1976, ii +169 pp.

12. A.N. KOLMOGOROV: Interpolation and Extrapolation von stationären zufälligen Folgen. Isz. Akad. Nauk SSSR. Ser. Math. 5 (1941), 3-14.

13. M.G. KREIN: On a problem of extrapolation of A.N. Kolmogorov. Dokl. Akad. Nauk SSSR 46 (1954), 306-309.

14. M.G. KREIN: On a fundamental approximation problem in the theory of extrapolation and filtration of stationary random processes. Dokl. Akad. Nauk SSSR

94 (1954), 13-16 [Engl. transl.: Amer. Math. Soc. Selected Transl. Math. Statist. Prob. 4 (1964), 127-131].

15. Y.W. LEE: Statistical Theory of Communication. John Wiley, New York, 1960, xvii + 509 pp.

16. J. MAKHOUL: Linear prediction: A tutorial review. Proc. IEEE 63 (1975), 561-580.

17. J.D. MARKEL and H.H. GRAY, JR.: Linear Prediction of Speech. Springer, New York, 1982.

18. D.H. MUGLER and W. SPLETTSTOESSER: Difference methods and round-off error bounds for the prediction of bandlimited functions from past samples. Frequenz 39 (1985), 182-187.

19. D.H. MUGLER and W. SPLETTSTOESSER: Linear prediction from samples of a function and its derivatives. IEEE Trans. Inform. Theory IT-33 (1987), 360-366.

20. B. NOBLE: Methods based on the Wiener-Hopf Technique for the Solution of Partial Differential Equations. Pergamon, London, 1958, X + 246 pp.

21. S. RIES and R.L. STENS: Approximation of generalized sampling series. In: Constructive Theory of Functions (Proc. Conf. Varna, Bulgaria, May 27 - June 2, 1984; Eds. Bl. Sendov et al.). Publ. House Bulg. Acad. Sci., Sofia, 1984, (939 pp.), pp. 746-756.

22. W. SPLETTSTOESSER: Bandbegrenzte und effektiv bandbegrenzte Funktionen und ihre Praediktion aus Abtastwerten. Habilitationsschrift, RWTH Aachen, 1981, 65 pp.

23. W. SPLETTSTOESSER: On the prediction of bandlimited signals from past samples. Information Sci. 28 (1982), 115-130.

24. W. SPLETTSTOESSER: Lineare Praediktion von nicht bandbegrenzten Funktionen. Z. Angew. Math. Mech. 64 (1984), T 939 - T 395.

25. L.A. WAINSTEIN and V.D. ZUBAKOV: Extraction of Signals from Noise. Prentice-Hall, Englewood Cliffs, N.J., 1962.

26. N. WIENER: Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. M.I.T. Press, Cambridge, MA., 1949.

27. N. WIENER: Collected Works, Vol. III. (ed. by P.R. Masani), MIT Press, Cambridge, MA., 1981.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

SOME NEW INEQUALITIES FOR THE ZEROS OF LAGUERRE POLYNOMIALS*

LUIGI GATTESCHI

ABSTRACT: *It is shown that certain approximations for the zeros* $\lambda_{n,k}^{(\alpha)}$ *of the Laguerre polynomials* $L_n^{(\alpha)}(x)$, $\alpha > -1$, *are upper or lower bounds. These bounds involve the zeros of the Bessel function* $J_\alpha(x)$ *or the zeros of the Airy function* Ai (x) *and are obtained by using the Sturm comparison theorem.*

1. INTRODUCTION

In a recent paper [3] we have obtained some inequalities for the zeros of Jacobi polynomials. In this paper we will apply the same technique to derive bounds for the zeros $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, of the Laguerre polynomials $L_n^{(\alpha)}(x)$, $\alpha > -1$.

To this purpose we need the well-known Sturm comparison theorem in the following form given by Szegö [5, p. 19].

THEOREM 1.1 (Sturm's comparison theorem). *Let* $f(x)$ *and* $F(x)$ *be functions continuous in* $x_0 < x < X_0$, *with* $f(x) \leqslant F(x)$. *Let the functions* $y(x)$ *and* $Y(x)$, *both not identically zero, satisfy the differential equations*

$$y'' + f(x)\, y = 0 \quad , \quad Y'' + F(x)Y = 0, \tag{1.1}$$

respectively. Let x' *and* x'', $x' < x''$, *be two consecutive zeros of* $y(x)$. *Then the function* $Y(x)$ *has at least one zero in the interval* $x' < x < x''$ *provided* $f(x) \not\equiv F(x)$ *in* $[x', x'']$.

The statement also holds for $x' = x_0$ $[y(x_0 + 0) = 0]$ *if the additional condition*

* *This work was supported by the Consiglio Nazionale delle Ricerche of Italy and by the Ministero della Pubblica Istruzione of Italy.*

(1.2) $$\lim_{x \to x_0 + 0} [y'(x)\, Y(x) - y(x)\, Y'(x)] = 0$$

is satisfied (similarly for $x'' = X_0$*).*

The differential equations that we shall use as comparison equations are the ones used by Erdélyi [2] in deriving uniform asymptotic approximations for the Laguerre polynomials. Such equations can also be obtained by applying Olver's theory [4] to the asymptotic study of the Laguerre differential equation near the singularity $x = 0$ and near the turning point $x = 4n + 2\alpha + 2$.

Let us recall the following inequalities and asymptotic results.

THEOREM 1.2 (see Szegö [5], p. 127). *Let* $\alpha > -1$ *and let* $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, *be the zeros of* $L_n^{(\alpha)}(x)$ *in increasing order. Then*

(1.3) $$\lambda_{n,k}^{(\alpha)} > \frac{j_{\alpha,k}^2}{\nu}, \quad \nu = 4n + 2\alpha + 2,$$

for $k = 1, 2, \ldots, n$ *and where* $j_{\alpha,k}$ *is the* k*-th positive zero of the Bessel functions* $J_\alpha(x)$. *Furthermore, we have for a fixed* k, *as* $n \to \infty$,

(1.4) $$\lambda_{n,k}^{(\alpha)} = \frac{j_{\alpha,k}^2}{\nu} + O(n^{-2})$$

Tricomi [7] gave an improvement of (1.4), but its validity remains still restricted to the case of a fixed k.

THEOREM 1.3 (see Szegö [5], p. 131). *Let* a_k, $k = 1, 2, \ldots$, *be the zeros in decreasing order* $0 > a_1 > a_2 \ldots$, *of the Airy function* Ai (x).
If $|\alpha| \geqslant 1/4$, $\alpha > -1$, *then*

(1.5) $$\lambda_{n,k}^{(\alpha)} < [\nu^{1/2} + 2^{-1/3}\, \nu^{-1/6}\, a_{n-k+1}]^2,$$

for $k = 1, 2, \ldots, n$ *and where* ν *has the same meaning as in* (1.3). *Furthermore, we have for fixed* $n-k$, *as* $n \to \infty$,

(1.6) $$\lambda_{n,k}^{(\alpha)} = [\nu^{1/2} + 2^{-1/3}\, \nu^{-1/6}\, (a_{n-k+1} + \varepsilon_n)]^2,$$

where $\lim_{n \to \infty} \varepsilon_n = 0$.

Here the notations for the Airy function Ai(x) and for the zeros a_k are different from the ones used by Szegö; he uses $i_k = -3^{1/3}\, a_k$ instead of a_k.

A simplified form of a formula due to Tricomi [8] is given by the following:

THEOREM 1.4. *Let $\alpha > -1$ and let $x_{n,k}^{(\alpha)}$ be the root of the equation*

$$x - \sin x = \frac{4n - 4k + 3}{\nu}\,\pi \tag{1.7}$$

Then we have

$$\lambda_{n,k}^{(\alpha)} = \nu \cos^2 (x_{n,k}^{(\alpha)}/2) + O(n^{-1}), \tag{1.8}$$

for the zeros which belong to the interval $(a\nu, b\nu)$, where a and b, $0 < a < b < 1$, are fixed positive constants.

Recently, Temme [6] has obtained an interesting asymptotic representation of $\lambda_{n,k}^{(\alpha)}$ which involves the zeros of the Hermite polynomial $H_n(x)$. This representation gives good numerical results expecially for large values of the parameter α.

2. AN UPPER BOUND FOR THE ZEROS OF $L_n^{(\alpha)}(x)$

We shall refer throughout this paper to the differential equation

$$\frac{d^2y}{dt^2} + \left[\frac{\nu^2}{4}\left(\frac{1}{t} - 1\right) + \frac{1-\alpha^2}{4t^2}\right] y = 0, \tag{2.1}$$

$$\nu = 4n + 2\alpha + 2,\ \alpha > -1,$$

which is satisfied by

$$y(t) = e^{-\frac{1}{2}\nu t}\,(\nu t)^{\frac{1}{2}(\alpha+1)}\,L_n^{(\alpha)}(\nu t). \tag{2.2}$$

Now we observe that the function

$$z(t) = \left(\frac{f}{f'}\right)^{1/2} J_\alpha [f(t)] \tag{2.3}$$

satisfies the differential equation

$$\frac{d^2z}{dt^2} + F(t)\,z = 0, \tag{2.4}$$

where

$$F(t) = \frac{1}{2}\frac{f'''}{f'} - \frac{3}{4}\left(\frac{f''}{f'}\right)^2 + \left(\frac{1}{4} - \alpha^2\right)\left(\frac{f'}{f}\right)^2 + f'^2. \tag{2.5}$$

The equation (2.4) can be used, by assuming

$$(2.6)\qquad f(t) = \frac{\nu}{2}\,[(t-t^2)^{1/2} + \arcsin t^{1/2}],\ 0 < t < 1,$$

as a comparison equation to derive, by means of Sturm's method, inequalities for the zeros $\lambda_{n,k}^{(\alpha)}, k = 1, 2, \ldots, n$, of $L_n^{(\alpha)}(x)$.

This requires the study of the function

$$(2.7)\qquad G(t,\alpha) = F(t) - \left[\frac{\nu^2}{4}\left(\frac{1}{t} - 1\right) + \frac{1-\alpha^2}{4t^2}\right],$$

for $0 < t \leqslant 1$, or, more simply, of the function

$$(2.8)\qquad G^*(t,\alpha) = \frac{t\,G(t,\alpha)}{1-t},$$

which is analytic at $t = 0$. Indeed, it is easily seen that

$$(2.9)\qquad G^*(t,\alpha) = \frac{1/4 - \alpha^2}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} + \frac{3-8t-4(1-\alpha^2)(1-t)^2}{16t(1-t)^3}$$

and that

$$(2.10)\qquad G^*(t,\alpha) = \frac{\alpha^2 - 1}{6} + \frac{13\alpha^2 - 37}{60}\,t + O(t^2).$$

LEMMA 2.1. *Let* $\alpha^2 = 1$. *Then* $G^*(t, \pm 1) \leqslant 0$ *for* $0 \leqslant t < 1$. *The equality sign holds if and only if* $t = 0$.

We have

$$(2.11)\qquad 4t\,G^*(t, \pm 1) = \frac{-3t}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} + \frac{3-8t}{4(1-t)^3}.$$

First we prove that the property $G^*(t, \pm 1) < 0$, which is trivial for $3/8 \leqslant t < 1$, holds in the interval $1/16 \leqslant t < 1$. Indeed, by observing that the function

$$u(t) = \frac{t}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2}$$

increases in $0 < t < 1$ and that the function

$$v(t) = \frac{3 - 8t}{(1 - t)^3}$$

decreases in $1/16 < t < 1$, we obtain for $1/16 \leqslant t < 1$,

$$4t\, G^*(t, \pm 1) \leqslant -3\, u\left(\frac{1}{16}\right) + \frac{1}{4} v\left(\frac{1}{16}\right) < 0.$$

For the remaining interval $0 < t < 1/16$ we use the inequality

$$(2.12) \qquad \frac{1}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} > \frac{1}{4t}\left(1 + \frac{t}{3}\right), \quad 0 < t < 1,$$

and we set

$$(2.13) \qquad \frac{1}{(1 - t)^3} = 1 + 3t + a(t)\, t^2,$$

where

$$a(t) = \left[\frac{1}{(1-t)^3} - 1 - 3t\right] \frac{1}{t^2}$$

is an increasing function in $0 < t < 1$. Then from (2.11) we obtain

$$4\, G^*(t, \pm 1) < \frac{-3}{4t}\left(1 + \frac{t}{3}\right) + \left(\frac{3}{4t} - 2\right)\, [1+3t+a(t)t^2]$$

$$= 3\left[\frac{a(t)}{4} - 2\right] t - 2\, a(t)\, t^2,$$

i.e.

$$4\, G^*(t, \pm 1) < 3\left[\frac{a(t)}{4} - 2\right] t,$$

which, being $a(t) < a(1/16) = 6.689\ldots$ if $0 < t < 1/16$, completes the proof of the lemma.

LEMMA 2.2. *Let $G(t, \alpha)$ be the function defined by* (2.7). *In the interval $0 < t < 1$, $G(t, \alpha)$ has at least one zero if $\alpha^2 > 1$ and is negative if $\alpha^2 \leqslant 1$.*

For the proof we use the function $G^*(t, \alpha)$, defined by (2.8), which is continuous on $0 \leqslant t < 1$. From (2.9) and (2.10) we obtain, if $\alpha^2 > 1$,

$$\lim_{t \to 1-0} G^*(t, \alpha) = -\infty$$

and

$$\lim_{t \to 0+0} G^*(t, \alpha) = \frac{\alpha^2 - 1}{6} > 0,$$

respectively. Therefore, the first part of the lemma is proved.

We now observe that $G^*(t, \alpha)$ increases with respect to the parameter α^2. Indeed from (2.8) we have

$$\frac{\partial G^*}{\partial(\alpha^2)} = \frac{-1}{[(t-t^2)^{1/2} + \arcsin t^{1/2}]^2} + \frac{1}{4t(1-t)}$$

and setting $t^{1/2} = \sin\vartheta$ we find

$$\frac{\partial G^*}{\partial(\alpha^2)} = \frac{-1}{[\sin\vartheta\cos\vartheta + \vartheta]^2} + \frac{1}{4\sin^2\vartheta\cos^2\vartheta}$$

$$= \frac{-1}{\sin^2 2\vartheta\left[\frac{1}{2} + \frac{1}{2}\frac{2\vartheta}{\sin 2\vartheta}\right]^2} + \frac{1}{\sin^2 2\vartheta} > 0,$$

for $0 < \vartheta < \pi/2$. Hence, by using Lemma 2.1,

$$G^*(t, \alpha) \leqslant G^*(t, \pm 1) < 0, \quad 0 < t < 1,$$

when $\alpha^2 \leqslant 1$.

The property $G(t, \alpha) < 0$, *if* $0 < t < 1$ and $-1 < \alpha \leqslant 1$, established by Lemma 2.2, enables us to compare the zeros of the solution $y(t)$ of the equation (2.1) with the positive zeros of the function $z(t)$ defined by (2.3) and (2.6).

We notice that

$$\left(\frac{f}{f'}\right)^{1/2} = (2t)^{1/2}\left(1 + \frac{1}{6}t + \ldots\right), \quad 0 < t < 1.$$

Therefore, by means of the series representation of $J_\alpha(z)$, we obtain

(2.14) $\quad z(t) = t^{\frac{1}{2}(\alpha+1)}(a_0 + a_1 t + \ldots), \quad 0 < t < 1,$

with $a_0 \neq 0$.

Now, let $-1 < \alpha \leqslant 1$ and let $\tau_{n,k} \equiv \tau_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots$, be the zeros of $z(t)$ in $0 \leqslant t < 1$. We have

(2.15) $\tau_{n,0} = 0, \quad \frac{\nu}{2}\,[(\tau_{n,k} - \tau_{n,k}^2)^{1/2} + \arcsin \tau_{n,k}] = j_{\alpha,k},$

$k = 1, 2, \ldots, n.$

This follows by observing that $f(t)$ is a positive increasing function in $0 \leqslant t \leqslant 1$ varying from 0 to $\nu\pi/4$ and that (see Watson [9], p. 497) the number of the positive zeros of $x^{-\alpha}J_\alpha(x)$ between 0 and $n\pi + \alpha/2 + \pi/4$ is exactly n.

The condition (1.2), which is required when we apply Theorem 1.1 to the interval $[0, \tau_{n,1}]$, is satisfied if $\alpha > -1$ since, from (2.2),

$$y(t) = t^{\frac{1}{2}(\alpha+1)}\,(b_0 + b_1 t + \ldots)$$

and consequently, by using (2.14), we find that

$$y(t)\,z'(t) - z(t)\,y'(t) = 0\,(t^{\alpha+1}), \qquad t \to 0.$$

We may conclude that each interval

$$\tau_{n,k-1} < t < \tau_{n,k}, \quad k = 1, 2, \ldots, n,$$

contains exactly one zero $t_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, of $L_n^{(\alpha)}(\nu t)$.
Or, in other words: for the zeros $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, of $L_n^{(\alpha)}(x)$, if $-1 < \alpha \leqslant 1$, we have

(2.16) $\lambda_{n,k}^{(\alpha)} < \nu\,\tau_{n,k}^{(\alpha)}, \quad k = 1, 2, \ldots, n.$

This is the main result of this section. It can be stated in the following form.

THEOREM 2.1. *Let $-1 < \alpha \leqslant 1$. Let $x_{n,k}^{(\alpha)}$ be the root of the equation*

$$(2.17) \qquad x - \sin x = \pi - \frac{4\,j_{\alpha,k}}{\nu}, \quad \nu = 4n + 2\alpha + 2,$$

where $j_{\alpha,k}$ is the k-th positive zero of $J_\alpha(x)$. Then the k-th zero $\lambda_{n,k}^{(\alpha)}$ of $L_n^{(\alpha)}(x)$ satisfies the inequality

$$(2.18) \qquad \lambda_{n,k}^{(\alpha)} < \nu \cos^2 (x_{n,k}^{(\alpha)}/2), \quad k = 1, 2, \ldots, n.$$

Indeed, by setting $t^{1/2} = \cos\vartheta$, the equation $f(t) = j_{\alpha,k}$ becomes

$$2\vartheta - \sin 2\vartheta = \pi - \frac{4\,j_{\alpha,k}}{\nu}.$$

Thus, for the zeros $\tau_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$, defined by (2.15), we have

$$\tau_{n,k}^{(\alpha)} = \cos^2 (x_{n,k}^{(\alpha)}/2)$$

3. INEQUALITIES INVOLVING THE ZEROS OF THE AIRY FUNCTION

We shall use in this section, as comparison equation, the differential equation

$$(3.1) \qquad \frac{d^2u}{dt^2} + H(t)\, u = 0,$$

with

$$(3.2) \qquad H(t) = \frac{1}{2}\frac{h'''}{h'} - \frac{3}{4}\left(\frac{h''}{h'}\right)^2 - h\,h'^2,$$

which is satisfied by

$$(3.3) \qquad u(t) = [h'(t)]^{-1/2}\, \mathrm{Ai}\,[h(t)],$$

where $\mathrm{Ai}(x)$ is the Airy function of first kind.

It will be useful to recall some properties of $\mathrm{Ai}\,(x)$ and their zeros.

The function $\mathrm{Ai}\,(x)$ has no positive zero and infinitely many negative zeros, it is positive for $x > 0$ and $\mathrm{Ai}'(x) \to 0$ as $x \to \infty$. More precisely, we have as $x \to +\infty$.

$$(3.4) \quad \begin{cases} \mathrm{Ai}\,(x) \sim \dfrac{1}{2}\,\pi^{-1/2}\, x^{-1/4} \exp\left(-\dfrac{2}{3}\,x^{3/2}\right), \\ \mathrm{Ai}'(x) \sim \dfrac{1}{2}\,\pi^{-1/2}\, x^{1/4} \exp\left(-\dfrac{2}{3}\,x^{3/2}\right). \end{cases}$$

LEMMA 3.1. *Let a_k, $k = 1, 2, \ldots$, be the zeros in decreasing order of* $\mathrm{Ai}\,(x)$. *Then*

$$(3.5) \qquad -\left[\frac{3}{8}\left(4k - \frac{5}{6}\right)\pi\right]^{2/3} < a_k < -\left[\frac{3}{8}\,(4k-1)\,\pi\right]^{2/3},$$

$$k = 1, 2, \ldots .$$

For the proof we first consider the cylinder function

$$C_\alpha(x) = J_\alpha(x)\cos\varphi - Y_\alpha(x)\sin\varphi,$$

with $0 \leqslant \varphi < \pi$ and where $Y_\alpha(x)$ is the Bessel function of second kind. The positive zeros $c_{\alpha,k}$, $k = 1, 2, \ldots$, of $C_\alpha(x)$ satisfy, when $-1/2 < \alpha \leqslant 1/2$, the inequalities of Schafheitlin [9, p. 490].

$$(3.6)\qquad k\,\pi-\frac{\pi}{4}+\frac{1}{2}\,\alpha\,\pi-\varphi<c_{\alpha,k}<k\,\pi-\frac{\pi}{8}+\frac{1}{4}\,\alpha\,\pi-\varphi,$$

$$-\frac{1}{2}<\alpha\leqslant\frac{1}{2},\quad k=1,2,\ldots.$$

Next, by using the representation of Airy's function in terms of Bessel functions

$$\mathrm{Ai}\,(-x)=\frac{1}{3}\sqrt{x}\,[J_{1/3}(\xi)+J_{-1/3}(\xi)],\ \xi=\frac{2}{3}\,x^{3/2},$$

and the formula

$$J_{-1/3}(z)=J_{1/3}(z)\cos\pi/3-Y_{1/3}(z)\sin\pi/3,$$

we obtain

$$\mathrm{Ai}\,(-x)=\sqrt{\frac{x}{3}}\,[J_{1/3}(\xi)\cos\pi/6-Y_{1/3}(\xi)\sin\pi/6],\ \xi=\frac{2}{3}\,x^{3/2}.$$

Then, (3.6) with $\varphi=\pi/6$ yields

$$\frac{4k-1}{4}\,\pi<\frac{2}{3}\,(-a_k)^{3/2}<\frac{24k-5}{24}\,\pi,\quad k=1,2,\ldots,$$

that is the inequalities (3.5).

In order to compare the equation (3.1) with the Laguerre equation (2.1), we assume in (3.2)

$$(3.7)\qquad h(t)=\begin{cases}-\nu^{2/3}\left[\dfrac{3}{4}\,[\arccos t^{1/2}-(t-t^2)^{1/2}]\right]^{2/3}, & 0\leqslant t\leqslant 1,\\[2ex] +\nu^{2/3}\left[\dfrac{3}{4}\,[(t^2-t)^{1/2}-\operatorname{arccosh} t^{1/2}]\right]^{2/3}, & t>1.\end{cases}$$

We find

$$(3.8)\qquad H(t)=\frac{5}{36}\,\frac{1-t}{t\,\psi(t)}+\frac{3-8t}{16\,t^2\,(1-t)^2}+\frac{\nu^2\,(1-t)}{4t},\quad t>0$$

where

$$(3.9)\qquad \psi(t)=\begin{cases}[\arccos t^{1/2}-(t-t^2)^{1/2}]^2, & 0<t\leqslant 1,\\[2ex] -[(t^2-t)^{1/2}-\operatorname{arccosh} t^{1/2}]^2, & t>1.\end{cases}$$

The following lemma holds.

LEMMA 3.2. *In the interval* $0 < t < +\infty$, *the function*

$$(3.10)\quad Q(t,\alpha) = H(t) - \left[\frac{\nu^2}{4}\left(\frac{1}{t}-1\right) + \frac{1-\alpha^2}{4t^2}\right],$$

with $H(t)$ *defined by* (3.8) *and* (3.9), *is negative if* $\alpha^2 \leqslant 1/4$, *is positive if* $\alpha^2 \geqslant 4/9$ *and has exactly one zero if* $1/4 < \alpha^2 < 4/9$.

For the proof we write $Q(t,\alpha)$ in the form

$$Q(t,\alpha) = \frac{1}{4t^2}\left[q(t) + \alpha^2 - \frac{1}{4}\right],$$

where

$$q(t) = \frac{5}{9}\,\frac{t(1-t)}{\psi(t)} - \frac{t(2+3t)}{4(1-t)^2}$$

with $\psi(t)$ given by (3.9). Then it is easily seen that the function $q(t)$ is negative and decreasing for $0 < t < \infty$. Further, we have

$$q(0) = 0, \qquad \lim_{t\to\infty} q(t) = \frac{-7}{36};$$

whence the lemma readily follows.

We can now derive this final result.

THEOREM 3.1. *Let* $x_{n,k}^{(\alpha)}$ *be the root of the equation*

$$(3.11)\quad x - \sin x = \frac{8}{3\nu}\,(-a_{n-k+1})^{3/2},\quad \nu = 4n+2\alpha+2,$$

where a_j *is the* j*-th zero of the Airy function* Ai (x). *Then, for the* k*-th zero* $\lambda_{n,k}^{(\alpha)}$ *of* $L_n^{(\alpha)}(x)$ *we have*

$$(3.12)\quad \lambda_{n,k}^{(\alpha)} > \nu\cos^2(x_{n,k}^{(\alpha)}/2),\ \text{if}\ -\frac{1}{2} \leqslant \alpha \leqslant \frac{1}{2},$$

$$(3.13)\quad \lambda_{n,k}^{(\alpha)} < \nu\cos^2(x_{n,k}^{(\alpha)}/2),\ \text{if}\ -1 < \alpha \leqslant -\frac{2}{3}\ \text{ or }\ \alpha \geqslant \frac{2}{3},$$

where $k = 1, 2, \ldots, n$.

We give only the essential steps of the proof.

In the case $-1/2 \leqslant \alpha \leqslant 1/2$ the function $u(t)$, defined by (3.3) and (3.7), has exactly n real zeros. Moreover, these zeros belong to the interval $(0,1)$ and can be obtained by solving the equations

(3.14) $$\arccos t^{1/2} - (t-t^2)^{1/2} = \frac{4}{3\nu}(-a_m)^{3/2},\quad \nu = 4n+2\alpha+2,$$

for $m = 1, 2, \ldots, n$, with respect to t. Indeed, the inequalities (3.5) show that

$$0 < \frac{4(-a_m)^{3/2}}{3\nu} < \frac{4n-5/6}{4n+2\alpha+2}\,\frac{\pi}{2} < \frac{\pi}{2}\ ,\quad -\frac{1}{2} \leqslant \alpha \leqslant \frac{1}{2}\ ,$$

for $m = 1, 2, \ldots, n$, while

$$\frac{4(-a_m)^{3/2}}{3\nu} > \frac{\pi}{2}\ ,$$

for $m > n$. Since the function $h(t)$ is negative and decreasing, the statement easily follows.

Now, let

$$u_{n,1}^{(\alpha)} > u_{n,2}^{(\alpha)} > \ldots > u_{n,n}^{(\alpha)}$$

be the zeros, in decreasing order, of $u(t)$ and let $u_{n,0}^{(\alpha)} = +\infty$. According to Lemma 3.2, $Q(t, \alpha)$ is negative for $0 < t < \infty$ if $-1/2 \leqslant \alpha \leqslant 1/2$. Therefore, we can apply Theorem 1.1 to the interval $(0, \infty)$. By using (3.4) we see that the condition (1.2) is satisfied at $t = \infty$ and we conclude that each interval

$$u_{n,n-k+1}^{(\alpha)} < t < u_{n,k}^{(\alpha)}\ ,\ k = 1, 2, \ldots n,$$

contains exactly one zero $t_{n,m}^{(\alpha)}$, $m = 1, 2, \ldots, n$. More precisely, we have

$$t_{n,k}^{(\alpha)} > u_{n,n-k+1}^{(\alpha)},\ k = 1, 2, \ldots, n.$$

By setting $t^{1/2} = \cos(x/2)$ in (3.14) with $m = n - k + 1$ we derive (3.11) and (3.12).

For the proof of (3.13) we use the same interval $(0, \infty)$. Since $Q(t, \alpha)$ is positive if $-1 < \alpha \leqslant -2/3$ or $\alpha \geqslant 2/3$, we find that each interval

$$t_{n,k}^{(\alpha)} < t < t_{n,k+1}^{(\alpha)},\quad k = 1, 2, \ldots, n,$$

where $t_{n,n+1}^{(\alpha)} = +\infty$, contains at least one zero $u_{n,m}^{(\alpha)}$, $m = 1, 2, \ldots$ That is, we have

$$t_{n,k}^{(\alpha)} < u_{n,n-k+1}^{(\alpha)},\quad k = 1, 2, \ldots, n.$$

Whence (3.13) follows.

4. NUMERICAL RESULTS AND CONCLUDING REMARKS

The inequalities given in Theorem 2.1 and in Theorem 3.1 furnish very sharp results, which are generally better than those we can obtain by using previously known inequalities.

TABLE 1 - *Bounds for some zeros* $\lambda_{20,k}^{(0)}$.

k	*Lower bound (1.3)*	*Exact value*	*Upper bound (2.18)*
1	0.070527	0.070540	0.070547
2	0.371601	0.372127	0.372164
10	11.444867	12.038803	12.040338
20	46.951357	66.524416	66.642245

The Table 1, which refers to some few values of k in the case $\alpha = 0$ and $n = 20$, shows, in the first column, the lower bounds given by the old inequality (1.3) and, in the third column, the upper bounds furnished by applying the new inequality (2.18).

The case $-1/2 \leqslant \alpha \leqslant 1/2$ is particularly interesting. Indeed, in this case, Theorem 2.1 and Theorem 3.1 give upper and lower bounds for $\lambda_{n,k}^{(\alpha)}$ respectively and the following corollary holds.

COROLLARY 4.1. *Let* $-1/2 \leqslant \alpha \leqslant 1/2$ *and let* $X(y)$ *be the function that we obtain upon inverting*

$$y = \sin x - x. \tag{4.1}$$

Then

$$\nu\cos^2\left[\frac{1}{2}X\left(\frac{8}{3\nu}(-a_{n-k+1})^{3/2}\right)\right] < \lambda_{n,k}^{(\alpha)} < \nu\cos^2\left[\frac{1}{2}X\left(\pi - \frac{4j_{\alpha,k}}{\nu}\right)\right], \tag{4.2}$$

for $k = 1, 2, \ldots, n$ *and where* ν, a_s *and* $j_{\alpha,s}$ *have the previous meaning.*

The lower and upper bounds in Theorem 4.1 furnish very sharp results when they are used as approximations, say $1_{n,k}^{(\alpha)}$, of $\lambda_{n,k}^{(\alpha)}$, $k = 1, 2, \ldots, n$. This is shown in Figure 1 where the number of exact significant digits in the approximation of $\lambda_{n,k}^{(\alpha)}$, i.e. the *digits of accuracy* represented by the function

$$r_{n,k}(\alpha) = -\log_{10}\left|\frac{\lambda_{n,k}^{(\alpha)} - l_{n,k}^{(\alpha)}}{\lambda_{n,k}^{(\alpha)}}\right|$$

is plotted for $\alpha = 1/2$ and $n = 20$.

The curves Γ_1 and Γ_2 refer to the approximations

$$l_{20,k}^{(1/2)} = 83\cos^2\left[\frac{1}{2}X\left(\frac{8}{249}(-a_{21-k})^{3/2}\right)\right], \quad \textit{(lower bound)}$$

and

$$l_{20,k}^{(1/2)} = 83\cos^2\left[\frac{1}{2}X\left(\pi - \frac{4j_{1/2,k}}{83}\right)\right], \quad \textit{(upper bound)}$$

respectively.

In the same figure we have plotted (see the dashed curves γ_1 and γ_2) the digits of accuracy corresponding to the approximations

$$l_{20,k}^{(1/2)} = \frac{1}{83}\,j_{1/2,k}^2, \quad \textit{(lower bound)}$$

$$l_{20,k}^{(1/2)} = [(83)^{1/2} + 2^{-1/3}(83)^{-1/6}\,a_{21-k}]^2, \quad \textit{(upper bound)}$$

obtained by using the old bounds (1.3) and (1.5) respectively.

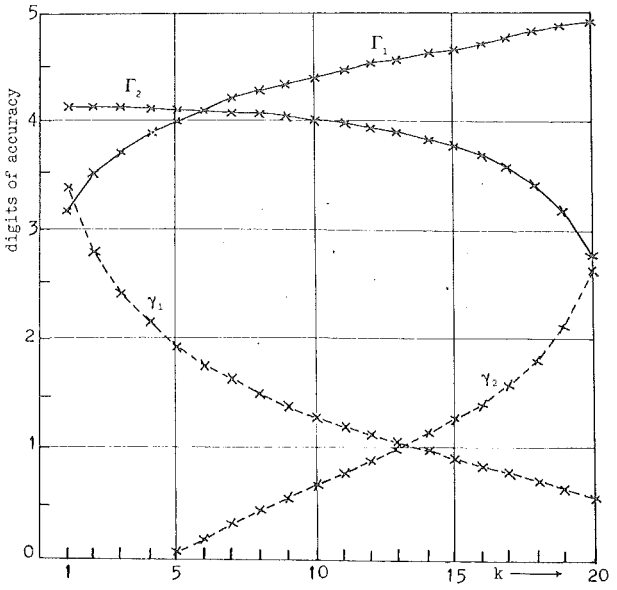


FIG. 1 - $r_{20,k}(1/2)$ versus k = 1(1)20.

The inequalities (4.2) can be used to derive upper and lower bounds for the zeros of the Hermite polynomials $H_n(x)$. Indeed, by taking into account that

$$H_{2m}(x) = (-1)^m 2^{2m} m!\, L_m^{(-1/2)}(x^2);\quad H_{2m+1}(x) = (-1)^m 2^{2m+1} m!\, x\, L_m^{(1/2)}(x^2),$$

and that

$$j_{1/2,k} = k\pi,\quad j_{-1/2,k} = (2k-1)\frac{\pi}{2},\qquad k = 1, 2, \ldots,$$

we obtain the following result:

COROLLARY 4.2. *Let $h_{n,k}$, $k = 1, 2, \ldots, [n/2]$, be the positive zeros in increasing order of the Hermite polynomial $H_n(x)$. Then*

$$\begin{aligned} h_{n,k} &< \sqrt{2n+1}\cos\left[\frac{1}{2}X\left(\frac{2n-4k+3}{2n+1}\pi\right)\right], \text{ if } n \text{ is even},\\ h_{n,k} &< \sqrt{2n+1}\cos\left[\frac{1}{2}X\left(\frac{2n-4k+3}{2n+1}\pi\right)\right], \text{ if } n \text{ is odd}, \end{aligned} \tag{4.3}$$

Furthermore, we have

$$h_{n,k} > \sqrt{2n+1}\cos\left[\frac{1}{2}X\left(\frac{8}{3(2n+1)}(-a_{[n/2]-k+1})^{3/2}\right)\right]. \tag{4.4}$$

Here $X(y)$ has the same meaning as in Corollary 4.1.

We remark that the Tricomi asymptotic formula (1.8) can be written, as $n \to \infty$,

$$\lambda_{n,k}^{(a)} \sim \nu\cos^2\left[\frac{1}{2}X\left(\frac{4n-4k+3}{\nu}\pi\right)\right],$$

for all the zeros $\lambda_{n,k}^{(a)}$, belonging to the interval $(a\nu, b\nu)$ with a and b fixed positive constants, $0 < a < b < 1$.

Now, by using the asymptotic expansions (see Abramowitz and Stegun [1], p. 371 and p. 450)

$$j_{a,s} = \left(s + \frac{\alpha}{2} - \frac{1}{4} \right) \pi \, [1+0(s^{-2})], \quad s \to \infty,$$

$$-a_s = \left[\frac{3\pi}{8} (4s-1) \right]^{2/3} [1+0(s^{-2})], \quad s \to \infty,$$

it is easily seen that

$$\pi - \frac{4j_{a,k}}{\nu} \sim \frac{4n-4k+3}{\nu} \pi, \text{ for } k \to \infty,$$

$$\frac{8}{3\nu}(-a_{n-k+1})^{3/2} \sim \frac{4n-4k+3}{\nu} \pi, \text{ for } n-k \to \infty.$$

Hence, the bounds for $\lambda_{n,k}^{(\alpha)}$ that we have considered in this paper are in fact approximations which coincide with the Tricomi approximation as $n \to \infty$, uniformly for all values of $k = [pn], [pn]+1, \ldots, [qn]$, where $p, q \in (0,1)$, $p < q$. More precisely, taking into account the results of Erdélyi [2] on the asymptotics for Laguerre polynomials, we have

$$\lambda_{n,k}^{(\alpha)} \sim \nu \cos^2 \left| \frac{1}{2} X \left(\pi - \frac{4j^{\alpha,k}}{\nu} \right) \right|, \quad n \to \infty,$$

$k = 1, 2, \ldots [qn],$

and

$$\lambda_{n,k}^{(\alpha)} \sim \nu \cos^2 \left| \frac{1}{2} X \left(\frac{8}{3\nu}(-a_{n-k+1})^{3/2} \right) \right|, \quad n \to \infty,$$

$k = [pn], [pn]+1, \ldots, n-1, n.$

longer assumed to be supported on $\mathbb{R}_+$ and to have constant sign. What, for example, would happen if one took a typical oscillatory measure, like $ds(t) = P_n(t)dt$ on $[-1,1]$, where P_n is the Legendre polynomial of degree n?

In a letter to Hermite, dated November 8, 1894 (in fact, his last letter in the life-long correspondence with Hermite; see Baillaud and Bourget [1905, v.2, pp. 439–441]), Stieltjes indeed looks at (what is now called) Legendre's function of the second kind

$$Q_n(z) = \int_{-1}^{1} \frac{P_n(t)}{z-t}\, dt, \tag{1.2}$$

expands it into descending powers of z (beginning with $z^{-(n+1)}$ by orthogonality of P_n) and then has the fortunate idea of expanding the reciprocal of Q_n,

$$\frac{1}{Q_n(z)} = z^{n+1}(\mu_0^{(n)} + \mu_1^{(n)} z^{-1} + \cdots), \quad \mu_0^{(n)} \neq 0. \tag{1.3}$$

This led him naturally to consider the polynomial part in (1.3),

$$E_{n+1}(z) = z^{n+1}(\mu_0^{(n)} + \mu_1^{(n)} z^{-1} + \cdots + \mu_{n+1}^{(n)} z^{-(n+1)}), \tag{1.4}$$

a polynomial of exact degree $n + 1$, now appropriately called *Stieltjes' polynomial*, and to investigate its properties. By a residue calculation, he first observes that

$$E_{n+1}(t) = \frac{1}{2\pi i} \oint_C \frac{dz}{(z-t)Q_n(z)}, \tag{1.5}$$

where C is a sufficiently large contour, and then goes on to multiply (1.5) by $t^k P_n(t)dt$, $k = 0, 1, \ldots, n$, and to integrate, obtaining

$$\int_{-1}^{1} E_{n+1}(t) t^k P_n(t)dt = \frac{1}{2\pi i} \oint_C \frac{dz}{Q_n(z)} \int_{-1}^{1} \frac{t^k P_n(t)}{z-t}\, dt$$

$$= \frac{1}{2\pi i} \oint_C \frac{dz}{Q_n(z)} \int_{-1}^{1} \frac{z^k - (z^k - t^k)}{z-t} P_n(t)dt$$

$$= \frac{1}{2\pi i} \oint_C \frac{z^k dz}{Q_n(z)} \int_{-1}^{1} \frac{P_n(t)}{z-t}\, dt$$

$$= \frac{1}{2\pi i} \oint_C z^k\, dz = 0,$$

where orthogonality of P_n is used in the third equality. Thus,

$$\int_{-1}^{1} E_{n+1}(t) p(t) P_n(t) dt = 0, \quad \text{all} \quad p \in \mathbb{P}_n, \tag{1.6}$$

that is, Stieltjes' polynomial E_{n+1} *is orthogonal to all lower-degree polynomials relative to the* (sign-variable) *measure* $ds(t) = P_n(t)dt$.

At this point, Stieltjes conjectures (1) that E_{n+1} has $n + 1$ real simple zeros, all contained in $(-1,1)$ and (2) that they separate those of P_n. He presents a numerical example with $n = 4$. He furthermore believes (strongly so in the case of reality and simplicity of the roots, less so for the separation property) that this is a special case of "a much more general theorem".

In his reply (of November 10, 1894), Hermite expressed his delight in the polynomials E_{n+1} and "the beautiful properties" conjectured for it and encouraged Stieltjes to look for a differential equation as a possible key to these properties. Stieltjes may have already been too ill to respond. Neither he, nor anybody else after him was able to give an affirmative answer to Hermite's suggestion. (It has been found, nevertheless, that the Stieltjes polynomials, at least in the realm of Jacobi measures $d\sigma^{(\alpha,\beta)}(t) = (1-t)^{\alpha}(1+t)^{\beta}\, dt$, do *not* satisfy a three-term recurrence relation unless $|\alpha| = |\beta| = \frac{1}{2}$, in which case they do, and in fact also satisfy a differential equation; cf. Monegato [1982].)

Stieltjes' ideas seem to have gone unnoticed for many years. Geronimus in 1930, however, developed similar ideas, considering in place of (1.3) the expansion of $[Q_n(z)\sqrt{z^2-1}]^{-1}$, where $Q_n(z) = \int_{-1}^{1} \pi_n(t\,; wdt) w(t) dt/(z-t)$ and $\pi_n(\cdot\,; wdt)$ is the nth degree orthogonal polynomial associated with the weight function $w(t) = (1-t)^{\alpha}(1+t)^{\beta} h(t)$, h being continuous and positive on $[-1,1]$ (Geronimus [1930]). Although this approach does not lead to a perfect orthogonality result, like the one in (1.6), it nevertheless has relevance to the subject at hand; see the beginning of Subsection 3.5 below.

The first who has taken up Stieltjes' challenge in earnest was Szegö in 1935. He expresses (Szegö [1935]) Stieltjes' polynomial on the circle as a cosine polynomial,

$$E_{n+1}(\cos\theta) = \lambda_0^{(n)} \cos(n+1)\theta + \lambda_1^{(n)} \cos(n-1)\theta + \cdots , \tag{1.7}$$

and relates an extended (infinite) sequence $\lambda_\nu = \lambda_\nu^{(n)}$ to an explicitly known sequence $f_\nu = f_\nu^{(n)}$ via a reciprocity identity for the respective power series. From this he proves $\lambda_0 > 0$ and the negativity of all λ_ν, $\nu \geq 1$, as well as $\sum_{\nu=0}^{\infty} \lambda_\nu = 0$. It follows from this that the polynomial $\lambda_0 z^{n+1} + \lambda_1 z^{n-1} + \cdots$ has all its zeros in $|z| < 1$, which implies, via the argument principle, that (1.7) vanishes at least $2n + 2$ times. This proves Stieltjes' first conjecture. Szegö also proves the second conjecture, but this requires a deeper analysis involving, in particular, Legendre functions on the cut.

Szegö's analysis is not peculiar to Legendre polynomials. Indeed, he himself extends it to Gegenbauer polynomials $P_n^{(\lambda)}$, orthogonal on $[-1,1]$ with respect to the measure $d\sigma(t) = (1-t^2)^{\lambda-1/2}dt$, $\lambda > -1/2$. If $E_{n+1}^{(\lambda)}$ denotes the corresponding Stieltjes polynomial,

$$\int_{-1}^{1} E_{n+1}^{(\lambda)}(t)p(t)P_n^{(\lambda)}(t)(1-t^2)^{\lambda-1/2}dt = 0, \quad \text{all } p \in \mathbb{P}_n, \tag{1.8}$$

which (up to a multiplicative constant) is uniquely defined, then Szegö shows that both conjectures of Stieltjes continue to hold for $0 < \lambda \leq 2$. When $\lambda=0$, two zeros of $E_{n+1}^{(\lambda)}$ move into the endpoints ± 1; they move outside of $[-1,1]$ for $\lambda<0$, as is shown by the example $n=2$. The question of whether the same can happen for $\lambda>2$ is left unanswered by Szegö. (The answer is still unknown today, but, according to Table 3.1 below, is probably "no", at least as long as the interlacing property holds.)

Szegö concludes by considering the Gaussian quadrature formula for the (sign-variable) measure $ds(t) = P_n(t)dt$ and shows that its weights alternate in sign.

This brings us naturally to the work of Kronrod in 1964, which is also concerned with quadrature. Motivated by a desire to economically estimate the error in the classical Gaussian quadrature formula

$$\int_{-1}^{1} f(t)dt \approx \sum_{\nu=1}^{n} \gamma_\nu f(\tau_\nu), \tag{1.9}$$

where $\tau_\nu = \tau_\nu^{(n)}$ are the zeros of the Legendre polynomial P_n and $\gamma_\nu = \gamma_\nu^{(n)}$ the corresponding Christoffel numbers, Kronrod [1964a,b] proposes to extend the n-point formula (1.9) to a $(2n + 1)$-point formula

$$\int_{-1}^{1} f(t)dt = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \tag{1.10}$$

in which the τ_ν are the same as in (1.9), but new nodes τ_μ^* and new weights σ_ν, σ_μ^* have been introduced and chosen to increase the degree of exactness from $2n - 1$ (for (1.9)) to $3n + 1$ (for (1.10)), i.e.,

$$R_n(f) = 0, \quad \text{all } f \in \mathbb{P}_{3n+1}. \tag{1.11}$$

It turns out that the nodes τ_μ^* must be precisely the zeros of Stieltjes' polynomial E_{n+1}. With all nodes τ_ν, τ_μ^* at hand, it is then easy to determine the weights σ_ν, σ_μ^* by interpolation.

In the same manner, one can try to extend the Gauss-Gegenbauer quadrature formula to a formula of the type

$$\int_{-1}^{1} f(t)(1-t^2)^{\lambda-\frac{1}{2}}\, dt = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \quad \lambda > -\tfrac{1}{2}, \tag{1.12}$$

and, more generally, to do the same for an integral with arbitrary (positive) measure $d\sigma$,

$$\int_{-1}^{1} f(t)d\sigma(t) = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \quad R_n(\mathbb{P}_{3n+1}) = 0. \tag{1.13}$$

(The dependence of the nodes and weights on n and $d\sigma$ will from now on be suppressed in our notation.) The new nodes τ_μ^*, similarly as before, are then the zeros of the (unique, monic) polynomial $\pi_{n+1}^*(\cdot) = \pi_{n+1}^*(\cdot\,; d\sigma)$ satisfying the orthogonality property

$$\int_{\mathbb{R}} \pi_{n+1}^*(t)p(t)\pi_n(t)d\sigma(t) = 0, \quad \text{all } p \in \mathbb{P}_n, \tag{1.14}$$

where $\pi_n(\cdot) = \pi_n(\cdot\,; d\sigma)$ is the orthogonal polynomial of degree n associated with the measure $d\sigma$. To be useful in practice, the formulae (1.12), (1.13) should have nodes τ_μ^* which are all contained in the support interval of $d\sigma$ and are different from the τ_ν, and they should have weights σ_ν, σ_μ^* which, if at all possible, are all positive. By Szegö's theory, we know that the former is

true for (1.12), if $0 < \lambda \leq 2$, while the latter has been proven true by Monegato [1978a] if $0 \leq \lambda \leq 1$, hence, in particular, for the original Gauss-Kronrod formula (1.10) (which corresponds to $\lambda = 1/2$).

Soon after Kronrod's work, it has occurred to a number of people (probably first to Patterson [1968a]) that other quadrature rules can be similarly extended, for example, the Gauss-Lobatto rule. In addition, it is not unreasonable to also consider the interpolatory quadrature rule based solely on the nodes τ_μ^* in (1.13). In the case of (1.10), numerical results suggest that these quadrature rules also have all weights positive and enjoy an interlacing property of their own: the zeros of E_{n+1} alternate with those of E_n; cf. Monegato [1982]. Indeed, having three quadrature rules at disposal – the one just mentioned, the Gauss rule (1.9), and (1.10) – with degrees of exactness roughly equal to n, $2n$ and $3n$, respectively, might well be an attractive feature that could be useful in automatic integration schemes (Kahaner [1987]).

Orthogonality with respect to sign-variable measures and related quadrature rules have independently been studied by Struble [1963], who develops a general theory. It might be interesting to explore this theory in the framework of more general indefinite inner product spaces (cf., e.g., Bognár [1974]).

The merit of discovering the connection between Kronrod's work and the earlier work of Stieltjes and Szegö is due to Mysovskih [1964], although it has been noted, independently, in the Western literature, by Barrucand [1970]. The relevance of Geronimus' work to Gauss-Kronrod quadrature is pointed out by Monegato [1982] and Monegato and Palamara Orsi [1985].

Brief accounts of the Kronrod and Patterson methods can be found in Davis and Rabinowitz [1984, pp. 106–109, 426] and Atkinson [1978, pp. 243–248].

2. Extended quadrature formulae. We now give a more systematic treatment of the problem of extending quadrature rules. We begin with a general theorem, which has become part of "folklore" in numerical quadrature and is difficult to attribute to any one in particular. In its key ingredients, it goes back to Jacobi [1826].

Let $d\sigma$ be a nonnegative measure on the real line $\mathbb{R}$, with bounded or unbounded support and with infinitely many points of increase. Assume that all its moments $\mu_k = \int_{\mathbb{R}} t^k d\sigma(t)$ exist and are finite. We consider quadrature rules of the form

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{N} \sigma_\nu f(\tau_\nu) + R_N(f), \tag{2.1}$$

where τ_ν, σ_ν are real and $N \geq 1$ an integer. We say that (2.1) has *degree of exactness* d if $R_N(f) = 0$ for every $f \in \mathbb{P}_d$, the class of polynomials of degree $\leq d$. We associate with (2.1) the polynomial

$$\omega(t) = \prod_{\nu=1}^{N} (t - \tau_\nu) \tag{2.2}$$

and call it the *node polynomial*. The theorem in question then reads as follows.

Theorem. *The quadrature rule* (2.1) *has degree of exactness* $d = N - 1 + k$, $k \geq 0$, *if and only if both of the following conditions are satisfied:*

(i) (2.1) *is interpolatory* (i.e., $d = N-1$);

(ii) $\int_{\mathbb{R}} \omega(t)p(t)d\sigma(t) = 0$ *for all* $p \in \mathbb{P}_{k-1}$.

We remark that polynomial degree of exactness $N-1$ (the case $k=0$ of the theorem) can always be achieved, simply by interpolating at the nodes τ_ν; this is condition (i) of the theorem. To get higher degree of exactness ($k>0$), the node polynomial, according to (ii), has to be orthogonal (relative to the measure $d\sigma$) to sufficiently many polynomials. If we have complete freedom in the choice of τ_ν and σ_ν, we can take k as large as $k=N$, in which case (ii) identifies $\omega(\cdot)$ with the (monic) orthogonal polynomial $\pi_N(\cdot\,; d\sigma)$ of degree N associated with the measure $d\sigma$, and the nodes τ_ν in (2.1) with its zeros. This, of course, is the well-known Gauss-Christoffel quadrature rule (cf., e.g., Gautschi [1981]).

The situation we are going to consider here is somewhat different: We shall assume that some of the nodes are prescribed and the rest variable. Let

$$N = N^\circ + N^*, \tag{2.3}$$

and suppose the prescribed (distinct) nodes are $\tau_1, \tau_2, \ldots, \tau_{N^\circ}$; we denote the remaining ones by

$$\tau_\mu^* = \tau_{N^\circ+\mu}, \quad \mu = 1, 2, \ldots, N^*. \tag{2.4}$$

Correspondingly, we let $\sigma_\mu^* = \sigma_{N^\circ+\mu}$ and write (2.1) in the form

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{N^\circ} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{N^*} \sigma_\mu^* f(\tau_\mu^*) + R_N(f). \tag{2.5}$$

We may interpret (2.5) as an "extension" of some quadrature rule

$$\int_{\mathbb{R}} f(t)d\sigma(t) \approx \sum_{\nu=1}^{N^\circ} \gamma_\nu f(\tau_\nu). \tag{2.6}$$

The degree of exactness of (2.6) is quite irrelevant for what follows, as the weights γ_ν are being discarded.

Putting

$$\pi_{N^\circ}^\circ(t) = \prod_{\nu=1}^{N^\circ} (t - \tau_\nu), \quad \pi_{N^*}^*(t) = \prod_{\mu=1}^{N^*} (t - t_\mu^*), \tag{2.7}$$

the theorem above, since $\omega(t) = \pi_{N^\circ}^\circ(t)\pi_{N^*}^*(t)$, becomes:

Corollary. *The quadrature formula* (2.5) *has degree of exactness* $d = N - 1 + k$, $k \geq 0$, *with N given by* (2.3), *if and only if it is interpolatory and the polynomial* $\pi_{N^*}^*$ *satisfies*

$$\int_{\mathbb{R}} \pi_{N^*}^*(t)p(t)\pi_{N^\circ}^\circ(t)d\sigma(t) = 0, \quad \textit{all} \quad p \in \mathbb{P}_{k-1}. \tag{2.8_k}$$

One expects the maximum degree of exactness to be realized for $k = N^*$ (there are $N + N^*$ degrees of freedom!), in which case (2.8_k) becomes

$$\int_{\mathbb{R}} \pi_{N^*}^*(t)p(t)\pi_{N^\circ}^\circ(t)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_{N^*-1}. \tag{2.8_{N^*}}$$

We call (2.5) an *optimal extension* of (2.6) if $k = N^*$, i.e., if (2.8_{N^*}) holds, and a *nonoptimal* [*interpolatory*] *extension* if (2.8_k) holds with $0 \leq k < N^*$ [k=0]. (We assume $p \equiv 0$ in (2.8_k) if k=0.) Thus, (2.5) is an optimal extension of (2.6) if and only if $\pi_{N^*}^*$ *is orthogonal to all lower-degree polynomials with respect to the* (sign-variable) *measure* $d\sigma^*(t) = \pi_{N^\circ}^\circ(t)d\sigma(t)$. Here is how sign-variable measures enter into the process of extending quadrature rules.

We now discuss a number of specific examples.

Example 2.1: Gauss-Kronrod formulae.

This is the case $N^\circ = n$, $\pi_{N^\circ}^\circ(\cdot) = \pi_n(\cdot\,; d\sigma)$, $N^* = n+1$, so that $N = 2n + 1$, $d = 3n + 1$, and (2.8_{N^*}) takes the form

$$\int_{\mathbb{R}} \pi_{n+1}^*(t)p(t)\pi_n(t\,; d\sigma)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_n. \tag{2.9}$$

(We must necessarily have $N^* \geq n + 1$ in this case; cf. Monegato [1980].) In other words, the classical n-point Gauss-Christoffel formula is optimally extended to a $(2n + 1)$-point formula of the form

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f). \tag{2.10}$$

The measure involved in the orthogonality relation (2.9) is $d\sigma^*(t) = \pi_n(t\,; d\sigma)d\sigma(t)$, which for $d\sigma(t) = dt$ is precisely the one considered by Stieltjes. We call π_{n+1}^* in (2.9) the *Stieltjes polynomial* associated with $d\sigma$ and denote it by $\pi_{n+1}^*(\cdot) = \pi_{n+1}^*(\cdot\,; d\sigma)$. It is easily seen that π_{n+1}^* (assumed monic of degree $n+1$) is uniquely determined by (2.9).

For the weights in (2.10) one finds (see, e.g., Monegato [1976])

$$\begin{aligned} \sigma_\nu &= \gamma_\nu + \frac{\|\pi_n\|_{d\sigma}^2}{\pi_{n+1}^*(\tau_\nu)\,\pi_n'(\tau_\nu)}, \qquad \nu = 1, 2, \ldots, n; \\ \sigma_\mu^* &= \frac{\|\pi_n\|_{d\sigma}^2}{\pi_n(\tau_\mu^*)\,\pi_{n+1}^{*\prime}(\tau_\mu^*)}, \qquad \mu = 1, 2, \ldots, n+1, \end{aligned} \tag{2.11}$$

where $\gamma_\nu = \gamma_\nu^{(n)}(d\sigma)$ are the Christoffel numbers, and $\|\cdot\|_{d\sigma}$ the L_2-norm for the measure $d\sigma$.

For symmetric measures, i.e., $d\sigma(-t) = d\sigma(t)$ and the support of $d\sigma$ symmetric with respect to the origin, it follows easily from uniqueness that

$$\pi_n(-t\,; d\sigma) = (-1)^n\,\pi_n(t\,; d\sigma), \quad \pi_{n+1}^*(-t\,; d\sigma) = (-1)^{n+1}\,\pi_{n+1}^*(t\,; d\sigma) \quad (d\sigma \text{ symmetric}), \tag{2.12}$$

so that (2.9) holds trivially for even polynomials p and is therefore valid for all $p \in \mathbb{P}_{n+1}$ if n is odd. Thus, $d = 3n + 1$ if n is even, and $d = 3n + 2$ if n is odd. (In special cases, the degree of exactness can be even higher; see Subsections 3.3 and 3.5 for examples.)

Example 2.2: Kronrod extension of Gauss-Radau formulae.

For definiteness we consider only the Radau formula with fixed node τ_0 at -1. The case $\tau_0 = 1$ is treated similarly.

We assume that $d\sigma$ is supported on $[-1,1]$ and that the measure $(1+t)d\sigma(t)$ allows $(2n + 1)$-point Kronrod extension, i.e., the Stieltjes polynomial $\pi^*_{n+1}(\cdot\,; (1+t)d\sigma)$ has distinct real zeros, all in $(-1,1)$ and all different from the zeros of $\pi_n(\cdot\,; (1+t)d\sigma)$. Then there exists a unique optimal extension of the Gauss-Radau formula for the measure $d\sigma$. It has the form

$$\int_{-1}^{1} f(t)d\sigma(t) = \sigma_0 f(-1) + \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma^*_\mu f(\tau^*_\mu) + R_n(f) \tag{2.13}$$

and corresponds to the case $N^\circ = n+1$, $\pi^\circ_{N^\circ}(t) = (1+t)\pi_n(t\,; (1+t)d\sigma)$, $N^* = n+1$, hence has degree of exactness (at least) $d = 3n + 2$. The orthogonality condition (2.8_{N^*}) assumes the form

$$\int_{-1}^{1} \pi^*_{n+1}(t)p(t)\pi_n(t\,; (1+t)d\sigma)(1+t)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_n. \tag{2.14}$$

Thus, as far as the nodes τ^*_μ are concerned, we can obtain them exactly as if we were to extend the Gauss formula for the measure $(1+t)d\sigma(t)$. Also, the quantities $(1 + \tau_\nu)\sigma_\nu$ and $(1 + \tau^*_\mu)\sigma^*_\mu$ can be obtained by expressions which are identical to the ones on the right-hand sides of (2.11), where the Christoffel numbers and norm refer to the measure $(1+t)d\sigma(t)$. The weight σ_0 then follows from $\sigma_0 + \sum_{\nu=1}^{n} \sigma_\nu + \sum_{\mu=1}^{n+1} \sigma^*_\mu = \mu_0, \quad \mu_0 = \int_{\mathbb{R}} d\sigma(t)$.

Example 2.3: Kronrod extension of Gauss-Lobatto formulae.

We assume, similarly as in Example 2.2, that the measure $(1 - t^2)d\sigma(t)$, supported on $[-1,1]$, allows Kronrod extension. Then the unique optimal extension of the $(n+2)$-point Gauss-Lobatto formula for the measure $d\sigma$ is given by

$$\int_{-1}^{1} f(t)d\sigma(t) = \sigma_0 f(-1) + \sigma_{n+1} f(1) + \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma^*_\mu f(\tau^*_\mu) + R_n(f), \tag{2.15}$$

and is the case $N^\circ = n+2$, $\pi^\circ_{N^\circ}(t) = (1-t^2)\,\pi_n(t\,; (1-t^2)d\sigma)$, $N^* = n+1$ of (2.5), with the degree of exactness now being (at least) $d = 3n + 3$. The orthogonality condition (2.8_{N^*}) becomes

$$\int_{-1}^{1} \pi^*_{n+1}(t)p(t)\pi_n(t\,; (1-t^2)d\sigma)(1-t^2)d\sigma(t) = 0, \quad \text{all} \quad p \in \mathbb{P}_n, \tag{2.16}$$

and is the same as for Kronrod extension of the n-point Gauss formula for the measure $(1-t^2)d\sigma(t)$. Again, the quantities $(1-\tau_\nu^2)\sigma_\nu$ and $(1-\tau_\mu^{*2})\sigma_\mu^*$ have representations identical to those on the right of (2.11), the measure being $(1-t^2)d\sigma(t)$ throughout. The remaining weights σ_0, σ_{n+1} are most easily obtained by solving the system of two linear equations expressing exactness of (2.15) for $f(t)=1$ and $f(t)=t$.

We remark that in the special case of Jacobi measures $d\sigma^{(\alpha,\beta)}(t)=(1-t)^\alpha(1+t)^\beta dt$, $\alpha>-1$, $\beta>-1$, we have

$$\pi_n(\cdot\,;(1-t^2)d\sigma^{(\alpha,\beta)})=\frac{1}{n+1}\,\pi'_{n+1}(\cdot\,;d\sigma^{(\alpha,\beta)}), \tag{2.17}$$

as follows readily from the identity $(d/dt)P_{n+1}^{(\alpha,\beta)}(t)=\tfrac{1}{2}(n+\alpha+\beta+2)\,P_n^{(\alpha+1,\beta+1)}(t)$ for Jacobi polynomials.

Example 2.4: "Kronrod-heavy" extension of Gauss formulae.

The "Kronrod nodes" τ_μ^* and "Gauss nodes" τ_ν in the Gauss-Kronrod formula (2.10) are nicely balanced, in that exactly one Kronrod node fits into the space between two consecutive Gauss nodes and between the extreme Gauss nodes and the respective endpoints (possibly $\pm\infty$) of the support interval of $d\sigma$. There are, however, occasions (for example, in cases of nonexistence; cf. Subsection 3.4) where it might be necessary to forgo this balance in favor of more Kronrod nodes; we call such extensions *Kronrod-heavy*. These also fit into the general scheme (2.5), where $N^\circ=n$, $\pi_{N^\circ}^\circ(\cdot)=\pi_n(\cdot\,;d\sigma)$, $N^*=n+q$ with $q>1$, and give rise to the orthogonality condition

$$\int_{\mathbb{R}}\pi_{n+q}^*(t)p(t)\pi_n(t\,;d\sigma)d\sigma(t)=0,\quad\text{all}\quad p\in\mathbb{P}_{n+q-1}. \tag{2.18}$$

In contrast to Gauss-Kronrod formulae, the unique existence of π_{n+q}^*, let alone the reality of its zeros, is no longer assured. Starting with the unique $\pi_{n+1}^*(\cdot\,;d\sigma)=\pi_{n+1,n}^*(\cdot)$, however, there is an infinite sequence $\{\pi_{n+q_m,n}^*\}_{m=1}^\infty$ of uniquely determined polynomials $\pi_{n+q_m}^*=\pi_{n+q_m,n}^*$ of exact degree $n+q_m$, $1=q_1<q_2<q_3<\cdots$, such that (2.18) holds with $q=q_m$, and such that no polynomial $\pi_{n+q_m}^*$ of degree $<n+q_m$ exists for which (2.18) holds with $q=q_m$ (Monegato [1980]).

One can try, of course, to extend in this manner other quadrature formulae, e.g., the Gauss-Radau or Gauss-Lobatto formulae.

Example 2.5: Repeated Kronrod extension of Gauss formulae.

Given an n-point Gauss formula, one can try to extend it optimally to a $(2n + 1)$-point formula as in Example 2.1, then extend this formula once again to a $(4n + 3)$-point formula (by optimally adding $2n + 2$ new nodes), and so on. The likelihood of such repeated extensions to all exist (i.e., have real distinct nodes) is probably quite small. Remarkably, however, for n=3 and $d\sigma(t) = dt$ on [−1,1], such extensions, even with all weights positive, have been successfully computed by Patterson [1968a], [1973] up to the 255-point formula.

For the second extension, for example, the node polynomial π^*_{2n+2} must be orthogonal to all lower-degree polynomials with respect to the measure $d\sigma^*(t) = \pi_n(t\,; d\sigma)\,\pi^*_{n+1}(t\,; d\sigma)d\sigma(t)$.

Example 2.6: Extension by contraction.

As contradictory as this may sound, the point here is that one starts with a "base formula" containing a sufficiently large number of nodes, then successively removes subsets of nodes to generate a sequence of quadrature rules having fewer and fewer nodes. Looking at this sequence in the opposite direction then turns it into a sequence of (finitely often) extended quadrature rules.

More specifically, following Patterson [1968b], one takes as base formula any $(2^r + 1)$-point formula and then defines r subsets of points by successively deleting alternate points from the preceding subset (keeping the first and the last). For example, if r=3, the successive three subsets of the original points with index set {1,2,3,4,5,6,7,8,9} contain the points with indices {1,3,5,7,9}, {1,5,9} and {1,9}, respectively. A sequence of r+1 quadrature formulae can now be defined by taking the interpolatory formulae for the original node set and all r subsets of nodes. (A slightly different procedure is proposed by Rabinowitz, Kautsky and Elhay; see Rabinowitz, Kautsky, Elhay and Butcher [1987, Appendix A, p.125].)

The reality of the nodes is thereby trivially guaranteed, but not necessarily the positivity of the weights. Patterson [1968b], nevertheless, finds by computation that all weights remain posi-

tive if one starts with the 33-point, or 65-point Gauss-Legendre formula (r=5 and r=6, respectively), or with the 65-point Lobatto formula (r=6) as base formulae.

Another example of a suitable base formula, which in fact (Imhof [1963], Brass [1977, Satz 77]) has positivity of all weights built in, is the Clenshaw-Curtis formula (Clenshaw and Curtis [1960]) based on the initial point set $\tau_\nu = \cos(\nu\pi/2^r)$, $\nu = 0, 1, 2, \ldots, 2^r$.

If one is willing to delete successively one point at a time, then the following result of Rabinowitz, Kautsky, Elhay and Butcher [1987] is of interest: Given any interpolatory quadrature rule with all weights positive, it is possible to delete one of its points such that the interpolatory rule based on the reduced point set has all weights nonnegative.

All sequences of extended quadrature rules in Example 2.6 are examples of nonoptimal, in fact interpolatory, extensions. Other examples of nonoptimal, even subinterpolatory, extensions are those of product rules given by Dagnino [1983] (see also Dagnino [1986]). The severe sacrifice in polynomial degree of exactness is justified in this reference in terms of a simplified convergence and stability theory.

We restricted our discussion here to quadrature rules of the simplest type (2.1). There is little work in the literature on the extension of quadrature rules involving derivatives. Bellen and Guerra [1982], however, extend Turán-type formulae, but work them out only in very simple special cases.

3. Existence, nonexistence and remainder term. We consider here mainly the Gauss-Kronrod formula as defined in Example 2.1, that is,

$$\int_{\mathbb{R}} f(t)d\sigma(t) = \sum_{\nu=1}^{n} \sigma_\nu f(\tau_\nu) + \sum_{\mu=1}^{n+1} \sigma_\mu^* f(\tau_\mu^*) + R_n(f), \quad R_n(\mathbb{P}_{3n+1}) = 0. \tag{3.1}$$

We say that the nodes τ_ν, τ_μ^* in (3.1) *interlace* if they are all real and, when arranged decreasingly, satisfy

$$-\infty < \tau_{n+1}^* < \tau_n < \tau_n^* < \cdots < \tau_2^* < \tau_1 < \tau_1^* < \infty. \tag{3.2}$$

For any given $n \geq 1$, the following properties are of interest:

(a) The nodes τ_ν, τ_μ^* interlace.

(b) All nodes τ_ν, τ_μ^*, in addition to interlacing, are contained in the interior of the smallest interval containing the support of $d\sigma$.

(c) The nodes interlace and each weight σ_ν is positive. (It is known, cf. Monegato [1976], that the interlacing property is equivalent to $\sigma_\mu^* > 0$, all μ.)

(d) All nodes, without necessarily satisfying (a) and/or (b), are real.

Little has been *proved* with regard to these properties; any new piece of information, from whatever source – computational or otherwise – should therefore be greeted with appreciation. In this section, we give an account of what is known, or what can be conjectured, for some classical and nonclassical measures.

3.1 *Gegenbauer measures* $d\sigma^{(\lambda)}(t) = (1-t^2)^{\lambda-\frac{1}{2}}dt$ *on* $[-1,1]$, $\lambda > -\frac{1}{2}$. Properties (a) and (b), as already mentioned in Section 1, have been proved for all $n \geq 1$ by Szegö [1935], when $0 < \lambda \leq 2$, and property (c) by Monegato [1978a], when $0 \leq \lambda \leq 1$. Properties (a) and (b) also hold for the extension of Lobatto formulae, if $-\frac{1}{2} < \lambda \leq 1$ (cf. Example 2.3), but nothing as yet has been proved concerning property (c). This, then, is the extent of what is known rigorously, for arbitrary n, at this time.

A good deal more, however, can be uncovered for specific values of n, if we let the parameter λ move continuously away from the above intervals and observe the resulting motion of the nodes τ_ν, τ_μ^* and the movement of the weights σ_ν, σ_μ^*. Given n, property (a) will cease to hold at the very moment a node τ_ν collides (for the first time) with a node τ_μ^*. This event is coincident with the vanishing of the resultant of $\pi_n(\cdot\,; d\sigma^{(\lambda)})$ and $\pi_{n+1}^*(\cdot\,; d\sigma^{(\lambda)})$. When λ has moved beyond this critical value, the nodes τ_ν and τ_μ^* involved in the collision have likely crossed each other, so that two Kronrod nodes now lie between consecutive Gauss nodes. Only now is it possible that two Kronrod nodes may collide and split into a pair of complex nodes, an event that is signaled by the vanishing of the discriminant of $\pi_{n+1}^*(\cdot\,; d\sigma^{(\lambda)})$. By using purely algebraic methods, it is thus possible to delineate parameter intervals in which properties (a) and (d) are valid. The

subintervals of the first of these, in which properties (b) and (c) hold, can be determined rather more easily, in an obvious manner.

This has been carried out computationally in Gautschi and Notaris [submitted] for values of n up to 40. Based on these results it is conjectured (and proved for $n \leq 4$) that property (p) holds for $\lambda_n^p < \lambda < \Lambda_n^p$, where the bounds λ_n^p and Λ_n^p for $p = a, b, c, d$ are as shown in Table 3.1.

n	λ_n^a	Λ_n^a	λ_n^b	Λ_n^b	λ_n^c	Λ_n^c	λ_n^d	Λ_n^d
1	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞
2	$-\frac{1}{2}$	∞	0	∞	$-\frac{1}{2}$	∞	$-\frac{1}{2}$	∞
3	$-\frac{1}{2}$	16	0	16	$-\frac{1}{2}$	6.552...	$-\frac{1}{2}$	16
4	$-\frac{1}{2}$	∞	0	∞	$-\frac{1}{2}$	51.78...	$-\frac{1}{2}$	∞
≥ 5	$-\frac{1}{2}$	Λ_n^a	0	Λ_n^a	$-\frac{1}{2}$	Λ_n^c	$-\frac{1}{2}$	Λ_n^d

Table 3.1. *Property (p) for Gegenbauer measures*

Here, Λ_n^a, Λ_n^c, Λ_n^d are certain constants satisfying $1 < \Lambda_n^a < \infty$, $1 < \Lambda_n^c < \Lambda_n^a$ and $\Lambda_n^d \geq \Lambda_n^a$ with equality precisely when $n = 4r - 1$, $r = 1, 2, 3, \ldots$. Numerical values of these constants, to 10 decimal places, are provided in the cited reference for $n = 5(1)20(4)40$.

The fact that Kronrod extension (satisfying properties (c) and (d)) cannot exist for all $n \geq 1$ when λ is sufficiently large, not even if the degree of exactness is lowered to $[2rn + l]$, $r > 1$, l an integer, is claimed by Monegato [1979]. (The proof given is erroneous, but can be repaired; Monegato [1987].)

3.2 *Jacobi measures* $d\sigma^{(\alpha,\beta)}(t) = (1-t)^\alpha(1+t)^\beta dt$ *on* $[-1,1]$. Since interchanging the parameters α and β has the effect of changing the signs of the nodes τ_ν and τ_μ^*, hence, if the order (3.2) is maintained, of renumbering them in reverse order, and the same renumbering applies to the weights σ_ν and σ_μ^*, the validity of property (p), $p = a, b, c, d$, is unaffected by such an interchange. We will assume, therefore, that $-1 < \alpha \leq \beta$.

Except for the cases $|\alpha| = |\beta| = \frac{1}{2}$ (considered in Subsection 3.3) and the transformations to Gegenbauer measures noted below, the only known proven result is that property (b) is false for $\alpha = -\frac{1}{2}, -\frac{1}{2} < \beta < \frac{1}{2}$ when n is even, and for $\alpha = -\frac{1}{2}, \frac{1}{2} < \beta \leq 3/2$ when n is odd (Rabinowitz [1983, p.75] †).

(†) There is a misprint on p.75 of this reference: The superscript $\mu + \frac{1}{2}$ should be replaced by $\mu - \frac{1}{2}$ twice in Eq. (68), and twice in the discussion immediately following Eq. (69).

Monegato [1982] notes that $\pi_{n+1}^{*(\alpha,-\frac{1}{2})}(2t^2-1) = 2^{n+1}\, t\, \pi_{2n+1}^{*(\alpha,\alpha)}(t) - d_n$, where d_n is an explicitly given constant, and similarly, $\pi_{n+1}^{*(\alpha,\frac{1}{2})}(2t^2-1) = 2^{n+1}\, \pi_{2n+2}^{*(\alpha,\alpha)}(t)$. In the latter case, there are also simple relationships between the weights σ_ν, σ_μ^* of the respective Gauss-Kronrod formulae (3.1); cf. Gautschi and Notaris [submitted, Thm. 5.1]. The cases $\alpha > -1$, $\beta = \pm \frac{1}{2}$ can thus be reduced to the Gegenbauer case, and appeal can be made to the empirical results of Subsection 3.1, at least when $\beta = \frac{1}{2}$. A similar reduction is possible in the case $\alpha > -1$, $\beta = \alpha + 1$ (Monegato [1982, Eq. (36)]), which is of interest in connection with Kronrod extension of Gauss-Radau formulae for Gegenbauer measures (cf. Example 2.2).

The algebraic methods described in Subsection 3.1 have also been applied to general Jacobi measures (Gautschi and Notaris [submitted]) and the results for $2 \le n \le 10$ displayed by means of graphs. There are marked qualitative differences for n even and n odd, as is shown in Figure 3.1 for the cases n=6 and n=7. The region of validity for property (p) is consistently below the curve labeled "p", except for p=b and n even, when it is above and to the right of the curve.

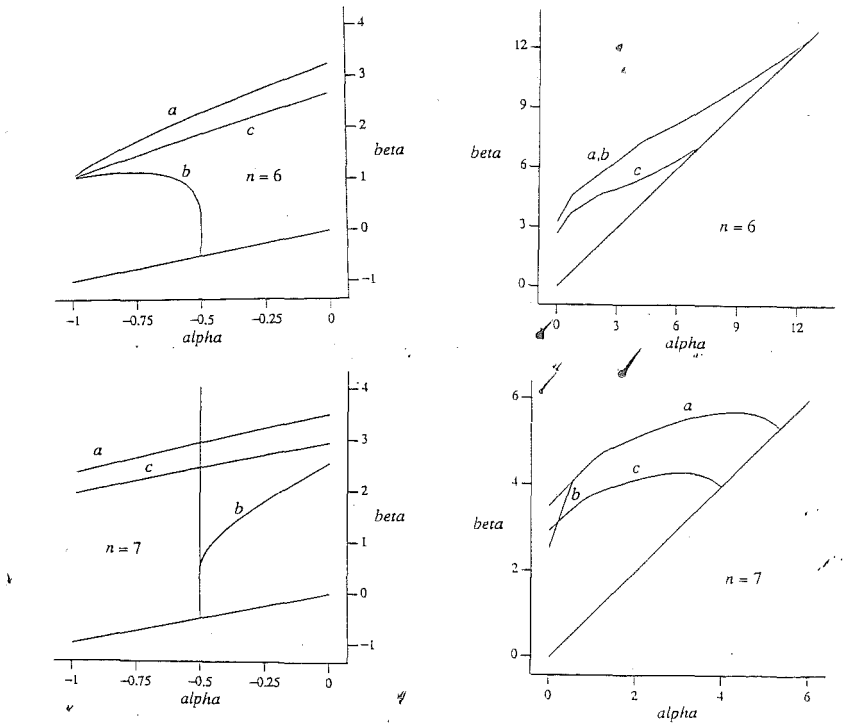


Figure 3.1. *Property (p), p = a, b, c for the Jacobi measure* $d\sigma^{(\alpha,\beta)}$ *when n=6 and n=7*

3.3 *Chebyshev measures of* 1st, 2nd *and* 3rd *kind.* These are the cases $|\alpha| = |\beta| = \frac{1}{2}$ of the Jacobi measure $d\sigma^{(\alpha,\beta)}$. They are the only known cases in which both the Gauss formulae and their Kronrod extensions can be written down explicitly (in terms of trigonometric functions). If $\alpha = \beta = -\frac{1}{2}$, the (optimal) extension of the n-point Gauss-Chebyshev formula of the first kind, when $n \geq 2$, is simply the $(2n + 1)$-point Lobatto formula for the same weight function. (For n=1, it is the 3-point Gauss-Chebyshev rule.) To get the Kronrod extension of the n-point Gauss-Chebyshev formula of the second kind ($\alpha = \beta = \frac{1}{2}$), it suffices to replace n by $2n + 1$ in the same formula. Finally, for $\alpha = -\frac{1}{2}$, $\beta = \frac{1}{2}$, the Kronrod extension is the Radau formula (with fixed node at 1) for the same weight function. All these extended formulae have elevated degrees of exactness, namely $4n - 1$, $4n + 1$ and $4n$, respectively, and enjoy property (p) for all $p = a, b, c$ (hence also d). These elegant relationships have been noted as early as 1964 by Mysovskih [1964]; see also Monegato [1982, p.147]. For the first two cases, Monegato [1976] points out that the formulae can be extended infinitely often in an explicit manner.

3.4 *Laguerre and Hermite measures.* Here is another instance in which a nonexistence result is known (Kahaner and Monegato [1978]): For the generalized Laguerre measure $d\sigma^{(\alpha)}(t) = t^{\alpha}e^{-t}\,dt$ on $[0,\infty]$, $-1 < \alpha \leq 1$, the Kronrod extension of the n-point Gauss-Laguerre formula, with real nodes and positive weights, does not exist when $n \geq 23$, and if α=0 not even for n>1. As a corollary, n-point Gauss-Hermite formulae cannot be so extended, unless $n = 1, 2$ or 4, confirming earlier empirical results of Ramskiĭ [1974]. These negative results led Kahaner, Waldvogel and Fullerton [1982], [1984] to explore the feasibility of Kronrod-heavy extensions for the Laguerre measure. Computational experience is reported for $n = 1(1)10$ and q=8 (11 for n=1 and 9 for n=2), where q is defined as in Example 2.4.

3.5 *Other measures.* At the heart of Geronimus' theory (Geronimus [1930]) is the measure $d\sigma_{\mu}(t) = (1-t^2)^{\frac{1}{2}}dt/(1 - \mu t^2)$ on $[-1,1]$, $-\infty < \mu \leq 1$. The corresponding polynomials $\pi_n(\cdot\,; d\sigma_{\mu})$ and $\pi^*_{n+1}(\cdot\,; d\sigma_{\mu})$ turn out to be linear combinations of Chebyshev polynomials U_n, U_{n-2} and T_{n+1}, T_{n-1}, respectively. This allows explicit construction of the associated Gauss-Kronrod extension and verification of all properties (a) – (c); cf. Gautschi and Rivlin [submitted]. In addition,

the degree of exactness is exceptionally high (Monegato [1982, p.146]). Similar expressions for π_n and π^*_{n+1} result if the denominator of $d\sigma_\mu$ is replaced by a positive, not necessarily even, polynomial of degree 2 (Monegato and Palamara Orsi [1985]).

Gautschi and Notaris [submitted, Thm. 5.2] observe that the problem of Kronrod extension for the measure $^\gamma d\sigma^{(\alpha)}(t) = |t|^\gamma (1-t^2)^\alpha dt$ on $[-1,1]$, $\alpha > -1$, $\gamma > -1$, can be reduced, when n is odd, to the analogous problem for the Jacobi measure $d\sigma^{(\alpha,(\gamma+1)/2)}$.

Very little is known for measures unrelated to classical measures. One that is likely to admit satisfactory Kronrod extension for every $n \geq 1$ (judging from numerical results of Caliò, Gautschi and Marchetti [1986]) is the logarithmic measure $d\sigma(t) = \ln(1/t)dt$ on $[0,1]$ for which properties (a), (b) and (c) appear to be all true. The same is conjectured for measures $d\sigma(t) = t^\alpha \ln(1/t)dt$, $\alpha = \pm \frac{1}{2}$, except for $\alpha = -\frac{1}{2}$ and n odd, in which case property (b), though not (d), fails, the polynomial $\pi^*_{n+1}(\cdot\,; d\sigma)$ having exactly one negative zero.

3.6 *Remainder term.* The Gauss-Kronrod formula (3.1) can be characterized in the manner of Markov [1885] as the unique quadrature formula (if it exists) obtained by integrating the interpolation polynomial $p_{3n+1}(f\,; \tau_\nu, \tau^*_\mu, \tau^*_\mu; \cdot)$ (with simple knots τ_ν and double knots τ^*_μ) of degree $\leq 3n + 1$ and by requiring (if possible) that the coefficients of all derivative terms in the resulting quadrature sum be zero. The elementary Hermite interpolation polynomials g_ν, h_μ, k_μ associated with this interpolation process can be easily expressed in terms of the fundamental Lagrange polynomials l_ν and l^*_μ for the nodes $\tau_1, \tau_2, \ldots, \tau_n$ and $\tau^*_1, \tau^*_2, \ldots, \tau^*_{n+1}$, respectively (see, e.g., Caliò, Gautschi and Marchetti [1986, Eq. (3.13)]). The coefficients $\sigma^{*\prime}_\mu$ required to be zero are then

$$\sigma^{*\prime}_\mu = \int_{\mathbb{R}} k_\mu(t) d\sigma(t), \quad \mu = 1, 2, \ldots, n+1, \tag{3.3}$$

where

$$k_\mu(t) = \frac{\pi_n(t)}{\pi_n(\tau^*_\mu)} [l^*_\mu(t)]^2 (t - \tau^*_\mu), \quad \pi_n(\cdot) = \pi_n(\cdot\,; d\sigma). \tag{3.4}$$

Thus we must have

$$\pi^*_{n+1}{}'(\tau^*_\mu) \int_{\mathbb{R}} \pi_n(t)[l^*_\mu(t)]^2 (t - \tau^*_\mu) d\sigma(t) = \int_{\mathbb{R}} \pi^*_{n+1}(t) l^*_\mu(t) \pi_n(t) d\sigma(t) = 0, \quad \mu = 1, 2, \ldots, n+1, \tag{3.5}$$

which, by the linear independence of the l_μ^*, is equivalent to the orthogonality condition (2.9).

From interpolation theory there follows that

$$R_n(f) = \frac{1}{(3n+2)!} \int_R [\pi_{n+1}^*(t)]^2 f^{(3n+2)}(\tau(t))\pi_n(t)d\sigma(t), \tag{3.6}$$

provided $f \in C^{3n+2}$ on an interval containing $\text{supp}(d\sigma)$. For Gegenbauer measures $d\sigma(t) = (1-t^2)^{\lambda-\frac{1}{2}}dt$ on $[-1,1]$, with $0 < \lambda < 1$, Monegato [1978b], relying heavily on Szegö's theory, shows that $|\pi_{n+1}^*(t\,; d\sigma)| < 2^{-n}$ on $[-1,1]$, which in combination with known bounds for $|\pi_n(\cdot\,; d\sigma)|$ yields an explicit upper bound for $|R_n(f)|$ in terms of $||f^{(3n+2)}||_\infty$. Rabinowitz [1980] improves this bound slightly and extends it to the case $1 < \lambda < 2$, as well as to Kronrod extensions of Gauss-Lobatto rules for $-\frac{1}{2} < \lambda \leq 1$, $\lambda \neq 0$. He also proves that for $0 < \lambda \leq 2$, $\lambda \neq 1$ the degrees $d = 3n+1$ and $d = 3n+2$ for n even and odd, respectively, are indeed the exact degrees of precision. (When $\lambda = 1$, one has exact degree $4n + 1$, and when $\lambda = 0$ exact degree $4n - 1$.) Analogous statements are proved for the Kronrod extension of the Gauss-Lobatto rule. Szegö's work, again, proves invaluable for this analysis, as it does, in combination with a result of Akrivis and Förster [1984, Proposition 1], to show that the remainder term $R_n(f)$ is indefinite if $0 < \lambda < 1$ and $n \geq 2$ (Rabinowitz [1986b]). For $\lambda > 1$, the question of definiteness is still open; it is also open for Kronrod extensions of Gauss-Lobatto rules for any λ (with the obvious exceptions).

Error constants in Davis-Rabinowitz type estimates of the remainder (Davis and Rabinowitz [1954]) for functions analytic on elliptic domains are given by Patterson [1968a] for his repeated extensions of the 3-point Gauss formula. They are compared with the corresponding constants for the Gauss and Clenshaw-Curtis formulae having the same number of points.

4. Computational methods, numerical tables, computer programs and applications.

4.1 *Computational methods.* Kronrod originally computed the Stieltjes polynomial $\pi_{n+1}^*(\cdot\,; dt)$ in power form, requiring it to be orthogonal (in the sense of (1.14)) to all monomials of degree $\leq n$. The zeros of π_{n+1}^* are then obtained by a rootfinding procedure, and the weights

σ_ν, σ_μ^* from a system of linear equations expressing exactness of (1.10) for the first $2n+1$ monomials. (Symmetry, of course, was used throughout.) As he himself observes, the procedure is subject to considerable loss of accuracy and therefore requires elevated precision. Patterson [1968a] achieves better stability by expanding π_{n+1}^* in Legendre polynomials and computing the coefficients recursively. He does so not only for the Kronrod extension of the Gauss formula, but likewise for the extension of the Lobatto formula. Further improvements and simplifications result from expansion in Chebyshev polynomials; cf. Piessens and Branders [1974]. Their procedure, even somewhat simplified and generalized to Gegenbauer measures, actually can be extracted from the work of Szegö [1935], as is pointed out by Monegato [1978b]; see also Monegato [1979], [1982]. For Gegenbauer measures, then, this seems to be the method of choice. Once the nodes have been computed, the weights can be obtained, e.g., by the formulae in (2.11).

Expansion of $\pi_{n+1}^*(\cdot\,; d\sigma)$ in orthogonal polynomials $\pi_k(\cdot\,; d\sigma)$, $k = 0, 1, \ldots, n+1$, however, is possible for arbitrary measures $d\sigma$. Replacing $p(\cdot)$ in (2.9) successively by $\pi_i(\cdot\,; d\sigma)$, $i = 0, 1, \ldots, n$, indeed yields a triangular system of equations which can be readily solved. Its coefficients can be computed, e.g., by Gauss-Christoffel quadrature relative to the measure $d\sigma$, using $[(3n + 3)/2]$ points; cf. Caliò, Gautschi and Marchetti [1986, Sec. 4]. (For another method, see Caliò, Marchetti and Pizzi [1984] and Caliò and Marchetti [1987].)

A rather different approach, resembling (in fact, generalizing) the well-known Golub-Welsch procedure (Golub and Welsch [1969]) for computing Gauss-Christoffel quadrature formulae is developed by Kautsky and Elhay [1984] and Elhay and Kautsky [1984] and relies on eigenvalues of suitably constructed matrices. For the weights, these authors use their own methods and software for generating interpolatory quadrature rules (Kautsky and Elhay [1982], Elhay and Kautsky [1985]).

Instead of computing, as above, the Gauss-Kronrod formula piecemeal – first the Stieltjes polynomial, then its zeros, and finally the weights – it might be preferable to compute these components all at once, for example by applying Newton's method to the system of $3n + 2$ (non-

linear) equations expressing exactness of the quadrature rule (2.10) for some set of basis functions in $\mathbb{P}_{3n+1}$. The feasibility of this idea is demonstrated in Caliò, Gautschi and Marchetti [1986], where the numerical condition of the underlying problem, hence the stability of the procedure, is also analyzed. It appears, though, that this method runs into severe ill-conditioning when one attempts to use it for repeated Kronrod extension (Gautschi and Notaris [in preparation]).

4.2. *Numerical tables.* There are a number of places where Kronrod extensions of n-point Gauss formulae can be found tabulated: Kronrod himself (Kronrod [1964b]) has them (transformed to the interval [0,1]) for $n = 1(1)40$ to 16 decimals (also in binary form!). In addition, he tabulates errors incurred when the formulae are applied to monomials. Patterson [1968a] (on microfiche) gives 20 S values for $n = 65$, and Piessens [1973] 16 S values for $n = 10$. The most accurate are the 33-decimal tables for $n = 7$, 10(5)30 in Piessens et al. [1983, pp. 19–23]. Extensions of $(n+2)$-point Lobatto formulae, $n = 1(1)7$ and $n = 63$, can be found to 20 decimals in Patterson [1968a] (on microfiche), and extensions of the $(n+1)$-point Radau formula, $n = 2(2)16$ (but incomplete), to 15 decimals in Baratella [1979].

Repeated Gauss-Kronrod extensions of the 3-point Gauss formula, as far up as the 127-point formula, are given to 20 significant digits in Patterson [1968a] (on microfiche), and the 255-point formula to the same accuracy in Patterson [1973] (in a Fortran data statement). The repeatedly extended 10-point formula, through the one with 87 points, is given to 33 decimals in Piessens et al. [1983, pp. 19, 26–27]. Extensions in the sense of Example 2.6 are tabulated to 20 decimals in Patterson [1968b] (on microfiche), using the 33-point and 65-point Gauss formula, as well as the 65-point Lobatto formula as "base formulae".

For measures other than the constant weight measure, there are 25 S tables of $(2n+1)$-point Gauss-Kronrod formulae for $d\sigma(t) = t^{\alpha} \ln(1/t)dt$ on [0,1], $\alpha = 0, \pm\frac{1}{2}$, where $n = 5(5)25$ for $\alpha = 0, \frac{1}{2}$, and $n = 4(4)24$ for $\alpha = -\frac{1}{2}$ (Caliò, Gautschi and Marchetti [1986, Suppl. S57–S63]). 15 S tables for the same weight functions, but with $n = 4$ and 12 for $\alpha = 0, \frac{1}{2}$, and $n = 6$ and 12 for $\alpha = -\frac{1}{2}$, are given in Caliò and Marchetti [1987]. Kahaner, Waldvogel and Fullerton [1984]

provide 15–18 S tables of Kronrod-heavy extensions of the Gauss-Laguerre formula ($d\sigma(t) = e^{-t}\,dt$ on $[0, \infty]$) with $n = 1$, $q = 3(1)6$ and $n = 10$, $q = 18$ (in the notation of Example 2.4).

We finally mention the 16 S tables of Piessens [1969] of the complex Gauss-Kronrod formulae, with $n = 2(1)12$, for the Bromwich integral, and the 15 S table of the interpolatory $(n+1)$-point formula based solely on the Kronrod nodes, given by Monegato [1982] for $d\sigma(t) = dt$ and $n = 2(1)9$.

4.3. *Computer programs.* Fortran programs for Kronrod extension of the n-point Gauss formula are provided in Squire [1970, p. 279] for $n = 20$, and in Piessens and Branders [1974] for arbitrary n. Dagnino and Fiorentino [1984] describe a Fortran program (listed in Dagnino and Fiorentino [1983]) generating Gauss-Kronrod formulae for Gegenbauer measures $d\sigma(t) = (1-t^2)^{\lambda - \frac{1}{2}}dt$ on $[-1,1]$, $0 \leq \lambda \leq 2$, $\lambda \neq 1$, using the recursive algorithm of Szegö as resurrected by Monegato (cf. Subsection 4.1). Programs for more general measures are described and listed in Caliò and Marchetti [1987], [1985], respectively.

A number of routines employing Gauss-Kronrod quadrature in the context of automatic integration are discussed and listed in Piessens et al. [1983].

4.4. *Applications.* The original motivation came from a desire to estimate the error of Gaussian, or other quadrature formulae (taking the more accurate Kronrod extension as a substitute for the exact answer). The need for such error estimates has recently been highlighted in connection with the development of automatic integration schemes; see, e.g., Cranley and Patterson [1971], Patterson [1973], Piessens [1973] and Piessens et al. [1983]. For an interesting interpretation of the Kronrod scheme of error estimation, see Laurie [1985]. A rather different estimation procedure is proposed in Berntsen and Espelid [1984].

Patterson's repeated extensions of the 3-point Gauss-Legendre rule (cf. Example 2.5) has been used with some success in certain methods to compute improper integrals arising in weakly singular integral equations. One method employs the ε – algorithm to accelerate a sequence of approximants (Evans, Hyslop and Morgan [1983]), another suitable transformations of variables to attenuate the singularity (Evans, Forbes and Hyslop [1983]).

Kronrod's idea has been applied to other types of integrals, for example, as already mentioned, to the Bromwich integral for the inversion of Laplace transforms (Piessens [1969]), and to Cauchy type singular integrals involving Gegenbauer measures (Rabinowitz [1983]). These applications, especially the latter, are not entirely straightforward, as the occurrence of numerical cancellation, or derivative values, may present difficulties. They can be surmounted, to some extent, by more stable implementations (Rabinowitz [1986a]), using, in part, Kronrod-heavy extensions (with $q = 2$; see Example 2.4). For an application of Kronrod's idea to cubature formulae, see Malik [1980], Genz and Malik [1980], [1983], Laurie [1982], Neumann [1982], Cools and Haegemans [1986], [1987] and Berntsen and Espelid [1987].

An interesting application, first noted by Barrucand [1970], is the use of Gauss-Kronrod formulae for computing Fourier coefficients in orthogonal expansions,

$$c_n(f) = \|\pi_n\|_{d\sigma}^{-1} \int_{\mathbb{R}} \pi_n(t) f(t) d\sigma(t), \quad n = 0, 1, 2, \ldots, \tag{4.1}$$

where $\pi_n(\cdot) = \pi_n(\cdot; d\sigma)$ is the nth degree orthogonal polynomial associated with the measure $d\sigma$. The $(2n+1)$-point Gauss-Kronrod formula (for the coefficient c_n), in this case, reduces to an $(n+1)$-point formula,

$$c_n(f) = \|\pi_n\|_{d\sigma}^{-1} \left[\sum_{\mu=1}^{n+1} \sigma_\mu^* \pi_n(\tau_\mu^*) f(\tau_\mu^*) + R_n(\pi_n f) \right], \tag{4.2}$$

but still has degree of exactness (at least) $2n + 1$. The new weights, $\sigma_\mu^* \pi_n(\tau_\mu^*)$, however, even if all σ_μ^* are positive, alternate in sign, which somewhat detracts from the usefulness of these formulae. For Gegenbauer measures $d\sigma^{(\lambda)} = (1-t^2)^{\lambda - \frac{1}{2}}$, $\lambda \neq 0,1$, Rabinowitz [1980] shows that the degree of exactness $2n + 1$ ($2n + 2$ if n is odd) is best possible. (4.2) is exact for polynomials of degree $3n - 1$, when $\lambda = 0$, and of degree $3n + 1$, when $\lambda = 1$, both of which is again best possible. The highest precision is thus obtained for Fourier-Chebyshev coefficients of the second kind.

Finite element and projection methods frequently rely on numerical integration but so far, Gauss-Kronrod formulae, unlike the Gauss formulae, have been shunned. An exception is Bellen [1980], who uses them in his "extended collocation-least squares" method.

Acknowledgment. The author is indebted to Professor P. Rabinowitz for providing additional references, particularly on multidimensional integration.

References

Akrivis, G. and Förster, K.–J. [1984]: *On the definiteness of quadrature formulae of Clenshaw-Curtis type,* Computing *33,* 363–366.

Atkinson, K.E. [1978]: *An Introduction to Numerical Analysis,* Wiley, New York, 1978.

Baillaud, B. and Bourget, H. [1905]: *Correspondance d'Hermite et de Stieltjes* I, II. Gauthier-Villars, Paris.

Baratella, P. [1979]: *Un' estensione ottimale della formula di quadratura di Radau,* Rend. Sem. Mat. Univ. e Politec. Torino *37,* 147–158.

Barrucand, P. [1970]: *Intégration numérique, abscisse de Kronrod-Patterson et polynomes de Szegö,* C.R. Acad. Sci. Paris *270,* 336–338.

Bellen, A. [1980]: *Metodi di proiezioni estesi,* Boll. Un. Mat. Ital. Suppl., no. *1,* 239–251.

Bellen, A. and Guerra, S. [1982]: *Su alcune possibili estensioni delle formule di quadratura gaussiane,* Calcolo *19,* 87–97.

Berntsen, J. and Espelid, T.O. [1984]: *On the use of Gauss quadratures in adaptive automatic integration schemes,* BIT *24,* 239–242.

_____, _____ [1987]: *On the construction of higher degree three dimensional embedded integration rules* (abstract), in: Numerical Integration – Recent Developments, Software and Applications (P. Keast and G. Fairweather, eds.), NATO Advanced Science Institute Series, Series C: Mathematical and Physical Sciences, Vol. *203,* 173–174. D. Reidel Publ. Co., Dordrecht.

Bognár, J. [1974]: *Indefinite Inner Product Spaces,* Ergebnisse der Mathematik und ihrer Grenzgebiete, Vol. 78, Springer, New York.

Brass, H. [1977]: *Quadraturverfahren,* Vandenhoeck and Ruprecht, Göttingen.

Caliò, F. and Marchetti, E. [1985]: *A program code of an algorithm to evaluate singular integrals,* Internal Report, Dipartimento di Matematica, Politecnico di Milano.

_____, _____ [1987]: *Derivation and implementation of an algorithm for singular integrals,* Computing *38,* 235–245.

_____, Gautschi, W. and Marchetti, E. [1986]: *On computing Gauss-Kronrod quadrature formulae,* Math. Comp. *47,* 639–650.

_____, Marchetti, E. and Pizzi, G. [1984]: *Valutazione numerica di alcuni integrali con singolarità di tipo logaritmico,* Rend. Sem. Fac. Sci. Univ. Cagliari *54,* 31–40.

Clenshaw, C.W. and Curtis, A.R. [1960]: *A method for numerical integration on an automatic computer,* Numer. Math. *2,* 197–205.

Cools, R. and Haegemans, A. [1986]: *Optimal addition of knots to cubature formulae for planar regions,* Numer. Math. *49,* 269–274.

_____, _____ [1987]: *Construction of sequences of embedded cubature formulae for circular symmetric planar regions*, in: Numerical Integration – Recent Developments, Software and Applications (P. Keast and G. Fairweather, eds.), NATO Advanced Science Institute Series, Series C: Mathematical and Physical Sciences, Vol. *203*, 165–172. D. Reidel Publ. Co., Dordrecht.

Cranley, R. and Patterson, T.N.L. [1971]: *On the automatic numerical evaluation of definite integrals*, Comp. J. *14*, 189–198.

Dagnino, C. [1983]: *Extended product integration rules*, BIT *23*, 488–499.

_____ [1986]: *Extensions of some results for interpolatory product integration rules to rules not necessarily of interpolatory type*, SIAM J. Numer. Anal. *23*, 1284–1289.

_____ and Fiorentino, C. [1983]: *A Fortran code for the computation of nodes and weights of extended Gaussian rules*, Internal Report, Dipartimento di Matematica, Politecnico di Torino.

_____, _____ [1984]: *Computation of nodes and weights of extended Gaussian rules*, Computing *23*, 271–278.

Davis, P.J. and Rabinowitz, P. [1954]: *On the estimation of quadrature errors for analytic functions*, Math. Tables Aids Comput. *8*, 193–203.

_____, _____ [1984]: *Methods of Numerical Integration*, 2nd ed., Academic Press, Orlando, Florida.

Elhay, S. and Kautsky, J. [1984]: *A method for computing quadratures of the Kronrod Patterson type*, Austral. Comput. Sci. Comm. *6*, no. *1*, 15.1–15.9. Department of Computer Science, University of Adelaide, Adelaide, South Australia.

_____, _____ [1985]: *IQPACK – Fortran subroutines for the weights of interpolatory quadratures*, School of Mathematical Sciences, The Flinders University of South Australia.

Evans, G.A., Hyslop, J. and Morgan, A.P.G. [1983]: *An extrapolation procedure for the evaluation of singular integrals*, Internat. J. Comput. Math. *12*, 251–265.

_____, Forbes, R.C. and Hyslop, J. [1983]: *Polynomial transformations for singular integrals*, Internat. J. Comput. Math. *14*, 157–170.

Gautschi, W. [1981]: *A survey of Gauss-Christoffel quadrature formulae*, in: E.B. Christoffel (P.L. Butzer and F. Fehér, eds.), 72–147. Birkhäuser, Basel.

_____, Notaris, S. [submitted]: *An algebraic study of Gauss-Kronrod quadrature formulae for Jacobi weight functions*.

_____, _____ [in preparation]: *Newton's method and Gauss-Kronrod quadrature*.

_____, Rivlin, T.J. [submitted]: *A family of Gauss-Kronrod quadrature formulae*.

Genz, A.C. and Malik, A.A. [1980]: *Algorithm 019 – Remarks on algorithm 006: An adaptive algorithm for numerical integration over an N-dimensional rectangular region*, J. Comput. Appl. Math. *6*, 295–302.

_____, _____ [1983]: *An imbedded family of fully symmetric numerical integration rules*, SIAM J. Numer. Anal. *20*, 580–588.

Geronimus, J. [1930]: *On a set of polynomials*, Ann. of Math. *31*, 681–686.

Golub, G.H. and Welsch, J.H. [1969]: *Calculation of Gauss quadrature rules*, Math. Comp. *23*, 221–230.

Imhof, J.P. [1963]: *On the method for numerical integration of Clenshaw and Curtis*, Numer. Math. *5*, 138–141.

Jacobi, C.G.J. [1826]: *Ueber Gauβ neue Methode, die Werthe der Integrale näherungsweise zu finden*, J. Reine Angew. Math. *1*, 301–308.

Kahaner, D.K. [1987]: *Personal communication.*

_____, Monegato, G. [1978]: *Nonexistence of extended Gauss-Laguerre and Gauss-Hermite quadrature rules with positive weights*, Z. Angew. Math. Phys. *29*, 983–986.

_____, Waldvogel, J. and Fullerton, L.W. [1982]: *Addition of points to Gauss-Laguerre quadrature formulas*, IMSL Tech. Rep. Series, 8205. IMSL, Houston.

_____, _____, _____ [1984]: *Addition of points to Gauss-Laguerre quadrature formulas*, SIAM J. Sci. Stat. Comput. *5*, 42–55.

Kautsky, J. and Elhay S. [1982]: *Calculation of the weights of interpolatory quadratures*, Numer. Math. *40*, 407–422.

_____, _____ [1984]: *Gauss quadratures and Jacobi matrices for weight functions not of one sign*, Math. Comp. *43*, 543–550.

Kronrod, A.S. [1964a]: *Integration with control of accuracy* (Russian), Dokl. Akad. Nauk SSSR *154*, 283–286.

_____ [1964b]: *Nodes and Weights for Quadrature Formulae. Sixteen-place Tables* (Russian). Izdat "Nauka", Moscow. [English translation: Consultants Bureau, New York, 1965.]

Laurie, D.P. [1982]: *Algorithm 584 – CUBTRI: Automatic cubature over a triangle*, ACM Trans. Math. Software *8*, 210–218.

_____ [1985]: *Practical error estimation in numerical integration*, J. Comput. Appl. Math. *12 & 13*, 425–431.

Malik, A.A. [1980]: *Some new fully symmetric rules for multiple integrals with a variable order adaptive algorithm*, Ph.D. thesis, University of Kent, Canterbury.

Markov, A. [1885]: *Sur la méthode de Gauss pour le calcul approché des intégrales*, Math. Ann. *25*, 427–432.

Monegato, G. [1976]: *A note on extended Gaussian quadrature rules*, Math. Comp. *30*, 812–817.

_____ [1978a]: *Positivity of the weights of extended Gauss-Legendre quadrature rules*, Math. Comp. *32*, 243–245.

_____ [1978b]: *Some remarks on the construction of extended Gaussian quadrature rules*, Math. Comp. *32*, 247–252.

_____ [1979]: *An overview of results and questions related to Kronrod schemes*, in: Numerische Integration (G. Hämmerlin, ed.), ISNM *45*, 231–240. Birkhäuser, Basel.

_____ [1980]: *On polynomials orthogonal with respect to particular variable-signed weight functions*, Z. Angew. Math. Phys. *31*, 549–555.

_____ [1982]: *Stieltjes polynomials and related quadrature rules*, SIAM Review *24*, 137–158.

_____ [1987]: *Personal communication.*

_____, Palamara Orsi, A. [1985]: *On a set of polynomials of Geronimus*, Boll. Un. Mat. Ital. *B* (6) *4*, 491–501.

Mysovskih, I.P. [1964]: *A special case of quadrature formulae containing preassigned nodes* (Russian), Vesci Akad. Navuk BSSR Ser. Fiz.-Tehn. Navuk, No. *4*, 125–127.

Neumann, G. [1982]: *Boolesche interpolatorische Kubatur*, Ph.D. thesis, Universität GH Siegen.

Patterson, T.N.L. [1968a]: *The optimum addition of points to quadrature formulae*, Math. Comp. *22*, 847–856. Loose microfiche suppl. C1–C11. [Errata, ibid. *23* (1969), 892.]

_____ [1968b]: *On some Gauss and Lobatto based integration formulae*, Math. Comp. *22*, 877–881. Loose microfiche suppl. D1–D5.

_____ [1973]: *Algorithm 468 – Algorithm for automatic numerical integration over a finite interval*, Comm. ACM *16*, 694–699.

Piessens, R. [1969]: *New quadrature formulas for the numerical inversion of the Laplace transform*, BIT *9*, 351–361.

_____ [1973]: *An algorithm for automatic integration*, Angew. Informatik, Heft *9*, 399–401.

_____, Branders, M. [1974]: *A note on the optimal addition of abscissas to quadrature formulas of Gauss and Lobatto type*, Math. Comp. *28*, 135–139. Suppl., ibid., 344–347.

_____, de Doncker-Kapenga, E., Überhuber, C.W. and Kahaner, D.K. [1983]: *QUADPACK: A Subroutine Package for Automatic Integration.* Springer Series in Computational Mathematics *1*. Springer, Berlin.

Rabinowitz, P. [1980]: *The exact degree of precision of generalized Gauss-Kronrod integration rules*, Math. Comp. *35*, 1275–1283. [Corrigendum: ibid. *46* (1986), 226 footnote.]

_____ [1983]: *Gauss-Kronrod integration rules for Cauchy principal value integrals,* Math. Comp. *41,* 63–78. [Corrigenda: ibid. *45* (1985), 277.]

_____ [1986a]: *A stable Gauss-Kronrod algorithm for Cauchy principal-value integrals,* Comput. Math. Appl. *12B,* 1249–1254.

_____ [1986b]: *On the definiteness of Gauss-Kronrod integration rules,* Math. Comp. *46,* 225–227.

_____, Kautsky, J., Elhay, S. and Butcher, J.C. [1987]: *On sequences of imbedded integration rules,* in: Numerical Integration – Recent Developments, Software and Applications (P. Keast and G. Fairweather, eds.), NATO Advanced Science Institute Series, Series C: Mathematical and Physical Sciences, Vol. *203,* 113–139. D. Reidel Publ. Co., Dordrecht.

Ramskiĭ, Ju. S. [1974]: *The improvement of a certain quadrature formula of Gauss type* (Russian), Vyčisl. Prikl. Mat. (Kiev) Vyp. *22,* 143–146.

Squire, W. [1970]: *Integration for Engineers and Scientists,* American Elsevier, New York.

Struble, G.W. [1963]: *Orthogonal polynomials – Variable-signed weight functions,* Numer. Math. *5,* 88–94.

Szegö, G. [1935]: *Über gewisse orthogonale Polynome, die zu einer oszillierenden Belegungsfunktion gehören,* Math. Ann. *110,* 501–513. [Collected Papers (R. Askey, ed.), Vol. 2, 545–557.]

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

THE HOLOGRAPHIC TRANSFORM

WALTER SCHEMPP

ABSTRACT: The basic idea of holography is to record analog signals as complex-valued functions on the (complexified) time-frequency plane. The holographic transform is a sesquilinear integral transformation which performs a planar encoding of the time and the frequency domains of signals simultaneously by means of interference patterns in the holographic plane. The 'frozen' interference patterns are recorded in the holographic plane by the hologram. The phase differences between the reference wave and the signal waves may be decoded by the coherent light of a laser beam in order to reconstruct the three-dimensional object from the planar hologram. - The present paper establishes an analog of the Paley-Wiener theorem for the holographic transform. Moreover, the holographic transform of the Hermite (or oscillator wave) functions is calculated explicitly in terms of Laguerre and Poisson-Charlier polynomials, and a series of holographic identities for digital signals are established. As a result, new identities for theta-null values are popping up. The energy preserving invariants of the holographic identities are classified by the ornamental groups (= dihedral groups D_m under the crystallographic restriction $m \in \{1,2,3,4,6\}$) via the elliptic Möbius transforms of the holographic plane $\mathbf{C}$. The orbits of the plane crystallographic groups D_m ($m \in \{2,3,4,6\}$) in the holographic plane $\mathbf{C}$ admit far-reaching applications to computerized holography, information theory, and neuromathematics.

0. CONTENTS

1. THE PERFECT LOW-PASS FILTER SINC

Recall the Paley-Wiener theorem which is at the basis of the classical sampling theorem.

Theorem 1 (Paley-Wiener). Let ψ denote an entire holomorphic function such that

$$\int_{\mathbf{R}} |\psi(x)|^2 dx < +\infty$$

and the estimate

$$|\psi(z)| \leq C\, e^{2\pi A|z|} \qquad (z \in \mathbf{C})$$

holds for positive constants A and C. Then there exists a function $\Psi \in L^2(-A,+A)$ such that

$$\psi(z) = \int_{-A}^{+A} \Psi(t) e^{2\pi i z t} dt \qquad (z \in \mathbf{C})$$

("finite Fourier cotransform" of Ψ).

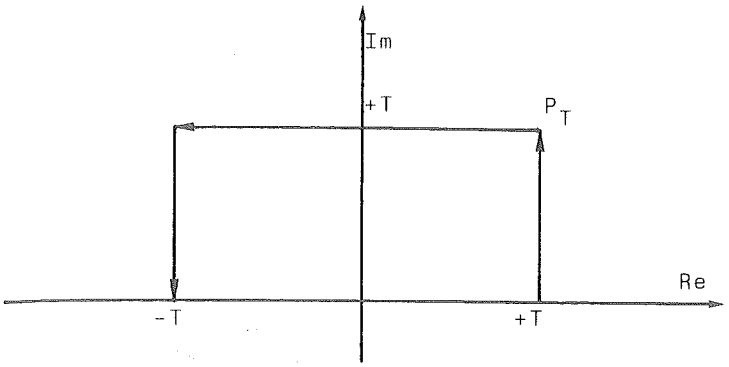
Proof. In order to establish that $\Psi = \mathscr{F}_{\mathbf{R}}\psi$ vanishes almost everywhere outside the compact interval $[-A,+A]$ of the real line $\mathbf{R}$, it will be sufficient by Cauchy's theorem to prove

$$\lim_{T \to +\infty} I_t = 0 \qquad (|t| > A)$$

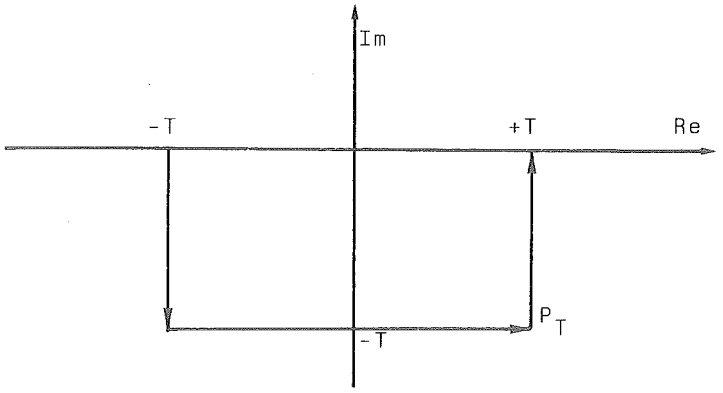
where the compact path P_T of the complex contour integral

$$I_t = \int_{P_T} \psi(z) e^{-2\pi i t z} dz \qquad (T > 0)$$

is defined in the following way:



$t < -A$



$t > +A$

The Phragmén-Lindelöf principle (a far-reaching generalization of the maximum modulus principle) implies that an entire holomorphic function of exponential type that is bounded on a line must be bounded on every parallel line in $\mathbf{C}$. It follows

$$|\psi(x+iT)| \leq Me^{2\pi AT} \qquad (x \in \mathbf{R})$$

where $M > 0$ is an appropriate constant. Without loss of generality, suppose $t < -A$. Then this estimate shows that the part of the complex contour integral I_t that belongs to the horizontal line of P_T vanishes as $T \to +\infty$. The line integrals belonging to the vertical parts of P_T can be

handled in a similar way. Indeed, consider the right vertical line of the path P_T. Then the corresponding line integral admits in absolute value an estimate by

$$\int_0^T e^{2\pi ty} |\psi(T+iy)| \, dy = \int_0^{T'} e^{2\pi ty} |\psi(T+iy)| \, dy + \int_{T'}^T e^{2\pi ty} |\psi(T+iy)| \, dy$$

where $T' \in]0,T[$. Another Phragmén-Lindelöf argument shows that

$$\lim_{T \to +\infty} \psi(T+iy) = 0$$

holds uniformly in $y \in [0,T']$. Consequently, we have

$$\lim_{T \to +\infty} \int_0^{T'} e^{2\pi ty} |\psi(T+iy)| \, dy = 0 \qquad (T' > 0).$$

Again by appealing to the Phragmén-Lindelöf principle, we conclude the estimate

$$\int_{T'}^T e^{2\pi ty} |\psi(T+iy)| \, dy \leq M \int_{T'}^T e^{2\pi(t+A)y} dy = \frac{M}{2\pi(t+A)} \left(e^{2\pi(t+A)T} - e^{2\pi(t+A)T'}\right)$$

Since $t < -A$, the last terms approach zero as $T' \to +\infty$ and $T \to +\infty$.-

The preceding proof can be traced back to lectures given by G.H. Hardy. For details of the simplified version, see the monograph by Boas [1].

The complex vector space $\mathcal{PW}(\mathbf{C})$ of all entire holomorphic functions of exponential type at most $(A \leq \frac{1}{2})$ that are square integrable along the real axis $\mathbf{R}$ forms a complex Hilbert space under the standard scalar product

$$\langle \psi | \phi \rangle = \int_{\mathbf{R}} \psi(x) \bar{\phi}(x) dx .$$

Let $\mathbf{T}$ denote the compact circle group. Then the Fourier transform $\mathcal{F}_{\mathbf{R}}$ is an isometric isomorphism of the Paley-Wiener space $\mathcal{PW}(\mathbf{C})$ onto the complex Hilbert space $L^2(\mathbf{T})$. By taking the Fourier cotransform $\bar{\mathcal{F}}_{\mathbf{R}}$ of the modes $e^{2\pi i \mu t}$ $(\mu \in \mathbf{Z})$ it follows that the sequence of functions

$$\operatorname{sinc}(z-\mu) = \begin{cases} \dfrac{\sin \pi(z-\mu)}{\pi(z-\mu)} & (z \neq \mu) \\ 1 & (z = \mu) \end{cases}$$

forms a Hilbert basis of $\mathcal{PW}(\mathbf{C})$. Accordingly each function $\psi \in \mathcal{PW}(\mathbf{C})$ admits a unique expansion of the form

$$\psi(z) = \sum_{\mu \in \mathbf{Z}} c_\mu \operatorname{sinc}(z-\mu) \quad (z \in \mathbf{C})$$

with $\|\psi\|^2 = \sum_{\mu \in \mathbf{Z}} |c_\mu|^2$. It follows $c_\mu = \psi(\mu)$ for all $\mu \in \mathbf{Z}$ and therefore we established the so-called sampling theorem.

Theorem 2 (Whittaker-Nyquist-Shannon-Kotel'nikov). A function $\psi \in \mathcal{PW}(\mathbf{C})$ can be recaptured from its values at the integers by the cardinal series:

$$\psi(z) = \sum_{\mu \in \mathbf{Z}} \psi(\mu) \operatorname{sinc}(z-\mu) \qquad (z \in \mathbf{C}).$$

The cardinal series is uniformly convergent in each horizontal strip in $\mathbf{C}$.

In terms of electrical engineering, a band-limited function ψ can be recovered from its equidistant samples by passing the data samples $(\psi(\mu))_{\mu \in Z}$ through a perfect low-pass filter. Since voice and video form band-limited signals, the sampling theorem is at the basis of digital signal processing. The scaled sinc-function serves as a perfect low-pass filter.

Example: CD-ROM (=Compact Disc Read Only Memory) for linear sequential digital signal processing. The encoding process is normally based on CIRC (= Cross-Interleaved Reed-Solomon Code).

Corollary 1. For all functions ψ and ϕ in $\mathcal{PW}(\mathbf{C})$ the sesquilinear quadrature formula

$$\sum_{n \in Z} \psi(n)\bar{\phi}(n) = \int_R \psi(x)\bar{\phi}(x)dx$$

holds.

Corollary 2. The complex Hilbert space $\mathcal{PW}(\mathbf{C})$ admits the reproducing kernel

$$(z,w) \longrightarrow \text{sinc } (z-\bar{w}).$$

For all functions $\psi \in \mathcal{PW}(\mathbf{C})$ the integral representation

$$\psi(z) = \int_R \psi(t) \text{ sinc } (t-z)\ dt$$

is valid for all $z \in \mathbf{C}$.

For a survey of the Whittaker-Nyquist-Shannon-Kotel'nikov sampling theorem, the reader is referred to the articles by Butzer [3], and Higgins [7]. Higgins also reviews some of the mathematics connected with the cardinal series and traces the origins of the result to before Whittaker. Also see the paper [21] for a proof of the sampling theorem via harmonic analysis on the compact Heisenberg nilmanifold.

As a final application of the Paley-Wiener theorem, we establish the following result due to S.N. Bernstein (1923).

Theorem 3 (Bernstein's inequality). Let $\psi \in \mathcal{PW}(\mathbf{C})$ - then

$$\|\psi' | \mathbf{R}\|_\infty \leq \pi \|\psi|\mathbf{R}\|_\infty .$$

Proof. Apply Theorem 1 to the entire holomorphic function

$$\phi_\varepsilon : z \longrightarrow \psi(z)\operatorname{sinc} \varepsilon z \qquad (\varepsilon > 0)$$

and observe that $\lim_{\varepsilon \to 0+} \phi_\varepsilon'(x) = \psi(x)$ holds for all $x \in \mathbf{R}$.-

Thus the derivative of a band-limited function on the real line $\mathbf{R}$ cannot get too large compared with the value of the function. This constraint is a fundamental one which has strong impact to vision. See Marr [12].

2. HOLOGRAPHY

The reasoning of the preceding section depends upon the duality of the complex Hilbert spaces

$$L^2(\mathbf{R}) \text{ and } L^2(\mathbf{R})$$

or

$$\mathcal{PW}(\mathbf{C}) \text{ and } L^2(\mathbf{T})$$

performed by the (linear) Fourier transform

$$\psi \longrightarrow \mathcal{F}_{\mathbf{R}}\psi.$$

From the physical point of view, however, the separation of the time and the frequency domains of (band-limited) signals is artificial. Moreover, it leads to serial algorithms which are not very efficient ways of signal processing.

The holography or wave front reconstruction (cf. Gabor [5]) is based on the following main idea: Consider for parallel signal processing the wave functions $\psi \in L^2(\mathbf{R})$ and their Fourier transformed versions $\mathscr{F}_{\mathbf{R}}\psi \in L^2(\mathbf{R})$ simultaneously.

From the mathematical point of view, the simultaneous encoding of time and frequency in the holographic plane can be performed by introducing the quadratic Fourier transform $H(\psi;.,.)$ of $\psi \in L^2(\mathbf{R})$ according to the prescription

$$H(\psi;x,y) = \int_{\mathbf{R}} \psi(t+x)\bar{\psi}(t)e^{2\pi iyt}dt$$

with $(x,y) \in \mathbf{R} \oplus \mathbf{R}$. If $\phi \in L^2(\mathbf{R})$, the sesquilinear analog reads as follows:

$$H(\psi,\phi;x,y) = \int_{\mathbf{R}} \psi(t+x)\bar{\phi}(t)e^{2\pi iyt}dt$$

Definition. The cross-correlator

$$L^2(\mathbf{R}) \times L^2(\mathbf{R}) \ni (\psi,\phi) \rightsquigarrow H(\psi,\phi;.,.)$$

is called the sesquilinear holographic transform. Its restriction to the diagonal, i.e., the corresponding autocorrelator, is called the quadratic holographic transform.

Key observation: Let $A(\mathbf{R})$ denote the three-dimensional real Heisenberg two-step nilpotent Lie group with one-dimensional center Z [23]. The projection $A(\mathbf{R})/Z$ of $A(\mathbf{R})$ along Z induces a symplectic structure on the plane $\mathbf{R} \oplus \mathbf{R}$ and a twisted convolution product on $L^2(\mathbf{R} \oplus \mathbf{R})$. The infinite dimensional, topologically irreducible, unitary, linear representations of $A(\mathbf{R})$ are square integrable mod Z. The sesquilinear holographic transform $H(\psi,\phi;.,.)$ coincides with the projection of the matrix coefficient of the linear Schrödinger re-

presentation U_1 of $A(\mathbf{R})$ defined by $\psi \in L^2(\mathbf{R})$ and $\phi \in L^2(\mathbf{R})$ along Z to the holographic plane [24],[25],[26]. The coadjoint orbit associated with U_1 under the Kirillov correspondence carries the symplectic form $(X,X') \rightarrowtail \det(X,X')$ and is isomorphic to the holographic plane by the exponential mapping.

Obviously the quadratic holographic transform satisfies the "peak property"

$$H(\psi;0,0) = \| \psi \|^2.$$

By virtue of the Cauchy-Schwarz-Bunjakovsky inequality, the sesquilinear holographic transform satisfies the estimate

$$H(\psi;\phi;x,y) \leq \| \psi \|.\| \phi \| \qquad ((x,y) \in \mathbf{R} \oplus \mathbf{R})$$

for all $\psi, \phi \in L^2(\mathbf{R})$. More important is the following result:

Theorem 4. For all functions ψ', ϕ' and ψ, ϕ in $L^2(\mathbf{R})$ the orthogonality relations

$$\int_{\mathbf{R}\oplus\mathbf{R}} \int H(\psi',\phi';x,y)\bar{H}(\psi,\phi;x,y)dxdy = \langle\psi'|\psi\rangle\langle\phi|\phi'\rangle$$

are valid.

As a consequence the following analog of the classical Paley-Wiener theorem (Theorem 1 supra) obtains.

Corollary. The sesquilinear holographic transform

$$\psi \otimes \phi \rightarrowtail H(\psi,\phi;.,.)$$

extends to an isometry of $L^2(\mathbf{R}) \hat{\otimes}_2 L^2(\mathbf{R})$ to the complex Hilbert space of Hilbert-Schmidt operators K on $L^2(\mathbf{R})$ realized as kernel operators

$$K\psi(x) = \int_{\mathbf{R}} k(x,y)\psi(y)dy \qquad (\psi \in L^2(\mathbf{R}))$$

with kernels $k \in L^2(\mathbf{R} \oplus \mathbf{R})$.

It is known (see Segal [27]) that the kernel k takes the form

$$k_f(x,y) = ({}_2\bar{\mathscr{F}}_{\mathbf{R}} f)(x-y,y) \qquad ((x,y) \in \mathbf{R} \oplus \mathbf{R})$$

where $f \in L^2(\mathbf{R} \oplus \mathbf{R})$ and ${}_2\bar{\mathscr{F}}_{\mathbf{R}}$ denotes the partial Fourier co-transform with respect to the second variable of the holographic plane. The bijective linear mapping

$$L^2(\mathbf{R} \oplus \mathbf{R}) \ni f \rightarrowtail k_f \in L^2(\mathbf{R} \oplus \mathbf{R})$$

is the Weyl transform. It gives rise to the natural Hilbert-Schmidt extension of the sesquilinear holographic transform and hence to the following result:

Theorem 5. A hologram generated on the holographic plane $\mathbf{R} \oplus \mathbf{R}$ by the Weyl transform

$$f \rightarrowtail k_f$$

acts by the Hilbert-Schmidt extension of the holographic transform as a linear spatial filter in a coherent optical system.

In Section 6 infra the preceding result will be used to point out an algorithm for generating sampled Fourier transform holograms.

3. RADIALITY

The property of the quadratic holographic transform $H(\psi;.,.)$ to form a radial function on the holographic plane $\mathbf{R} \oplus \mathbf{R}$ implies a serious restriction on the wave function $\psi \in L^2(\mathbf{R})$.

Theorem 6. Let $\psi \in L^2(\mathbf{R})$ be given and suppose that its quadratic holographic transform $H(\psi;.,.)$ is a radial function on the holographic plane $\mathbf{R} \oplus \mathbf{R}$. Then

$$\psi = \xi_n H_n$$

where $\xi_n \in \mathbf{C}$ is a constant and H_n is the Hermite function of degree $n \geq 0$.

4. SOME ORTHOGONAL POLYNOMIALS

a) Recall the definition of the Hermite functions

$$H_n(x) = e^{-\frac{1}{2}x^2} h_n(x) \qquad (x \in \mathbf{R})$$

where h_n denotes the Hermite polynomial of degree $n \geq 0$ satisfying the orthogonality relation

$$\int_{\mathbf{R}} h_n(x) h_m(x) e^{-x^2} dx = \delta_{nm}.$$

b) Let $L_n^{(\alpha)}$ denote the Laguerre function, i.e.,

$$L_n^{(\alpha)}(x) = e^{-\frac{1}{2}x} l_n^{(\alpha)}(x) \qquad (x \in \mathbf{R})$$

where $l_n^{(\alpha)}$ denotes the Laguerre polynomial of degree $n \geq 0$ and order $\alpha > -1$ satisfying the orthogonality relations

$$\int_0^\infty l_n^{(\alpha)}(x) l_m^{(\alpha)}(x) x^\alpha e^{-\frac{1}{2}x} dx = \delta_{nm}.$$

c) Finally, the Charlier-Poisson polynomials $c_n(.;a)$ on $\mathbf{N}$ of degree $n \geq 0$ and parameter value $a > 0$ are needed. The polynomials $c_n(.;a)$ satisfy the discrete orthogonality relations

$$\sum_{x \in \mathbf{N}} c_n(x;a)c_m(x;a)\frac{a^x}{x!} = e^a a^n n!\,\delta_{nm}$$

where $a > 0$.

Using the preceding orthogonality relations, we get the following result:

Theorem 7. The holographic transform of the Hermite functions reads in terms of Laguerre functions and Charlier-Poisson polynomials as follows:

$$H(H_m,H_n;x,y) = \sqrt{\frac{n!}{m!}}\,(\sqrt{\pi}(x+iy))^{m-n} L_n^{(m-n)}(\pi(x^2+y^2))$$

$$= \frac{(-1)^n}{\sqrt{m!n!}}\, z^{m-n}\,|z|^{2n}\, e^{-\frac{1}{2}|z|^2}\, c_n(m;|z|^2)$$

where $m \geq n \geq 0$ and $z = \sqrt{\pi}(x+iy) \in \mathbf{C}$.

5. THE HOLOGRAPHIC IDENTITIES

In the preceding theorem we identified the holographic plane $\mathbf{R} \oplus \mathbf{R}$ with the complex plane $\mathbf{C}$. If we restrict the holographic transform $H(\psi,\phi;.,.)$ to the quadratic lattice $\mathbf{Z} \oplus \mathbf{Z}$ in $\mathbf{R} \oplus \mathbf{R}$, i.e., to the lattice $\mathbf{Z}[i]$ of Gaussian integers in $\mathbf{C}$ we get

Theorem 8. Let ψ and ϕ be elements of $L^2(\mathbf{R})$ then the holographic identity

$$\sum_{(\mu,\nu) \in \mathbf{Z}\oplus\mathbf{Z}} H(\psi;\mu,\nu).\bar{H}(\phi;\mu,\nu) = \sum_{(\mu,\nu) \in \mathbf{Z}\oplus\mathbf{Z}} |H(\psi,\phi;\mu,\nu)|^2$$

is valid.

On the left hand side the signal terms occur whereas the right hand side encompass the interference terms. This explains the name. It can be established that the holographic identity implies the classical sampling theorem as a special case. However, it implies more.

In view of Theorem 7 we get by choosing for ψ and ϕ the Hermite functions:

Theorem 9. Let m,n be integers such that $m \geq n \geq 0$ - then the identity

$$\sum_{(\mu,\nu) \in Z \oplus Z} L_m^{(0)}(\pi(\mu^2+\nu^2)) \cdot L_n^{(0)}(\pi(\mu^2+\nu^2)) =$$

$$\frac{n!}{m!}\pi^{m-n} \sum_{(\mu,\nu) \in Z \oplus Z} (\mu^2+\nu^2)^{m-n} (L_n^{(m-n)}(\pi(\mu^2+\nu^2)))^2$$

holds.

The theta function is defined by means of the Fourier series

$$\vartheta(z,\tau) = \sum_{\mu \in Z} e^{-\pi\mu^2\tau} e^{2\pi i\mu z}$$

which is normally convergent in the domain $\{(z,\tau) \in \mathbb{C}^2 \mid \mathrm{Re}\,\tau > 0\}$. It was C.G.J. Jacobi (1804-1851) who invented the theta-function in the 1820s. Since then it has been used in many investigations by generations of number theorists. It is involved in many fascinating identities of number-theoretical and combinatorial import, and it provides one of the most effective ways to construct automorphic forms. According to D. Newman (Lecture in honour of A. Sharma, Edmonton 1986) the theta-function actually belongs to theology, and not to mathematics. In the early 1960s André Weil, inspired especially by the work of C.L. Siegel, provided a representation-

theoretic foundation for the theory of theta-function. See the classical paper by Weil [30]. He found that the theta-function is intimately connected with the metaplectic (or oscillator) representation, which forms a most singular projective unitary linear representation of the symplectic group. This representation arises by virtue of the existence of an action by automorphisms of the symplectic group on the Heisenberg two-step nilpotent Lie group A(R) mentioned in Section 2 supra. Moreover, André Weil showed the intimate relationship to the law of quadratic reciprocity (cf.[22]). The preceding theorem implies the following identities for the odd powers of π which can be considered as identities for the classical theta-function ("theta-null value")

$$\vartheta(\tau) = \vartheta(0,\tau) = \sum_{\mu \in Z} e^{-\pi\mu^2\tau} \qquad (\mathrm{Re}\,\tau > 0)$$

at the point $\tau = 1$ of the right half-plane.

m = 1, n = 0

$$\pi = \frac{\sum_{\mu \in Z} e^{-\pi\mu^2}}{4\sum_{\mu \in Z} \mu^2 e^{-\pi\mu^2}}$$

See Advanced Problem # 6491, Amer. Math. Monthly 92 (1985), 217.

m = 2, n = 1

$$\pi^3 = \frac{15\sum_{\mu \in Z} (8\pi^2\mu^4 - 1)e^{-\pi\mu^2}}{32\sum_{\mu \in Z} \mu^6 e^{-\pi\mu^2}}$$

$m = 3, \quad n = 2$

$$\pi^5 = \frac{45 \sum_{\mu \in Z} (16\pi^4\mu^8 - 140\pi^2\mu^4 + 21) e^{-\pi\mu^2}}{64 \sum_{\mu \in Z} \mu^{10} e^{-\pi\mu^2}}$$

$m = 4, \quad n = 3$

$$\pi^7 = \frac{91 \sum_{\mu \in Z} (256\pi^6\mu^{12} - 15840\pi^4\mu^8 + 166320\pi^2\mu^4 - 25245) e^{-\pi\mu^2}}{1024 \sum_{\mu \in Z} \mu^{14} e^{-\pi\mu^2}}$$

.

.

.

See Amer. Math. Monthly 93 (1986), 822-823 and Proc. Amer. Math. Soc. 92 (1984), 103-110.

6. HOLOGRAPHIC ENCODING

A mapping of the holographic plane

$$\sigma: \mathbf{R} \oplus \mathbf{R} \longrightarrow \mathbf{R} \oplus \mathbf{R}$$

is said to be an invariant of the quadratic holographic transform $H(\psi;.,.)$, if the identity

$$H(\psi;x,y) = H(\psi_\sigma;\sigma(x,y))$$

holds for all pairs $(x,y) \in \mathbf{R} \oplus \mathbf{R}$ and all functions $\psi \in L^2(\mathbf{R})$

in such a way that the assignment $\psi \longrightarrow \psi_\sigma$ defines a unitary operator in $L^2(\mathbb{R})$.

Theorem 10. A mapping of the holographic plane

$$\sigma : \mathbb{R} \oplus \mathbb{R} \longrightarrow \mathbb{R} \oplus \mathbb{R}$$

is an invariant of H, if and only if

$$\sigma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad \det \sigma = 1$$

with real coefficients a,b,c,d, i.e., $\sigma \in SL(2,\mathbb{R})$.

In the case when σ preserves the lattice $\mathbb{Z}[i]$ and the radiality of H, the choices of σ are drastically reduced.

Theorem 11. Let σ be an invariant of the holographic identity displayed in Theorem 8 supra. Then

$$\sigma = \begin{bmatrix} \cos 2\pi\frac{k}{m} & \sin 2\pi\frac{k}{m} \\ -\sin 2\pi\frac{k}{m} & \cos 2\pi\frac{k}{m} \end{bmatrix} \qquad (0 \leq |k| \leq m-1)$$

and m satisfies the cristallographic restriction

$$m \in \{1,2,3,4,6\}.$$

Proof. Since σ preserves the lattice $\mathbb{Z}[i]$, the coefficients a,b,c,d are integers. If

$$\sigma \notin \{-id_{\mathbb{R}\oplus\mathbb{R}},\ id_{\mathbb{R}\oplus\mathbb{R}}\}$$

preserves the radiality of H, the mapping

$$z \longrightarrow \frac{az+b}{cz+d}$$

defines an elliptic Möbius transformation of the upper complex half-plane preserving $\mathbf{R}$. It follows

$$|\mathrm{tr}\ \sigma| < 2$$

and since $\mathrm{tr}\ \sigma \in \mathbf{Z}$ obviously

$$\mathrm{tr}\ \sigma \in \{-1,0,+1\}.$$

Therefore σ is a turn through $\pm\pi/3$, $\pm\pi/2$ or $\pm 2\pi/3$, and no other turn is allowed.-

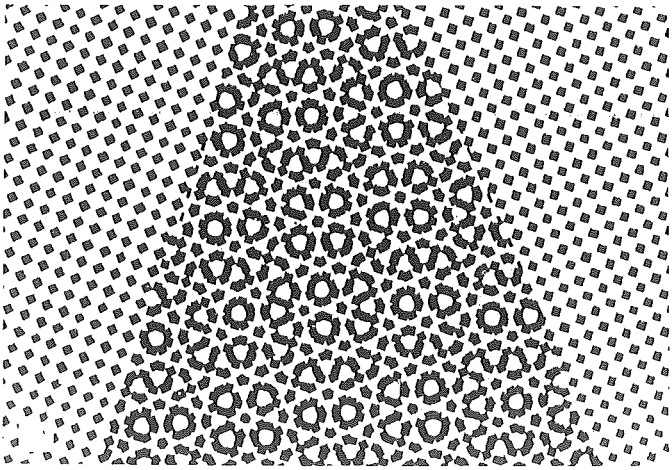
It follows that the holographic identities have the dihedral groups D_m ($m \in \{1,2,3,4,6\}$) as their groups of invariants. Nothing like a turn through $\pm\pi/5$ is possible. Only the classical planar crystal symmetries (or ornamental groups) and none of the forbidden fivefold symmetries, well-known from the theory of quasi-crystals, are allowed. For similar patterns arising in long crested wave models, see the paper by Schachter [20].

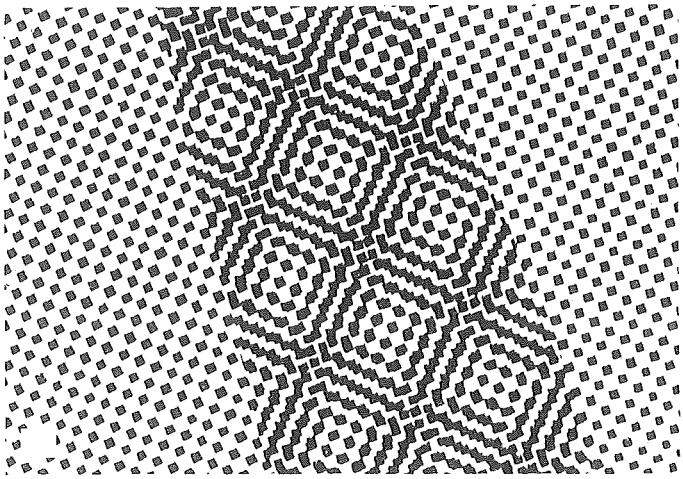
It should be observed that the dihedral groups D_m have order $2m$ and not the order m of the cyclic groups $\mathbf{Z}/m\mathbf{Z}$. Actually this fact reflects that a hologram generates two images, a real pseudoscopic image and a virtual orthoscopic image. It can be shown that the generation of orthoscopic and pseudoscopic images is at the basis of non-linear laser optics and in particular of non-linear optical phase-conjugation [24].

The figures on the next page show two superpositions of patterns formed by squares ($m = 4$).

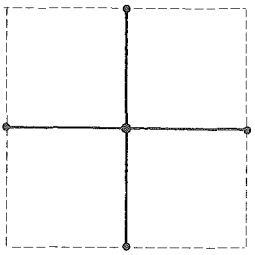
7. COMPUTERIZED HOLOGRAPHY

The periodic tilings of the holographic plane $\mathbf{R} \oplus \mathbf{R}$ enable to implement numerically various discretizations of the kernel

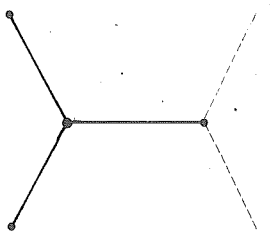




function k_f, the image of $f \in L^2(\mathbb{R} \oplus \mathbb{R})$ under the Weyl transform. In this way an algorithm arises by Theorem 5 supra to generate computer holograms. One way to do this is to compute in a first step by the FFT algorithm the Fourier transform $f = \mathcal{F}_{\mathbb{R} \oplus \mathbb{R}}\, g$ of the "two-dimensional image" g on the lattice with group D_m ($m \in \{2,3,4,6\}$) of invariants and then the second step is to compute the kernel k_f on the grid. In the case m = 4 we get Lee's encoding scheme of generating sampled Fourier transform holograms by decomposing the complex-valued functions to be synthesized into four components [10], [11]. Four times more samples are used along one direction than the other are required by this encoding technique.



In the case m = 6 we get Burckhardt's encoding scheme of generating sampled Fourier transform holograms by decomposing the complex valued functions to be synthesized into three components [2]. Also see Yaroslavskii [29].



For processing hexagonally sampled two-dimensional signals, the reader should consult Mersereau [13]. A similar procedure is possible in the cases m = 3 and m = 2.

3. THE NEURAL HOLOGRAPHIC MODEL

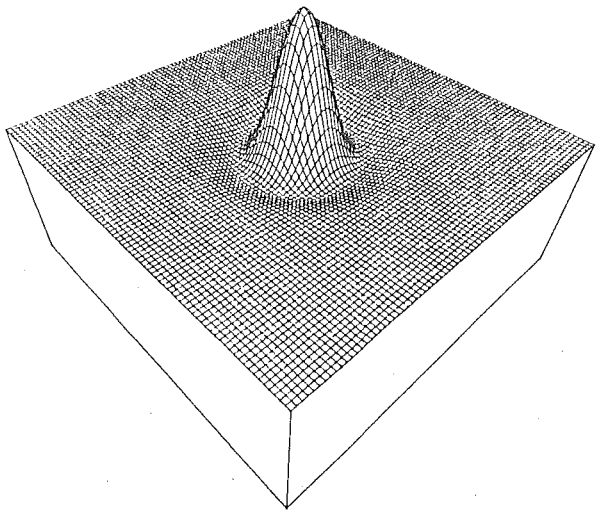
A growing number of theorists in the field of neurophysiology have invoked the principles of holography to explain certain aspects of brain function. One of the best established facts about brain mechanisms and memory is that large destructions within a neural system do not seriously impair its function. Indeed, the pioneering experiments by Lashley [9] showed that 80% or more of the visual cortex of a rat could be damaged without loss of the ability to correctly respond to patterns. Moreover, Robert Galambos (see Galambos, Norton, and Frommer [6]) has surgically removed as much as 98% of the optic tracts of cats with little effect on visual recognition behaviour. These and similar tests on monkeys and even men (performed during neurosurgery) have been interpreted to indicate that the neural elements necessary to the recognition and recall processes must be distributed throughout the brain systems involved. The problem that then confronts neurophysiologists is essential this: how can the relationships between neural activity become distributed and stored (temporarily or permanently) by a neural network. The neural holographic model developed by P.R. Westlake, K.H. Pribram and co-workers (Pribram [15], [16], [17]; also see Pribram, Nuwer, and Baron [18], Ferguson [4]) explains the property of distributed storage. Indeed, what makes the hologram unique as a storage device is that every element of the original image is distributed by the holographic transform and the Weyl transform (cf. Theorem 5 supra) over the entire holographic plane. Aside from this property, holographic memories show large capacities, parallel processing, and content addressability for rapid recognition, associative storage for perceptual completion, and for associative recall. The holographic hypothesis is in agreement with the experimental results of Rodieck [19] who found circularly symmetric excitability profiles of visual receptive fields which are conformal to Theorems 6 and 7 supra and also with the

thematical results by Marr [12]. See the figure on the following page and also Kronauer and Zeevi [8]. Moreover, Theorem 11 supra is in agreement with the results by Welt, Aschoff, Kameda, and Brooks [28] who found that "sensory convergence into the motor (sensory) cortex is superimposed on topographically uniform output organization in radial arrays, the diameter of which is estimated to be 0.1 to 0.4 mm. Thus, neurons with fixed local receptive fields provide a radially oriented framework (a reference system) for common peripheral inputs..." More precisely, Nicolis [14] concludes from his model of thalamocortical pacemaker that "specifically cognition is manifested at the cortex as a result of a matching process between pairs of spatial-temporal patterns, each containing a great number of elemental units (neurons). In each pair, one pattern (the same for all pairs) is the unknown information; it is embodied in incoming triggers, coded either in sequences of pulses from the peripheral nervous system, or, if it comes from other areas of the central nervous system, encoded in strings of macromolecular (neuro-transmitter/hormonal) releases from presynaptic endings. The other pattern of the pair is one of the pattern/attractors created by the processor; it constitutes a prestored spatial-temporal "mosaic" embodied in a set of partly synchronized post-synaptic membrane potentials or a spatial-temporal pattern of post-synaptic membrane receptors. The coupling or cross-correlation between the above two patterns of each pair takes place dynamically via energy exchanges between equal or neighbouring frequency pairs shared by both spectra... The result of the cross-correlation in phase and amplitude determines the "degree of cognition" between the incoming and the preset or the unknown and the expected patterns..."

It follows that the holographic transform provides a rigorous basis of neuromathematics. It includes the transference of phase informations to bijective linear transformations of the

holographic plane by the metaplectic representation of the symplectic group which explains the neural encoding of signal pulses emphasized by D.H. Hubel and T.N. Wiesel as well as the parallel processing of information emphasized by F.W. Campbell and D.A. Pollen.

Finally, let us quote P. Greguss (Lecture presented at the International Conference on Holography Applications, Beijing 1986): "I would like to express my belief that the holographic concept of Gabor is as fundamental as the general relativity theorem of Einstein, and it has to be explored further for a better understanding of nature in which we live."



Acknowledgments. The author is grateful to Professors Pál Greguss (Technical University Budapest) and Yehoshua Y. Zeevi (Harvard and Technion) for stimulating discussions. Moreover, he acknowledges the constant support and constructive criticisms by Miklôs Nyari (Technical University Budapest). Finally, the hospitality of the Mathematical Research Institute at Oberwolfach is gratefully acknowledged, where parts of this work has been done.

REFERENCES

1. R.P. BOAS, JR.: Entire functions. Academic Press, New York, N.Y., 1954.

2. C.B. BURCKHARDT: A simplification of Lee's method of generating holograms by computer. Applied Optics **9** (1970), 1949, 2813.

3. P.L. BUTZER: A survey of the Whittaker-Shannon sampling theorem and some of its extensions. J. Math. Research Exposition **3** (1983), 185-212.

4. M. FERGUSON: Wirklichkeit und Wandel - Karl Pribram als Pionier der Gehirn- und Bewußtseinsforschung. In: Das holographische Weltbild (K. Wilber, Hrsg.), Scherz-Verlag, Bern, München, Wien, 1986, pp. 12-26.

5. D. GABOR: Associative holographic memories. IBM J. of Research and Development **13** (1969), 156-159.

6. R. GALAMBOS, T.T. NORTON and C.P. FROMMER: Optic tract lesions sparing pattern vision in cats. Experimental Neurology **18** (1967), 8-25.

7. J.R. HIGGINS: Five short stories about the cardinal series. Bull. (New Series) Amer. Math. Soc. **12** (1985), 45-89.

8. R.E. KRONAUER and Y.Y. ZEEVI: Reorganization and diversification of signals in vision. IEEE Trans. Syst., Man, Cybern. **13** (1985), 91-101.

9. K.S. LASHLEY: In search of the engram. In: Physiological Mechanisms in Animal Behaviour. Academic Press, New York, N.Y. 1951, pp. 112-146.

10. W.H. LEE: Sampled Fourier transform hologram generated by computer. Applied Optics **9** (1970), 639-643.

11. W.H. LEE: Computer-generated holograms: Techniques and applications. In: Progress in Optics, Vol. XVI (E. Wolf, ed.), North-Holland, Amsterdam, New York, Oxford, 1978, pp. 119-232.

12. D. MARR: Vision. W.H. Freeman, San Francisco, 1982.

13. R.M. MERSEREAU: The processing of hexagonally sampled two-dimensional signals. Proc. IEEE **67** (1979), 930-949.

14. J.S. NICOLIS: Dynamics of hierarchical systems. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1986.

15. K.H. PRIBRAM: Languages of the brain: Experimental paradoxes and principles in neuropsychology. 5th ed. Brandon House, Bronx, N.Y., 1982.

16. K.H. PRIBRAM: Worum geht es beim holographischen Paradigma? In: Das holographische Weltbild (K. Wilber, Hrsg.), Scherz-Verlag Bern, München, Wien, 1986, pp. 27-36.

17. K.H. PRIBRAM: Holography and brain function. In: Encyclopedia of neuroscience (G. Adelman, ed.), Vol. I. Birkhäuser Verlag, Boston, Basel, Stuttgart, 1987, pp. 499-500.

18. K.H. PRIBRAM, M. NUWER and R.J. BARON: The holographic hypothesis of memory structure in brain function and perception. In: Measurement, Psychophysics, and Neural Information Processing (D.H. Krantz, R.D. Luce, R.C. Atkinson, P. Suppes, eds.), W.H. Freeman, San Francisco, 1974, pp. 416-457.

19. R.W. RODIECK: Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vison Res. **5** (1965), 583-601.

20. B. SCHACHTER: Long crested wave models. Computer Graphics and Image Processing **12** (1980), 187-201. Also in: Image Modeling (A. Rosenfeld, ed.), Academic Press, New York, London, Toronto, Sydney, San Francisco, 1981, pp. 327-341.

21. W. SCHEMPP: Gruppentheoretische Aspekte der Signalübertragung und der kardinalen Interpolationssplines I. Math. Methods Appl. Sci. **5** (1983), 195-215.

22. W. SCHEMPP: Group theoretical methods in approximation theory, elementary number theory, and computational signal geometry. In: Approximation Theory V (C.K. Chui, L.L. Schumaker, J.D. Ward, eds.), Academic Press, Boston, Orlando, San Diego, New York, Austin, London, Sydney, Tokyo, Toronto, 1986, pp. 129-171.

23. W. SCHEMPP: Harmonic analysis on the Heisenberg nilpotent Lie group, with applications to signal theory. Pitman Research Notes in Math., Vol. **147**. Longman Scientific and Technical, Harlow, Essex, 1986.

24. W. SCHEMPP: Signal geometry (to appear).

25. W. SCHEMPP: The holographic transformation (to appear).

26. W. SCHEMPP: The holographic plane (to appear).

27. I.E. SEGAL: Transforms for operators and symplectic automorphisms over a locally compact abelian group. Math. Scand. **13** (1963), 31-43.

28. C. WELT, J.C. ASCHOFF, K. KAMEDA and V.B. BROOKS: Intracortical organization of cat's motorsensory neurons. In: Neuropysiological Basis of Normal and Abnormal Motor Activities (M.D. Yahr, D.P. Purpura, eds.), Raven Press, Hewlett, N.Y., 1967, pp. 255-294.

29. L.P. YAROSLAVSKII: Applied problems of digital optics. In: Advances in Electronics and Electron Physics (P.W. Hawkes, ed.), Academic Press, Orlando, San Diego, New York, Austin, London, Montreal, Sydney, Tokyo, Toronto, 1986, pp. 1-140.

30. A. WEIL: Sur certains groupes d'opérateurs unitaires. Acta Math. **111** (1964), 143-211. Also in: Collected papers, Vol. III. Springer-Verlag, New York, Heidelberg, Berlin, 1980, pp. 1-69.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

THE MOVING GRID METHOD FOR BLN PROBLEM

M. ALIĆ and R. MANGER

ABSTRACT. *We consider Godunov method for the Bardos, Leraux and Nedelec initial-boundary problem in the case of nonuniform grids. Computer code and results are also included.*

Let $f \in C^2(R)$, a_0, $a_L \in R$ and let $Q_T =]0,L[\times]0,T[$, $\tilde{Q}_T = [0,L]\times]0,T[$ for $T > 0$, $L > 0$. For $u \in BV(Q_T)$, $c \in R$ and $\varphi \in C_0^1(\tilde{Q}_T)$ we introduce the notation

$$E(u,\varphi,c) = -\int_0^L \int_0^T \{|u-c| \frac{\partial \varphi}{\partial t} + \operatorname{sign}(u-c)[f(u)-f(c)] \frac{\partial \varphi}{\partial x}\} dx\, dt +$$

$$+ \int_0^T \operatorname{sign}(c-a_0)[f(T_0u)-f(c)]\varphi(0,t)dt -$$

$$- \int_0^T \operatorname{sign}(c-a_L)[f(T_Lu) - f(c)]\varphi(L,t)dt ,$$

where $(T_0u)(t) = \lim_{x\to 0+} u(x,t)$ in $L^1(0,T)$ and $(T_Lu)(t) = \lim_{x\to 1-} u(x,t)$ in $L^1(0,T)$.

For a given $u_0 \in BV(]0,L[)$ we consider Bardos, Leroux and Nedelec problem

$$(1) \quad \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \qquad \text{in } Q_T$$

$$(2) \quad u(x,0) = u_0(x) \qquad \text{in }]0,1[$$

$$(3) \quad \min_{c\in J[T_0u(t),a_0]} \operatorname{sign}(a_0-T_0u(t))[f(T_0u(t))-f(c)] = 0$$

$$(4) \quad \min_{c\in J[T_Lu(t),a_L]} \operatorname{sign}(T_Lu(t)-a_L)[f(T_Lu(t))-f(c)] = 0 ,$$

for $t \in]0,T[$ where

$$J[\alpha,\beta] = [\min\{\alpha,\beta\}, \max\{\alpha,\beta\}].$$

DEFINITION. A function $u \in BV(Q_T)$ is a solution of the problem (1)-(4) if it satisfies the initial condition (2) almost everywhere in $]0,T[$ and if

(5) $$E(u,\varphi,c) \leq 0$$

for all $c \in \mathbf{R}$ and all non negative $\varphi \in C_0^1(Q_T)$.

Bardos, Leroux and Nedelec have proven in [1] the existence and uniqueness theorem for the above problem.

The following lemma is a fundamental one for our consideration:

LEMMA. Let $u_0 \in BV(]0,L[)$ be a step function. If $u \in BV(Q_{T'})$ is the solution of (1)-(4) and if $w \in BV(\mathbf{R}\times]0,T'[)$ is the solution of the Cauchy problem

(6) $$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x} f(w) = 0 \quad \text{in} \quad \mathbf{R}\times]0,T'[$$

(7) $$w(x,0) = \begin{cases} a_0, & x \leq 0 \\ u_0(x), & x \in]0,L[\\ a_L, & x \geq L \end{cases}$$

then

$$u = v|_{Q_T},$$

for a short time $T' > 0$.

The proof of this lemma follows from the fact that if w is short time solution of Cauchy problem (with short time T given by some Courant condition, see [6]) then w satisfies boundary conditions (3) and (4) as a solution of a Riemann problem (see [2]).

For $\delta \in]0,\delta_0[$ we consider a set of grids $\{G_\delta\}$ in Q_T where $G_\delta = \{(x_i^j, t^j)\}$ and where

$$0 = t^0 < t < \ldots < t^n = T$$

$$0 = x_0^j < x_1^j < \ldots < x_{m_j}^j = L \ .$$

We suppose that there exist positive constants C_0, C_1, k_0, k_1 such that

$$(8) \qquad k_0 \cdot \delta \leq x_{i+1}^j - x_i^j = \Delta x_i^j \leq k_1 \delta$$

$$(9) \qquad C_0 \leq \frac{\Delta^+ t^j}{\Delta x_i^j} = \frac{t^{j+1} - t^j}{\Delta x_i^j} \leq C_1$$

where

$$(10) \qquad C_1 = \frac{1}{2 \max\limits_{n \in I} |f'(u)|} \ ,$$

$I = [\min\{a_0, a_L, \inf u_0(x)\}, \max\{a_0, a_L, \sup u_0(x)\}$

(this is Courant condition!).

It follows from (8) and (9) that

$$(11) \qquad n\delta \leq C_2$$

where $C_2 = \frac{T}{C_0 k_0}$. We define a regular set of grids as a set of grids with properties (8), (9) and (10).

For the formulation of Godunov method we use the solution operator $S(t)$ for BLN problem (1)-(4) and the averaging operator A_j, $j=0,\ldots,n-1$. The operator A_j is defined for $u \in L^1(0,1)$ by the formula

$$(A_j u)(x) = \frac{1}{\Delta x_i^j} \int_{x_i^j}^{x_{i+1}^j} u(\xi) d\xi$$

for $x \in [x_i^j, x_{i+1}^j[$. We define an approximation v^δ by

$$v^\delta(x,t) = v^j(x,t) \ ,$$

for $(x,t) \in]0,L[\times]t^j, t^{j+1}[$ and $j=0,\ldots,n-1$, where

$$v^0(x,0) = A_0 u_0(x)$$

$$v^j(x,t^j) = A_j v^{j-1}(x,t^j), \quad j=1,\dots,n-1,$$

and

$$v^j(x,t) = S(t-t^j)v^j(x,t^j)$$

for $t \in [t^j,t^{j+1}]$ and $j=0,\dots,n-1$.

THEOREM. If $u \in BV(Q_T)$ is the solution of the problem (1)-(4) and $\{G_\delta\}$ a regular set of grids in Q_T then $u = \lim_{\delta\to 0} v^\delta$ in $L^\infty(0,T:L(0,L))$.

Proof. Let w^j be the solution of Cauchy problem

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \ , \text{ in } R \times [t^j,t^{j+1}] \ , \tag{12}$$

$$w(x,t^j) = \begin{cases} a_0 \ , & x \le 0 \\ v^j(x,t^j), & x \in]0,L[\\ a_L \ , & x \ge L \ . \end{cases} \tag{13}$$

For w^j and $t \in [t^j,t^{j+1}]$ we have fundamental Kružkov estimates:

$$\inf_x w^j(x,t^j) \le w^j(x,t) \le \sup_x w^j(x,t^j) \ , \tag{14}$$

$$Var[w^j(\cdot,t);R] \le Var\,[w^j(\cdot,t^j);R] \tag{15}$$

and

$$\|w^j(\cdot,t+\tau)-w^j(\cdot,t)\|_{L^1(R)} \le |\tau| \cdot L \cdot Var[w^j(\cdot\, t^j) \ ; \ R \ ] \tag{16}$$

for $t+\tau \in [t^j,t^{j+1}]$ where L is the Lipschitz constant of f on I. We define the function $\overline{v}^\delta$ by the formula

$$\overline{v}^\delta = \sum_{j=0}^{n-1} w^j \cdot \chi_{]0,L]\times[t^j,t^{j+1}[}$$

such that

$$v^\delta = \overline{v}^\delta|_{Q_T}.$$

By using the results of Lemma 3.1. and Lemma 3.2. from [5]

we obtain that for some positive constant C_1, C_2 and C_3

$$\|\bar{v}^{\delta}\|_{L^{\infty}(R^2)} \leq C_1 \,, \tag{17}$$

$$\mathrm{Var}[\,\bar{v}^{\delta}(\cdot,t);\ R] \leq C_2 \,, \tag{18}$$

and

$$\|\,\bar{v}^{\delta}(\cdot,t+\tau)-\bar{v}^{\delta}(\cdot,t)\|_{L^{\infty}(R^2)} \leq C_3(|\tau|+\delta). \tag{19}$$

Estimates (17), (18) and (19) imply that every sequence (v^{δ}) with δ tending to zero has a subsequence converging to a limit in $L^{\infty}(0,T;L(0,L))$. This limit is also in $BV(Q_T)$, by some integral criterium (see [7], C.IV.§3). By the inequality (17) there exists a subsequence (v^{δ_r}) such that $v^{\delta_r} \to v$ in $L^1(Q_T)$, $f(T_0 v^{\delta_r}) \to p$ and $f(T_L v^{\delta_r}) \to q$ weak star in $L^{\infty}(0,T)$. Similarly as in [5] it follows that $p = f(T_0 v)$ and $q = f(T_L v)$ if $\lim_{r\to\infty} \bar{t}(v^{\delta_r},\varphi,C) \leq 0$. Indeed, from inequalities $E(v^j,\psi,C) \leq 0$ for

$$\psi \in C_0([0,T]\times]t^j,t^{j+1}[\,,$$

$\psi \geq 0$, $j=0,\ldots,n-1$, $C \in R$

we obtain the inequality

$$E(v^{\delta},\varphi,c) \leq \sum_{j=0}^{n-1} \{\int_0^L |v^j(x,t^j)-c|\varphi(x,t^j)dx\} -$$
$$- \int_0^L |v^j(x,t^{j+1})- c|\varphi(x,t^{j+1})dx\}\,, \tag{20}$$

for $\varphi \in C_0(\tilde{Q}_T)$, $\varphi \geq 0$. The inequality (20) implies, as in [6] the inequality

$$E(v^{\delta},\varphi,c) \leq K\delta$$

for a positive K and we finally have

$$E(v,\varphi,c) \leq 0$$

which, because of uniqueness theorem, completes the proof of Theorem.

Define $v_i^j = v^\delta(x,t^j)$ for $x \in [x_i^j, x_{i+1}^j[$. Using the fact that v^δ is the exact solution on the strip $]0,L[\times]t^j,t^{j+1}[$ and the divergence theorem on the trapezium with vertices (t^{j+1},x_{i+1}^{j+1}), (t^{j+1},x_i^{j+1}), (t^j,x_k^j), (t^j,x_l^j) we obtain the Godunov sheme

$$(21) \qquad v_i^{j+1}\,\Delta x_i^j = \sum_{r=k}^{l-1} v_r^j \Delta x_r^j - \Delta^+ t^j [Y_{l,i+1}^j - Y_{k,i}^j] \ ,$$

where

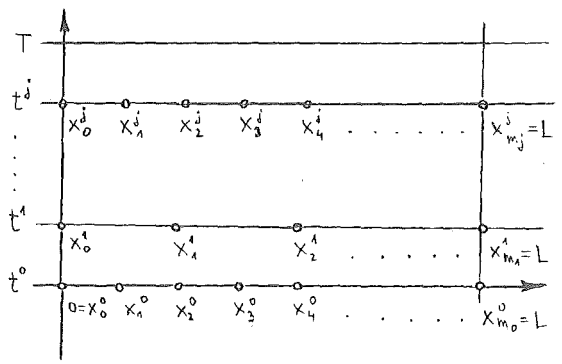
$$Y_{k,i}^j = \begin{cases} \min\limits_{u\in[v_{K-1}^j, v_K^j]} [f(u) - X_{K,i}^j u] \ , & \text{if} \quad v_{K-1}^j < v_K^j \\ \max\limits_{u\in[v_K^j, v_{K-1}^j]} [f(u) - X_{K,i}^j u] \ , & \text{if} \quad v_K^j < v_{k-1}^j \end{cases} ,$$

and where

$$X_{k,i}^j = \frac{\Delta x_i^j}{\Delta^+ t^j} \ .$$

The index k (or l) is chosen such that the line segment $(x_i^{j+1},t^{j+1}),(x_k^j,t^j)$ (or $(x_{i+1}^{j+1},t^{j+1}),(x_l^j,t^j)$) nowhere transversaly crosses any Riemann fan.

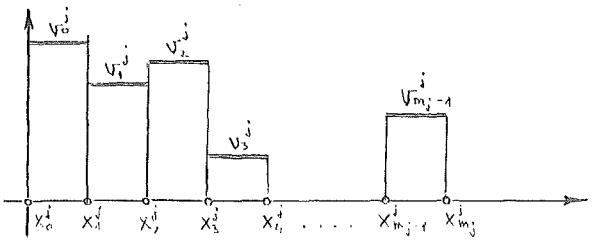
In order to test the method numerically, we have made a computer program which solves the BLN problem. Our program is only one of many possible implementations of the method. A grid of points (x_i^j,t^j), $j=0,\ldots,n$, $i=0,\ldots,m_j$ is automatically constructed: The grid covers the domain of our problem, i.e. $x_j^0=0, x_{m_j}^j=L$ $t^0=0, t^n \geq T$



The parameters n, m_j $(j=1,\ldots,n)$, $t^j (j=1,\ldots,n)$ are chosen by the program during the computation. On one time layer (for fixed j) the grid is uniform, i.e.:

$$x_1^j - x_0^j = x_2^j - x_1^j = \ldots = x_{m_j}^j - x_{m_j-1}^j = L/m_j = :h_j$$

Yet, the whole grid is still nonuniform, since m_j and $\Delta t^j = t^{j+1} - t^j$ depend on j. There is even stronger regularity among the numbers m_j: m_{j+1} can be either equal to m_j or two times greater than m_j or two times less than m_j. For given t^j, the solution $u(x,t^j)$ of the BLN problem is approximated by a step function:

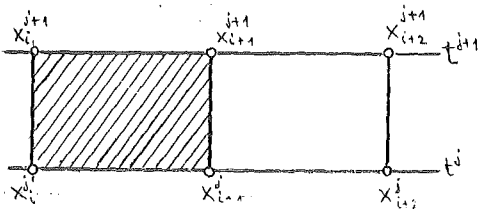


The set of all computed step functions is stored in random-access files. Additional modules of the program use these files to produce printed or plotted reports containing approximate versions of the functions $u(x,t^j)$ (for some user-specified values t^j).

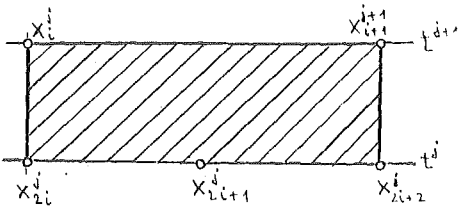
Now, we will describe essential features of the algorithm used in our program:

- The computation is advancing time layer by time layer. The grid is being constructed in the course of computation
- In order to construct the next time layer, the program selects one of three possible patterns:

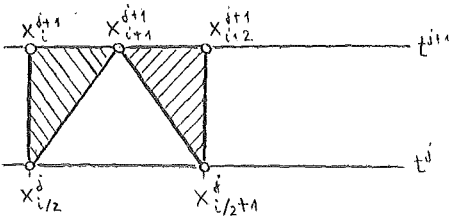
1^o the grid has equal number of x-intervals on j-th and on (j+1)-th time layer



2^o the grid is two times sparser on (j+1)-th time layer



3^o the grid is two times denser on (j+1)-th time layer



- The decision which pattern to choose is based on the following simple heuristics:

"If the function $u(x,t^j)$ is oscillating very much and/or has big discontinuities, then it should be computed more precisely (i.e. with denser grid)"

To measure oscillations and discontinuities of the function $u(x,t^j)$, the following variational norm is introduced:

$$d^j = \sum_{j=0}^{m_j} (v_i^j - v_{i-1}^j)^2, \text{ where } v_{-1}^j = a_0,\ v_{m_j}^j = a_L$$

If d^j (computed using the grid resulting from pattern 1^o) is significantly greater than d^j, then pattern 3^o, is rather used. Else if d^{j+1} is significantly less than d^j, then pattern 2^o is rather used.

- Time step Δt^j is chosen so as to keep the quotient $\Delta t^j/h^j$ constant through the whole computation. Initial value $\Delta t^0/h^0$ is determined as to satisfy the Courant condition.

- To compute the next step-function (on j+1-th time layer) the program uses formulae for v_i^{j+1} which are derived from more general formula (21). The trapesium (used in formula (21)) is substituted by a rectangle (pattern 1^o, pattern 2^o) or by one of the triangles (pattern 3^o). On figures illustrating the patterns, rectangles and triangles used are shaded.
- Since the quotient $\Delta t^j/h^j$ is kept constant, the "Godunov function" (used in formula (21)) can be replaced by three simpler functions:

$$g_0(v,\overline{v}) = \begin{cases} \min\limits_{u\in[v,\overline{v}]} f(u) \ , & \text{for } v \le \overline{v} \\ \max\limits_{u\in[\overline{v},v]} f(u) \ , & \text{for } v > \overline{v} \end{cases}$$

$$g_{-1}(v,\overline{v}) = \begin{cases} \min\limits_{u\in[v,\overline{v}]} \left(f(u) - \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v \le \overline{v} \\ \max\limits_{u\in[\overline{v},v]} \left(f(u) - \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v > \overline{v} \end{cases}$$

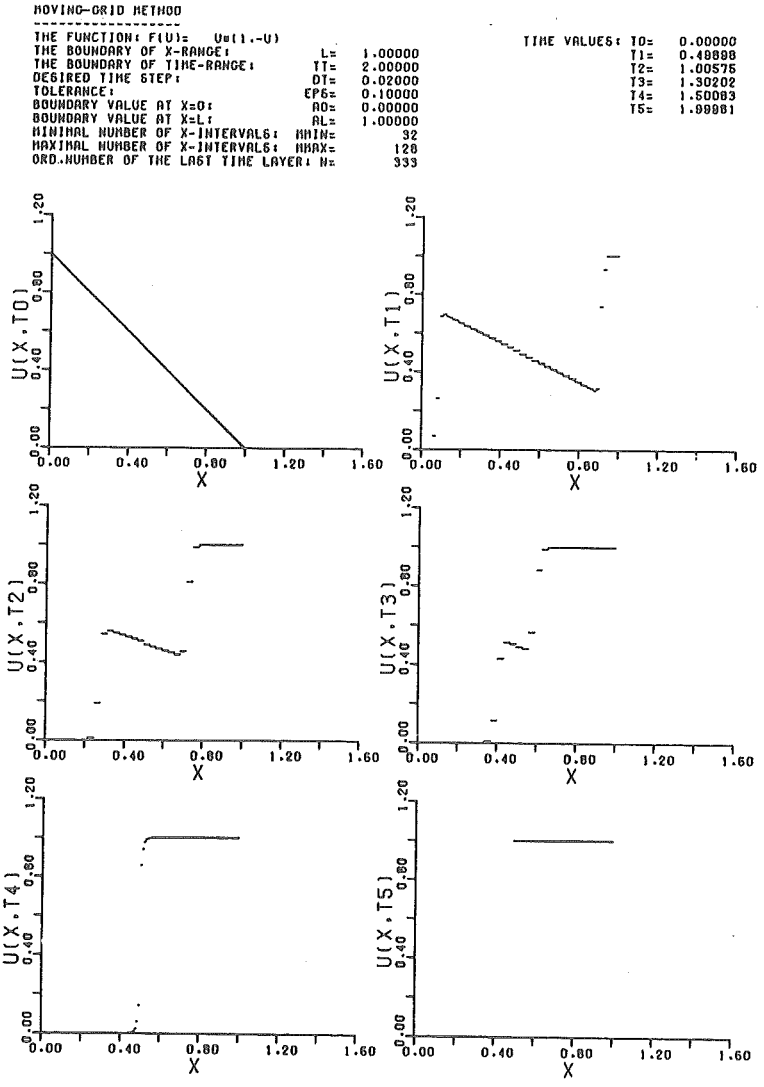
$$g_1(v,\overline{v}) = \begin{cases} \min\limits_{u\in[v,\overline{v}]} \left(f(u) + \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v \le \overline{v} \\ \max\limits_{u\in[\overline{v},v]} \left(f(u) + \frac{h^o}{2\Delta t^o} u\right) \ , & \text{for } v > \overline{v} \end{cases}$$

- Each evaluation of any of the functions g_0, g_{-1}, g_1 involves a constrained optimization problem. In order to solve these optimization problems efficiently, our program initially finds all local extrema of functions

$$f(u) \ , \ f(u) - \frac{h^o}{2\Delta t^o} u \ , \ f(u) + \frac{h^o}{2\Delta t^o} u \ .$$

The table of local extrema is used whenever one of the functions g_0, g_{-1}, g_1 is being evaluated.

The program was tested on a number of examples involving three different f-functions. The results were compared with known exact solutions (or numerical solutions obtained by a different method) as given in papers [3], [4]. There is a good accordance between the computed and expected values. On the following page a plotted report generated by our program is reproduced. All data describing the corresponding BLN problem are quoted. Approximate versions of the functions



$u(x,t^j)$ for six different values of t^j are plotted. Since the value t^o is equal to 0, the initial condition is also visible.

REMARKS. The work described in this paper was carried out as a part of authors' collaboration with INA Naftaplin petroleum Industry from Zagreb, Yugoslavia. The considered BLN problem has an interpretation which arises in the study of one-dimensional flow of two immiscible fluids (i.e. oil and water) through a porous medium. By solving a series of BLN problems and comparing the solutions with experimental results, one can estimate parameters describing physical properties of a given porous material. This is an important step leading to a reliable petroleum reservoir simulation.

REFERENCES:

1. C.BARDOS, A.Y.LEROUX and J.C.NEDELEC: First order quasilinear equations with boundary conditions, Rapp.Intervue CMA 38(1978)
2. Y.BRENIER and S.OSHER: Approximate Riemann solvers and numerical flux functions, SIAM J.Numer.Anal.2 (1986), 259-273.
3. G.CHAVENT and G.SALZANO: 1-D water flooding problem with gravity, J.Comput.Phys. 45(1982),307-343.
4. P.CONCUS and W.PROSKUROWSKI: Numerical solutions of a nonlinear hyperbolic equation by the random choice method, J.Comput.Phys. 30(1979),153-166.
5. A.Y.Le ROUX: Etude de probleme mixte pour une equation quasi-lineaire du premier ordre, C.R.Acad.Sc.Paris 285(1977), 351-354
6. R.SANDERS: The moving grid method for nonlinear hyperbolic conservation laws, SIAM J.Numer.Anal. 4(1985), 713-728.
7. A.I.VOLJPERT, S.I.HUDJAJEV: Analiz v klassah razryvnyh funkcij i uravnenija matematičeskoj fiziki, Nauka, Moskva, 1975.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

THE SPLINE TRANSFORM AND ITS APPLICATION IN THE PROBLEMS OF SIGNALS' DIGITAL TREATMENT

A.H. ARAKELIAN and M.R. VOSKANIAN

ABSTRACT: In this work the calculating formulas for computing the spectral characteristics are brought, obtained by the application of wide class of splines. The programmes of computing the spectral characteristics were used for investigating the medical-biological curves.

The practical application of splines shows, that for obtaining a considerable degree of closeness of a spline to the interpolating function, it is sufficient that the degree of splines be limited by four.

INTRODUCTION

A great number of papers is devoted to the treatment and prophylaxis of postcholecystectomy syndrome. But the number of research works on usage of differentiated health resort treatment complexes depending upon clinical variations of postcholecystectomy syndrome course is as far extremely insufficient. In a number of papers the significance of sanatoria and health resort treatment using mineral waters at early period after cholecystectomy is especially emphasized as a prophylactic method of serious complications after cholecystectomy.

Relative to the problem we have supposed that it would be timely to make clear the possibility and expedience of the usage at early periods after operation on bilare tract the health resort factors in particular, mineral water "Jermuk" of complex chemical composition, with the purpose of prophylaxis of postcholecystectomy syndrome and most rapid restoration of working capacity. There are no information concerning effect of Armenian mineral waters both under health resort and under common conditions on patient rehabilitation at early periods after cholecystectomy.

The paper presented is devoted to the investigation of the effect of complex treatment method developed for the patients after cholecystectomy operation, on the disease course. The investigation was conducted by means of rheopatography namely, utilization of registration of hepatic blood supply regularities by means of rheohepatograms (RHG). Fig.1 presents a typical RHG-registration.

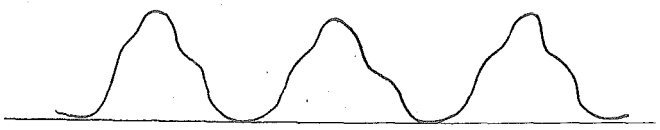


Fig. 1.

1. THE AIM OF INVESTIGATIONS

The basic aim of investigations conducted consists in the development of effective methods of patients early rehabilitation after cholecystectomy by physical factors depending upon the character of pathologic process in hepatobiliary system, in the estimation of efficiency, and in the recommendations for these methods usage.

The pecularities of clinical course of the patients condition after cholecystectomy, the effect of mineral water "Jermuk" combined with pine bath, massage of portal fissure zone muscles, remedial gymnastics on the patient condition after cholecystectomy, laboratory indices characterizing the hepatobiliary system condition, the alteration of liver hemodynamics according to rheohepatography data and modification dynamics of RGH-curve spectral characteristics were investigated.

2. METHODS OF SPECTRAL ANALYSIS

75 patients were observed. 50 of them have been transferred from surgical clinic to gastroenterological department in 2-3 weeks after cholecystectomy 25 patients had a period from 6 months to 6 years after operation. Almost all the patients investigated had a pain in the right hypochondrium, discomfort in epigastric region after eating, and at times nausea, heartburn, bitter taste and xerostomia. Often the patients have mentioned disorders of intestine emptying function.

In the first group of patients the phenomen of asthenic syndrome has been observed. The patients from both groups have received the same treatment complex during 24-26 days of hospital treatment. The analysis of data received has shown that for 95,8% patients in first group and 76,5% patients in second group the pain in the right hypochondrium has disappeared and for the others the pain intensity has essentially decreased. An analogous effect was observed for the pain in epigastric region. But, unlike the pain a number of patients have continued to complain of gastric dyspepsia effects in particular, heartburn, eructations although with decreased frequency of their appearance. For almost all the patients bitter taste, gastric flatulence, asthenic effects (weakness, erethism) have disappeared, defecation has normalized. For the registration of hepatic blood supply character the active electrode was placed across the medioclavicular line to the right, in the region of its intersection with the costal arch, and the passive one across the medioscapular line to the right, in the center between the angle of the scapulae and the crest of the iliac bone.

For calculation of integral Fourier transformation a method based on the approximate representation of integrand function by means of Hermite spline was used [1,3]. The spectral processing of the signals was conducted on a computer in real-time.

Special attention was drawn to functional condition of liver affected mostly by cholelithiasis. The liver condition was investigated by means of rheohepatography. Analysis of used treatment complex results was conducted by means of RHG-curve registration received by 4RG1A apparatus.

The interesting RHG characteristics are splash values m=0,038 Hz before and n=0,07 Hz after treatment. Fig.2 gives the typical shape of RHG spectrum before and after treatment with distinguished peaks. The typical frequency values are about 0,036-0,04 Hz for m and about 0,067-0,071 Hz for n.

3. ALGORITHM OF RHG SPECTRAL ANALYSIS

In the investigations conducted a calculation procedure based on RHG-curve Fourier transform represented by Hermite spline is used. The method permits in contrast with the visual one not only to reduce the investigation time but also to free the investigator from elements of subjectively peculiar to the visual method [2,3].

The calculation procedure of amplitude-frequency characteristic determination proceeds as follows:

Let $f(x) \in C^m[a,b]$ function be analysed where $C^m[a,b]$ is a space of real functions continuous on [a,b] interval and has continuous derivatives of m-degree.

Let

$$\Delta_n : a = x_0 < x_1 < x_2 < \ldots < x_n = b, \quad n \in N$$

is a net given on a finite interval [a,b].

Let us denote $(x-y_i)_+^{2m} = max[0; (x-y_i)]^m$.

Definition [1]. $S_{2m}(x) = S_{2m}(x; f)$, $m \geq 1$ function is called Hermite interpolation spline for a function $f(x) \in C^m[a,b]$, $m \in N$, if

a) $S_{2m}(x; f) = C^m[a,b]$ and

$$S_{2m}(x; f) = \sum_{s=0}^{2m} \frac{a_s^{(i)}}{s!}(x-x_i)^s + \frac{a_{2m+1}^{(i)}}{(2m)!}(x-y_i)_+^{2m}$$

for every $x \in [x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$.

b) $S_{2m}^{(k)}(x_i) = f^k(x_i)$,

$k = 0, 1, \ldots, m$; $i = 0, 1, \ldots, n$.

If $f(x) \in C^m[a,b]$, then [1,4] spline transform of Fourier for Hermite spline representation of $f(x)$ function is equal to $\bar{H}(j\omega) = \frac{1}{2\pi}\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} S_{2m}(x; f) =$

$$\frac{1}{2\pi}\left[\sum_{i=0}^{n-1}\sum_{s=0}^{2m}\frac{a_s^{(i)}}{s!}\int_{x_i}^{x_{i+1}}(x-x_i)^s e^{-j\omega x}dx + \int_{y_i}^{x_{i+1}}(x-y_i)^{2m}e^{-j\omega x}dx\right].$$

If we denote

$$Q^{(i)}(s,\omega) = \int_{x_i}^{x_{i+1}}(x-x_i)^s e^{-j\omega x}dx; \quad L^i(2m,\omega) = \int_{y_i}^{x_{i+1}}(x-y_i)^{2m}e^{-j\omega x}dx,$$

then it is possible to construct the following iteration procedures for their definition

$$Q^{(i)}(0,\omega) = (e^{-j\omega x_i} - e^{-j\omega x_{i+1}})/j\omega; \quad Q^{(i)}(1,\omega) = [Q^{(i)}(0,\omega) - (x_{i+1}-x_i)e^{-j\omega x_{i+1}}]/j\omega;$$

$$Q^{(i)}(s,\omega) = [sQ(s-1,\omega) - (x_{i+1}-x_i)^s e^{-j\omega x_{i+1}}]/j\omega$$

and $L^i(0,\omega) = -(e^{-j\omega x_{i+1}} - e^{-j\omega y_i})/j\omega$; $L^{(i)}(1,\omega) = [L^{(i)}(0,\omega) - (x_{i+1}-y_i)e^{-j\omega x_{i+1}}]/j\omega$;

$$L^{(\kappa)}(\kappa,\omega) = [\kappa Q(\kappa-1,\omega) - (x_{i+1} - y_i)^{\kappa} e^{-j\omega x_{i+1}}]/j\omega$$

respectively.

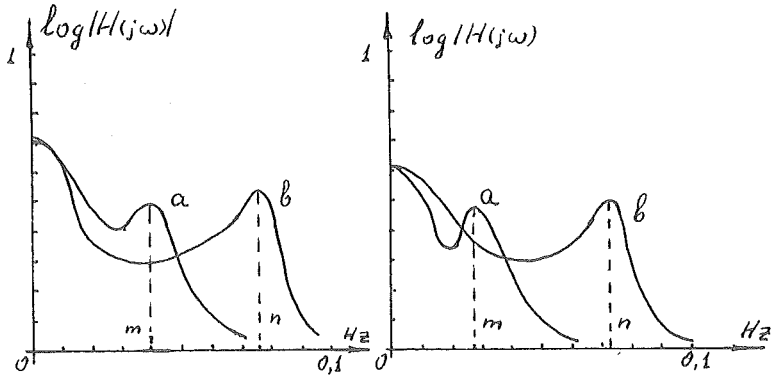
If we denote the Fourier transform of $f(x)$ function by $H(j\omega)$, then $|H(j\omega) - \bar{H}(j\omega)| = \frac{1}{2\pi} \left| \int_a^b (f(x) - S_{2m}(x; f)) e^{-j\omega x} dx \right| \leq$

$$\frac{1}{2\pi} (2m+1) \frac{\|\Delta_n\|^{2m} \tilde{\omega}(f^{(2m)}, \|\Delta_n\|)}{2^{2m}(2m)!},$$

where $\|\Delta_n\| = \max_{0 \leq i \leq n-1} (x_{i+1} - x_i)$, $\tilde{\omega}(f^{(\kappa)}, \|\Delta_n\|) = \max_{0 \leq i \leq n-1} \max_{y \in [x_i, x_{i+1}]} |f^{(\kappa)}(x) - f^{(\kappa)}(y)|$.

4. RHG SIGNAL SPECTRAL ANALYSIS

It is known that RHG represents an electrical signal generated by an non-stationary source, the liver. Fig.2 and 3 represent the results of RHG-curve spectral analysis for the first and second groups of patients, respectively.



The results of investigations conducted have shown that the increase speed of frequency value at which the spectrum peak is observed, is conditioned by intensity increase of hepatic blood flow at the expense of both arterial inflow and venous outflow. Besides, the first group of patients has larger values of spectral characteristics than second one.

Thus, the observations have shown that the used treatment complex including inner dose of carbonate-hydrocarbonate-sulphate-chlorine-natrium-calcium-magnesium mineral water "Jermuk" permits to improve essentially the liver

hemodynamics for both groups of patients. At the same time information obtained testifies more evidently expressed decrease of liver hypoxia, for the first group of patients for which the rehabilitating treatment began at earlier period after operation.

5. CONCLUSIONS

1. The rehabilitating treatment involving the balneologic factors and conducted at early period after cholecystectomy promotes the favourable dynamics of post-operational syndromes, the normalization of lives functional condition, circulation of the blood in liver, the most rapid restoration of working capacity.

2. The use of signal digital processing methods and algorithms and their realizing programs stipulates for possibility of objective quantitative estimation of RHG-curve. The problem of determining when and under what conditions RHG-curve changes was so far not solved. The RHG spectral processing method used in the paper permits to receive the quantitative information about violations of hepatic blood supply character.

References

1. ARAKELIAN A.H.: Optimization methods of dialogue information systems for testing, Acad. Science of Armenian SSR, Yerevan, 1985.
2. ARAKELIAN A.H., AGAIAN S.S.: On an algorithm of spectral analysis, Cyber. and Syst. Res. 2, 1984. R. Trappl. North-Holl., pp.273-275.
3. VOSKANIAN M.R., ARAKELIAN A.H.: The Spectr. Analy. of EGH and its Application. IN: Proceedings of a IX All-Union Conf. on Operational Problems. M.1983, p.394
4. VOSKANIAN M.R., KHONDKARIAN N.S.: Spline transform of Fourier in the Problems of RHG Spectral Analysis. In: Proceedings of a II All-Union Conference. "The Realization of Mathematical Methods in Clinical and Experimental Medicine", Moscow, 1986, pp.92-93.

AN IMPLEMENTATION OF A SEMI-DEFINITE PROGRAMMING METHOD TO CHEBYSHEV APPROXIMATION PROBLEMS

M.D. AŠIĆ and V.V. KOVAČEVIĆ-VUJČIĆ

ABSTRACT: A discretization method for solving linear semi-infinite programming problems arising from Chebyshev approximation is presented. The method is based on selective refinement of the initial coarse grid, which enables an efficient treatment of multidimensional problems. Numerical examples from Chebyshev approximation are also presented.

1. INTRODUCTION

This paper is a natural extension of a sequence of papers on semi-infinite programming methods ([2] , [3] , [4] , [5] , [6]). We consider here the following Chebyshev approximation problem:

Let $C=[p_1,q_1]\times\ldots\times[p_r,q_r]$ and let $g_i:C\to R$, $i=1,\ldots,m$ and $f:C\to R$ be given functions. Find $x_1,\ldots,x_m$ such that

$$\max_{t\in C} |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \tag{1}$$

is minimized.

It is easy to see that this problem can be reformulated as the linear semi-infinite programming problem:

$$\min x_{m+1}$$
$$x_{m+1}\geq |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \quad \text{for all } t\in C. \tag{2}$$

For brevity, let $x=(x_1,\ldots,x_{m+1})$,

$$c_1(x,t)=x_1g_1(t)+\ldots+x_mg_m(t)-x_{m+1}$$
$$c_2(x,t)=-x_1g_1(t)-\ldots-x_mg_m(t)-x_{m+1}.$$

Then (2) becomes

$$\min x_{m+1}$$
$$x\in X,\ X=\{x\in R^{m+1} \mid c_1(x,t)\leq f(t),\ c_2(x,t)\leq -f(t) \quad \text{for all } t\in C\}. \tag{3}$$

In the sequel we shall use the following:

Assumption 1. (i) There exists an $\bar{x}\in X$ such that the set

hemodynamics for both groups of patients. At the same time information obtained testifies more evidently expressed decrease of liver hypoxia, for the first group of patients for which the rehabilitating treatment began at earlier period after operation.

5. CONCLUSIONS

1. The rehabilitating treatment involving the balneologic factors and conducted at early period after cholecystectomy promotes the favourable dynamics of post-operational syndromes, the normalization of lives functional condition, circulation of the blood in liver, the most rapid restoration of working capacity.

2. The use of signal digital processing methods and algorithms and their realizing programs stipulates for possibility of objective quantitative estimation of RHG-curve. The problem of determining when and under what conditions RHG-curve changes was so far not solved. The RHG spectral processing method used in the paper permits to receive the quantitative information about violations of hepatic blood supply character.

References

1. ARAKELIAN A.H.: Optimization methods of dialogue information systems for testing, Acad. Science of Armenian SSR, Yerevan, 1985.
2. ARAKELIAN A.H., AGAIAN S.S.: On an algorithm of spectral analysis, Cyber. and Syst. Res. 2, 1984. R. Trappl. North-Holl., pp.273-275.
3. VOSKANIAN M.R., ARAKELIAN A.H.: The Spectr. Analy. of EGH and its Application. IN: Proceedings of a IX All-Union Conf. on Operational Problems. M.1983, p.394
4. VOSKANIAN M.R., KHONDKARIAN N.S.: Spline transform of Fourier in the Problems of RHG Spectral Analysis. In: Proceedings of a II All-Union Conference. "The Realization of Mathematical Methods in Clinical and Experimental Medicine", Moscow, 1986, pp.92-93.

AN IMPLEMENTATION OF A SEMI-DEFINITE PROGRAMMING METHOD TO CHEBYSHEV APPROXIMATION PROBLEMS

M.D. AŠIĆ and V.V. KOVAČEVIĆ-VUJČIĆ

ABSTRACT: A discretization method for solving linear semi-infinite programming problems arising from Chebyshev approximation is presented. The method is based on selective refinement of the initial coarse grid, which enables an efficient treatment of multidimensional problems. Numerical examples from Chebyshev approximation are also presented.

1. INTRODUCTION

This paper is a natural extension of a sequence of papers on semi-infinite programming methods ([2] , [3] , [4] , [5] , [6]). We consider here the following Chebyshev approximation problem:

Let $C=[p_1,q_1]\times\ldots\times[p_r,q_r]$ and let $g_i:C\to R$, $i=1,\ldots,m$ and $f:C\to R$ be given functions. Find $x_1,\ldots,x_m$ such that

$$\max_{t\in C} |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \tag{1}$$

is minimized.

It is easy to see that this problem can be reformulated as the linear semi-infinite programming problem:

$$\min x_{m+1}$$
$$x_{m+1}\geq |f(t)-x_1g_1(t)-\ldots-x_mg_m(t)| \quad \text{for all } t\in C. \tag{2}$$

For brevity, let $x=(x_1,\ldots,x_{m+1})$,

$$c_1(x,t)=x_1g_1(t)+\ldots+x_mg_m(t)-x_{m+1}$$
$$c_2(x,t)=-x_1g_1(t)-\ldots-x_mg_m(t)-x_{m+1}.$$

Then (2) becomes

$$\min x_{m+1}$$
$$x\in X,\ X=\{x\in R^{m+1} \mid c_1(x,t)\leq f(t),\ c_2(x,t)\leq -f(t) \quad \text{for all } t\in C\}. \tag{3}$$

In the sequel we shall use the following:

Assumption 1. (i) There exists an $\bar{x}\in X$ such that the set

$$\bar{X} = X \cap \{x \in R^{m+1} \mid x_{m+1} \leq \bar{x}_{m+1}\}$$

is bounded.

(ii) The functions $g_1,\ldots,g_m$ and f satisfy the Lipschitz condition.

It is clear that Assumption 1 implies the existence of a uniform Lipschitz constant L for functions $c_1(x,t)$, $c_2(x,t)$ and $f(t)$, i.e.

$$|c_i(x,t')-c_i(x,t")| \leq L\|t'-t"\| \quad , \quad x \in \bar{X}, \quad i=1,2$$
$$|f(t')-f(t")| \leq L\|t'-t"\| .$$

The main idea of the method which will be described in Section 2 is to use selective discretization of the index set C in order to replace semi-infinite programming problem (3) by a sequence of linear programming problems. The method starts with a uniform grid which depends on the Lipschitz constant L and suc cessive refinements are made in such a way to ensure linear grow of the number of grid points, while retaining the usual converge ce properties.

2. THE METHOD

In order to describe the algorithm of the method we need th following notation:

Let (M_j) denote the sequence of uniform discretizations of the set C defined by

$$M_j = \{(p_1+k_1h_1^j,\ldots,p_r+k_rh_r^j) \mid k_i \in \{0,1,\ldots,2^j m_i\} \ , \ i=1,\ldots,r\} \ ,$$

where $h_i^j=(q_i-p_i)/(2^j m_i)$, and m_i are appropriately chosen posit ve integers. Furthermore, for given $y \in R^{m+1}$, $t \in C$, $h_1>0,\ldots,h_r>0$ let q_1 and q_2 be the functions defined by

$$q_1(s) = c_1(y,t)-f(t)+ \sum_{i=1}^{r} \bar{A}_{i1}(s_i-t_i) + \sum_{i=1}^{r} \bar{B}_{i1}|s_i-t_i|$$

(4)

$$q_2(s) = c_2(y,t)+f(t)+ \sum_{i=1}^{r} \bar{A}_{i2}(s_i-t_i) + \sum_{i=1}^{r} \bar{B}_{i2}|s_i-t_i| \ ,$$

where $\bar{A}_{i1},\bar{A}_{i2},\bar{B}_{i1},\bar{B}_{i2}$ are such that

$$q_1(s) \geq c_1(y,s)-f(s), \quad q_2(s) \geq c_2(y,s)+f(s)$$

for all s satisfying $|s_i-t_i| \leq h_i$, $i=1,\ldots,r$. It is obvious that and q_2 depend also on $y,t,h_1,\ldots,h_r$ and that they are actually piecewise linear majorants of c_1-f and c_2+f, respectively.

lgorithm 1. Input parameters: $\beta_0>0$, Lipschitz constant L, integers $m_1,\ldots,m_r$ satisfying

$$m_i>(q_i-p_i)L\sqrt{r}/(2\beta_0),\quad i=1,\ldots,r.$$

tep 0. Set $\gamma_0=\beta_0/L$, $C_0=M_0$, $k=0$.

tep 1. Solve the linear programming problem:

$$\min x_{m+1}$$

$$x\in Y_k,\quad Y_k=\{x\in R^{m+1} \mid c_1(x,t)\le f(t)-\beta_k,\ c_2(x,t)\le -f(t)-\beta_k,\ t\in C_k\}$$

nd let y be a solution. Set $j=0$, $E_0=C_0$.

tep 2. If $j=k$ go to Step 4. Otherwise, for each $t\in E_j$ find unctions q_1, q_2 such that

$$q_1(s)\ge c_1(y,s)-f(s),\quad q_2(s)\ge c_2(y,s)+f(s),$$

or all s satisfying $|s_i-t_i|\le h_i^j$, $i=1,\ldots,r$.

Let E'_{j+1} be the set of points $t\in E_j$ for which either q_1 or q_2 s greater than or equal to $-\beta_k$ at some extreme point of the set

$$([t_1-h_1^j,t_1+h_1^j]\times\ldots\times[t_r-h_r^j,t_r+h_r^j])\cap C.$$

et E_{j+1} be the set of points in M_{j+1} whose distance from the set , $j+1$ does not exceed γ_j. Replace j by j+1.

tep 3. If for all $t\in E_j$

$$c_1(y,t)\le f(t)-\beta_k,\quad c_2(y,t)\le -f(t)-\beta_k,$$

o to Step 2. Otherwise, let

$$T=\{t\in E_j \mid c_1(y,t)>f(t)-\beta_k \text{ or } c_2(y,t)>-f(t)-\beta_k\},$$

eplace C_k by $C_k\cup T$ and go to Step 1.

tep 4. Set $x^{k+1}=y$, $C_{k+1}=C_k$, $\beta_{k+1}=\beta_k/2$, $\gamma_{k+1}=\gamma_k/2$, replace k by +1 and go to Step 1.

It is easy to see that Algorithm 1 belongs to the class of ethods defined in [4]. The result in [4] implies that the Alorithm is well defined and that each cluster point of the sequnce (x^k) generated by the Algorithm is a solution to (3). Moreover, $x^k\in X$ for all $k=0,1,\ldots$.

The efficiency of the Algorithm depends on the cardinalities of the sets E_j and on the number of inner cycles of the type Step 1 → Step 2 → Step 3 → Step 1 at the k-th iteration, which determines the cardinality of the set C_k. For further analysis we need the following:

Assumption 2. (i) x^* is the unique solution to (3).

(ii) $(c_1(x^*,t)-f(t))(c_2(x^*,t)+f(t))=0$ for finitely many t's. Let T^* be a complete list.

(iii) Functions $f,g_1,\ldots,g_m$ are twice continuously differentiable on C.

(iv) For each $t^* \in T^*$ the following property holds: If t_i^* is an endpoint of $[p_i,q_i]$ and, say, $c_1(x^*,t^*)-f(t^*)=0$ then $\partial(c_1(x^*,t^*)-f(t^*))/\partial t_i \neq 0$. Moreover, the Hessian matrix with respect to the remaining t_i's is negative definite at t^*. Similar property holds if $c_2(x^*,t^*)+f(t^*)=0$.

The following result on the cardinalities of the sets E_j ho

Theorem 1. Let Assumptions 1 and 2 be satisfied and let q_1 a q_2 have the form (4). Moreover, assume that $\bar{B}_{i1} \to 0$, $\bar{B}_{i2} \to 0$ as $h_1 \to 0,\ldots,h_r \to 0$ for all $i=1,\ldots,r$. Then the cardinalities of the sets E_j generated by Algorithm 1 are bounded above by a constant independent on j and k.

The proof is similar to that of Theorem 3.1 in [5] and is t omitted. Let us only point out that under Assumption 2 functions q_1 and q_2 of the type (4) satisfying $\bar{B}_{i1} \to 0$, $\bar{B}_{i2} \to 0$ as $h_1 \to 0,..$ $h_r \to 0$, $i=1,\ldots,r$ can be obtained using the first order Taylor e pansion of c_1-f and c_2+f and rounding up the remainder term.

It remains to analyze the number of inner cycles at the k- iteration, which is done by the following theorem:

Theorem 2. Assume that the functions $g_1,\ldots,g_m$ satisfy the condition on the set C and $f \notin \operatorname{span}\{g_1,\ldots,g_m\}$. Suppose furth more that Assumptions 1 and 2 hold. Then the number of cycles the type Step 1 → Step 2 → Step 3 → Step 1 in Algorithm 1 is bounde a constant independent on k.

The proof follows directly from the following three Lemma

Lemma 1. Suppose that the assumptions of Theorem 2 are sati fied. Let y be one of the points generated in Step 1 of Algori during the (k+1)-th iteration. Then there is a positive consta independent on k such that

$$\bar{y}_{m+1} \leq y_{m+1} \leq \bar{y}_{m+1} + D\beta_{k+1}^2 \quad , \text{ where } \bar{y}=(x_1^k,\ldots,x_m^k,x_{m+1}^k-\beta_{k+1}$$

The proof of Lemma 1 is similar to that of Lemma 3.1 in [6 and is omitted.

Lemma 2. Suppose that the assumptions of Theorem 2 are satisfied Let y be one of the points generated in Step 1 of Algorithm 1 during the iteration k+1. Then there is a positive constant E independent on k such that

$$\|y-\bar{y}\| \leq E\beta_{k+1}^2, \text{ where } \bar{y}=(x_1^k,\ldots,x_m^k,x_{m+1}^k-\beta_{k+1}).$$

Proof : Note first that $\bar{y}$ is a solution to the problem

$$\min x_{m+1}$$

$$c_1(x,t)\leq f(t)-\beta_k,\ c_2(x,t)\leq -f(t)-\beta_k,\ t\in C_k.$$

By duality theorem there are nonpositive numbers $d_1(t)$, $d_2(t)$, $t\in C_1$ such that

$$\sum_{t\in C_k} d_1(t)g_i(t)-\sum_{t\in C_k} d_2(t)g_i(t) = 0,\ i=1,\ldots,m$$

$$\sum_{t\in C_k} d_1(t)+\sum_{t\in C_k} d_2(t) = -1.$$

By Caratheodory's theorem we may assume that

$$\sum_{j=1}^{m'} d_1(t^j)g_i(t^j) - \sum_{j=m'+1}^{m''} d_2(t^j)g_i(t^j) = 0,\ i=1,\ldots,m$$

$$\sum_{j=1}^{m'} d_1(t^j) + \sum_{j=m'+1}^{m''} d_2(t^j) = -1, \tag{5}$$

where $d_1(t^j)<0$, $j=1,\ldots,m'$, $d_2(t^j)<0$, $j=m'+1,\ldots,m''$ and $m''\leq m+1$. It easily follows that for k large enough each t^j is in the neigh borhood of some point in T^*. Without loss of generality we may as sume that these points are $t^{*1},\ldots,t^{*\bar{m}'},t^{*\bar{m}'+1},\ldots,t^{*\bar{m}''}$ and that the corresponding neighborhoods are disjoint. It is clear that $\bar{m}''\leq m''\leq$ We will show that $\bar{m}''=m+1$.

Assume the contrary and let

$$F = \begin{bmatrix} g_1(t^{*1}) & \ldots & g_1(t^{*\bar{m}'}) & -g_1(t^{*\bar{m}'+1}) & \ldots & -g_1(t^{*\bar{m}''}) \\ \vdots & & & & & \\ g_m(t^{*1}) & \ldots & g_m(t^{*\bar{m}'}) & -g_m(t^{*\bar{m}'+1}) & \ldots & -g_m(t^{*\bar{m}''}) \\ 1 & \ldots & 1 & 1 & \ldots & 1 \end{bmatrix}.$$

Due to the Haar condition, the system

$$F^T u = [1\ \ldots\ 1]^T$$

$$u_{m+1} = -1$$

has a solution $\bar{u}_1,\ldots,\bar{u}_m,\bar{u}_{m+1}$. Moreover, for k large enough,

$$g_1(t^j)\bar{u}_1+\ldots+g_m(t^j)\bar{u}_m+\bar{u}_{m+1}>0\ ,$$
$$-g_1(t^j)\bar{u}_1-\ldots-g_m(t^j)\bar{u}_m+\bar{u}_{m+1}>0\ .$$

Multiplying the equalities (5) by $\bar{u}_i$'s and adding we obtain:

$$0>\sum_{j=1}^{m'} d_1(t^j)(g_1(t^j)\bar{u}_1+\ldots+g_m(t^j)\bar{u}_m+\bar{u}_{m+1}) +$$
$$+\sum_{j=m'+1}^{m''} d_2(t^j)(-g_1(t^j)\bar{u}_1-\ldots-g_m(t^j)\bar{u}_m+\bar{u}_{m+1}) = -\bar{u}_{m+1} = 1,$$

which is a contradiction.

Hence, $\bar{m}''=m+1$ and (5) holds with $m''=m+1$. It is easy to see now that $d_1(t^j)$ and $d_2(t^j)$ in (5) are bounded above by a negative constant G. Multiplying the equalities (5) by $\bar{y}_1,\ldots,\bar{y}_m,-\bar{y}_{m+1}$, respectively, and adding, we obtain

$$x^k_{m+1}-\beta_{k+1}=\sum_{j=1}^{m'} d_1(t^j)c_1(\bar{y},t^j)+\sum_{j=m'+1}^{m+1} d_2(t^j)c_2(\bar{y},t^j)\ ,$$

which implies

$$x^k_{m+1}-\beta_{k+1}=\sum_{j=1}^{m'} d_1(t^j)f(t^j)-\sum_{j=m'+1}^{m+1} d_2(t^j)f(t^j)+\beta_{k+1}. \tag{6}$$

Let y be an arbitrary point generated at Step 1 during the iteration k+1. Then

$$\begin{aligned} c_1(y,t^j) &= f(t^j)-\beta_{k+1}+v_j,\ j=1,\ldots,m' \\ c_2(y,t^j) &= -f(t^j)-\beta_{k+1}+v_j,\ j=m'+1,\ldots,m+1, \end{aligned} \tag{7}$$

where $v_j\leq 0$, $j=1,\ldots,m+1$. Multiplying the equalities (5) by $y_1,\ldots,y_m,-y_{m+1}$, respectively, and adding, we obtain

$$y_{m+1}=\sum_{j=1}^{m'} d_1(t^j)c_1(y,t^j)+\sum_{j=m'+1}^{m+1} d_2(t^j)c_2(y,t^j),$$

which by (7) implies

$$y_{m+1}=\sum_{j=1}^{m'} d_1(t^j)(f(t^j)-\beta_{k+1}+v_j)+\sum_{j=m'+1}^{m+1} d_2(t^j)(-f(t^j)-\beta_{k+1}+v_j).$$

Now using (6) we obtain

$$y_{m+1}=x^k_{m+1}-\beta_{k+1}+\sum_{j=1}^{m'} d_1(t^j)v_j+\sum_{j=m'+1}^{m+1} d_2(t^j)v_j\leq x^k_{m+1}-\beta_{k+1}+D\beta^2_{k+1}$$

where the inequality follows from Lemma 1. Hence,

$$\sum_{j=1}^{m'} d_1(t^j)v_j+\sum_{j=m'+1}^{m+1} d_2(t^j)v_j\leq D\beta^2_{k+1}\ ,$$

so that

(8) $$0 \geq v_i \geq D\beta_{k+1}^2/G \ .$$

Note that y and $\bar{y}$ can be thought of as solutions to system of linear equations (7) and the corresponding system when v_i's are replaced by 0. By Cramer's rule we obtain

$$\|y-\bar{y}\| \leq D_1(|v_1|+\ldots+|v_n|) \ ,$$

where D_1 does not depend on k. Finally, (8) yields

$$\|y-\bar{y}\| \leq E\beta_{k+1}^2 .$$

Lemma 3. Suppose that the assumptions of Theorem 2 are fulfille and let $\bar{y}=(x_1^k,\ldots,x_m^k,x_{m+1}^k-\beta_{k+1})$, $w_1(t)=c_1(\bar{y},t)-f(t)+\beta_{k+1}$, $w_2(t)=c_2(\bar{y},t)+f(t)+\beta_{k+1}$. Let $\hat{t}$ be any point added to C_k during the iteration k+1. Then for k large enough either $w_1(t)$ or $w_2(t)$ has a local maximum $\bar{t}$ such that $\|\hat{t}-\bar{t}\| \leq F\beta_{k+1}$, where F does not depend on k

The proof of Lemma 3 and Theorem 2 is analogous to the proof of Lemma 3.3 and Theorem 3.2 in [6].

Let us note that an immediate consequence of Theorem 2 is that the cardinality of the sets C_k generated by Algorithm 1 grows at most linearly with k. Theorems 1 and 2 also imply that the total number of points generated by the algorithm at the k-th iteration is bounded above by a function linear in k, while at the same time the cardinality of the uniform grid M_k depends exponentially on k. Numerical experience seems to indicate that this linear behaviour is retained also when Haar's condition is omitted. It should be pointed out that the existing discretization methods (see e.g. [9], see also [8]) have an exponential upper bound on the number of points generated at the k-th step.

3. NUMERICAL EXPERIENCE

The method described in Section 2 was tested on a number of test problems, mostly taken from [1] and [7] . The obtained result agree very well with the data in the literature. Here we give the details for three examples. In the corresponding tables $N(C_k)$ and $N(E_j)$ stand for the cardinality of C_k and the average cardinality of E_j at the k-th iteration, respectively.

Example 1.[1]. Approximate $(t_1)^2t_2$ by $v_1=1$, $v_2=t_1$, $v_3=(t_1)^2$, $v_4=t_2$, $v_5=(t_2)^2$, $v_6=t_1t_2$ on $(t_1)^2+(t_2)^2 \leq 1$.

Following the authors in [1] the problem is reduced to the approximation problem on $[0,1] \times [0,2\pi]$. Input parameters are : $\beta_0=4.4$, $L=10.5$, $m_1=2$, $m_2=11$.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$N(C_k)$	36	42	46	50	54	58	62	66	70	74	78	82	86	90
$N(E_j)$	86	64	61	60	53	47	47	47	41	45	46	41	42	41

x^{13}=(0.0000,0.0000,0.0000,0.2500,0.0000,0.0000).

Exact solution x^*=(0,0,0,1/4,0,0).

Example 2.[7] . Approximate $\exp(-(t_1)^2-t_2)$ by functions $v_1=1$, $v_2=t_1$, $v_3=t_2$, $v_4=2(t_1)^2-1$, $v_5=t_1t_2$, $v_6=2(t_2)^2-1$ on the set $[0,1] \times [0,1]$.

Input parameters are: $\beta_0=2.5$, $L=24.2$, $m_1=m_2=7$.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$N(C_k)$	64	70	76	79	83	85	88	93	97	100	103	106	108	111
$N(E_j)$	0	46	44	40	40	40	42	39	36	37	35	36	34	36

x^{13}=(0.9858,-0.3480,-0.9027,-0.1446,0.4246,0.1129).

Solution in [7] : (0.9858,-0.3480,-0.9027,-0.1446,0.4246,0.1129).

Example 3. [1] . Approximate t^2 by functions $v_1=t$, $v_2=\exp(t)$ on the interval $[0,2]$.

Input parameters are: $\beta_0=2$, $L=13$, $m_1=8$.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$N(C_k)$	9	11	12	13	14	15	16	17	18	19	20	21	22	23
$N(E_j)$	0	7	7	6	6	6	6	6	6	6	6	6	5	5

x^{13}=(0.1842,0.4186).

Solution in [1] : (0.1842,0.4186).

REFERENCES

1. D.D. ANDREASSEN and G.A. WATSON : Linear Chebyshev approximati without Chebyshev sets. BIT 16 (1976), 349-362.

2. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : An application of semi-infinite programming to approximation theory. In: Proceedings of the XI Yugoslavian Symposium on Operations Research, Herceg Novi, 1984, pp. 55-64.

3. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : A semi-infinite programming method and its application to boundary value problems. ZAMM 66 (1986), 403-405.

4. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : An interior semi-infinite programming method. JOTA 59 (1988) .

5. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : Computational complexity of some semi-infinite programming methods. In: System Modelling and Optimization (A. Prekopa, B. Strazicky, J.Szelezsan, eds.), Springer-Verlag, Berlin, 1986, pp. 34-42.

6. M.D. ASIC and V.V. KOVACEVIC-VUJCIC : Linear semiinfinite programming problem: A discretization method with linearly growing number of points. In: Proceedings of 17. Jahrestagung "Mathematische Optimierung" (K. Lommatzsch ed .), Seminarbericht 85, Humboldt Universitat zu Berlin, 1986, pp. 1-10.

7. K. GLASHOFF and S.A. GUSTAFSON : Linear Optimization and Approximation. Springer-Verlag, Berlin, 1983.

8. R. HETTICH : A review of numerical methods for semi-infinite optimization. In: Semi-Infinite Programming and Applications (A.V. Fiacco, K.O. Kortanek, eds.), Springer-Verlag, Berlin, 1983, pp. 158-178.

9. R. HETTICH: An implementation of a discretization method for semi-infinite programming. Mathematical Programming 34 (1986), 354-361.

ON THE ZEROS OF A POLYNOMIAL

M. BIDKHAM and K.K. DEWAN

ABSTRACT: In this paper we have considered the problem of finding the maximum number of zeros in a prescribed region.

1. INTRODUCTION AND STATEMENT OF RESULTS

The following result is due to Mohammad [4]

THEOREM A. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ *be a polynomial of degree* n *such that*

$$a_n \geq a_{n-1} \geq \cdots \geq a_1 \geq a_0 > 0 ,$$

then the number of zeros of $p(z)$ *in* $|z| \leq \frac{1}{2}$ *does not exceed*

$$1 + \frac{1}{\log 2} \log \frac{a_n}{a_0}$$

As a generalization of Theorem A, Dewan [1] proved

THEOREM B. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ *be a polynomial of degree* n *with complex coefficients such that*

$$|\arg a_k - \beta| \leq \alpha \leq \pi/2 , \quad k = 0, 1, \ldots, n$$

for some real β , *and*

$$|a_n| \geq |a_{n-1}| \geq \cdots \geq |a_1| \geq |a_0| ,$$

then the number of zeros of $p(z)$ *in* $|z| \leq \frac{1}{2}$ *does not exceed*

$$\frac{1}{\log 2} \log \frac{|a_n|(\cos\alpha + \sin\alpha + 1) + 2\sin\alpha \sum_{k=0}^{n-1} |a_k|}{|a_0|}$$

THEOREM C. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ *be a polynomial of degree* n *with complex coefficients. If* $\operatorname{Re} a_k = \alpha_k$, $\operatorname{Im} a_k = \beta_k$, *for*

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

$$\alpha_n \geq \alpha_{n-1} \geq \ldots \geq \alpha_1 \geq \alpha_0 > 0 ,$$

then the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed

$$1 + \frac{1}{\log 2} \log \frac{\alpha_n + \sum_{k=0}^{n} |\beta_k|}{|a_0|} .$$

In this paper, we generalize Theorems A, B and C for different classes of polynomials which in turn also refine upon them. More precisely, we prove the following.

THEOREM 1. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n with complex coefficients. If for some real β, $|\arg a_i - \beta| \leq \alpha \leq \pi/2$, $0 \leq i \leq n$ and for some $0 < t \leq 1$

$$|a_0| \leq t|a_1| \leq \ldots \leq t^k|a_k| \geq t^{k+1}|a_{k+1}| \geq \ldots \geq t^n|a_n|, \quad 0 \leq k \leq n;$$

then the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed.

$$\frac{1}{\log 2} \log \frac{2t^{k+1}|a_k|\cos\alpha + 2t\sin\alpha \sum_{i=0}^{n} t^i|a_i| - (\cos\alpha + \sin\alpha - 1)\, t^{n+1}|a_n|}{t|a_0|} .$$

REMARK 1. For $t = 1$ and $k = n$ the above theorem reduces to Theorem B. If in addition to $t = 1$ and $k = n$, $\alpha = \beta = 0$ then it reduces to Theorem A.

THEOREM 2. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n with compex coefficients. If $\operatorname{Re} a_i = \alpha_i$, $\operatorname{Im} a_i = \beta_i$, for $i = 0, 1, \ldots, n$ and for some $0 < t \leq 1$

$$0 < \alpha_0 \leq t\alpha_1 \leq \ldots \leq t^k \alpha_k \geq t^{k+1} \alpha_{k+1} \geq \ldots \geq t^n \alpha_n$$

then the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed

$$1 + \frac{1}{\log 2} \log \left\{ \frac{t^{k+1}\alpha_k + t \sum_{i=0}^{n} |\beta_i| t^i}{t|\alpha_0|} \right\}$$

REMARK 2. For $k = n$ and $t = 1$, Theorem 2 reduces to Theorem C nd for $k = n$, $t = 1$ and $\beta_i = 0$, $0 \leq i \leq n$, it reduces to Theorem A.

The proof of next theorem follows on combining Theorem B and Lemma 2.

THEOREM 3. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n with complex coefficients such that

$$|\arg a_i - \beta| \leq \alpha \leq \pi/2 , \quad i = 0, 1, \ldots, n$$

for some real β, and

$$|a_n| \geq |a_{n-1}| \geq \cdots \geq |a_1| \geq |a_0|,$$

then the number of zeros of $p(z)$ in $R_2 \leq |z| \leq \frac{1}{2}$ does not exceed

$$\frac{1}{\log 2} \log \frac{|a_n|(\cos \alpha + \sin \alpha + 1) + 2 \sin \alpha \sum_{i=0}^{n} |a_i|}{|a_0|}$$

where R_2 is the same as defined in Lemma 2.

The above Thoerem is a refinement of Theorem B. In particular, if $\alpha = \beta = 0$, then it gives a refinement of Theorem A.

THEOREM 4. Let $p(z) = \sum_{i=0}^{n} a_i z^i$. If $\operatorname{Re} a_i = \alpha_i$, $\operatorname{Im} a_i = \beta_i$, for $i = 0, 1, \ldots, n$, and

$$\alpha_n \geq \alpha_{n-1} \geq \cdots \geq \alpha_1 \geq \alpha_0 > 0 , \quad \alpha_n > 0 ,$$

then the number of zeros of $p(z)$ in $R_4 \leq |z| \leq \frac{1}{2}$ does not exceed

$$1 + \frac{1}{\log 2} \log \frac{\alpha_n + \sum_{i=0}^{n} |\beta_i|}{|a_0|}$$

where R_4 is the same as defined in Lemma 3 .

Theorem 4, follows from Theorem C and Lemma 3. If $\beta_i = 0$ for $i = 0, 1, \ldots, n$ then it gives a refinement of Theorem A otherwise it is a refinement of Theorem C.

2. LEMMAS

LEMMA 1. Let $p(z) = \sum_{i=0}^{n} a_i z^i$ be a polynomial of degree n such that $|\arg a_i - \beta| \leq \alpha \leq \pi/2$ for $i = 0, 1, \ldots, n$ and for some real β , then for some $t > 0$

$$|ta_i - a_{i-1}| \leq |t|a_i| - |a_{i-1}||\text{Cos } \alpha + (t|a_i| + |a_{i-1}|)\text{Sin } \alpha .$$

The proof of the above lemma is omitted as it follows immediately from the Lemma in [3] .

LEMMA 2. Let $p(z)$ be the same as defined in Theorem B. Then $p(z)$ has all its zeros in the ring shaped region given by

$$R_2 \leq |z| \leq R_1 .$$

Here

$$R_1 = \frac{c}{2}\left(\frac{1}{|a_n|} - \frac{1}{M_1}\right) + \left\{ \frac{c^2}{4}\left(\frac{1}{|a_n|} - \frac{1}{M_1}\right)^2 + \frac{M_1}{|a_n|} \right\}^{\frac{1}{2}}$$

and

$$R_2 = \frac{1}{2M_2^2}\left[-R_1^2|b|(M_2 - |a_0|) + \left\{4|a_0|R_1^2M_2^3 + R_1^4|b|^2(M_2 - |a_0|)^2\right\}^{\frac{1}{2}}\right]$$

where

$$M_1 = |a_n|(\text{Cos } \alpha + \text{Sin } \alpha) + 2 \text{ Sin } \alpha \sum_{k=0}^{n-1} |a_k| ,$$

$$M_2 = |a_n|R_1^n\left[\frac{2 \text{ Sin } \alpha}{|a_n|}\sum_{k=0}^{n-1} |a_k| + R_1\left(1 - \frac{|a_n|}{|a_n|}\right)(\text{Cos } \alpha + \text{Sin } \alpha)\right] ,$$

$$c = |a_n - a_{n-1}| ,$$

$$b = a_1 - a_0$$

LEMMA 3. Let $p(z)$ be defined as in Theorem C . Then $p(z)$ has all its zeros in the ring shaped region given by

$$R_4 \leq |z| \leq R_3 .$$

Here

$$R_3 = \frac{c}{2}\left(\frac{1}{\alpha_n} - \frac{1}{M_3}\right) + \left\{ \frac{c^2}{4}\left(\frac{1}{\alpha_n} - \frac{1}{M_3}\right)^2 + \frac{M_3}{\alpha_n} \right\}^{\frac{1}{2}}$$

and

$$R_4 = \frac{1}{2M_4^2}\left[-R_3^2|b|(M_4 - |a_0|) + \left\{ 4|a_0|R_3^2 M_4^3 + R_3^4 |b|^2 (M_4 - |a_0|)^2 \right\}^{\frac{1}{2}} \right] ,$$

where

$$M_3 = \alpha_n R ,$$

$$R = 1 + \frac{1}{\alpha_n}\left[2 \sum_{k=0}^{n-1} |\beta_k| + |\beta_n| \right] ,$$

$$M_4 = R_3^n \left[(\alpha_n + |\beta_n|)R_3 + \alpha_n R - (\alpha_0 + |\beta_0|) \right] ,$$

$$c = |a_n - a_{n-1}| ,$$

$$b = a_1 - a_0 .$$

Lemmas 2 and 3 are due to Govil and Jain [2] .

3. PROOFS OF THE THEOREMS

Proof of Theorem 1. Consider

$$F(z) = (t - z)p(z)$$

$$= (t - z)(a_0 + a_1 z + \ldots + a_n z^n)$$

$$= -a_n z^{n+1} + ta_0 + \sum_{i=1}^{n} (ta_i - a_{i-1}) z^i$$

For $|z| \leq t \leq 1$,

$$|F(z)| \leq t^{n+1} |a_n| + t|a_0| + \sum_{i=1}^{n} |t|a_i| - |a_{i-1}||t^i \text{ Cos } \alpha$$

$$+ \sum_{i=1}^{n} (t|a_i| + |a_{i-1}|)\, t^i \text{ Sin } \alpha \text{ , (by Lemma 1)}$$

$$\leq t^{n+1} |a_n| + t|a_0| + \sum_{i=1}^{k} (t|a_i| - |a_{i-1}|)t^i \text{ Cos } \alpha$$

$$+ \sum_{i=k+1}^{n} (|a_{i-1}| - t|a_i|)t^i \text{ Cos } \alpha$$

$$+ \sum_{i=1}^{n} (t|a_i| + |a_{i-1}|)t^i \text{ Sin } \alpha$$

$$= 2t^{k+1} |a_k| \text{ Cos } \alpha + 2t \text{ Sin } \alpha \sum_{i=0}^{n} t^i |a_i|$$

$$- t|a_0|(\text{Cos } \alpha + \text{Sin } \alpha - 1) - t^{n+1} |a_n|(\text{Cos } \alpha + \text{Sin } \alpha - 1)$$

$$\leq 2t^{k+1}|a_k| \text{ Cos } \alpha + 2t \text{ Sin } \alpha \sum_{i=0}^{n} t^i |a_i|$$

$$- t^{n+1} |a_n|(\text{Cos } \alpha + \text{Sin } \alpha - 1).$$

Now it is known (see [5], p. 171) that if $P(z)$ is regular, $P(0) \neq 0$ and $|F(z)| \leq M$ in $|z| \leq 1$ then the number of zeros of $P(z)$ in $|z| \leq \frac{1}{2}$ does not exceed $\frac{1}{\log 2} \{\log \frac{M}{|P(0)|}\}$. Applying this result to $F(z)$, we get that the number of zeros of $F(z)$ in $|z| \leq \frac{1}{2}$ does not exceed

$$\frac{1}{\log 2}\left\{\log \frac{2t^{k+1}|a_k|\cos\alpha + 2t\sin\alpha \sum_{i=0}^{n} t^i|a_i| - (\cos\alpha \quad \sin\alpha - 1)|a_n|t^{n+1}}{t|a_0|}\right\}.$$

As the number of zeros of $p(z)$ in $|z| \leq \frac{1}{2}$ does not exceed the number of zeros of $F(z)$ in $|z| \leq \frac{1}{2}$, the theorem follows.

Proof of Theorem 2. Consider

$$F(z) = (t - z)\,p(z)$$

$$= -a_n z^{n+1} + ta_0 + \sum_{i=1}^{n}(ta_i - a_{i-1})z^i$$

For $|z| \leq t \leq 1$,

$$|F(z)| \leq t^{n+1}|a_n| + t|a_0| + \sum_{i=1}^{n}|ta_i - a_{i-1}|t^i$$

$$\leq t^{n+1}|a_n| + t|a_0| + \sum_{i=1}^{n}|t\alpha_i - \alpha_{i-1}|t^i$$

$$+ \sum_{i=1}^{n}(|\beta_{i-1}| + t|\beta_i|)t^i$$

$$\leq t^{n+1}|a_n| + t|a_0| + \sum_{i=1}^{k}(t\alpha_i - \alpha_{i-1})t^i$$

$$+ \sum_{i=k+1}^{n}(\alpha_{i-1} - t\alpha_i)t^i + \sum_{i=1}^{n}(|\beta_{i-1}| + t|\beta_i|)t^i$$

$$\leq t^{n+1}|a_n| + t|a_0| + 2t^{k+1}\alpha_k - t\alpha_0 - t^{n+1}\alpha_n + t|\beta_0|$$

$$- t^{n+1}|\beta_n| + 2t\sum_{i=1}^{n} t^i|\beta_i|$$

$$\leq 2\left(t^{k+1}\alpha_k + t\sum_{i=0}^{n} t^i|\beta_i|\right)$$

and following on the lines of the proof of Theorem 1, the proof of Theorem 2 can be completed.

REFERENCES

1. K.K. Dewan: Extremal properties and coefficient estimates for polynomials with restricted zeros and on location of zeros of polynomials. Ph.D. Thesis, I.I.T., Delhi, New Delhi, 1980.

2. N.K. Govil and V.K. Jain: On the Eneström-Kakeya Theorem II. Jour. of Approx. Theory 22 (1978), 1-10.

3. N.K. Govil and Q.I. Rahman: On the Eneström-Kakeya Theorem. Tôhoku Math. J. 20 (1968), 126-136.

4. Q.G. Mohammad: On the zeros of the polynomials. Amer. Math. Monthly, 72 (1965), 631-633.

5. E.C. Titchmarsh: The theory of functions, 2nd ed., Oxford University Press, London, 1939.

A POSTERIORI ERROR BOUNDS FOR EIGENSYSTEMS OF MATRICES

Z. BOHTE

ABSTRACT: In this paper an a posteriori error bound for approximate eigenvectors corresponding to simple eigenvalues of non-defective matrices is obtained. Under some additional assumptions the computable bound for the condition number is derived. Some illustrative numerical examples are given.

1. INTRODUCTION

A posteriori error bounds for computed eigenvalues of non-defective matrices and for computed eigenvectors of normal matrices are well-known (see [4]).

Let us summarize some of these known results.

Throughout this paper let A be a non-defective square complex matrix of order n and denote its eigenpairs by (λ_i, x_i), so that

$$Ax_i = \lambda_i x_i , \quad i = 1, \ldots , n \tag{1}$$

Denote by X the matrix of eigenvectors

$$X = [x_1, \ldots , x_n]$$

which is by assumption non-singular.

Let (λ, x) be an approximate eigenpair, usually computed by some numerical method, and let

$$r = Ax - \lambda x \tag{2}$$

be the corresponding residual vector. Then there exists an eigenvalue of the matrix A such that

$$\min_{1 \leq i \leq n} |\lambda_i - \lambda| \leq k(A) \|r\| / \|x\| \tag{3}$$

where

$$k(A) = \|X\| \, \|X^{-1}\| \tag{4}$$

The bound (3) holds for any of the norms 1, 2 or ∞. The number $k(A)$ is called the condition number of the matrix A with respect to the eigenvalue problem. For normal matrices $k_2(A) = 1$ and (3) gives the most satisfactory and easily computable a posteriori bound.

For normal matrices Wilkinson [4] gives the corresponding a posteriori error bound for the approximate eigenvector x. Let λ be an approximation to λ_1, let x_1 and x be normalized so that

$$\|x_1\|_2 = \|x\|_2 = 1 \tag{5}$$

and suppose that x_1 is multiplied by such a complex factor of modulus 1 that in

$$x = a_1 x_1 + \ldots + a_n x_n \tag{6}$$

the coefficient a_1 is non-negative:

$$a_1 \geq 0 \tag{7}$$

Further, let

$$d = \min_{2 \leq i \leq n} |\lambda_i - \lambda| \neq 0 \tag{8}$$

then

$$\|x - x_1\|_2 \leq (c/d)(1 + (c/d)^2)^{1/2} \tag{9}$$

where

$$c = \|r\|_2 \tag{10}$$

To use (9) in practice we need some information about other eigenvalues so that we can estimate the distance d from bel Unless c is significantly less then d, (9) provides no useful boun

Let us now consider a general non-defective matrix.

2. ERROR BOUND FOR APPROXIMATE EIGENVECTOR

Under the same conditions as above we shall prove that for the general non-defective matrix the bound for the error in the approximate eigenvector x is

$$\|x - x_1\|_2 \leq 2k_2(A)c/d \tag{11}$$

From (1), (2) and (6) it follows

$$r = (\lambda_1 - \lambda)a_1x_1 + \ldots + (\lambda_n - \lambda)a_nx_n$$

If we define

$$D = \mathrm{diag}(0, (\lambda_2 - \lambda)^{-1}, \ldots, (\lambda_n - \lambda)^{-1})$$

we have

$$XDX^{-1}r = a_2x_2 + \ldots + a_nx_n = u \tag{12}$$

and

$$\|u\|_2 \leq \|r\|_2 \|D\|_2 k_2(A) \tag{13}$$

Clearly,

$$\|D\|_2 = 1/d$$

where d is defined by (8). Using the notation (4) and (10) we can write the bound (13) in the form

$$\|u\|_2 \leq k_2(A)c/d \tag{14}$$

Further, under the assumptions (5) - (7) we have

$$\|a_1x_1\|_2 = a_1 \tag{15}$$

Since

$$x - x_1 = (a_1 - 1)x_1 + u$$

where u is defined by (12), we have

$$\|x - x_1\|_2 \leq |a_1 - 1| + \|u\|_2 \tag{16}$$

On the other hand

$$a_1x_1 = x - u$$

and using (15) and (5) we have

$$1 - \|u\|_2 \leq a_1 \leq 1 + \|u\|_2$$

From this two-sided inequality it follows

$$|a_1 - 1| \leq \|u\|_2 \tag{17}$$

The bound (11) follows directly from (16), (17) and (14).

For normal matrices the bound (11) is slightly weaker then the bound (9) where the orthogonality of eigenvectors has been taken into account.

In order to be able to use the bound (11) in practice we need also approximations to all other eigenvectors. The practical difficulty is that we must calculate an upper bound for $k_2(A)$ and a lower bound for d from an approximate eigensystem.

3. THE COMPUTABLE UPPER BOUND FOR THE CONDITION NUMBER

Let us denote by k the spectral condition number

$$k = k_2(A) = \|X\|_2 \|X^{-1}\|_2$$

In order to be able to compute a reliable upper bound for k we shall make a number of additional assumptions.

First, suppose that all the eigenvalues λ_i are simple and that we have calculated an approximate eigensystem (μ_i, y_i), $i = 1, \ldots, n$. Let all eigenvectors x_i and their approximations y_i be normalized

$$\|x_i\|_2 = \|y_i\|_2 = 1 , \quad i = 1, \ldots, n$$

and similarly to (6) and (7) we suppose that x_i are such that in

$$y_i = a_1^{(i)} x_1 + \ldots + a_i^{(i)} x_i + \ldots + a_n^{(i)} x_n$$

all

$$a_i^{(i)} \geq 0$$

Denote the matrix of approximate eigenvectors by

$$Y = [y_1, \ldots, y_n]$$

Then, clearly an approximation to k is the number

$$a = \|Y\|_E \|Z\|_E = \sqrt{n}\, \|Z\|_E \tag{18}$$

but it may not be an upper bound for it. This may happen because Y is only an approximation to X and it may be ill-conditioned and Z may be a poor approximation to Y^{-1}.

We shall have to calculate all the residual vectors

$$r_i = Ay_i - \mu_i y_i , \quad i = 1, \ldots , n$$

Denote

$$r = \max_{1 \leq i \leq n} \|r_i\|_2$$

and

$$m = \min_{i \neq j} |\mu_i - \mu_j|$$

Now, let us make the main assumption, that all the circles

$$C_i = (\mu_i, rk) , \quad i = 1, \ldots , n$$

with the centres μ_i and radii rk in the complex plane are disjoint. This means that in every one of them lies exactly one eigenvalue of the matrix A. We call λ_i the eigenvalue of A lying in C_i and from (3) we have the bounds

$$|\lambda_i - \mu_i| \leq rk , \quad i = 1, \ldots , n \tag{19}$$

To obtain the bounds for the errors in y_i we need a lower bound for

$$d_i = \min_{j \neq i} |\mu_i - \lambda_j| , \quad i = 1, \ldots , n$$

and clearly it follows from (19) that

$$d_i \geq m - rk = e , \quad i = 1, \ldots , n \tag{20}$$

From the assumption that all C_i are disjoint it is obvious that

$$e > 0$$

Therefore it follows from (11)

$$\|x_i - y_i\|_2 \leq 2 \|r_i\|_2 k/e , \quad i = 1, \ldots , n$$

These inequalities may be written in the form

$$\|X - Y\|_E \leq 2 \|R\|_E k/e \tag{21}$$

where R is the residual matrix

$$R = [r_1, \ldots , r_n]$$

Because

(22) $$k \leq \|X\|_E \|X^{-1}\|_E$$

and

(23) $$\|X\|_E = \sqrt{n}$$

we need only a bound for $\|X^{-1}\|_E$.

Let us denote

$$F = YZ - I$$

Then,

(24) $$E = XZ - I = F + (X - Y)Z$$

and using (21) we have the bound

(25) $$\|E\|_E \leq \|F\|_E + 2\|R\|_E \|Z\|_E k/e = g$$

Suppose that

(26) $$g < 1$$

This means that the matrix A should not be too ill-conditioned with respect to other terms in the right-hand side of (25). From (24) - (26) it follows directly

(27) $$\|X^{-1}\|_E \leq \|Z\|_E/(1 - g)$$

and we have the final inequality from (22), (23), (27), and (20)

(28) $$k \leq \sqrt{n}\|Z\|_E/(1 - \|F\|_E - 2\|R\|_E \|Z\|_E k/(m - rk))$$

Under the assumptions (20) and (26) both denominators on the right-hand side of (28) are positive.

Denoting

$$b = 1 - \|F\|_E , \quad c = 2\|R\|_E \|Z\|_E$$

and recalling (18) we can write (28) in the form

$$k \leq a/(b - ck/(m - rk)) = a(m - rk)/(bm - (c + br)k)$$

leading to the quadratic inequality

(29) $$(c + br)k^2 - (ar + bm)k + am \geq 0$$

For the exact eigensystem $r = c = 0$ and we obtain from (29) an

obvious bound

$$k \leq a/b$$

where the only errors are made in the computation of the inverse of the matrix of eigenvectors.

From (29) we obtain the bound

$$k \leq (p - (p^2 - 4amq)^{1/2})/(2q) = K \tag{30}$$

where

$$p = ar + bm, \quad q = c + br$$

This bound can be computed directly from approximate eigensystems. It can be shown that for sufficiently small r the number K is greater then 1 and gives therefore a useful bound for the condition number $k_2(A)$. It may happen, of course, that the number K is complex and then we have no bound for k.

The bound (30) can be used in the bounds (3) and (9) for individual eigenpairs. For the errors in eigenvalues we have from (3)

$$|\lambda_i - \mu_i| \leq K \|r_i\|_2 \tag{31}$$

and for the errors in eigenvectors we have from (9)

$$\|x_i - y_i\|_2 \leq 2K \|r_i\|_2 / g_i \tag{32}$$

where

$$g_i = \min_{j \neq i} (|\mu_i - \mu_j| - K \|r_j\|_2)$$

We must remember that these bounds hold provided all the above assumptions are fulfilled.

4. NUMERICAL EXAMPLES

Let us illustrate the obtained bounds by some simple examples of matrices of order $n = 3$.

(i) The matrix

$$A = \begin{bmatrix} 3 & 5 & -16 \\ 6 & 12 & -36 \\ 2 & 5 & -15 \end{bmatrix}$$

has eigenvalues

$$\lambda_1 = 1, \quad \lambda_2 = 2, \quad \lambda_3 = -3$$

and the spectral condition number

$$k_2(A) = 32{\cdot}14\ldots$$

If we take

$$\mu_1 = 1 + 10^{-5}, \quad \mu_2 = 2 + 2.10^{-5}, \quad \mu_3 = 3 - 3.10^{-5}$$

and for eigenvectors the correct eigenvectors rounded to 5 decimal places and also the correct inverse to 5 places, we obtain the bound

$$K = 52{\cdot}44\ldots$$

By the way, the number

$$a = 33{\cdot}45\ldots$$

is a very good approximation to $k_2(A)$. The bounds (31) and (32) are severe overestimates in this case. For instance,

$$|\lambda_1 - \mu_1| = 10^{-5}$$

but

$$K \|r_1\|_2 = 910.10^{-5}$$

and

$$\|x_1 - y_1\|_2 = 0{\cdot}52.10^{-5}$$

but

$$2K \|r_1\|_2 / g_1 = 1836.10^{-5}$$

(ii) The matrix

$$A = \begin{bmatrix} 19 & 7 & -2 \\ 10 & 16 & -2 \\ 4 & -8 & 19 \end{bmatrix}$$

has eigenvalues

$$\lambda_1 = 9, \quad \lambda_2 = 18, \quad \lambda_3 = 27$$

and

$$k_2(A) = 1 + \sqrt{2} = 2{\cdot}41\ldots$$

If we take

$$\mu_1 = 9 + 3.10^{-5}, \quad \mu_2 = 18 + 5.10^{-5}, \quad \mu_3 = 27 + 6.10^{-5}$$

and similarly round the eigenvectors and the inverse matrix, we obtain the bounds for the condition number

$$K = 3{\cdot}87\ldots$$

for the error in the third eigenvalue

$$K \|r_3\|_2 = 24.10^{-5}$$

and for the error in the third eigenvector

$$2K \|r_3\|_2 / g_3 = 5.10^{-5}$$

which is very satisfactory. The approximate condition number a is almost the same 3·87.

(iii) The upper triangular matrix

$$A = \begin{bmatrix} 1 & 100 & 0 \\ & 2 & 0 \\ & & 100 \end{bmatrix}$$

with eigenvalues

$$\lambda_1 = 1, \quad \lambda_2 = 2, \quad \lambda_3 = 100$$

is very ill-conditioned, namely

$$k_2(A) = 200{\cdot}05\ldots$$

With

$$\mu_1 = 1 + 10^{-5}, \quad \mu_2 = 2 + 2.10^{-5}, \quad \mu_3 = 100 + 10^{-3}$$

and rounded eigenvectors we get a complex number K and cannot use any of the bounds for the errors. The approximate condition number a is equal to 245.

5. CONCLUSIONS

It is rarely justified to use expensive a posteriori bounds which are usually too pessimistic. But compared to a posteriori bounds for the solution of the system of linear algebraic equations where the bound is approximately 6 times more expensive as the solution by the

Gauss elimination, here, even with the most economic methods (e.g. the QR method), the additional number of arithmetic operations $6n^3$ is not worrying. Of course, for practical use, some sort of iterative improvement of the approximate eigensystem is more desirable (see [3]).

Recently Chu [2] generalized the Bauer-Fike theorem [1] to defective matrices. Along these lines it would be worthwhile to attempt finding a posteriori bounds for the computed eigenvalues and eigenvectors using the Schur form.

Acknowledgement. I wish to express my sincere thanks to I. Vidav who proved the bound (11).

REFERENCES:

1. F. L. Bauer and C. T. Fike: Norms and exclusion theorems, Numer. Math. 2 (1960), 42 - 53.

2. K. E. Chu: Generalization of the Bauer-Fike theorem, Numer. Math. 9 (1986), 685 - 691.

3. H. J. Symm and J. H. Wilkinson: Realistic error bounds for a simple eigenvalue and its associated eigenvector, Numer. Math. 35 (1980), 113 - 126.

4. J. H. Wilkinson: The algebraic eigenvalue problem, Clarendon Press, Oxford 1965.

ON THE UNIFORM CONVERGENCE OF MODIFIED GAUSSIAN RULES FOR THE NUMERICAL EVALUATION OF DERIVATIVES OF PRINCIPAL VALUE INTEGRALS

G. CRISCUOLO and G. MASTROIANNI

ABSTRACT: *The authors prove some convergence theorems of a modified gaussian rule for the evaluation of the derivatives of Cauchy principal value integrals.*

1. INTRODUCTION

Let $\Phi(wf;t)$ denote the integral in the Cauchy principal value sense of the function f, associated with the weight w and defined by

$$(1)\quad \Phi(wf;t)=\int_{-1}^{1} \frac{f(x)}{x-t}\, w(x)dx=\lim_{\varepsilon\to 0^+}\int_{|x-t|\geq\varepsilon} \frac{f(x)}{x-t}\, w(x)dx\ ,\ -1<t<1.$$

In order to approximate the integral (1) we may consider the gaussian rule

$$\Phi_m(wf;t)=f(x)\int_{-1}^{1} \frac{w(x)}{x-t}\,dx+\sum_{i=1}^{m}\lambda_{m,i}\frac{f(x_{m,i})-f(t)}{x_{m,i}-t}\ ,\ t\neq x_{m,i},\ i=1,2,\ldots,m\ ,$$

where $x_{m,i}$, i=1,2,...m, are the zeros of the m-th orthogonal polynomial associated with the function w and $\lambda_{m,i}$, i=1,2,...m, are the Christoffel constants.

If the function f is "sufficiently smooth", then the sequence $\{\Phi_m(wf;t)\}$converges to $\Phi(wf;t)$. Furthermore, it is easy to prove that the inequality

$$|\Phi(wf;t)-\Phi_m(wf;t)| \leq \text{const } m^{-k}\ \omega(f^{(k)};m^{-1})\log m\ ,\quad f\in C^k(I),\ k\geq 1,$$

hold on every closed interval $\Delta\subset(-1,1)$.

Unfortunately, in the general rule, if f is an Hölder continuous function, then $\{\Phi_m(wf;t)\}$ does not converge to $\Phi(wf;t)$ almost everywhere in (-1,1), (see[4]).

In order to avoid this problem, the authors introduced in [1] a new formula $\Phi_m^*(wf;t)$; this is defined by

$$(2)\qquad \Phi_m^*(wf;t)=f(t)\int_{-1}^{1}\frac{w(x)}{x-t}\,dx+\sum_{\substack{i=1\\ i\neq c}}^{m}\lambda_{m,i}\,\frac{f(x_{m,i})-f(t)}{x_{m,i}-t}\ ,\quad m\in N\ ,$$

where c denotes the index corresponding to the "closest knot" $x_{c(m)}=x_{m,c}$ to the singularity t, defined by $|t-x_{m,c}|=\min\{t-x_{m,d},x_{m,d+1}-t\}$, $x_{m,d}\leq t\leq x_{m,d+1}$ for some $d\in\{0,1,\ldots,m\}$ with $x_{m,o}=-1$, $x_{m,m+1}=1$.

The "modified gaussian rule" $\Phi_m^*(wf;t)$ has degree of exactness 0; nevertheless the hypothesis $x_{m,i}\neq t$, $i=1,2,\ldots,m$ becomes unnecessary.

Notice that the derivative $\frac{d}{dt}\,\Phi(wf;t)$ appears in some integrodifferential equations concerning several branches of physics and engineering.

Further, the analytic solution of the integral equations with logarithmic singularities in the kernel may be represented by the derivatives of Cauchy principal value integrals.

In this paper we study the uniform convergence of the sequence

$$\left\{\frac{d^p}{dt^p}\,\Phi_m^*(wf;t)\right\}\ \text{to}\ \frac{d^p}{dt^p}\,\Phi(wf;t)\ \text{on}\ (-1,1)\ \text{for}\ p\geq 0.$$

This is of interest in solving singular integral equations with a collocation method too. Indeed, uniform convergence results of a quadrature rule on the whole interval (-1,1) are necessary to study the convergence of the method when, for example, the collocation points are zeros of orthogonal polynomials in [-1,1].

The convergence theorems are stated in the Section 2; they generalize

previous results [2] and are proved in the Section 3.

2. CONVERGENCE THEOREMS AND ESTIMATES OF THE REMAINDER

We start with some notation. Throughout this paper DT denotes the space of the continuous functions in I:=[-1,1] satisfying a "Dini type" condition, and $\mathrm{Lip}_M\,\lambda$ the space of the Hölder continuous functions; i.e.:

$$DT:=\{f\in C(I)/\ \int_0^1 \delta^{-1}\,\omega(f;\delta)d\delta < \infty\}$$

$$\mathrm{Lip}_M\lambda:=\{f\in C(I)/\ \omega(f;\delta) \leq M\delta^{\lambda}\ ,\ M>0,\ 0<\lambda\leq 1\}$$

where $\omega(f;\delta)=\max_{|x-y|<\delta} |f(x)-f(y)|$, $x,y\in I$, $\delta\geq 0$, is the modulus of continuity of the function f. We ought to remark that $DT\supset \mathrm{Lip}_M\lambda$.

In the computation of the integral $\Phi(wf;t)$ defined by (1) we suppose that the weight function w can be written in the form $w(x)=\psi(x)\,u^{\alpha,\beta}(x)$, $x\in I$, with $u^{\alpha,\beta}(x)=(1-x)^{\alpha}(1+x)^{\beta}$, $\alpha,\beta>-1$ and $0<\psi\in DT$.

Let $\{P_m(w)\}$ be the sequence of the orthonormal polynomials on I associated with the weight function w; we denote the zeros of

$$P_m(x)=P_m(w;x)=\alpha_m x^m + \text{lower degree terms}\ ,\quad \alpha_m>0,$$

by $x_{m,i}=x_{m,i}(w)=\cos\,\Theta_{m,i}$, $i=1,2,\ldots,m$, so that

$$0=\Theta_{m,m+1}<\Theta_{m,m}<\ldots<\Theta_{m,2}<\Theta_{m,1}<\Theta_{m,o}=\pi .$$

Furthermore, the numbers $\lambda_{m,i}=\lambda_{m,i}(w)$, $i=1,2,\ldots m$, are the Christoffel constants defined by $\lambda_{m,i}(w)=\lambda_m(w;x_{m,i})$ where $\lambda_m(w;x)=\left[\sum_{k=o}^{m-1} P_k^2(w;x)\right]^{-1}$ is the m-th Christoffel function.

Denoting by $E_m^*(wf)=\Phi(wf)-\Phi_m^*(wf)$ the remainder term of the formula $\Phi_m^*(wf)$ defined by (2), we can state the following

THEOREM 1.

If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $\alpha,\beta\geq 0$, then for any function f such that $f^{(p)}\in DT$ the sequence $\left\{\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)\right\}$ converges uniformly to $\frac{d^p}{dt^p}\ \Phi(wf;t)$ on the whole open interval $(-1,1)$ $p\geq 0$.

Moreover, if $f^{(p)}\in Lip_M\lambda$, $0<\lambda\leq 1$, it is also

$$\left|\frac{d^p}{dt^p}\ E_m^*(wf;t)\right| \leq \text{const } m^{-\lambda}\ \log m, \quad -1<t<1,\ p\geq 0$$

THEOREM 2.

If $w=\psi u^{\alpha\ \beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-1<\alpha,\beta<0$, then for any function f such that $f^{(p)}\in Lip_M\lambda$, $0<\lambda\leq 1$, it results

$$(3)\qquad \left|\frac{d^p}{dt^p}\ E_m^*(wf;t)\right| \leq \text{const } m^{-\lambda} u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\ \log\frac{m}{\sqrt{1-t^2}},\quad -1<t<1,\ p\geq 0$$

In particular, if $\alpha+\lambda/2,\beta+\lambda/2\geq 0$ then by Theorem 2 it follows

Corollary 3. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-\frac{1}{2}<\alpha,\beta<0$, then for any function f such that $f^{(p)}\in Lip_M\lambda$, $-2\ \min(\alpha,\beta)<\lambda\leq 1$, the sequence $\left\{\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)\right\}$ converges uniformly to $\frac{d^p}{dt^p}\ \Phi(wf;t)$ on the whole open interval $(-1,1)$, $p\geq 0$.

Moreover, taking into account (3), it seems that the sequence $\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)$ can not converge uniformly on $(-1,1)$ for $\alpha,\beta\leq -\frac{1}{2}$, generally. Nevertheless, a favourable case of interest in the applications comes true when $t\in\Delta_m:=[-1+\text{const } m^{-2},\ 1-\text{const } m^{-2}]$. In fact, by Theorem 2 we deduce also

Corollary 4. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-1<\alpha,\beta<0$, then for any function f such that $f^{(p)}\in Lip_M\lambda$, $-2\min(\alpha,\beta)\leq\lambda\leq 1$, the sequence $\left\{\frac{d^p}{dt^p}\ \Phi_m^*(wf;t)\right\}$ converges uniformly to $\frac{d^p}{dt^p}\ \Phi(wf;t)$ on Δ_m, $p\geq 0$.

Moreover, it results

$$\left|\frac{d^p}{dt^p}\ E_m^*(wf;t)\right| \leq \text{const } m^{-\lambda}\ u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\ \log m,\quad t\in\Delta_m,\ p\geq 0.$$

orollary 5. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, $-1<\alpha,\beta<0$, then for any function f uch that $f^{(p)}\in Lip_M\lambda$, $-\min(\alpha,\beta)=\gamma<\lambda<-2\min(\alpha,\beta)$, the sequence $\left\{\frac{d^p}{dt^p}\phi_m^*(wf;t)\right\}$ onverges uniformly to $\frac{d^p}{dt^p}\Phi(wf;t)$ on Δ_m, $p\geq 0$.

loreover, it results

$$\left|\frac{d^p}{dt^p}E_m^*(wf;t)\right| \leq \text{const}\ \frac{\log m}{m^{2(\lambda-\gamma)}}\ , \quad t\in\Delta_m, \quad p\geq 0.$$

3. PROOF SKETCH OF THE MAIN RESULTS

For the convenience of the reader, we collect some properties of the orthonormal polynomials $P_m(w)$ with $w(x)=\psi(x)u^{\alpha,\beta}(x)$, $-1\leq x\leq 1$, $u^{\alpha,\beta}(x)=(1-x)^\alpha(1+x)^\beta$, $\alpha,\beta>-1$, $0<\psi\in DT$, which will be used in the following.

The equivalence

(4) $\quad \Theta_{m,k} - \Theta_{m,k+1} \sim m^{-1}$, uniformly for $0\leq k\leq m$, $m\in N$,

(5) $\quad \lambda_{m,k} \sim m^{-1}\ u^{\alpha+1/2,\beta+1/2}(x_{m,k})$, uniformly for $1\leq k\leq m$, $m\in N$

holds for the zeros of $P_m(w)$ and for the Christoffel constants respectively. One can find the relations (4) and (5) in [3].

Furthermore, it follows from (4) that

(6) $\quad u^{\alpha,\beta}(t) \sim u^{\alpha,\beta}(x_{m,k})\ , \qquad x_{m,k} \leq t \leq x_{m,k+1}$,

for $k=2,3,\ldots,m-1$ (see[3,p.48]).

To derive the proofs of the theorems stated in the previous section, the following lemmas are needed.

Lemma 1. If $f\in C^r(I)$, $r\geq 0$, then for each $m\in N$ there exists a polynomial t_m of degree at most $m\geq 4(r+1)$ such that

$$\left|f^{(k)}(x)-t_m^{(k)}(x)\right|\leq \text{const}\left[m^{-1}\sqrt{1-x^2}\right]^{r-k}\omega(f^{(r)};m^{-1}\sqrt{1-x^2}),0\leq k\leq r,\ -1\leq x\leq 1,$$

$$\left|t_m^{(p)}(x)\right|\leq \text{const}\,[\Delta_m(x)]^{r-p}\omega(f^{(r)};\ \Delta_m(x)),\ p>r,\quad -1\leq x\leq 1,$$

where $\Delta_m(x)=m^{-1}\sqrt{1-x^2}+m^{-2}$.

Lemma 2. If $w=\psi u^{\alpha,\beta}$, $0<\psi\in DT$, $\alpha,\beta>-1$ then for any function $f\in C(I)$ the inequality

$$\lambda_{m,c}\left|\frac{t_m(x_{m,c})-t_m(t)}{x_{m,c}-t}\right|\leq \text{const}\left[\sqrt{1-t}+m^{-1}\right]^{2\alpha}\left[\sqrt{1+t}+m^{-1}\right]^{2\beta}\omega(f;\Delta_m(t)),$$

holds uniformly for $t\in(-1,1)$, where t_m is the polynomial of Lemma 1, $x_{m,c}$ is the closest knot to the point t, and $\Delta_m(t)=m^{-1}\sqrt{1-t^2}+m^{-2}$.

Setting $\sigma_m^*(t)=\sum\limits_{\substack{i=1\\ i\neq c}}^{m}\frac{\lambda_{m,i}}{|x_{m,i}-t|}$, we can state

Lemma 3. If $w=\psi u^{\alpha,\beta}$, $0<\psi\in DT$, then the inequalities

$$\sigma_m^*(t)\leq \text{const}\ \log m\ ,\quad \text{if } \alpha,\beta\geq 0,$$

$$\sigma_m^*(t)\leq \text{const}\ u^{\alpha,\beta}(t)\ \log m\ ,\quad \text{if } -1<\alpha,\beta<0,$$

hold uniformly for $t\in(-1,1)$.

Lemma 4. If $w=\psi u^{\alpha,\beta}$, $0<\psi\in DT$, then the inequalities

$$\sum_{\substack{i=1\\ i\neq c}}^{m}\lambda_{m,i}\frac{|r_m(x_{m,i})-r_m(t)|}{|x_{m,i}-t|}\leq \text{const}\begin{cases}\omega(t;m^{-1})\log m, & \text{if } \alpha,\beta\geq 0,\quad f\in C(I)\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log m, & \\ \text{if } -1<\alpha,\beta<0, & f\in \text{Lip}_M\lambda,\ 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\varepsilon(-1,1)$, where $r_m=f-t_m$, being t_m the polynomial of Lemma 1.

Lemma 5. If $w=\psi u^{\alpha,\beta}$, $0<\psi\varepsilon DT$, then the inequalities

$$\int_{-1}^{1}\left|\frac{r_m(x)-r_m(t)}{x-t}\right|w(x)dx\leq const\begin{cases}\omega(f;m^{-1})\log m\ , & \text{if } \alpha,\beta\geq 0,\quad f\varepsilon C(I),\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log\frac{m}{\sqrt{1-t^2}} & \\ \text{if } -1<\alpha,\beta<0,\quad f\varepsilon Lip_M\lambda,\ 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\varepsilon(-1,1)$, where $r_m=f-t_m$ being t_m the polynomial of Lemma 1.

Lemma 1 can be found in [5]; instead, the other previous lemmas are proved in [2].

Lemma 6. Let $v\varepsilon L_1[-1,1]$, i.e. $\int_{-1}^{1}|v(x)|dx<\infty$, possibly having singularities at $v_1,\ldots,v_s$, and assume that v is continuous on each closed interval enclosed in $I-\{v_1,\ldots,v_1\}$. If g is a function such that $g^{(p)}\varepsilon DT$, $p\geq 1$, then the integral $\int_{-1}^{1}\frac{d^p}{dt^p}\left[\frac{g(x)-g(t)}{x-t}\right]v(x)dx$ exists and the identity

$$\frac{d^p}{dt^p}\int_{-1}^{1}\frac{g(x)-g(t)}{x-t}v(x)dx=\int_{-1}^{1}\frac{d^p}{dt^p}\left[\frac{g(x)-g(t)}{x-t}\right]v(x)dx,$$

holds whenever t is in a closed set enclosed in $I-\{v_1,\ldots,v_s\}$ and $p\geq 1$.

The proof of Lemma 6 is based on known results of classical analysis and elementary inequalities for the modulus of continuity.

Lemma 7. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\varepsilon DT$, $\alpha,\beta>-1$, then for any function $f\varepsilon C^p(I)$ the inequality

$$\lambda_{m,c}\left|\frac{d^p}{dt^p}\,\frac{t_m(x_{m,c})-t_m(t)}{x_{m,c}-t}\right|\leq\text{const}\left[\sqrt{1-t}+m^{-1}\right]^{2\alpha}\left[\sqrt{1+t}+m^{-1}\right]^{2\beta}\omega(f^{(p)};\,\Delta_m(t)),$$

holds uniformly for $t\in(-1,1)$, $p\geq 0$, where t_m is the polynomial of Lemma 1, $x_{m,c}$ is the closest knot to the point t, and $\Delta_m(t)=m^{-1}\sqrt{1-t^2}+m^{-2}$.

The proof of this lemma can be deduced by Lemmas 1,2,6 and applying the inequality (6).

Lemma 8. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, then the inequalities

$$\sum_{\substack{i=1\\ i\neq c}}^{m}\lambda_{m,i}\left|\frac{d^p}{dt^p}\,\frac{r_m(x_{m,i})-r_m(t)}{x_{m,i}-t}\right|\leq\text{const}\begin{cases}\omega(f^{(p)};m^{-1})\log m, & \text{if } \alpha,\beta\geq 0,\ f\in C^p(I)\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log m, & \text{if } -1<\alpha,\beta<0,\\ f^{(p)}\in \mathrm{Lip}_M\lambda, & 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\in(-1,1)$, $p\geq 0$, where $r_m=f-t_m$, being t_m the polynomial of Lemma 1.

Lemma 8 follows from Lemmas 1,3,4,6 and taking into account the relation (6).

Lemma 9. If $w=\psi u^{\alpha,\beta}$, $\psi>0$, $\psi^{(p)}\in DT$, then the inequalities

$$\left|\frac{d^p}{dt^p}\int_{-1}^{1}\frac{r_m(x)-r_m(t)}{x-t}\,w(x)dx\right|\leq\text{const}\begin{cases}\omega(f^{(p)};m^{-1})\log m, & \text{if } \alpha,\beta\geq 0,\ f\in C^p(I)\\ m^{-\lambda}u^{\alpha+\lambda/2,\beta+\lambda/2}(t)\log\frac{m}{\sqrt{1-t^2}}, & \\ \text{if } -1<\alpha,\beta<0, & f^{(p)}\in \mathrm{Lip}_M\lambda,\quad 0<\lambda\leq 1,\end{cases}$$

hold uniformly for $t\in(-1,1)$, $p\geq 0$, where $r_m=f-t_m$ being t_m, the polynomial of Lemma

Applying again the inequality (6), the proof of Lemma 9 follows from the results of Lemma 1,5,6.

Now, since rule (2) has degree of exactness 0, we have

$$\left|\frac{d^p}{dt^p} E_m^*(wf;t)\right| \leq \lambda_{m,c}\left|\frac{t_m(x_{m,c})-t_m(t)}{x_{m,c}-t}\right| + \sum_{\substack{i=1 \\ i\neq c}}^{m} \lambda_{m,i}\left|\frac{d^p}{dt^p}\,\frac{r_m(x_{m,i})-r_m(t)}{x_{m,i}-t}\right| +$$

$$+ \left|\frac{d^p}{dt^p}\int_{-1}^{1} \frac{r_m(x)-r_m(t)}{x-t}\, w(x)dx\right| ,$$

where $r_m = f - t_m$, being t_m the polynomial of Lemma 1. Thus, Theorem 1 and Theorem 2 follow from Lemmas 7,8 and 9.

REFERENCES

[1] G.CRISCUOLO and G.MASTROIANNI - On the convergence of the Gauss quadrature rules for the Cauchy Principal Value integrals.- Ricerche di Matematica XXXV (1986), 45-60.

[2] G.CRISCUOLO and G.MASTROIANNI - On the uniform convergence of gaussian quadrature rules for Cauchy Principal Value integrals.- To appear on Numeriscke Mathematik.

[3] P.NEVAI and P.VERTESI - Mean convergence of Hermite-Fejér Interpolation.- J.Math. Anal. and Appl. 105 (1985), 26-58.

[4] P.RABINOWITZ - On the convergence of Hunter's method for Cauchy Principal Value integrals. In: Numerical Solution of Singular Integral Equations (A.Gerasoulis, R.Vichmevetsky, eds.) IMACS, 1984.

[5] P.O.RUNCK - Bemerkungen zu den approximatiossätzen von Jackson und Jackson-Timan.- ISNM 10 (1969), 303-308.

APPROXIMATE EXPANSIONS OF DIFFERENTIABLE FUNCTIONS IN POLYNOMIAL SERIES

M.R. DA SILVA

ABSTRACT : One of the most important tools in applied analysis is the expansion of a given function $y = y(x)$ in a series of polynomials. If these are orthogonal, then there are explicit, well-known formulas for the expansion coefficients, but they involve quadratures which are generally difficult to perform. To avoid the evaluation of those integrals, we use a simple approximation principle which leads naturally to good polynomial approximants of y in the sense of the τ - method and is much more amenable to computer programming than Lanczos' original perturbation idea.

I. INTRODUCTION

It is well-known how important it is in applied analysis to be able to develop a given function in a series of algebraic polynomials. If these are orthogonal, then there are explicit, well-known formulas for the expansion coefficients, but they are often unsuitable for numerical evaluation, as they involve quadratures which are generally difficult to perform. For some important particular orthogonal polynomial systems we may approximate the corresponding expansion coefficients recursively, with and without numerical quadratures.

1.1. A RECURSIVE METHOD FOR THE EXPANSION COEFFICIENTS OF DIFFERENTIABLE FUNCTIONS IN SERIES OF JACOBI POLYNOMIALS IN $[0, 1]$.

Let $P_k^*(x) = P_k^{(\alpha,\beta)}(2x - 1)$, $k = 0, 1\ldots,$ $0 \le x \le 1$, $\alpha,\beta > -1$, be the standard shifted Jacobi orthogonal polynomials and assume, formally, that

$$y(x) = \sum_{k=0}^{\infty} a_k P_k^*(x) . \tag{1.1}$$

Multiplying both sides of (1.1) by $(1-x)^{\alpha} x^{\beta} P_n^*(x)$ and integrating from 0 to 1, we get

$$a_n = \frac{1}{\gamma_n} \int_0^1 (1-x)^{\alpha} x^{\beta} P_n^*(x)\, y(x)\, dx \qquad n = 0, 1, \ldots$$

$$\gamma_n = \int_0^1 (1-x)^{\alpha} x^{\beta} (P_n^*(x))^2\, dx \tag{1.2}$$

It is well-known that the fact that we can calculate a_n, $n = 0, 1, \ldots$, does not guarantee that the series in (1.1) converges, or, if the series converges, that its sum is $y(x)$.

Using Rodrigues' formula for $P_n^*(x)$,

$$P_n^*(x) = \frac{(-1)^n}{n!} \frac{D^n\{(1-x)^{\alpha+n} x^{\beta+n}\}}{(1-x)^{\alpha} x^{\beta}}, \quad D = \frac{d}{dx},$$

we obtain

$$a_n = \frac{(-1)^n}{\gamma_n n!} \int_0^1 D^n\{(1-x)^{\alpha+n} x^{\beta+n}\}\, y(x)\, dx \tag{1.3}$$

and after integrating (1.3) by parts n times,

$$a_n = \frac{(-1)^n}{\gamma_n n!} \int_0^1 (1-x)^{\alpha+n} x^{\beta+n} y^{(n)}(x)\, dx. \tag{1.4}$$

Instead of the integral transforms (1.2) - (1.4) we can solve a linear lower triangular system and obtain the coefficients a_n, $n = 0, 1, \ldots$, recursively.

Inserting Rodrigues' formula for $P_k^*(x)$ in (1.1) gives

$$(1-x)^{\alpha} x^{\beta} y(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} a_k D^k\{(1-x)^{\alpha+k} x^{\beta+k}\}. \tag{1.5}$$

Defining

$$I^{n+1} g(\xi) \equiv \int_0^{\xi} \cdots \int_0^{\xi} g(\xi)\, d\xi = \int_0^{\xi} \frac{(\xi - x)^n}{n!} g(x)\, dx,$$

then

$$I^{n+1}\{(1-\xi)^{\alpha} \xi^{\beta} y(\xi)\} = \int_0^{\xi} \frac{(\xi - x)^n}{n!} (1-x)^{\alpha} x^{\beta} y(x)\, dx. \tag{1.6}$$

Also, from (1.5),

$$I^{n+1}\{(1-\xi)^{\alpha}\,\xi^{\beta}\,y(\xi)\} = \sum_{k=0}^{n}\frac{(-1)^k}{k!}\,a_k\,I^{n-k+1}\{(1-\xi)^{\alpha+k}\,\xi^{\beta+k}\}$$

$$+\sum_{k=n+1}^{\infty}\frac{(-1)^k}{k!}\,a_k\,D^{k-n-1}\{(1-\xi^{\alpha+k}\,\xi^{\beta+k}\}$$

$$(1.7)\qquad = \sum_{k=0}^{n}\frac{(-1)^k}{k!}\,a_k\int_0^{\xi}\frac{(\xi-x)^{n-k}}{(n-k)!}\,(1-x)^{\alpha+k}.\,x^{\beta+k}\,dx+\ldots$$

Taking $\xi = 1$ and comparing (1.6) with (1.7) we get the following seemingly new formulas for the coefficients of $y(x)$ in series of Jacobi polynomials

$$(1.8)\qquad \sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{\Gamma(\alpha+n+1)\,\Gamma(\beta+k+1)}{\Gamma(\alpha+\beta+n+k+2)}\,a_k = \int_0^1(1-x)^{\alpha+n}\,x^{\beta}\,y(x)\,dx\ ,$$

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

In particular, for the shifted Legendre polynomials, $\alpha = \beta = 0$ and (1.8) leads to

$$\sum_{k=0}^{n}(-1)^k\frac{\binom{n}{k}}{\binom{n+k+1}{k}}\,a_k = (n+1)\int_0^1(1-x)^n\,y(x)\,dx\ ,\quad n = 0,\ 1,\ldots$$

1.2. POLYNOMIAL SERIES DEVELOPMENTS AND BASIC IDEA OF THE LANCZOS'S τ - METHOD

To avoid the evaluation of the integrals in (1.2), Lanczos [7,8,9] conceived the following perturbation technique, which has long been known as the τ - method.

Given an equation of the form

$$(1.9)\qquad Dy(x) = f(x)\ ,\quad a \le x \le b\ ,\quad |a|, |b| < \infty\ ,$$

where $f(x)$ is an Nth degree algebraic polynomial and D a νth order linear differential operator with polynomial coefficients,

$$(1.10)\qquad D \equiv \sum_{r=0}^{\nu} p_r(x)\,\frac{d^r}{dx^r}\ ,$$

together with ν supplementary (initial, boundary or mixed) conditions through linear combinations of function and derivative values of y, which we may write as

$$g_j(y) = \sigma_j \quad , \quad j = 1(1)\nu \quad , \tag{1.11}$$

where the g_j's are given linear functionals, the basic idea of the Lanczos' τ-method for the construction of a polynomial approximation y_n to the solution y of the problem in (1.9) - (1.11) in a form suitable for numerical evaluation is to perturb the given equation (1.9) through the addition to its r.h.s. of an algebraic polynomial H_n, usually chosen to be a linear combination of Chebyshev or Legendre polynomials with free coefficients, called the τ-parameters, which are to be determined so that y_n is the unique polynomial solution of the perturbed problem

$$Dy_n(x) = f(x) + H_n(x) \quad , \quad a \leq x \leq b \quad , \tag{1.12}$$

$$g_j(y_n) = \sigma_j \quad , \quad j = 1(1)\nu \quad , \tag{1.13}$$

to be called the τ-problem in the sequel.

The choice of H_n is essentially made on the basis of

i) The given supplementary conditions, so H_n is to contain ν τ-parameters to be determined to ensure satisfaction of conditions (1.13);

ii) Intrinsic properties of D, namely its range R_D, which is not, in general, the whole space $\mathbb{P}$ of algebraic polynomials, and its height h, $h = \sup_{n \in \mathbb{N}_0} \{\partial(Dx^n) - n\}$, where $\mathbb{N}_0$ is the set of nonnegative integers, and ∂ stands for "degree of", so H_n is to have degree $\leq n + h$ and to be in R_D. To be more precise, there generally exists for D an s-dimensional residual subspace R_S complementary to R_D,

$$\mathbb{P} = R_D \oplus R_S \quad , \quad R_D \cap R_S = \{0\} \quad , \tag{1.14}$$

$$R_S = \text{span}\,\{x^k : k \in S\} \quad , \quad S = \{k \in \mathbb{N}_0 : x^k \notin R_D\} \quad ,$$

and so H_n is also to contain s τ-parameters to be determined to ensure compatibility of the τ-problem, i.e., that no component of H_n lies in R_S [22];

iii) Approximation properties that y_n is required to possess. Clearly, the quality of y_n as an approximation of y depends on H_n, as follows from the fact that the τ-error function $\varepsilon_n = y - y_n$ is such that

$$D\varepsilon_n(x) = -H_n(x) \quad , \quad a \leq x \leq b \tag{1.15}$$

$$g_j(\varepsilon_n) = 0 \quad , \quad j = 1(1)\nu \quad ,$$

hence

$$\varepsilon_n(x) = -\int_a^b G(x, t)\, H_n(t)\, dt \quad , \tag{1.16}$$

where $G(x, t)$ is the corresponding Green's function, so H_n should be small, e.g., in the sense of the uniform norm, $\| H_n \| = \max_{a \leq x \leq b} | H_n(x) |$.

In the hope that the smallness of H_n will imply that of ε_n, one usually chooses

$$H_n(x) = \sum_{i=0}^{r} \tau_{m-i}^{(n)}\, v_{m-i}(x) \quad , \quad r = \nu + s - 1 \quad , \quad m = n + h \quad ,$$

where the $\tau_j^{(n)}$'s are the τ-parameters to be determined and the v_j's are Chebyshev or Legendre polynomials according as $y_n(x)$ is required to be a good (nearly uniform) global or endpoint approximation of $y(n)$, respectively on $a \leq x \leq b$ or at $x = b$ (see [10] and [16,17] for details and applications).

The existence, uniqueness, and convergence questions for the Lanczos' τ-approximation problem are reduced to the corresponding questions for the τ-parameters. These are uniquely determined by the supplementary and compatibility conditions referred to above and tend exponentially to zero as $n \to \infty$ [2,3].

As for the questions of existence, uniqueness, and characterization of perturbations leading to τ-approximants endowed with prescribed properties, they are still open, as far as we are aware, and we conclude that, in general, the choice of a suitable perturbation is not a simple matter.

1.3. AN ALTERNATIVE PRINCIPLE FOR τ-METHOD APPROXIMATION

As an alternative to the Lanczos' original perturbation idea, the following approximation principle [23,24] has evolved.

Choose a basis $v = \{v_k\}_{k=0}^{n}$ for $\mathbb{P}_n = \{P \in \mathbb{P} : \partial(P) \leq n\}$,

preferably orthogonal for rapid convergence, express y_n in it,

$$y_n = \sum_{k=0}^{n} \alpha_k v_k \quad ,$$

and determine the α_k's by making y_n satisfy the supplementary conditions in (1.13) and Dy_n agree with Dy as far as possible or desired.

This alternative principle, which emerges from [6] and [18], is shown in [23] to lead naturally to an approximation of the solution y of the problem in (1.9) - (1.11) in the sense of the τ - method, to be much more amenable to computer programming than Lanczos' original idea, and to be applicable to any kind of equation involving a linear (algebraic, differential, or integral) operator mapping $\mathbb{P}$ into itself, such as D in (1.10) or its integrated forms. There are, however, important differences to be considered between this approximation principle and Lanczos' original idea. For instance, in the Lanczos' τ - method the perturbation H_n is chosen in advance, whereas here it is not.

1.4. NUMERICAL SOLUTION OF THE τ - APPROXIMATION PROBLEM

There are essentially two approaches to the numerical solution of the τ - problem (1.12) - (1.13), one in terms of the matrix operator representation of D acting on v [23], to be described next for completeness and the other in terms of the sequence $Q = \{Q_k(x)\}_{k \in \mathbb{N}_0}$ of canonical polynomials associated with D and v, which are given by the functional equation

$$DQ_k(x) = v_k(x) \quad , \quad k = 0, 1, \ldots \quad , \tag{1.17}$$

(cf. [8,9]) obviously inconsistent for $k \in S$, or by the Ortiz' [14,15] redefining equation

$$\begin{aligned} DQ_k(x) &= v_k(x) + r_k(x) \quad , \quad r_k \in R_S \quad , \quad k \notin S \quad , \\ r_k(x) &= -v_k(x) \qquad\qquad\quad , \quad k \in S \quad , \end{aligned} \tag{1.18}$$

based on the fact in (1.14) that every element of $\mathbb{P}$ is uniquely decomposed into the sum of an element in R_D with another in R_S.

Canonical polynomials are, in fact, equivalence classes modulo $K_D = \{P \in \mathbb{P} : DP = 0\}$, the set of exact polynomial solutions of the given

equation, but this is a technical point, the details of which are to be found in [14].

Needless to say, the above definitions (1.17) - (1.18) extend immediately to any linear operator L mapping $\mathbb{P}$ into itself.

2. POLYNOMIAL τ - METHOD APPROXIMATION IN MATRIX OPERATIONAL TERMS

To show that the polynomial y_n in (1.12) - (1.13) is a τ - approximant of the solution y of the problem in (1.9) - (1.11), we review and extend some basic definitions and notation relative to the principle of using matrix operations in the Lanczos' τ - method, which has been developed in [13], [21], and [19,20].

By furnishing zero components, if need be, all vectors in the sequel are infinite - dimensional. Columnvectors are underlined once and row - vectors twice.

Let $\underline{v}$ and $D\underline{v}$ stand for the vectors with components v_k and Dv_k respectively, $k \geq 0$, then

$$D\underline{v} = \Pi_v \underline{v} \quad ,$$

Π_v being the matrix operator representation of D acting on $\mathbb{P}$ when we take for $\mathbb{P}$ the basis v. Π_v may be obtained directly whenever the laws of differentiation and multiplication in v are simple enough, otherwise we may work in the basis $\{x^k\}_{k=0,1\ldots}$ to get

$$D\underline{x} = \Pi_x \underline{x} \quad , \qquad \Pi_x = \sum_{r=0}^{\nu} \eta^r p_r(\mu) \quad ,$$

$$\eta = [\underline{e}_1, 2\underline{e}_2, 3\underline{e}_3, \ldots] \quad , \qquad \mu = [\underline{0}, \underline{e}_0, \underline{e}_1, \ldots] \quad ,$$

$\underline{e}_k$ being the vector with 1 in the kth position and 0 elsewhere, and switch to the basis v to get $\Pi_v = V \Pi_x V^{-1}$, V being such that $\underline{v} = V\underline{x}$. We refer to [18] and [21] for computational details and for structural properties of Π_x and Π_v. Π_x is a band matrix operator and its band width is $\leq \nu + h + 1$. Π_v is no longer banded from below, but is still banded from above.

If we let $y = \underline{\alpha}\,\underline{v}$ be the formal v - series expansion of the solution of the problem in (1.9) - (1.11), express $f(x)$ in the basis v, $f(x) = \underline{\underline{F}}\,\underline{v}$, introduce the vector $\underline{\sigma} = (\sigma_1, \ldots, \sigma_\nu, 0, 0, \ldots)$ and the matrix

$$B_\nu = (b_{ij}) \ , \quad b_{ij} = \begin{cases} g_j(v_i) \ , & j = 1(1)\nu \ ; \ i = 0, 1, \ldots \\ 0 \ , & j > \nu \ , \end{cases}$$

to express the supplementary conditions in (1.11) in the form

$$\underline{\underline{\alpha}} \, B_\nu = \underline{\underline{\sigma}} \ ,$$

and define the matrix

$$\Gamma_\nu = B_\nu + \Pi_\nu \, \mu^\nu$$

and the vector

$$\underline{\underline{\beta}} = \underline{\underline{\sigma}} + \underline{\underline{F}} \, \mu^\nu \ .$$

(postmultiplication of Π_ν and $\underline{\underline{F}}$ by μ^ν shifts their column entries ν places to the right), then $\underline{\underline{\alpha}}$ satisfies the infinite system of linear algebraic equations

$$\underline{\underline{\alpha}} \, \Gamma_\nu = \underline{\underline{\beta}} \ ,$$

whose truncation to its first $n+1$ equations, $n \geq N + \nu$,

$$\underline{\underline{\alpha}}^{(n)} \, \Gamma_\nu = \underline{\underline{\beta}} \ , \tag{2.1}$$

leads to the coefficient vector $\underline{\underline{\alpha}}^{(n)} = (\alpha_0^{(n)}, \ldots, \alpha_n^{(n)}, 0, 0, \ldots)$ of $y_n = \underline{\underline{\alpha}}^{(n)} \, \underline{v}$, which is a polynomial approximation of y in the sense of the τ-method. Indeed, with $\underline{\underline{\alpha}}^{(n)}$ given by

$$\underline{\underline{\alpha}}^{(n)} \, B_\nu = \underline{\underline{\beta}} \ , \quad \underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_j = F_j \ , \quad j = 0(1) \ n - \nu \ ,$$

the first ν equations representing the supplementary conditions in (1.13), then

$$\begin{aligned} Dy_n(x) &\equiv \underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{v} \\ &= \sum_{j=0}^{n-\nu} (\underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_j) \, v_j + \sum_{j=n-\nu+1}^{n+h} (\underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_j) \, v_j \\ &= f(x) + \sum_{j=1}^{\nu+h} (\underline{\underline{\alpha}}^{(n)} \, \Pi_\nu \, \underline{e}_{n-\nu+i}) \, v_{n-\nu+i} \end{aligned}$$

agrees with $Dy(x)$ as far as possible, i.e., up to $v_{n-\nu}$, and we have

solved the τ - problem in (1.12) - (1.13) with the perturbation

$$(2.2)\quad H_n = \sum_{i=1}^{\nu+h} \tau_i^{(n)} v_{n-\nu+i} \ , \quad \tau_i^{(n)} = \underline{\alpha}^{(n)} \Pi_\nu \underline{e}_{n-\nu+i} \ , \quad i = 1(1)\, \nu+h \ .$$

For rapid convergence, the components of $\underline{\alpha}^{(n)}$ should tend to zero fast. To achieve this, a convenient orthogonal basis has to be chosen. Lanczos would have taken a perturbation like that in (2.2), with $v_k = T_k^*(x)$, $k = 0, 1, \ldots$, the Chebyshev polynomials shifted to $a \le x \le b$, to get a good global polynomial approximation y_n of y . There are, however, important differences to be considered between the above approximation principle and Lanczos' original perturbation idea. For instance, in the Lanczos' τ - method, the perturbation is chosen in advance, whereas here it is not. On the other hand, the evaluation of the Lanczos' τ - parameters is not generally amenable to computer programming, whereas (2.2) gives them immediately, the moment $\underline{\alpha}^{(n)}$ is obtained.

While error analysis for the general polynomial τ - method approximation is undoubtedly difficult, an upper bound for $\| \varepsilon_n \|$ may be easily obtained from (1.16). Also, from (1.15), an efficient estimation of ε_n may be obtained as follows (see [23] and references given there for details, applications, and numerical examples).

Let $\varepsilon_{n,m}$, $m > n$, be an m th order polynomial approximation of ε_n , then $\varepsilon_{n,m}$ satisfies the perturbed ODE

$$D\, \varepsilon_{n,m}(x) = - H_n(x) + H_m(x) \ , \qquad a \le x \le b \ ,$$

is such that

$$\varepsilon_{n,m}(x) = y_m(x) - y_n(x) \ ,$$

and thus, every time two τ - approximants $y_n(x)$ and $y_m(x)$ are computed, an estimation of $\varepsilon_n(x)$ is obtained.

Clearly, all we have said about the differential operator D extends immediately to any linear operator L mapping $\mathbb{P}$ into itself. As pointed out in [6], the only embarrassment of the method is the accidental singularity of the system (2.1). If this happens, we merely change n to $n+1$ and keep doing this until we find a nonsingular system. This must ultimately exist if the solution of the given problem has a convergent ν - series expansion.

3. RECURSIVE CONSTRUCTION OF POLYNOMIAL τ - APPROXIMANTS IN TERMS OF CANONICAL POLYNOMIALS

The canonical and residual polynomials associated with a given linear operator $L : \mathbb{P} \to \mathbb{P}$ and a given basis v may be computed recursively [15].

Writing

$$L v_n = \sum_{j=0}^{m} \Pi_{nj} v_j \quad , \quad m = n + h \quad ,$$

$$= \Pi_{nm} v_m + \sum_{j=0}^{m-1} \Pi_{nj} (L Q_j - r_j) \quad ,$$

$$L \left(v_n - \sum_{j=0}^{m-1} \Pi_{nj} Q_j\right) = \Pi_{nm} v_m - \sum_{j=0}^{m-1} \Pi_{nj} r_j \quad ,$$

assuming that $\Pi_{nm} \neq 0$ and that Q_j and r_j , $j = 0(1)\ m-1$, have already been computed, then

$$Q_m = \left(v_n - \sum_{j=0}^{m-1} \Pi_{nj} Q_j\right) / \Pi_{nm}$$

$$r_m = -\left(\sum_{j=0}^{m-1} \Pi_{nj} r_j\right) / \Pi_{nm} \qquad n = 0,\ 1, \ldots$$

Once the first $M = \max \{m , N\}$ canonical and residual polynomials are known, the τ - solution y_n of the equation

$$Ly = f \quad , \qquad f = \sum_{k=0}^{N} F_k v_k \quad ,$$

which is the exact polynomial solution of the perturbed equation

$$Ly_n = f + H_n \quad , \quad H_n = \sum_{k=0}^{m} h_k v_k \quad ,$$

$$= Z_M \quad , \quad Z_M = \sum_{k=0}^{M} z_k v_k \quad , \quad z_k = F_k + h_k \quad ,$$

is immediately at hand,

$$y_n = \sum_{k=0}^{M} z_k Q_k \quad ,$$

as well as the corresponding compatibility conditions,

$$\sum_{k=0}^{M} z_k \, r_k \equiv 0 \quad ,$$

which are easily seen to be equivalent to the formal use of the undefined canonical polynomials (see (1.17)) and the subsequent cancellation of their coefficients.

In addition to satisfying a self-starting recurrence relation and being an efficient basis for the representation of τ - solutions, canonical polynomials have a number of other useful properties, namely, they are

i) Permanent, and so, if we need y_{n+1} after y_n has been constructed, namely to improve the approximation accuracy, only Q_{m+1} has to be computed ;

ii) Independent of the given supplementary conditions, hence initial and boundary value problems are treated alike ;

iii) Independent of the approximation interval, and so piecewise polynomial and rational τ - approximants are easily constructed (see, e.g., [1], [2], [4,5], [11,12], and [16,17], where they choose a convenient perturbation in advance, and [24], where we just accept the perturbation the given problem leads to).

4. THE LINK BETWEEN THE TWO FOREGOING APPROACHES TO THE NUMERICAL SOLUTION OF THE τ - APPROXIMATION PROBLEM

Let Π_v be the matrix operator representation of a given linear operator $L : \mathbb{P} \to \mathbb{P}$ when we take for $\mathbb{P}$ the basis v, i.e.,

$$L\underline{v} = \Pi_v \underline{v} \quad ,$$

let $Q = \{Q_k\}_{k \in \mathbb{N}_0 - S}$ and $r = \{r_k\}_{k \in \mathbb{N}_0 - S}$ be the sequences of canonical and residual polynomials associated with L and v, and assume, with no loss of generality, that $S = \{0, 1, \ldots, s-1\}$. Following [20], we define the vectors $\underline{r} = (r_s, r_{s+1}, \ldots)$ and $\underline{Q} = (Q_s, Q_{s+1}, \ldots)$ and the matrices $[R;0]$ and C such that

$$\underline{r} = [R;0] \, \underline{v} \quad , \quad \underline{Q} = C \, \underline{v} \quad ,$$

then Ortiz' funcional equations (1.18) may be written as

$$L\underline{Q} = [R;I]\ \underline{v} \quad ,$$

but $L\underline{Q} = C\,\Pi_v\,\underline{v}$, so $C\,\Pi_v = [R;I]$, and if we set $\Pi_v = [\Pi_S;\Pi_Q]$, then

$$C\,\Pi_Q = I \quad , \quad C\,\Pi_S = R \ .$$

The calculation of the canonical polynomials Q_k , $k \notin S$, is, therefore, equivalent to the inversion of the matrix Π_v stripped of its columns of order $k \in S$.

5. EXAMPLES OF APPROXIMATE EXPANSIONS IN SERIES OF ORTHOGONAL POLYNOMIALS

1) To construct approximate expansions of the function

$$y(x) = ((1-x)/2)^{1/2} \quad , \qquad -1 \le x \le 1 \quad ,$$

in the Legendre basis $v = \{P_n(x)\}_{n=0,1,\ldots}$,

$$y(x) = \underline{\underline{\alpha}}\,\underline{v} = \frac{2}{3}P_0(x) - 2\sum_{n=1}^{\infty}\frac{P_n(x)}{(2n-1)(2n+3)} \quad , \tag{5.1}$$

we choose a definition of $y(x)$ in terms of a linear operator $L: \mathbb{P} \to \mathbb{P}$, e.g.,

$$L\,y(x) \equiv (1-x)\,y(x) + \frac{3}{2}\int_{-1}^{x} y(t)\,dt = 2 \quad ,$$

integrated form of the IVP

$$D\,y(x) \equiv 2(1-x)\,y'(x) + y(x) = 0 \quad , \quad y(-1) = 1 \quad ,$$

and construct the matrix operator Π_v such that $L\underline{v} = \Pi_v\,\underline{v}$, i.e.,

$$L\,P_0 = \frac{5}{2}P_0 + \frac{1}{2}P_1$$

$$L\,P_n = -\frac{2n+3}{2(2n+1)}P_{n-1} + P_n - \frac{2n-1}{2(2n+1)}P_{n+1} \quad , \quad n = 1, 2, \ldots$$

Thanks to the structure of Π_v , the polynomial approximation $y_n = \underline{\underline{\alpha}}^{(n)}\,\underline{v}$ of y is such that

$$L\,y_n = 2 + \alpha_n^{(n)}\,P_{n+1} \quad , \quad n = 0, 1, \ldots \tag{5.2}$$

and its coefficient vector $\underline{\alpha}^{(n)}$ may be obtained either by solving the following system of linear algebraic equations

$$\underline{\alpha}^{(n)} \Pi_v \underline{e}_0 = 2 , \quad \underline{\alpha}^{(n)} \Pi_v \underline{e}_j = 0 , \quad j = 1(1) n ,$$

or by using the canonical and residual polynomials associated with L and v ,

$$Q_0 = 0 , \quad r_0 = -1 ; \quad Q_1 = 2 P_0 , \quad r_1 = 5 ;$$

$$Q_{n+1} = -\frac{1}{2n-1} \left[2(2n+1)(P_n - Q_n) + (2n+3) Q_{n-1}\right] ,$$

$$r_{n+1} = \frac{1}{2n-1} \left[2(2n+1) r_n - (2n+3) r_{n-1}\right] , \quad n = 1, 2, \ldots$$

From (5.2) and (1.18) we obtain

$$y_n = \alpha_n^{(n)} Q_{n+1} , \quad \alpha_n^{(n)} = \frac{2}{r_{n+1}} , \quad n = 0, 1, \ldots ,$$

successive approximations of the corresponding partial sums of the Legendre series in (5.1). In particular, the coefficients of y_4 ,

$$y_4 = \frac{1}{45045} (7525 P_0 - 4452 P_1 - 985 P_2 - 378 P_3 - 135 P_4) ,$$

approximate the corresponding coefficients of y with absolute error $\leq 1.4 \times 10^{-2}$.

2) To construct approximate expansions of

$$y(x) = e^{-x} , \quad 0 \leq x \leq 1 ,$$

in the Hermite basis $v = \{H_n^*(x)\}_{n=0,1\ldots}$,

$$y(x) = \underline{\alpha}\, \underline{v} = e^{1/4} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n n!} H_n^*(x) ,$$

let us define $y(x)$ in terms of the following linear operator

$$L y(x) \equiv y(x) + \int_0^x y(t)\, dt = 1 .$$

Recalling that

$$\int_0^x H_n^*(t)\,dt = \frac{1}{2(n+1)}\left[H_{n+1}^*(x) - H_{n+1}^*(0)\right] ,$$

$$H_{2n}^*(0) = (-1)^n \frac{(2n)!}{n!} , \quad H_{2n+1}^*(0) = 0 , \quad n = 0, 1,\ldots ,$$

we get

$$L H_n^*(x) = \frac{1}{2(n+1)}\left[H_{n+1}^*(x) + 2(n+1)\, H_n^*(x) - H_{n+1}^*(0)\right] ,$$

hence

$$Q_0 = 0 , \quad r_0 = -1$$

$$Q_n = 2n(H_{n-1}^* - Q_{n-1}) , \quad r_n = -2n\, r_{n-1} - H_n^*(0) , \quad n = 1, 2,\ldots$$

The polynomial approximation $y_n = \underline{\alpha}^{(n)} \underline{v}$ of $y = \underline{\alpha}\, \underline{v}$ satisfies the perturbed equation

$$L y_n = 1 + \alpha_n^{(n)} H_{n+1}^* , \quad n = 0, 1,\ldots ,$$

and is such that

$$y_n = \frac{Q_{n+1}}{r_{n+1}} , \quad n = 0, 1,\ldots .$$

In particular, the coefficients $\alpha_k^{(n)}$, $n = 0(1)\,5$, $k = 0(1)\,n$, of the first six y_n's ,

$$y_0 = H_0$$

$$y_1 = \frac{2}{3}(2H_0^* - H_1^*)$$

$$y_2 = \frac{1}{6}(8H_0^* - 4H_1^* + H_2^*)$$

$$y_3 = \frac{2}{75}(48H_0^* - 24H_1^* + 6H_2^* - H_3^*)$$

$$y_4 = \frac{1}{300}(384H_0^* - 192H_1^* + 48H_2^* - 8H_3^* + H_4^*)$$

$$y_5 = \frac{1}{2990}(3840H_0^* - 1920H_1^* + 480H_2^* - 80H_3^* + 10H_4^* - H_5^*) ,$$

approximate the corresponding coefficients α_k of y with the errors given below

n \ k	$\alpha_k - \alpha_k^{(n)}$					
0	2.84×10^{-1}					
1	-4.93×10^{-2}	2.47×10^{-2}				
2	-4.93×10^{-2}	2.47×10^{-2}	-6.16×10^{-3}			
3	4.03×10^{-3}	-2.01×10^{-3}	5.03×10^{-4}	-8.39×10^{-5}		
4	4.03×10^{-3}	-2.01×10^{-3}	5.03×10^{-4}	-8.39×10^{-5}	1.05×10^{-5}	
5	-2.56×10^{-4}	1.28×10^{-4}	-3.19×10^{-5}	5.32×10^{-6}	-6.67×10^{-7}	6.60×10^{-8}

REFERENCES

1. J.P. COLEMAN: The Lanczos tau-method. J. Inst. Maths. Applics 17 (1976), 85-97.

2. M.R. CRISCI: The tau method with perturbation term depending on the differential operator. J. Comp. Appl. Maths. 15 (1986), 123-136.

3. M.R. CRISCI and E.L. ORTIZ: Existence and convergence results for the numerical solution of differential equations with the tau method. Imperial College NAS Res. Rep., University of London, London 1981.

4. M.R. CRISCI and E. RUSSO: A stability of a class of methods for the numerical integration of certain linear systems of ordinary differential equations. Math. Comp. 38 (1982), 431-435.

5. M.R. CRISCI and E. RUSSO: An extension of Ortiz' recursive formulation of the tau method to certain linear systems of ordinary differential equations. Math. Comp. 41 (1983), 27-42.

6. L. FOX and I.B. PARKER: Chebyshev polynomials in numerical analysis. Oxford Univ. Press, London 1968.

7. C. LANCZOS: Trigonometric interpolation of empirical and analytical functions. J. Math. Phys. 17 (1938), 123-199.

8. C. LANCZOS: Tables of Chebyshev polynomials $S_n(x)$ and $C_n(x)$, Introduction. Nat. Bur. Standards Appl. Math. Ser. 9, U.S. Govt. Printing Office, Washington 1952.

9. C. LANCZOS: Applied analysis. Pitman, London 1957.

10. C. LANCZOS: Legendre versus Chebyshev polynomials. In: Topics in numerical analysis (J. J. H. Miller, ed.), Academic Press, London 1973, pp. 191-201.

11. Y.L. LUKE: The special functions and their approximations. Academic Press, New York 1969.

12. Y.L. LUKE: Mathematical functions and their approximations. Academic Press, London 1975.

13. P. ONUMANYI, E.L. ORTIZ, and H. SAMARA: Software for a method of finite approximations for the numerical solution of differential equations. Appl. Math. Modelling 5 (1981), 282 - 286.

14. E.L. ORTIZ: The tau method. SIAM J. Numer. Anal. 6 (1969), 480 - 492.

15. E.L. ORTIZ: Canonical polynomials in the Lanczos tau method. In: Studies in numerical analysis (B.K.P. Scaife, ed.), Academic Press, London 1974, pp. 73 - 93.

16. E.L. ORTIZ: Step by step tau method - part I. Piecewise polynomials approximations. Comp. & Maths. with Appls. 1 (1975), 381 - 392.

17. E.L. ORTIZ: Sur quelques nouvelles applications de la methode tau. In: Séminaires IRIA analyse et contrôle de systèmes, IRIA, Paris 1975, pp. 247 - 257.

18. E.L. ORTIZ and H. SAMARA: An operational approach to the tau method for the numerical solution of nonlinear differential equations. Computing 27 (1981), 15 - 25.

19. E.L. ORTIZ and H. SAMARA: Matrix displacement mappings in the numerical solution of functional and nonlinear differential equations with the tau method. Num. Funct. Anal. and Optimiz. 6 (1983), 379 - 398.

20. E.L. ORTIZ and H. SAMARA: Numerical solution of differential eigenvalue problems with an operational approach to the tau method. Computing 31 (1983), 95 - 103.

21. M.R. da SILVA: LACALGEBRA versions of Lanczos' tau method for the numerical solution of differential equations. Port. Math. 41 (1982), 295 - 316.

22. M.R. da SILVA: A quick survey of recent developments and applications of the τ - method. In: Numerical approximation of partial differential equations (E.L. Ortiz, ed.), North-Holland, Amsterdam 1987, pp. 297-308.

23. M.R. da SILVA: Numerical treatment of differential equations with the τ - method. J. Comp. Appl. Math. 20 (1987), 1 - 7.

24. M.R. da SILVA and M.J. RODRIGUES: A simple alternative principle for rational τ - method approximation. To appear in the proceedings of the conference on Nonlinear Numerical Methods and Rational Approximation, Antwerp 1987.

ON MONOTONICITY OF SOME LINEAR POSITIVE OPERATORS

B. DELLA VECCHIA

ABSTRACT: In this paper we study the monotonicity of the sequences of some linear positive operators, to which we apply the iterated Nörlund operator. As particular cases, we find the results established by D.D. Stancu for the sequence $\{(B_n f)^{(m)}(x)\}_n$ and by the author for the sequences $\{(M_n f)^{(m)}(x)\}_n$ and $\{(P_n f)^{(m)}(x)\}_n$ where M_n and P_n are the Favard-Szasz-Mirakyan and Baskakov operator respectively.

1. INTRODUCTION

It is well known that the Bernstein Polynomials corresponding to functions convex in $[0,1]$ verify the following monotonicity relationship [1,18,28]:

$$(1)\quad (B_{n+1}f)(x) = B_{n+1}f(x) \leqq B_n f(x), \quad 0 \leqq x \leqq 1$$

This property has been extended to other linear positive operators [3,8,9,10,11,12,19].

Later Stancu in [24] studied the derivatives of the sequence of Bernstein polynomials and obtained interesting monotonicity properties for this sequence.

Then Horová in [7] established a relation of type (1) for the first derivatives of Favard-Szasz-Mirakyan operator [5,15,16].

Recently in [4] we have proved monotonicity properties for the derivatives of order s, $s > 1$, of the sequence of Favard-Szasz-Mirakyan operator and for the derivatives of order s, $s \geqq 1$, of the sequence of Baskakov operator [2,6,27].

On the other hand some authors introduced separately discrete type operators generalizing Bernstein, Favard-Szasz-Mirakyan and Baskakov operators.

The main purpose of this paper is to extend the procedure given in [24] to the sequence of these operators, to which we apply the iterated Nörlund difference operator, instead of the differentiation operator.

As a particular case, we find the results established in [4].

2. PRELIMINARY RESULTS

Let f be a function defined on an interval I of the real axis. As usual, we denote by $[t_0,t_1,\dots,t_n;f]$ the divided difference of order n, of the function f, with respect to the distinct nodes $t_0,t_1,\dots,t_n \in I$.

We recall also that f is called convex, non-concave, polynomial, non-convex respectively concave of n-order on an interval $I=[a,b]$, if all its divided differences of order n+1, on n+2 distinct nodes from I, are >0, $\geqq 0$, $=0$, $\leqq 0$, resp. <0.

We use in the sequel also the formula

$$(2.1)\quad D_\alpha^m[f(x)g(x)] = \sum_{i=0}^{m} \binom{m}{i} D_\alpha^i f(x+(m-i)\alpha) D_\alpha^{m-i} g(x)$$

with $m \in N$, f and g defined on I, $x \in I$, $\alpha \in R^+$ and

$$D_\alpha g(x) = [g(x+\alpha)-g(x)]\alpha^{-1}, \quad D_\alpha^m = D_\alpha(D_\alpha^{m-1}), \quad D_\alpha^0 g(x) = g(x).$$

Then let r and n be two integers, with $0 \leqq r \leqq n$, and consider the following points of the interval I: $a_i=a+ih$, $i=0,1,\dots,n$ and $b_j=a+j\ell$, $j=1,2,\dots,n$, where $0<h\leqq\frac{b-a}{n}$, $0<\ell<\frac{b-a}{n}$.

Now we denote by $T_k^{(\nu)}$, $0\leqq k\leqq n$, $1<\nu\leqq r+1$, the linear functionals defined recursively as follows

$$T_k^{(2)} f = [a_k, a_{k+1}, b_{k+1}; f], \quad 0\leqq k\leqq n-1$$

$$(2.2)\quad T_k^{(\nu+1)} f = T_{k+1}^{(\nu)} f - T_k^{(\nu)} f, \quad 1<\nu\leqq r, \quad 0\leqq k\leqq n-r$$

where f is a function defined on [a,b].

This functional has been introduced in [24] by Stancu, who proved there that

$$\text{2.3)}\quad \text{f convex (non-concave) of order } r+1 \Longrightarrow T_k^{(r+2)} f > 0 \quad (T_k^{(r+2)} f \geqq 0), \quad 0 \leqq k \leqq n-r$$

We consider now the class of linear and positive operators V_n^α defined by

$$\text{2.4)}\quad V_n^\alpha f(x) = \sum_{k=0}^{\infty} (-1)^k D_\alpha^k \phi_n^\alpha(x) \frac{x^{(k,-\alpha)}}{k!} f\left(\frac{k}{n}\right)$$

where: α is a non-negative parameter that can depend only on $n \in N$; $x^{(k,-\alpha)} = x(x+\alpha)\ldots(x+(k-1)\alpha)$; the functions ϕ_n^α $(n \in N)$ are defined on R and verify the following conditions:

i) $\phi_n^\alpha(0) = 1$;

ii) $\forall k \in N$ and $\forall x \in R\ (-1)^k D_\alpha^k \phi_n^\alpha(x) \geqq 0$;

iii) $\sum_{k=0}^{\infty} (-1)^k D_\alpha^k \phi_n^\alpha(x) \frac{x^{(k,-\alpha)}}{k!} = 1$;

$f \in C^*$, where C^* denotes the set of functions defined on $[0,+\infty[$ and such that (2.4) has meaning.

This operator has been introduced and studied in [11,12,23].

Setting

$$\gamma_{n,i}^\alpha = \frac{(-1)^i D_{-\alpha}^i \phi_n^\alpha(0)}{n^i}, \quad \gamma_{n,0}^\alpha = 1$$

it is known [11] that

Theorem 2.1. If

$$\lim_n \gamma_{n,r+i}^\alpha = 1, \quad i=0,1,2, \quad r \in N$$

with $0<\alpha=\alpha(n)\to 0$, when $n\to\infty$, then we have

$$\lim_n \left\| f^{(r)} - D_\alpha^r V_n^\alpha f \right\| = 0,$$

$\forall f^{(r)} \in \bar{C}^\circ$, where $\bar{C}^\circ$ denotes the set of functions defined on $[0,+\infty[$, there bounded and uniformly continuous, and $\forall r \in N$.

We notice that, by choosing suitably ϕ_n^α functions, V_n^α beco-

mes well-known linear positive operators, studied separately by some authors in [13,14,17,19,20-22,25].

Here we want to consider the following three particular cases.

1) If

$$\phi_n^{\alpha}(x) = \frac{(1-x)^{(n,-\alpha)}}{1^{(n,-\alpha)}}, \qquad 0 \leqq x \leqq 1$$

V_n^{α} coincides with the Stancu operator S_n^{α} defined by

$$(2.5)\quad S_n^{\alpha} f(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{(1-x)^{(n-k,-\alpha)}}{1^{(n,-\alpha)}} x^{(k,-\alpha)} f(\frac{k}{n}), \qquad \begin{matrix} 0 \leqq x \leqq 1 \\ f \in C([0,1]). \end{matrix}$$

Stancu introduced in [19] this operator, which later has been studied in [11,13,14,20,22,23].

We notice that, for $\alpha=0$, S_n^{α} becomes equal to Bernstein operator B_n.

2) If

$$\phi_n^{\alpha}(x) = (1+n\alpha)^{-x/\alpha}, \qquad x \geqq 0, \qquad 0 \leqq n\alpha \leqq 1$$

V_n^{α} coincides with the M_n^{α} operator defined by

$$(2.6)\quad M_n^{\alpha} f(x) = (1+n\alpha)^{-x/\alpha} \sum_{k=0}^{\infty} (\alpha+\frac{1}{n})^{-k} \frac{x^{(k,-\alpha)}}{k!} f(\frac{k}{n})$$

$f \in E_{\beta,B}$, where $E_{\beta,B}$ denotes the set of functions defined on $[0,+\infty[$, continuous in $[0,B]$, $(B>0)$ and such that $f(x)=O(2^{\beta x})$ $(x\to\infty)$, with β a positive fixed number.

This operator was proposed by different approaches in [12,17,23,25]. We notice that, for $\alpha=0$, (2.6) becomes

$$(2.7)\quad M_n f(x) = e^{-nx} \sum_{k=0}^{\infty} \frac{(nx)^k}{k!} f(\frac{k}{n})$$

where M_n is Favard-Szasz-Mirakyan operator [3,5,7,9,15,19,26

3) If

$$\phi_n^{\alpha}(x) = (1+x)^{(-n,-\alpha)}\, 1^{(n,-\alpha)} = \frac{1^{(n,-\alpha)}}{(1+x)^{(n,-\alpha)}} \qquad \begin{matrix} x \geqq 0 \\ 0 \leqq \alpha < 1/2 \end{matrix}$$

V_n^{α} coincides with Baskakov-Stancu operator P_n^{α} defined by

$$(2.8)\quad P_n^{\alpha} f(x) = 1^{(n,-\alpha)} \sum_{k=0}^{\infty} \binom{n+k-1}{k} \frac{x^{(k,-\alpha)}}{(1+x)^{(n+k,-\alpha)}} f(\frac{k}{n}), \quad f \in C^*$$

This operator was introduced by Stancu in [21,23] and later studied by Mastroianni in [11]. As a particular case, for $\alpha=0$, P_n^{α} becomes equal to Baskakov operator [2,6,9,27]

$$2.9)\quad P_n f(x) = \sum_{k=0}^{\infty} \binom{n+k-1}{k} \frac{x^k}{(1+x)^{n+k}} f\left(\frac{k}{n}\right)$$

. ON THE MONOTONICITY OF THE SEQUENCE $\{D_{\alpha}^{m} S_n^{\alpha} f(x)\}_n$

Let $S_n^{\alpha} f$ be the Stancu operator defined by (2.5). It is ell known [19] that two consecutive terms of the sequence $S_n^{\alpha} f(x)\}_n$ verify the following relationship:

$$(3.1)\quad \begin{aligned} (S_{n+1}^{\alpha} - S_n^{\alpha}) f(x) &= -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot \\ &\cdot \sum_{\nu=0}^{n-1} \frac{(-1)^{\nu} D_{\alpha}^{\nu} \phi_{n-1}^{\alpha}(x-\alpha)(x+\alpha)^{(\nu,-\alpha)}}{\nu!} \left[\frac{\nu}{n}, \frac{\nu+1}{n+1}, \frac{\nu+1}{n}; f\right] \end{aligned} \qquad \begin{matrix} f\in C([0,1]) \\ x\in[0,1] \end{matrix}$$

If we choose the following points

$$a_{\nu+i} = \frac{\nu+i}{n+i}, \qquad b_n = \frac{\nu}{n}, \qquad \nu=0,1,\ldots,n-1, \qquad i=0,1$$

using the functional $T_{\nu}^{(2)}$ defined by (2.2), (3.1) becomes

$$(3.2)\quad \begin{aligned} (S_{n+1}^{\alpha} - S_n^{\alpha}) f(x) &= -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot \\ &\cdot \sum_{\nu=0}^{n-1} (-1)^{\nu} D_{\alpha}^{\nu} \phi_{n-1}^{\alpha}(x-\alpha)(x+\alpha)^{(\nu,-\alpha)} T_{\nu}^{(2)} f \end{aligned}$$

Now we introduce, $\forall r\in N$ and $\forall f\in C([0,1])$, the linear positive operator $S_n^{\alpha,r}$

$$S_n^{\alpha,r} f(x) = \sum_{\nu=0}^{n-r} (-1)^{\nu+r-1} \frac{D_{\alpha}^{\nu+r-1} \phi_{n-1}^{\alpha}(x-\alpha)(x+r\alpha)^{(\nu,-\alpha)}}{\nu!} f\left(\frac{\nu}{n}\right)$$

From (3.2) it follows that

$$(3.3)\quad (S_{n+1}^{\alpha}-S_n^{\alpha})f(x) = -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} S_n^{\alpha,1} T_{\nu}^{(2)} f(x)$$

Now, in order to study the monotonicity of the sequence $\{D_{\alpha}^{m} S_n^{\alpha} f(x)\}_n$, we prove

Theorem 3.1. The following relationship holds:

$$D_{\alpha}^{m}(S_{n+1}^{\alpha}-S_n^{\alpha})f(x) = -\frac{1}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot$$

$$(3.4)\quad \cdot\Big[(x+m\alpha)(1-x-m\alpha)S_n^{\alpha,m+1} T_{\nu}^{m+2} f(x) +$$

$$+ m(1-2x-(2m-1)\alpha)S_n^{\alpha,m} T_{\nu}^{(m+1)} f(x) - m(m-1)S_n^{\alpha,m-1} T_{\nu}^{(m)} f(x)\Big]$$

with $m \leqq n$ and $\alpha \geqq 0$

Proof.

In fact, using (2.1) in (3.3), with

$$f(x) = -\frac{x(1-x)}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \text{ and } g(x) = S_n^{\alpha,1} T_{\nu}^{(2)} f(x),$$

we have

$$D_{\alpha}^{m}(S_{n+1}^{\alpha}-S_n^{\alpha})f(x) = -\frac{1}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)} \cdot$$

$$\cdot \Big\{(x+m\alpha)(1-x-m\alpha)D_{\alpha}^{m} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) +$$

$$+ mD_{\alpha}^{1}[(x+(m-1)\alpha)(1-x-(m-1)\alpha)]\, D_{\alpha}^{m-1} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) +$$

(3.5)

$$+\frac{m(m-1)}{2} D_{\alpha}^{2}[(x+(m-2)\alpha)(1-x-(m-2)\alpha)]\, D_{\alpha}^{m-2} S_n^{\alpha,1} T_{\nu}^{(2)} f(x)\Big\} =$$

$$= -\frac{1}{n(n+1)(1+(n-1)\alpha)(1+n\alpha)}\Big\{(x+m\alpha)(1-x-m\alpha)D_{\alpha}^{m} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) +$$

$$+m(1-2x-(2m-1)\alpha)D_{\alpha}^{m-1} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) - m(m-1)D_{\alpha}^{m-2} S_n^{\alpha,1} T_{\nu}^{(2)} f(x)\Big\}$$

On the other hand, one can easily verify by induction that

$$D_{\alpha}^{m} S_n^{\alpha,1} T_{\nu}^{(2)} f(x) = S_n^{\alpha,m+1} T_{\nu}^{(m+2)} f(x)$$

and making use of this last relation in (3.5), the theorem follows.

From Theorem 3.1, by (2.3), we obtain

Corollary 3.2. The sequence

$$(3.6)\quad \{D_{\alpha}^{m}S_{n}^{\alpha}f(x)\}_{n}, \qquad 0\leq m\alpha\leq 1, \qquad m\leq n$$

verifies the following monotonicity properties:

i) for m=1

a) If on the interval $\left[0,\frac{1-\alpha}{2}\right]$ the function f is convex (concave) of first and second order, then the sequence (3.6) is decreasing (increasing) on $\left[0,\frac{1-\alpha}{2}\right]$;

b) If on the interval $\left[\frac{1-\alpha}{2},1-\alpha\right]$ the function f is concave (convex) of first order and convex (concave) of second order, then the sequence (3.6) is decreasing (increasing) on the interval $\left[\frac{1-\alpha}{2},1-\alpha\right]$;

c) If on the interval $[1-\alpha,1]$ the function f is concave (convex) of first and second order, then the sequence (3.6) is decreasing (increasing) on $[1-\alpha,1]$.

ii) for m≥2

a) If on the interval $\left[0,\frac{1-(2m-1)\alpha}{2}\right]$ the function f is concave (convex) of order m-1 and convex (concave) of order m and m+1, then the sequence (3.6) is decreasing (increasing) on $\left[0,\frac{1-(2m-1)\alpha}{2}\right]$;

b) If on the interval $\left[\frac{1-(2m-1)\alpha}{2},1-m\alpha\right]$ the function f is concave (convex)of order m-1 and m and convex (concave) of order m+1, then the sequence (3.6) is decreasing (increasing) on $\left[\frac{1-(2m-1)\alpha}{2},1-m\alpha\right]$;

c) If on the interval $[1-m\alpha,1]$ the function f is concave (convex) of order m-1, m and m+1, then the sequence (3.6) is decreasing (increasing) on $[1-m\alpha,1]$.

We notice that, for α=0, Corollary 3.2 gives us a monotonicity result for the derivatives of the sequence of Bernstein polynomials, obtained by Stancu in [24].

4. ON THE MONOTONICITY OF THE SEQUENCE $\{D_\alpha^m M_n^\alpha f(x)\}_n$

Let $M_n^\alpha f$ be the operator introduced in (2.6).
It is known [12] that the following relationship holds for two consecutive terms of the sequence $\{M_n^\alpha f(x)\}_n$:

$$(4.1)\qquad (M_{n+1}^\alpha - M_n^\alpha) f(x) = -\frac{x}{n(n+1)} \cdot \sum_{k=0}^{\infty} (\alpha - \frac{1}{n})^{k+1} D_\alpha^{k+1} \phi_n^\alpha(x) \frac{(x+\alpha)^{(k,-\alpha)}}{k!} \left[\frac{k}{n}, \frac{k+1}{n+1}, \frac{k+1}{n}; f\right] \quad \begin{matrix} f \in E_{\beta,B} \\ x \geq 0 \end{matrix}$$

We introduce now, $\forall r \in N$ and $\forall f \in E_{\beta,B}$, the operator $M_n^{\alpha,r}$ defined as follows

$$M_n^{\alpha,r} f(x) = \frac{1}{n} \sum_{k=0}^{\infty} (-1)^{k+r} D_\alpha^{k+r} \phi_n^\alpha(x) \frac{(x+r\alpha)^{(k,-\alpha)}}{k!} f(\frac{k}{n})$$

One can easily verify that this operator is linear and positive.
Letting then

$$a_{k+i} = \frac{k+i}{n+i} \quad \text{and} \quad b_k = \frac{k}{n}, \qquad k = 0,1,\ldots, \qquad i = 0,1$$

and recalling the definition of the functional $T_k^{(2)}$ introduced in (2.2), (4.1) becomes

$$(4.2)\qquad (M_{n+1}^\alpha - M_n^\alpha) f(x) = -\frac{x}{n(n+1)} M_n^{\alpha,1} T_k^{(2)} f(x)$$

Now, in order to study the monotonicity of the sequence $\{D_\alpha^m M_n^\alpha f(x)\}_n$, we prove

Theorem 4.1. The following relationship holds

$$(4.3)\qquad D_\alpha^m [M_{n+1}^\alpha - M_n^\alpha] f(x) = -\frac{x+m\alpha}{n(n+1)} M_n^{\alpha,m+1} T_k^{(m+2)} f(x) + \\ - \frac{m}{n(n+1)} M_n^{\alpha,m} T_k^{(m+1)} f(x), \qquad m \in N \quad \text{and} \quad 0 \leq n\alpha \leq 1$$

Proof.
Indeed, by applying (2.1) to (4.2), with

$$f(x) = -\frac{x}{n(n+1)} \quad \text{and} \quad g(x) = M_n^{\alpha,1} T_k^{(2)} f(x),$$

we have

$$D_\alpha^m [M_{n+1}^\alpha - M_n^\alpha] f(x) = -\frac{1}{n(n+1)} \sum_{i=0}^{1} \binom{m}{i} D_\alpha^i [x+(m-i)\alpha] \cdot$$

$$(4.4) \quad \cdot D_\alpha^{m-i} M_n^{\alpha,1} T_k^{(2)} f(x) =$$

$$= -\frac{1}{n(n+1)} \left[(x+m\alpha) D_\alpha^m M_n^{\alpha,1} T_k^{(2)} f(x) + m D_\alpha^{m-1} M_n^{\alpha,1} T_k^{(2)} f(x) \right]$$

Moreover, one can easily prove by induction that

$$D_\alpha^m M_n^{\alpha,1} T_k^{(2)} f(x) = M_n^{\alpha,m+1} T_k^{(m+2)} f(x)$$

and, by using this last relationship in (4.4), the theorem follows.

From Theorem 4.1, by (2.3), we have

Corollary 4.2. For the sequence

$$(4.5) \quad \{D_\alpha^m M_n^\alpha f(x)\}_n, \qquad m \in N \quad 0 \leqq n\alpha \leqq 1$$

the following monotonicity property holds:
if on the interval $[0,+\infty[$ the function f is convex (concave) of order m and m+1, then the sequence (4.5) is decreasing (increasing) on $[0,+\infty[$.

We recall that, for $\alpha=0$, M_n^α operator coincides with M_n operator defined by (2.7); so, Corollary 4.2 represents an extension of a result previously established in [4] for the sequence $\{(M_n f)^{(m)}(x)\}_n$.

5. ON THE MONOTONICITY OF THE SEQUENCE $\{D_\alpha^m P_n^\alpha f(x)\}_n$

Let $P_n^\alpha f$ be the Baskakov-Stancu operator defined by (2.8). It is well-known [11] that the difference between two consecutive terms of the sequence $\{P_n^\alpha f(x)\}_n$ can be expressed as follows:

$$(P^{\alpha}_{n+1}-P^{\alpha}_{n})\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}x}{n(n+1)}\;\cdot$$

(5.1)
$$\cdot\sum_{k=0}^{\infty}\frac{(n+k+1)\ldots(n+2)}{(1+x)^{(n+k+1,-\alpha)}}\,\frac{(x+\alpha)^{(k,-\alpha)}}{k!}\left[\frac{k}{n+1},\frac{k+1}{n+1},\frac{k+1}{n};f\right],\quad x\geqq 0$$

Letting then

$$a_k=\frac{k}{n+1}\qquad\text{and}\qquad b_k=\frac{k}{n},\qquad k=0,1,\ldots$$

taking (2.2) into account, we have

$$(P^{\alpha}_{n+1}-P^{\alpha}_{n})\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}x}{n(n+1)}\;\cdot$$

(5.2)
$$\cdot\sum_{k=0}^{\infty}\frac{(n+k+1)\ldots(n+2)}{k!}\;\frac{(x+\alpha)^{(k,-\alpha)}}{(1+x)^{(n+k+1,-\alpha)}}\;T^{(2)}_{k}f$$

We introduce now, $\forall r\in N$, the linear positive operator $P^{\alpha,r}_{n}$

$$P^{\alpha,r}_{n}\,f(x) \;=\; \sum_{k=0}^{\infty}\frac{(n+k+1)\ldots(n+2)(x+r\alpha)^{(k,-\alpha)}}{(1+x)^{(n+k+r,-\alpha)}k!}\;f\left(\frac{k}{n}\right)$$

So (5.2) can be written as follows

$$(5.3)\quad (P^{\alpha}_{n+1}-P^{\alpha}_{n})\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}x}{n(n+1)}\;P^{\alpha,1}_{n}\;T^{(2)}_{k}f\,(x)$$

Now, in order to study the monotonicity of the sequence $\{D^{m}_{\alpha}P^{\alpha}_{n}\,f(x)\}_{n}$, we prove

Theorem 5.1. The following relationship holds

$$D^{m}_{\alpha}[P^{\alpha}_{n+1}-P^{\alpha}_{n}]\,f(x) \;=\; -\,\frac{1^{(n,-\alpha)}}{n(n+1)}\;m!\;\Big\{(x+m\alpha)\cdot$$

$$(5.4)\quad \cdot\Big[P^{\alpha,\,m+1}_{n}\binom{n+k+m}{m}T^{(m+2)}_{k}f(x)+P^{\alpha,m+1}_{n}\binom{n+k+m}{m-1}T^{(m+1)}_{k+1}f(x)\Big]+$$

$$+\Big[P^{\alpha,m}_{n}\binom{n+k+m-1}{m-1}T^{(m+1)}_{k}f(x)+P^{\alpha,m}_{n}\binom{n+k+m-1}{m-2}T^{(m)}_{k+1}f(x)\Big]\Big\}$$

$\forall m\in N$ and $0\leqq\alpha<\frac{1}{2}$

:oof.

, using (2.1) in (5.3), with

$$f(x) = -\frac{1^{(n,-\alpha)}x}{n(n+1)} \quad \text{and} \quad g(x) = P_n^{\alpha}, {}_1T_k^{(2)} f(x)$$

e obtain

$$D_\alpha^m[P_{n+1}^\alpha - P_n^\alpha] f(x) = -\frac{1^{(n,-\alpha)}}{n(n+1)} \cdot$$

$$\cdot \sum_{i=0}^{1} \binom{m}{i} D_\alpha^i (x+(m-i)\alpha) D_\alpha^{m-i} P_n^\alpha, {}_1T_k^{(2)} f(x) =$$

$$= -\frac{1^{(n,-\alpha)}}{n(n+1)} \left[(x+m\alpha) D_\alpha^m P_n^\alpha, {}_1T_k^{(2)} f(x) + m D_\alpha^{m-1} P_n^\alpha, {}_1T_k^{(2)} f(x) \right]$$

n the other hand, we can easily verify by induction that

$$D_\alpha^m P_n^\alpha, {}_1T_k^{(2)} f(x) = m! \left[P_n^{\alpha,m+1} \binom{n+k+m}{m} T_k^{(m+2)} f(x) + P_n^{\alpha,m+1} \binom{n+k+m}{m-1} T_{k+1}^{(m+1)} f(x) \right]$$

nd, by taking this last relation into account in (5.5), the heorem follows.

rom Theorem 5.1, by (2.3), we have

orollary 5.2. The sequence

$$5.6) \quad \{ D_\alpha^m P_n^\alpha f(x) \}_n , \quad m \in \mathbb{N}, \quad 0 \leq \alpha < \frac{1}{2}$$

erifies the following monotonicity properties

) for m=1

if on the interval $[0,+\infty[$ the function f is convex (concave) of first and second order, then the sequence (5.6) is decreasing (increasing) on $[0,+\infty[$;

i) for $m \geq 2$

if on the interval $[0,+\infty[$ the function f is convex (concave) of order m-1, m and m+1, then the sequence (5.6) is decreasing (increasing) on $[0,+\infty[$.

We notice that for $\alpha=0$, P_n^α operator becomes equal to P_n operator defined by (2.9); so, Corollary 5.2 generalizes a re-

sult established in [4] for the sequence $\{(P_n f)^{(m)}(x)\}_n$.

REFERENCES

1. O. ARAMĂ: Proprietăţi privind monotonia şirului polinoamelor de interpolare ale lui S.N. Bernstein şi aplicarea lor la studiul aproximării funcţiilor. Stud. Cerc. Mat. (Cluj) 8 (1957), 195-210.
2. V.A. BASKAKOV: An instance of a sequence of linear positive operators in the space of continuous functions. Dokl. Akad. Nauk. SSSR 113 (1957), 249-251.
3. E.W. CHENEY and A. SHARMA: Bernstein power series. Canad. J. Math. XVI, 2 (1964), 241-252.
4. B. Della Vecchia: On the monotonicity of the derivatives of the sequences of Favard and Baskakov operators. Submitted to Ricerche di Matematica (1987).
5. J. FAVARD: Sur les multiplicateurs d'interpolation. J. Math. Pures Appl. 23 (1944), 219-247.
6. T. HERMANN: On Baskakov-type operators. Acta Math. Sci. Hungar. 31, (3-4) (1978), 307-316.
7. I. HOROVÁ: A note on the sequence formed by the first order derivatives of the Szász-Mirakyan operators. Mathematica 24 (47) (1982), 49-52.
8. I. HOROVÁ: Linear positive operators and their applications to differnetial equations. Arch. Math. (Brno) 20 (1984), 1-8.
9. A. LUPAŞ: On Bernstein power series. Mathematica 8 (31) (1966), 287-296.
10. A. LUPAŞ and M.W. MÜLLER: Approximation properties of the M_n-Operators. Aeq. Math. 5 (1970), 19-37.
11. G. MASTROIANNI: Su una classe di operatori lineari e positivi. Rend. Accad. Scien. M.F.N. Serie IV XLVIII (1980).
12. G. MASTROIANNI: Una generalizzazione dell'operatore di

Mirakyan. Rend. Accad. Sci. M.F.N. Serie IV XLVIII (1980).

3. G. MASTROIANNI and M.R. OCCORSIO: Sulle derivate dei polinomi di Stancu. Rend. Accad. Sci. M.F.N. Serie IV XLV (1978).

4. G. MASTROIANNI and M.R. OCCORSIO: Una generalizzazione dell'operatore di Stancu. Rend. Accad. Sci. M.F.N. Serie IV XLV (1978).

5. G. MIRAKYAN: Approximation des fonctions continues au moyen de polynomes de la forme ... Dokl. Akad. Nauk. SSSR 31 (1941), 201-205.

6. O. SZASZ: Generalization of Bernstein's polynomials to the infinite interval. J. Res. Nat. Bur. Standards 45 (1950), 239-245.

7. S.P. PETHE and G.C. JAIN: Approximation of functions by a Bernstein-type operator. Canad. Math. Bull. 15 (4) (1972), 551-557.

8. D.D. STANCU: On the monotonicity of the sequence formed by the first order derivatives of the Bernstein polynomials. Math. Z. 98 (1967), 46-51.

9. D.D. STANCU: Approximation of functions by a new class of linear polynomial operators. Rev. Roumanie Math. Pures Appl. 13 (1968), 1173-1194.

10. D.D. STANCU: Approximation properties of a class of linear positive operators. Studia Univ. Babes-Bolyai, Cluj, Ser. Mat. 2 (1970), 33-38.

11. D.D. STANCU: Two classes of positive linear operators. Analele Univ. Timişoara, Ser. Mat. 8 (1970), 213-220.

12. D.D. STANCU: On the remainder of approximation of functions by means of parameter-dependent linear polynomial operator. Studia Univ. Babes-Bolyai, Cluj 16 (1971), 59-66.

23. D.D. STANCU: Approximation of functions by means of some new classes of positive linear operators. In: Proc. Conf.

Math. Res. Inst. Oberwolfach (L. Collatz, G. Meinardus eds.), 1972.

24. D.D. STANCU: Application of divided differences to the study of monotonicity of the derivatives of the sequences of Bernstein polynomials. Calcolo 16 (1979), 431-445.
25. D.D. STANCU: A study of the remainder in an approximation formula using a Favard-Szasz type operator. Studia Univ. Babes-Bolyai, Mathematica XXV (1980), 70-76.
26. F. STANCU: Asupra restului în formulele de aproximare prin operatorii lui Mirakian de una şi două variabile. Analele Şt. Univ. "Al. I. Cuza", Iaşi 14 (1968), 415-422.
27. F. STANCU: Asupra aproximării funcţiilor de una şi două variabile cu ajutorul operatorilor lui Baskakov. St. Cerc. Mat. 22 (1970), 531-542.
28. W.B. TEMPLE: Stieltjes integral representation of convex functions. Duke Math. J. 21 (1954), 527-531.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

OPTIMAL PERIODIC INTERPOLATION IN THE MEAN

F.-J. DELVOS

ABSTRACT: *The concept of periodic Hilbert spaces was introduced by Babuska in connection with universally optimal quadrature formulas. It was shown by Prager, Locher, Knauff - Kress, and the author that periodic Hilbert spacesform an appropriate tool for constructing periodic interpolation splines and some of its extensions such as rational trigonometric interpolation. It was pointed out by Subbotin that it is natural to approximate functions from L^1 via interpolation in the mean splines. In this paper we will develop the method of optimal interpolation in the mean in periodic Hilbert spaces. Applications to periodic splines are presented.*

1. TRIGONOMETRIC INTERPOLATION IN THE MEAN

We denote by $\tau_{0,n-1}$ the n-dimensional space of trigonometric polynomials spanned by the functions

$$e_k(t) = \exp(ikt) \quad (0 \le k < n) .$$

Recall that $\tau_{0,n-1}$ is the appropriate space for discussing the discrete Fourier transform method. Assume that there are n real numbers $t_0,\ldots,t_{n-1}$ and a positive real number h satisfying

$$0 = t_0 < t_1 < t_2 < \ldots < t_{n-1} < 2\pi$$

and

$$0 < h \le 1/n , \quad h \le t_{j+1}-t_j \quad (0 \le j < n)$$

with $t_n = 2\pi$. The *interpolation in the mean functionals* are given by

$$L_{j,h}(f) = \frac{1}{h} \int_{t_j}^{t_j+h} f(t)\, dt \quad (0 \le j < n) .$$

Note that the interpolation functionals

$$L_j(f) = f(t_j) \quad (0 \le j < n)$$

are obtained formally by setting $h = 0$.

Proposition 1.1

For any $f \in L^1_{2\pi}$ there is a unique trigonometric polynomial $H_n(f) \in \tau_{0,n-1}$ satisfying the interpolation conditions

$$L_{j,h}(H_n(f)) = L_{j,h}(f) \quad (0 \le j < n) \ .$$

Proof. It is sufficient to show that

$$A = (a_{j,k}) := (L_{j,h}(e_k))_{0\le j,k<n} \ .$$

is a regular matrix. It is easily seen that

$$L_{j,h}(e_0) = 1 \quad (0 \le j < n) \ ,$$
$$L_{j,h}(e_k) = \exp(ikt_j)(-1+\exp(ikh))/(ikh) \quad (0\le j<n \ , \ 0<k<n) \ .$$

This implies $A = VD$ with

$$V = (\exp(ikt_j))_{0\le j,k<n}$$

and

$$D = \mathrm{diag}(d_0,\ldots,d_{n-1}) \ ,$$

$$d_0 = 1 \ , \ d_k = (\exp(ikh) - 1)/(ikh) \quad (0 < k < n) \ .$$

V is a Vandermonde matrix and D is a regular diagonal matrixin view of $\exp(ikh) \neq 1$. This completes the proof of Proposition 1.1

2. PERIODIC HILBERT SPACES

In this section we recall the properties of periodic Hilbert spaces as described in Prager [10] . Let

$$d_k \quad (k \in \mathbb{Z})$$

be a biinfinite sequence of real numbers from l^1 satisfying

(2.1) $$d_k = d_{-k} > 0 \quad (k \in \mathbb{Z}) \ , \ d_0 = 1 \ .$$

Then there exists a unique function ψ from the Wiener algebra $A_{2\pi}$ satisfying

$$d_k = (\psi, e_k) = \frac{1}{2\pi}\int_0^{2\pi} \psi(t)\exp(-ikt)\,dt \quad (k \in \mathbb{Z}) \ .$$

It is obvious that ψ is real-valued and even :

$$\psi(t) \quad = \quad 1 + 2 \sum_{k=1}^{\infty} d_k \cos(kt) \quad .$$

The periodic Hilbert space H_d related to $d = (d_k)$, respectively ψ, is defined by

$$H_d \quad = \quad \{ f \in L^2_{2\pi} : \sum_{k=-\infty}^{\infty} (f,e_k)(e_k,f)/d_k \; < \; \infty \} \quad .$$

The inner product of H_d is given by

$$(f,g)_d \quad = \quad \sum_{k=-\infty}^{\infty} (f,e_k)(e_k,g)/d_k \quad .$$

Obviously, H_d contains the algebra τ of trigonometric polynomials. Moreover, Prager showed that

(2.2) $\qquad H_d \quad \subseteq \quad A_{2\pi} \quad .$

It is easily seen that for $f \in H_d$ and $a \in \mathbb{R}$ we have

$$f(.-a) \in H_d \quad ,$$

i. e., H_d is closed with respect to translation. Moreover, we have

(2.3) $\qquad (f(.-a),g(.-a))_d \quad = \quad (f,g)_d$

for all functions $f,g \in H_d$.

Proposition 2.2

Let $f \in H_d$ and $x \in \mathbb{R}$. Then we have

$$f(x) \quad = \quad (f,\psi(.-x))_d \quad .$$

Thus, H_d is a reproducing kernel Hilbert space of periodic functions with kernel $K(y,x) = \psi(y-x)$.

The function $\psi(.-x)$ is the representer of the Dirac measure δ_x which is a bounded linear functional on H_d . Let L be a bounded linear functional on H_d and $u \in H_d$ be its representer. Then we have

(2.4) $\qquad L(f) \quad = \quad (f,u)_d$

for all functions $f \in H_d$. It was shown by Prager [10] that the Fourier series of u is given by the formula

$$(2.5) \qquad u(t) = \sum_{k=-\infty}^{\infty} d_k \overline{L(e_k)}\, e_k(t) \quad .$$

As an example we consider the construction of the periodic Sobolev space $W_{2\pi}^r$ with $r \in \mathbb{N}$. The defining sequence d is given by

$$d_0 = 1 \quad , \quad d_k = k^{-2r} \quad (k \neq 0)$$

and we have $H_d = W_{2\pi}^r$. The reproducing kernel of $W_{2\pi}^r$ satisfies the relation

$$(2.6) \qquad K_r(y,x) = 1 + (-1)^r B_{2r}(y-x) = \psi(y-x) \quad ,$$

$$B_q(x) = \sum_{k\neq 0} (ik)^{-q} e_k(x) \quad .$$

$B_q(x)$ is the periodic Bernoulli function of order q which is defined uniquely by the relations

$$(2.7) \qquad B_1(x) = \pi - x \quad , \quad B'_{q+1}(x) = B_q(x) \quad , \quad (B_{q+1}, e_0) = 0$$

$$(0 < x < 2\pi) \quad .$$

3. OPTIMAL INTERPOLATION IN THE MEAN

In this section we will study interpolation in the mean as a minimum norm interpolation problem in the reproducing kernel Hilbert space H_d [8].

Proposition 3.1

The linear functionals $L_{0,h}, \ldots\ldots, L_{n-1,h}$ are bounded and linearly independent over H_d.

Proof. It follows from Proposition 2.2 that the following estimate

$$(3.1) \qquad \|f\|_\infty \leq \sqrt{\psi(0)}\, \|f\|_d$$

holds for all functions $f \in H_d$. Thus, the interpolation in the mean functionals $L_{j,h}$, $0 \leq j < n$, are bounded. Since the trigonometric polynomials are contained in H_d it follows from Proposition 1.1 that $L_{0,h}, \ldots\ldots, L_{n-1,h}$ are linearly independent over H_d.

we will determine the representers u_j of $L_{j,h}$ for $j =$ $,\ldots,n-1$. For this construction we need the *periodic integral* Ψ of . Recall that Ψ is the unique function from $C^1_{2\pi}$ such that

3.2) $\Psi'(t) = \psi(t) - (\psi,e_0)$, $(\Psi,e_0) = 0$.

he Fourier series of Ψ is given by

3.3) $\Psi(t) = \sum_{k\neq 0} d_k (ik)^{-1} e_k(t) = \sum_{k=1}^{\infty} 2d_k \sin(kt)/k$.

roposition 3.2

The representer u_j of $L_{j,h}$ is given by the formula

3.4) $u_j(t) = 1 + (\Psi(t-t_j) - \Psi(t-t_j-h))/h$

ith $0 \leq j < n$.

roof. Recall that

$L_{j,h}(e_0) = 1$, $L_{j,h}(e_k) = e_k(t_j)(e_k(h)-1)/(ikh)$ ($k \neq 0$).

Jsing relation (2.5) and relation (3.3) we can conclude

$$u_j(t) = \overline{L_{j,h}(e_0)} + \sum_{k\neq 0} d_k \overline{L_{j,h}(e_k)}\, e_k(t)$$

$$= 1 - \sum_{k\neq 0} (d_k/(ikh))\, e_k(-t_j-h)\, e_k(t)$$

$$+ \sum_{k\neq 0} (d_k/(ikh))\, e_k(-t_j)\, e_k(t)$$

$$= 1 - (\Psi(t-t_j) - \Psi(t-t_j-h))/h \quad .$$

This completes the proof of Proposition 3.2 .

_et U_n be the linear span of the representers $u_0,\ldots,u_{n-1}$ and let S_n be the orthogonal projector in H_d with

$$\mathfrak{R}(S_n) = U_n = \langle u_0,\ldots,u_{n-1} \rangle \quad .$$

The following result is a consequence of the method of minimum norm interpolation in Hilbert spaces (Prager [10]).

Proposition 3.3

Let $f \in H_d$ be given. Then $S_n(f) \in U_n$ is the unique function in H_d having the characteristic properties

$$(i) \quad \frac{1}{h}\int_{t_j}^{t_j+h} S_n(f)(t)\,dt = \frac{1}{h}\int_{t_j}^{t_j+h} f(t)\,dt \quad (0 \le j < n) ,$$

$$(ii) \quad \|S_n(f)\|_d \le \|g\|_d \text{ if } L_{j,h}(g) = L_{j,h}(f) \quad (0 \le j < n) .$$

As an example we determine the *periodic interpolation in the mean splines* which are obtained by choosing $H_d = W^r_{2\pi}$ [2,6,8,11] . It follows from Proposition 3.2 and relations (2.6) and (2.7) that the representers $u_0,\dots,u_{n-1}$ are given by

$$(3.5) \quad u_j(t) = 1 + (-1)^r(B_{2r+1}(t-t_j) - B_{2r+1}(t-t_j-h))/h \quad (0 \le j < n) .$$

The properties of the Bernoulli functions imply that u_j is a periodic spline of degree $2r$ with spline knots $t_j+2\pi k$, $t_j+h+2\pi k$ $(k \in \mathbb{Z})$. As a consequence U_n is an n-dimensional space of periodic splines of degree $2r$ with splines knots

$$t_j + 2\pi k \ , \ t_j + h + 2\pi k \quad (0 \le j < n \ , \ k \in \mathbb{Z}) .$$

Let us consider the case of a uniform mesh, i. e. ,

$$t_j = 2\pi j/n \quad (j \in \mathbb{Z}) .$$

Then the space of interpolation in the mean splines is generated by translation from the *generating function*

$$(3.6) \quad u(t) = 1 + (B_{2r+1}(t) - B_{2r+1}(t-h))/h ,$$

i. e., we have

$$(3.7) \quad U_n = \langle u(.-t_0),\dots,u(.-t_{n-1})\rangle =: V_n(u) .$$

It should be noted that for the special case $h = t_1$ the space $V_n(u)$ is just the space of periodic splines of degree $2r$ with knots t_j. This follows from the fact that

$$1 = (u(t-t_0) + u(t-t_1) + \dots + u(t-t_{n-1}))/n .$$

Let $v \in U_n = V_n(u)$ be the unique function satisfying

3.8) $L_{j,h}(v) = \delta_{0,j} \quad (0 \le j < n)$.

Since U_n is translation invariant with respect to t_1 it follows from the relation

3.9) $$L_{j,h}(f) = \frac{1}{h} \int_0^h f(t+t_j)\, dt = L_{0,h}(f(.+t_j))$$
$(0 \le j < n)$

that the interpolation in the mean spline $S_n(f)$ is given by the formula

3.10) $$S_n(f)(t) = \sum_{j=0}^{n-1} L_{j,h}(f)\, v(t-t_j) \quad .$$

(*Interpolation in the mean by translation*).

4. THE CONSTRUCTION OF THE FUNDAMENTAL FUNCTION

For the case of a uniform mesh we will apply the method of the discrete Fourier transform to derive an explicit formula for *interpolation in the mean fundamental function* v . Recall that

(4.1) $$u(t) = 1 + (\Psi(t) - \Psi(t-h))/h \quad ,$$

$$u_k(t) = u(t-t_k) \quad (0 \le k < n) .$$

Furthermore, let ϕ be the periodic integral of Ψ , i. e.,

(4.2) $$\phi'(t) = \Psi(t) \quad , \quad (\phi, e_0) = 0 \quad .$$

Since $\Psi(-t) = -\Psi(t)$ it follows

(4.3) $$\phi(-t) = \phi(t) \quad .$$

Then we can conclude

$$L_{j,h}(u_k)$$

$$= 1 + h^{-2}\left(\int_{t_j}^{t_j+h} \Psi(t-t_k)\, dt - \int_{t_j}^{t_j+h} \Psi(t-t_k-h)\, dt \right)$$

$$= 1 + h^{-2} \left(\phi(t_{j-k}+h) + \phi(t_{j-k}-h) - 2\phi(t_{j-k}) \right) \quad ,$$

i. e., we have

$$(4.4)\quad L_{j,h}(u_k) = 1 + h^{-2}(\phi(t_{j-k}+h)+\phi(t_{j-k}-h)-2\phi(t_{j-k})) =: w(j-k)$$
$$(\ 0 \le j,k < n\ )\quad .$$

It follows from the definition of u_k that the matrix

$$(4.5)\quad T = (L_{j,h}(u_k))_{0\le j,k<n}$$

is a circulant Toeplitz matrix which is also positive definite. The interpolation in the mean fundamental function v is given by

$$(4.6)\quad v(t) = \sum_{k=0}^{n-1} c_k\, u(t-t_k)$$

with

$$(4.7)\quad \sum_{k=0}^{n-1} w(j-k)\, c_k = \delta_{0,k} \quad (\ 0 \le j < n\ )\quad .$$

Using discrete Fourier transform methods [3] we obtain the following explicit formulas

$$(4.8)\quad c_k = \frac{1}{n}\sum_{j=0}^{n-1} e_k(t_j)/a_j \quad (\ 0 \le k < n\ )\quad ,$$

$$a_j = \sum_{k=0}^{n-1} w(k)\, e_j(-t_k) \quad (\ 0 \le j < n\ )\quad .$$

For the practically important case $n = 2m$ these formulas reduce to

$$(4.9)\quad a_j = w(0) + w(m)\cos(\pi j) + 2\sum_{k=1}^{m-1} w(k)\cos(jt_k)\quad ,$$

$$c_k = \frac{1}{n}(a_0^{-1} + \cos(\pi k)a_m^{-1} + 2\sum_{j=1}^{m-1}\cos(kt_j)a_j^{-1}) \quad (\ 0 \le j,k < n\ )\ .$$

5. EXAMPLES

In this section we determine for two special choices of ψ the related functions Ψ and ϕ . The first example is concerned with the function

$$(5.1)\quad \psi(t) = 1 + (-1)^r B_{2r}(t)\quad ,\ r \in \mathbb{N}\ .$$

The related periodic Hilbert space H_d is the periodic Sobolev space

$w_{2\pi}$. Using the properties of the Bernoulli functions we obtain the following formulas:

$$(5.1') \quad \Psi(t) = (-1)^r B_{2r+1}(t) \quad , \quad \phi(t) = (-1)^r B_{2r+2}(t) \quad .$$

The space of optimal periodic interpolants in the mean $\mathfrak{K}(S_n)$ consists of periodic splines of even degree.

The second example is characterized by the function

$$(5.2) \quad \psi(t) = 1 - (\cosh(b) - 1)/(\cosh(b) - \cos(t))^2 \quad , \; b > 0.$$

In this case H_d is a space of periodic functions having holomorphic extensions in the strip $|\mathrm{Im}(z)| < b$:

$$H_d = \{ f \in L^2_{2\pi} : \sum_{k\neq 0} |(f,e_k)|^2 |k|^{-1} \exp(b|k|) < \infty \} \quad .$$

It is easily seen that the following relations hold :

$$(5.2') \quad \Psi(t) = -\sin(t)/(\cosh(b) - \cos(t)) \quad ,$$

$$\phi(t) = -\ln(\cosh(b) - \cos(t)) \quad .$$

In this case $\mathfrak{K}(S_n)$ is a linear space of rational trigonometric functions.

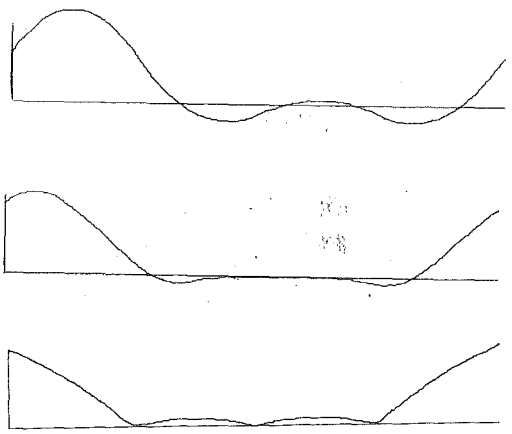


Fig. 1 Fundamental interpolation in the mean function v for $\psi(t) = 1 - B_2(t)$ with $n = 4$ and $h = t_1, t_1/2, t_1/128$

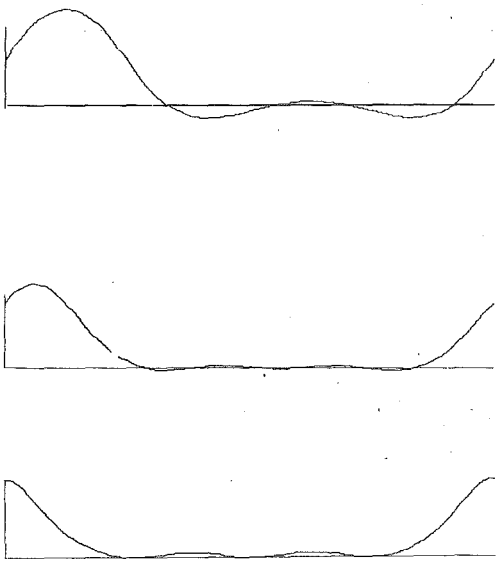


Fig. 2 Fundamental interpolation in the mean function v for $\psi(t) = 1-(\cosh(1)-1)(\cosh(1)-\cos(t))^{-2}$, $n = 4$ and $h = t_1,\ t_1/2,\ t_1/128$.

References

1. I. BABUSKA : Ueber universal optimale Quadraturformeln, Teil 1, Apl. mat. 13(1968), 304-338, Teil 2, Apl. mat. 13(1968), 388-404 .

2. G. BASZENSKI and L.L. SCHUMAKER : On a method for fitting an unknown function based on mean value measurements . SIAM J. Numer. Anal. 24(1987), 725-736 .

3. D.R. BRILLINGER: "Time series: data analysis and theory". Holt, Rinehardt and Winston, Inc., New York, 1975.

4. F.-J. DELVOS: Periodic interpolation on uniform meshes . J. Approx. Theory 49(1987) (to appear) .

5. F.-J. DELVOS: Convergence on interpolation by translation. Colloquia Mathematica János Bolyai 49, In: Alfred Haar Memorial Conference , Budapest (Hungary), 1985, pp. 273-287.

6. F.-J. DELVOS: Periodic area matching interpolation.
In: Numerical Methods of Approximation Theory (L. Collatz, G. Meinardus, G. Nürnberger , eds.), ISNM 81, Birkhäuser Verlag, Basel 1987, pp. 54-66.

7. W.KNAUFF and R. KRESS: Optimale Approximation linearer Funktionale auf periodischen Funktionen. Numer. Math. 22(1974), 187-205.

8. P.-J. LAURENT: Approximation et optimisation. Herrmann, Paris, 1972.

9. F. LOCHER: Interpolation on uniform meshes by translates of one function and related attentuation factors. Mathematics of Computation 37(1981), 403-416 .

10. M. PRAGER: Universally optimal approximation of functionals. Apl. mat. 24(1979), 406-420 .

11. Yu. SUBBOTIN: Extremal problems of functional interpolation and splines for interpolation in the mean. Proc Steklov Inst. Math. 138(1975), 127-185.

Numerical Methods and Approximation Theory III (Niš, August 18-21, 1987)

ACCURATE EXPLICIT FINITE DIFFERENCE SOLUTION OF THE SHOCK TUBE PROBLEM

S.K. DEY and CHARLIE DEY

Abstract

A simple predictor-corrector algorithm has been developed in [1] for numerical solution of initial-value problems. In this article we will discuss computer experimentation of this method for solution of the shock tube problem.

Introduction

One dimensional motion of compressible flow is described by:

$$U_t + F_x = 0 \tag{1}$$

where

$$U = (\rho,\ \rho u,\ e)^T$$

$$F = (a,\ b,\ c)^T$$

$$a = \rho u, \quad b = (\gamma-1)e + ((3-\gamma)/2)\ \rho u^2$$

$$c = \gamma e u - ((\gamma-1)2)\ \rho u^3. \tag{2}$$

ρ = density, u = velocity, p = pressure, e = total energy per unit volume, given by $e = \rho\varepsilon + u^2/2$, ε = internal energy per unit mass. For a perfect gas pressure p is defined by

$$p = (\gamma-1)(e - \rho u^2/2), \quad \gamma = 1.4$$

subject to various initial/boundary conditions, this set of equations will describe various compressible flow models.

Now let us consider the shock tube problem. Let two gases separated by a diaphragm be in equilibrium in a tube. Let the densities of them be unequal. If the diaphragm is suddenly broken, the gas molecules start mixing. This mixing phenomenon is often referred to as the shock tube problem. Here we will use the following initial conditions:

At $t = 0$, $\rho = 1$, $u = 0$, $e = 1/(\gamma-1)$ for $0 \leq x \leq 1.9$ and $\rho = 0.1$, $u = 0$, $e = 0.1/(\gamma-1)$ for $1.9 < x \leq 5$.

As the gas molecules start mixing, sharp changes of density, pressure, velocity and energy take place at several points along the x-axis. To describe this phenomenon appropriately by a numerical algorithm is often a challenge for researchers in computational fluid dynamics.

This challenge has been undertaken by many researchers in the past [2, 3, 4], and in some cases excellent results have been found. In this work an explicit finite difference scheme, whose algorithm is much simpler than all the above methods, has been successfully applied to obtain quite accurate numerical solutions of the shock tube problem. The algorithm has been developed by the second author and applied extensively by him to solve several linear and nonlinear models in Engineering and Applied Mathematics [1]. Let us briefly describe the algorithm and some of its properties.

The Algorithm

Let us consider an initial-value problem

$$du/dt = f(u, t), \quad u(t_o) = u_o \tag{3}$$

A predictor-corrector algorithm to solve (3) may be expressed as:

$$\hat{U} = U_n + \Delta t\, f(U_n, t_n) \qquad \text{predictor} \tag{4a}$$

$$U_{n+1} = (1-\gamma)\,\hat{U} + \gamma\,\{U_n + \Delta t\, f(\hat{U}, t_{n+1})\} \qquad \text{corrector} \tag{4b}$$

where $U_n = U(t_n)$, Δt = step size, $\hat{U}$ is the predicted value of U at t_{n+1}, U_{n+1}, is the corrected value of U at t_{n+1}.

γ is called a filtering parameter, and it is assumed that $0 < \gamma < 1$. The predictor is Euler's forward difference approximation. If $\gamma = 0.5$, then (4a) and (4b) are reduced to a second-order Runge-Kutta scheme. It is expected that the corrector should filter most errors generated by the predictor. But one must choose a value of γ before the algorithm may be used. Such a choice for the value of γ may be obtained if we do the stability analysis of this numerical method [1].

Linearized Stability Analysis

If we linearize (3), write $du/dt = \lambda u$ and use (4a) and (4b), we get the combined form of (4a) and (4b) as,

$$U_{n+1} = \sigma U_n, \text{ where } \sigma = 1+z+\gamma z^2, \; z = \lambda \Delta t.$$

For stability, $|\sigma| \leq 1$. Let us consider the following example to look into an interesting property of this method. $du/dt = -80u$, $u(0) = 1$. The analytical solution is $u(t) = e^{-80t}$. Here $\sigma = 1-80h + 6400\gamma h^2$. With $\gamma = 0.1$, if $h = 0.01$, $|\sigma| < 1$ (stable), if $h = 0.07$, $|\sigma| > 1$ (unstable) and if $h = 0.1$, $|\sigma| < 1$ (stable). When the algorithm is stable, it gives quite accurate steady-state solutions. But often time accurate solutions given by this scheme are not up to expectations. This is true for one equation or a system of equations.

Since λ may be complex, z may be taken to be a complex variable, and hence $\sigma(z)$ is a complex function. Lomax [5] developed a computer code such that for a given γ, $\sigma(z)$ may be plotted in a complex plane and the region of stability may be found. Some of these contours have been described in [1].

Difference Approximation of (1)

The equation (1) may be approximated as follows:

$$U_j^{n+1} = U_j^n + \Delta t \, (F_{j-1}^n - F_{j+1}^n)/(2\Delta x) \qquad (5)$$

or

$$U_j^{n+1} = U_j^n + \Delta t\ (F_{j-1}^n - F_j^n)/\Delta x \qquad (6)$$

where $U_j^n = U(x_j, t_n)$. To stabilize the numerical process, an artificial viscosity term was introduced. This second-order derivative was approximated by central differences. Using (5), the predictor-corrector algorithm is:

$$\hat{U}_j = U_j^n + (\Delta t/2\Delta x)(F_{j-1}^n - F_{j+1}^n)$$

$$U_j^{n+1} = (1-\gamma)\ \hat{U}_j + \gamma\cdot\{U_j^n + (\Delta t/2\Delta x)(\hat{F}_{j-1}^{n+1} - \hat{F}_{j+1})\} \qquad (7)$$

Since there are three components of U, for each component a unique value of γ may be chosen.

Discussions

Figures 1, 2, and 3 which describe distributions of density, pressure and velocity were obtained by using the predictor-corrector with (5) as predictor. Figures 4, 5, and 6 describe distributions of the same and were obtained using the same algorithm with (6) as predictor. If the filtering parameters are not selected properly the results may not be reasonably correct. This often poses a problem, since (1) is a nonlinear model. For nonlinear Burgers' equation the model was linearized and γ was computed using the contours of stability [1]. This has not yet been done for the Euler's equation (1). We hope that such a stability analysis may resolve the problem in the future.

Acknowledgement

This research was conducted at the University of Siedlce, Poland, where both authors received research fellowships to undertake this work.

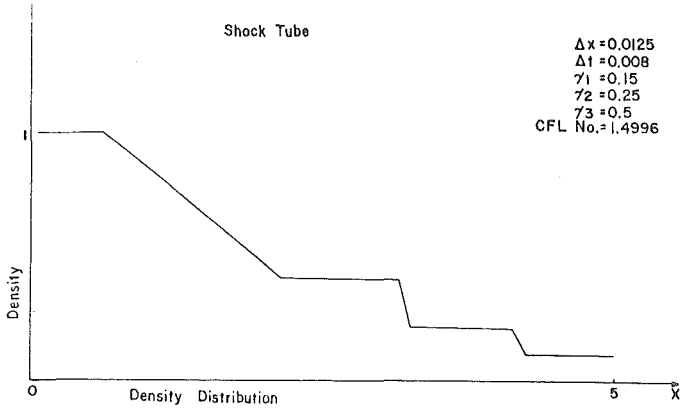


FIGURE 1

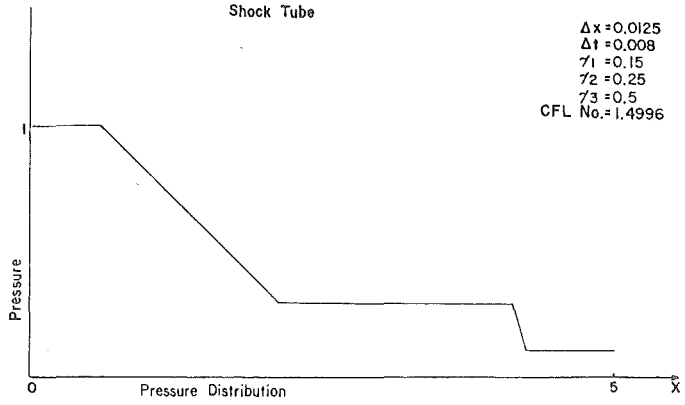


FIGURE 2

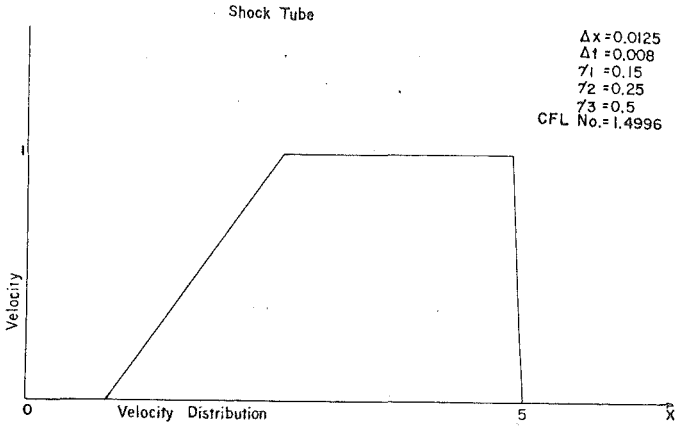


FIGURE 3

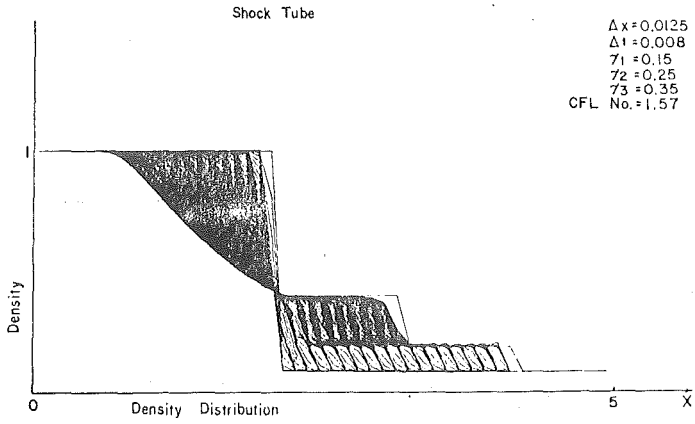


FIGURE 4

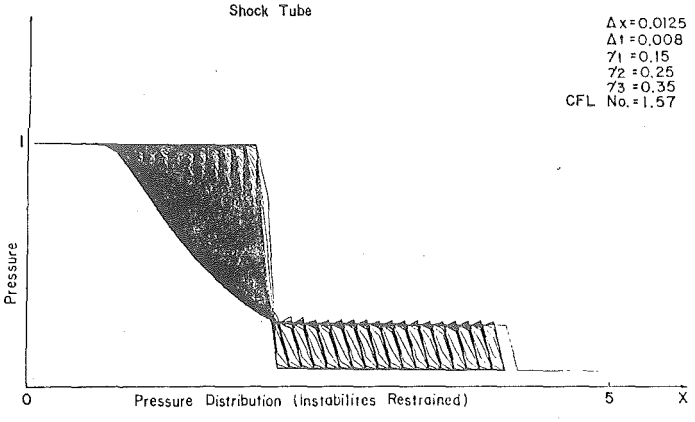


FIGURE 5

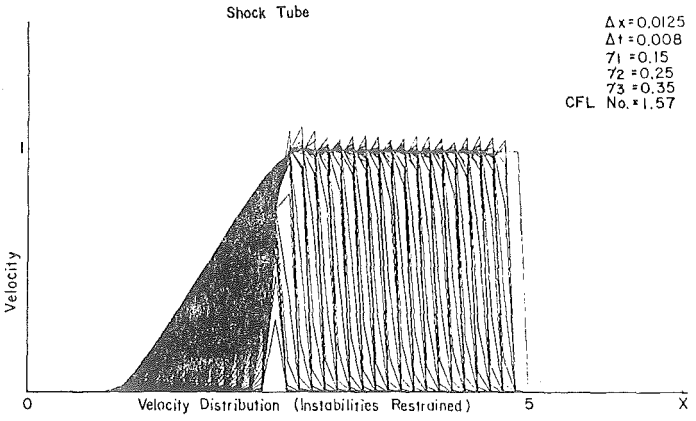


FIGURE 6

References

1. S. K. Dey and Charlie Dey, "An Explicit Predictor-Corrector Solver with Applications to Burgers' Equation." NASA Technical Memo 84402, September 1983. Ames Research Center, Moffett Field, CA 94035.

2. P. L. Roe, "Approximate Riemann Solvers, Parameter Vectors and Difference Schemes." J. Comp. Physics, Vol. 43, 1981.

3. E. J. Kansa, "Highly Accurate Shock Flow Calculations with Moving Grids and Mesh Refinement." Pro. Int. Congress on Sci. Comp. IMAC, Oslo, Norway, 1985.

4. S. K. Dey, "Numerical Solution of Euler's Equation by Perturbed Functionals." Lectures in Applied Math, AMS, Vol. 22, 1985.

5. H. Lomax, NASA Ames Research Center, Private Communication.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

SOME ASPECTS OF AUTOMATIC DIFFERENTIATION

HERBERT FISCHER

Abstract: Gradient and Hessian matrix of an explicitely given function can be computed automatically and straightforward by way of "automatic differentiation". This method is applicable to a broad class of functions. No quotients of differences are used. And no symbolic manipulation of symbols is involved. Complexity considerations show that "automatic differentiation" is competitive and efficient.

1. INTRODUCTION

Gradient and Hessian matrix of a real function of several variables play an important role in many numerical methods, especially in Nonlinear Optimization. But little effort has been devoted to the computation of these entities so far. "The Hessian matrix is not available." This statement used to be an axiom in the optimization folklore for decades. It led to the construction of well-known algorithms for nonlinear optimization problems, where the Hessian matrix respectively its inverse is approximated. Nevertheless, we will show how to obtain gradient and Hessian matrix "automatically" in an easy and straightforward manner. No manipulation of symbols is involved, we deal with numbers, not with formulas. This complies with the fact that, within the implementation of a relevant numerical method, gradient and Hessian matrix themselves are of interest rather than formulas for them.

Let us revisit the Hessian situation. Assume f is a twice differentiable function of several variables and $\bar{x}$ is a point in the domain of f. Assume further, we need the Hessian matrix $H(\bar{x})$ of f at $\bar{x}$. There are various approaches to compute $H(\bar{x})$, for instance

)

(1) derive a formula for H(x) and evaluate this formula for the specific $\bar{x}$,

(2) approximate $H(\bar{x})$ by a matrix of quotients of differences, using the gradient of f or an approximation thereof,

(3) "update" somehow a previously obtained approximation to $H(\bar{y})$, where $\bar{y}$ is "near" $\bar{x}$,

(4) use Automatic Differentiation.

The approach (1) is cumbersome, time consuming and prone to error, even if an outside computer-program for manipulation of symbols is used.
The numerical differentiation mentioned in (2) inevitably leads into the well-known predicament: a large stepsize yields inaccurate values and a small stepsize makes the computational process instable.
The way (3) may be considered an emergency measure.
The approach (4) seems to be the easiest one. It is astonishing that for a long time the idea of Automatic Differentiation has been overseen or ignored, despite quite a number of publications in this direction. We should mention that most of the protagonists' papers were too intermingled with programming languages or had to establish their own differentiation programming system. This may have hindered general recognition and delayed use. The breakthrough in Automatic Differentiation came with the work of L.B. Rall.
The automatic generation of gradient and Hessian matrix for a broad class of functions $\mathbb{R}^n \to \mathbb{R}$ may well restrict the above "axiom" to very expensive functions and to functions which are defined implicitely.

2. THE IDEA

In this section we sketch the basic idea of Automatic Differentiation as far as gradient and Hessian matrix are concerned.
Assume we have a rational function

$$r: D \subseteq \mathbb{R}^n \to \mathbb{R}$$

in explicit form. This means, for r(x) we have a formula which only contains the components $x_1, x_2, \ldots, x_n$ of x, some real numbers, the four arithmetic operations addition, subtraction, multiplication and division, and parentheses at proper places.

Let $r_G(x)$ denote the gradient of r at $x \in D$ and

let $r_H(x)$ denote the Hessian matrix of r at $x \in D$.

case: r is primitive

$r(x) = x_i$ = i-th component of x, for some $i \in \{1,2,\ldots,n\}$.

For any $x \in D$ we have $r_G(x)$ = i-th unit-vector, $r_H(x)$ = zero-matrix.

case: r is constant

$r(x) = \text{const} = c$, for some $c \in \mathbb{R}$.

For any $x \in D$ we have $r_G(x)$ = zero-vector , $r_H(x)$ = zero-matrix.

case: r is neither primitive nor constant

We employ Cesar's rule *divide et impera*, which of course in our situation reads

split and differentiate!

In splitting the formula for r(x), we obtain one of the four cases

(A) $r(x) = a(x) + b(x)$

(S) $r(x) = a(x) - b(x)$

(M) $r(x) = a(x) \cdot b(x)$

(D) $r(x) = a(x) / b(x)$

where a and b are rational functions $D \subseteq \mathbb{R}^n \to \mathbb{R}$. Furthermore, the functions a and b are available in explicit form and the formulas for a(x) and b(x) are shorter than the formula for r(x).

To follow the rule, we differentiate the function r. This yields

(A') $r_G = a_G + b_G$

(S') $r_G = a_G - b_G$

(M') $r_G = b \cdot a_G + a \cdot b_G$

(D') $r_G = (a_G - r \cdot b_G) / b$

where a_G = gradient of the function a and b_G = gradient of the function b.

From the formulas A, S, M, D and A', S', M', D' we conclude

> For any $x \in D$, the pair $r(x), r_G(x)$ can be computed from the pairs $a(x), a_G(x)$ and $b(x), b_G(x)$.

Notice that for a given $x \in D$, the pair $r(x), r_G(x)$ is not a pair of formulas, nor is it a pair of functions, it is an element of $\mathbb{R} \times \mathbb{R}^n$.

We differentiate the function r_G: $D \subseteq \mathbb{R}^n \to \mathbb{R}^n$. This yields

(A") $r_H = a_H + b_H$

(S") $r_H = a_H - b_H$

(M") $r_H = b \cdot a_H + b_G \cdot a_G^t + a \cdot b_H + a_G \cdot b_G^t$

(D") $r_H = (a_H - b_G \cdot r_G^t - r \cdot b_H - r_G \cdot b_G^t) / b$

where a_H = Hessian of the function a and b_H = Hessian of the function b.

Now we already know what to conclude

> For any $x \in D$, the triple $r(x), r_G(x), r_H(x)$ can be computed from the triples $a(x), a_G(x), a_H(x)$ and $b(x), b_G(x), b_H(x)$.

3. EXAMPLE

We consider the rational function

$$f: D \subseteq \mathbb{R}^3 \to \mathbb{R} \quad \text{with} \quad f(x) = x_1 - x_2 \cdot x_3 + x_1/(x_2 \cdot x_3 \cdot x_3).$$

First we split the formula for $f(x)$, then we split the parts, and so on. We obtain the tree shown in figure 1. Now we identify equivalent parts of the tree and get the graph shown in figure 2. This graph is a guide-line to compute $f(x)$.

code-list for $f(x)$

$f_1(x) = x_1$ = given
$f_2(x) = x_2$ = given
$f_3(x) = x_3$ = given
$f_4(x) = f_2(x) \cdot f_3(x)$
$f_5(x) = f_4(x) \cdot f_3(x)$
$f_6(x) = f_1(x) - f_4(x)$
$f_7(x) = f_1(x) / f_5(x)$
$f(x) = f_6(x) + f_7(x)$

For convenience we define

$$\overline{f}_i(x) = (f_i(x), f_{iG}(x), f_{iH}(x)) \quad \text{for } i = 1,2,\ldots,7$$
$$\overline{f}(x) = (f(x), f_G(x), f_H(x)) .$$

Now we know from section 2, that we can compute

$\overline{f}_4(x)$ from $\overline{f}_2(x)$ and $\overline{f}_3(x)$
$\overline{f}_5(x)$ from $\overline{f}_4(x)$ and $\overline{f}_3(x)$
$\overline{f}_6(x)$ from $\overline{f}_1(x)$ and $\overline{f}_4(x)$
$\overline{f}_7(x)$ from $\overline{f}_1(x)$ and $\overline{f}_5(x)$
$\overline{f}(x)$ from $\overline{f}_6(x)$ and $\overline{f}_7(x)$

This information allows to draw a graph to compute $\overline{f}(x)$, see figure 3. Notice that there is little difference between the graph to compute $f(x)$ and the graph to compute $\overline{f}(x)$.

The computational activities to get the value $f(x)$, the gradient $f_G(x)$ and the Hessian $f_H(x)$ for some $x \in D$ are obvious:

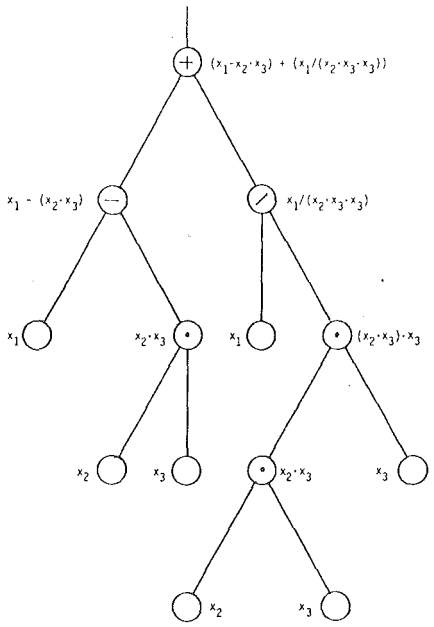


Figure 1: Tree for $f(x) = x_1 - x_2 \cdot x_3 + x_1/(x_2 \cdot x_3 \cdot x_3)$

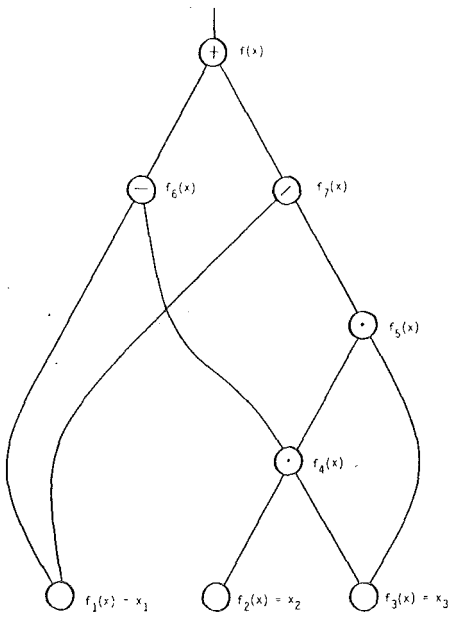


Figure 2: Graph to compute $f(x)$

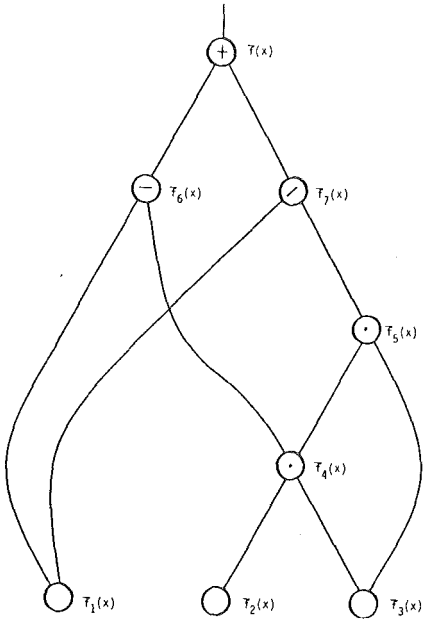


Figure 3: Graph to compute $\bar{f}(x) = (f(x), f_6(x), f_4(x))$

Set $\overline{f}_1(x)$, $\overline{f}_2(x)$, $\overline{f}_3(x)$ to their known values.

Then compute $\overline{f}_4(x)$, $\overline{f}_5(x)$, $\overline{f}_6(x)$, $\overline{f}_7(x)$, $\overline{f}(x)$ in this order.

The final triple $\overline{f}(x)$ provides the desired entities.

4. COMPLEXITY

Assume that $f: D \subseteq \mathbb{R}^n \to \mathbb{R}$ is an explicitely given rational function. This means, we can evaluate $f(x)$ using only addition, subtraction, multiplication and division of real numbers. Hence, we can set up a *characterizing sequence* $f_1, f_2, \ldots, f_s$ of functions $f_i: D \to \mathbb{R}$ such that

(1) for $i \in \{1,2,\ldots,n\}$

$f_i(x) = x_i$ = i-th component of x

(2) for $i \in \{n+1, n+2, \ldots, n+d\}$ with some $d \in \{0,1,\ldots\}$

$f_i(x) = c_i$ = real constant

(3) for $i \in \{n+d+1, n+d+2, \ldots, s\}$

$f_i(x) = a_i(x) * b_i(x)$ with $* \in \{+,-,\cdot,/\}$, and $a_i, b_i \in \{f_1, f_2, \ldots, f_{i-1}\}$

(4) $f_s(x) = f(x)$

In order to avoid superfluous operations, we assume that for $i \in \{n+d+1, \ldots, s-1\}$ the function f_i shows up as operand in at least one of the subsequent functions $f_{i+1}, \ldots, f_s$.

Prior to complexity investigations we have to specify the methods considered. Let us indicate gradient and Hessian matrix of a function by the subscript G resp. H.

Gradient-Method

(0) choose an $x \in D$

(1) for $i = 1, \ldots, n$ set $f_i(x), f_{iG}(x)$ to their values

(2) for $i = n+1, \ldots, n+d$ set $f_i(x), f_{iG}(x)$ to their values

(3) for $i = n+d+1, \ldots, s$ compute $f_i(x), f_{iG}(x)$ according to section 2

(4) then $f_s(x) = f(x)$, $f_{sG}(x) = f_G(x)$

Hessian-Method

(0) choose an $x \in D$

(1) for $i = 1,\ldots,n$ set $f_i(x), f_{iG}(x), f_{iH}(x)$ to their values

(2) for $i = n+1,\ldots,n+d$ set $f_i(x), f_{iG}(x), f_{iH}(x)$ to their values

(3) for $i = n+d+1,\ldots,s$ compute $f_i(x), f_{iG}(x), f_{iH}(x)$
according to section 2

(4) then $f_s(x) = f(x)$, $f_{sG}(x) = f_G(x)$, $f_{sH}(x) = f_H(x)$

What does it cost to compute the gradient $f_G(x)$ and the Hessian $f_H(x)$ of f at some $x \in D$? We answer this question in terms of arithmetic operations. Let us define

$\#(f)$:= number of arithmetic operations to compute $f(x)$ using the characterizing sequence $f_1, f_2, \ldots, f_s$

$\#(f, f_G)$:= number of arithmetic operations to compute $f(x)$ and $f_G(x)$

$\#(f, f_G, f_H)$:= number of arithmetic operations to compute $f(x)$, $f_G(x)$ and $f_H(x)$

Of course, $\#(f, f_G)$ and $\#(f, f_G, f_H)$ depend on the method used.

For the Gradient-Method we obtain

$$\#(f, f_G) \leq (3n + 1)\cdot\#(f) ,$$

and for the Hessian-Method we get

$$\#(f, f_G, f_H) \leq (\frac{7}{2}n^2 + \frac{13}{2}n + 1)\cdot\#(f) .$$

These bounds show that Automatic Differentiation is competitive if compared with numerical methods which approximate components of gradient and Hessian matrix by quotients of differences. Furthermore, it should be mentioned that, by some sophisticated organization, we are able to establish Automatic Differentiation methods with

$$\#(f, f_G) \leq 4\cdot\#(f) \quad \text{and} \quad \#(f, f_G, f_H) \leq (12n+8)\cdot\#(f) .$$

5. REMARKS

a) In section 2 we used a rational function r to point out the basic idea of Automatic Differentiation as far as gradient and Hessian matrix are concerned. But the formulas A', S', M', D' and A", S", M", D" mentioned there are not restricted to rational functions. The crucial point is the *rational composition* of r, rather than the rational character of the parts of r.

b) In section 4 we assumed that f is a *rational* function. This restriction was set only for didactic reasons. In the more general case where the formula for f(x) also involves some library functions like sin, cos, ..., the key is: Assume that the functions

$$a\colon D \subseteq \mathbb{R}^n \to \mathbb{R} \quad \text{and} \quad \lambda\colon E \subseteq \mathbb{R} \to \mathbb{R}$$

are twice differentiable. Under the provision $a(D) \subseteq E$ define the function

$$r\colon D \subseteq \mathbb{R}^n \to \mathbb{R} \quad \text{with} \quad r(x) := \lambda(a(x)) \ .$$

Then r is twice differentiable and

$$r_G(x) = \lambda'(a(x)) \cdot a_G(x)$$
$$r_H(x) = \lambda''(a(x)) \cdot a_G(x) \cdot a_G^t(x) + \lambda'(a(x)) \cdot a_H(x)$$

These formulas reveal the following fact:

> For any $x \in D$, the triple $r(x), r_G(x), r_H(x)$ can be computed from the triple $a(x), a_G(x), a_H(x)$ using λ, λ', λ''.

c) An efficient implementation of the Gradient- and the Hessian-Method has to take care of system-zeros (zeros in gradient and Hessian of primitive and constant functions).

REFERENCES

1. W. BAUR and V. STRASSEN: The complexity of partial derivatives. Theoretical Computer Science 22 (1983), 317-330.

2. H. KAGIWADA et al.: Numerical Derivatives and Nonlinear Analysis. Plenum Press, New York and London 1986.

3. L.B. RALL: Automatic Differentiation: Techniques and Applications. Lecture Notes in Computer Science No. 120, Springer-Verlag, Berlin-Heidelberg-New York 1981.

4. L.B. RALL: Differentiation in Pascal-SC: Type GRADIENT. ACM Trans. Math. Software 10 (1984), 161-184.

5. L.B. RALL: Global Optimization Using Automatic Differentiation And Interval Iteration. MRC Technical Summary Report #2832, University of Wisconsin, Madison 1985.

Numerical Methods and Approximation Theory III (Niš, August 18-21, 1987)

ON TWO SIDED APPROXIMATION FOR SOME SECOND ORDER VALUE BOUNDARY PROBLEMS*

P. GHELARDONI, G. GHERI and P. MARZULLI

ABSTRACT - This paper is concerned with boundary value problems for second order systems of the form $y'' = f(x,y)$. From a theorem proving under suitable conditions the existence of a solution, using Picard's iterations, a numerical procedure is derived to find actually two sided approximations of the solution. To this purpose a class of linear two-step methods is shown to be efficient, when two formulas of the class, with error constants of opposite sign, are alternatively used.
As a numerical application three test problems are developped.

1. INTRODUCTION.

Let the two point boundary value problem

(1) $y'' = f(x,y)$, $y(0) = \alpha$, $y(1) = \beta$,

be given with $f, y, \alpha, \beta \in R^m$, $x \in [0,1]$ of R^1 and let a function $y_1(x)$ be chosen so that $y_1(0) = \alpha$, $y_1(1) = \beta$; it is well known that under suitable hypotheses, Picard's iterations defined by

$$y_n'' = f(x,y_{n-1}), \quad y_n(0) = \alpha, \quad y_n(1) = \beta, \quad n>1,$$

can generate monotone sequences converging to a solution of (1). In a different way monotone sequences bounding the solution of

(1) Work supported by the Italian Ministero della Pubblica Istruzione.

(1) can be obtained by quasi-linearization ([8], ch. 5). Both those methods require to solve a linear problem at each step, bu the numerical solutions of these linear problems can bound the solution of (1) in the same fashion as the theoretical solution only if certain additional assumptions are verified ([2], p. p. 98-100 and corresponding references). Abiding by this frame, this paper we present a numerical method based on Picard's iter which give a theoretical two sided approximation to the solutio (1). The numerical method is suitable to solve the linear probl arising at each step, under assumptions sufficient to preserve property of a two sided approximation.

2. A PICARD'S SOLUTION OF THE PROBLEM.

Among several theorems assuring the existence of a solutio the problem (1), we are interested in the following formulation in [3], concerned with a more general problem than (1). Such formulation, which is given for a scalar equation, is a particu case of that reported in [1] (theor. 3.3., p. 34).

THEOREM 1. Given the boundary value problem

(1') $y'' = f(x,y,y')$, $y(0) = \alpha$, $y(1) = \beta$,

f, y, α, $\beta \in R$, let $f(x,y,y')$ be continous on

$$\{(x,y,y') \mid x \in [0,1],\ |y|<\infty,\ |y'|<\infty\}$$

and satisfy the Lipschitz conditions

$$|f(x,y,y')-f(x,y^*,y')| \leq L_1|y-y^*|,$$

$$|f(x,y,y')-f(x,y,y'^*)| \leq L_2|y'-y'^*|.$$

Then if

$$L_1 + 4L_2 < 8$$

there exists at least one solution of the problem (1').

The proof is given by constructing the sequence $\{F_n(x)\}$, $x \in [0,1]$,

$$F_1(x) = (\beta-\alpha)x + \alpha,$$
$$F_n''(x) = f(x,F_{n-1}(x),F_{n-1}'(x)), \quad n>1,$$
$$F_n(0) = \alpha, \quad F_n(1) = \beta;$$

this sequence is shown to converge uniformly on $[0,1]$ to a solution $y(x)$ of the problem (1').

This theorem can be extended with some slight modification to the problem (1).

We denote, for simplicity, $y_B(x) = (\beta-\alpha)x + \alpha$ the linear function satisfying the boundary conditions and

$$S = \{z(x) \mid z(x) \in C^0[0,1],\ z(0) = \alpha,\ z(1) = \beta\},$$

where the vector norm $\| \cdot \|$ is given by

$$\| z(x) \| = \sum_{i=1}^{m} \max_{0\leq x\leq 1} |z_i(x)|.$$

Then the following result holds.

THEOREM 2. Consider the problem (1) where $y(x) \in S$. Assuming that

(2) $\quad f_i(x,y) \in C^0([0,1]\times S),\ i = 1,2,\ldots,m,$

and, for any $y, y^* \in S$,

(3) $\quad |f_i(x,y)-f_i(x,y^*)| \leq L_i \sum_{j=1}^{m} |y_j-y_j^*|,$

(4) $\quad L_i < 8/m,\ i = 1,2,\ldots,m,$

then the Picard's iterations of S

$$
\begin{aligned}
&F_1(x) = y_B(x),\\
(5)\qquad &F_n''(x) = f(x,F_{n-1}(x)),\ n>1,\\
&F_n(0) = \alpha,\ F_n(1) = \beta,
\end{aligned}
$$

converge uniformly on $[\,0,1\,]$ to a solution of (1).

PROOF. The main feature of the proof is to prove that the sum $F_1(x) + (F_2(x)-F_1(x)) + \ldots + (F_n(x)-F_{n-1}(x))$ for $n\to\infty$ converges uniformly to a function which is shown to be a solution of (1). Since $F_n(x)$ is a solution of (5) we can write ([2] p. 42-43, [7])

$$
(5')\qquad F_n(x) = y_B(x) + \int_0^1 g(x,t)f(t,F_{n-1}(t))\,dt\ ,
$$

where

$$
g(x,t) = \begin{cases} t(x-1) & \text{for } t\le x\\ x(t-1) & \text{for } t>x \end{cases}
$$

is the Green's function. Thus we have

$$
|F_{2i}(x)-F_{1i}(x)| \le \int_0^1 |g(x,t)|\,|f_i(t,y_B(t))|\,dt\ ,\ i = 1,2,\ldots
$$

Now we observe that, for $i= 1,2,\ldots,m$, it results $\alpha_i\le y_{Bi}(x)\le\beta_i$ or $\alpha_i\ge y_{Bi}(x)\ge\beta_i$ on $[\,0,1\,]$, so that the functions $f_i(x,y_B)$ are defined on a closed and bounded domain D. Thus, from (2), we can set $l_i = \max_D |f_i(x,y_B)|$ and it results, for any $x\in[\,0,1\,]$,

$$
|F_{2i}(x)-F_{1i}(x)| \le \frac{1}{8}l_i,\ i = 1,2,\ldots,m.
$$

Then it follows $\|F_2-F_1\| \le l/8$, where $l = \sum_{i=1}^{m} l_i$.
Analogously, by (3), we obtain

$$
|F_{ni}(x)-F_{n-1,i}(x)| \le \frac{1}{8}L_i\,\|F_{n-1}-F_{n-2}\|\ ,\ i = 1,2,\ldots,m
$$

for any $x\in[\,0,1\,]$. Adding all these inequalities and denoting

$L = \sum_{i=1}^{m} L_i$, we have $\| F_n - F_{n-1} \| \leq \frac{1}{8} L \| F_{n-1} - F_{n-2} \|$, or, by recurrence

$$\| F_n - F_{n-1} \| \leq \frac{1}{8} 1 (\frac{L}{8})^{n-2}.$$

Owing to (4) the uniform convergence of $\{F_n(x)\}$ on $[\,0,1\,]$ is proved.

Let $F(x) = \lim_{n\to\infty} F_n(x)$ and

$$s(x) = F(x) - \int_0^1 g(x,t) f(t,F(t))\, dt - y_B(x).$$

To prove that $F(x)$ is a solution of (1) it is sufficient to verify that $s(x) = 0$ identically.

In fact from (5') we have

$$s(x) = F(x) - F_n(x) - \int_0^1 g(x,t)(f(t,F(t)) - f(t,F_{n-1}(t)))\, dt,$$

and from (3) we can write, for any $x \in [\,0,1\,]$,

$$|f_i(x,F(x)) - f_i(x,F_n(x))| \leq L_i \| F - F_n \|, \quad i = 1,2,\ldots,m.$$

Then, because of $\lim_{n\to\infty} \| F - F_n \| = 0$, it can be seen that for any arbitrary positive δ we can choose n so large that $|s_i(x)| < \delta$, $i = 1,2,\ldots,m$, for any $x \in [\,0,1\,]$. Then $s(x) = 0$ identically, which completes the proof.

We introduce now a partial ordering in R^m defining that $v \geq w$ means $v_i \geq w_i$ for $i = 1,2,\ldots,m$; so we can prove the following theorem giving sufficient conditions for the two sided approximat: to the solution $y(x)$ of (1) by means of the sequence $\{F_n(x)\}$.

THEOREM 3. Let the problem (1) satisfy the hypotheses of the theorem 2 and let the inequalities

(6) $f(x,y(x)) \geq 0$,

(7) $J(x,y(x)) = \partial f/\partial y \geq 0$,

hold in $[0,1] \times S$; then the Picard's sequence $\{F_n(x)\}$, as defined in (5) and converging to $y(x)$, satisfies, for $n \geq 1$, the inequaliti

(8) $F_{2n}(x) \leq y(x) \leq F_{2n-1}(x)$.

This statement is also true if inequalities (6), (7) and (8) are reversed.

PROOF. Since the limit vector $F(x) = y(x)$ is a solution of (1) we have

(9) $$y(x) = y_B(x) + \int_0^1 g(x,t) f(t,y(t))\, dt .$$

As previously observed, we can write

(10) $$F_{n+1}(x) = y_B(x) + \int_0^1 g(x,t) f(t,F_n(t))\, dt .$$

Writing (9) in the equivalent form

$$y(x) - y_B(x) = (x-1)\int_0^x tf(t,y(t))\, dt + x\int_x^1 (t-1)f(t,y(t))\, d$$

we have

(11) $y(x) \leq y_B(x) = F_1(x)$.

Subtracting (10) from (9) and using the mean value theorem we obtain

(12) $$y(x) - F_{n+1}(x) = (x-1)\int_0^x t\hat{J}(y(t)-F_n(t))\, dt + x\int_x^1 (t-1)\hat{\hat{J}}(y(t)-F_n(t))\, dt$$

where $\hat{J}$ and $\hat{\hat{J}}$ are two suitable evaluations of the jacobian matri

From (7) and (12) it follows that

$$y(x) \gtreqless F_n(x) \quad \text{implies} \quad y(x) \lesseqgtr F_{n+1}(x) \ ,$$

and taking into account (11), the inequalities (8) are proved.

3. METHODS FOR TWO SIDED APPROXIMATION.

Consider the well known class of linear q-step methods of the form

$$\sum_{i=0}^{q} \alpha_i y_{k+i-1} = h^2 \sum_{i=0}^{q} \gamma_i f_{k+i-1}, \quad q \geq 2$$

(see [6] p. 27-28 and [5] p. 252-256) and limit our attention to the family of methods

$$(13) \qquad y_{k-1} - 2y_k + y_{k+1} = h^2(\gamma_0 f_{k-1} + \gamma_1 f_k + \gamma_2 f_{k+1}).$$

Moreover we consider for γ_0, γ_1, γ_2 only non negative values guaranteeing among other things to avoid operations where exact significant figures may be lost.

It is easy to verify that: if $\gamma_0 + \gamma_1 + \gamma_2 = 1$, from (13) we have a class of first order formulas at least; if in addition we impose $\gamma_1 + 2\gamma_2 = 1$ we have a class of second order formulas; by adding the further condition $\gamma_1 + 4\gamma_2 = 7/6$ we obtain a unique fourth order formula; finally in the family of formulas (13) with non negative γ_i there are two sub-classes of formulas having error constants of opposite sign.

Denoting M_1 and M_2 a couple of formulas with non negative γ_i and error constants of opposite sign, we can set respectively for the truncation errors

$$\tau^{(1)} = C_1 h^{p_1 - 2} y^{(p_1)}(\zeta),$$

$$\tau^{(2)} = -C_2 h^{p_2 - 2} y^{(p_2)}(\eta),$$

where p_1, $p_2 \geq 3$, ζ and η are suitable values on $(0,1)$ even depending on k, C_1 and C_2 are positive constants.

We want to apply M_1 and M_2 to the problem (5). Discretizing the interval $[0,1]$ with the mesh points $x_k = kh$, $k = 0,1,\ldots,K+1$, $h = 1/(K+1)$, and using M_1 or M_2 to approximate the solution of t problem (5), we have to solve at each step an algebraic linear s In fact, denoting by $g_k^{(n)}$ the m-vector giving the approximate di solution at the current step and at the point x_k, supposing to u an exact arithmetic, from (13) we have

$$g_{k-1}^{(n)} - 2g_k^{(n)} + g_{k+1}^{(n)} = h^2(\gamma_0 f(x_{k-1}, g_{k-1}^{(n-1)}) + \gamma_1 f(x_k, g_k^{(n-1)}) + \gamma_2 f(x_{k+1}, g_{k+1}^{(n-1)})).$$

Collecting all these equations for k varying from 1 to K we have

$$Hg_n = c^{(i)}(g_{n-1}) \quad , \quad n = 2,3,\ldots,$$

where g_n is the mK-vector whose components are $(g_k^{(n)})_j$ $((j=1,2,.$ $k=1,2,\ldots,K)$; (i) means we have used the method M_i $(i=1,2)$; $c^{(i)}$ is a non decreasing mK-vector function of its argument $z \in R^{mK}$ a consequence of the condition (7) and of the non-negativity of t selected γ_0, γ_1, γ_2 in (13); H is an mK-order T-matrix ([9], c of the tridiagonal block form

$$H = \begin{bmatrix} -2I & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & -2I \end{bmatrix}$$

where I is the m-order identity matrix. Moreover we observe that is a M-matrix ([9], p. 42-45) and consequently a monotone matri for which $-Hu \leq -Hv$ implies $u \leq v$ or $Hu \geq Hv$ implies $u \leq v$ as we being u, v, real mK-vectors.

Now let $Y_k^{(n)}$ be the m-vector formed with the exact values of the solution $F_n(x)$ of the current problem (5) at the mesh-point x_k and let $G_k^{(n)}$ be the m-vector approximating $Y_k^{(n)}$ and formed with the corresponding values obtained using M_1 or M_2 in presence of round-of errors.

If $\rho_k^{(n)}$ and $\tau_k^{(n)}$ denote the m-vectors giving respectively the local round-off error and the local truncation error, we have, by definiti for each x_k, $k = 1,2,\ldots,K$,

$$(14)\quad G_{k-1}^{(n)} - 2G_k^{(n)} + G_{k+1}^{(n)} = h^2(\gamma_0 f(x_{k-1},G_{k-1}^{(n-1)}) + \gamma_1 f(x_k,G_k^{(n-1)}) + \gamma_2 f(x_{k+1},G_{k+1}^{(n-1)})) + \rho_k^{(n)}$$

$$(14')\quad Y_{k-1}^{(n)} - 2Y_k^{(n)} + Y_{k+1}^{(n)} = h^2(\gamma_0 f(x_{k-1},Y_{k-1}^{(n-1)}) + \gamma_1 f(x_k,Y_k^{(n-1)}) + \gamma_2 f(x_{k+1},Y_{k+1}^{(n-1)})) + h^2\tau_k$$

Subtracting (14') from (14) and setting $e_k^{(n)} = G_k^{(n)} - Y_k^{(n)}$, we obta

$$e_{k-1}^{(n)} - 2e_k^{(n)} + e_{k+1}^{(n)} = h^2(\gamma_0 J_{k-1} e_{k-1}^{(n-1)} + \gamma_1 J_k e_k^{(n-1)} + \gamma_2 J_{k+1} e_{k+1}^{(n-} + \rho_k^{(n)} - h^2\tau_k^{(n)},$$

where J_{k-1}, J_k, J_{k+1} are suitable determinations of the jacobian matrix of the function f.

Defining now the mK-vectors G_n, Y_n, e_n, ρ_n, τ_n having the componen respectively given by $(G_k^{(n)})_j$, $(Y_k^{(n)})_j$, $(e_k^{(n)})_j$, $(\rho_k^{(n)})_j$, $(\tau_k^{(n)})_j$ $((j=1,2,\ldots,m)$, $k=1,2,\ldots,K)$, the last equation can be written as

$$(15)\quad -He_n = -Qe_{n-1} + h^2\tau_n - \rho_n$$

where Q is a mK-order non-negative matrix. Note that $e_1 = 0$ becaus we assume $G_1 = Y_1$.

Denoting with $|v|$ the vector whose components are the absolute val of the corresponding components of v, the following theorem holds.

THEOREM 4. Let the solution of the problem (5) have both p_1-th and p_2-th derivatives ≥ 0 and let the methods M_1 and M_2 be applied to (5) with

$$|\rho_n| < h^2|\tau_n|. \tag{16}$$

Then the method M_1 gives $G_n > Y_n$ if $G_{n-1} \leq Y_{n-1}$ and the method M_2 gives $G_n < Y_n$ if $G_{n-1} \geq Y_{n-1}$. This statement is also true if all inequalities, but (16), are reversed.

PROOF. Since $-H$ is an M-matrix, $(-H)^{-1}$ is a non-negative matri so that from (15) and (16) and taking into account the present s of τ_n the proof is obtained.

Theorems 3 and 4 enable to carry out a procedure for a two sided approximation of the solution of a problem of the type (1) satisf the condition (2), (3), (4), (6), (7) of the section 2 and having solution $y(x)$ with non-negative p_1-th and p_2-th derivatives (the with reversed inequalities is analogously obtained).

Denoting with Y the mK-vector whose components are the exact val of the solution $y(x)$ of the problem (1) at the mesh-points x_1, ..,x_K, the procedure can be described as follows.

From (8) we have

$$Y_{2n} \leq Y \leq Y_{2n-1}, \quad n \geq 1;$$

on the other hand, according to theorem 4, starting with $G_1 = Y_1$ applying the method M_2 to the problem (5), with $n = 2$, we obtain $G_2 < Y_2$ and this implies that the further application of the met M_1 with $n = 3$ will give $G_3 > Y_3$. The entire process can be repea using alternatively M_2 and M_1 to obtain, in general

$$G_{2n} < Y_{2n} \leq Y \leq Y_{2n-1} < G_{2n-1}.$$

We observe that this result can be also obtained using the monot of $-H$ and because $c^{(i)}(z)$ is a not decreasing function.

Because of the convergence of the sequence $\{Y_n\}$ to Y, it can be expected the couples G_{2n}, G_{2n-1} give, for suitable n, good two sided approximations to Y, according to the accuracy of the methods M_1 and M_2.

4. NUMERICAL TESTS.

We have considered the following formulas of the kind (13) and the corresponding truncation errors:

$$(M_1) \qquad \begin{aligned} & y_{k-1} - 2y_k + y_{k+1} = h^2 f_k, \\ & \tau^{(1)} = \frac{1}{12} h^2 y^{(4)}(\zeta); \end{aligned}$$

$$(M_2) \qquad \begin{aligned} & y_{k-1} - 2y_k + y_{k+1} = h^2 \left(\frac{1}{2} f_{k-1} + \frac{1}{2} f_{k+1}\right), \\ & \tau^{(2)} = -\frac{1}{24} h^2 y^{(4)}(\theta), \end{aligned}$$

$$(M_{2'}) \qquad \begin{aligned} & y_{k-1} - 2y_k + y_{k+1} = h^2 \left(\frac{1}{12} f_{k-1} + \frac{5}{6} f_k + \frac{1}{12} f_{k+1}\right), \\ & \tau^{(2')} = -\frac{1}{240} h^4 y^{(6)}(\eta). \end{aligned}$$

As the results are concerned with three test problems, we have quoted in the tables 1, 2, 3, the m-vectors $e_k^{(n)} = e^{(n)}(x_k)$ with n=8 and n=9.

Coupling the preceding formulas in the two fashions (M_1, M_2) and $(M_1, M_{2'})$, an application has been made, in the test 1 and 2, to a class of linear problems like (1) considered in [4]:

$$(17) \qquad y'' = A(x)y(x),$$

where the matrix $A(x) \equiv A \in R^{m\times m}$ is given by

$$A = W''W + (W')^2 + 2W'DW + WDW' + (W'W)^2 + WDWW'W + WD^2W,$$

$W=W(x) \in R^{m\times m}$ is such that $W^2=I$, and $D=\mathrm{diag}(\lambda_i)$ is a diagonal

matrix of order m with λ_i, i = 1,2,...,m, real parameters not depending on x. For this problem we have the exact solution $y(x) = W(x)z(x)$, where

$$z(x) = (e^{\lambda_1 x}, e^{\lambda_2 x}, \ldots\ldots, e^{\lambda_m x})^T.$$

Test 1. We have selected for W the constant binomial matrix of order m

$$W = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \ldots\ldots \\ 1 & -1 & 0 & 0 & 0 & \ldots\ldots \\ 1 & -2 & 1 & 0 & 0 & \ldots\ldots \\ 1 & -3 & 3 & -1 & 0 & \ldots\ldots \\ 1 & -4 & 6 & -4 & 1 & \ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots & & & & & \end{bmatrix},$$

and chosen $\lambda_i = a^{ib/2}$, i = 1,2,...,m, with a>0 and b real numbers Then we have $A = WD^2W$ whose elements are given by

$$(A)_{ij} = \binom{i-1}{j-1} a^{jb}(1-a^b)^{i-j}, \quad 1 \leq j \leq i \leq m.$$

Choosing a and b so that $0 \leq a^b \leq 1$ it is not difficult to verify that the following properties hold:

i) $A \geq 0$;

ii) $\|A\|_\infty = a^b$, so that the condition (4) of the theorem (2) is equivalent to $m \leq 8/a^b$;

iii) $y^{(p)}(x) \geq 0$ if $m \leq 1+1/a^{\frac{1}{2}pb}$.

Then the test problem is

$$y'' = Ay,$$

$$y(0) = \alpha = (1, 0, \ldots, 0)^T,$$

$$y_i(1) = \beta_i = \sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^k e^{\lambda_k}, \quad i = 1,2,\ldots,m.$$

ssuming m=4, b=2, a=0.5, the conditions of the theorems 2 and 3 re satisfied and we have $y^{(4)}(x)\geq 0$, $y^{(6)}(x)\geq 0$, $x \in [0,1]$, so nat the theorem 4, related to the couples (M_1,M_2) and $(M_1,M_2,)$ olds. In the table 1 we have listed the values of $e^{(8)}(x)$ and $^{(9)}(x)$ corresponding to the odd mesh-points with a step-lenght =0.1.

Table 1

	(M_1,M_2)			$(M_1,M_2,)$	
	$e^{(8)}(x)$	$e^{(9)}(x)$	$(\times 10^{-5})$	$e^{(8)}(x)$	$e^{(9)}(x)$
1	-0.63	0.69		-0.13	0.17
	-0.58	0.67		-0.11	0.16
	-0.57	0.66		-0.11	0.15
	-0.55	0.63		-0.10	0.14
.3	-1.04	1.17		-0.21	0.24
	-0.96	1.09		-0.20	0.23
	-0.87	0.99		-0.19	0.21
	-0.81	0.92		-0.18	0.20
.5	-1.22	1.45		-0.32	0.35
	-1.18	1.33		-0.30	0.33
	-1.05	1.20		-0.26	0.31
	-0.98	1.11		-0.24	0.27
.7	-1.18	1.35		-0.30	0.35
	-1.11	1.27		-0.27	0.33
	-1.03	1.17		-0.26	0.30
	-0.98	1.11		-0.24	0.27
.9	-0.74	0.84		-0.15	0.20
	-0.96	0.79		-0.13	0.19
	-0.66	0.75		-0.12	0.17
	-0.63	0.72		-0.11	0.15

est 2. We consider again the problem (17) with m=2 and

$$W = \begin{bmatrix} a_{11}x+b_{11} & a_{12}x+b_{12} \\ a_{21}x+b_{21} & a_{22}x+b_{22} \end{bmatrix}$$

where a_{ij}, b_{ij}, $1\leq i,j\leq 2$ are real numbers chosen according to the condition $W^2 = I$.

It is easy to verify that $(W')^2 = 0$, $W'W=-WW'$ and, obviously, $W''= 0$; so we have $A = 2W'DW+WD^2W$.

Furthermore the class of the matrices like W for which $W'\neq 0$ is defined by the following conditions

$$a_{11}^2 + a_{12}a_{21} = 0 ,$$
$$a_{12}b_{21} + b_{12}a_{21} + 2a_{11}b_{11} = 0 ,$$
$$b_{11}^2 + b_{12}b_{21} = \pm 1 .$$

Selecting, for example, $a_{11}= a_{12}= b_{12}= a_{22}= 0$, $b_{11}= 1$ and $b_{21}= b_{22} = -1$, we obtain the test problem

$$y'' = (2W'DW+WD^2W)y ,$$
$$y(0) = (1,-2)^T, \quad y(1) = (e^{\lambda_1},-e^{\lambda_2})^T .$$

If $0<\lambda_2\leqq\lambda_1\leqq 2$ and $0\leqq\lambda_1-\lambda_2\leqq 1$, the conditions of the theorems 2 and 3 hold, and $y^{(s)}(x)\geqq 0$ for $s\geqq 2$ so that the theorem 4 is applicable as well.

In the table 2 are displayed the values of $e^{(8)}(x)$ and $e^{(9)}(x)$ relatively to the odd mesh points with step-lenght h=0.1 and $\lambda_1=2$, $\lambda_2=1$.

Table 2

	(M_1,M_2)			$(M_1,M_2,)$	
x	$e^{(8)}(x)$	$e^{(9)}(x)$	$(\times 10^{-4})$	$e^{(8)}(x)$	$e^{(9)}(x)$
0.1	-0.74	1.05		-0.39	0.59
	-0.96	1.03		-0.38	0.58
0.3	-1.56	2.21		-0.82	1.24
	-1.07	1.51		-0.56	0.87
0.5	-3.31	4.80		-1.80	2.73
	-1.57	2.20		-0.80	1.22
0.7	-7.76	10.92		-4.15	6.28
	-2.15	3.03		-1.19	1.70
0.9	-6.49	9.10		-3.37	5.12
	-1.40	1.94		-0.74	1.13

Test 3. As a case different from the type (17), we consider the nonlinear problem

$$y_i''(x) = e^{y_{m-i+1}} , \quad i = 1,2,\ldots,m.$$

n this equation the conditions (6) and (7) are verified, and $^{(4)}(x)>0$.

urthermore, choosing suitably the boundary values of y(x) such hat $0 \geq \alpha > \beta$, we find $y(x) \leq 0$ and $y'(x) \leq 0$ on $[0,1]$.
hen even the conditions $y^{(6)}(x)>0$ and $\|\partial f/\partial y\|_\infty \leq 1$ hold: thus it s possible to apply the two sided approximation process for ≤ 8.

n the table 3 results for $e^{(8)}(x)$ and $e^{(9)}(x)$ at some odd mesh oints with h=0.1 are displayed for a problem with m=4 and boundary onditions given by

$$y(0) = 0,$$
$$y_1(1) = -0.4, \qquad y_3(1) = -0.8,$$
$$y_2(1) = -0.6, \qquad y_4(1) = -1.0.$$

Table 3

	(M_1,M_2)			$(M_1,M_2,)$	
x	$e^{(8)}(x)$	$e^{(9)}(x)$	$(\times 10^{-4})$	$e^{(8)}(x)$	$e^{(9)}(x)$
0.1	-2.49	3.41		-0.35	0.51
	-1.85	2.57		-0.28	0.41
	-1.25	1.71		-0.21	0.30
	-0.83	1.13		-0.12	0.18
0.3	-4.44	6.05		-0.76	1.13
	-3.66	4.98		-0.56	0.82
	-2.58	3.50		-0.30	0.45
	-1.72	2.33		-0.24	0.34
0.5	-4.39	6.00		-0.68	1.00
	-3.57	4.91		-0.56	0.81
	-2.92	3.99		-0.37	0.55
	-1.78	2.44		-0.27	0.39
0.7	-3.20	4.35		-0.49	0.71
	-2.66	3.61		-0.37	0.54
	-1.93	2.62		-0.29	0.43
	-1.28	1.75		-0.20	0.29
0.9	-1.16	1.61		-0.17	0.25
	-0.97	1.33		-0.14	0.21
	-0.80	1.10		-0.09	0.16
	-0.52	0.70		-0.07	0.12

All calculations have been performed on the IBM 370 at the CNUCE of Pisa using a double precision arithmetic (8bytes).

REFERENCES

[1] P.B. BAILEY - L.F. SHAMPINE - P.E. WALTMAN, Nonlinear two point boundary value problems, Academic Press, New York, 1968.

[2] J.W. DANIEL - R.E. MOORE, Computation and theory in ordinary differential equations, W.H. Freeman and Company, San Francisco, 1970.

[3] G. GHELARDONI, Sul problema di valori al contorno per l'equazione differenziale y''=f(x,y,y'), Accademia Nazionale dei Lincei, Serie VIII, vol. XXXIII, fasc. 5, 1962, p. 237-243.

[4] G. GHERI - P. MARZULLI, Collocation for initial value problems based on Hermite interpolation, CALCOLO, vol. XXIII, 1986, p. 115-130.

[5] J.D. LAMBERT, Computational methods in ordinary differen equations, Jhon Wiley, London, 1972.

[6] L. LAPIDUS - J.H. SEINFELD, Numerical solution of ordina differential equations, Academic Press, New York, 197

[7] G.F. ROACH, Green's functions, Cambridge University Pres Cambridge, 1982.

[8] S.M. ROBERTS - J.S. SHIPMAN, Two-point boundary value problems: shooting methods, American Elsevier, New Yor 1972.

[9] D.M. YOUNG, Iterative solution of large linear systems, Academic Press, New York, 1971.

ON THE APPROXIMATE CALCULATION OF INTEGRALS ON A POLYGON IN R^2

ALLAL GUESSAB

Abstract : We will consider the problem of approximating a double integral on a polygon in $\mathbb{R}^2$ as a linear combination of integrals on the real line. Cubature formulas are obtained in such a way as to minimize the exact error bounds of the formulas for a given class of functions.

1. INTRODUCTION : NOTATIONS AND DEFINITIONS

The theory of numeric cubature formulas for functions of one variable is well developed. We refer to Davis-Rabinowitz [1], Stroud-Secrest [23] and Krylov [11] .

In this work, we will consider the problem of approximating a double integral on a polygon K in $\mathbb{R}^2$ as a linear combination of integrals on the real lir

Let us first fix a few notations. For $\alpha = (\alpha_1,\alpha_2)$ in $\mathbb{N}^2$, we denote by X^α the monomial defined by :

$$X^\alpha = x^{\alpha_1} y^{\alpha_2} .$$

For $k = (k_1,k_2)$ in $\mathbb{N}^*.\mathbb{N}^*$, and $m = (m_1,m_2)$ in $\mathbb{N}^*.\mathbb{N}^*$ (such that $k_1 < m_1$ and $k_2 < m_2$) . We have the following definitions :

$$(1.1) \qquad L_k = \left\{ \alpha = (\alpha_1,\alpha_2) \in \mathbb{N}^2 , \quad \alpha_i \le k_i , \quad i = 1,2 \right\} ,$$

$$(1.2) \qquad R_k = \left\{ R \equiv \sum_{\alpha \in L_k} a_\alpha X^\alpha , \quad a_\alpha \in \mathbb{R} \right\} ,$$

$$(1.3) \qquad V_k = \left\{ R \equiv X^k + \sum_{\alpha \in L_{(k_1-1,k_2-1)}} a_\alpha X^\alpha , \ a_\alpha \in \mathbb{R} \right\} .$$

Finally, let c be a real number, which is assumed to be fixed. Let us introduce the following notation :

$M^{k}_{0,K}(c)$ is the set of all functions $f(x,y)$ which have piecewise continuous derivatives

$$\frac{\partial^{i+j}}{\partial x^{i}\partial y^{j}} f(x,y) \ , \quad i = 0,1,\ldots,k_1 \ ; \quad j = 0,1,\ldots,k_2 \tag{1.4}$$

on K and satisfy the conditions

$$\left\| \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f \right\|_{0,K} \leq c \tag{1.5}$$

where :

$$\left\| \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f \right\|_{0,K} = \left(\int_K \left(\frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f(x,y) \right)^2 dxdy \right)^{1/2} \tag{1.6}$$

Let K be a polygon in $\mathbb{R}^2$ and g_k an arbitrary polynomial in V_k. Consequently the following equality is true :

$$I_K(f) = \int_K f(x,y)dxdy = \frac{1}{k_1!k_2!} \int_K f(x,y) \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} g_k(x,y)dxdy \ .$$

Then if $f \in M^{k}_{0,K}(c)$, applying Green's formula to the right-hand side, we obtain the following cubature formula :

$$I_K(f) = Q_K(f,g_k) + E_K(f,g_k) \ , \tag{1.7}$$

where

$$Q_K(f,g_k) = \frac{1}{k_1!k_2!} \left(\sum_{j=0}^{k_1-1} (-1)^j \int_\Gamma \frac{\partial^j}{\partial x^j} f(x,y) \frac{\partial^{k_1-1-j+k_2}}{\partial x^{k_1-1-j}\partial y^{k_2}} g_k(x,y)\vartheta_1 \right.$$
$$\left. + \sum_{j=0}^{k_2-1} (-1)^{k_1+j} \int_\Gamma \frac{\partial^{k_1+j}}{\partial x^{k_1}\partial y^{j}} f(x,y) \frac{\partial^{k_2-1-j}}{\partial y^{k_2-1-j}} g_k(x,y) \ \vartheta_2 d\sigma \right) ,$$

where ϑ_i , is the i-th component of the outer normal vector along K , and

$$E_K(f,g_k) = \frac{(-1)^{k_1+k_2}}{k_1!k_2!} \int_K g_k(x,y) \frac{\partial^{k_1+k_2}}{\partial x^{k_1}\partial y^{k_2}} f(x,y)dxdy \ . \tag{1.8}$$

APPROACH TO THE PROBLEM

Our goal in this work is to derive optimal cubature formulas of the type (1.7) in the space $M^k_{0,K}(c)$. The formula (1.7) will be optimal in the space $M^k_{0,K}(c)$, if the polynomial g_k is chosen so that the quantity

$$E(M^k_{0,K}(c)) = \sup_{f\in M^k_{0,K}(c)} |E(f,g_k)| .$$

has the minimal value.

. PRELIMINARY RESULTS

Remark 2.1. It is clear that the cubature formula (1.7) is exact on F, i.e.

$$E_K(f,g_k) = 0 , \quad \text{for } f \text{ in } F ,$$

where F is the vector space generated by the family of monomials X^α, $\alpha = (\alpha_1,\alpha_2) \in A$, with

$$A = (\{0,1,2,\ldots,k_1-1\} \cdot \mathbb{N}) \cup (\mathbb{N} \cdot \{0,1,2,\ldots,k_2-1\})$$

Remark 2.2. Assuming $f \in M^k_{0,K}(c)$. By Hölder's inequality, we obtain from (1.8) that :

$$(2.1) \qquad \sup_{f\in M^k_{0,K}(c)} |E_K(f,g_k)| \le \frac{c}{k_1!k_2!} \|g_k\|_{0,K}$$

Proposition 2.1. Assume $Q_K(f,g_k)$ is a formula of type (1.7). Then :

$$(2.2) \qquad \sup_{f\in M^k_{0,K}(c)} |E_K(f,g_k)| = \frac{c}{k_1!k_2!} \|g_k\|_{0,K}$$

Proof : For the function

$$R(x,y) = \frac{(-1)^{k_1+k_2}}{(k_1-1)!(k_2-1)!} \frac{c}{\|g_k\|_{0,K}} \int_0^x \int_0^y (x-u)^{k_1-1} (y-v)^{k_2-1} g_k(u,v)$$

belonging to $M^k_{0,K}(c)$, it follows from (1.8) that

$$|E_K(R,g_k)| = \frac{c}{k_1!k_2!} \|g_k\|_{0,K} .$$

Then we have from (2.1) the equality (2.2).

In the sequel, we will use the following terminology :
K is a polygon in $\mathbb{R}^2$. We establish a triangulation $\mathcal{C}_h$ ((cf. Fig.2.1)) over K, i.e. K is expressed as a finite union

$$K = \bigcup_{i=1}^{n} K_{h,i} \quad , \tag{2.3}$$

of triangles $K_{h,i}$ in such a way that these triangles are non-overlapping, and are all interior to K .

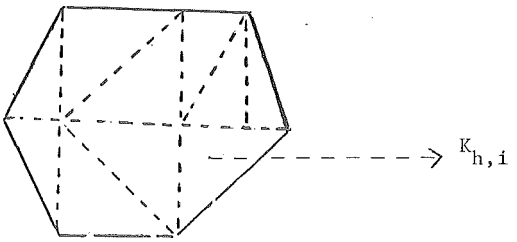


Fig.2.1 Subdivision of K into triangles.

APPROACH TO THE PROBLEM IF K IS A TRIANGLE.

Let K be a triangle with vertices (a,b), $(a+h,b)$ and $(a,b+h')$, where $h, h' \in \mathbb{R}_+^*$. We denote by Γ_i , $i = 1,2,3$ the three sides of K , and

$$\Gamma = \bigcup_{i=1}^{3} \Gamma_i \quad , \tag{3.1}$$

where Γ_1 (resp. Γ_2) is a piece of a line parallel to the y-axis (resp. x-axis) :

$$\Gamma_1 = \{(x,y) \in K \ , \ d_1(x,y) = x - a = 0 \}$$

$$\Gamma_2 = \{(x,y) \in K \ , \ d_2(x,y) = y - b = 0 \} \quad .$$

Definition 3.1. type (1.7) cubature formula $Q_K(f,g_k)$ is said to be optimal with respect to K of order $\tilde{r} = (r_1,r_2) \in \mathbb{N}^* \mathbb{N}^*$ such that $r_1-1 \le k_1$ and $r_2-1 \le k_2$, if and only if the following properties hold :

i) g_k lies in $V_k^{(1)}$, where $V_k^{(1)}$ is the set of polynomials p_k in V_k such that :

$$p_k = p^*_{k-\tilde{r}} d_1^{r_1} d_2^{r_2} \quad , \quad p^*_{k-\tilde{r}} \in V_{k-\tilde{r}} \ .$$

ii) $$E_K^*(f,g_k) = \sup_{f \in M^k_{0,K}(c)} |E_K(f,g_k)| = \inf_{p_k \in V_k^{(1)}} \sup_{f \in M^k_{0,K}(c)} |E_K(f,p_k)| \ .$$

Theorem 3.1. Let K be a triangle of the type (3.1). Then, there exists an optimal cubature formula of the type (1.7) with respect to K, of order $\tilde{r} = (r_1,r_2)$, such that :

$$g_k = g^*_{k-\tilde{r}} d_1^{r_1} d_2^{r_2} \quad ,$$

where $g^*_{k-\tilde{r}}$ is in $V_{k-\tilde{r}}$, and is orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(K)$, when considering the inner product associated to integration on K and weight function $d_1^{2r_1} d_2^{2r_2}$.

Proof : From proposition 2.1, we have :

$$\sup_{f \in M^k_{0,K}(c)} |E_K(f,p_k)| = \frac{c}{k_1!k_2!} \|p_k\|_{0,K} \quad , \text{ for all } p_k \in V_k \ .$$

Then

$$E_K^*(f,g_k) = \inf_{p_k \in V_k^{(1)}} \sup_{f \in M^k_{0,K}(c)} |E_K(f,p_k)| = \frac{c}{k_1!k_2!} \inf_{p_k \in V_k^{(1)}} \|p_k\|_{0,K} \ .$$

So :

$$E_K^*(f,g_k) = \frac{c}{k_1!k_2!} \inf_{p \in V_{k-\tilde{r}}} \left[I_K (d_1^{2r_1} d_2^{2r_2} p^2) \right]^{1/2} \ .$$

It is shown in [12] that this problem has one and only one solution $g^*_{k-\tilde{r}} \in V_{k-\tilde{r}}$, also characterized by

$$I_K\left(d_1^{2r_1} d_2^{2r_2} g^*_{k-\tilde{r}} R\right) = 0 \ , \ \forall\, R \in R_{(k_1-r_1-1,k_2-r_2-1)}(K) \ .$$

Example 3.1. $K = \{(x,y) \in \mathbb{R}^2 \ , \ x + y \le 1 \ , \ x \ge 0 \ , \ y \ge 0\}$, $k_1 = k_2 = 3$, and : $r_1 = r_2 = 2$. In this case : $d_1(x,y) = x$, $d_2(x,y) = y$, and

$g_{(3,3)}(x,y) = x^2y^2 \left(xy - \frac{25}{132}\right)$.

We then get :

$$E^*_K(f,g_{(3,3)}) = \frac{c}{36} \|g_{(3,3)}\|_{0,K} \simeq 1.4\ 10^{-5} c \ .$$

Finally, the optimal cubature formula reads as :

$$\begin{aligned} Q_K(f,g_{(3,3)}) = \frac{1}{36} \Bigg(& 36 \int_0^1 x f(x,1-x)dx - 18 \int_0^1 x^2 \frac{\partial}{\partial x} f(x,1-x)dx \\ & + 6 \int_0^1 x^3 \frac{\partial^2}{\partial x^2} f(x,1-x)dx - \frac{25}{66} \int_0^1 x^2 \frac{\partial^3}{\partial x^3} f(x,0)dx \\ & - \int_0^1 x^2 \left(6x(1-x) - \frac{25}{66}\right) \frac{\partial^3}{\partial x^3} f(x,1-x)dx \\ & + \int_0^1 x^2(1-x)\left(3x(1-x) - \frac{25}{66}\right) \frac{\partial^4}{\partial x^3 \partial y} f(x,1-x)dx \\ & - \int_0^1 x^2(1-x)^2 \left(x(1-x) - \frac{25}{132}\right) \frac{\partial^5}{\partial x^3 \partial y^2} f(x,1-x)dx \Bigg) . \end{aligned} \tag{3.2}$$

which is exact on F , where F is the vector space generated by the family of monomials X^α , $\alpha = (\alpha_1, \alpha_2) \in A$, with

$$A = (\{0,1,2\} \times \mathbb{N}) \cup (\mathbb{N} \times \{0,1,2\}) . \tag{3.3}$$

Example 3.2. $K = \{(x,y) \in \mathbb{R}^2 \ , \ x + y \le 1 \ , \ x \ge 0 \ , \ y \ge 0\}$, $k_1 = k_2 = 5$, and : $r_1 = r_2 = 4$. In this case : $d_1(x,y) = x$, $d_2(x,y) = y$, and

$$g_{(5,5)}(x,y) = x^4y^4 \left(xy - \frac{81}{380}\right) .$$

We then get :

$$E_K^*(f,g_{(5,5)}) = \frac{c}{5!5!}\,\|g_{(5,5)}\|_{0,K} \simeq 10^{-9}c \quad .$$

Finally, the optimal cubature formula reads as :

$$\begin{aligned} Q_K(f,g_{(5,5)}) = \frac{1}{14400}\Bigg(& 14400 \int_0^1 xf(x,1-x)dx - 7200 \int_0^1 x^2 \frac{\partial}{\partial x} f(x,1-x)dx \\ & + 2400 \int_0^1 x^3 \frac{\partial^2}{\partial x^2} f(x,1-x)dx - 600 \int_0^1 x^4 \frac{\partial^3}{\partial x^3} f(x,1-x)dx \\ & + 120 \int_0^1 x^5 \frac{\partial^4}{\partial x^4} f(x,1-x)dx - \frac{486}{95} \int_0^1 x^4 \frac{\partial^5}{\partial x^5} f(x,0)dx \\ & - \int_0^1 x^4(120x(1-x) - \frac{486}{95}) \frac{\partial^5}{\partial x^5} f(x,1-x)dx \\ & + \int_0^1 x^4(1-x)(60x(1-x) - \frac{486}{95}) \frac{\partial^6}{\partial x^5 \partial y} f(x,1-x)dx \\ & - \int_0^1 x^4(1-x)^2(20x(1-x) - \frac{243}{95}) \frac{\partial^7}{\partial X^5 \partial y^2} f(x,1-x)dx \\ & + \int_0^1 x^4(1-x)^3(5x(1-x) - \frac{81}{95}) \frac{\partial^8}{\partial x^5 \partial y^3} f(x,1-x)dx \\ & - \int_0^1 x^4(1-x)^4(x(1-x) - \frac{81}{380}) \frac{\partial^9}{\partial x^5 \partial y^4} f(x,1-x)dx \Bigg) \end{aligned} \tag{3.4}$$

which is exact on F , where F is the vector space generated by the family of monomials X^α , $\alpha = (\alpha_1,\alpha_2) \in A$, with

$$A = (\{0,1,2,3,4\} \,.\, \mathbb{N}) \cup (\mathbb{N} \,.\, \{0,1,2,3,4\}) \quad .$$

Example 3.3 : $K = \{(x,y) \in \mathbb{R}^2 \,,\, x + y \le 1 \,,\, x \ge 0 \,,\, y \ge 0\}$, $k_1 \in \mathbb{N}^*$, $k_2 \in \mathbb{N}^*$ and : $r_1 = k_1 - 1$, $r_2 = k_2 - 1$. In this case : $d_1(x,y)=x$, $d_2(x,y)=y$, and $g_{(k_1,k_2)}(x,y) = x^{k_1-1}y^{k_2-1}(xy - c_k)$, where

$$c_k = \frac{(2k_1-1)(2k_2-1)}{(2k_1+2k_2-1)(2k_1+2k_2)} \quad ,$$

We then get :

$$E_K^*(f,g_{(k_1,k_2)}) = \frac{c}{k_1!k_2!} \left\| g_{(k_1,k_2)} \right\|_{0,K} =$$

$$\frac{1}{k_1!k_2!}\left(\frac{(2k_1)!(2k_2)!}{(2k_1+2k_2+2)!} - 2c_k \frac{(2k_1-1)!(2k_2-1)!}{(2k_1+2k_2)!} + c_k^2 \frac{(2k_1-2)!(2k_2-2)!}{(2k_1+2k_2-2)!} \right)$$

Finally, the optimal cubature formula reads as :

$$Q_k(f,g_k) = \frac{1}{k_1!k_2!} \left(\sum_{j=0}^{k_1-1} (-1)^j A_{j,k} \int_0^1 x^{j+1} \frac{\partial^j}{\partial x^j} f(x,1-x)dx \right.$$

$$\left. + (-1)^{k_2} B_k \int_0^1 x^{k_1-1} \frac{\partial^{k_1}}{\partial x^{k_1}} f(x,0)dx \right.$$

$$\left. + \sum_{j=0}^{k_2-1} (-1)^{k_1+j} \int_0^1 x^{k_1-1} (1-x)^j (C_{j,k}\, x(1-x) - D_{j,k}) \frac{\partial^{k_1+j}}{\partial x^{k_1} \partial y^j} f(x,1-x) \right.$$

where

$$A_{j,k} = k_2!(k_1-j-1)! \binom{j+1}{k_1}$$

$$B_k = (k_2-1)!C_k$$

$$C_{j,k} = (k_2-j-1)! \binom{j+1}{k_2}$$

$$D_{j,k} = (k_2-j-1)! \binom{j}{k_2-1} C_k$$

Remark 3.1. If K is a triangle with vertices (e,f), $(e+h_1,f)$, $(e,f+h_2)$ then using the change of variables $u = \frac{x-e}{h_1}$, $v = \frac{y-f}{h_2}$, we are led back to the situation $\hat{D} = \{(x,y) \in \mathbb{R}^2,\ x \geq 0,\ y \geq 0,\ x+y \leq 1\}$, with :

$$E_k^*(f,g_k) = h_1^{k_1+\frac{1}{2}} h_2^{k_2+\frac{1}{2}} E_{\hat{D}}^*(f,p_k). \quad ,$$

where

$$p_k = x^{r_1} y^{r_2} p_{k-\tilde{r}},$$

and $p_{k-\tilde{r}} \in V_{k-\tilde{r}}$ and is orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(\hat{D})$, with respect to $\hat{D}$ and the weight function $x^{2r_1} y^{2r_2}$.

Remark 3.2. If K is a triangle with vertices (x_1,y_1), (x_2,y_2), (x_3,y_3), then, using the change of variables

$$u(x,y) = (x_2-x_1)x + (x_3-x_1)y + x_1$$
$$v(x,y) = (y_2-y_1)x + (y_3-y_1)y + y_1 \quad ,$$

we have

$$\int_K f(u,v)dudv = |J| \int_{\hat{D}} f(u(x,y), v(x,y))\,dxdy \quad , \tag{3.5}$$

where :

$$\hat{D} = \{ (x,y) \in \mathbb{R}^2 \ , \ x \geq 0 \ , \ y \geq 0 \ , \ x + y \leq 1 \} \quad ,$$

and

$$J = \ell_2 \ell_3 - \ell_1 \ell_4 \quad ,$$

with

$$\ell_1 = x_3-x_1 \ , \ \ell_2 = x_2-x_1 \ , \ \ell_3 = y_3-y_1 \ , \ \ell_4 = y_2-y_1$$

From (3.2) and (3.5), we have the cubature formula :

(3.6)

$$Q_K(f,g_{(3,3)}) = \frac{|J|}{36}\Big(36 \int_0^1 xf(\ell_1 x+\ell_2(1-x),\ell_3 x+\ell_4(1-x)+y_1)dx$$
$$- 18(\ell_1+\ell_3) \int_0^1 x^2 \frac{\partial}{\partial x} f(\ell_1 x+\ell_2(1-x),\ell_3 x+\ell_4(1-x)+y_1)dx$$

$$+ 6(\ell_1+\ell_3)^2 \int_0^1 x^3 \frac{\partial^2}{\partial x^2} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1)\,dx$$

$$+ \frac{25}{66} (\ell_1+\ell_3)^3 \int_0^1 x^2 \frac{\partial^3}{\partial x^3} f(\ell_1 x+x_1 , \ell_3 x+y_1)\,dx$$

$$-(\ell_1+\ell_3)^3 \int_0^1 x^2(6x(1-x) - \frac{25}{66}) \frac{\partial^3}{\partial x^3} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1)\,dx$$

$$+(\ell_1+\ell_3)^3(\ell_2+\ell_4) \int_0^1 x^2(1-x)(3x(1-x)- \frac{25}{66}) \frac{\partial^4}{\partial x^3 \partial y} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1$$

$$- (\ell_1+\ell_3)^3(\ell_2+\ell_4)^2 \int_0^1 x^2(1-x)^2(x(1-x)- \frac{24}{132}) \frac{\partial^5}{\partial x^3 \partial y^2} f(\ell_1 x+\ell_2(1-x), \ell_3 x+\ell_4(1-x)+y_1)$$

4. GENERAL PROBLEM

Let K be a polygon in $\mathbb{R}^2$, from (2.3) we can write

$$I_K(f) = \sum_{i=1}^{n} I_{i,h}(f) \quad , \tag{4.1}$$

where

$$I_{i,h}(f) = \int_{K_{i,h}} f(x,y)\,dxdy \quad ,$$

where $K_{i,h}$, are defined in (2.3).

For each $K_{h,i}$, we assume that the boundary of $K_{i,h}$

$\Gamma_{h,i,j} = \bigcup_{i=1}^{3} \Gamma_{h,i,j}$

where $\Gamma_{h,i,j}$, $j = 1,2$ are defined by :

$$\Gamma_{h,i,1} = \{(x,y) \in K_{h,i} , d_{h,i,1}(x,y) = x-e_{h,i} = 0 \}$$

$$\Gamma_{h,i,2} = \{(x,y) \in K_{h,i} , d_{h,i,2}(x,y) = y-f_{h,i} = 0 \}$$

with $e_{h,i} \in \mathbb{R}$, $f_{h,i} \in \mathbb{R}$.

Hence, we obtain the following equality ,

$$I_K(f) = \frac{1}{k_1!k_2!} \left(\sum_{i=1}^{n} I_{i,h} \left(f \frac{\partial^{k_1+k_2}}{\partial x^{k_1} \partial y^{k_2}} g_{k,i} \right) \right)$$

for each $g_{k,i}$ in $V_k(K_{h,i})$.

Now, we may apply (1.7) for each $K_{h,i}$, and it follows :

$$I_K(f) = \sum_{i=1}^{n} Q_{h,i}(f,g_{k,i}) + \sum_{i=1}^{n} E_{h,i}(f,g_{k,i}) \tag{4.2}$$

$$= Q_K(f) + E_K(f) \quad .$$

Assuming $f \in M^k_{0,K}(K)$, from (4.2), we have :

$$\sup_{f \in M^k_{0,K}(c)} |E_K(f)| \leq \frac{c}{k_1!k_2!} \left(\sum_{i=1}^{n} \|g_{k,i}\|_{0,K_{h,i}} \right) .$$

If $g_{k,i}$ is of the form :

$$g_{k,i} = g^{(1)}_{h,i} d^{r_1}_{h,i,1} \cdot d^{r_2}_{h,i,2} \quad ,$$

with $g^{(1)}_{k,i} \in V_{k-\tilde{r}}$, where $\tilde{r} = (r_1,r_2)$.

By Hölder's inequality we obtain from (4.2) that

$$\sup_{f \in M^k_{0,K}(c)} |E_K(f)| \leq \frac{c}{k_1!k_2!} \left(\sum_{i=1}^{n} \|g^*_{k,i}\|_{0,K_{h,i}} \right) \tag{4.3}$$

$$= \inf_{g_{k,i} \in V^{(1)}_k} \left(\sum_{i=1}^{n} \|g_{k,i}\|_{0,K_{h,i}} \right) ,$$

where

$$g^*_{k,i} = g^*_{h,i} d^{r_1}_{h,i,1} d^{r_2}_{h,i,2} \quad , \tag{4.4}$$

with $g^{*}_{h,i} \in V_{k-\tilde{r}}(K_{h,i})$, and orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(K_{h,i})$ when considering the weight function $d^{2r_1}_{h,i,1} d^{2r_2}_{h,i,2}$. We summarize the results of this section by :

Theorem 4.1. Let K be a polygon of $\mathbb{R}^2$. Then, there exists an optimal cubature formula of the type (4.2), where

$$g_{k,i} = g^{*}_{h,i} d^{r_1}_{h,i,1} d^{r_2}_{h,i,2} ,$$

and

$g^{*}_{h,i} \in V_{k-\tilde{r}}(K_{h,i})$, and orthogonal to $R_{(k_1-r_1-1,k_2-r_2-1)}(K_{h,i})$ when considering the weight function $d^{2r_1}_{h,i,1} d^{2r_2}_{h,i,2}$.

And the remainding term satisfies the relation (4.3).

5. NUMERICAL EXAMPLES.

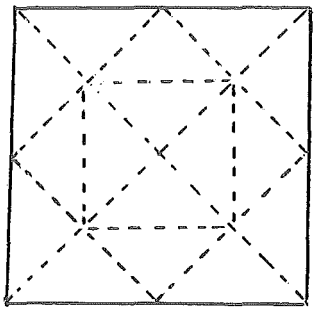
In this section we present some numerical examples in order to demonstrate the performance of the optimal formula (3.6) for various choices of K .

We compare the evaluation of the integral

$$I(f) = \int_K \frac{1}{4+x+y} \, dx \, dy$$

by the optimal formula (3.6). For each of the simple integrals of formula (3.6), we use an 5-point Gauss formula.

Example 5.1 : If $K = [-1.1] \cdot [-1.1]$, and $f(x,y) = \frac{1}{4+x+y}$,

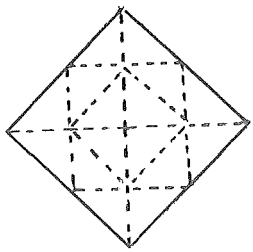


Subdividion of K into triangles.

In this case, we have :

Number of triangles	Approximate value of I(f)
16	1.04659549549661
64	1.04649884004569
256	1.04649633382633
1024	1.04649628828193
4096	1.04649628754098
Exacte value	1.0464962 8752910

Example 5.2 : If K is the Lozenge $(\pm 1,0)$, $(0,\pm 1)$, and $f(x,y) = \frac{1}{4+x+y}$



Subdivision of K into triangles.

In this case, we have :

Number of triangles	Approximate value of I(f)
16	0.509837941482940
64	0.510825848299233
256	0.510825627424528
1024	0.510825623823773
4096	0.510825623762376
Exact value	0.510825623765991

The above calculation were carried out in turbo-pascal (about sixteen significant digits) on the IBM PC AT.

References

[1] P.J. DAVIS & P. RABINOWITZ : "Methods of numerical integration". Academic press, New York, 1975.

[2] I. GANSCA : "Best quadrature formulas with relatively few terms". Revue roumaine de Math. Pures et Appliquées XXI, Vol.2, 1977, pp.143-151.

[3] J.L. GOUT & A. GUESSAB : "Sur les formules de quadrature numérique à nombre minimal de noeuds d'intégration". Numer. Math. Vol.49,1986, pp.439-455.

[4] J.L. GOUT & A. GUESSAB : "Exemples de formules de quadrature numérique à nombre minimal de noeuds sur des domaines à double symétrie axial R.A.I.R.O., vol.20, 1986, pp.287-314.

[5] A. GUESSAB : "Formules de quadrature numérique dans un compact de $\mathbb{R}^n$" . Thèse de 3ème cycle, 1983.

[6] A. GUESSAB : "Cubature formulae which are exact on space P , intermediat between P_k and Q_k". Numer. Math., Vol.49, 1986, pp.561-576.

[7] A. GUESSAB : "Numerical cubature formulas with preassigned knots". to appear. Numer. Math. 1987.

[8] A. GUESSAB : "Numerical cubature with multiple knots" to appear IMA journal or Numer. Anal. 1987.

[9] A. GUESSAB : "Sur les formules de quadrature numérique dans $\mathbb{R}^n$ avec certains noeuds ayant une compoante connue" to appear Applicable Analysis 1987.

[10] C.B. HUELSMAN:"Quadrature formulas over fully symmetric planar regions". Numer. Math. Vol.10, pp.539-552, 1973.

[11] H.I. KRYLOV : "Approximate calculation of integral". Mac. Millan New Yor London 1962.

[12] P.J. LAURENT : "Approximation et Optimation" Hermann, Paris, 1972.

[13] M. LEVIN : "On the approximate calculation of double integrals". Math. Comp. Vol.40, 1983, pp.273-282.

[14] M. LEVIN & J. GIRSHOVICH "Extremal problems for cubature formulas" Soviet Math. dokl., Vol.18, 1977, pp.1355-1358.

[15] H.M. MÖLLER : "Kubaturformeln mit minimaler knotenzhl" Num. Math. Vol.25, 1976, pp.185-200.

[16] H.M. MÖLLER : "Lower Bounds for the number of nodes in cubature formula Birlchöuser Verlag ISNM Vol.45, 1978, pp.221-230.

[17] F.W.J. OLVER : "Asymptotics and special functions" Academic Press, New York San Francisco London 1974.

18] J. PIESSENS & HAEGEMANS : "Cubature formulas of degree seven for symmetric planar regions" Journal of Comp. and applied Math. Vol.1, 1975, pp.79-83.

19] P. RABINOWITZ & N. RICHTER : "Perfectly symmetric two-dimensional integration formulas with minimal numbers of points" Math. Comp., Vol.23, 1969, pp.767-779.

20] P. RABINOWITZ & N. RICHTER : "Asymptotic properties of minimal integration rules" Math. Comp. Vol.24, 1970, pp.593-609.

21] H.J. SCHMIDT : "On cubature formulae with a minimal number of knots". Numer. Math., vol.31, 1978, pp.281-297.

22] A.H. STROUD : "Approximate calculation of multiple integrals". Prentice Hall, Englewood cliffs, N.J. 1971.

23] A.H. STROUD & D. SECREST : "Gaussian quadrature formulas". Prentice Hall, Englewood cliffs, N.J. 1966.

24] G. SZEGÖ : "Orthogonal polynomials" 3rd ed. Amer. Math. Soc. Colloq. Publ., Vol.VVIII. Amer. Math. Soc., New York (1960).

A COMBINATION OF RELAXATION METHODS AND METHOD OF AVERAGING FUNCTIONAL CORRECTIONS

DRAGOSLAV HERCEG and LJILJANA CVETKOVIĆ

ABSTRACT: *We consider a combination of the Accelerated Overrelaxation method for solving linear systems (basic method), introduced by A. Hadjidimos, with the method of Averaging Functional Corrections in order to form the composite method, which is in some cases faster than the basic method. Sufficient conditions for the convergence of this method are obtained. Several numerical examples demonstrate the efficiency of our method.*

1. INTRODUCTION

If we want to solve a system of linear equations

$$x = Mx + d,\ M = [m_{ij}] \in \mathbb{R}^{n,n},\ d = [d_1, \dots, d_n]^T \in \mathbb{R}^n, \tag{1}$$

instead of the basic iterative method

$$x^{k+1} = Mx^k + d, \qquad k = 0, 1, \dots,$$

in order to accelerate the convergence, we can use AFC (method of averaging functional corrections). This method was introduced by Sirenko [5], where it was given in the following form:

Algorithm:

Step 0: Calculate $m = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}$;

Step 1: If $n \leq m$ stop, otherwise go to step 2;

Step 2: Choose $x^0 \in \mathbb{R}^n$;

Step 3: Calculate $s_0 = \frac{1}{n-m}\sum_{i=1}^{n} d_i$; $k = 0$;

Step 4: Calculate $x^{k+1} = M(x^k + s_k\delta) + d$, $\delta = [1, \ldots, 1]^T \in \mathbb{R}^n$;

Step 5: Calculate $s_{k+1} = \frac{1}{n-m}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\left(x_j^{k+1} - x_j^k - s_k\right)$;

Step 6: Take $k = k + 1$ and return to step 4.

Numerical examples show that, very often, AFC method converges faster than the basic method. Because of that, it was the subject of our investigations, [2] the results of which we shall give in section 2.

In this paper, as the basic iterative method, we shall use AOR (Accelerated Overrelaxation) method introduced by A. Hadjidimos [4]. It means that we consider a system of linear equations

$$Ax = b, \quad A = [a_{ij}] \in \mathbb{R}^{n,n}, \quad b \in \mathbb{R}, \tag{2}$$

which we solve by using AOR method

$$x^{k+1} = M_{\sigma,\omega}x^k + d, \quad k = 0,1, \ldots,$$

and, after that, by AFC method. Here we denote by $M_{\sigma,\omega} = (D - \sigma T)^{-1}((1 - \omega)D + (\omega - \sigma)T + \omega S)$, $d = \omega(D - \sigma T)^{-1}b$, where $A = D - T - S$ is the standard splitting of the matrix A into diagonal (D), strictly lower (T) and strictly upper (S) triangular parts, σ and ω are real parameters, $\omega \neq 0$.

2. CONVERGENCE OF THE AFC METHOD

In [2] we proved that the AFC method for solving system (1) can be written in the following form

3) $$x^{k+1} = Bx^k + d', \quad k = 0, 1, \ldots$$

where

$$B = \left(E + \frac{1}{n-m} MP\right) M\left(E - \frac{1}{n}P\right), \quad d' = \left(E + \frac{1}{n-m}MP\right)d,$$

$M = [m_{i,j}] \in \mathbb{R}^{n,n}$ is the iterative matrix of the basic method, $m = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}$, P is the nxn matrix all entries of which are equal to 1 and E is identity matrix. Also, we showed that

4) $$(B)_{ij} = m_{ij} - \frac{1}{n-m} m_i(1 - m_j^*), \quad i,j = 1, 2, \ldots, n,$$

where

$$m_i = \sum_{j=1}^{n} m_{ij}, \qquad m_i^* = \sum_{j=1}^{n} m_{ji}, \qquad i = 1, 2, \ldots, n.$$

Now, it is easy to see that AFC method converges if $\rho(B) < 1$.

3. AOR + AFC METHOD

AOR + AFC method has the matrix form (3), where $M = M_{\sigma,\omega}$. Some sufficient conditions for the convergence of this method we can obtain by analysing the condition $\rho(B) < 1$. So, we obtain the following theorem.

THEOREM 1. *Let* $M_{\sigma,\omega} \geq 0$, $L_1 = \|D^{-1}T\|_1$, $\|D^{-1}(T + S)\|_\infty < 1$, $U_1 = \|D^{-1}S\|_1$, $0 < L_1 < 1$, $L_1 + U_1 \leq 1$,

5) $$\begin{cases} 0 < \omega \leq 1, & -\dfrac{\omega(1 - L_1 - U_1)}{2L_1} \leq \sigma \leq \dfrac{\omega(1 + L_1 - U_1)}{2L_1} \quad or \\ 1 \leq \omega \leq \dfrac{2}{1 + L_1 + U_1}, & \dfrac{-2 + \omega(1 + L_1 + U_1)}{2L_1} \leq \sigma \leq \dfrac{2 - \omega(1 - L_1 + U_1)}{2L_1}, \end{cases}$$

and

(6.1) $$0 < \omega \le 1, \quad -\min_{1\le i\le n} \frac{1 - l_i - u_i}{2l_i} \le \sigma \le \min_{1\le i\le n} \frac{1 + l_i - u_i}{2l_i}$$

or

(6.2) $$1 < \omega < \frac{2}{1 + \max\limits_{1\le i\le n}(l_i + u_i)}, \quad \max_{1\le i\le n} \frac{\omega(1 + l_i + u_i) - 2}{2l_i} < \sigma < 0,$$

where $$l_i = \sum_{j=1}^{n} \left|(D^{-1}T)_{ij}\right|, \quad u_i = \sum_{j=1}^{n} \left|(D^{-1}S)_{ij}\right|, \quad i = 1,2, \dots, n.$$

Then AOR + AFC method converges for any start vector.

P r o o f. If ω,σ are choosen as in (6), then $\|M_{\sigma,\omega}\|_\infty < 1$, (see [3]). From (5) we shall prove that $\|M_{\sigma,\omega}\|_1 \le 1$. Obviously,

$$\|M_{\sigma,\omega}\|_1 \le \|(D - \sigma T)^{-1}\|_1 \|(1 - \omega)D + (\omega - \sigma)T + \omega S\|_1,$$

$$\|M_{\sigma,\omega}\|_1 \le \|(E - \sigma L)^{-1}\|_1 \|(1 - \omega)E + (\omega - \sigma)L + \omega U\|_1.$$

If $|\sigma|L_1 < 1$, we have

$$\|M_{\sigma,\omega}\|_1 \le \frac{1}{1 - |\sigma|L_1}\left(|1 - \omega| + |\omega - \sigma|L_1 + |\omega|U_1\right).$$

If σ and ω are choosen as in (5), then it can be verified that the two following conditions

$$|\sigma|L_1 < 1 \quad \text{and} \quad \frac{1}{1 - |\sigma|L_1}\left(|1 - \omega| + |\omega - \sigma|L_1 + |\omega|U_1\right) \le 1$$

are satisfied. Hence $\|M_{\sigma,\omega}\|_1 \le 1$. Because of that, since $M_{\sigma,\omega} \ge 0$, we have $m_i < 1$, $m_i^* \le 1$, $i = 1, 2, \dots, m$. Now, for the AOR + AFC matrix B it holds:

$$\sum_{s=1}^{n} \left|(B)_{is} - (B)_{js}\right| = \sum_{s=1}^{n} \left| m_{is} - \frac{1}{n-m} m_i(1 - m_s^*) - m_{js} + \frac{1}{n-m} m_j (1 - m_s^*)\right| \le$$

$$\leq \sum_{s=1}^{n} |m_{is} - m_{js}| + \frac{1}{n-m} |m_i - m_j| \sum_{s=1}^{n} (1 - m_s^*) =$$

$$= \sum_{s=1}^{n} |m_{is} - m_{js}| + |m_i - m_j| \leq m_i + m_j + |m_i - m_j| =$$

$$= 2 \max (m_i, m_j) < 2.$$

Matrix B has constant row sums equal to 0. Now, we construct the matrix $C = [c_{ij}] \in \mathbb{R}^{n,n}$ as follows:

$$c_{ij} = (B)_{ij} - b_{j,min}, \quad i, j = 1, 2, \ldots, n,$$

where

$$b_{j,min} = \min_{1 \leq i \leq n} (B)_{i,j}.$$

It follows that $C \geq 0$ and

$$\gamma = \sum_{j=1}^{n} c_{ij} = - \sum_{j=1}^{n} b_{j,min} \geq 0.$$

It holds $\gamma > 0$, except in the trivial case $B = 0$. Now, the matrix $\frac{1}{\gamma} C$ is a stochastic matrix, for which (see [6]) we know that h

$$\rho\left(\frac{1}{\gamma} C\right) \leq \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} \frac{1}{\gamma} |c_{is} - c_{js}|.$$

Now, it follows

$$\rho(C) \leq \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} |c_{is} - c_{js}| = \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} |(B)_{is} - b_{s,min} - (B)_{js} + b_{s,min}|$$

$$\rho(C) \leq \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} |(B)_{is} - (B)_{js}| < 1.$$

It remains to show that $\rho(B) \leq \rho(C)$. Let λ denote an eigenvalue of the matrix B^T, $\lambda \neq 0$, and let y be a corresponding eigenvector. For i = 1, 2, ...,n,

we have

$$(B^T y)_i = \sum_{j=1}^{n} (B)_{ji} y_j = \lambda y_i$$

and

$$\lambda \sum_{i=1}^{n} y_i = \sum_{j=1}^{n} y_j \sum_{i=1}^{n} (B)_{ji} = 0.$$

Since $\lambda \neq 0$, we obtain $\sum_{i=1}^{n} y_i = 0$. Using this, we have for $i = 1, 2, \ldots, n$,

$$(B^T y)_i = \sum_{j=1}^{n} (B)_{ji} y_j = \sum_{j=1}^{n} (B)_{ji} y_j - b_{i,min} \sum_{j=1}^{n} y_j = \sum_{j=1}^{n} \left((B)_{ji} - b_{i,min} \right) y_j = (C^T y)_i.$$

So, λ is an eigenvalue of the matrix C^T, too. Since B and B^T, as well as C and C^T have the same eigenvalues, we can conclude that all eigenvalues of the matrix B (except, might be, $\lambda = 0$) are eigenvalues of the matrix C. Hence, $\rho(B) \leq \rho(C) < 1$, which completes the proof. □

Some of the conditions from Theorem 1 are very restrictive. For example, the absolute and maximum norm of the Jacobi matrix ($M_{0,1}$) have to be less than 1. The condition $M_{\sigma,\omega} \geq 0$ is satisfied, for example, when A is an L-matrix and $0 \leq \sigma \leq \omega \leq 1$, while intervals for σ and ω, given by (5) and (6) are always nonempty (always it is possible to choose $\sigma = 0$, $\omega = 1$). The convergence intervals given in this theorem are always wider than the ones from [1].

From the above discussion we can not conclude when the combined AOR+AFC method converges faster than the basic AOR method. But, numerical examples show that in some cases the AOR+AFC method has better convergence than the AOR method. So, for the simple example 3 the graphs of Figure 2 give the behaviour of the actual error as a function of the iteration k for both AOR and AOR+AFC methods.

4. NUMERICAL EXAMPLES

Example 1. We consider a system of linear equations with the matrix

$$A = \begin{bmatrix} -0.875 & 0.05 & 0.0125 & 0.0125 & 0.0625 & 0.0624 \\ 0.024 & -0.75 & 0.0125 & 0.125 & 0.00624 & 0.01325 \\ 0.12 & 0.0125 & -0.875 & 0.0625 & 0.05 & 0.0625 \\ 0.0125 & 0.12 & 0.0125 & -0.5 & 0.0625 & 0.0624 \\ 0.00625 & 0 & 0.0025 & 0.0625 & -0.9375 & 0.0049 \\ 0.00327 & 0 & 0.024 & 0.0124 & 0.025 & -0.837 \end{bmatrix}$$

The convergence area for the parameters of AOR + AFC method is given in the figure 1.

Example 2. In the Table 1 we can see that AOR + AFC method converges even if the basic method diverges, as well as the convergence is very fast, where the matrix of linear system is:

$$A = \begin{bmatrix} -0.875 & 0.5 & 0.125 & 0.125 & 0.0625 & 0.0624 \\ 0.24 & -0.75 & 0.125 & 0.125 & 0.0624 & 0.1325 \\ 0.12 & 0.125 & -0.875 & 0.0625 & 0.5 & 0.0625 \\ 0.125 & 0.12 & 0.125 & -0.5 & 0.0625 & 0.0624 \\ 0.625 & 0 & 0.25 & 0.0625 & -0.9375 & 0.49 \\ 0.327 & -0.001 & 0.24 & 0.0124 & 0.25 & -0.837 \end{bmatrix}$$

Table 1.

(σ,ω)	$\rho(M_{\sigma,\omega})$	$\rho(B)$
(0,1)	1.070	0.531
(1,1)	1.128	0.650
(0.255,0.882)	1.070	0.252

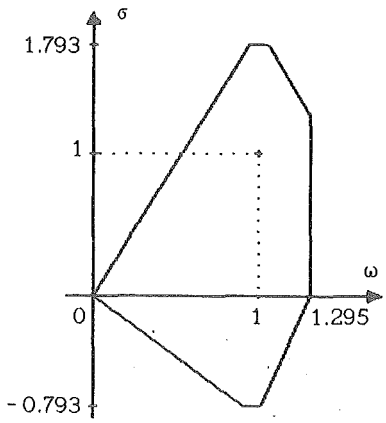


Figure 1.

Example 3. Let

$$A = \begin{bmatrix} -0.8 & 0.22 & 0.26 & 0.24 \\ 1 & -1.02 & -0.06 & -0.04 \\ 0.5 & 0.4 & -1.03 & -0.02 \\ 0.5 & -0.5 & 1 & -0.98 \end{bmatrix} \qquad b = \begin{bmatrix} 0.42 \\ -0.22 \\ -1.20 \\ 0.40 \end{bmatrix}$$

In the following Figure 2 we present the value -logE as a function of iteration k, where

$$E := \frac{\|x - x^k\|_\infty}{\|x\|_\infty},$$

and $x = [1,1,2,2]^T$ is the exact solution of the system $Ax = b$, and x^k is the k-th iteration obrtained by AOR or AOR + AFC method with $x^0 = [100,0,0,-100]$.

The graphs are denoted as follows:

1 - Jacobi method;

2 - Gauss-Seidel method;

3 - AOR method with $\sigma = 0.9875$ and $\omega = 1.27$;

4 - AOR method with $\sigma = 0.972$ and $\omega = 0.965$;

5 - Jacobi + AFC method;

6 - Gauss-Seidel + AFC method;

7 - AOR + AFC method with $\sigma = 0.9875$ and $\omega = 1.27$;

8 - AOR + AFC method with $\sigma = 0.972$ and $\omega = 0.965$.

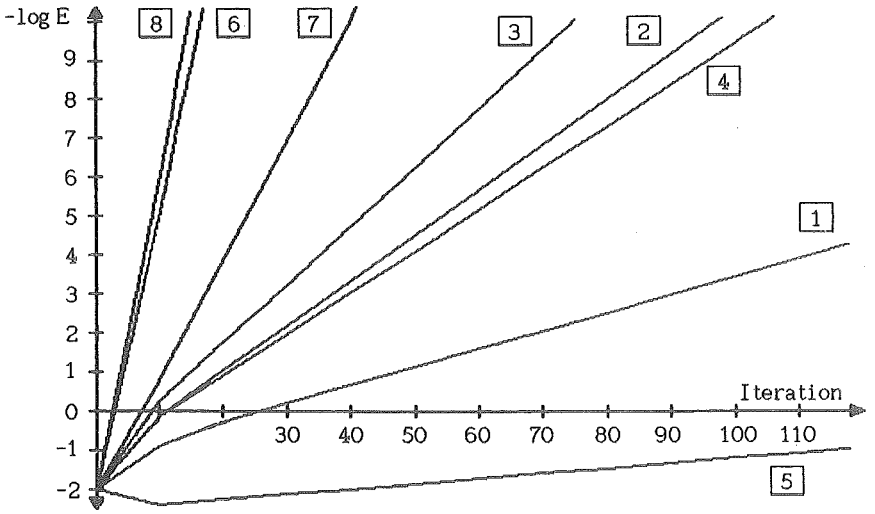


Figure 2.

Acknowledgments. We would like to thank the referee for having improved the presentation of the paper.

REFERENCES

[1] Lj. Cvetković, D. Herceg: Eine Modifikation des AOR-Verfahrens, Z. Angew. Math. Mech. Bd. 67,N.5(1987), 913-914.

[2] Lj. Cvetković, D. Herceg: On the method of averaging functional corrections applied to linear systems, Zb. Rad. Prir.Mat.Fak. Ser.Mat. 17,2(1987) (in print).

[3] Lj. Cvetković, D. Herceg: An improvement for the area of convergence of the AOR method, Anal. Numer. Theor. Approx. 16(1987), 109-115.

[4] A. Hadjidimos: Accelerated overrelaxation method, Math. Comp., 32(1978), 149-157.

[5] В.Х. Сиренко: О численной реализации метода осреднения функциональных поправок, УМЖ 13(1961), 51-66.

[6] Ch. Zenger, A comparation of some bounds for the nontrivial eigenvalues of stochastic matrices, Numer. Math. 19(1972), 209-211.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

ON THE EFFICIENCY OF ITERATIVE METHODS FOR BOUNDING THE INVERSE MATRIX

J. HERZBERGER

ABSTRACT: *In this note we are considering the higher-order interval Schulz methods for improving bounds for the inverse matrix. First we give a different computation scheme for the iteration formula which is more efficient especially for the higher-order formulas. Next, we derive a modification of the methods which has the same properties as the original ones but compares favourably for the higher-order cases. For both versions presented here some efficiency indices are listed and compared with those of the original formulas.*

0. INTRODUCTION

Let $\underset{\cdot}{A}$ be an m×m nonsingular real matrix and $X^{(0)}$ be an m×m interval matrix with $\underset{\cdot}{A}^{-1} \in X^{(0)}$. In [1], Chapter 18 there are described iteration methods which improve $X^{(0)}$ iteratively. These formulas are the following ones:

$$(1) \qquad X^{(n+1)} = m(X^{(n)}) \sum_{i=0}^{r} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^i + X^{(n)} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^{r+1} ,$$

$$(2) \qquad X^{(n+1)} = \{m(X^{(n)}) \sum_{i=0}^{r} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^i + X^{(n)} (\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)}))^{r+1}\} \cap X^{(n)} .$$

where

$$\underset{\cdot}{I} - \underset{\cdot}{A} m(X^{(n)})$$

and

$$m(X) = m(([x_{ij}^1, x_{ij}^2])) = ((x_{ij}^1 + x_{ij}^2)/2)$$

$(r \geq 0)$.

In Theorem 1 and Theorem 2 of Chapter 18 in [1] it is shown that for the methods (1) and (2)

$$\underset{\cdot}{A}^{-1} \in \underset{\cdot}{X}^{(k)} \quad , \quad k \geq 0$$

holds true. Furthermore, for the R-order of convergence (see [1]) we have the estimations

$$O_R((1),\underset{\cdot}{A}^{-1}) \geq r+2 \quad \text{and} \quad O_R((2),\underset{\cdot}{A}^{-1}) \geq r+2 \ .$$

Since the convergence criterion of (1) is weaker than that of (2), we usually start with (1) and after some contracting iterations switch to method (2) as soon as its convergence criterion is fulfilled. Then method (2) produces a nested sequence of inclusions for $\underset{\cdot}{A}^{-1}$ and thus allows to establish a quite natural stopping rule. For more details see [1], Chapter 18. Formulas (1) and (2) are computed by means of the Horner scheme and require $r+2$ matrix multiplications each of them. Now, the efficiency index (see [4] Appendix C) E_H can be estimated by

$$E_H \geq (r+2)^{\frac{1}{r+2}} \ .$$

1. MODIFIED SCHEMES

We consider the iteration formulas

$$(3) \qquad \underset{\cdot}{X}^{(n+1)} = m(\underset{\cdot}{X}^{(n)}) \prod_{j=0}^{k-1} \Big(\sum_{i=0}^{r} (\underset{\cdot}{T}^{(n)})^{i(r+1)^j} \Big) + \underset{\cdot}{X}^{(n)} (\underset{\cdot}{T}^{(n)})^{(r+1)^k}$$

and

$$(4) \qquad \underset{\cdot}{X}^{(n+1)} = \Big\{ m(\underset{\cdot}{X}^{(n)}) \prod_{j=0}^{k-1} \Big(\sum_{i=0}^{r} (\underset{\cdot}{T}^{(n)})^{i(r+1)^j} \Big) + \underset{\cdot}{X}^{(n)} (\underset{\cdot}{T}^{(n)})^{(r+1)^k} \Big\} \cap \underset{\cdot}{X}^{(n)}$$

($k \geq 1$, $r \geq 0$). Setting $k=1$ in (3) and (4) we formally get the methods (1) and (2) as special cases. On the other hand, by virtue of the equality for real matrices $\underset{\cdot}{C}$

$$\prod_{j=0}^{k-1} \Big(\sum_{i=0}^{r} \underset{\cdot}{C}^{i(r+1)^j} \Big) = \sum_{i=0}^{(r+1)^k-1} \underset{\cdot}{C}^{i}$$

which can be proved by complete induction using a proper rearrangement of the summation terms, we get formulas equivalent to (3) and (4) by

$$(3)' \qquad X^{(n+1)} = m(X^{(n)}) \sum_{i=0}^{(r+1)^k-1} (\dot{T}^{(n)})^i + X^{(n)} (\dot{T}^{(n)})^{(r+1)^k} ,$$

$$(4)' \qquad X^{(n+1)} = \{m(X^{(n)}) \sum_{i=0}^{(r+1)^k-1} (\dot{T}^{(n)})^i + X^{(n)} (\dot{T}^{(n)})^{(r+1)^k}\} \cap X^{(n)} .$$

This shows that (3) and (4) are also just methods of the kind of (1) and (2) and therefore all have the same properties. In particular the R-order of convergence is

$$O_R((3), \dot{A}^{-1}) \geq (r+1)^k + 1 \quad \text{and} \quad O_R((4), \dot{A}^{-1}) \geq (r+1)^k + 1 .$$

Again we measure the amount of work by the necessary matrix multiplications which count exactly $k(r + (1 - \delta_{1k})\varphi(r+1)) + 2$ when using the Horner scheme for the occuring matrix polynomial factors. Here $\varphi(u)$ denotes the number of multiplications required for computing the u-th power. So, we get for the efficiency index E_M the estimation

$$E_M \geq ((r+1)^k+1)^{\frac{1}{k(r+(1-\delta_{1k})\varphi(r+1))+2}} .$$

In comparison to this the efficiency index for the original formulas (1) and (2) of the corresponding order was

$$E_H \geq ((r+1)^k+1)^{\frac{1}{(r+1)^k+1}} .$$

The following tables show for some selected values of the parameters r and k bounds for the efficiency indices.

r=1	k=1	2	3	4	5	6
E_H	1.442	1.379	1.277	1.181	1.112	1.066
E_M	1.442	1.308	1.316	1.328	1.338	1.347

r=2	k=1	2	3
E_H	1.414	1.259	1.126
E_M	1.414	1.259	1.269

Remarks: As the tables show, the modifications (3) and (4) are considerably more efficient for greater values of parameters or - with other words - for higher orders.
The bounds for the efficiency indices E_H and E_M achieve its maximum $\sqrt[3]{3}$ for the formulas of order three as an easy analysis shows.

2. MODIFIED METHODS

Now, we consider the iteration methods

$$y^{(n+1,0)} = m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n)} (I - Am(X^{(n)}))^r ,$$

$$y^{(n+1,i+1)} = m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n+1,i)} (I - Am(X^{(n)}))^r , \tag{5}$$

$$X^{(n+1)} = y^{(n+1,s)} , \qquad 0 \leq i < s ,$$

and

$$y^{(n+1,0)} = \{m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n)} (I - Am(X^{(n)}))^r\} \cap X^{(n)} ,$$

$$y^{(n+1,i+1)} = \{m(X^{(n)}) \sum_{j=0}^{r-1} (I - Am(X^{(n)}))^j + \\ + X^{(n+1,i)} (I - Am(X^{(n)}))^r\} \cap y^{(n+1,i)} , \tag{6}$$

$$X^{(n+1)} = y^{(n+1,s)} , \qquad 0 \leq i < s ,$$

where $r \geq 1$ and $s \geq 0$. (In case $s = 0$ the second statements of the iteration formulas are to be empty.) Setting $s = 0$ we again get the methods (1) and (2) as special cases.

A simple backward substitution of the quantities $y^{(n,i)}$ in (5) leads to the equivalent formula

$$(5)' \qquad X^{(n+1)} = m(X^{(n)}) \sum_{i=0}^{(s+1)r-1} (I - A\,m(X^{(n)}))^i + \\ + (\ldots((X^{(n)}(I - A\,m(X^{(n)}))^r)(I - A\,m(X^{(n)}))^r)\ldots$$

which is again a method like (1). Such a transformation is, however, not possible for (6).

From the equality

$$m(X^{(n)}) \sum_{i=0}^{r-1} (I - A\,m(X^{(n)}))^i = A^{-1} - A^{-1}(I - A\,m(X^{(n)}))^r$$

together with the inclusion monotonicity of the interval operations we get for (5) and (6) by complete induction the property

$$A^{-1} \in X^{(n+1)}, \quad y^{(n+1,i)}, \quad (0 \leq i \leq s), \quad n \geq 0 .$$

Similarly like in [1] Chapter 18 we can prove by a straight forward analysis the same convergence criteria for (5) and (6) as for (1) and (2). As for the R-order of convergence, we immediately get from the representation (5)' of (5)

$$O_R((5), A^{-1}) \geq (s+1)r + 1 .$$

By a similar analysis for the sequences $\{d(X^{(n)})\}$, where d is the width operator, we get in addition to this estimation

$$O_R((6), A^{-1}) \geq (s+1)r + 1 .$$

The amount of work in terms of matrix multiplications is in case of (5) or (6) $r + s + (1 - \delta_{0,s})\varphi(r) + 1$. Thus we get for the efficiency index E_{MM} the estimation

$$E_{MM} \geq ((s+1)r + 1)^{\frac{1}{r+s+(1-\delta_{0,s})\varphi(r)+1}} = \alpha(r,s)$$

in contrast to the corresponding efficiency index E_H

$$E_H \geq ((s+1)r+1)^{\frac{1}{(s+1)r+1}} = \beta(r,s) .$$

It is easy to see that the inequality

$$\beta(r,s) \leq \alpha(r,s)$$

holds true. This means that the unmodified methods are not so efficient as the modified methods.
The following tables give the bounds for E_{MM} and E_H for some selected values of parameters r and s.

r=2	s=1	2	3	4	5
E_H	1.380	1.320	1.277	1.244	1.218
E_{MM}	1.380	1.383	1.367	1.350	1.330

r=3	s=1	2	3	4	5
E_H	1.320	1.259	1.218	1.189	1.168
E_{MM}	1.320	1.344	1.330	1.320	1.307

Remarks: The bounds for the efficiency index E_{MM} achieve its maximum $\sqrt[3]{3}$ for s=0, r=2 or r=s=1 with methods of order three. A direct comparison between the given bounds for E_M and E_{MM} is not possible because the orders of convergence of the produced methods for different values of parameters do not coincide.

REFERENCES

1. G. ALEFELD and J. HERZBERGER: Introduction to Interval Computations. Academic Press, New York 1983.

2. J. HERZBERGER: Some aspects of iterative methods for bounding the inverse matrix. Colloquia Mathematica Societas János Bolyai, 50. Numerical Methods, 1987, 185-199.

3. J. HERZBERGER: Zur Effizienz von intervallmäßigen Schulz-Verfahren höherer Ordnung. Z. angew. Math. Mech. (to appear).

4. J.F. TRAUB: Iterative methods for the solution of equations. Prentice-Hall, Englewood Cliffs N.J. 1964

PROCESS IDENTIFICATION USING B-SPLINES

LJ. M. KOCIĆ and B. DANKOVIĆ

Abstract: An application of B-splines in computation of Laplace transform of the unit step response (and then the transfer function) of a given system of automatic control is considered. The algorithm suggested is based upon de Boor-Cox algorithm for numerically stable calculation of B-splines. Error estimation is given, and a numerical experiment is performed.

1. INTRODUCTION

Suppose that we are given the system (process) of automatic control in a "black box" form (Fig. 1). To identify this process means to find its transfer function $W(s)=Y(s)/X(s)$, where $s \mapsto X(s)$ $(s \mapsto Y(s))$ is the Laplace transform of input (output) signal as function of time: $t \mapsto x(t)$ $(t \mapsto y(t))$. Namely, we shall write $X(s)=L(x(t))=\int_0^\infty e^{-st}x(t)dt$, where we supposed $x, y \in L_2[0, \infty)$, and $x(t)=y(t)=0$ for $t<0$, which is satisfied in a great number of practical situations. When $x(t)$ is the unit step function, we have $W(s)=sY(s)=sL(y(t))$, where $y(t)$ will be regarded as the unit step response of the "black box", and we can gather the information on $y(t)$ only by measuring it in some discrete set of points $\tau=\{\tau_0, \tau_1, \ldots, \tau_N\}$ $(\tau_i<\tau_{i+1}$, $i=0, \ldots, N-1)$. As the result, we should get the set of data $d=\{y_i=y(t_i)\}_{i=1}^N$. Based on τ and d we can calculate the function $y^*(t)$ which approximate $y(t)$ on $[\tau_0, \tau_N]$ so that $\|y-y^*\|<E_N$, where E_N is the prescribed error of appro-

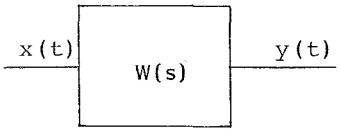


Fig. 1.

ximation and $\|\cdot\|$ is one of the usual norms, taken over the interval $[\tau_0, \tau_N]$. Now, we can find $Y^*(s) = L(y^*(t))$ and then $W^*(s) = sY^*(s)$. But, we must pay attention to an important detail. The unit step response $y(t)$, always (for real systems) approaches to a fixed value, say, y^∞, after a long enough interval of time. It is convenient to suppose that $y^*(t)$ approximates $y(t)$ on $[\tau_0, \tau_N]$ and $y^*(t) = 0$ outside, so with $y^*(t) + y^\infty$ we have the approximation to $y(t)$ completed. Then, we can put $Y^*(s) = L[y^*(t) + y^\infty] = L[y^*(t)] + s^{-1}y^\infty$, and therefore $W^*(s) = sY^*(s) = s\,L[y^*(t)] + y^\infty$. Since $W(s)$ has the similar form, namely $W(s) = sL[y(t)] + y^\infty$, the error (see section 3) will not contain y^∞.

So, the problem is to find the approximation y^* for the unit step response y which has to be "good" in the following sense:

$$\|y^* - y\| \to \min \; ; \quad \left\| \frac{d}{dt} y^* - \frac{d}{dt} y \right\| \to \min .$$

This two requests arises in natural way in the theory of adaptive processes (see for example, [10]).

In this paper we investigate the most convenient way to use polynomial splines in order to compute $L[y^*(t)]$ and to estimate the error $|W(s) - W^*(s)|$.

Let $S_{k,\xi}$ be the space of polynomial splines of order k, with the knot sequence $\xi = \{\xi_1, \xi_2, \ldots, \xi_{n+1}\}$ (which is strictly increasing one). Then, if, for example, $y^* \in S_{k,\xi}$, we have representation via truncated power basis

$$y^*(t) = \sum_{j=0}^{k-1} a_j (t - \xi_1)^j + \sum_{i=2}^{n} b_i (t - \xi_i)_+^{k-1} ,$$

where

$$a_j = \frac{y^*(\xi_1 + 0)^{(j)}}{j!} , \quad b_j = \frac{D^{k-1}y^*(\xi_i + 0) - D^{k-1}y^*(\xi_i - 0)}{(k-1)!} .$$

This representation is very convenient for applying L-transform, but, unfortunately, very unstable for numerical calculations (see [1]). So, we will turn to B-splines which provide very stable numerical process.

Some of known methods (for example the Aizerman's method) use piecewise constant approximation of y. In terms of splines, this means that $y^* \in S_{1,\xi}$ (see [7]). In [3], the function y is approximated by parabolic segments performed to fit the set of data $\mathbf{d}$. Of course, such interpolant, y^* sufferes from low smootheness.

2. B-SPLINE AND ITS L-TRANSFORM

As it is already known, the space $S_{k,\xi}$ has so called B-spline base $\{ B_{i,k} \}_{i=1}^{n}$, where $B_{i,k}$ is defined for the set of knots $t = \{ t_1, \ldots, t_{n+k} \}$ which can be derived from ξ by adding $2k$ new knots $t_1 = \ldots = t_k = \xi_1$ and $t_{n+1} = \ldots = t_{n+k} = \xi_{n+1}$ and so that $\xi_i = t_i$ $(i=k+1,\ldots,n)$. Then, the i-th B-spline of order k is given by

$$(1) \quad B_{i,k}(t) = (t_{i+k} - t_i)[t_i, \ldots, t_{i+k}](\cdot - t)_+^{k-1} ,$$

for $i = 1, \ldots, n$. Now, for $y^* \in S_{k,t}$ we have

$$(2) \qquad y^*(t) = \sum_{i=1}^{n} c_i B_{i,k}(t).$$

Computation of the coefficients c_i depands on approximation scheme we want to use. For example, we can interpolate the data $\mathbf{d}$ in the nodes τ. If we have the freedom of choosing the nodes τ_i it is advisable to take

$$\tau_i = \frac{1}{k-1} (t_{i+1} + \ldots + t_{i+k-1}).$$

We also can calculate c_i in order to smooth the data $\mathbf{d}$. According to [1] we can do that by minimizing the quantity

$$p \sum_{i=1}^{N} \left(\frac{y_i - y^*(\tau_i)}{\delta y_i} \right)^2 + (1-p) \int_{\tau_1}^{\tau_N} (D^m y^*(t))^2 \, dt$$

where δy_i is an estimate of the variance in y_i, and $p \in [0,1]$ is a given parameter. The role of p is to emphasize the closeness to data (when $p \to 1_-$) or smootheness of y^* (when $p \to 0_+$). In this way, we get so called Whittaker spline, and the corresponding package SMOOTH is given in [1]. By the another procedure from [1], named L2APPR, we can calculate c_i from (2), and the spline y^* is then the approximation of data in the sense of least squares.

Another interesting procedures for smoothong data via splines from $S_{k,t}$ can be found in [4] and [5].

Suppose that we have all c_i $(i=1,\ldots,n)$ calculated, so we have y^* completely defined. Now, if we apply the L operator on both sides of (2), we shall get

$$(3) \qquad L(y^*(t)) = \sum_{i=1}^{n} c_i \, L(B_{i,k}(t)) = \sum_{i=1}^{n} c_i \, L_{i,k}(s).$$

So, we must calculate $L_{i,k}(s) = L(B_{i,k}(t))$ $(i=1,\ldots,n)$, and we have to do that in the most efficient way. For example, we do not recommend using the explicit formula

$$B_{i,k}(t) = (t_{i+k} - t_i) \sum_{j=i}^{i+k} \frac{(t_j - t)_+^{k-1}}{\pi'_{k,i}(t_j)} \quad , \qquad \pi_{k,i}(t) = \prod_{j=i}^{i+k} (t - t_j)$$

from the reason of its low accuracy. Instead of that, we shall start from de Boor-Cox algorithm for stable calculation of B-splines ([1] , [2]):

$$(4) \qquad B_{i,1}(t) = \begin{cases} 1 \, , & t_i \leq t < t_{i+1} \, , \\ 0 \, , & \text{otherwise} \, , \end{cases}$$

$$(5) \qquad B_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t).$$

Can we find the similar reccurence formula for $L_{i,k}(s)$? The answer is affirmative. Namely, from (4) we can compute directly

$$(6) \qquad L_{i,1}(s) = \frac{1}{s}\left(e^{-t_i s} - e^{-t_{i+1} s} \right), \quad i=1,\dots,n.$$

Now, we can use the definition relation (1). Owing to the obvious identity

$$(t - x)_+^{k-1} = (t - x)^{k-1} + (-1)^k (x - t)_+^{k-1} ,$$

and the fact that $[t_i,\dots,t_{i+k}]\,(t - \cdot)^{k-1} = 0$, we have

$$(7) \qquad B_{i,k}(t) = (-1)^k (t_{i+k} - t_i)\,[t_i,\dots,t_{i+k}]\,(t - \cdot)_+^{k-1} .$$

The point, named "placeholder", states instead of the variable which the divided difference is applying on. If we apply the L-operator on both sides of (7), we get

$$(8) \qquad L_{i,k}(s) = (-1)^k (t_{i+k} - t_i)\,[t_i,\dots,t_{i+k}]\{L(t - \cdot)_+^{k-1}\}$$

$$= \frac{(k-1)!}{s^k} (-1)^k (t_{i+k} - t_i)[t_i,\dots,t_{i+k}] e^{-s(\cdot)} .$$

The authors of [8] have derived the same formula beginning from Schoenberg's identity.

Now, due to the recurrence relation for divided differences we have

$$L_{i,k}(s) = \frac{(k-1)!}{s^k}(-1)^k (t_{i+k} - t_i)\, x$$

$$x \frac{[t_{i+1},\dots,t_i] e^{-s(\cdot)} - [t_i,\dots,t_{i+k-1}] e^{-s(\cdot)}}{t_{i+k} - t_i}$$

$$= \frac{k-1}{s} \left\{ \frac{1}{t_{i+k-1} - t_i} \frac{(k-2)!}{s^{k-1}} (-1)^k (t_{i+k-1} - t_i)[t_i,\dots,t_{i+k-1}] e^{-s(\cdot)} \right.$$

$$\left. - \frac{1}{t_{i+k} - t_{i+1}} \frac{(k-2)!}{s^{k-1}} (-1)^{k-1} (t_{i+k} - t_{i+1})[t_{i+1},\dots,t_{i+k}] e^{-s(\cdot)} \right\}$$

and thus

$$(9)\quad L_{i,k}(s) = \frac{k-1}{s}\left(\frac{L_{i,k-1}(s)}{t_{i+k-1} - t_i} - \frac{L_{i+1,k-1}(s)}{t_{i+k} - t_{i+1}}\right) \quad i=1,\ldots,n,\ k \geq 2$$

where we have taken into account (8). So, the set $\{L_{i,k}(s)\}_{i=1}^{n}$ is completely defined by (6) and (9). This commpletes the procedure of finding $Y^*(s)$ (see (3)) and then $W^*(s)$ as well.

We must underline that the computation of (9) can also become unstable for small $|s|$, and then we recommand technique given in [6]. For further study of L-transform of B-splines, see [9].

3. ERROR ESTIMATION

The distance between $W(s)$ and $W^*(s)$ in the complex plane $s = \sigma + j\omega$ is given by

$$|W(s) - W^*(s)| = |s|\,|Y(s)-Y^*(s)| = |s|\left|\int_0^{+\infty}\{y(t)-y^*(t)\}e^{-st}dt\right|$$

$$\leq |s| \int_{\tau_0}^{\tau_N} |y(t)-y^*(t)|\, e^{-\sigma t}\,dt \leq |s|\, E_N \int_{\tau_0}^{\tau_N} e^{-\sigma t}\,dt\ ,$$

so we have

$$|W(s) - W^*(s)| \leq \begin{cases} \left[1 + \left(\frac{\omega}{\sigma}\right)^2\right]^{1/2}\left(e^{-\sigma\tau_0} - e^{-\sigma\tau_N}\right)\cdot E_N\ , & \sigma \neq 0, \\ \left[\sigma^2 + \omega^2\right]^{1/2}(\tau_N - \tau_0)\cdot E_N\ , & \sigma = 0. \end{cases}$$

The case $\sigma = 0$ is of especially importance in analysis of stability of automatic control systems. The E_N is the C-norm error of spline approximation. For example, if y^* is a cubic spline interpolant for the data d, we have

$$E_N \leq \frac{5}{385}|\tau|^4 \max_{[\tau_0,\tau_N]}\left|y^{(4)}(t)\right|\ ,\ |\tau| = \max_i \Delta\tau_i$$

and so on.

4. EXAMPLE

For a test-example we can take an ideal system with the unit step response $y(t) = 1 - e^{-t}$ (Fig.2). The graph of its tran-

sfer function, in the case $\sigma = 0$ is a half-circle. It is represented in (P, Q)-plane, where $P + jQ = W(j\omega)$ (Fig. 3). The cubic spline function $y^*(t)$ that interpolates $y(t)$ in the nodes (0., 2., 4., 6.) is constructed and its graph is shown in Fig. 4. The equidistant set of nodes we used is the worst choice we can made. This results in oscilations of y^*. However, the resulting graph of $W^*(j\omega) = P^*(\omega) + jQ^*(\omega)$, shown in Fig.4, has very satisfied form.

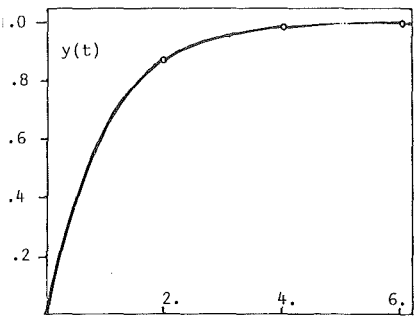


Fig. 2. Ideal system step response $y(t)$, $0. < t < 6.$;

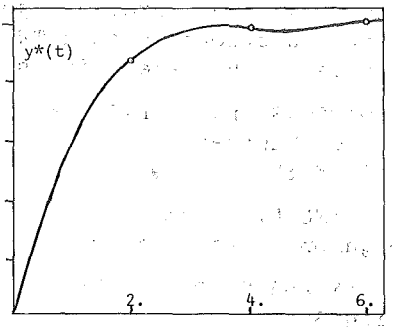


Fig. 3. Cubic spline $y^*(t)$ with the nodes 0.,2.,4.,6.;

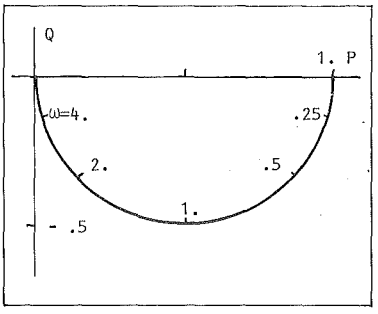


Fig. 4. Transfer function $W(j\omega)$ for the ideal system;

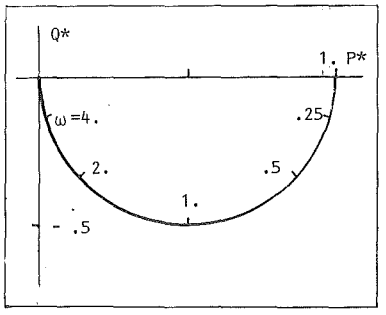


Fig. 5. Approximation of the transfer functiom based on spline approximation y^*.

The authors are grateful to the referees and professor G.V. Milovanović for a number of valuable suggestions which led to improvement of this paper.

REFERENCES

1. C. de BOOR: A Practical Guide to Splines. Springer-Verlag, New York, 1978.
2. M. G. Cox: The Numerical Evaluation of B-Splines. J. Inst. applics., 10 (1972), 132-149.
3. B. Danković, D. Ignjatović: Korišćenje računara pri identifikaciji tehnoloških procesa. In: Informacijski sistemi zasnovani na primeni računara, Zbornik radova, Niš 1985,69-76.
4. P. DIERCXS: An algorithm for smoothing, differentiation and integration of experimental data using spline functions. J. Comput. Appl. Math., 1 (1975), no. 3, 165-184.
5. P. DIERCXS: An Improved Algorithm for Curve Fitting with Spline Functions. Report TW54, July 1981, Dept. of Computer Science, Katholieke Universitet, Leuven (Belgium).
6. P. DIERCXS and R. PIESSENS: Calculation of Fourier Coefficients of Discrete Functions Using Cubic Splines. J. Comp. Appl. Maths. 3 (1977), 207-209.
7. P. EYKNOFF: Trends and progress in system identification. Pergamon Press, Oxford 1981.
8. M. LAX and G. P. AGRAWAL: Evaluation of Fourier Integrals Using B-Splines. Maths. Computation 39 (1982),no.160,535-548
9. W. SCHEMPP: Complex Contour Integral Representation of Cardinal Spline Functions. Contemporary Mathematics, Vol. 7, Amer. Math. Soc., Providence 1982.
10. V. TARAN, S. BRUDNIK, J. KOFANOV: Matematicheskie voprosi avtomatizacii priozvodstvenyh processov. Vyshaya shkola, Moskva 1978.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

ON CALCULATING QUADRATIC B-SPLINES IN TWO VARIABLES*

J. KOZAK and M. LOKAR

ABSTRACT: *One of the encountered problems in practical use of multivariate splines is a stable and efficient evaluation of a spline given as a linear combination of B-splines. No generalization has been found for the well-known univariate recursion scheme. Thus the only way to compute the value of a spline is to compute the values of all B-splines incident at a given point. In the paper we propose a special scheme for calculating all quadratic B-splines (in two dimensions) incident at a point in a certain subregion of the original domain. Our discusion can be viewed as a refinement of the work done by Meyling ([7,8]). We show that our scheme requires minimal constant B-splines evaluations.*

1. Introduction

Multivariate (simplex) splines have attracted quite a lot of attention in the past ten years. However, the theoretical work was not so widely followed by practical applications as one might expect for such a powerfull and flexible tool. There are several reasons for this fact. Perhaps one of the main obstacles is algorithmic and computational complexity of the computer procedures. The purpose of this paper is to tackle one practical aspect in dealing with bivariate quadratic splines, i.e. an evaluation of a spline. Though this is the simplest nontrivial case, several computational problems will be revealed.

---

*Supported by Research Council of Slovenija.

In order to construct a bivariate spline we recall its basis function first. There are several ways to define a B-spline. Perhaps the most apparent is the geometric one ([1]). A bivariate B-spline of degree n-2 is given by

$$M(\underline{x}|\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^n) := \frac{\mathrm{vol}_{n-2}(\{\underline{v}\in\sigma:\underline{v}|_{\mathbb{R}^2}=\underline{x}\})}{vol_n(\sigma)},$$

independently of σ, where $\sigma := [\underline{v}^0,\underline{v}^1,\ldots,\underline{v}^n]$ is n-simplex in $\mathbb{R}^n$ such that

i) $\mathrm{vol}_n(\sigma) > 0$,

ii) $\underline{v}^i|_{\mathbb{R}^2} = \underline{x}^i$, *for* $i = 0,1,\ldots,n$.

In other words, the set $\{\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^n\}$ *contains the orthogonal projections of the vertices* $\underline{v}^i$ *onto* $\mathbb{R}^2$. *Quite clearly M is a piecewise polynomial function of total degree n-2. If* $\underline{x}^i$ *are in general position (no triple lies on a line) then*

$$M(\underline{x}|\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^n) \in C^{n-3}(\mathbb{R}^2).$$

Consider now a given domain $\Omega \subset \mathbb{R}^2$. *Let* Δ *be its triangulation,* $T := |\Delta|$ *and* $V := \{\underline{x}^0,\underline{x}^1,\ldots,\underline{x}^N\}$ *the set of knots. A spline space is derived by the following procedure ([4],[6]): Each knot* $\underline{x}^j \in V$ *is pulled apart in*

$$\underline{x}^{j,0} := \underline{x}^j,\ \underline{x}^{j,1},\ldots,\underline{x}^{j,n-2}.$$

Further, for any triangle

$$\rho_j := [\underline{x}^{j_0},\underline{x}^{j_1},\underline{x}^{j_2}] \in \Delta,\ j_0 < j_1 < j_2,$$

a set C_{ρ_j} of $\binom{n}{2}$ B-splines supports $K_{j,r}$ is constructed. If we associate with every point $\underline{x}^{j}{}_{m,q}$ *an element* (m,q) *of the lattice* $(0,1,2) \times (0,1,\ldots,n)$, *then the knot sets* $K_{j,r}$ *can be identified with nondescending paths along grid lines from* $(0,0)$ *to* $(2,n-2)$.

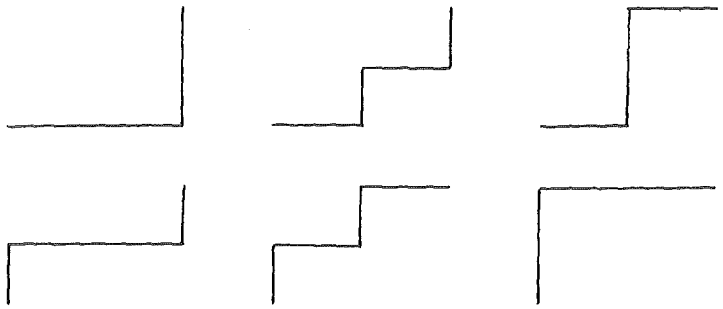


Fig. 1

In the quadratic case this would read

$$C_{\rho_j} = \{K_{j,r},\ r = 1,2,\ldots,6\}$$

where

$$\begin{aligned}
K_{j,1} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_1,0},\ \underline{x}^{j_2,0},\ \underline{x}^{j_2,1},\ \underline{x}^{j_2,2}\},\\
K_{j,2} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_1,0},\ \underline{x}^{j_1,1},\ \underline{x}^{j_2,1},\ \underline{x}^{j_2,2}\},\\
K_{j,3} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_1,0},\ \underline{x}^{j_1,1},\ \underline{x}^{j_1,2},\ \underline{x}^{j_2,2}\},\\
K_{j,4} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_0,1},\ \underline{x}^{j_1,1},\ \underline{x}^{j_2,1},\ \underline{x}^{j_2,2}\},\\
K_{j,5} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_0,1},\ \underline{x}^{j_1,1},\ \underline{x}^{j_1,2},\ \underline{x}^{j_2,2}\},\\
K_{j,6} &= \{\underline{x}^{j_0,0},\ \underline{x}^{j_0,1},\ \underline{x}^{j_0,2},\ \underline{x}^{j_1,2},\ \underline{x}^{j_2,2}\}.
\end{aligned}$$

Let $C := \bigcup_{j=1}^{T} C_{\rho_j}$. *The B-splines*

$$M(\underline{x}|K_{j,r}),\ j = 1,2,\ldots,T;\quad r = 1,2,\ldots,\binom{n}{2}$$

are the basis functions of the spline space

$$S(C) := span\{M(\underline{x}|K_{j,r})\}$$

over C with

$$dim\ S(C) = \binom{n}{2} T$$

Thus any $s \in S(C)$ can be expressed as

$$s(\underline{x}) = \sum_{K \in C} c_k M(\underline{x} | K).$$

2. Evaluation of a spline

There is no known analog of the univariate algorithm that computes a value of a spline by repeatedly forming convex combinations of its B-spline coefficients. Thus the value

$$s(\underline{x}) = \sum_{K \in C} c_K M(\underline{x} | K).$$

can be computed only by computing all nonzero B-splines $M(\underline{x} | \underline{x}^0, \underline{x}^1, \ldots, \underline{x}^n)$ incident at a given $\underline{x}$. A far reaching application of Stokes theorem reveals that ([4],[9])

$$M(\underline{x} | \underline{x}^0, \underline{x}^1, \ldots, \underline{x}^n) = \frac{n}{n-2} \sum_{m=0}^{n} \lambda_m M(\underline{x} | \underline{x}^0, \underline{x}^1, \ldots, \underline{x}^{m-1}, \underline{x}^{m+1}, \ldots, \underline{x}^n), \qquad n > 2$$

where $\underline{x}$ is expressed as any affine combination of $\underline{x}^i$,

$$\sum_{m=0}^{n} \lambda_m \underline{x}^m = \underline{x}, \qquad \sum_{m=0}^{n} \lambda_m = 1.$$

Put

$$M(\underline{x} | \underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2) = \begin{cases} \dfrac{1}{\mathrm{vol}_2([\underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2])}, & \underline{x} \in int[\underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2], \\ 0 & \underline{x} \notin [\underline{x}^i 0, \underline{x}^i 1, \underline{x}^i 2] \end{cases}$$

and general degree B-spline can be computed by the previous recurrence relation, at least for the points that do not lie

on any of the mesh lines. A constant B-spline on its boundary must be defined on a slightly different way. A simple remedy is as follows: let $\underline{\gamma}$ be a direction that is not parallel to any of mesh lines. A constant B-spline at a boundary point $\underline{x}$ is defined by

$$M(\underline{x}|\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2) := M(\underline{x}+O(\underline{\gamma})|\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2)$$

where $O(\underline{\gamma})$ is a small perturbation in the direction $\underline{\gamma}$. In a special case when all the mesh points lie in general position, one can avoid this difficulty by stopping recurrence when $n=3$ since linear B-splines are in this case continuous ([8]).

Quite clear, λ_i have to be nonnegative in order to assure numerical stability. Further, computational complexity implies that as many λ_i as possible should be zero. Thus in practice the recurrence step for a B-spline with support $[K]$ reads

$$M(\underline{x}|K) = (n-2)\sum_{m=0}^{n}\lambda_m(\underline{x}|\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2)M(\underline{x}|K\setminus\{\underline{x}^{i}m\})$$

where λ_m are nonnegative barycentric coordinates of $\underline{x}$ in a triangle $[\underline{x}^{i}0,\underline{x}^{i}1,\underline{x}^{i}2] \subset K$. The choice of i_0, i_1, i_2 is in general not unique since $\underline{x}$ may belong to several triangle parts of $[K]$.

3. The quadratic case

Let us start by defining subregions of the region supported by C_{ρ_j} of a particular interest ([5],[7]), i.e.

$$B_{\rho_j} := \bigcup_{(q_0,q_1,q_2)\in Q}[\underline{x}^{j}0,q_0,\underline{x}^{j}1,q_1,\underline{x}^{j}2,q_2]$$

and

$$Q := \{(q_0,q_1,q_2)\in \mathbb{Z}_+^{3}: 0\leq q_0\leq q_1\leq q_2\leq k\}.$$

Here k denotes degree of a B-spline. One can show that only $\binom{k+2}{2}$ B-splines $M(\underline{x}|K)$, $K \in C$ are incident at a point $\underline{x} \in B_{\rho_j}$, all j. Here B_{ρ_j} is a subregion of ρ_j, a and all of these $\binom{k+2}{2}$ B-splines are constructed over ρ_j. On the other hand, at points $\underline{x} \notin B_{\rho_j}$, all j, the number of B-splines $I_{\underline{x}}(C)$ incident at $\underline{x}$ can be quite large ([7]). This is a consequence of the fact that B-splines from adjecent knot set configurations may overlap the triangle ρ_j. $I_{\underline{x}}(C)$ depends on the original triangulation as well as on the pulling-apart procedure. For k = 2 (and general knot position) the following statement can be proved

$$I_{\underline{x}}(C) = \binom{2+2}{2} = 6, \qquad \underline{x} \in B_{\rho_j} \text{ for some } j,$$

$$I(C) := \max_{\underline{x} \in \Omega} I_{\underline{x}}(C) > 6.$$

There is very little hope that a general strategy which minimizes the computational complexity can be found when $\underline{x} \notin B_{\rho_j}$, all j. We shall therefore restrict ourselves to the case $\underline{x} \in B_{\rho_j}$, for some j. If $\underline{x} \notin B_{\rho_j}$, all j, we shall assume that the recurrence is applied in a straightforward way, and no attempt is made to reduce the number of operations. If the steps in pulling-appart procedure are small, then

$$C \setminus \cup B_{\rho_j}$$

is small compared to

$$\cup B_{\rho_j}$$

and care for evaluation in B_{ρ_j} justified.

Let $\rho_j = [\underline{x}^{j_0}, \underline{x}^{j_1}, \underline{x}^{j_2}]$, with $j_0 < j_1 < j_2$, be a triangle in Δ. For calculating all 6 quadratic B-splines incident in B_{ρ_j}, we propose the recursion scheme, which involves common lower order B-splines. A complete recursion

forest is shown in Fig.2.

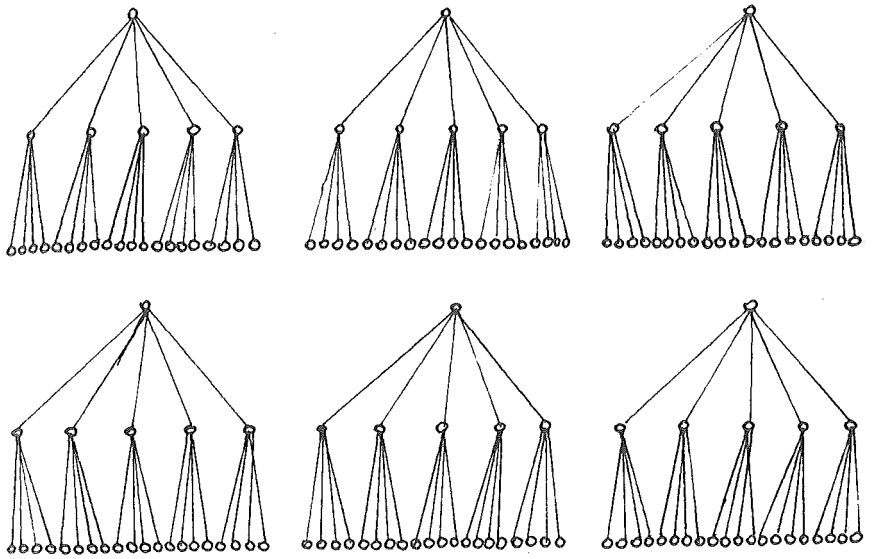


Fig. 2

As already pointed out there is no need to evaluate all of the tree. Even for general k there is enough to compute three tree knots at each tree level (except the root one).

Each quadratic spline is calculated from at most three linear splines that are chosen among five. Also at calculating linear splines we have four possible selections of three constant splines. Thus we have $(\binom{5}{3}\ 4^3)^6$ possible evaluation schemes.

In order to count necessary constant B-spline evaluations let us define a reduced knot set. Let $K \in C_{\rho_j}$ for some j, and denote by K^r the reduced knot set as any of its subsets of cardinality 5 - r. Quite clear, the supports of lower order B-splines in figure 2 are obtained as reduced sets of the root ones. Note also that the reduced set may belong to different tree knots at the same level.

All linear B-splines involved when evaluating the value of a quadratic spline are of type $M(\underline{x}|K^1)$, where K^1 is a reduced knot set consisting of four points. Further, all

constant splines that appear in the scheme are of type $M(\underline{x}|K^2)$ with K^2 a reduced knot set of cardinality 3. A straightforward calculation reveals that there are 24 different reduced sets of cardinality 4 and 37 of cardinality 3. As the full evaluation forest involves 30 sets with 4 points and 120 with 3 points, some reduced knot sets have to appear several times. In fact, an exact upper bound for necessary constant B-spline evaluations can be stated.

Theorem In order to evaluate all $\binom{k+2}{2}$ B-splines of degree k at the point $\underline{x} \in B_{\rho_j}$ it is necessary to compute at most

$$\varphi(k) := \frac{k+1}{3}(5k^2 + 7k + 3)$$

constant B-splines.

The proof is based upon carefull counting of the number of reduced knot sets of order 3. Note that

$$\varphi(0) = 1, \quad \varphi(1) = 10, \quad \varphi(2) = 37, \quad \varphi(3) = 92,$$

etc.

Number of appearence of sets with three points varies from 2 to 8. It is not obvious which reduced knot sets are to be choosen to minimise in general the overall scheme. A heuristic approach that uses more frequently appearing reduced sets can reduce the computational effort significiantly. However, by posing additional requirement on the pulling-apart procedure an optimal algorithm can be found.

If the knots $\underline{x}^{i,q}$, q = 1,2 are chosen in the polygon R_i (the convex polygon which contains $\underline{x}^i$ and is bounded, but not intersected, by the lines passing through the midpoints of any edges belonging to the same triangle with vertex $\underline{x}^i$) 9 linear B-splines vanish in B_{ρ_j}. Using this fact, in [8] an algorithm was presented that computes all 6 C^1 quadratic B-splines by evaluating 6 linear B-splines. The following scheme improves the result to five linear B-splines

evaluations, and all of these five can be computed from 10 constant B-splines. Let K_j^r denotes j-th reduced set at level r of the evaluation forest, counted from the left.

$$M(\underline{x}|K_{j,1}) = 2\lambda_2(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_2,0})M(\underline{x}|K^1_{15})$$

$$M(\underline{x}|K_{j,2}) = 2\,[\lambda_1(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_1,1})M(\underline{x}|K^1_{17}) + \lambda_2(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_1,1})M(\underline{x}|K^1_{15})$$

$$M(\underline{x}|K_{j,3}) = 2\lambda_1(\underline{x}|\underline{x}^{j_0,0},\underline{x}^{j_1,0},\underline{x}^{j_2,2})M(\underline{x}|K^1_{16})$$

$$M(\underline{x}|K_{j,4}) = 2\,[\lambda_0(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,1},\underline{x}^{j_2,1})M(\underline{x}|K^1_{17}) + \lambda_2(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,1},\underline{x}^{j_2,1})M(\underline{x}|K^1_{5})\,]$$

$$M(\underline{x}|K_{j,5}) = 2\,[\lambda_0(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^1_{6}) + \lambda_1(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^1_{5})\,]$$

$$M(\underline{x}|K_{j,6}) = 2\lambda_0(\underline{x}|\underline{x}^{j_0,2},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^1_{6})$$

Furtner, linear splines are computed as

$$M(\underline{x}|K^1_6) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{16}) + \lambda_1(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{5}) + \lambda_2(\underline{x}|\underline{x}^{j_0,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{3})$$

$$M(\underline{x}|K^1_5) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_0,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{5}) + \lambda_1(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_0,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{15}) + \lambda_2(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_0,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{24})\,]$$

$$M(\underline{x}|K^1_{17}) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_2,1},\underline{x}^{j_1,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{15}) + \lambda_1(\underline{x}|\underline{x}^{j_2,1},\underline{x}^{j_1,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{19}) + \lambda_2(\underline{x}|\underline{x}^{j_2,1},\underline{x}^{j_1,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{36})\,]$$

$$M(\underline{x}|K^1_{15}) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_1,0},\underline{x}^{j_2,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{19}) + \lambda_1(\underline{x}|\underline{x}^{j_1,0},\underline{x}^{j_2,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{12}) + \lambda_2(\underline{x}|\underline{x}^{j_1,0},\underline{x}^{j_2,1},\underline{x}^{j_0,0})M(\underline{x}|K^2_{34})\,]$$

$$M(\underline{x}|K^1_{16}) = 3\,[\lambda_0(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{16}) + \lambda_1(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{15}) + \lambda_2(\underline{x}|\underline{x}^{j_1,1},\underline{x}^{j_1,2},\underline{x}^{j_2,2})M(\underline{x}|K^2_{13})\,]$$

For simplicty let us abreviate $\underline{x}^{j_i,q}$ by iq. The sets K_i then read as follows

$K^1_5=\{00,01,11,22\}$ $K^1_6=\{00,01,12,22\}$ $K^1_{15}=\{00,10,21,22\}$
$K^1_{16}=\{00,11,12,22\}$ $K^1_{17}=\{00,11,21,22\}$

and

$K^2_3 = \{00,01,12\}$ $K^2_5 = \{00,01,22\}$ $K^2_{12} = \{00,10,22\}$
$K^2_{13} = \{00,11,12\}$ $K^2_{15} = \{00,11,22\}$ $K^2_{16} = \{00,12,22\}$
$K^2_{19} = \{00,21,22\}$ $K^2_{24} = \{01,11,22\}$ $K^2_{34} = \{10,21,22\}$
$K^2_{36} = \{11,21,22\}$

Observe that the B-spline evaluation is numericaly stable, since at each point $\underline{x}$ in B_{ρ_j} all barycentric coordinates $\lambda_m(\underline{x}|\underline{x}^{j_0,q_0},\underline{x}^{j_1,q_1},\underline{x}^{j_2,q_2})$, $m = 0,1,2$; $0 \leq q_0 \leq q_1 \leq q_2 \leq 2$ are nonnegative.

We now proceed by showing that this evaluation scheme is the best as far as constant B-spline evaluations are concerned.

Theorem Evaluation of all 6 quadratic B-splines, incident at $\underline{x} \in B_{\rho_j}$, requires at least 10 constant B-spline evaluations.

roof.

Supose that only 9 constant B-spline evaluations are eeded.

In the Fig. 3 linear splines that vanish are denoted y □ .

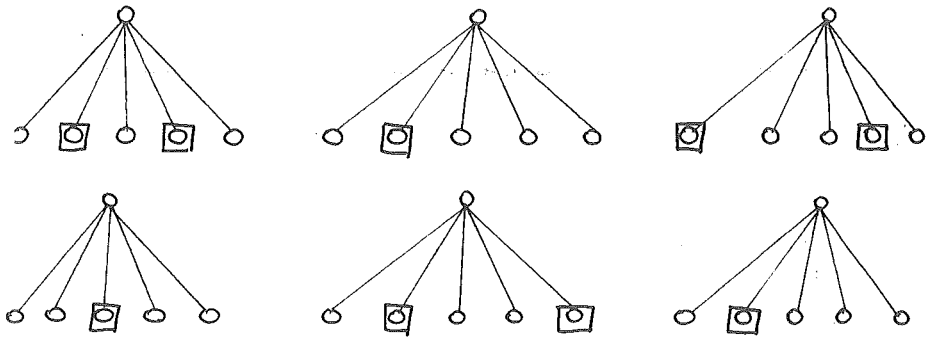


Fig.3

Thus with our 9 constant splines we have to determine :he value of at least one linear B-spline in the first, :hird and fifth tree and at least two in the other trees. Vith a straightforward computer program we can check all :ombinations. It shows up that no combination satisfies requirement presented.■

Since evaluation of all linear B-splines incident at $\underline{x} \in B_{\rho_j}$ requires four constant B-spline evaluations we are :empted to conjucture

$$\binom{k+3}{3}$$

as the lower bound in general case.

4. References

[1] de BOOR, C., Splines as linear combinations of B-splines, in Approximation Theory II, G.G. Lorenz, C.K.Chui & L.L.Schumaker eds., Academic Press, 1976, 1-47.

[2] DAHMEN, W., Polynomials as linear combinations of multivariate B-splines, Math. Z., 169 (1979), 93 - 98.

[3] DAHMEN, W., On multivariate B-splines, SIAM J. Num. Anal., 17 (1980), 179 - 191.

[4] DAHMEN, W., MICCHELLI, C.A., On the linear independence of multivariate B-splines, I:Triangulations of simploids, SIAM J. Num. Anal., 19 (1982), 993 - 1012.

[5] DAHMEN, W., MICCHELLI, C.A., Multivariate splines - a new constructive approach, In Surfaces in Computer Aided Geometric Design, ed. R.E. Barnhill in W. Böhm, North Holland, Amsterdam (1983), 191 - 215.

[6] GMELIG MEYLING, R.H.J., An algorithm for constructing configurations of knots for bivariate B-splines, SIAM J. Num. Anal., xx, xx-xx.

[7] GMELIG MEYLING, R.H.J., On algoritmes and applications for bivariate B-splines, Proc. Conf. Algorithms for the Approximation of Functions and Data, ed. J.C. Mason and M.G.Cox, Shrivenham (1985), xx-xx.

[8] GMELIG MEYLING, R.H.J., Least squares approximation by linear combinations of bivariate B-splines, In Ph.D. Thesis, University of Amsterdam, 1986.

[9] HØLLIG, K., Multivariate splines, SIAM J. Num. Anal., 19 (1982), 1013 - 1031.

ON BOUNDED TENSION INTERPOLATION *

J. KOZAK and M. LOKAR

1. Introduction

The spline in tension goes back to [9], and it was followed by [3], [10], [4], [6], and many others. It was introduced as a (single additional parameter) tool in the shape preserving interpolation, i.e. interpolation that preserves convexity of data. However, it was applied also in other problems (singular perturbation differential equation problems etc.). Thus splines in tension have attracted quite a lot of attention, but they are not very popular in practical computations. Two main reasons for this fact are:

(1) Their use is more time consuming compared to the use of polynomial or rational splines.

(2) The choice od tension parameters is not always apparent.

We shall not bother about (1) since computational complexity for all these spline classes is quite clearly of

ABSTRACT: *Splines in tension are not very popular in practical computations. One of the reasons is obviously the choice of tension parameters that is not always apparent. In this talk, we tackle this problem and we consider the spline in tension as a tool in several approximation problems. In particular, we describe a complete interpolatory tension spline that can be bounded independently of the partition.*

*Supported by Research Council of Slovenija.

the same order, but we shall discuss a remedy for (2): in some applications of splines in tension a natural (and simply computable) choice of tension parameters can be found. However, we shall keep in mind that this talk is far from being meant to show that splines in tension are more useful than for example polynomial ones. They can compete with them only in special circumstances.

Let us recall the definition of a spline in tension. Let $\underline{\tau} := (\tau_i)$ be a strictly increasing partition of $[0,1]$,

$$0 =: \tau_1 < \tau_2 < \ldots < \tau_{n+1} := 1.$$

A spline in tension, with tension parameters $\underline{p} := (p_i)$, $p_i \geq 0$, and breakpoint sequence $\underline{\tau}$ is a function that belongs to

$$S_{4,\underline{\tau},\underline{p}} := C^{(2)}(0,1) \cap N(L)$$

where $L := L_{\underline{p}}$ (and its tension part $M := M_{\underline{p}}$) is piecewise defined by

$$L_{\underline{p}} := \frac{d^4}{dx^4} - \left(\frac{p}{\Delta\tau_i}\right)^2 \frac{d^2}{dx^2} = \frac{d^2}{dx^2}\left(\frac{d^2}{dx^2} - \left(\frac{p}{\Delta\tau_i}\right)^2\right) =:$$

$$=: \frac{d^2}{dx^2}(M_{\underline{p}}) =: \frac{d^2}{dx^2}(M), \qquad \text{on } [\tau_i, \tau_{i+1}).$$

Quite clearly the choice $\underline{p} = (0)$ reproduces the cubic spline space $S_{4,\underline{\tau}}$ as well as $\underline{p} = (\infty)$ the piecewise linear functions $S_{2,\underline{\tau}}$. It is a special case of exponential (and hyperbolic) spline, or more generally L-spline. Some of the ideas presented here could be carried over to more general case. Let us now turn to the examples of practical applications of splines in tension.

2. Shape preserving interpolation

Suppose that a given function g is known at points $\underline{\tau}$. Assume that the partition is extended by

$$\tau_0 := \tau_1, \quad \tau_{n+1} =: \tau_{n+2}$$

in order to simplify the discussion. A shape preserving interpolant is a function that agrees with g at $\underline{\tau}$ and (locally) preserves convexity of the function. However, since g is known only at certain points, an approximation of the second derivative

$$d_j := [\tau_{j-1}, \tau_j, \tau_{j+1}]g,$$

takes over the role. Thus if $d_i d_{i+1} \leq 0$ the second derivative of the interpolant should not change sign in $[\tau_i, \tau_{i+1}]$.

A complete interpolatory spline in tension was the first tool to deal with this (generally nonlinear) problem ([9]). It is easy to see a complete tension interpolatory spline $I_{4,\underline{p}}g$, with interpolatory projector $I_{4,\underline{p}}$ defined by

$$I_{4,\underline{p}} : C(0,1) \longrightarrow S_{4,\underline{\tau},\underline{p}} : f \longrightarrow I_{4,\underline{p}}f := (I_{4,\underline{p}}f|_{\underline{\tau}} = f|_{\underline{\tau}}),$$

is uniquely defined for any tension parameters $\underline{p}$. Thus these additional parameters can be used for smoothing out extraneous inflection points of the interpolant. Quite clearly, a choice $\underline{p} = (\infty)$ would smooth out all inflection points, but produce at most second order approximation. Thus a natural choice ([6]) suggests to choose tension parameters as small as possible in order to preserve the approximation power of the cubic polynomial spline. As it turns out, on each of the subintervals $[\tau_{i-1}, \tau_i]$ it is enough to consider approximate modified derivatives

$$s_0 := \frac{\Delta\tau_i}{\Delta\tau_{i-1}+\Delta\tau_i}([\tau_{i-1}, \tau_i]g - [\tau_i, \tau_{i+1}]g),$$

$$s_1 := \frac{\Delta\tau_i}{\Delta\tau_i+\Delta\tau_{i+1}}([\tau_{i+1}, \tau_{i+2}]g - [\tau_i, \tau_{i+1}]g).$$

If they are of the opposite sign, data indicate that there is no inflexion point, and the interpolant should preserve convexity on the given interval. In this case a quantity $\omega := s_1/(s_1-s_0)$ is studied. A short analysis shows that a spline in tension will not have an inflection point if ω is

trapped in a certain interval. This leads to a simple nonlinear equation for p_i.

3. Tension spline collocation

To start more generaly, consider m-th order linear (to make the discussion simpler) ordinary differential equation for unknown u,

$$Au = f, \text{ on } [0,1]$$

with boundary or initial conditions

$$\beta_i u = c_i, \quad i = 1,2,\ldots,m.$$

Here

$$A := \sum_{i=0}^{m} a_i D^i, \quad a_m \neq 0.$$

We shall assume that the equation has a unique solution u for any f, i.e. there exists a unique Green's function G.

In a simple outfit a collocation approximation to u is constructed as follows:

1) The interval $[0,1]$ is partitioned by a strictly increasing breakpoint sequence $\underline{\tau}$.

2) An approximate solution $u_{\underline{\tau}}$ is looked for as

$$u_{\underline{\tau}} \in C^{(m-1)}(0,1) \cap B^{-1}(P_{k,\underline{\tau}}).$$

where $P_{k,\underline{\tau}}$ denotes as usually the space of piecewise polynomial functions of order k, and B playes the role of an approximation of A. It is usually taken as

$$B = D^m,$$

i.e. the leading part of A. $u_{\underline{\tau}}$ has to satisfy differential equation at collocation points

$$\xi_{ij} \in [\tau_i, \tau_{i+1}), \quad j = 1,2,\ldots,k; \; i = 1,2,\ldots n,$$

$$\xi_{ij} < \xi_{i,j+1}, \text{ all } j$$

and additional conditions

$$\beta_i u_{\underline{\tau}} = c_i, \quad i = 1,2,\ldots,m.$$

Error analysis reveals pointwise error e as

$$e(x) := u(x) - u_{\underline{\tau}}(x) = \int_0^1 G(x,\cdot)(f - Au_{\underline{\tau}})$$

where the Green's function satisifies zero additional conditions,

$$\beta_i G(\cdot,y) = 0, \quad i = 1,2,\ldots,m.$$

The factor $r := f - Au_{\underline{\tau}}$ vanishes at the collocation points ξ_{ij} in $[\tau_i,\tau_{i+1})$. As a consequence, this contributes a factor

$$|\underline{\tau}|^k := \max_i \Delta\tau_i$$

to the L_∞ error bound if $r^{(k)}$ can be properly bounded. The choice of collocation points, based upon orthogonality relations, can further raise the order of approximation up to

$$O(|\underline{\tau}|^{k+m}),$$

and at the breakpoints τ_i even up to

$$O(|\underline{\tau}|^{2k}).$$

For a smooth f, behaviour of r depends on $Au_{\underline{\tau}}$, thus more or less on the quality of approximation of A by $\bar{B}$. Improper choice of $\bar{B}$ could introduce a large error by the method of solution (the choice of collocation functions), regardless of the nature of G that is inherent to the problem, and cannot be avoided. In such a case we can conclude that the nullspaces

$$N(A), \quad N(\bar{B})$$

differ significantly. The usuall choice $B = D^m$ quite clearly fails if the behaviour of solution depends heavily on all of A, not only on its leading term. Consider an example, a second order singular perturbation problem of the form

$$A = -\varepsilon D^2 + a_0 I, \quad 0 < \varepsilon << 1, \; a_0 \neq 0.$$

A solution depends heavily on the sign of a_0, and an approximation B has to take this into account. The best choice would be $B = A$, since then Au_τ reduces to a polynomial. However, such a B cannot be always practically computed. Assume now $a_0 > 0$. A piecewise constant $O(|\tau|)$ approximation

$$B = -\varepsilon D^2 + a_0(\frac{\tau_i + \tau_{i+1}}{2}), \text{ on } [\tau_i, \tau_{i+1}),$$

depends on the sign of a_0 too. Further, $k = 2$ brings us back to the splines in tension as collocation functions, with natural choice of tension parameters

$$p_i = \Delta\tau_i \sqrt{a_0(\frac{\tau_i + \tau_{i+1}}{2})/\varepsilon}, \quad \text{all } i.$$

4. Bounded tension interpolation

It is customary to study approximation power of linear interpolation problems by analysing its bound, expressed as a product of two factors. The first depends on the interpolation scheme, the second on the best approximation of the given function in the space concerned. For a familiar complete spline projector I_4 this inequality, called Lebesgue, would read

$$\| I_4 f - f \| \leq (1 + \| I_4 \|) dist(f, S_{4,\underline{\tau}}).$$

Here $\| \cdot \| := \| \cdot \|_\infty$, and *dist* defined correspondingly. A properly bounded I_4 would quite clearly produce an optimal order approximation. On the other hand, the interpolation error can be bounded also from below by $\| I_4 f \| - \| f \|$. This shows that interpolation error for some

functions f has to be large if $\| I_4 \|$ is large though $\| f \| = 1$. But then one could expect that large norm of interpolation projector would significantly amplify errors in the measured, not accurate data. Thus we can conclude that the bounding of an interpolatory projector has its theoretical as well as practical importance.

It is well known that the projector I_4 can not be bounded independently of $\underline{\tau}$ ([2]), and various restrictions have been imposed on $\underline{\tau}$ in order to produce a bounded projector. One of the approaches (which is also of practical importance for partitions that are close to the geometric one) is to bound I_4 by considering local mesh ratio

$$m_i := \frac{\Delta\tau_i}{\Delta\tau_{i-1}}$$

and its bound

$$m_\Delta := \sup_{|i-j|=1} \frac{\Delta\tau_i}{\Delta\tau_j}.$$

The result that was quite a while looked for can be found in [1]: the complete cubic spline interpolation is bounded (independently of n) if

$$m_\Delta \leq m^* < m_4^* := \frac{3+\sqrt{5}}{2} \quad (m^* \text{ constant}).$$

If the partition $\underline{\tau}$ is too nonuniform one can shift to splines in tension ([7]). The idea is to choose large tension parameters p_i where the partition is changing too rapidly, but to stick to $p_i = 0$ if it is locally uniform. To be precise, a natural choice of $\underline{p}$ is as follows: $I_{4,\underline{p}}$ should be as close to I_4 as possible, but bounded independently of n by a given constant.

The tension parameters are obtained by looking at tension nullsplines. A nullspline $s \in S_{4,\underline{\tau},\underline{p}}$ satisfies

$$s(\tau_i) = 0, \text{ all } i.$$

A tension nullspline is described on each of the intervals by two values which are continuously carried over the interval boundary. A short computation yields,

$$\underline{s}_{i+1} = - A_i \underline{s}_i$$

where

$$A_i := A(m_i, p_i),$$

$$A(p,m) := \begin{pmatrix} \alpha m & 2\frac{\alpha^2 - 1}{\beta} m^2 \\ \frac{\beta}{2} m & \alpha m^2 \end{pmatrix},$$

$$\alpha := \alpha(p) := \frac{p\,\mathrm{ch}(p) - \mathrm{sh}(p)}{\mathrm{sh}(p) - p},$$

$$\beta := \beta(p) := \frac{p^2 \mathrm{sh}(p)}{\mathrm{sh}(p) - p},$$

and

$$\underline{s}_i := \begin{pmatrix} \Delta\tau_{i-1} s'(\tau_i) \\ \frac{\Delta\tau_{i-1}^2}{2} s''(\tau_i) \end{pmatrix}.$$

The choice of $\underline{p}$ has to guarantee that a nullspline increases exponentially in at least one direction. An argument in [5] reduces this to the inequalities (by elements)

$$|A(\frac{1}{m_\Delta}, p_\Delta)| \leq |A(m_i, p_i)| \leq |A(m_\Delta, p_\Delta)|,$$

$$|A^{-1}(m_\Delta, p_\Delta)| \leq |A^{-1}(m_i, p_i)| \leq |A^{-1}(\frac{1}{m_\Delta}, p_\Delta)|.$$

Here, p_Δ is chosen in advance in such a way that the largest eigenvalue λ_2 of the matrix A,

$$\lambda_1 := \lambda_1(m,p) < \lambda_2 := \lambda_2(m,p) < 0$$

satisfies

$$\omega := |\lambda_2(m_\Delta, p_\Delta)| < 1.$$

This assures that the fundamental splines decay by at least a factor ω, and as consequence produces a bounded interpolation. The matrix inequalities are further by a somewhat tedious argument reduced to a single nonlinear equation that determines p_i.

Let us conclude with a brief mention on the approximation power of splines in tension, with the emphasis on its dependence on tension parameters. A general result, with tension parameters hidden in a constant, can be found in [8]. Let us state a refined conclusion ([7]:

Let $f \in C^{(4)}(0,1)$. Then

$$dist_i(f, S_{4,\underline{\tau},\underline{p}}) \leq \| f - I_{4,\underline{p}} f \|_i \leq \frac{\Delta\tau_i^4}{2} C(p_i) \| Mf'' \|_i$$

with

$$C(p) := \frac{1}{p^2}\left(1 - \frac{1}{\mathrm{ch}(p/2)}\right).$$

Here $\| \cdot \|_i$ denotes the sup norm on $[\tau_i, \tau_{i+1}]$, and $dist_i$ is defined correspondingly. Note that $p_i = 0$, all i, reduces the bound to

$$\frac{\Delta\tau_i^4}{16} \| f^{(4)} \|_i,$$

as well as $p_i \longrightarrow \infty$, all i to

$$\frac{\Delta\tau_i^2}{2} \| f'' \|_i.$$

This is (up to the constant) expected.

5. References

1. C. de BOOR: On cubic spline functions which vanish at all knots. Advances in Mathematics 20(1976), 1-17.

2. C. de BOOR: A Practical Guide to Splines. Springer Verlag, New York, 1978.

3. A. CLINE: Scalar- and planar-valued curve fitting in one and two dimensional spaces using splines under tension. Comm. ACM 17(1974), 218-223.

4. J. E. FLAHERTY, W. MATHON: Collocation with polynomial and tension splines for singularly-perturbed boundary value problems. SIAM J. Sci. Stat. Comp. 1(1980), 260-289.

5. S. FRIEDLAND, C.A. MICCHELLI: Bounds of the solutions of difference equations and spline interpolation at knots. Lin. Alg. and its Appl., 20(1978), 219-251.

6. J. KOZAK: Shape preserving approximation. Computers in Industry 7 (1986), 435-440.

7. Y.Y.FENG, J. KOZAK: An approach to the interpolation of nonuniformly spaced data. to appear.

8. L.L. SCHUMAKER: Spline Functions: Basic Theory. John Wiley & Sons, New York, 1981.

9. D.G. SCHWEIKERT: An interpolating curve using a spline in tension. J.Math. Physics 45 (1966), 312-317.

10. H. SPÄTH: Spline-Algorithmen zur Konstruktion glatter Kurven und Flächen. R. Oldenbourg Verlag, München, 1973.

NUMERICAL METHODS IN SEMICONDUCTOR DEVICE SIMULATION

P.A. MARKOWICH, C. SCHMEISER* and S. SELBERHERR

Abstract: The simulation of the electrical behavior of semiconductor devices involves the solution of initial-boundary value problems for a nonlinear elliptic-parabolic system. Two major difficulties in the numerical solution of these problems are discussed:

a) The construction of discretisations is not obvious as the equations are singularly perturbed.
b) The discretised problems are very large systems of nonlinear algebraic equations which have to be solved iteratively.

1. INTRODUCTION

The electrical behavior of a semiconductor device is determined by the flow of two types of free charge carriers, the electrons in the conduction band (density $n(x,t)$) and the defect electrons or holes in the valence band (density $p(x,t)$). Well accepted models for the flow of electrons and holes are the Boltzmann transport equations, but their complexity is prohibitive for the numerical simulation of complicated devices. Perturbation arguments lead to the simplified drift-diffusion approximation of the current densities:

$$J_n = \mu_n(\nabla n + nE) \,, \qquad J_p = -\mu_p(\nabla p - pE) \,. \tag{1.1a}$$

(All the appearing variables and parameters are already in scaled dimensionless form.)

*The work of the second author was supported by "Österreichischer Fonds zur Förderung der wissenschaftlichen Forschung".

In (1.1a) the parameters μ_n, μ_p denote mobilities and E is the electric field which is related to the electrostatic potential ψ by

(1.1b) $\quad E = -\nabla\psi \ .$

Common models for the mobilities depend on n,p,E and the position x.

Maxwell's equations imply the continuity equations

$$\text{(1.1c)} \qquad \begin{aligned} \operatorname{div} J_n - n_t &= R \ , \\ \operatorname{div} J_p + p_t &= -R \end{aligned}$$

and Poisson's equation

(1.1d) $\quad \lambda^2 \Delta\psi = n - p - C(x) \ ,$

where the source term R, called the recombination-generation rate, is the number of electron-hole pairs which are generated $(R<0)$ or disappear $(R>0)$ per unit time. It is usually modelled as a given function of n,p,E and position. The function $C(x)$, the so called doping profile, denotes the concentration of impurity ions. The dimensionless parameter λ is the scaled minimal Debye length and takes small values for realistic semiconductor devices.

The unscaled equations (1.1) are due to Van Roosbroeck [21]. For a derivation from Maxwell's equations and the Boltzmann transport equation see Selberherr [18]. The scaling which leads to (1.1) can be found in Markowich [8].

Mathematically a semiconductor device is given by the doping profile $C(x)$ defined in a bounded domain $\Omega \subseteq \mathbb{R}^3$ which represents the semiconductor part of the device. For the purpose of simulation it often makes sense to reduce the dimension of Ω. Thus, we take $\Omega \subset \mathbb{R}^k$, $k = 1,2$ or 3. The boundary $\partial\Omega$ spli into the union of contact segments $\partial\Omega_D$ where Dirichlet boundary conditions for n,p and ψ are given

(1.2a) $\quad n\big|_{\partial\Omega_D} = n_D \ , \quad p\big|_{\partial\Omega_D} = p_D \ , \quad \psi\big|_{\partial\Omega_D} = \psi_D \ ,$

and the insulating part $\partial\Omega_N$ where the homogeneous Neumann conditions

(1.2b) $\quad (J_n,\nu)\big|_{\partial\Omega_N} = (J_p,\nu)\big|_{\partial\Omega_N} = (E,\nu)\big|_{\partial\Omega_N} = 0$

hold. In (1.2b) ν denotes the outward normal vector of $\partial\Omega$.

Substituting (1.1a) into (1.1c) shows that (1.1) is a system of two parabolic equations for n and p coupled to an elliptic equation for ψ. In order to complete the formulation of an initial-boundary value problem initial conditions for the densities

$$(1.3) \qquad n(x,0) = n_I(x) \ , \quad p(x,0) = p_I(x) \ , \quad x \in \Omega$$

have to be prescribed. The potential at $t = 0$ can be determined by solving Poisson's equation. Several existence and uniqueness results for (1.1)-(1.3) can be found in the literature (see e.g. Mock [12]). Existence results for the corresponding stationary problem are contained in [8] and [12]. Uniqueness cannot be expected in general (see Steinrück [19]).

For the construction and analysis of numerical methods some a priori knowledge of the solution structure is extremely important. This can be gained from a singular perturbation analysis by exploiting the smallness of the parameter λ^2 in (1.1d). In the stationary case such an analysis shows that the solution can be approximated by setting $\lambda = 0$ except in thin layer regions where it varies rapidly (see [8]). For the time dependent problem additionally an initial layer appears (see Ringhofer [14], Szmolyan [20], Markowich [9]). In this paper we will be concerned with the stationary problem. Its analysis is facilitated by the transformation

$$(1.4) \qquad n = e^{\psi}u \ , \quad p = e^{-\psi}v$$

which takes the stationary differential equations to the form

$$\begin{aligned} \lambda^2 \Delta\psi &= e^{\psi}u - e^{-\psi}v - C(x) \\ (1.5) \qquad \mathrm{div}(\mu_n e^{\psi} \nabla u) &= R \\ \mathrm{div}(\mu_p e^{-\psi} \nabla v) &= R \end{aligned}$$

The continuity equations are in self-adjoint form now. Besides u and v are so called slow variables which means that they do not exhibit layer behavior. As opposed to (1.1d) the potential can be determined from the reduced (λ=0) Poisson's equation. Subject to the appropriate boundary conditions each of the equations in (1.5) represents a well posed problem for the variable which appears with the highest differential order, when the other two variables are considered as known.

These properties make it much easier to design numerical methods which are well suited for (1.5) than for the original system. Unfortunately the potential becomes rather large in many applications such that u and v are so out of range that they are impossible to compute with (for different choices of variables and related conditioning questions see Bank et al. [3], Schmeiser et al. [17], Ascher et al. [1]). These facts led to the following approach: Methods are designed and analysed for (1.5). In computations the transformation (1.4) is applied on the discrete level to be able to compute with the original variables ψ,n and p.

2.DISCRETISATIONS

In this section we shall present discretisations for the steady state semiconductor equations which take into account the singular perturbation nature of the problem. The properties of system (1.5) allow for a decoupled approach, where each equation is treated separately.

2.1. Poisson's equation is a semilinear elliptic equation for the potential when u and v are considered to be known. The solution is approximated by a solution of the reduced equation except close to regions of rapid variation of the doping profile and possibly close to the boundaries where the solution varies rapidly. When trying to solve the problem numerically one would expect to be forced to use grids which are fine enough in the regions of rapid variation to resolve the solution structure. For the simulation of complex devices the cost of using such a grid is prohibitive. In order to get around this difficulty, discretisations are used which mimic the above described properties of the continuous problem by the use of lumping for the evaluation of the right hand side. A finite element of finite difference discretisation at node x_i then takes the form

$$\lambda^2(\Delta_h\psi_h)_i = e^{\psi_i}u_i - e^{-\psi_i}v_i - C(x_i) \tag{2.1}$$

where Δ_h is a discretised version of the Laplace-operator (see Markowich [8], Selberherr [18]). The effect of lumping is that the reduced equations in the continuous and the discrete

case are the same. For any discretisation which inherits the stability properties of the continuous operator (maximum principle) the solution structure is similar for the discrete and continuous problems even if a coarse mesh is used. The main difference is that layers in the discrete case may be wider $(O(h))$ than in the continuous case $(O(\lambda))$. This fact will be demonstrated in the following section. It has two effects of major importance. First, even when starting on a very coarse grid adaptive grid refinement will be able to detect the correct solution structure. Second, as the solution is approximated well away from the thin layer regions the approximation error will be small if measured in integral norms although large pointwise errors may occur. The importance of this effect will also be demonstrated in section 3.

2.2. The continuity equations. We shall only deal with the electron continuity equation as the necessary modifications for the hole continuity equation are obvious. Let us first consider the one-dimensional situation. As the variables u and J_n are slow variables - in the language of singular perturbation theory - in this case, the discretisation of

$$J_n' = R \ , \quad J_n = \mu_n e^{\psi} u' \tag{2.2}$$

is not very critical. For simplicity we assume an equidistant grid and replace the first equation at the gridpoint x_i by

$$J_{n,i+1/2} - J_{n,i-1/2} = h R_i \tag{2.3a}$$

where R_i denotes an approximation of the recombination-generation rate at x_i. The second equation is approximated between gridpoints by

$$J_{n,i+1/2} = \mu_{n,i+1/2} (e^{\psi})_{i+1/2} \frac{u_{i+1}-u_i}{h} \ , \tag{2.3b}$$

where the approximation $\mu_{n,i+1/2}$ for μ_n at $\frac{x_i+x_{i+1}}{2}$ depends on the model which is used. For the approximation $(e^{\psi})_{i+1/2}$ two obvious choices are

$$\frac{1}{2}(e^{\psi_i}+e^{\psi_{i+1}}), \quad \exp(\frac{\psi_i+\psi_{i+1}}{2}) \ .$$

A third possibility is obtained by replacing μ_n and J_n by constants and ψ by a linear function in $[x_i, x_{i+1}]$ and solving the second equation in (2.2) explicitely. This results in the approximation

$$(2.3c) \qquad (e^{\psi})_{i+1/2} = \frac{\psi_{i+1} - \psi_i}{e^{-\psi_i} - e^{-\psi_{i+1}}} .$$

This procedure could have also been applied to the equation (1.1a) in the original variable n. The so obtained discretisation which is equivalent to (2.3) is an example of an exponentially fitted method (see Doolan et al. [5]) and bears the names of the engineers Scharfetter and Gummel [16] in the semiconductor device simulation literature.

The difference between the above mentioned discretisations is an unsettled issue from the theoretical point of view, but in practically all of the existing device simulation software the Scharfetter-Gummel scheme is used.

Extensions to finite difference methods in the two- and three-dimensional cases are straightforward (see Selberherrr [18]). Finite element methods which are generalisations of the Scharfett Gummel method to the two-dimensional situation can be found in Buturla and Cottrell [4] and Markowich and Zlamal [10]. It can be shown that the errors only depend on the variation of the current density J_n (see [10], Mock [13]). The drawback in the multidimensional situation is that J_n is not a slow variable in general (see Markowich [8]) which makes it necessary to use fine grids in regions of rapid variation of J_n. However, in most practical situations J_n varies much less than ψ,n and p and the computational effort remains reasonable.

The above error considerations dealt with each equation separately. In order to prove convergence results for the full system one has to assume wellposedness of the problem. Then the error estimates for the single equations can be combined (see [8]).

3.A UNIFORM CONVERGENCE RESULT

When talking about numerical methods for singular perturbation problems uniform convergence means roughly that errors can be estimated independently of the singular perturbation parameter. In particular, errors are even small if the grid ignores layers. Results of this kind can be proven for pointwise errors when using exponentially fitted methods (see Doolan et al. [5]). Such a result cannot be expected for the discretisations of the semiconductor device equations discussed in the preceding section, but this is of minor importance when the goals of device modeling are considered. These goals are basically twofold. One aim is to reveal the solution structure inside the device, the second is to obtain the relation between applied voltages - which enter the Dirichlet boundary conditions - and outflow currents - which are computed by integrals of the current densities along contact segments. Only for the latter part the accuracy of the method is of decisive importance. In this section we prove for a model problem that both aims can be met with reasonable computational effort.

We consider a one-dimensional problem with constant mobilities and vanishing recombination-generation rate. System (1.5) reads

$$
\begin{aligned}
&\lambda^2\psi'' = e^{\psi}u - e^{-\psi}v - C(x) \ , \\
&(e^{\psi}u')' = 0 \ , \\
&(e^{-\psi}v')' = 0
\end{aligned}
\tag{3.1}
$$

in this case. The simulation domain is $\Omega = (0,1)$. System (3.1) is subject to Dirichlet boundary conditions at $x = 0$ and $x = 1$. We consider an equidistant grid on $[0,1]$. Poisson's equation is discretised by using the common three point formula for the approximation of ψ''. The approximate solution ψ_h is obtained by linear interpolation between the gridpoints.

The Scharfetter-Gummel method amounts to replacing ψ by ψ_h in the continuity equations and solving them explicitely because of the assumptions on μ_n, μ_p and R. Problem (3.1) can be written as a fixed point problem by denoting the solutions of the continuity equations for given ψ

$$
\begin{aligned}
u(x) &= u(0) + (u(1)-u(0)) \int_0^x e^{-\psi} \Big/ \int_0^1 e^{-\psi} \,, \\
v(x) &= v(0) + (v(1)-v(0)) \int_0^x e^{\psi} \Big/ \int_0^1 e^{\psi}
\end{aligned} \tag{3.2}
$$

by $u(\psi), v(\psi)$ and the solution of

$$\lambda^2 \phi'' = e^{\phi} u(\psi) - e^{-\phi} v(\psi) - C(x)$$

plus boundary conditions by $\phi = T(\psi)$. A fixed point of the operator T corresponds to a solution.

The discretised problem can be written as

$$
\begin{aligned}
&\frac{\lambda^2}{h^2} (\psi_{i+1} - 2\psi_i + \psi_{i-1}) = e^{\psi_i} u_i - e^{-\psi_i} v_i - C(x_i) \,, \\
&u_h = u(\psi_h) \,, \quad v_h = v(\psi_h) \,.
\end{aligned} \tag{3.3}
$$

Our convergence analysis will be based on the

Lemma 3.1: Let the Frechet derivative of the operator $(I-T)$ at ψ_h be invertible and the inverse be bounded as operator from $L^1(\Omega)$ to $L^1(\Omega)$ independently of λ and h.
Let $\| \psi_h - T(\psi_h) \|_1$ be sufficiently small, where $\|.\|_p$ denotes the L^p-Norm on $(0,1)$.
Then (3.1) has a locally unique solution ψ^* and

$$\| \psi^* - \psi_h \|_1 \leq K_1 \| \psi_h - T(\psi_h) \|_1$$

with K_1 independent of λ and h holds.

The proof is a straightforward application of the implicit function theorem (For similar results see [8],[12]).

Because of Lemma 3.1 we only have to estimate the L^1-Norm of the error in solving Poisson's equation. This is contained in

Lemma 3.2: Let $C(x)$ have a finite number of jump discontinuities in $[0,1]$ and Lipschitz-continuous first derivatives between those points. Let $u(0), u(1), v(0), v(1) > 0$ hold. Then

$$\| \psi_h - T(\psi_h) \|_1 \leq K_2 (\lambda + h)$$

holds with K_2 independent of λ and h.

Outline of a proof:A priori estimates (see [8],[12]) show that

$$e^{\psi_i} u_i + e^{-\psi_i} v_i \geq K > 0$$

holds for the derivative with respect to ψ_i of the right hand side of (3.3). Thus the discrete operator in (3.3) is of inverse monotone type (see Meis-Markowitz [11]). This allows the use of comparison functions for estimates of the solution. Comparison functions can be constructed which are roughly the sum of a solution of the reduced equation and of terms which decay exponentially away from the boundaries and the discontinuities of the doping profile. The L^1-Norm of the decaying terms can be computed and shown to be of the order $O(\lambda+h)$. The argument that the layer terms in the continuous solution are $O(\lambda)$ with respect to the L^1-Norm completes the proof.

A combination of the above lemmata yields the main result of this section

Theorem 3.3: Let the assumptions of the Lemmata 3.1 and 3.2 hold. If the total current density is denoted by $J = J_n + J_p$, the estimate

$$\|\psi^*-\psi_h\|_1 + \|u^*-u_h\|_\infty + \|v^*-v_h\|_\infty + |J-J_h| \leq K_3(\lambda+h)$$

holds with K_3 independent of λ and h.

Proof: The estimate for the error in the potential follows direct from the preceding lemmata. Considering the representation (3.2) for u and v and

$$J_n = (u(1)-u(0)) \Big/ \int_0^1 e^{-\psi}, \quad J_p = (v(0)-v(1)) \Big/ \int_0^1 e^{\psi}$$

for the current densities the proof of the remaining estimates is also immediate.

Supposedly the above result can be extended to one-dimensional problems with less stringent assumptions one the mobilities and the recombination-generation rate. In the multidimensional situation a similar result cannot be expected to hold because layers in the current densities have to be resolved which requires grid-spacings of the order $O(\lambda)$.

4. NONLINEAR ITERATION METHODS

By discretising (1.5) we obtain a large system of nonlinear algebraic equations. Their solution requires the use of apprc priate iteration methods. Although these methods are applied to the discrete problem we discuss them for the continuous eq tions for notational convenience. Assuming again constant mob lities and vanishing recombination-generation we have to solv

$$
\begin{aligned}
&\lambda^2\Delta\psi - e^{\psi}u + e^{-\psi}v + C(x) = b_1 = 0 \ , \\
&\operatorname{div}(e^{\psi}\nabla u) = b_2 = 0 \ , \\
&\operatorname{div}(e^{-\psi}\nabla v) = b_3 = 0 \ .
\end{aligned}
\tag{4.1}
$$

Newton's method for (4.1) reads

$$
\begin{aligned}
&\lambda^2\Delta d\psi - (e^{\psi}u+e^{-\psi}v)d\psi - e^{\psi}du + e^{-\psi}dv = - b_1 \ , \\
&\operatorname{div}(J_n d\psi + e^{\psi}\nabla du) = -b_2 \ , \\
&\operatorname{div}(J_p d\psi + e^{-\psi}\nabla dv) = -b_3 \ .
\end{aligned}
\tag{4.2}
$$

Its application requires the solution of a large linear syste in each iteration step. The computational cost can be reduce by "freezing" the Frechet-derivative for several iteration s For efficient strategies of this kind and their analysis see and Rose [2]. Their concept of approximate Newton methods al for perturbations in the Frechet-derivative. A worthwhile go to find perturbations which decouple the linear system (4.2) certain extent. One method of this kind relies on the assump that the current densities are comparatively small. Obviousl (4.2) is decoupled if J_n and J_p are replaced by zero. Th resulting method amounts to solving the continuity equations given ψ and then the linearized Poisson's equation with the dated u and v in each step. This method was first proposed b Gummel [6]. An alternative which also carries his name is to solve the nonlinear Poisson's equation in each step which ca also be seen as the Picard iteration for the fixed point pro $\psi = T(\psi)$ formulated in the preceding section. Convergence an ses of Gummel's method for small current densities are conta in Markowich [8] and Kerkhoven [7].

When the current densities take values of significant size the convergence of Gummel's method often deteriorates. In view of this situation a different kind of decoupling by approximating the Frechet derivative was introduced in Ringhofer and Schmeiser [15]. Here the singular perturbation character of the linearised problem (4.2) is used. As du and dv are slow variables they are approximated well by the solution of the reduced problem. Thus, we substitute

$$\overline{d\psi} = (-e^{\psi}\overline{du} + e^{-\psi}\overline{dv} + b_1)(e^{\psi}u + e^{-\psi}v)^{-1}$$

into the linearized continuity equations

$$\text{(4.3a)}\qquad \begin{aligned} &\operatorname{div}\Big(\frac{J_n}{e^{\psi}u+e^{-\psi}v}\,(-e^{\psi}\overline{du}+e^{-\psi}\overline{dv}+b_1) + e^{\psi}\nabla\overline{du}\Big) = -b_2\,,\\ &\operatorname{div}\Big(\frac{J_p}{e^{\psi}u+e^{-\psi}v}\,(-e^{\psi}\overline{du}+e^{-\psi}\overline{dv}+b_1) + e^{-\psi}\nabla\overline{dv}\Big) = -b_3\,. \end{aligned}$$

As $d\psi$ is a fast variable, $\overline{d\psi}$ is a good approximation only away from layers. In order to improve on that the full linearised Poisson's equation has to be solved:

$$\text{(4.3b)}\qquad \lambda^2\Delta\widetilde{d\psi} - (e^{\psi}u+e^{-\psi}v)\widetilde{d\psi} - e^{\psi}\overline{du} + e^{-\psi}\overline{dv} = -b_1$$

Instead of the Newton corrections $d\psi, du, dv$ we now use $\widetilde{d\psi}, \overline{du}, \overline{dv}$ In the perturbed problem (4.3) Poisson's equation is decoupled from the continuity equations which are coupled to each other by the terms multiplied by J_n and J_p.

Some of the most important semiconductor devices (e.g.MOSFETs) are so called unipolar devices. They are characterised by the property that only one type of charge carriers (i.e. electrons or holes) contributes significantly to the current flow. This means that one current density (for example J_p) is very small compared to the other. This motivates a further decoupling by replacing J_p by zero in (4.3). The resulting method was proven to converge linearly in [15] with a convergence rate of the form

$$\text{(4.4)}\qquad \text{const}\,(c(\lambda)\,\|J_n\| + \|J_p\|)$$

if the problem is well-posed. In (4.4) $c(\lambda)$ tends to zero as $\lambda \to 0$ and $\|.\|$ denotes a suitable norm. The value of (4.4) is so

small in many applications that the convergence behavior is dominated by the quadratic terms throughout the computations which suggests the use of the term "almost quadratic convergence". The performance of this method was examined in [15] by numerical tests which showed that - compared to Gummel's method - a significant improvement can be achieved.

References:

[1] U.Ascher, P.A.Markowich, C.Schmeiser, H.Steinrück, R.Weiss, Conditioning of the Steady State Semiconductor Device Problem, Techn.Rep.86-18, Comp.Sc., UBC, 1986.

[2] R.E.Banks, D.J.Rose, Global Approximate Newton Methods, Numer.Math.37, 279-295, 1981.

[3] R.E.Bank, D.J.Rose, W.Fichtner, Numerical Methods for Semiconductor Device Simulation, SIAM JSSC 4, No.3, 416-435, 1983.

[4] E.M.Buturla, P.E.Cottrell, Two-Dimensional Finite Element Analysis of Semiconductor Steady Transport Equations, Proc Int.Conf. "Computer Methods in Nonlinear Mechanics", Aust Texas, 512-530, 1974.

[5] E.P.Doolan, J.H.H.Miller, W.H.A.Schilders, Uniform Numeric Methods for Problems with Initial and Boundary Layers, Boc Press, Dublin, 1980.

[6] H.K.Gummel, A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations, IEEE Trans.El.Devices, ED-11, 455-465, 1964.

[7] T.Kerkhoven, Convergence of Gummel's Algorithm for Realist Device Geometries, Proc.NASECODE IV Conf., Boole Press, Dub 1985.

[8] P.A.Markowich, The Stationary Semiconductor Device Equatio Springer-Verlag, Wien, 1986.

[9] P.A.Markowich, Spatial-Temporal Structure of Solutions of the Semiconductor Device Problem, to appear in "Lecture Notes on Appl.Math.", 1987.

[10] P.A.Markowich, M.Zlamal, Inverse-Average-Type Finite Element Discretisations of Self-Adjoint Second Order Elliptic Problems, submitted to Math.Comp., 1987.

[11] T.Meis, U.Marcowitz, Numerische Behandlung Partieller Differentialgleichungen, Springer-Verlag, Berlin, 1978.

[12] M.S.Mock, Analysis of Mathematical Models of Semiconductor Devices, Boole Press, Dublin, 1983.

[13] M.S.Mock, Analysis of a Discretisation Algorithm for Stationary Continuity Equations in Semiconductor Device Models I, COMPEL 2, No.3, 117-139, 1983.

[14] C.A.Ringhofer, The Shape of Solutions to the Basic Semiconductor Equations, to appear in "Lecture Notes on Appl.Math.", 1987.

[15] C.A.Ringhofer, C.Schmeiser, An Approximate Newton Method for the Solution of the Basic Semiconductor Device Equations, submitted to SIAM J.Numer.Anal., 1987.

[16] D.L.Scharfetter, H.K.Gummel, Large Signal Analysis of a Silicon Read Diode Oscillator, IEEE Trans.El. Devices, ED-16, 64-77, 1969.

[17] C.Schmeiser, S.Selberherr, R.Weiss, On Scaling and Norms for Semiconductor Device Simulation, Proc. NASECODE IV Conf., Boole Press, Dublin, 1985.

[18] S.Selberherr, Analysis and Simulation of Semiconductor Devices, Springer-Verlag, Wien, 1984.

[19] H.Steinrück, A Bifurcation Analysis of the One Dimensional Steady State Semiconductor Device Equations, submitted to SIAM J.Appl.Math., 1987.

[20] P.Szmolyan, Ein hyperbolisches System aus der Halbleiterphysik, Thesis, Techn.Univ.Wien, 1987.

[21] W.V.Van Roosbroeck, Theory of Flow of Electrons and Holes in Germanium and Other Semiconductors, Bell Syst.Techn.J. 29, 560-607, 1950.

SOLUTION OF THE DIFFUSION EQUATION IN VLSI PROCESS MODELING BY A NONLINEAR MULTIGRID ALGORITHM

S. MIJALKOVIĆ and N. STOJADINOVIĆ

ABSTRACT: *An application of the nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm for solution of the diffusion equation in VLSI process modeling has been investigated. It is demonstrated that this approach shows high efficiency, which is essentialy independent of physical and numerical parameters of the problem.*

1. INTRODUCTION

For the present underlying physical models of processes used in VLSI process simulation programs, solution of the two-dimensional diffusion equation places heavy demands on computer resources. Moreover, further improvements in kinetic models of point defects, because of their important role in coupling oxidation and diffusion processes, will require at least a tenfold increase in computational throughput for the next generation of VLSI process simulation programs [8]. Therefore, it is clear that more emphasis should be put on numerical approaches that are more efficient than those currently used for diffusion process simulation.

In this view, multigrid methods, well known as the fastest solvers of discretized partial differential equations, seem to be a good choice for this application. Besides their computational efficiency, multigrid methods are fully parallelizable on multiprocessor computers. However, it should be noted that highly efficient and extendable multigrid solvers for more complex problems could be obtained only with a proper choice of a multigrid algorithm and various additional multigrid components [1].

In this paper an application of a nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm for solution of the diffusion equation in VLSI process modeling has been investigated. In the following section mathematical description of diffusion equation and discretization procedure are briefly outlined. The third section describes multigrid algorithm and related multigrid components used. Finally, the two last sections contain analysis of the empirical solution efficiency obtained and some practical examples of actual simulation.

2. PROBLEM DEFINITION

Simulation of the redistribution of impurities in semiconductors under practical processing conditions involves solution of the nonlinear diffusion equation in a domain where one of the boundaries (the silicon--oxide interface) is continually and nonuniformly moving in space as a function of time.

By ignoring diffusion in the oxide (which is justified in many cases), one can consider the diffusion equation [5]

$$\frac{\partial C}{\partial t} - \frac{\partial}{\partial x}(D\frac{\partial C}{\partial x}) - \frac{\partial}{\partial \eta}(D\frac{\partial C}{\partial \eta}) = 0 \tag{1}$$

as two-dimensional initial boundary-value problem in the bounded domain Ω with C being the impurity concentration and D=D(C) the concentration dependent diffusion coefficient.

The boundary conditions are

1) deep in the silicon substrate i.e. at the top of simulated region ($\eta=\eta_1+mU$): $C=C_{min}=10^{13}$ cm^{-3},

2) along the lines of symmetry ($x=0$ and $x=x_1$): $\partial C/\partial n=0$ and

3) on the silicon-oxide interface i.e. at the bottom of simulated region ($\eta=mU$):

$$D\frac{\partial C}{\partial n} - (k-m)\cdot\dot{U}\cdot C\cdot n = 0 \tag{2}$$

where x_1 and η_1 determine the domain extent, $U=U(x,t)$ is the local oxide thickness, k is the segregation coefficient, m is the ratio of silicon thickness consumed to oxide thickness produced and n is the unit vector normal to each boundary.

Since numerical treatment of the problem has been the main goal of the paper, electric field induced flux of impurities as well as coupled impurity diffusion have been excluded for simplicity. Most of physical parameters in (1) and (2) have been modeled according to the program SUPREM [2].

To avoid the problems with discretization at the moving silicon--oxide interface $\eta=B(x,t)=m\cdot U(x,t)$, coordinate transformation [5]

$$y = \eta - B(x,t)$$

which transforms physical domain into the time independent rectangular domain has been used. This yields the following transformed diffusion equation:

$$\frac{\partial C}{\partial t} - \frac{\partial}{\partial x}(D\frac{\partial C}{\partial x}) - (1+B'^2)\frac{\partial}{\partial y}(D\frac{\partial C}{\partial y}) -$$

3)

$$- (B-D\cdot B'')\frac{\partial C}{\partial y} + B'\{\frac{\partial}{\partial x}(D\frac{\partial C}{\partial y}) + \frac{\partial}{\partial y}(D\frac{\partial C}{\partial x})\} = 0$$

ıe boundary condition (2) is converted into

$$D\{(1+B'^2)\frac{\partial C}{\partial y} - B'\frac{\partial C}{\partial x}\} - (k-m)\cdot B\cdot C = 0$$

ıile the other boundary conditions remain unchanged for symmetry
ɜasons.

The multigrid solution of the diffusion equation (1) could be also ɜrformed in physical domain without coordinate transformation [3]. This ɔproach, because of need for special boundary discretization control t the moving oxide-silicon interface and additional modifications in ɹltigrid components used still requires advanced multigrid techniques ɹs the local coordinate transformation [1] near the moving boundary.

The time discretization of the diffusion equation (3) is performed ıy the implicit backward Euler scheme. An automatic time step selection ɔased on Milne´s device [8] has been used. That implies three integration ɜteps in each time step: a crude step and two finer steps with integra-ɜion time half that of the crude step.

The spatial derivatives of (3) are discretized by 9-point central ɟifferences for the second order terms and by upwind differences for the ʃirst order convection terms. For discretization of the Neuman boundary ɔonditions so-called "mirror imaging" [7] has been used. This is the same ɟiscretization as inside the domain substitutuing the missing quantities ɔutside the domain using Neuman boundary conditions and linear interpo-lation.

Two-dimensional ion-implanted concentration profiles based on LSS ʇheory [7] have been used as initial solution for the first time step, ẇhile the following time steps use the results of previous ones as their initial solution.

3. MULTIGRID APPROACH

The discretization of the diffusion equation leads to a nonlinear algebraic system of equations. To solve this system we use a nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm. This approach is in many respects advantageous to linear multigrid method with

Correction Scheme already used for this application [6]. Inherently non-linear, FAS algorithm does not require global linearization of equations. Hence, no extra storage for coefficients of linearized equations is needed and the programing is very convenient. FAS algorithm also gives a natural way to estimate a local truncation error which could be useful for making an efficient stopping criterion for iteration. Finally, in future developments one can benefit from various advanced multigrid techniques as local refinements for grid adaption and local coordinate transformation at the silicon-oxide interface.

The FAS algorithm used employs sequence $\{G_k,H_k\}_{1\leq k\leq M}$ of uniform, non-staggered, rectangulars grids with corresponding meshsizes ($H_{k-1}=2H_k$) where k is the grid level. Regarding the discretized diffusion equation as a discrete eliptic problem $L_MC_M=F_M$ on the finest grid (G_M), it can be solved for the unknown grid function C_M starting with k=M the following recursive procedure

```
procedure FAS (k,ν1,ν2: integer; Ck,Fk: array);
var  j: integer; τk-1,Ck-1,Fk-1: array
begin
  if k=1 then Ck:=L^-1(Fk) else
    begin
      for j:=1 to ν1 do Ck:=Sk(Ck,Fk);
      Ck-1:=R(Ck);
      τk-1:=Lk-1(Ck-1)-R(LkCk);
      Fk-1:=R(Fk)+τk-1;
      FAS(k-1,ν1,ν2,Ck-1,Fk-1);
      Ck:=Ck-1+P(Ck-1-R(Ck));
      for j:=1 to ν2 do Ck:=Sk(Ck,Fk);
    end;
  if (k=M) and (||Fk-Lk(Ck)||>1/3||τk-1||) then
    FAS(k,ν1,ν2,Ck,Fk);
end;
```

For the purpose of smoothing (S_k), succesive-displacement Gauss-Siedel relaxation with lexicographical (LEX) and red-black (RB) ordering of points were tested. RB ordering of points is in some way advantageous because it could readily be fully parallelized. The only linearization required in FAS algorithm is that in smoothing process local to the corresponding grid point

For the purpose of this linearization, so-called "principal linearization" [1], which confines the original form of differential operator has been used.

Normal full weighting (9-point symmetric) formula [1] has been used as the restriction operator (R) which is natural for highly varying grid functions like impurity doping concentration. The prolongation (P) has been performed with a bilinear interpolation. The use of cubic rather than bilinear interpolation has not significantly ameliorated the situation.

On the coarsest grid (G_1), the solution (L^{-1}) is obtained with 5 iterations of S_1 type. The fine-to-coarse defect correction (τ_k) is used to estimate the local truncation error as $\tau_{M-1}/(2^p-1)$ [1] where p is the local approximation order of differential operator. This feature gives an efficient stopping criterion for terminating FAS algorithm.

4. SOLUTION EFFICIENCY

Very important question when the application of a multigrid algorithm is considered is the efficiency of the obtained solution. As a measure of solution efficiency we have considered an average residual error reduction per work unit (WU) i.e. per work equivalent to one relaxation over the finest grid level, which is onwards referred to as convergence rate (ρ_{WU}).

Regarding the generallity and robustness of the algorithm as one of its most significant attributes, our main concerns were the empirical prediction of behaviour of the convergence rate over a wide range of problem parameters and choice of smoothing technique. As a reference for comparation of the FAS algorithm convergence rate, equivalent single-grid (SG) iteration solver on the finest grid level has been used. One time step simulation of highly nonlinear neutral diffusion process from initially implanted arsenic layer has been chosen as an exemplary problem of numerical testing of convergence rates.

The behaviour of residual error L_2 norm for various time steps (Ts1-Ts4) and the different finest grid levels (M_1 and M_2) during solution procedure is shown in Figs.1 and 2, respectively. The dashed lines represent levels of estimated local truncation error norms for the given finest grid level. Relaxation has been performed with RB ordering of points. Table I gives calculated convergence rates for corresponding time steps and grid levels from Figs.1 and 2 with comparation of RB and LEX ordering of points in relaxation.

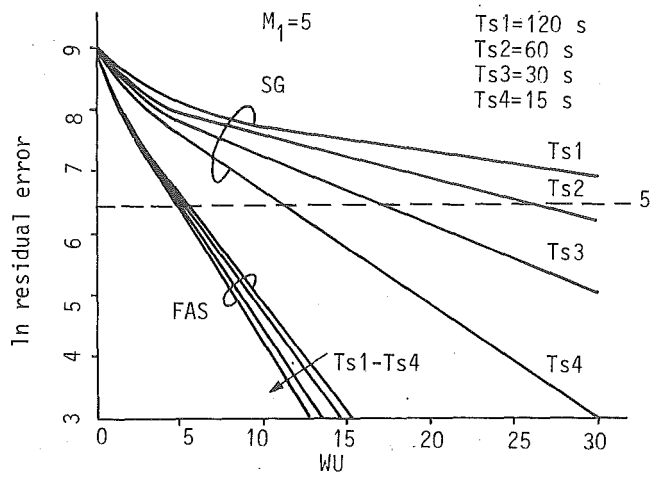


Fig.1 Single-grid and FAS residual restriction for different time step sizes

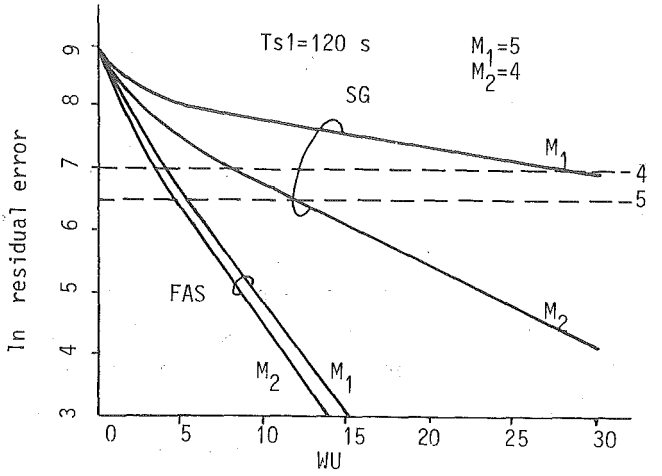


Fig.2 Single-grid and FAS residual restriction for the different finest grid levels

TABLE I Empirical convergence rates (ρ_{WU})

	M_1				Ts1	
	Ts_1	Ts_2	Ts_3	Ts_4	M_1	M_2
SG-RB	0.9019	0.8547	0.7711	0.6512	0.9019	0.6840
SG-LEX	0.9009	0.8574	0.7753	0.6623	0.9009	0.7185
FAS-RB	0.4364	0.4188	0.3987	0.3615	0.4364	0.3768
FAS-LEX	0.4827	0.4554	0.4204	0.3930	0.4827	0.4327

As an empirical observation, it is obvious that FAS algorithm solves a problem to the level of truncation error level in just a few work units which is commonly regarded as "normal" multigrid efficiency [4]. More important fact is that the FAS algorithm convergence rate is almost independent of time step size and choice of the finest grid level which means that for all practical simulation examples the computational cost of FAS algorithm is essentialy problem independent.

5. SAMPLE SIMULATION

Two diffusion process steps typical for fabrication of VLSI NMOS transistor structure are considered as practical examples of simulation. The first process step is the diffusion of boron channel-stop implant during local oxidation (LOCOS). The second process is a high-temperature anneal following arsenic implant for the source/drain region formation.

The boron channel-stop implant through a predefined field-oxide region, with the dose $5 \cdot 10^{12}$ cm^{-2} at 100 keV is followed by the 240 min field oxidation at $1000^{o}C$ in H_2O. On the other hand, arsenic was implanted with the dose 10^{16} cm^{-2} at 150 keV and driven-in for 15 min in neutral ambient at $1000^{o}C$.

The contour plots of the impurity concentrations at various stages of processing are shown in Figs.3 through 6. The boron channel-stop implant and source/drain arsenic implant distributions after the completion of the ion-implantation process step are shown in Figs.3 and 5, respectively. The silicon-nitride mask for boron channel-stop implantation extends from x=0 to x=1μm and for arsenic implantation from x=0 to x=0.25μm. Fig.4 shows the final boron profile after the local oxidation step. Initial oxide thickness for this process step was 0.05μm. Fig.5 shows the final arsenic distribution after the high-temperature anneal.

6. CONCLUSION

The main objective of the next generation VLSI process simulation programs is to reduce the design duration while simultaneously increasing the efficiency in achieving well-designed processes. Special attention should be paid on simulation of diffusion processes which are the most crucial and time consuming simulation steps.

In this paper we have presented an application of a nonlinear multigrid method with Full Approximation Scheme (FAS) algorithm for solution

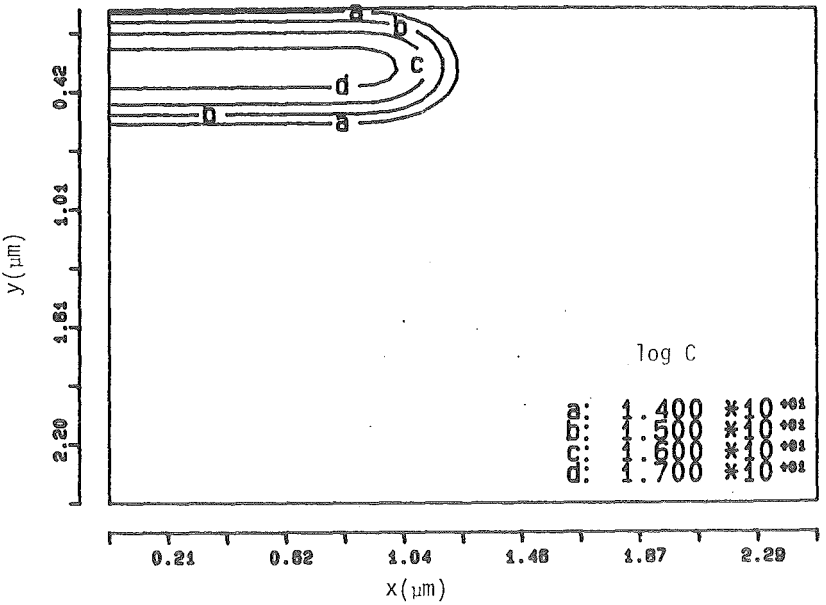


Fig.3 Boron channel-stop implant distribution

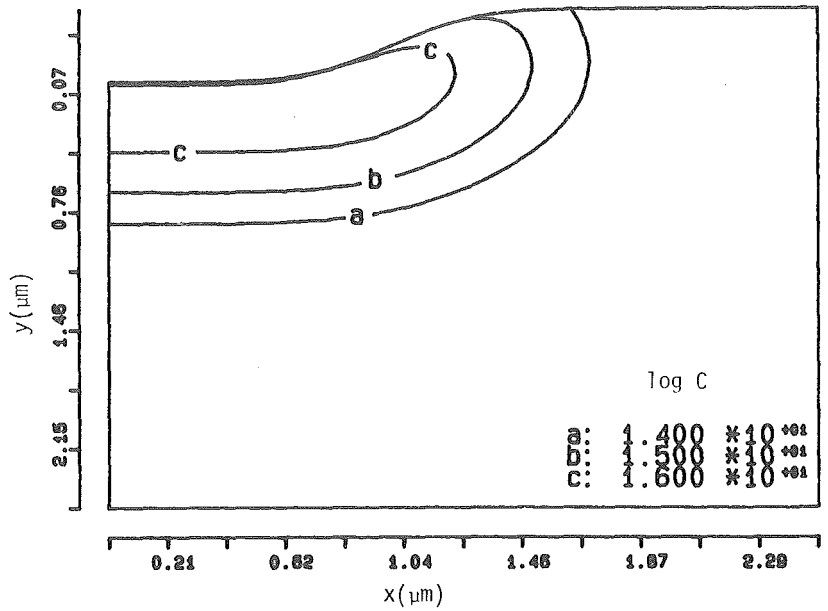


Fig.4 Boron distribution after local oxidation process step

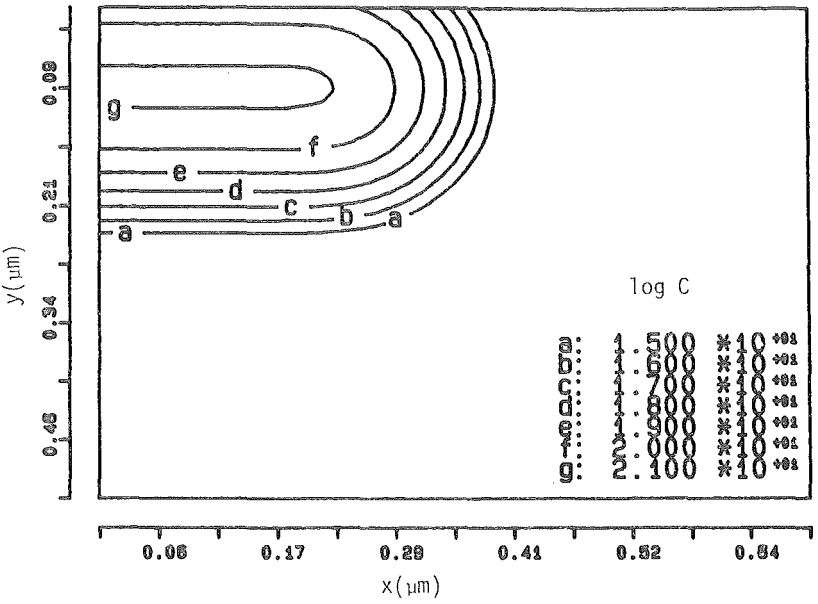


Fig.5 Arsenic source/drain implant distribution

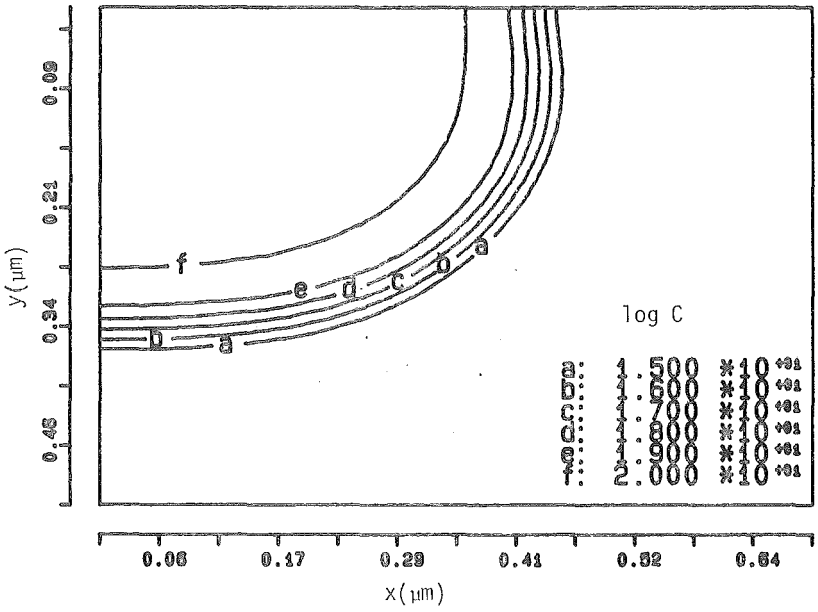


Fig.6 Arsenic distribution after high temperature anneal process step

of diffusion equation in VLSI process modeling. It has been demonstrated that nonlinear FAS algorithm shows high computational efficiency of solution which is almost independent of problem parameters in most practical applications.

Having in mind these features and possible extensions of FAS algorithm for more complex problems this numerical approach could be very effective for the next generation of process simulation programs.

7. REFERENCES

1. A. BRANDT: Multigrid Technique: 1984 Guide with Application to Fluid Dynamics. GMD-Studien Nr.85. Bonn, W. Germany.

2. C. HO, D. PLUMMER, S. HANSEN and R. DUTTON: VLSI Process Modeling-- SUPREM III. IEEE Trans. Electron Device ED-30 (1983), 1438-1453.

3. W. JOPPICH: A Multigrid Method for Solving the Nonlinear Diffusion Equation on a Time-Dependent Domain Using Rectangular Grids in Cartesian Coordinates. In: Proceedings of NASECODE V Conference. Dublin: Boole Press 1987.

4. S. MIJALKOVIĆ and N. STOJADINOVIĆ: Multigrid Method: An Efficient Numerical Tool in VLSI Process Modeling. In: Proceedings of First International Conference on Computer Technology, Systems and Application. Hamburg, 1987, 508-509.

5. B. PENUMALLI: A Comprehensive Two-Dimensional VLSI Process Simulation Program, BICEPS. IEEE Trans. Electron Device ED-30 (1983), 986-992.

6. A. SEIDL: A Multigrid Method for Solution of the Diffusion Equation in VLSI Process Modeling. IEEE Trans. Electron Device ED-30 (1983), 999-1004.

7. S. SELBERHERR: Analysis and Simulation of Semiconductor Devices. Springer-Verlag, Wien 1984.

8. H. YEAGER and R. DUTTON: An Approach to Solving Multiparticle Diffusion Exhibiting Nonlinear Stiff Coupling. IEEE Trans. Electron Device ED-32 (1985) 1964-1975.

CONSTRUCTION OF s - ORTHOGONAL POLYNOMIALS AND TURÁN QUADRATURE FORMULAE

GRADIMIR V. MILOVANOVIĆ

ABSTRACT: *A connection between Turán quadratures and s-orthogonal polynomials with respect to a nonnegative measure on the real line $\mathbb{R}$ is given. Using a discretized Stieltjes procedure and the Newton-Kantorowič method, an iterative method with quadratic convergence for the construction of s-orthogonal polynomials is formulated. Some numerical examples are included. Finally, some considerations about Turán quadrature formulae with Chebyshev measure are given.*

1. INTRODUCTION

In 1950 P. Turán investigated numerical quadratures of the type

$$\int_{-1}^{1} f(t)\,dt = \sum_{\nu=1}^{n} \sum_{i=0}^{k-1} A_{i,\nu} f^{(i)}(\tau_\nu) + R_{n,k}(f), \tag{1.1}$$

where

$$A_{i,\nu} = \int_{-1}^{1} \ell_{\nu,i}(t)\,dt \qquad (\nu=1,\ldots,n;\ i=0,1,\ldots,k-1)$$

and $\ell_{\nu,i}(t)$ are the fundamental functions of Hermite interpolation. The $A_{i,\nu}$ are Cotes numbers of higher order. The formula (1.1) is exact if f is a polynomial of degree at most kn-1 and the points $-1 \leq \tau_1 < \tau_2 < \ldots < \tau_n \leq 1$ are arbitrary.

For k=1 the formula (1.1), i.e.,

$$\int_{-1}^{1} f(t)\,dt = \sum_{\nu=1}^{n} A_{0,\nu} f(\tau_\nu) + R_{n,1}(f),$$

can be exact for all polynomials of degree $\leq 2n-1$ if the nodes τ_ν are the zeros of the Legendre polynomial P_n. That is the well-known Gauss-Legendre quadrature.

Because of the theorem of Gauss it is natural to ask whether knots τ_ν can be chosen so that the quadrature formula (1.1) will be exact for polynomials of degree not exceeding $(k+1)n-1$. P. Turán [17] showed that the answer is negative for $k=2$, and for $k=3$ it is positive. He proved that the knots τ_ν should be chosen as the zeros of the monic polynomial $\pi_n^*(t) = t^n + \dots$ which minimizes the following integral

$$\int_{-1}^{1} \pi_n(t)^4 dt,$$

where $\pi_n(t) = t^n + a_{n-1}t^{n-1} + \dots + a_1 t + a_0$.

More generally, the answer is negative for even, and positive for odd k, and then τ_ν are the zeros of the polynomial minimizing

$$\int_{-1}^{1} \pi_n(t)^{k+1} dt. \tag{1.2}$$

For $k=1$, π_n is the monic Legendre polynomial P_n.

Because of the above, we put $k=2s+1$. Instead of (1.1), it is also interesting to investigate the analogous formula with a weight function $t \to p(t)$,

$$\int_{-1}^{1} f(t)p(t)dt = \sum_{i=0}^{2s} \sum_{\nu=1}^{n} A_{i,\nu} f^{(i)}(\tau_\nu) + R(f),$$

or more generally, with some nonnegative measure $d\lambda(t)$ on the real line $\mathbb{R}$,

$$\int_{\mathbb{R}} f(t)d\lambda(t) = \sum_{i=0}^{2s} \sum_{\nu=1}^{n} A_{i,\nu} f^{(i)}(\tau_\nu) + R(f). \tag{1.3}$$

This paper is organized as follows. In Section 2 we give a connection between Turán quadratures and s-orthogonal polynomials, which were studied extensively by several Italian mathematicians [12], [7], [13], [14]. Also, in this section we mention a recent method of Vincenti [20] for the computation of the coefficients of s-orthogonal polynomials with respect to an even function. In Section 3 we develop a new method for the numerical construction of s-orthogonal polynomials with respect to an arbitrary weight function. Numerical examples are given in Section 4. Section 5 deals with Turán quadratures with Chebyshev measure.

2. TURAN QUADRATURES AND s-ORTHOGONAL POLYNOMIALS

We consider the Turán quadrature formula (1.3), where $d\lambda(t)$ is a nonnegative measure on the real line $\mathbb{R}$, with compact or infinite support, for which all moments $\mu_k = \int_{\mathbb{R}} t^k d\lambda(t)$, $k=0,1,\ldots$, exist and are finite, and $\mu_0 > 0$. The formula (1.3) must be exact for all polynomials of degree at most $2(s+1)n-1$. The role of the integral (1.2) is taken over by

$$F = \int_{\mathbb{R}} \pi_n(t)^{2s+2} d\lambda(t),$$

where $F \equiv F(a_0,\ldots,a_{n-1})$, $\pi_n(t) = \sum_{k=0}^{n} a_k t^k$, $a_n = 1$. In order to minimize F we must have

$$\int_{\mathbb{R}} \pi_n(t)^{2s+1} t^k d\lambda(t) = 0, \quad k=0,1,\ldots,n-1. \tag{2.1}$$

Usually, instead of $\pi_n(t)$ we write $P_{s,n}(t)$.

The case $d\lambda(t) = p(t)dt$ on $[a,b]$ has been considered by the Italian mathematicians A.Ossicini [12], A.Ghizzetti and A.Ossicini [7], S.Guerra [8], [9]. It is known that there exists a unique

$P_{s,n}(t) = \prod_{\nu=1}^{n} (t-\tau_\nu)$, whose zeros τ_ν are real, distinct and located in the interior of the interval [a,b]. These polynomials are known as s-orthogonal (or s-self associated) polynomials in the interval [a,b] with respect to the weight function p.

For s=0 we have the standard case of orthogonal polynomials. The case when s>0 is very difficult. It requires the use of a method with special numerical treatment.

Recently G.Vincenti [20] has considered an iterative process to compute the coefficients of s-orthogonal polynomials in a special case when the interval [a,b] is symmetric with respect to origin, say, [-b,b], and the weight function p is an even function p(-t)=p(t). Then $P_{s,n}(-t) = (-1)^n P_{s,n}(t)$. He considered two cases: when n is even and when n is odd.

In the first case n=2m, $P_{s,n}(t) = \sum_{i=0}^{m} a_i t^{2m-2i}$, $a_0 = 1$. From (2.1) Vincenti obtained a nonlinear system of equations of the form

$$\sum_{i=0}^{m} C_{m+r-i} a_i = -C_{m+r} \qquad (r=0,1,\ldots,m-1),$$

where

$$C_j^{(0)} = \int_0^b p(t) t^{2j} dt, \qquad C_j^{(h)} = \sum_{p,q=0}^{m} C_{j+2m-p-q}^{(h-1)} a_p a_q,$$

and $C_j^{(s)} \equiv C_j$. Then he has solved this system by some iterative method like Newton's method. For n=2m+1, a similar system of equations was obtained.

Vincenti applied his process to the Legendre case. When n and s increase, the process becomes ill-conditioned. So, the author gave numerical results in the following cases: n=2,3, $1 \leq s \leq 10$; n=4,5, $1 \leq s \leq 5$; n=6,7, $1 \leq s \leq 3$; n=8,9, $1 \leq s \leq 2$; n=10,11, s=1. The results were obtained with 18 correct decimal digits,

but using an arithmetic with 36 decimal digits.

From (2.1) we can see that this procedure needs the first $2(s+1)n$ moments of the weight function: $\mu_0, \mu_1, \cdots, \mu_{2(s+1)n-1}$. We see that $c_j^{(0)} = \mu_{2j}/2$. Of course, in this special case, the moments of odd order are zero. Here, we have a nonlinear map $V_{n,s}: \mathbb{R}^{2(s+1)n} \to \mathbb{R}^n$, given by $[\mu_0, \mu_1, \ldots, \mu_{2(s+1)n-1}]^T \to [a_0, a_1, \ldots, a_{n-1}]^T$. The problem itself is highly sensitive to small perturbations in the moments, so that any algorithm which theoretically solves the problem using the moments will be subject to severe growth of errors when executed in an arithmetic of finite precision ([4],[5]). It would be useful to find a numerical condition number of the map $V_{n,s}$, but that will not be our aim here.

3. CONSTRUCTION OF s-ORTHOGONAL POLYNOMIALS

In this section we will give a stable procedure for the numerical construction of s-orthogonal polynomials with respect to $d\lambda(t)$ on $\mathbb{R}$. Namely, we will reduce our problem to the standard theory of orthogonal polynomials, and then we will use the Stieltjes procedure ([3],[5]). The main idea is an interpretation of the "orthogonality conditions" (2.1), i.e.,

$$\int_{\mathbb{R}} \pi_n(t)\, t^k\, \pi_n(t)^{2s}\, d\lambda(t) = 0, \qquad k=0,1,\ldots,n-1.$$

For given n and s, we put $d\mu(t) = d\mu^{s,n}(t) = (\pi_n(t))^{2s} d\lambda(t)$. These conditions can be interpreted as

$$\int_{\mathbb{R}} \pi_k^{s,n}(t)\, t^\nu\, d\mu(t) = 0, \qquad \nu=0,1,\ldots,k-1,$$

where $(\pi_k^{s,n})$ is a sequence of monic orthogonal polynomials with respect to the new measure $d\mu(t)$. Of course, $P_{s,n}(.) = \pi_n^{s,n}(.)$.

As we can see, the polynomials $\pi_k^{s,n}$, $k=0,1,\ldots$, are implicitly defined, because the measure $d\mu(t)$ depends of $\pi_n^{s,n}(t)$. The general class of such polynomials was introduced by H.Engels (see [2, pp. 214-226]).

We will write only $\pi_k(.)$ instead of $\pi_k^{s,n}(.)$. These polynomials satisfy a three-term recurrence relation

$$\pi_{k+1} = (t-\alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \quad k=0,1,\ldots, \tag{3.1}$$

$$\pi_{-1}(t) = 0, \quad \pi_0(t) = 1,$$

where, because of orthogonality,

$$\alpha_k = \alpha_k(s,n) = \frac{\langle t\pi_k, \pi_k\rangle}{\langle \pi_k, \pi_k\rangle} = \frac{\int_{\mathbb{R}} t\pi_k^2(t)\,d\mu(t)}{\int_{\mathbb{R}} \pi_k^2(t)\,d\mu(t)}, \tag{3.2}$$

$$\beta_k = \beta_k(s,n) = \frac{\langle \pi_k, \pi_k\rangle}{\langle \pi_{k-1}, \pi_{k-1}\rangle} = \frac{\int_{\mathbb{R}} \pi_k^2(t)\,d\mu(t)}{\int_{\mathbb{R}} \pi_{k-1}^2(t)\,d\mu(t)},$$

and, for example, $\beta_0 = \int_{\mathbb{R}} d\mu(t)$.

The coefficients α_k and β_k are the fundamental quantities in the constructive theory of orthogonal polynomials. They provide a compact way of representing orthogonal polynomials, requiring only a linear array of parameters. The coefficients of orthogonal polynomials, or their zeros, in contrast need two-dimensional arrays.

Finding the coefficients α_k, β_k $(k=0,1,\ldots,n-1)$ gives us access to the first n+1 orthogonal polynomials $\pi_0, \pi_1, \ldots, \pi_n$. Of course, for a given n, we are interested only in the last of them i.e., $\pi_n (\equiv \pi_n^{s,n})$. So, for $n=0,1,\ldots$, the diagonal (boxed) elements

In the following table are our s-orthogonal polynomials $\pi_n^{s,n}$.

TABLE 3.1

n	$d\mu^{s,n}(t)$	Orthogonal Polynomials
0	$(\pi_0^{s,0}(t))^{2s} d\lambda(t)$	$\boxed{\pi_0^{s,0}}$
1	$(\pi_1^{s,1}(t))^{2s} d\lambda(t)$	$\pi_0^{s,1}$ $\boxed{\pi_1^{s,1}}$
2	$(\pi_2^{s,2}(t))^{2s} d\lambda(t)$	$\pi_0^{s,2}$ $\pi_1^{s,2}$ $\boxed{\pi_2^{s,2}}$
3	$(\pi_3^{s,3}(t))^{2s} d\lambda(t)$	$\pi_0^{s,3}$ $\pi_1^{s,3}$ $\pi_2^{s,3}$ $\boxed{\pi_3^{s,3}}$
⋮		

A stable procedure for finding the coefficients α_k, β_k is the discretized Stieltjes procedure, especially for infinite intervals of orthogonality (see Gautschi [5], and Gautschi, Milovanović [6]). Unfortunately, in our case this procedure cannot be used directly, because the measure $d\mu(t)$ involves an unknown polynomial $\pi_n^{s,n}$. Consequently, we consider the system of nonlinear equations

$$f_0 \equiv \beta_0 - \int_{\mathbf{R}} \pi_n^{2s}(t)\, d\lambda(t) = 0,$$

$$f_{2k+1} \equiv \int_{\mathbf{R}} (\alpha_k - t)\pi_k^2(t)\pi_n^{2s}(t)\, d\lambda(t) = 0, \quad k=0,1,\ldots,n-1, \tag{3.3}$$

$$f_{2k} \equiv \int_{\mathbf{R}} (\beta_k \pi_{k-1}^2(t) - \pi_k^2(t))\pi_n^{2s}(t)\, d\lambda(t) = 0, \quad k=1,\ldots,n-1,$$

which follows from (3.2).

Let x be a $(2n)$-dimensional column vector with components $\alpha_0, \beta_0, \ldots, \alpha_{n-1}, \beta_{n-1}$ and $f(x)$ a $(2n)$-dimensional vector with components $f_0, f_1, \ldots, f_{2n-1}$, given by (3.3). If $W = W(x)$ is the corresponding Jacobi matrix of $f(x)$, then we can apply Newton-Kantorovič's method

$$x^{[\nu+1]} = x^{[\nu]} - W^{-1}(x^{[\nu]})\, f(x^{[\nu]}), \qquad \nu = 0,1,\ldots, \tag{3.4}$$

for determining the coefficients of the recurrence relation (3.1) Starting with a reasonable good approximation $x^{[0]}$, the convergence of the method (3.4) is quadratic.

It is interesting that the elements of Jacobi matrix can be easily computed in the following way:

First, we have to determine the partial derivatives $a_{k,i} = \frac{\partial \pi_k}{\partial \alpha_i}$ and $b_{k,i} = \frac{\partial \pi_k}{\partial \beta_i}$. Differentiating the recurrence relation (3.1) with respect to α_i and β_i we obtain

$$a_{k+1,i} = (t - \alpha_k) a_{k,i} - \beta_k a_{k-1,i} ,$$

and

$$b_{k+1,i} = (t - \alpha_k) b_{k,i} - \beta_k b_{k-1,i} ,$$

where

$$a_{k,i} = 0, \quad b_{k,i} = 0, \quad k \le i,$$

$$a_{i+1,i} = -\pi_i(t), \quad b_{i+1,i} = -\pi_{i-1}(t).$$

These relations are the same as those for π_k, but with other initial values. The elements of the Jacobi matrix are

$$\frac{\partial f_{2k+1}}{\partial \alpha_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t) \left[(\alpha_k - t) p_{k,i}(t) + \frac{1}{2} \delta_{ki} \pi_k^2(t) \pi_n(t) \right] d\lambda(t),$$

$$\frac{\partial f_{2k+1}}{\partial \beta_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t)(\alpha_k - t) q_{k,i}(t) d\lambda(t),$$

3.5)
$$\frac{\partial f_{2k}}{\partial \alpha_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t)(\beta_k p_{k-1,i}(t) - p_{k,i}(t)) d\lambda(t),$$

$$\frac{\partial f_{2k}}{\partial \beta_i} = 2 \int_{\mathbb{R}} \pi_n^{2s-1}(t)\{(\beta_k q_{k-1,i}(t) - q_{k,i}(t)) + \frac{1}{2}\delta_{ki}\pi_{k-1}^2(t)\pi_n(t)\} d\lambda(t),$$

here $p_{k,i}(t) = \pi_k(t)(a_{k,i}\pi_n(t) + sa_{n,i}\pi_k(t))$ and $q_{k,i}(t) = k(t)(b_{k,i}\pi_n(t) + sb_{n,i}\pi_k(t))$, and δ_{ki} is Kronecker's delta.

All of the above integrals in (3.3) and (3.5) can be found xactly, except for rounding errors, by using a Gauss-Christoffel uadrature formula with respect to the measure $d\lambda(t)$,

3.6)
$$\int_{\mathbb{R}} g(t) d\lambda(t) = \sum_{k=1}^{N} A_k^{(N)} g(\tau_k^{(N)}) + R_N(g),$$

aking $N=(s+1)n$ knots. This formula is exact for all polynomials f degree at most $2N-1 = 2(s+1)n - 1 = 2(n-1) + 2ns + 1$.

Thus, for all calculations we use only the fundamental three-erm recurrence relation and the Gauss-Christoffel quadrature 3.6). As initial values $\alpha_k^{[0]} = \alpha_k^{[0]}(s,n)$ and $\beta_k^{[0]} = \beta_k^{[0]}(s,n)$ we ta-e the values obtained for n-1, i.e. $\alpha_k^{[0]} = \alpha_k(s,n-1)$, $\beta_k^{[0]} = \beta_k(s,n-1)$, ≦ n-2. For α_{n-1} and β_{n-1} we use the corresponding extrapolated alues.

In the case n=1 we solve the equation

$$\Phi(\alpha_0) = \Phi(\alpha_0(s,1)) = \int_{\mathbb{R}} (t-\alpha_0)^{2s+1} d\lambda(t) = 0,$$

and then determine

$$\beta_0 = \beta_0(s,1) = \int_{\mathbb{R}} (t-\alpha_0)^{2s} d\lambda(t).$$

4. NUMERICAL EXAMPLES

We will consider two examples, involving Laguerre and Legendre measures.

Example 4.1. $d\lambda(t) = e^{-t}dt$ on $(0,\infty)$.

Using the presented method, we determined the recursion coefficients $\alpha_k(s,n)$ and $\beta_k(s,n)$, $k=0,1,\ldots,n-1$, for $s = 1(1)5$ and $n = 1(1)10$. These coefficients and zeros of $\pi_n^{s,n}$, $\tau_k(s,n)$, $k=1,\ldots,n$, for some selected values of s and n, are given in Table 4.1. Numbers in parentheses denote decimal exponents.The zeros $\tau_k(s,n)$, $k=1,\ldots,n$, were obtained as eigenvalues of the (symmetric tridiagonal) Jacobi matrix

$$J_n = \begin{bmatrix} \alpha_0(s,n) & \sqrt{\beta_1(s,n)} & & & O \\ \sqrt{\beta_1(s,n)} & \alpha_1(s,n) & \sqrt{\beta_2(s,n)} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \sqrt{\beta_{n-1}(s,n)} \\ O & & & \sqrt{\beta_{n-1}(s,n)} & \alpha_{n-1}(s,n) \end{bmatrix},$$

using the QR algorithm.

Example 4.2. $d\lambda(t) = dt$ on $(-1,1)$. In this (Legendre) cas the coefficients $\alpha_k(s,n)$ are equal to zero, so the computatio can be simplified. The system of equations (3.3) becomes

$$g_0 = f_0 = \beta_0 - \int_{-1}^{1} \pi_n^{2s}(t)dt = 0,$$

$$g_k = f_{2k} = \int_{-1}^{1} (\beta_k \pi_{k-1}^2(t) - \pi_k^2(t))\pi_n^{2s}(t)dt = 0, \quad k=1,\ldots,n$$

TABLE 4.1

s,n)	k	$\alpha_k(s,n)$	$\beta_k(s,n)$	$\tau_{k+1}(s,n)$
(1,5)	0	1.53297437454020(0)	1.95429735674308(6)	3.8619211523014(-1)
	1	5.58879530809235(0)	3.09769990936949(0)	2.5326808971664(0)
	2	9.67825960904726(0)	1.44873867444755(1)	6.8055964648137(0)
	3	1.38195768909663(1)	3.44094720328124(1)	1.3770543148954(1)
	4	1.79144743230187(1)	6.31554867162230(1)	2.5039067879500(1)
1,10)	0	1.51947559720794(0)	1.15245141095965(18)	1.9845989648554(-1)
	1	5.54285984605682(0)	3.04910058102535(0)	1.2852724641604(0)
	2	9.57433648078956(0)	1.42156585447179(1)	3.3633782573586(0)
	3	1.36134600304330(1)	3.35373242373804(1)	6.4866460030154(0)
	4	1.76626196755034(1)	6.10680235778704(1)	1.0743607524688(1)
	5	2.17262963633088(1)	9.68925818155147(1)	1.6274303555431(1)
	6	2.58125737578751(1)	1.41154607302825(2)	2.3303521691882(1)
	7	2.99352863173826(1)	1.94114444641607(2)	3.2216061440735(1)
	8	3.40937486293287(1)	2.56171126328495(2)	4.3764898673766(1)
	9	3.80755964787433(1)	3.26547484073315(2)	5.9920103669108(1)
(2,5)	0	3.06241261660323(0)	1.11900724691562(16)	5.1108081782716(-1)
	1	8.17357215072018(0)	6.27220780166492(0)	3.6504048515689(0)
	2	1.43542025111386(1)	3.14187808183856(1)	1.0011553444478(1)
	3	2.06411614818251(1)	7.61775799352481(1)	2.0452776123775(1)
	4	2.68361238086797(1)	1.41467716850165(2)	3.7441657331318(1)
(3,5)	0	2.58905931144849(0)	5.71776101144993(27)	6.3593164870754(-1)
	1	1.07564139072170(1)	1.05185172722828(1)	4.7669589415140(0)
	2	1.90289971948242(1)	5.47855138833478(1)	1.3215882166030(1)
	3	2.74628973371076(1)	1.34292199752058(2)	2.7133552841620(1)
	4	3.57594914086306(1)	2.50922121763235(2)	4.9844533561357(1)
(4,5)	0	3.11368201971988(0)	6.65045548992180(40)	7.6048752765420(-1)
	1	1.33381242208130(1)	1.58333974393260(1)	5.8827138815968(0)
	2	2.37031258589862(1)	8.45825858503624(1)	1.6419218171525(1)
	3	3.42845650702239(1)	2.08746777684076(2)	3.3813401673707(1)
	4	4.46834996793661(1)	3.91510787488863(2)	6.2247175594627(1)
(5,5)	0	3.63680292296229(0)	8.46508537128994(54)	8.8474548516636(-1)
	1	1.59190911806156(1)	2.22147113900203(1)	6.9978980073417(0)
	2	2.83768214565758(1)	1.20806800183997(2)	1.9621882995226(1)
	3	4.11061379266411(1)	2.99537220959448(2)	4.0492610101210(1)
	4	5.36078046528124(1)	5.63228814211952(2)	7.4649521550663(1)

Table 4.2 shows the numerical results for s = 1,3,5 and n = 3,5,10. The corresponding zeros $\tau_\nu(s,n)$, k=1,...,n, are given in Table 4.3.

TABLE 4.2

n	ν	$\beta_\nu(1,n)$	$\beta_\nu(3,n)$	$\beta_\nu(5,n)$
3	0	0.483864899809040(-1)	0.999799077102820(-4)	0.284169237312933(-6)
	1	0.396390615424778	0.438361519822241	0.455125737914133
	2	0.266920571579793	0.262372968797798	0.259637334393080
5	0	0.313354730979678(-2)	0.264465724288258(-7)	0.301618113315945(-13)
	1	0.397514379556632	0.440125755974452	0.456936553362545
	2	0.266421480435867	0.261489083023563	0.258693332791772
	3	0.256509353896241	0.254475851257394	0.253414689828449
	4	0.253674592138278	0.252629769731300	0.252061944536419
10	0	0.314536690060498(-5)	0.261903853328827(-16)	0.290667534992279(-27)
	1	0.398771414276302	0.442152192689833	0.459032427879297
	2	0.266409589288295	0.261261065487811	0.258382986818575
	3	0.256307280251967	0.254101849534999	0.253013674028616
	4	0.253361155621508	0.252167886595534	0.251600165348871
	5	0.252110174900276	0.251373736923891	0.251025244228691
	6	0.251467087710631	0.250973891152692	0.250737300372080
	7	0.251096334167793	0.250747641830448	0.250575234680807
	8	0.250866757894766	0.250611009696396	0.250478257846294
	9	0.250718964459874	0.250526857803099	0.250419693077896

TABLE 4.3

n	ν	$\tau_\nu(1,n)$	$\tau_\nu(3,n)$	$\tau_\nu(5,n)$
3	1,3	±0.81443918557776	±0.83709885235857	±0.84543661637477
	2	0.	0.	0.
5	1,5	±0.92711786960989	±0.93810619284349	±0.94197468869998
	2,4	±0.56086741916164	±0.57330378590709	±0.57774579736053
	3	0.	0.	0.
10	1,10	±0.98066259593659	±0.98398991804138	±0.98512298236202
	2,9	±0.87750022098482	±0.88396182054293	±0.88618806147381
	3,8	±0.69262442514005	±0.69957700233546	±0.70197668437523
	4,7	±0.44320099195064	±0.44838741280314	±0.45017897460267
	5,6	±0.15247058767942	±0.15437687188524	±0.15503560566469

5. TURÁN QUADRATURES WITH CHEBYSHEV WEIGHT

Now, we will consider again the quadrature formula of Turán (1.3). If we define ω_ν, by

$$\omega_\nu(t) = \left(\frac{\pi_n(t)}{t-\tau_\nu}\right)^{2s+1}, \quad \nu=1,\dots,n,$$

where $\pi_n(t)=\pi_n^{s,n}(t)$ and $\tau_\nu=\tau_\nu(s,n)$, then the coefficients $A_{i,\nu}$ in Turán quadrature (1.3) can be expressed in the form [16]

$$A_{i,\nu} = \frac{1}{i!(2s-i)!}\left[D^{2s-i}\,\frac{1}{\omega_\nu(t)}\int_{\mathbb{R}}\frac{\pi_n(x)^{2s+1}-\pi_n(t)^{2s+1}}{x-t}\,d\lambda(x)\right]_{t=\tau_\nu},$$

where D is the standard differentiation operator. Especially, for i=2s, we have

$$A_{2s,\nu} = \frac{1}{(2s)!\,(\pi_n'(\tau_\nu))^{2s+1}}\int_{\mathbb{R}}\frac{\pi_n(x)^{2s+1}}{t-\tau_\nu}\,d\lambda(x),$$

i.e.,

$$A_{2s,\nu} = \frac{B_\nu^{(s)}}{(2s)!\,(\pi_n'(\tau_\nu))^{2s}}, \quad \nu=1,\dots,n, \tag{5.1}$$

where $B_\nu^{(s)}$ are the Christoffel numbers of the following quadrature (with respect to the measure $d\mu(t)=\pi_n^{2s}(t)d\lambda(t)$)

$$\int_{\mathbb{R}} g(t)\,d\mu(t) = \sum_{\nu=1}^{n} B_\nu^{(s)} g(\tau_\nu) + R_n(g), \quad R_n(\mathbb{P}_{2n-1})=0, \tag{5.2}$$

So we have $A_{2s,\nu}>0$.

The expressions for the other coefficients (i<2s) become very complicated.

For the numerical calculation we can use a triangular system of linear equations obtained from the formula (1.3) by replacing f

with the Newton polynomials: $1,\ t-\tau_1,\ \ldots,\ (t-\tau_1)^{2s+1}$,

$(t-\tau_1)^{2s+1}(t-\tau_2),\ldots,\ (t-\tau_1)^{2s+1}(t-\tau_2)^{2s+1}\ldots(t-\tau_n)^{2s}$.

Particularly interesting is the case of the Chebyshev weight

$$p(t) = (1-t^2)^{-1/2}.$$

In 1930, S. Bernstein [1] showed that $2^{1-n}T_n(t)$ minimizes all integrals of the form

$$\int_{-1}^{1} \frac{|\pi_n(t)|^{k+1}}{\sqrt{1-t^2}}\,dt, \qquad k \geq 0.$$

So the Turán-Chebyshev formula

$$(5.3)\qquad \int_{-1}^{1} (1-t^2)^{-1/2} f(t)\,dt = \sum_{i=0}^{2s}\sum_{\nu=1}^{n} A_{i,\nu} f^{(i)}(\tau_\nu) + R(f),$$

with $\tau_\nu = \cos\frac{(2\nu-1)\pi}{2n}$, $\nu=1,\ldots,n$, is exact for polynomials of degree not exceeding $2(s+1)n-1$. Turán has stated a problem of explicit determination of $A_{i,\nu}$ and its asymptotic behavior as $n\to\infty$ (Problem XXVI in [18]). In this regard, Micchelli and Rivlin ([11]) have proved the following characterization: If $f\in\mathbb{P}_{2(s+1)n-1}$ then

$$\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}\,dt = \frac{\pi}{n}\left\{\sum_{\nu=1}^{n} f(\tau_\nu) + \sum_{j=1}^{s} \alpha_j f'[\tau_1^{2j},\ldots,\tau_n^{2j}]\right\},$$

where

$$\alpha_j = (-1)^j \frac{\binom{-1/2}{j}}{2j\,4^{(n-1)j}}, \qquad j=1,2,\ldots,$$

and $g[y_1^r,\ldots,y_m^r]$ designate the devided difference of the function g, where each y_j is repeated r times.

For s=1, the solution of the Turán problem XXVI is given by

$$A_{0,\nu} = \frac{\pi}{n}\ ,\ A_{1,\nu} = -\frac{\pi\tau_\nu}{4n^3}\ ,\quad A_{2,\nu} = \frac{\pi}{4n^3}(1-\tau_\nu^2)\,.$$

In 1975 R.D. Riess [15],and in 1984 A.K. Varma[19], using very different methods, obtained the explicit solution of the Turán problem for s=2:

$$A_{0,\nu} = \frac{\pi}{n}\ ,\quad A_{1,\nu} = -\frac{\pi\tau_\nu}{64n^5}(20n^2-1)\,,\ A_{2,\nu} = \frac{\pi}{64n^5}[3+(20n^2-7)(1-\tau_\nu^2)],$$

$$A_{3,\nu} = -\frac{6\pi\tau_\nu}{64n^5}(1-\tau_\nu^2)\,,\quad A_{4,\nu} = \frac{\pi}{64n^5}(1-\tau_\nu^2)^2\,.$$

Notice that (5.1), for the Chebyshev weight, reduces to

$$A_{2s,\nu} = \frac{\pi}{4^s n^{2s+1}(s!)^2}(1-\tau_\nu^2)^s\,,\qquad \nu=1,\dots,n.$$

One simple answer to Turán question was given by O. Kis [10]. His result can be stated in the following form: If g is an even trigonometric polynomial of degree at most 2(s+1)n-1, then

$$\int_0^\pi g(\theta)\,d\theta = \frac{\pi}{n(s!)^2}\sum_{j=0}^{s}\frac{S_j}{4^j n^{2j}}\sum_{\nu=1}^{n} g^{(2j)}\left(\frac{2\nu-1}{2n}\pi\right),$$

where S_{s-j} $(j=0,1,\dots,s)$ denotes the symmetric elementary polynomials with respect to the numbers $1^2, 2^2, \dots, s^2$, i.e.,

$$S_s = 1,\quad S_{s-1} = 1^2+2^2+\dots+s^2,\ \dots,\ S_0 = 1^2\cdot 2^2\cdots s^2.$$

Consequently,

$$\int_{-1}^{1}(1-t^2)^{-1/2}f(t)\,dt = \frac{\pi}{n(s!)^2}\sum_{j=0}^{s}\frac{S_j}{4^j n^{2j}}\sum_{\nu=1}^{n}\left[D^{2j}f(\cos\theta)\right]_{\theta=\frac{2\nu-1}{2n}\pi}.$$

Using the expansion

$$D^{2k} f(\cos\theta) = \sum_{i=1}^{2k} a_{k,i}(t) f^{(i)}(t), \qquad \cos\theta = t, \quad k>0,$$

where the functions $a_{i,j} \equiv a_{i,j}(t)$ are given recursively by

$$a_{k+1,1} = (1-t^2)a'_{k,1} - ta_{k,1},$$

$$a_{k+1,2} = (1-t^2)a'_{k,2} - ta_{k,2} + 2(1-t^2)a_{k,1} - ta_{k,1},$$

$$a_{k+1,i} = (1-t^2)a'_{k,i} - ta_{k,i} + 2(1-t^2)a_{k,i-1} - ta_{k,i-1} + (1-t^2)a_{k,i-2}$$

$$(k = 3, \ldots, 2k),$$

$$a_{k+1,2k+1} = 2(1-t^2)a_{k,2k} - ta_{k,2k} + (1-t^2)a_{k,2k-1},$$

$$a_{k+1,2k+2} = (1-t^2)a_{k,2k},$$

with $a_{1,1} = -t$ and $a_{1,2} = 1-t^2$, we obtain the formula (5.3). For example, when s=3, we have

$$A_{0,\nu} = \frac{\pi}{n}, \quad A_{1,\nu} = \frac{\pi\tau_\nu}{2304n^7}(784n^4+56n^2-1),$$

$$A_{2,\nu} = \frac{\pi}{2304n^7}\left\{(784n^4-392n^2+31)(1-\tau_\nu^2) + 168n^2-15\right\},$$

$$A_{3,\nu} = -\frac{\pi\tau_\nu}{2304n^7}\left\{(336n^2-89)(1-\tau_\nu^2) + 15\right\},$$

$$A_{4,\nu} = \frac{\pi}{2304n^7}\left\{(56n^2-65)(1-\tau_\nu^2)^2 + 45(1-\tau_\nu^2)\right\},$$

$$A_{5,\nu} = \frac{\pi\tau_\nu}{2304n^7}\left\{674(1-\tau_\nu^2)^2 - 240(1-\tau_\nu^2)\right\}, \quad A_{6,\nu} = \frac{\pi}{2304n^7}(1-\tau_\nu^2)^3.$$

To conclude, we mention the corresponding formula (5.2) for the Chebyshev weight,

$$(5.4) \qquad \int_{-1}^{1} g(t)\frac{\hat{T}_n^{2s}(t)}{\sqrt{1-t^2}}dt = \frac{\pi}{4^s n}\binom{2s}{s}\sum_{\nu=1}^{n} g(\tau_\nu) + R_n(g),$$

where $\tau_\nu = \cos(2\nu-1)\frac{\pi}{2n}$, $\nu=1,\dots,n$. Note that all weights are equal, that is, the formula (5.4) is one of Chebyshev type.

The last formula can be reduced to a "cosinus" formula

$$\int_0^\pi f(\cos x)\cos^{2s}(nx)\,dx = \frac{\pi}{n4^s}\binom{2s}{s}\sum_{\nu=1}^{n} f(\cos(2\nu-1)\tfrac{\pi}{2n}) + R_n(f),$$

where $R_n(f) \equiv 0$ if $f\in\mathbb{P}_{2(s+1)n-1}$.

Acknowledgment. The author is grateful to Professor Walter Gautschi for his careful reading of the paper and useful suggestions for better formulations of the material.

REFERENCES

1. S. BERNSTEIN: Sur les polynomes orthogonaux relatifs à un segment fini. J. Math. Pures Appl. (9)9(1930), 127-177.

2. H. ENGELS: Numerical quadrature and cubature. Academic Press, London, 1980.

3. W. GAUTSCHI: Construction of Gauss-Christoffel quadrature formulas. Math. Comp. 22(1968), 251-270.

4. W. GAUTSCHI: A survey of Gauss-Christoffel quadrature formulae. In: E.B. Christoffel - The Influence of his Work on Mathematics and the Physical Sciences (P.L. Butzer and F. Fehèr, eds.), Birkhäuser Verlag, Basel, 1981, pp. 72-147.

5. W. GAUTSCHI: On generating orthogonal polynomials. SIAM J. Sci. Statist. Comput. 3(1982), 289-317.

6. W. GAUTSCHI and G.V. MILOVANOVIĆ: Gaussian quadrature involving Einstein and Fermi functions with an application to summation of series. Math. Comp. 44(1985), 177-190.

7. A. GHIZZETTI and A. OSSICINI: Su un nuovo tipo di sviluppo di una funzione in serie di polinomi. Rend. Accad. Naz. Lincei (8)43(1967), 21-29.

8. S. GUERRA: Polinomi generati da successioni peso e teoremi di rappresentazione di una funzione in serie di tali polinomi. Rend. Ist. Mat. Univ. Trieste 8(1976), 172-194.

9. S. GUERRA: Su un determinante collegato ad un sistema di polinomi ortogonali. Rend. Ist. Mat. Univ. Trieste 10(1978), 66-79.

10. O. KIS: Remark on mechanical quadrature *(Russian)*. Acta Math. Acad. Sci. Hungar. 8(1957), 473-476.

11. C.A. MICCHELLI and T.J. RIVLIN: Turán formulae highest precision quadrature rules for Chebyshev coefficients. IBM J. Res. Develop. 16(1972), 372-379.

12. A. OSSICINI: Costruzione di formule di quadratura di tipo Gaussiano. Ann. Mat. Pura Appl. (4)72(1966), 213-238.

13. A. OSSICINI and F. ROSATI: Funzioni caratteristiche nelle formule di quadratura gaussiane con nodi multipli. Bull. Un. Mat. Ital. (4)11(1975), 224-237.

14. A. OSSICINI and F. ROSATI: Sulla convergenza dei funzionali ipergaussiani. Rend. Mat. (6)11(1978), 97-108.

15. R.D. RIESS: Gauss-Turán quadratures of Chebyshev type and error formulae. Computing 15(1975), 173-179.

16. D.D. STANCU: Asupra unor formule generale de integrare numerica. Acad. R. P. Romîne. Stud. Cerc. Mat. 9(1958), 209-216.

17. P. TURÁN: On the theory of the mechanical quadrature. Acta Sci. Math. Szeged. 12(1950), 30-37.

18. P. TURÁN: On some open problems of approximation theory. J. Approx. Theory 29(1980), 23-85.

19. A.K. VARMA: On optimal quadrature formulae. Studia Sci. Math. Hungar. 19(1984), 437-446.

20. G. VINCENTI: On the computation of the coefficients of s-orthogonal polynomials. SIAM J. Numer. Anal. 23(1986), 1290-1294.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

ON SOME PARALLEL HIGHER-ORDER METHODS OF HALLEY'S TYPE FOR FINDING MULTIPLE POLYNOMIAL ZEROS

M.S. PETKOVIĆ and L.V. STEFANOVIĆ

ABSTRACT: *Using Newton's and Halley's corrections, some modifications of the iterative method for the simultaneous finding multiple complex zeros of a polynomial, based on the Halley-like algorithm, are obtained. The convergence order of the proposed methods is five and six, respectively. Further improvements of these methods are performed by applying the Gauss-Seidel approach. The lower bounds of the R-order of convergence for the accelerated (single-step) methods are also given. Faster convergence is attained without additional calculations which makes the proposed methods be very efficient. The considered iterative procedures are illustrated numerically in the example of an algebraic equation.*

1. ITERATION SCHEMES

Consider a monic polynomial of degre $n \geq 3$

$$P(z) = \prod_{i=1}^{\nu} (z - r_i)^{\mu_i}$$

with real or complex zeros $r_1,\dots,r_\nu$ having the order of multiplicity $\mu_1,\dots,\mu_\nu$ respectively, where $\mu_1 + \cdots + \mu_\nu = n$ $(\nu > 1)$. Let $z \in \mathbb{C}$ and

$$f_k(z) = \frac{P^{(k)}(z)}{P(z)} \qquad (k=1,2) ,$$

$$g(z) = \frac{f_1(z)}{2}\left(1 + \frac{1}{\mu}\right) - \frac{f_2(z)}{2 f_1(z)} ,$$

$$S_i(a,b) = \frac{1}{\mu_i}\left[\sum_{j=1}^{i-1} \mu_j (z-a_j)^{-1} + \sum_{j=i+1}^{\nu} \mu_j (z-b_j)^{-1}\right]^2 + \sum_{j=1}^{i-1} \mu_j (z-b_j)^{-2} + \sum_{j=i+1}^{\nu} \mu_j (z-b_j)^{-2} ,$$

where $a = (a_1,\dots,a_\nu)^T$ and $b = (b_1,\dots,b_\nu)^T$ are some vectors. In particular, according to the above, we have, for example,

$$S_i(a,a) = \frac{1}{\mu_i}\left[\sum_{\substack{j=1\\ j\neq i}}^{\nu} \mu_j (z-a_j)^{-1}\right]^2 + \sum_{\substack{j=1\\ j\neq i}}^{\nu} \mu_j (z-a_j)^{-2} .$$

Using the Bell's polynomials, X. Wang and S. Zheng have derived in [10] the following relations

$$(1) \qquad r_i = z - \left[g(z) - \frac{1}{2f_1(z)} S_i(r,r)\right]^{-1} \qquad (i=1,\ldots,\nu) ,$$

where $r = (r_1,\ldots,r_\nu)^T$ and $\mu=\mu_i$.

Assume that reasonably good approximations $z_1,\ldots,z_\nu$ of the zeros $r_1,\ldots,r_\nu$ were found. Letting $z=z_i$ and $r_i:=\hat{z}_i$ in (1), where $\hat{z}_i$ is the new approximation of the zero r_i, and taking certain approximations of r_j in S_i on the right-hand side of the relation (1), we obtain some iterative methods for simultaneous finding all zeros of the polynomial P.

We shall first define

$$N(z) = \mu/f_1(z) \qquad \text{(the Newton's correction)} ,$$

$$H(z) = 1/g(z) = 2\left[\frac{f_2(z)}{f_1(z)} - (1+\frac{1}{\mu})f_1(z)\right]^{-1}$$

(the Halley's correction)

and introduce the vectors

$$z = (z_1,\ldots,z_\nu)^T \qquad \text{(the former approximations)} ,$$

$$z_N = (z_{N,1},\ldots,z_{N,\nu})^T , \quad z_{N,i} = z_i - N(z_i)$$

(the Newton's approximations),

$$z_H = (z_{H,1},\ldots,z_{H,\nu})^T , \quad z_{H,i} = z_i - H(z_i)$$

(the Halley's approximations),

$$\hat{z} = (\hat{z}_1,\ldots,\hat{z}_\nu)^T \qquad \text{(the new approximations)} .$$

In calculating the approximations $z_{N,i}$ and $z_{H,i}$, as well in all formulas where the function $g(z)$ appears, one has to take $\mu=\mu_i$.

(TS) For $r_j:=z_j$ $(j\neq i)$ we obtain the total-step iteration (TS)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(z,z) \right]^{-1} \qquad (i=1,\ldots,\nu). \tag{2}$$

This method has been discussed in [10] (see, also [11]) as a special case obtained from the family of iterative methods. The formula (2) can be rewritten in the form

$$\hat{z}_i = z_i - H(z_i) \left[1 - \frac{H(z_i)}{2f_1(z_i)} S_i(z,z) \right]^{-1},$$

wherefrom we observe the similarity of the iterative method (2) (which has the convergence order equal to *four*) with the Halley's method (with *cubic* convergence)

$$\hat{z}_i = z_i - H(z_i)$$

for improvement of multiple zero r_i (see [2]). Thus, the correction term in the form of summs provides (i) the increase of convergence order and (ii) the determination of all zeros of a polynomial. Furthermore, we note that the formula (2) is more complicated to the square root iteration (which also has the convergence order equal to *four*, see, e.g. [7]), but (2) does not require the extraction of a root and the selection of appropriate value (of two values) of the square root.

(SS) Let $r_j:=\hat{z}_j$ $(j<i)$ and $r_j:=z_j$ $(j>i)$ (the Gauss-Seidel approach), then we obtain from (1) the single-step iteration (SS)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(\hat{z},z) \right]^{-1} \qquad (i=1,\ldots,\nu). \tag{3}$$

(TSN) Letting $r_j := z_{N,j} = z_j - N(z_j)$ $(j \neq i)$ in (1), one obtains the total-step method with Newton's correction (TSN)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(z_N, z_N) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{4}$$

(SSN) Substituting $r_j := \hat{z}_j$ $(j < i)$, $r_j := z_{N,j} = z_j - N(z_j)$ $(j > i)$ in (1), we obtain the single-step method with Newton's correction (SSN)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(\hat{z}, z_N) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{5}$$

(TSH) Putting $r_j := z_{H,j} = z_j - H(z_j)$ $(j \neq i)$ in (1), similar as for TSN method, we obtain the total-step method with Halley's correction (TSH)

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(z_H, z_H) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{6}$$

(SSH) TSH method can be accelerated using the Gauss-Seidel approach: setting $r_j := \hat{z}_j$ $(j < i)$, $r_j := z_{H,j} = z_j - H(z_j)$ $(j > i)$ in (1), we get the single-step method with Halley's correction

$$\hat{z}_i = z_i - \left[g(z_i) - \frac{1}{2f_1(z_i)} S_i(\hat{z}, z_H) \right]^{-1} \quad (i=1,\ldots,\nu). \tag{7}$$

2. CONVERGENCE ORDER

In this section we shall consider the convergence order of the iterative schemes (2)-(7). For the single-step methods, where the new approximations are used immediately they become available, we shall apply the definition of the R-order of convergence (see [4]). The R-order of convergence of the iterative process IP with the limit point $r = (r_1,\ldots,r_\nu)^T$, where $r_1,\ldots,r_\nu$ are the polynomial zeros, will be denoted by $O_R(IP,r)$.

Let $u_i^{(m)}$ be a multiple of $|z_i^{(m)} - r_i|$ $(i=1,\ldots,\nu)$, where $m=0,1,\ldots$ is the iteration index. Using the technique applied in [1], [6] or that presented in [7], it can be shown that the iterative methods (2)-(7) belong to a class of iterative simul-

taneous methods for which the following relations can be derived under suitable initial conditions

$$(8)\qquad u_i^{(m+1)} = \frac{1}{\nu - 1}\, u_i^{(m)p}\left(\alpha \sum_{j<i} u_j^{(m+1)} + \sum_{j>i} u_j^{(m)q}\right)$$

$$(i=1,\ldots,\nu;\ p,q\in\mathbb{N},\ \alpha\in\{0,1\}).$$

As in [5], we introduce the order triplet $U(IP) = (p,\alpha,q)$ as a characteristic of the relations (8) for the iterative process IP. The integers p and q are the exponents of $u_i^{(m)}$, while $\alpha = 0$ for a TS method and $\alpha = 1$ in the case of an SS method.

In order to determine the convergence order of the algorithms (2)-(7), we shall use the following assertion proved in [5]. Assume that the starting approximations $z_1^{(0)},\ldots,z_\nu^{(0)}$ are chosen sufficiently close to the zeros $r_1,\ldots,r_\nu$ so that $u_i^{(0)} < 1$ $(i=1,\ldots,\nu)$. Then, for the iterative process IP with $U(IP) = (p,\alpha,q)$ we have

$$(9)\qquad \begin{aligned} O_R(IP,\mathbf{r}) &= p + q \qquad \text{if } \alpha = 0 \quad \text{(total-step method)},\\ O_R(IP,\mathbf{r}) &= p + t_\nu \qquad \text{if } \alpha = 1 \quad \text{(single-step method)}, \end{aligned}$$

where t_ν is the unique positive root of the equation

$$(10)\qquad t^\nu - tq^{\nu-1} - pq^{\nu-1} = 0 \quad .$$

Using the results presented in [5] and [8] we can find the following bounds for t_ν:

$$(11)\qquad q + \frac{pq}{(\nu-1)(p+q)} < t_\nu \leq q + \frac{2p}{1+\sqrt{1+\frac{4p}{q}}} \quad .$$

An extensive but elementary analysis, similar as in [1] or [5]-[7], shows that the iterative schemes (2)-(7) have the following characteristics:

$$U(TS) = (2,0,2)\ ,\ U(TSN) = (3,0,2)\ ,\ U(TSH) = (3,0,3)\ ,$$

$$U(SS) = (3,1,1)\ ,\ U(SSN) = (3,1,2)\ ,\ U(SSH) = (3,1,3)\ .$$

According to this and (9) we have the assertions:

THEOREM 1. *The convergence order of the total-step methods TS(2), TSN(4) and TSH(6) is four, five and six, respectively.*

THEOREM 2. *The R-order of convergence of the single-step methods SS(3), SSN(5) and SSH(7) is given by*

$$O_R(SS, r) \geq 3 + \tau_\nu \ ,$$

$$O_R(SSN, r) \geq 3 + x_\nu$$

and

$$O_R(SSH, r) \geq 3 + y_\nu \ ,$$

where τ_ν, x_ν *and* y_ν *are the unique positive roots of the equations*

$$\tau^\nu - \tau - 3 = 0 \ ,$$

$$x^\nu - x \cdot 2^{\nu-1} - 3 \cdot 2^{\nu-1} = 0$$

and

$$y^\nu - y \cdot 3^{\nu-1} - 3^\nu = 0 \ ,$$

respectively.

The values of the lower bounds of the R-order of convergence in the case of the single-step methods can be easily obtained solving the algebraic equation (10) starting from the interval given by (11). These values are displyed in Table 1 and coincide with that concerning the corresponding modifications of square-root iterations (the single-step versions, without or with the Newton's and Halley's corrections) (see [7]).

method \ ν	2	3	4	5	6	7	8	9	10
SS(3)	5.303	4.672	4.453	4.341	4.274	4.229	4.196	4.172	4.153
SSN(5)	6.646	5.862	5.585	5.443	5.357	5.299	5.257	5.225	5.200
SSH(7)	7.854	6.974	6.662	6.502	6.404	6.338	6.291	6.255	6.227

Table 1

3. NUMERICAL RESULTS

In order to test the presented iterative schemes, a FORTRAN routine was realised on a HONEYWELL 66 system in double-precision arithmetic (about 18 significant decimal digits). In realising the TSN, SSN, TSH and SSH methods with Newton's and Halley's corrections, before calculating new approximations $z_i^{(m+1)}$, the values $f_k(z_i^{(m)})$ $(k=1,2\ ;\ m=0,1,\ldots)$ were calculated. The same values are used for calculating the function

$$g(z_i^{(m)}) = \frac{1}{2}\ (1+\frac{1}{\mu_i})\ f_1(z_i^{(m)}) - \frac{1}{2}\cdot\frac{f_2(z_i^{(m)})}{f_1(z_i^{(m)})}\quad ,$$

the Newton's correction

$$N(z_i^{(m)}) = \frac{\mu_i}{f_1(z_i^{(m)})}$$

and Halley's correction

$$H(z_i^{(m)}) = \frac{1}{g(z_i^{(m)})} = 2\left[\frac{f_2(z_i^{(m)})}{f_1(z_i^{(m)})} - (1+\frac{1}{\mu_i})\ f_1(z_i^{(m)})\right]^{-1}.$$

Thus, the proposed iterative methods with Newton's and Halley's correction terms require slightly more numerical opperations in relation to the basic methods (algorithms (2) and (3)). Taking into account the significantly increased order of convergence, it is obvious that the proposed methods have a greater efficiency.

In order to illustrate numerically the efficiency of the modified methods, the algorithms TS(2), SS(3), TSN(4), SSN(5), TSH(6) and SSH(7) were applied for the improvement of zeros of the polynomial

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

The exact zeros of this polynomial are $r_1 = 1+2i$, $r_2 = 1-2i$, $r_3 = -1$ and $r_4 = 3$, with multiplicities $\mu_1 = 2$, $\mu_2 = 2$, $\mu_3 = 3$ and $\mu_4 = 2$. As initial approximations to these zeros the

following complex numbers were taken:

$$z_1^{(0)} = 1.7 + 2.7i, \quad z_2^{(0)} = 1.7 - 2.7i,$$
$$z_3^{(0)} = -0.3 - 0.7i, \quad z_4^{(0)} = 2.4 - 0.6i.$$

In spite of crude initial approximations ($\min_i |z_i^{(0)} - r_i| \cong 1$) the modified methods demonstrate very fast convergence. Numerical results, obtained in the second iteration, are given in Table 2.

method	i	Re $\{z_i^{(2)}\}$	Im $\{z_i^{(2)}\}$
TS(2)	1	0.999999703872727	1.999999577023530
	2	1.000004966234449	-1.999858354626263
	3	-1.000001724263487	1.28×10^{-6}
	4	3.000175153200852	4.58×10^{-5}
SS(3)	1	0.999999603833368	2.000000538829041
	2	0.999997513035036	-2.000168291520113
	3	-1.000001434643141	8.31×10^{-7}
	4	3.000000000400398	4.03×10^{-9}
TSN(4)	1	1.000005463270708	1.999990357789566
	2	1.000000009930465	-2.000000025656453
	3	-0.999999370541218	-2.44×10^{-7}
	4	2.999980969476169	5.24×10^{-6}
SSN(5)	1	0.999998904155992	1.999998927299469
	2	0.999999988521851	-1.999999982758255
	3	-1.000000001576391	5.07×10^{-9}
	4	3.000000000001254	-3.26×10^{-12}
TSH(6)	1	1.000000002444691	2.000000000565806
	2	1.000000002639924	-2.000000001014728
	3	-0.999999999964674	-2.81×10^{-12}
	4	3.000000003876174	-3.25×10^{-10}
SSH(7)	1	1.000000000020514	2.000000000101261
	2	1.000000000012086	-1.999999999988034
	3	-1.000000000000157	-2.86×10^{-13}
	4	3.000000000000029	-2.88×10^{-14}

Table 2

REFERENCES

. G. ALEFELD and J. HERZBERGER: *On the convergence speed of some algorithms for the simultaneous approximation of polynomial roots*. SIAM J. Numer. Anal. 11 (1974), 237-243.

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

. G. V. MILOVANOVIĆ and M. S. PETKOVIĆ: *On the convergence order of a modified method for simultaneous finding polynomial zeros*. Computing 30 (1983), 171-178.

. J. M. ORTEGA and W. C. RHEINBOLDT: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York 1970.

. M. S. PETKOVIĆ, G. V. MILOVANOVIĆ and L. V. STEFANOVIĆ: *Some higher-order methods for the simultaneous approximation of multiple polynomial zeros*. Comput. Math. Appls. 9 (1986), 951-962.

. M. S. PETKOVIĆ and L. V. STEFANOVIĆ: *On the convergence order of accelerated root iterations*. Numer. Math. 44 (1984), 463-476.

'. M. S. PETKOVIĆ and L. V. STEFANOVIĆ: *On some improvements of square root iteration for polynomial complex zeros*. J. Comput. Appl. Math. 15 (1986), 13-25.

3. L. V. STEFANOVIĆ: *Some iterative methods for the simultaneos finding of polynomial zeros* (in Serbo-Croatian). Ph. D. Thesis, University of Niš, Niš 1986.

). D. WANG and Y. WU: *Some modifications of the parallel Halley iteration method and their convergence*. Computing 38 (1987), 75-87.

). X. WANG and S. ZHENG: *A family of parallel and interval iteration for finding simultaneously all roots of a polynomial with rapid convergence (I)*. J. Comput. Math. 2 (1984), 70-76.

L. X. WANG and S. ZHENG: *A family of parallel and interval iteration for finding simultaneously all roots of a polynomial with rapid convergence (II)*. (in Chinese). J. Comput. Math. 4 (1985), 433-444.

Numerical Methods and Approximation Theory III (Niš, August 18-21, 1987)

PADÉ-APPROXIMATION AND BAND-LIMITED PROCESSES

TIBOR K. POGÁNY

ABSTRACT: In the paper we apply the Padé-approximation method to the approximation of spectral densities which are analytical at the origin.The observed densities are positive on a finite interval $I = [-w,w] \subset \mathbb{R}$ and vanish otherwise. Some results are given on the lower and upper bound of Padé-approximants on I and the convergence for some approximant sequences of the observed density was investigated.Related convergence results are given for the sequences of Padé-processes.

1. INTRODUCTION

The estimation theory of wide-sense stationary stochastic processes use the so called Wiener-Hopf equation,which Yaglom has solved explicitly for the class of processes with rational spectral densities.The Wiener-Hopf equation can be solved only in this class.

The concept of a band-limited process is an important one in practice.Many processes in applied sciences have a spectrum f(u) which is concetrated on a finite interval I.Practically these processes are band-limited: The harmonic oscillations $f(u)e^{iut}$ with frequencies u outside I have very small energy.

Because some Padé-approximants of the spectral density of a band-limited process have identical properties as the spectral densities,with the help of the convergence results for approximant sequences and Padé-process sequences,we may

map the rational approximation problem into a stochastic process class.

The final step is: solving the estimative problems for a rational, Padé-density class.

2. PRELIMINARIES AND SOME DEFINITIONS

Let $f(u)$ be a real function analytic at the origin. The Padé-approximant (in further PA) of order (L,M) of the function $f(u)$ is the rational expression $(L/M)_f(u) = P_L(u)/Q_M(u)$, $Q_M(o) = 1$, which has $L+M$-order contact with $f(u)$ at the origin. We can write:

$$Q_M(u)f(u) - P_L(u) = O(u^{L+M+1}) \tag{1}$$

or equivalently

$$Q_M(u)f(u) - P_L(u) = u^{L+M+1}\, h_{L,M}(u) \tag{2}$$

where $h_{L,M}(o) \neq o$. The coefficients of the polynomials $P_L(u) = \sum_{o}^{L} p_k u^k$, $Q_M(u) = \sum_{o}^{M} q_k u^k$ can be computed from (1), see for example [1].

The formal power series of $f(u)$ is

$$f(u) = \sum_{k=o}^{\infty} f_k u^k = \sum_{k=o}^{\infty} \frac{f^{(k)}(o)}{k!}\, u^k \quad , \tag{3}$$

where the series on the right side of (3) converges uniformly on the interval $(-r,r)$, $r^{-1} = \lim_{k\to\infty} \sup |f_k|^{1/k}$. When $r = +\infty$, $f(u)$ is an entire function; for $r = o$ the power series converges only at the origin and the power series (3) is formal.

The real function $f(u)$ is analytic on the interval $I =$

$= [a,b]$ if in some neighborhood $u_o - \delta < u < u_o + \delta$ of all points $u_o \in I$ there exists the expansion

$$f(u) = \sum_{k=o}^{\infty} f_k (u - u_o)^k \tag{4}$$

where the coefficients f_k are real.

Let $f(u)$ be a real function, analytic on I, and $\hat{f}(z)$ an analytic function on some region D which contains I, and $f(u) = \hat{f}(u)$ on I. Then $\hat{f}(z)$ said be the analytical continuation of $f(u)$ from I into the region D.

The series

$$\sum_{k=o}^{\infty} f_k (z - u_o)^k \tag{5}$$

we obtain from (4) with a complex value $z = u + iv$. It converges on the disk $|z - u_o| < \delta$ and its sum is $\hat{f}(z)$. Of course the sums (4) and (5) are identical on I. Finally, $f(u)$ can be analytical continued from I to some region D, which is symmetric with respect to the real axis: this fact follows from the Riemann-Schwartz principle of symmetry.

The function $h_{L,M}(z)$ has an integral representation:

$$\frac{1}{2\pi i} \oint_{\Gamma} \frac{f(t) Q_M(t)}{t^{L+M+1}(t - z)} \, dt \tag{6}$$

where Γ is a positively oriented contour in $\mathbb{C}$ which satisfies the following conditions:

(i) the origin and the point $t=z$ are inside Γ,

(ii) $f(u)$ is analytic on and within Γ.

Naturally we choose in the integral representation (6) the analytical continuation of the functions $f(u), Q_M(u)$ to the whole complex region G ($\partial G = \Gamma$). We shall lightly recognize, based on the context, the nature of the investigated functions. Instead of $\hat{f}, \hat{h}, \hat{Q}_M$ and $\hat{P}_L$ we shall write f, h, Q_M and P_L .

The existence of PAs was discussed for example in [1] .

A wide-sense stationary stochastic process has the spectral representation in the form:

$$X(t) = \int_{\mathbb{R}} e^{itu} \, dZ_X(u),$$

where $Z_X(u)$ is the so called spectral process of $X(t)$. The connection between the process and its spectral density $f(u)$ is given with the correlation function $K_X(t)$ and the so called Bochner-Khintchine's theorem:

$$EX(t)\overline{X(o)} = K_X(t) = \int_{\mathbb{R}} e^{itu} f(u) du \ .$$

From $K_X(t) = EX(t)\overline{X(o)} = E\,\overline{X(o)\overline{X(t)}} = \overline{K_X(-t)}$ and from

$$f(u) = \frac{1}{2} \int_{\mathbb{R}} e^{-itu} K_X(t) \, dt$$

it follows that a spectral density is nonnegative, selfconjugate and $L_1(\mathbb{R})$ - integrable. The quantity $K_X(o) = E|X(t)|^2$ is the variance of the process $X(t)$, we note $K_X(o) = DX(t)$.

A wide-sense stationary stochastic process is said to be band-limited if there exists a positive real number w, such that

$$K(t) = \int_{-W}^{W} e^{itu} f(u) \, du \ .$$

In other words, the spectral density vanishes outside of $I = [-w,w]$.

In our investigations we consider only the centered processes, i.e. $EX(t) = 0$.

3. PADÉ-APPROXIMATION OF SPECTRAL DENSITIES

Let $f(u)$ be a real function, analytic at the origin and

$$f(u) \begin{cases} >0 & u \in [-w,w] \\ =0 & \text{otherwise} \end{cases} . \tag{7}$$

Because $f(u)$ is a real function, the poles of $f(u)$ are complex when $f(u)$ is bounded or L_1-integrable. All functions which satisfy the condition (7), form the class $\mathbb{F}$. $\mathbb{F}$ is a subclass of the basic function class.

Theorem 1: Let $f \in \mathbb{F}$. If $P_L(u) > 0$ on $I = [-w,w]$, it follows that $(L/M)_f(u) \in L_1(\mathbb{R})$.

Proof: From relation (2) follows

$$Q_M(z)f(z) - P_L(z) = \frac{z^{L+M+1}}{2\pi i} \oint_{\Gamma} \frac{f(t)Q_M(t)}{t^{L+M+1}(t-z)}\, dt, \tag{8}$$

where Γ is a closed, positively oriented contour which contains I. We can now evaluate the quantity $|h_{L,M}(z)|$ through

$$|h_{L,M}(z)| \leq \max|f(z)| \left| \frac{1}{2\pi i} \oint_{\Gamma} \frac{Q_M(t)}{t^{L+M+1}(t-z)}\, dt \right|$$

$$\leq f^{+}\left(\sum_{k=1}^{L+M+1} \left| \frac{1}{2\pi i} \oint_{\Gamma} \frac{A_k}{t^k}\, dt \right| + \left| \frac{1}{2\pi i} \oint_{\Gamma} \frac{B}{t-z}\, dt \right| \right) = H,$$

where $f^+ = \max_G |f(z)|$, and

$$A_{L+M-k+1} = \begin{cases} -z^{-k-1} \sum_{j=0}^{k} q_j z^j & k = \overline{0,M-1} \\ -z^{-k-1} Q_M(z) & k = \overline{M,L+M} \end{cases} ,$$

$B = -A_1 = Q_M(z)/z^{L+M+1}$. From the Cauchy's integral formula it follows that

$$H = f^+(|A_1| + |B|) = 2f^+ |Q_M(z)| / |z|^{L+M+1} .$$

Finally

$$|h_{L,M}(z)| \leq 2f^+ |Q_M(z)| / |z|^{L+M+1} . \tag{9}$$

Now, from (2), (9) and $||a| - |b|| \leq |a - b|$ follows

$$\left| |Q_M(z)| |f(z)| - |P_L(z)| \right| \leq |Q_M(z)f(z) - P_L(z)| \leq 2f^+ |Q_M(z)|$$

for all $z \in G$. Further, we choose the restrictions of the investigated functions to the real axis. The last evaluation gives

$$|Q_M(u)| \geq |P_L(u)| / 3f^+ . \tag{10}$$

Let $P_L^- = \min_I |P_L(u)|$. From the positivity of P_L^- and (10) we have

$$0 < P_L^- / 3f^+ \leq \left| \frac{P_L(u)}{Q_M(u)} \right| = |(L/M)_f(u)| . \tag{11}$$

The upper bound of $(L/M)_f(u)$ there exists: (10) guarantees that

$(L/M)_f(u)$ has no poles on I.So $(L/M)_f(u)$ is a bounded,positive function on I.The positivity of $(L/M)_f(u)$ is a simple consequence of $(L/M)_f(o) = f(o) > o$.

The proof is complete. □

Consequence: Let f $\mathbb{F}$ .Then $(o/2m)_f(u)$ satisfies the inequality

$$(12) \qquad o < (Q_M^+)^{-1} \leq \left|(o/2m)_f(u)\right| \leq \frac{3f^+}{f(o)}$$

where Q_M^+ is the restriction to I of $\max_G |Q_M(z)|$.

Proof: Based on the maximum modulus principle for the closed region G is $|Q_M(z)| \leq \max_G |Q_M(z)| = Q_M^+ \cdot P_L^- = f(o)$ and from (1o) and (11) follows the statement of the consequence.□

Remark: From foregoing considerations it is clear that we observe the function $(L/M)_f(u)$ only on I.Exactly,we think that $(L/M)_f(u)$ vanishes outside of I.

A special type of Padé-approximants,the (o,2m) order PAs have a very interesting property: it can be written as

$$(13) \qquad (o/2m)_f(u) = f(o)/|A_m(u)|^2$$

for some complex coefficient polynomial $A_m(u)$ retaining the properties given in (12),if f(u) is an <u>even</u> function.

From (1) we have

$$(14) \qquad \sum_o^k q_j f_{k-j} = p_k \quad (k = \overline{o,L}); \quad \sum_o^k q_j f_{k-j} = o \quad (k = \overline{L+1,M+L}).$$

If f(u) is an even function the formal power series (3) contains only the eventh order elements: $f_{2k-1} = o$, $k \in \mathbb{N}$.

For the (o,2m) order PA $L = o, M = 2m$ and the system (14) reduces to

$$(15) \qquad q_o = 1 \ , \qquad \sum_{j=o}^{k} q_j f_{k-j} = o \qquad (k = \overline{1,2m})$$

and $q_1 = q_3 = \dots = q_{2m-1} = o$. The matrix form of (15) is

$$\begin{bmatrix} 1 & o & o & \cdot & \cdot & \cdot & o \\ f_2 & f(o) & o & \cdot & \cdot & \cdot & o \\ f_4 & f_2 & f(o) & \cdot & \cdot & \cdot & o \\ \cdot & \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & & \cdot & \cdot \\ f_{2m} & f_{2m-2} & f_{2m-4} & \cdot & \cdot & \cdot & f(o) \end{bmatrix} \cdot \begin{bmatrix} q_o \\ q_2 \\ q_4 \\ \cdot \\ \cdot \\ \cdot \\ q_{2m} \end{bmatrix} = \begin{bmatrix} 1 \\ o \\ o \\ \cdot \\ \cdot \\ \cdot \\ o \end{bmatrix}$$

The solution of the previous system is unique and nontrivial:

$$q_{2j} = \frac{(-1)^j}{f(o)^j} \begin{vmatrix} f_2 & f(o) & \cdot & \cdot & \cdot & o \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & f(o) \\ & & & & \cdot & \\ f_{2j} & f_{2j-2} & \cdot & \cdot & \cdot & f_2 \end{vmatrix}$$

where $j = \overline{o,m}$.

The connection between the coefficients of $Q_{2m}(u)$ and $A_m(u)$ must be from (13):

$$(16) \qquad q_k = \sum_{\substack{i+j=k \\ i,j \in \{o,\dots,m\}}} a_i \overline{a_j} \ ,$$

where $A_m(u) = \sum_{o}^{m} a_k u^k$, $a_k \in \mathbb{C}$. We can now solve (16) with

respect to a_k ,but this solution is not unique.For example:

(i) $|a_o|^2 = 1$ i.e. $a_o = e^{i\varphi_o}$.

(ii) Let us take $a_j = r_j e^{i\varphi_j}$, $j = \overline{1,m}$
$a_o\overline{a_1} + \overline{a_o}a_1 = 2\mathrm{Re}(\overline{a_o}a_1) = o$ give us $a_1 = a_o r_1 i(-1)^{k_1}$
for some integer k_1 and arbitrary $r_1 > o$.

(iii) $a_o\overline{a_2} + |a_1|^2 + \overline{a_o}a_2 = 2\mathrm{Re}(\overline{a_o}a_2) + r_1^2 = -f_2/f(o)$,
i.e. $r_2\cos(\varphi_2-\varphi_o) = -1/2(r_1^2 + f_2/f(o))$.The last equation has a solution but it is not unique etc.

In this way we prove the existence of the coefficients a_k. Hence,the relation (13) is valid,and $(o/2m)_f(u)$ is positive on I,therefore the function

$$(o/2m)_f(u) = \begin{cases} f(o)/|A_m(u)|^2 & u \in I = [-w,w] \\ o & \text{elsewhere} \end{cases}$$

is the rational spectral density of a band-limited process $(o/2m)_X(t)$ if the band-limited process X(t) has the spectral density f(u).

4. CONVERGENCE OF PA SEQUENCES

We consider a sequence $\{(o/2m)_f(u)\}$ of PAs of a band-limited density f(u).The uniform convergence of such sequences was investigated by many authors: De Montessus,Beardon, Pommerenke etc. in the following cases: $n/m \to \infty$; $n=am$, $a \in (o,1)$ etc..Now,we investigate the pointwise convergence on the possible largest interval on (-r,r).We prove first a result for

the sequence of $(L/M)_f(u)$ PAs.

Theorem 3: The truncation error for PA approximation of $f \in \mathbb{F}$ is

$$\left| f(u) - (L/M)_f(u) \right| = O\left(\left(\frac{|u|}{r} \right)^{L+M+1} \right) \text{ on } I \cap (-r,r).$$

Proof: From (2) follows

$$\left| f(z) - (L/M)_f(z) \right| \leq \frac{|z|^{L+M+1}}{2\pi |Q_M(z)|} \left| \oint_{\Gamma} \frac{f(t) Q_M(t)}{t^{L+M+1}(t - z)} \, dt \right| .$$

We choose a new integration contour $C_r = \left\{ re^{is} : 0 \leq s \leq 2\pi \right\}$, that contains the point z .It follows

$$\left| f(z) - (L/M)_f(z) \right| \leq \frac{|z|^{L+M+1}}{2\pi |Q_M(z)|} \int_0^{2\pi} \frac{\left| f(re^{is}) Q_M(re^{is}) \right|}{r^{L+M} \left| re^{is} - z \right|} \, ds \quad .$$

From the maximum modulus principle $|Q_M(re^{is})|$ has its maximum on the integration contour C_r , this value is $Q^+_{M,r}$ .Theorem 1 give us the estimates

$$\left| f(z) - (L/M)_f(z) \right| \leq \frac{|z|^{L+M+1}}{2\pi r^{L+M} P_L^-} 3 Q^+_{M,r} (f^+)^2 \int_0^{2\pi} \frac{ds}{\left| re^{is} - z \right|}$$

$$\leq \frac{|z|^{L+M+1} \, 3 (f^+)^2}{r^{L+M}(r - |z|) P_L^-} Q^+_{M,r} \quad .$$

Naturally,we retain the solution on the positivity of $P_L(u)$. Finally,we get the inequality:

$$\left| f(z) - (L/M)_f(z) \right| \leq 3 Q^+_{M,r} \frac{(f^+)^2}{P_L^-} (|z|/r)^{L+M+1} \left(1 - \frac{|z|}{r}\right)^{-1} .$$

ıoosing the restrictions of the functions in the last in-
ıuality to I $=[-w,w]\cap(-r,r)$ it is not hard to show that its
ɔal-valued variant is equivalent to the assertion of the
ıeorem.□

As a consequence of the previous theorem we can formu-
ıte the

Theorem 4: The sequence of rational spectral densities $\{(o/2m)_f(u)\}$ tends to an even band-limited density f(u) point-
ise on $I\cap(-r,r)$, when m tends to infinity.□

The elements of the sequence $\{(o/2m)_f(u)\}$ are spectral
.ensities.We can so approximate poitwise the even spectral
.ensity of a band-limited process with rational spectral den-
ities.This result has very interesting consequences.

3. PADÉ-PROCESSES

The PA of the spectral density of a band-limited process
.s a spectral density when a) f(u) is an even function, b)
;he PA is of the order (o,2m).The first condition is a sim-
)le consequence of the reality of X(t).The connected process
)f the density $(o/2m)_f(u)$ we note $(o/2m)_X(t)$ and it is the
so called Padé-process.What can be said about the mean square
convergence of the sequence $(o/2m)_X(t)$ to X(t) if m tends to
.nfinity?Before we give an answer to this question,we discuss
;he connection between w and r.

1. $w \leqq r$.The interval of the convergence of $(o/2m)_f(u)$ to f(u) is (-r,r).Viewing in the light of the m.s.

convergence this case is interesting : we cannot lose any information on the nature of $f(u)$ and $(o/2m)_f(u)$, moreover on the processes $X(t)$ and $(o/2m)_X(t)$ too.

2. $w > r$. Outside $(-r,r)$ we cannot consider the pointwise convergence of $(o/2m)_f(u)$, therefore the convergence in the mean of $(o/2m)_X(t)$ is senseless.

For example the class of the basic functions D_∞ (which are infinitely differentiable and vanish outside of a finite interval) of L. Schwartz satisfy the property 1.

Thus in the following considerations we suppose that $w \leq r$.

The linear transformation (or filter) of the process $X(t)$ is a transformation $A: X(t) \to Y(t)$ where:

$$(17) \qquad Y(t) = \int_{\mathbb{R}} e^{itu} h_Y(u)\, dZ_X(u) \quad .$$

$Z_X(u)$ is the spectral process of $X(t)$ (section 2.) and the $L_2(f_X(u)du)$-integrable function $h_Y(t)$ is the spectral characteristic function of the filter A. Some classical examples are: the differential operator $\mathbb{D}$ with the spectral characteristic $h_{\mathbb{D}}(u) = iu$, the integration operator $\mathbb{I}$ with $h_{\mathbb{I}}(u) = 1/iu$. Let now $\hat{h}_Y(s)$ be the inverse Fourier-transform of $h_Y(u)$. Another representation of $Y(t)$ is (equivalently to (17)):

$$(18) \qquad Y(t) = \int_{\mathbb{R}} \hat{h}_Y(s) X(t - s)\, ds \quad .$$

Consequently $Y(t)$ is the response of the process $X(t)$ on the input A.

A Rozanov theorem gives us the connection:

$$f_X(u)\left|h_Y(u)\right|^2 = f_Y(u) \quad ,$$

where $f_X(u)$ and $f_Y(u)$ are the spectral densities of X(t) and Y(t). In our case we consider the process $(o/2m)_X(t)$ as the response on the input $(o/2m)_{(\cdot)}$ to the band-limited process X(t). It has a spectral density like f(u) in formula (7). The characteristic function is

$$h_{(o/2m)}(u) = \frac{f_X(o)^{1/2}}{A_m(u)}\, f_X^{-1/2}(u) \tag{19}$$

from Rozanov's theorem, where $f_X^{1/2}$ is the positive root of the equation $(f_X^{1/2})^2 = f_X$. Naturally, $h_{(o/2m)}(u)$ is L_2-integrable on the measure $f_X(u)du$:

$$\int_{\mathbb{R}} \left|h_{(o/2m)}(u)\right|^2 f_X(u)\, du = \int_{\mathbb{R}} (o/2m)_f(u)\, du =$$

$$= E\left|(o/2m)_X(t)\right|^2 = D(o/2m)_X(t) < \infty \, ,$$

and the process $(o/2m)_X(t)$ has bounded second moment.

Theorem 5: $\left|h_{(o/2m)}(u)\right|^2 \longrightarrow 1$ pointwise on the interval (-r,r) when m tends to infinity.

Proof: the statement follows from Theorem 4 and (19). □

The cross-correlation function $K_{X,P}(t)$ of the process X(t) and $(o/2m)_X(t)$ was defined with

$$K_{X,P}(t) = EX(t)\overline{(o/2m)_X(o)} = \int_{\mathbb{R}} e^{itu} f_{X,P}(u)\, du$$

where $f_{X,P}(u)$ is the so called cross-spectral density. Another result by Rozanov states that $f_{X,P}(u) = f_X(u)\, h_{(o/2m)}(u)$, for $f_X(u) \in \mathbb{F}$. It is clear that

$$(2o) \qquad E|X(t) - (o/2m)_X(t)|^2 = \int_{\mathbb{R}} |1 - h_{(o/2m)}(u)|^2 f_X(u)du.$$

It is not hard to show that there exists a positive real number C' and a positive integer m_o for which is

$$|1 - h_{(o/2m)}(u)|^2 \leq C'(1 - |h_{(o/2m)}(u)|^2)^2$$

if $m > m_o$. Based on theorem 5 we give

Theorem 6: $E|X(t) - (o/2m)_X(t)|^2 \xrightarrow[m\to\infty]{} o$. □

Of course, we can state that the theorems 5 and 6 are valid on the whole of $\mathbb{R}$. Naturally, we choose only the positive-r densities from $\mathbb{F}$ for which $w \leq r$. The formal spectral densities have no practical importance.

REFERENCES

1. G.A.BAKER Jr. and P.R.GRAVES-MORRIS: Padé-Approximants, Addison-Wesley Publishing co., Reading, Massachusets, 1981.
2. D.ELLIOTT: Truncation error in Padé-approximants to certain functions: an alternative approach, Math.Comp.21(1967), 3o8-32
3. T.POGÁNY: Singuläre zufällige Prozesse und mittelquadratisch Konvergenz, Publ.Math.Debrecen 34(1987), 197-2o5.
4. YU.A.ROZANOV: Stationary Random Prozesses, Fizmatgiz, Moscow, 1963.
5. A.M.YAGLOM: An Intriduction to the Theory of Stationary Random Functions, Dover Publications, New York, 1973.

AN APPLICATION OF VARIATIONAL CALCULUS IN MECHANICS AND SOME PROPERTIES OF THE EIGENVALUES OF THE LAPLACIAN

THEMISTOCLES M. RASSIAS

\BSTRACT. In this survey paper we present:
I. The stability and oscillations or small motions of a soap film suspended between parallel coaxial rings. The solution to the problem relates the radius of the film r to the displacement z along the axis of symmetry by the equation of

$$r=a\cosh\frac{z-b}{a}.$$

The constants a and b are to be determined by requiring that r be equal to the fixed radii of the rings for z=0 and h, where h is the separation of the rings.
We study this equilibrium problem using eigenfunction methods and prove that the dynamical stability of the film is determined by the sign of the lowest eigenvalue λ_1 of an associated Sturm-Liouville problem, with the film stable for $\lambda_1>0$ and unstable for $\lambda_1<0$. This follows Durand [6].

II. Some of the most important properties of the eigenvalues of the Laplacian with some remarks on the smoothness of eigenfunctions and a generalization of Courant's nodal domain theorem (see [19],[2

I. An Application of Variational Calculus in Mechanics

In this section we consider the stability and oscillations or small motions of a soap film suspended between parallel coaxial rings, as this has been analyzed in Durand [6]. It is a standard problem used to introduce variational calculus in mechanics to determine the equilibrium shape of a soap film suspended between two parallel coaxial circular rings. The solution to the problem relates the radius of the film r to the displacement z along the axis of symmetry by the equation of the catenary

$$r=a\cosh\frac{z-b}{a}. \tag{1}$$

The constants a and b are to be determined by requiring that r be equal to the fixed radii of the rings for z=0 and h, where h is the separation of the rings. If the rings are of equal radius r_0, the surface is symmetrical about $z=\frac{h}{2}$, b is equal to $\frac{h}{2}$,and a,the minimum radius of the film, is to be found by solving the equation

$$(2) \qquad r_0 = a \cosh \frac{h}{2a} .$$

There are two solutions for $\frac{h}{2r_0} < 0.66274...$, *only one of which is stable, and no solutions at all for* $\frac{h}{2r_0} < 0.66274...$ In the second case, *the tubular configuration of the soap film is unstable*. From the *experimental point of view* this can be demostrated as follows (see [6]): We start with a stable tubular film with $\frac{h}{2r_0} < 0.66274$. and gradually increasing the separation between the rings until $\frac{h}{2r_0}$ approaches and then exceeds the critical value. For $\frac{h}{2r_0}$ greater than the critical value, the film collapses in the center and splits into two planar films, one on each ring. As $\frac{h}{2r_0}$ approaches the critical value, any perturbation results in a characteristic low-frequency oscillation of the film.

We shall give a mathematical analysis of this equilibrium problem (following [6]) using eigenfunction methods, and show that the *dynamical stability* of the film is determined by the sign of the lowest eigenvalue λ_1 of an associated *Sturm-Liouville problem* with the film *stable for* $\lambda_1 > 0$ *and unstable for* $\lambda_1 < 0$.

The energy of an ideal static soap film with surface area S and surface tension σ is given, neglecting gravity, by

$$(3) \qquad V[S] = 2\sigma S.$$

The possible equilibrium shapes of the film are determined by finding those surfaces for which V[S] has a local minimum. We require that the film be attached to two plane parallel coaxial rings with radii r_1 and r_2 separated by a distance h, and have no other boundaries. The equilibrium surfaces are axially symmetric, with a surface energy given by

$$(4) \qquad V[S] = 2\sigma \int_S dS = 4\pi\sigma \int_S r\sqrt{dr^2+dz^2}$$

If V[S] is to be an extremal for a surface S, there must be no first-order change in V[S] when S is varied slightly subject

to the fixed boundary conditions, i.e. $\delta V[S]=0$. We will specify the shape of the surface by giving its radius r as a function of z. Thus we get that $V[S]$ vanishes if r(z) satisfies the Euler equation

$$(5)\quad \frac{d}{dz}\left(\frac{rr_z}{\sqrt{1+r_z^2}}\right) - \sqrt{1+r_z^2} = 0,\quad r_z \equiv \frac{d}{dz}\, r(z),$$

or equivalently, if

$$(6)\quad \frac{1}{r_z}\,\frac{d}{dz}\left(\frac{r}{\sqrt{1+r_z^2}}\right) = 0.$$

Equation (6) is satisfied if either

$$(7)\quad \frac{r}{\sqrt{1+r_z^2}} = a,$$

where a is a positive constant, or r_z is infinite. Solving (7) we obtain the equation of a hollow tube,

$$(8)\quad r(z)=a\cosh\left(\frac{z-b}{a}\right),$$

where b is a constant of integration. In the second case, $z_r=\frac{1}{r_z}$ vanishes, thus z does not vary with r, and the surface S consists for our boundary conditions of two disconnected plane disks which fill the rings. Any variation of the disks about the plane configuration clearly increases their surface area. As a result $V[S]$ has at least a local minimum, and the double soap film is stable against small perturbations. From the topological point of view the double soap film is distinct from thehollow tube. The constants of integration a and b in (8) must be specified in such a way that $r(0)=r_1$ and $r(h)=r_2$. We will consider the case of rings of equal radius r_0. Similar methods can be applied for the asymmetrical case We obtain $b=\frac{h}{2}$,

$$(9)\quad r(z)=a\cosh\left(\frac{z}{a}-\frac{h}{2a}\right),$$

and the boundary value problem reduces to that of determining a from the equation

$$(10)\quad r_0=a\cosh\frac{h}{2a},\quad a>0,$$

which again can be rewriten as

$$(11)\qquad \frac{2r_0}{h}=\frac{2a}{h}\cosh\frac{h}{2a}=u_0^{-1}\cosh u_0,\ u_0=\frac{h}{2a}\ .$$

The function $u^{-1}\cosh u$ is positive, diverges for $u\to 0$ and $u\to\infty$ (as $a\to\infty,0$), and has a finite minimum value 1.5089...for utanhu=1, $u=u_c=1.1997...$It follows that there exist two solutions to the boundary value problem for $\frac{2r_0}{h}<1.509$, a single solution for $\frac{2r_0}{h}=1.509$, and no solutions at all for $\frac{2r}{h}<1.509$ ($\frac{h}{2r_0}>$ $>0.66274...$).

We can solve the boundary value problem (11) by iteration starting with $a=r_0$, and find that

$$(12)\quad a=\frac{r_0}{\cosh\frac{h}{2a}}\approx r_0\left(1-\frac{h^2}{8a^2}+...\right)$$

$$\approx r_0\left(1-\frac{h^2}{8r_0^2}+...\right),\ \frac{h}{2r_0}<<1.$$

The shape of the film is given in the same approximation

$$(13)\qquad r=r_0\left(1-\frac{1}{2r_0^2}\,z(h-z)+...\right),\ \frac{h}{2r_0}<<1,$$

and is cylindrical up to terms of order $\frac{h^2}{4r_0^2}$.

The area of the film is

$$(14)\qquad S=\pi a^2\left(\sinh\frac{h}{a}+\frac{h}{a}\right)=2\pi\left(r_0(\sqrt{r_0^2-a^2})+\frac{1}{2}ha\right).$$

For the nearly cylindrical film (12),

$$(15)\qquad S=2\pi r_0 h\left[1+0\left(\frac{h^2}{4r_0^2}\right)\right],$$

where the correction terms are negative. It follows that this co figuration is stable, therefore that S has at least a local min mum. The second limiting solution for closely spaced rings $\frac{2r_0}{h}>>1$ is that for which $u_0=\frac{h}{2a}$ is large, and a is small,a<<h<< In fact, if we rewrite (11) as

$$(16)\qquad \frac{r_0}{a}=\frac{2r_0}{h}\cosh^{-1}\frac{r_0}{a}=\frac{2r_0}{h}\left[\ln\frac{2r_0}{a}+0\left(\frac{a^2}{r_0^2}\right)\right],\ \frac{a}{r_0}\ll 1,$$

We get

$$(17)\qquad a \approx \frac{h}{2}\left\{\ln\left[\frac{4r_0}{h}\left(\ln\frac{4r_0}{h}(\ldots)\right)\right]\right\}^{-1} < \frac{h}{2}\,.$$

For very closely spaced rings the extremal surface consists of two nearly planar surfaces connected by a narrow neck with radius $a<<h<<r_0$.

The area of the surface is then

$$(18)\qquad 2\pi r_0^2\left[1+0\left(\frac{h^2}{4r_0^2}\right)\right],\quad \frac{a}{r_0}<<1.$$

The correction terms are positive. The *nearby* configuration of two separate plane disks has a smaller area $2\pi r_0^2$, and can be approached arbitrarily closely by letting $\frac{h}{2r_0}\to 0$ $(\frac{a}{r_0}\to 0)$.

We get that the narrow-necked surface is probably unstable, therefore that S probably has a local maximum for this configuration. Continuity arguments imply that the entire branch of the solution curve with $\frac{h}{2a} > 1.2$ *is unstable*, while that with $\frac{h}{2a} < 1.2$ *is stable*.

Stability of the soap film. Suppose $r(z)=f(z)$ describe an initial surface S_0 (not necessarily an extremal surface) and consider a perturbed surface described the equation

$$(19)\qquad r(z)=f(z)+g(z),$$

where $g(z)$ is an infinitesimal twice-differentiable function with $g(0)=g(h)=0$. Assume also that g_z is infinitesimal for $0\leq z\leq h$. Then $V[S]$ can be written as a power series in g, g_z as follows:

$$(20)\qquad V[S]=4\pi\sigma\int_0^h \sqrt{1+r_z^2}\; r\,dz$$

$$=4\pi\sigma\int_0^h \left(f\sqrt{1+f_z^2}+g\sqrt{1+f_z^2}+\frac{ff_zg_z}{\sqrt{1+f_z^2}}\right.$$

$$(21)\qquad \left.+\frac{gg_zf_z}{\sqrt{1+f_z^2}}+\frac{fg_z^2}{2(1+f_z)^{3/2}}+0(z^3)\right)dz$$

$$=V[S_0]+4\pi\sigma\int_0^h g\left(\sqrt{1+f_z^2}-\frac{d}{dz}\,\frac{ff_z}{\sqrt{1+f_z^2}}\right)dz\,.$$

$$+2\pi\sigma \int_0^h (fg_z^2 - f_{zz}g^2) \frac{dz}{(1+f_z^2)^{3/2}} + O(z^3)$$

(23) $$= V[S_0] + \delta V[S_0] + \delta^2 V[S_0] + \ldots$$

For S_0 an extremal surface, f(z) satisfies the Euler equation (5) and $\delta V[S_0]=0$.

Let us consider now the case of symmetrical rings.

We obtain

(24) $$f = a \cosh\left(\frac{z}{a} - \frac{h}{2a}\right) = a\cosh u, \quad u = \frac{z}{a} - \frac{h}{2a}$$

and

(25) $$\delta^2 V[S_0] = 2\pi\sigma a \int_0^h \left(g_z^2 - \frac{1}{a^2} g^2\right) \cosh^{-2}\left(\frac{z}{a} - \frac{h}{2a}\right) dz$$

$$= 2\pi\sigma \int_{-u_0}^{u_0} (g_u^2 - g^2) \frac{du}{\cosh^2 u}, \quad u_0 = \frac{h}{2a}.$$

It is known from the work of Legendre, Jacobi, and Weierstrass that an extremal curve will give a minimum of $V[S]$ if (i) the second derivative of the integrand in (20) with respect to r_z is positive for all z and r in a neighborhood of the curve and all finite r_z; and (ii) there is no point *conjugate to z=0* on the interval $0 \leq z \leq h$. It is easy to see that both these conditions are satisfied in our case.

Another way to be used in order to verify the conditions for minimum is to convert the weak stability problem into one of the determining *the sign of the lowest eigenvalue of an appropriate Sturm-Liouville operator*. The weak form of condition (i) will enter when we define the Sturm-Liouville operator. The conjugate points are just the nodes of the lowest eigenfunction of this problem, and the condition (ii) is replaced by the requirement that the lowest eigenvalue be positive. We will change at this point from the radial displacements g(z) used above to equivalent infinitesimal displacements $\xi(z)$ perpedicular to the initial surface of the film. The ξ's are the natural coordinates for the study of the oscillations. The vector displacement of a point $\bar{r}=(r,z)$ associated with a perpedicular displacement $\bar{\xi}(z)$ is

$$\bar{r}'-\bar{r}=\bar{\xi}(z)=\hat{n}(z)\xi(z), \tag{27}$$

where $\hat{n}(z)$ is the normal to the surface at $(r,z)=(f(z),z)$,

$$\hat{n}=(\hat{r}-f_z\,\hat{z})\sqrt{1+f_z^2} \tag{28}$$

Therefore

$$r'(z')=r+\xi_r\ f(z)\ +\frac{\xi(z)}{\sqrt{1+f_z^2}} \tag{29}$$

$$z'=z+\xi_z=z-\ \frac{f_z}{\sqrt{1+f_z^2}}\ \ \xi(z) \tag{30}$$

If we substitute for z' in r'(z') and expand, we obtain

$$r'(z)=f+\xi\ \sqrt{1+f_z^2}\ +0(\xi^2) \tag{31}$$

$$=f+\xi\cosh\ u\ +0(\xi^2), \tag{32}$$

thus g(z), the first order change in r at fixed z, is given by

$$g=\xi\cosh\ u. \tag{33}$$

After some standard computation we derive

$$\delta^2V[S_0]=2\pi\sigma\int_{-u_0}^{u_0}(\xi_u^2-\ \frac{2}{\cosh^2u}\ \ \xi^2)\ du \tag{34}$$

$$=2\pi\sigma\int_{-u_0}^{u_0}\xi(-\xi_{uu}-\ \frac{2}{\cosh^2u}\ \xi)\ du. \tag{35}$$

The problem now of whether or not the extremal configuration of a soap film specified by a given value of $u_0=\frac{h}{2a}$ is stable can be restated at this point in terms of the operator

$$L=-\ \frac{d^2}{du^2}\ -\ \frac{2}{\cosh^2u}\ . \tag{36}$$

We have

$$\delta^2V[S_0]=2\pi\sigma\ (\xi,L\xi), \tag{37}$$

where the inner product is defined by the integral in (35). If L is a positive operator, that is, if $(\xi,L\xi)$ is positive for any ξ, then δ^2V is positive for any variation of the extremal configuration, and the soap film is stable. Now L is a positive operator if and only if its lowest eigenvalue is positive. Consider the

Sturm-Liouville eigenvalue problem defined by the differential equation

$$(38) \qquad L\Psi_n=\lambda_n w(u)\Psi_n$$

with the boundary conditions $\Psi_n(u_0)=\Psi_n(-u_0)=0$. The weight functi w(u) must be strictly positive, but arbitrary.

Set $w(u)=\cosh^2 u$. The equation (38) becomes

$$(39) \qquad \frac{d^2\Psi_n(u)}{du^2}+(\lambda_n\cosh^2 u+\frac{2}{\cosh^2 u})\Psi_n(u)=0.$$

The eigenfunctions Ψ_n can be chosen to be real, and we will assume also that they have been normalized. The orthogonality relation for the Ψ's is then

$$(40) \qquad \int_{-u_0}^{u_0}\Psi_n(u)\Psi_m(u)\cosh^2 u\,du=\delta_{nm}.$$

The eigenvalues λ_n, n=1,2,... are real and discrete, and will be assumed to be

$$\lambda_1<\lambda_2<\lambda_3<\cdots$$

We expand $\xi(u)$ in (25) as a series in the complete set o eigenfunctions Ψ_n,

$$(41) \qquad \xi(u)=\sum_{n=1}^{\infty}c_n\Psi_n(u),\ \{c_n\}\ \text{real},$$

and we find that

$$(42) \qquad \delta^2V[S_0]=2\pi\sigma\sum_{n=1}^{\infty}\lambda_n c_n^2 .$$

We now see that a given extremal configuration of the soa film is stable (resp. unstable) if the lowest eigenvalue λ_1 is positive (resp. negative). If λ_1 is positive, all the eigenvalues are positive, and δ^2V is positive for any choice of the c_n. Thus the area of the film increases for any variation $\xi(u)$ which satisfies the boundary conditions. This is the condition for stability. If λ_1 is negative, the choice $c_1\neq 0$, $c_n=0$ for n>1 gives a va riation ξ which decreases the area of the film, $\delta^2V<0$, and the configuration is unstable. If $\lambda_1=0$, the configuration is in neutr equilibrium since an infinitesimal displacement

$$\xi(u)=c\ \Psi_1(u)$$

gives

$$\delta V=\delta^2 V=0.$$

It can be shown that λ_1 is exactly zero for the critical value of $u_0, u_0=u_c=1.20$, $\frac{h}{2r_0}=0.663$. It can also be shown that the soap film is stable for $u_0<u_c$ and unstable for $u_0>u_c$.

II. Eigenvalues of the Laplacian

II.1 Let $D\subset\mathbb{R}^m, m\geq 2$ be a domain with a smooth boundary. We consider solutions of

$$(1)\qquad \begin{cases} \Delta u+\lambda u=0 & \text{in } D \\ u=0 & \text{on } \partial D \end{cases}$$

where D is a region (and ∂D is the boundary of D) such that the spectrum is discrete. For $m=2, \Delta u+\lambda u=0$ in D is also known as the *Helmholtz equation* [10]. Someone reduces to it from separating the time variable out of the wave equation. The eigenvalue problem (1) for m=2 may represent the vibration of a *fixed membrame*, with the eigenvalue $\lambda=k^2$, where k is proportional to a *principal frequency of vibration*, and the eigenfunction u represents the shape of *a mode of vibration*. These are also the frequencies and modes of the *simply supported plate* of the same plate (see [12]).

Suppose that the spectrum i.e., those values of λ for which a non-trivial solution exists, is discrete. We order the eigenvalues

$$0<\lambda_1<\lambda_2\leq\lambda_3\leq\ldots\leq\lambda_n\leq\ldots$$

and we normalize the corresponding eigenfunctions $u_1, u_2, \ldots, u_n, \ldots$ such that

$$(2)\qquad \int_D u_i u_j=\delta_{ij},\quad i,j=1,2,\ldots$$

Theorem 1 ([9]). *Let* σ *be the unique solution on* (λ_n,∞) *of the equation*

$$(3)\qquad \sum_{i=1}^{n}\frac{\lambda_i}{\sigma-\lambda_i}=\frac{mn}{4}.$$

Then

(4) $\lambda_{n+1} \leq \sigma$.

To prove Theorem 1, Hile and Protter have first established the following proposition.

Proposition 2. I. *For each integer* l *with* $1 \leq l \leq n$, *the following inequality holds:*

$$(5) \qquad \lambda_{n+1} \leq \frac{1}{2}(\lambda_n+\lambda_1) + \frac{2}{mn} \sum_{i=1}^{n} \lambda_i + \frac{2}{mn}\left\{\left[\frac{mn}{4}(\lambda_n-\lambda_1) + \sum_{i=1}^{n}\lambda_i\right]^2 - mn(\lambda_n-\lambda_1) \sum_{i=1}^{l}\lambda_i\right\}^{1/2}$$

II. *The first* $(n+1)$ *eigenvalues of the Laplacian satisfy the inequality*

$$(6) \qquad \sum_{i=1}^{n} \frac{\lambda_i}{\lambda_{n+1}-\lambda_i} \geq \frac{1}{4} mn$$

Remark. Inequality (6) is of interest only when λ_{n+1} is strictly greater than λ_n.

Proof of Theorem 1 ([9]). Consider the n trial functions

$$(7) \qquad \phi_i = x_1 u_i - \sum_{j=1}^{n} a_{ij} u_j, \quad i=1,2,\ldots,n,$$

such that

$$(8) \qquad a_{ij} = \int_D x_1 u_i u_j, \quad i,j=1,2,\ldots,n.$$

It follows that each ϕ_i is orthogonal to $u_1, u_2, \ldots, u_n$ and because of the fact $\phi_i = 0$ on ∂D, we obtain

$$(9) \qquad \lambda_{n+1} \leq \frac{-\int \phi_i \Delta \phi_i}{\int \phi_i^2} , \quad i=1,2,\ldots,n$$

It follows easily that

$$(10) \qquad \lambda_{n+1} \int \phi_i^2 \leq \lambda_i \int \phi_i^2 - 2 \int u_{i,x_1} \phi_i, \quad i=1,2,\ldots,n$$

Thus

$$(11) \qquad \lambda_{n+1} \sum_{i=1}^{n} \int \phi_i^2 \leq \sum_{i=1}^{n} \lambda_i \int \phi_i^2 - 2 \sum_{i=1}^{n} \int u_{i,x_1} \phi_i .$$

Because of the fact a_{ij} are symmetric,

$$-2 \sum_{i=1}^{n} \int u_{i,x_1} \phi_i = -2 \sum_{i=1}^{n} \int x_1 u_i u_{i,x_1} + 2 \sum_{i,j=1}^{n}$$

$$a_{ij}\int u_j u_{i,x_1} = \sum_{i=1}^{n}\int u_i^2 = n.$$

Therefore (11) becomes

$$(12)\qquad \lambda_{n+1}\sum_{i=1}^{n}\int\phi_i^2 \leq \sum_{i=1}^{n}\lambda_i\int\phi_i^2 - 2(1+\beta)\sum_{i=1}^{n}\int u_{i,x_1}\phi_i - n\beta$$

where the real parameter β will be chosen later

Let $\tau_1,\tau_2,\ldots,\tau_n$ be positive constants and apply Cauchy's inequality, then (12) reduces to

$$(13)\ \lambda_{n+1}\sum_{i=1}^{n}\int\phi_i^2 \leq \sum_{i=1}^{n}(\lambda_i+\tau_i)\int\phi_i^2 + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int u_{i,x_1}^2 - n\beta\ .$$

Set $\tau_n=\tau$ and choose the τ_i, $i=1,2,\ldots,n-1$ so that

$$\tau_i=\tau+\lambda_n-\lambda_i,\quad i=1,2,\ldots,n-1$$

Define

$$S_1=\sum_{i=1}^{n}\int\phi_i^2$$

Then (13) can be written as follows

$$(14)\ \lambda_{n+1}\ S_1 \leq (\lambda_n+\tau)S_1 + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int u_{i,x_1}^2 - n\beta\ .$$

In place of the trial functions (7) we may choose the functions

$$(15)\ \phi_{ik}=x_k u_i - \sum_{j=1}^{n} a_{ijk}u_j, i=1,2,\ldots,n: k=1,2,\ldots,m$$

Peforming an analysis as above for $k=2,3,\ldots,m$ and,denotin

$$S_k=\sum_{i=1}^{n}\int\phi_{ik}^2\ ,\quad k=1,2,\ldots,m,$$

we obtain the m inequalities

$$(16)\ \lambda_{n+1}S_k \leq (\lambda_n+\tau)S_k + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int u_{i,x_k}^2 - n\beta, k=1,2,\ldots,m$$

Setting

$$S=\sum_{k=1}^{m}S_k,\ \text{we obtain}$$

$$\lambda_{n+1}S \leq (\lambda_n+\tau)S + (1+\beta)^2\sum_{i=1}^{n}\tau_i^{-1}\int|\nabla u_i|^2 - mn\beta,$$

or

$$(17)\quad \lambda_{n+1}S \leq (\lambda_n+\tau)S+(1+\beta)^2 \sum_{i=1}^{n} \lambda_i\tau_i^{-1}-mn\beta$$

The selection of τ such that

$$(18)\quad \sum_{i=1}^{n} \lambda_i\tau_i^{-1} \leq (1+\beta)^{-2}\, mn\beta$$

implies an inequality for the τ_i as a function of β. Therefore the condition on τ becomes

$$\sum_{i=1}^{n} \lambda_i(\tau+\lambda_n-\lambda_i)^{-1} \leq \frac{mn}{4}$$

We note that

$$f(\tau)=\sum_{i=1}^{n} \frac{\lambda_i}{\tau+\lambda_n-\lambda_i}$$

is a decreasing function of τ on (o,∞) and $\lim_{\tau\to 0^+} f(\tau)=+\infty, \lim_{\tau\to\infty} f(\tau)=0$.

Thus setting $\sigma=\tau+\lambda_n$ we observe that there is a unique solution of (3) on (λ_n,∞) and (17),(18) imply (4).

The equation (3) can be written also in the form

$$(19)\quad \prod_{i=1}^{n}(\sigma-\lambda_i)-\frac{4}{mn}\sum_{i=1}^{n}\lambda_i \prod_{\substack{j=1\\ j\neq i}}^{n}(\sigma-\lambda_i)=0.$$

We denote

$$P(\sigma)\equiv \prod_{i=1}^{n}(\sigma-\lambda_i)=\sigma^n+\sum_{k=1}^{n}(-1)^k a_k\sigma^{n-k},$$

where a_i is the i-th elementary symmetric function of $\lambda_1,\lambda_2,\ldots,$ and also denote

$$R(\sigma)\equiv \sum_{i=1}^{n}\lambda_i \prod_{\substack{j=1\\ j\neq i}}^{n}(\sigma-\lambda_j)=\sum_{k=1}^{n}(-1)^{k+1}\, k a_k\sigma^{n-k}$$

Then (19) reduces to the following form

$$(20)\quad \sigma^n+\sum_{k=1}^{n}(-1)^k a_k\left(1+\frac{4k}{mn}\right)\sigma^{n-k}.$$

Hence (4) is given by the unique root of (20) on the interval (λ_n,∞).

Q.E.D.

Remark. The above result of Hile-Protter generalizes the one given by Payne, Polya and Weinberger [14], which states that:

For domains in $\mathbb{R}^2$, *the inequality*

$$\lambda_{n+1} \leq \lambda_n + \frac{2}{n} \sum_{i=1}^{n} \lambda_i, \quad n=1,2,\ldots$$

holds if the spectrum is discrete.

In the following we outline a new method of Hile-Protter [9] which can be used to improve the upper bound estimates for λ_2. For this, let u denote the first normalized eigenfunction for

$$\begin{cases} \Delta u + \lambda_1 u = 0 & \text{in } D, \\ u = 0 & \text{on } \partial D, \end{cases}$$

and let f be any C^1 function in $D \cup \partial D$,

Theorem 3 [9]. *Let*

$$(21) \qquad c = \frac{1}{\lambda_1}\left(\frac{1}{2\lambda_1}\right)^{n-1} \frac{1}{(2n+1)^2}$$

and suppose $n \geq \lambda_1$. *Then for any domain* D *in* $\mathbb{R}^2$ *contained in the unit disk,*

$$(22) \qquad \lambda_2 \leq k\lambda_1,$$

with

$$k = \frac{(5-2c)+\sqrt{(5-2c)^2+8}}{4}$$

The proof of Theorem 3 has been based upon the following series of Lemmas [9]

Lemma 1. *Suppose*

$$\int_D f u^2 = 0$$

Then

$$(23) \qquad \lambda_2 \leq \lambda_1 + \frac{\int u^2 |\nabla f|^2}{\int f^2 u^2}.$$

Let

$$(24) \qquad A(\alpha) = \frac{\int u^{2\alpha}}{(\int u^{\alpha+1})^2}.$$

Lemma 2. *The following inequality holds:*

$$(25)\quad \frac{1}{\int x^2u^2}+\frac{1}{\int y^2u^2}\leq\frac{(\alpha+1)^2}{2\alpha-1}\lambda_1 A(\alpha),\ a\geq 1$$

Lemma 3. *Define* $\nu=\frac{\lambda_2}{\lambda_1}$; *then the inequality*

$$(26)\quad A(\alpha)\leq\frac{(2\alpha-1)(\nu-1)}{(2\alpha-1)\nu-\alpha^2}$$

holds for $1\leq\alpha<\nu+\sqrt{\nu^2-\nu}$

Lemma 4. *The following inequalities hold:*

$$(27)\quad \frac{1}{\int x^2u^2}+\frac{1}{\int y^2u^2}\leq\lambda_1\frac{3\nu+1}{\nu}$$

and, provided the axes are rotated properly,

$$\frac{1}{\int x^2u^2}\leq\frac{1}{2}\lambda_1\cdot\frac{3\nu+1}{\nu}$$

Lemma 5. *For* $\alpha\geq 1$ *define*

$$B_\alpha=\int x^\alpha u^2$$

and choose coordinate axes so that $B_1=\int xu^2=0$.

Let

$$J\equiv\frac{\left[(2n+1)B_{2n}-\lambda_1(\nu-1)B_{2n+2}\right]^2}{(2n+1)^2B_2B_{4n}}$$

with n *a positive integer. Then*

$$(28)\quad \lambda_2\leq\lambda_1+\frac{1}{B_2}-J$$

Lemma 6. *For* $n\geq\lambda_1$, *the following inequality holds:*

$$(29)\quad J\geq\left(\frac{1}{2\lambda_1}\right)^{n-1}\cdot\frac{1}{(2n+1)^2}$$

Remark. Applying Theorem 3, Hile and Protter were able to derive the following inequality of J.J.A.M. Brands [4]

$$(30)\quad \lambda_2\leq\frac{5+\sqrt{33}}{4}\lambda_1,$$

which is essentially the inequality (22) for c=0. The inequality of Brands dor $\mathbb{R}^m$ becomes

$$(31)\quad \frac{\lambda_2}{\lambda_1}\leq\frac{m+3+\sqrt{m^2+10m+9}}{2m}.$$

II.2. Smoothness of eigenfunctions. The eigenfunctions are chara-

cterized with the *unique continuation property*, that is, a function cannot satisfy $\Delta u+\lambda u=0$ in D and vanish on an open subset of D without vanishing identically in D. Each eigenfunction u_n is infinitely differentiable (i.e. $u_n \in C^\infty$) at the interior points of D(cf. [3]). At a straight line segment of the boundary, u_n can be reflected as an odd function accross the boundary. The resulting function satisfies $\Delta u+\lambda u=0$ in D in a whole neighborhood of that portion of the boundary and thus is C^∞ accross the boundary on straight line segments.

II.3. Nodal lines. The set of points in D where $u_n=0$ is the *nodal set* of u_n. Applying the unique continuation property, the nodal set consists of *curves* that are C^∞ in the interior of D. It is a very interesting property to be noted that where nodal lines cross, they form equal angles (cf [5]).

Courant's nodal line theorem [5] states that the nodal lines of the nth eigenfunction divide D into no more than (n-1) subregions which are called *nodal domains*. We note that u_1 has no interior nodes and thus λ_1 is an eigenvalue of multiplicity one. In the special case where D is a convex region, then u_1 has convex level curves (a fact which is not hard to be seen geometrically). Pleijel [16] has given an elegant proof of the nodal line theorem by applying the minimax property and unique continuation. It is an interesting fact to be noted that equality cannot hold for more than a finite number of n. This follows from the *Faber-Krahn inequality* ([7], [11]) for each nodal domain and *Weyl's law*, which is the asymptotic relation for the nth eigenvalue.

$$(32) \qquad \lambda_n \sim \frac{4\pi n}{A} \quad \text{as } n\to\infty$$

where A is the area of D.

It is a standard fact that the nth eigenvalue λ_n of D is the first eigenvalue for each of its nodal domains and a higher eigenvalue for a union of nodal domains.

A generalization of Courant's nodal domain Theorem.

In the following we outline J.Peetre's approach [15] for an extension of A.Pleijel's nodal domain theorem [16] to Riemannian manifolds.

Assume M is a 2-dimensional Riemannian manifold. The *Beltrami-Laplace operator* in M is

$$(33) \qquad \Delta = -g^{-\frac{1}{2}} \frac{\partial}{\partial x^j} \left(g^{\frac{1}{2}} g^{jk} \frac{\partial}{\partial x^k}\right),$$

where g_{kj} and g^{jk} are the covariant and contravariant components of the metric tensor in a local coordinate system and $g=\det g_{jk}$.

Assume now that D is a relatively compact connected domain in M. Consider the eigenvalue problem.

$$(34) \qquad \begin{aligned} \Delta u - \lambda u &= 0 \quad \text{in } D \\ u &= 0 \quad \text{on } \partial D \text{(boundary of D)} \end{aligned}$$

Our program is to compute the number of nodal domains N of the n-th eigenfunction of (34). We suppose that M is homeomorphic to a disk in the Euclidean plane.

Theorem 4.([15]). *Let* D_0 *be the least simply connected domain containing* D. *Suppose that*

$$(35) \qquad V_0 \sup_{D_0} K^+ \leq \pi,$$

where K *is the Caussian curvature,*

$K^+=\max(K,0)$, *and* V_0 *is the area of* D_0.

Then

$$(36) \qquad S^2 \geq 4\pi V \left(1 - \frac{1}{2\pi} \int_D K^+ dV\right),$$

where S *is the length of* ∂D *and* V *the area of* D. *Equality holds if and only if* $K=0$ *and* Ω *is a circle*

Proof ([15]). If D is simply connected ($D=D_0$) then (36) is a theorem of A.Huber (1954). If D is multiply connected then applying Huber's theorem to D_0 we obtain

$$(37) \qquad S_0^2 \geq 4\pi V_0 \left(1 - \frac{1}{2\pi} \int_{D_0} K^+ dV\right),$$

where S_0 measures the length of ∂D_0.

Suppose now that Σ is the interior of D_0-D and set $U=V_0-V$. Then we get

$$V_0 \int_{D_0} K^+ dV = V \int_D K^+ dV + V \int_\Sigma K^+ dV + U \int_{D_0} K^+ dV$$

$$\leq V \int_D K^+ dV + 2UV_0 \sup_{D_0} K^+$$

Therefore

$$(38)\qquad V(2\pi - \int_D K^+ dV) \leqq V_0(2\pi - \int_{D_0} K^+ dV).$$

Now (36) follows as a result of (37),(38) and $S \geqq S_0$. If equality holds in (36), then D must be simply connected and the last assertion of the theorem follows from Huber's theorem.

Q.E.D.

Theorem 5. ([15]). *Let* (36) *be satisfied and let* λ_1 *be the first eigenvalue of* (34).
Then

$$(39)\qquad \lambda_1 V \geqq \pi j^2 (1 - \frac{1}{2\pi} \int_D K^+ dV),$$

where j *is the first positive zero of the Bessel function* J_0. *Equality holds if and only if* K=0 *and* D *is a circle.*

Proof ([15]). Following the method of Faber [7] and Krahn [11] we can write:Let $u=u_1$ be the first eigenfunction. Set

$$(40)\qquad \begin{cases} D(\rho)=\{x \mid u(x)>\rho\}, \quad 0<\rho<\max u \\ \Delta(\rho)= \int_{D(\rho)} |\text{grad } u|^2 dV, \\ V(\rho)= \int_{D(\rho)} dV, \\ S(\rho)= \int_{\partial D(\rho)} dS, \\ H(\rho)= \int_{D(\rho)} u^2 dV. \end{cases}$$

Then

$$|\Delta'(\rho)| = -\Delta'(\rho) = \int_{\partial D(\rho)} |\text{grad } u| \, dS$$

and

$$|V'(\rho)| = -V'(\rho) = \int_{\partial D(\rho)} |\text{grad } u|^{-1} \, dS.$$

From Schwarz's inequality

$$(S(\rho))^2 \leqq |\Delta'(\rho)| \, |V'(\rho)|,$$

and from Theorem 4

$$(41)\qquad 4\pi(1 - \frac{1}{2\pi} \int_D K^+ dV) \; \frac{V(\rho)}{|V'(\rho)|} \leqq |\Delta'(\rho)|$$

If we apply a symmetrization process we get

$$(42) \qquad 4\pi\left(1-\frac{1}{2\pi}\int_D K^+ dV\right)\frac{\tilde{V}(\rho)}{|\tilde{V}'(\rho)|} \leq |\Delta'(\rho)|$$

Here we have replaced the domains $D(\rho)$ by concentric circles $\tilde{D}(\rho)$ with the same areas in the Euclidean plane, and the function u by a function $\tilde{u}$ which equals ρ on $\partial D(\rho)$.

It is true that $\tilde{V}(\rho)=V(\rho)$ and $\tilde{V}'(\rho)=V'(\rho)$. We also get

$$4\pi\frac{\tilde{V}(\rho)}{|\tilde{V}'(\rho)|} = |\tilde{\Delta}'(\rho)| \ ;$$

for $(\tilde{S}(\rho))^2=|\tilde{\Delta}'(\rho)||\tilde{V}'(\rho)|$ and $4\pi\tilde{V}(\rho)=|\tilde{S}(\rho)|^2$.

Therefore

$$\left(1-\frac{1}{2\pi}\int K^+ dV\right)|\tilde{\Delta}'(\rho)| \leq |\Delta'(\rho)|$$

Integrating over the interval $0<\rho<\max u$ we obtain

$$\left(1-\frac{1}{2\pi}\int_D K^+ dV\right)\tilde{\Delta} \leq \Delta$$

Also $\tilde{H}(\rho)=H(\rho)$ and $\tilde{H}=H$. It follows from Rayleigh's inequality that

$$\lambda_1 = \frac{D}{H} \ , \ \tilde{\lambda}_1 \leq \frac{\tilde{D}}{\tilde{H}};$$

and therefore (39) follows. If equality holds in (39), then the last assertion of the theorem follows from Theorem 4.

Q.E.D.

Theorem 6 ([15]). *There is a number $\alpha<1$ such that*

$$(43) \qquad \lim_{n\to\infty} \sup \frac{N}{n} \leq \alpha$$

Proof ([15]). Suppose now λ_n is the n-th eigenvalue and u_n is the n-th eigenfunction. Suppose also that $D_1, D_2, \ldots, D_N$ are the nodal domains of u_n. For each $D_l (l=1,2,\ldots,N)$ the value λ_n is the lowest eigenvalue. If we apply Theorem 5 to each D_l, we get

$$(44) \qquad \lambda_n V_l \geq \pi j^2 \left(1-\frac{1}{2\pi}\int_{D_l} K^+ dV\right)$$

If we take the sum of all inequalities (43) for $l=1,2,\ldots,N$ we obtain

$$\lambda_n V \geq \pi j^2 \left(N-\frac{1}{2\pi}\int_D K^+ dV\right)$$

But $\lim_{n\to\infty} n^{-1}\lambda_n V=4\pi$, therefore

$$\lim_{n\to\infty} \sup \frac{N}{n} \leq \left(\frac{2}{j}\right)^2<1. \tag{45}$$

Q.E.D.

Remark. It is easy to see that (43) remains true if (34) is replaced by an eigenvalue problem of the form

$$\Delta u+a(x)u=\lambda u,$$

where a(x) is a smooth bounded function.

Applying a similar argument, as in Theorem 5, we get [15]

$$(\lambda_1-\inf a(x))\, V \geq \pi j^2\left(1-\frac{1}{2\pi}\int_D K^+ dV\right), \tag{46}$$

and therefore (45) still follows.
It is now not difficult to extend the previous results to the case of a k-dimensional Riemannian manifold of constant curvature.

II.4. An orthogonal projection theorem for mappings and the Rayleigh quotient.

The Rayleigh quotient for the Jacobi operator $L[f]$ is defined by

$$R[f]=\frac{<L[f],f>}{<f,f>}$$

where L is defined in a Hilbert space H, with discrete point spectrum tending to infinity.

The eigenvalue λ_k of L, according to the classical principles of the Calculus of Variations (see for example [5]) of R.Courant and E.Fischer, can be written in te following form.

$$\lambda_k=\max_{W}\min_{f\perp W} R[f]=\min_{V}\max_{f\in V} R[f],$$

where W is any (k-1)-dimensional linear subspace of H and V is any k-dimensional linear subspace of H.Consider Σ_k to be a set (not a linear subspace), such that given any (k-1)-dimensional subspace W of H, there is a non-zero element of Σ_k that is orthogonal to W. The symbol k in this context is in order to know that Σ_k corresponds to the eigenvalue λ_k.

For any chosen $g\in\Sigma_k$ such that $g\perp W$, it follows that

$\min_{f \perp W} R[f] \leq R[g]$ and $R[g] \leq \max_{f \in \Sigma_k} R[f]$. Then $\min_{f \perp W} R[f] \leq \max_{f \in \Sigma_k} R[f]$

Then

$$\lambda_k = \max_{W} \min_{f \perp W} R[f] \leq \sup_{f \in \Sigma_k} R[f].$$

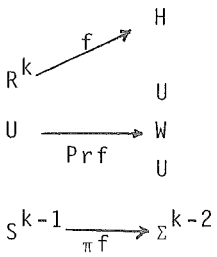
Therefore

$$\lambda_k \leq \sup_{f \in \Sigma_k} R[f]$$

This upper bound for the k^{th} eigenvalue λ_k, namely $\sup_{f \in \Sigma_k} R[f]$ may be a finite or infinite number and someone must be careful for a suitable choise of Σ_k in order for this upper bound to be a finite real number, and even more an accurate approximation of λ_k.

Proposition ([20]) *Let* H *be a Hilbert space and* $f:R^K \to H$ *a continuous mapping, homogeneous of odd degree (i.e.* $f(\lambda x)=\lambda^m f(x)$ *for some odd positive integer m) and satisfying* $f(x) \neq 0$ *for* $x \neq 0$. *Let* W *be a* (k-1)*-dimensional subspace of* H. *Then a vector* $x \neq 0$ *exists such that* $f(x) \perp W$

Proof. Assume that this is not the case and thus the mapping $f:R^K \to H$ has the property that for any W, a(k-1)-dimensional subspace of H, there is no vector $x \neq 0$ such that $f(x) \perp W$. Consider the orthogonal projection mapping Prf, $Prf:R^K \to W$, of $f:R^K \to H$, onto W.



Then $Prf(x) \neq 0$ for any $x \neq 0$, $x \in R^k$, and the mapping

$$\pi f = \frac{Prf}{||Prf||} : S^{K-1} \to \Sigma^{K-2}$$

is well defined, where

$$S^{K-1} = \{x \in R^K : ||x|| = 1\} \quad \text{and}$$

$$\Sigma^{K-2} = \{w \in W : ||w|| = 1\}.$$

Because of the fact $f:R^K \to H$ is a continuous mapping, homegeneous of

odd degree, i.e., $f(\lambda x)=\lambda^m f(x)$ for some odd positive integer $m, \lambda \in R$ and $x \in R^K$, it follows for $\lambda=-1$ that $f(-x)=-f(x)$ for $x \in R^K$. Therefore f is an odd mapping. However by the Borsuk-Ulam antipodal point theorem (see for example [21,p.266]) there is no such a mapping f, and thus we have proved that there exists a vector $x \neq 0$ such that $f(x) \perp W$.

Q.E.D.

Applications. Applying geometrical inequalities some very nice estimates can be deduced in function theory and in mathematical physics [1], [2].

We describe below a few results which are direct consequences of inequalities on two dimensional surfaces.

Suppose D is a simply-connected domain in the complex z- plane, $z_0 \in D$ an arbitrary point and

$$f(z)=(z-z_0)+a_2(z-z_0)^2+\ldots$$

a complex one-to one function mapping D conformally onto the circle $\{w: |w|<R_{z_0}\}$.
It follows from the Riemann mapping theorem that such a function exists and that R_{z_0} is uniquely defined. R_{z_0} *is called the conformal radius of* D *with respect to* z_0 *and*

$$\dot{R}=\sup \{R_{z_0} : z_0 \in D\}$$

is called the maximal conformal radius of D.

Pólya and Szegö [18] have discovered a fundamental inequality which relates the area A of D and the conformal radius:

$$\pi \dot{R}^2 \leq A$$

The equality sign being attained if and only if D is a circle. Consider in D a Riemannian metric $d\sigma^2=p ds^2$ of bounded Gaussian curvature K_0 and denote by A_σ the total area of D with respect to this metric. Then the following inequality (cf. [20]) holds:

$$R_z^2 \leq \frac{4A_\sigma}{p(z)(4\pi-K_0A_\sigma)} \quad \text{if } K_0A_\sigma<4\pi.$$

The above estimate holds for the maximal conformal radius if z is the point such that $R_z=\dot{R}$. Equality holds for the circle centered at the origin with the metric of constant Gaussian curvature K that is

$$d\sigma^2 = \frac{b}{\left(1+\frac{bK_o}{4}|z|^2\right)^2}\, ds^2 = e^{\hat{u}(r;b,K_o)}\, ds^2.$$

Because of the variational characterization of the eigenvalues upper bounds are relatively easier to construct and there are several isoperimetric inequalities providing such bounds (cf. [1], [2]). We would also like to mention the Pólya-Schiffers's inequality [17] concerning the connection of the maximal conformal radius with the sum of the reciprocal first n eigenvalues. Thi can be stated in the following way:

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ *be the first* n *eigenvalues of the fixed membrane equation in a simply connected domain* D *and let* $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n$, *be the corresponding eigenvalues of the circle of radius* 1. *Then*

$$\sum_{i=1}^{n} \lambda_i^{-1} \geq \dot{R}^2 \sum_{i=1}^{n} \hat{\lambda}_i^{-1},$$

where $\dot{R}$ *denotes the maximal conformal radius of* D.

This inequality has a natural extension to non-homogeneous membran [1], which can be stated as follows:

Theorem 7. ([1]). *Let* D *be a simply connected domain,* $z_0 \in D$ *an arbitrary point and* p *a mass density satisfying*

$$\Delta \log p + 2K_o p \geq 0, \text{ and}$$

$$K_o \int_D p\, dx \leq 2\pi$$

Set

$$\beta = p(z_0) R_{z_0}^2, \quad \text{and}$$

$$e^{\hat{u}(r,\beta;K_o)} = \frac{\beta}{\left(1+\frac{\beta K_o r^2}{4}\right)^2}.$$

Note that β *is a conformal invariant. If* $\hat{\lambda}_i$ *is the ith eigenvalu of*

$$\begin{cases} \Delta\hat{\phi} + \hat{\lambda}\, e^{\hat{u}(r,\beta;K_o)} \hat{\phi} = 0 & \text{in } \{x: |x|<1\} \\ \hat{\phi} = 0 & \text{on } \{x: |x|=1\} \end{cases}$$

then

$$\sum_{i=1}^{n} \lambda_i^{-1} \geq \sum_{i=1}^{n} \hat{\lambda}_i^{-1}.$$

Some sharper versions of Pólya and Schiffer's result for symmetric regions and an extension to multiply connected domains can be found in the very nice book of C.Bandle [1].

Very little is known for the *free membrane* described by the eigenvalue problem

$$\begin{cases} \Delta\psi+\nu\psi=0 & \text{in } D\subset\mathbb{R}^2, \\ \frac{\partial\psi}{\partial n}=0 & \text{on } \partial D, \end{cases}$$

where $\frac{\partial}{\partial n}$ denotes the outer normal derivative. By standard results there exists a countable number of eigenvalues $0=\nu_1<\nu_2\leq\ldots$ The following extremal property holds for a circle:
Among all domains of given area the circle yields the highest second eigenvalue ν_2.

This result can take the form of an inequality in the following way:

$$\nu_2\leq\frac{\pi p_1^2}{A},$$

where $p_1=1.841\ldots$ zero of the Bessel function J_1.
This result can easily be extended to the problem

$$\begin{cases} \Delta_S\psi+\nu\psi=0 & \text{in } D\subset\mathbb{R}^2 \\ \frac{\partial\psi}{\partial n}=0 & \text{on } \partial D \end{cases}$$

Theorem 8 ([1]). *Let* D *be a simply connected domain on* S *whose Gaussian curvature is bounded from above by* K_0. *If the total area* A_σ *of* D *satisfies* $K_0A_\sigma\leq 2\pi$, *then the value of*

$$\frac{1}{\mu_{\nu_2}}+\frac{1}{\mu_{\nu_3}}$$

takes its minimum for a geodesic circle on a surface of constant curvature K_0.

Nehari [13] considered membranes with *mixed boundary conditions*

$$\begin{cases} \Delta\phi+\mu\phi=0 & \text{in } D\subset\mathbb{R}^2 \\ \phi=0 & \text{on } \Gamma \\ \frac{\partial\phi}{\partial n}=0 & \text{on } \gamma \end{cases}$$

where $\Gamma\cup\gamma=\partial D$ and $\Gamma\cap\gamma=\phi$. Nehari proved the following theorem

Theorem. *Let γ be a concave arc. Then* $\mu_1 \geq \frac{\pi j_0^2}{2A}$ (μ_1 = *the lowest eigenvalue). Equality holds for semi-circles with* Γ *as circular arc and* γ *as the straight segment.*

Bandle [1] has generalized Nehari's theorem in various ways. In fact the concavity of γ has been dropped and extensions to inhomogeneous membranes have been considered. Then terms involving the curvature of γ enter into the inequalities.

From the topological and geometrical point of view the spectrum of the Laplacian on a Riemannian manifold has been studied, and some very useful estimates for the first non-trivial eigenvalue μ_1 have been investigated. Lichnerowicz and Obata have proved the interesting result that *for compact 2-dimensional manifolds of positive Gaussian curvature* $K(x) \geq \kappa_0 > 0$ *the following holds:*

$$\mu_1 \geq 2\kappa_0$$

The equality holds only for surface isometric to the sphere of radius $\frac{1}{\sqrt{\kappa_0}}$.

Hersch [8] has obtained some upper bounds for μ_1 in the case of a surface homeomorphic to the sphere. He has obtained among other results that

$$\left(\frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} \right) \frac{1}{A} \geq \frac{3}{8\pi} ,$$

where A denotes the area of the surface, with equality holding for the sphere.

In [20] we have investigated main topological and stability properties of some of the most important examples of complete minimal surfaces in R^3, by making use of the Morse-Smale index theorem (see also [19]) which we have formulated in terms of eigenvalues. This way we have completed a global analysis of the index for the stability of a complete minimal surface in R^3.

REFERENCES

1. C. Bandle, *Isoperimetric Inequalities and Applications*, Pitman Publ. London (1980).

2. C. Bandle, *Isoperimetric inequalities*, Convexity and its Applications (eds:P.M. Gruber and J.M. Wills), Birkhäser Verlag, Basel, 1983, pp 30-48.

3. D.L. Bernstein *Existence Theorems in Partial Differential Equations*, Annals of Math. Studies 23, Princeton Univ.Press, Princeton, NJ,1950.

4. J.J.A.M. Brands, Bounds for the ratios of the first three membrane eigenvalues, Arch. Rational Mech.Anal.16(1964), 265-268.

5. R.Courant and D.Hilbert, *Methods of Mathematical Physics*, Vol.I, Interscience, New York,1953.

6. L.Durand, *Stability and oscillations of a soap film:An analytic treatment*, Amer. J.Phys. 49 (1981),334-343.

7. G.Faber, *Beweis dass unter allen homogenen Membrane von gleicher Fläche und gleicher Spannung die kreisformige den tiefsten Grundton gibt*, Sitz. bayer. Akad.,Wiss.(1923), 169-172.

8. J.Hersch, *Quatre propriétés isopérimetriques de membranes spériques homogenes*, C.R. Acad. Sci. Paris A 270(1970), 1645-1648.

9. G.N.Hile and M.H. Protter, *Inequalities for eigenvalues of the Laplacian*, Indiana University Mathematics Journal 29 (1980), 523-538.

10. H.Von Helmholtz, *Die Lehre von den Tonempfindungen*, 1862.

11. E.Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaf des Kreises*, Math. Ann. 94(1924), 97-100.

12. J.R.Kuttler and V.G. Sigillito, *Eigenvalues of the Laplacian in two dimensions*, SIAM Review, 26(1984), 163-193.

13. L.Nehari, *On the principal frequency of a membrane*, Pac.J. Math. 8(1958),285-293.

14. L.E. Payne, G.Polya and H.F.Weinberger, *On the ratio of cons*

cutive eigenvalues, Journal of Math. and Physics 35(1956), 289-298.

15. J.Peetre, *A generalization of Courant's nodal domain theorem*, Math. Scand. 5(1957), 15-20.

16. A.Pleijel, *Remarks on Courant's nodal line theorem*, Comm. Pure Appl. Math. 9(1956), 543-550.

17. G.Pólya and M.Schiffer, *Convexity of functionals by transplantation*, J.d'Anal.Math. 3(1954), 245-345.

18. G.Pólya and G.Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Princeton University Press (1951).

19. Th.M.Rassias, *Sur la multiplicité du premier bord conjugué d'une hypersurface minimale de* R^n, $n \geq 3$, C.R. Acad.Sciences Paris 284(1977), 497-499.

20. Th. M.Rassias, *Foundations of Global Nonlinear Analysis*, Teubner-Texte zur Mathematik, Band 86, Leipzig,1986.

21. E.Spanier, *Algebraic Topology*, McGraw-Hill, New York,1966.

CLOSED FORM EXPRESSIONS FOR SOME SERIES INVOLVING BESSEL FUNCTIONS OF THE FIRST KIND

M.S. STANKOVIĆ, D.M. PETKOVIĆ and M.V. DJURIĆ

ABSTRACT: *During a few last years a large number of papers have been written on the summation of series of Bessel functions. Most of these works dealt with some particular cases of series (1) and (2). There are only two notable exceptions to these; works by M.L. Glasser,* [7] *and B. C. Berndt,* [2], [4], *that serves as excellent background for the advanced material discussed here. In this paper we evaluate and represent the series (1), (2) as the series over Riemann zeta and related functions, which degenerate in closed form formulas in certain cases.*

1. INTRODUCTION

In mathematical physics, particularly in certain problems of telecommunication theory, electrostatics, etc., one often requires numerical values of sums involving Bessel functions, (1) and product of Bessel functions, (2). So, it is useful to have closed form expressions of as many of these as possible.

$$(1) \qquad S_{\nu,\alpha} = \sum_{n=1}^{\infty} \frac{(s)^{n-1} J_{\nu}((an-b)x)}{(an-b)^{\alpha}} \qquad s=1 \text{ or } -1, \quad \mu,\nu,\alpha \in R, \quad \alpha > 0$$

$$(2) \qquad S_{\mu,\nu,\alpha} = \sum_{n=1}^{\infty} \frac{(s)^{n-1} J_{\mu}((an-b)x) J_{\nu}((an-b)x)}{(an-b)^{\alpha}}$$

$J_{\nu}(x)$ are Bessel functions of the first kind and of order ν.

Various special cases can be derived from the general forms (1),(2) and have been treated in [3], [17], [19], [22], [25] and [5], [6], [7], [21], [23] respectively. It seems unli-

kely that these series can be expressed in closed form when the only restrictions are those which are essential to secure the convergence.

Motivated by impossibility to obtain closed form formulas in the general cases, we find them, under some restrictions, for the most frequently occuring class of series, i.e. for $a=1$, $b=0$ and $a=2$, $b=1$, in terms of Riemann zeta functions and other known sums of reciprocal powers.

Inspired by closed form expressions of trigonometric series, the general terms of which are reciprocal powers of integral variable, [15], [20], [21], [22], we expanded an analytical procedure in order to obtain the formulas of interest.

2. PRELIMINARIES

This section deals with some results connected with trigonometric series (3), [20]

$$(3)\qquad \sum_{n=1}^{\infty} \frac{(s)^{n-1} f((an-b)x)}{(an-b)^{\alpha}} = c\,\frac{\pi}{2\Gamma(\alpha) f(\frac{\pi\alpha}{2})}\, x^{\alpha-1} + \sum_{i=o}^{\infty} \frac{(-)^{i} F(\alpha-2i-\delta)}{(2i+\delta)!}\, x^{2i+\delta},$$

$$\alpha \in R$$

$$\alpha > 0$$

where $f = \{\begin{smallmatrix}\sin\\ \cos\end{smallmatrix}\}$, $\delta = \{\begin{smallmatrix}1\\ 0\end{smallmatrix}\}$ and where all relevant parameters are given in the table I. $\zeta(\alpha)$, $\eta(\alpha)$, $\lambda(\alpha)$ and $\beta(\alpha)$ are Riemann zeta functions and other sums of reciprocal powers, [1], [8].

Note that when $f(x)=\sin x$ and $\alpha \to 2m$ or $f(x)=\cos x$ and $\alpha \to 2m+1$, $m \in N_o$, the limiting value of the right-hand side of (3) should be taken into account, [14], [20].

Another important occurence of (3) is when the right-hand side series truncate due to the vanishing of F functions, so in the completely different way one can get closed form

Table I: corresponding F and c

a	b	s	c	F	for
1	0	1	1	ζ	$0 < x < 2\pi$
		-1	0	η	$-\pi < x < \pi$
2	1	1	$\frac{1}{2}$	λ	$0 < x < \pi$
		-1	0	β	$-\frac{\pi}{2} < x < \frac{\pi}{2}$

Table II: closed form cases

F	f	α
ζ, η, λ	sin	$2m+1$
	cos	$2m$
β	sin	$2m$
	cos	$2m+1$

formulas as in [20], [21]. These cases, which are of great importance for our further discussion, are pointed out in the table II. Two formulas of that type are known from Cesaro´s work, 1936., see e.g. [20] and, as always, some really particular cases can be found in [1], [10], [18], [25]. If the paper [15] is not the compilation, then the author rediscovered the results from [20], [21].

It should be mentioned that when (3) has the closed form, it is a simple matter to obtain the following recursion formulas, [16]:

$$F(2m+\delta) = c\,\frac{(-)^{m+1}\pi^{2m}}{2(2m)!} + \sum_{i=1}^{m} \frac{(-)^{i+1}F(2m-2i+\delta)\pi^{2i}}{2^{2i\delta}(2i+1-\delta)!}, \quad m \geq 1,$$

Table III:
Corresponding c and δ

F	ζ	η	λ	β
c	1	0	$\frac{1}{2}$	0
δ	0	0	0	1

Namely, $\zeta(2m)$, $\eta(2m)$, $\lambda(2m)$ are in proportion to π^{2m} and $\beta(2m+1)$ to π^{2m+1}. This fact is very useful in all closed form formulas we discuss here. In [11] one can find corresponding formula for $\zeta(2m)$.

3. OUTLINE OF THE BASIC PROCEDURE

The procedure we shall use is based on undoubtedly well known integral representation of Bessel functions:

$$(4) \qquad J_\nu(z) = 2\,\frac{(\frac{z}{2})^\nu}{\Gamma(\frac{1}{2})\Gamma(\nu+\frac{1}{2})}\int_0^{\frac{\pi}{2}} \sin^{2\nu}\theta\cos(z\cos\theta)\,d\theta, \quad \mathrm{Re}\,\nu > -\frac{1}{2}\,.$$

We shall substitute (4) in (1). It also states that it is possible to interchange the order of summation and integration. When we use this fact, the series (1) can be presented as follows:

$$S_{\nu,\alpha} = \frac{2(\frac{x}{2})^\nu}{\Gamma(\frac{1}{2})\Gamma(\nu+\frac{1}{2})}\int_0^{\frac{\pi}{2}} \sin^{2\nu}\theta \sum_{n=1}^{\infty} \frac{(s)^{n-1}\cos((an-b)x\cos\theta)}{(an-b)^{\alpha-\nu}}\,d\theta, \quad \alpha-\nu>0$$

Obviously, the part of the integrand is the series of the type (3). Further, we use (3) and this procedure leads to the integral the type of which is:

$$\int_0^{\frac{\pi}{2}} \sin^{\mu-1}x\cos^{\nu-1}x\;dx = \frac{1}{2}B(\frac{\mu}{2},\frac{\nu}{2}), \qquad \mathrm{Re}\,\mu>0, \quad \mathrm{Re}\,\nu>0.$$

We shall not go into details and instead merely state the final result (6). The condition $\alpha-\nu>0$ restricts this result to be of the most general character. That's why we recall the integral representation of Bessel functions, but of integral order this time:

$$J_n(z) = \frac{1}{\pi}\int_0^{\pi}\cos(z\sin\theta-n\theta)\,d\theta, \qquad n\in\mathbb{N}_0\,.$$

The same procedure as above leads to the integrals of the type

$$\int_0^{\pi} \sin^{\mu}x\, f(\nu x)\,dx = \frac{\pi}{2^{\mu}} f\left(\frac{\nu\pi}{2}\right) \frac{\Gamma(\mu+1)}{\Gamma\left(\frac{\mu+\nu}{2}+1\right)\Gamma\left(\frac{\mu-\nu}{2}+1\right)}, \quad f=\begin{Bmatrix}\sin\\ \cos\end{Bmatrix}, \quad \mathrm{Re}\,\mu>-1$$

and finally to the result (7).

The treatment of the series over product of Bessel functions, (2), demands the different integral representation, [24]

$$(5) \qquad J_{\mu}(z)J_{\nu}(z) = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} J_{\mu+\nu}(2z\cos\theta)\cos(\mu-\nu)\theta\,d\theta, \qquad \mu,\nu \in R,\ \mu+\nu>-1$$

As it was done previosly, we insert the integral representation into the series under consideration. Changing the order of summation and integration gives

$$S_{\mu,\nu,\alpha} = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \cos(\mu-\nu)\theta \sum_{n=1}^{\infty} \frac{(s)^{n-1} J_{\mu+\nu}(2(an-b)x\cos\theta)}{(an-b)^{\alpha}}\,d\theta, \qquad \alpha>0\ .$$

The series in the integrand is of the type (1) and for $\alpha>\mu+\nu>-\frac{1}{2}$ has the sum (6) and for $\mu+\nu\in N_0$ has the sum (7). In this way we easily obtain the final results (8) and (9), where the integral of the type

$$\int_0^{\frac{\pi}{2}} \cos^{\mu}x\,\cos\nu x\,dx = \frac{\pi}{2^{\mu+1}} \frac{\Gamma(\mu+1)}{\Gamma\left(\frac{\mu+\nu}{2}+1\right)\Gamma\left(\frac{\mu-\nu}{2}+1\right)}, \qquad \mathrm{Re}\,\mu>-1$$

is tacitly used.

4. RESULTS AND DISCUSSION

We are now in position to give the sum of the series (1):

$$(6)\qquad S_{\nu,\alpha} = c\frac{\Gamma\left(\frac{\nu-\alpha+1}{2}\right)}{2\Gamma\left(\frac{\alpha+\nu+1}{2}\right)}\left(\frac{x}{2}\right)^{\alpha-1} + \sum_{i=0}^{\infty}\frac{(-)^i F(\alpha-\nu-2i)}{i!\,\Gamma(\nu+i+1)}\left(\frac{x}{2}\right)^{2i+\nu},$$

$$\alpha,\nu \in R,\qquad \alpha>0,$$

$$\alpha>\nu>-\frac{1}{2},$$

$$(7)\qquad S_{m,\alpha} = c\frac{(-)^{\frac{m-\delta}{2}}\pi}{2\Gamma\left(\frac{\alpha+m+1}{2}\right)\Gamma\left(\frac{\alpha-m+1}{2}\right)f\left(\frac{\pi\alpha}{2}\right)}\left(\frac{x}{2}\right)^{\alpha-1} + \sum_{i=0}^{\infty}\frac{(-)^i F(\alpha-m-2i)}{i!\,(m+i)!}\left(\frac{x}{2}\right)^{2i+m},$$

$$\alpha\in R,$$

$$\alpha>0,$$

where $m=\{{}^{2k+1}_{2k}\}$, $f=\{{}^{\sin}_{\cos}\}$, $\delta=\{{}^{1}_{0}\}$, $k\in N_o$ and where c and F are readable from the table I.

In the case $\alpha-\nu=2k+1$ in (6) and $m-\alpha=2k+1$ in (7), $k\in N_o$, one should work either with limiting values or with principal values of gamma functions.

The chief disadvantage of the formula (6) is unvalidity for $\nu>\alpha$ and therefore (7) is derived, but only for $m\in N_o$. Even in the case $\alpha=\nu=1$ formula (6) holds true, although it does not seem possible, and gives the same result as (7).

Based on Mellin transform, one can find in [17] slightly different and less general ($a=1$, $b=0$, $s=1$, $\alpha-\nu\neq 2k+1$, $\max\{1-\alpha,-\nu\}>\frac{3}{2}$) result than (6).

In spite of the simplicity of the applied procedure it seems that (6) and (7) are the best published results and have not been noticed until now, as far as the authors are informed.

Special, but very useful cases of (6) and (7), [17], [19], [22], [25], we get for $\alpha-\nu-\delta$ even, where δ is given in the table III. Then the right-hand series terminate due to the vanishing of F functions, as it is already pointed out. Particularly, for $s=-1$ and $m>\alpha\in N$ the sum (7) is equal to zero and

it is very useful in accelerating the convergence of certain class of Bessel series.

It is almost obvious that for $\nu=k+\frac{1}{2}$, $k\in N_o$, (6) reduces to:

$$\sum_{n=1}^{\infty}\frac{(s)^{n-1}j_k((an-b)x)}{(an-b)^{\alpha}}=c\,\frac{\sqrt{\pi}}{4}\,\frac{\Gamma(\frac{k-\alpha+1}{2})}{\Gamma(\frac{k+\alpha}{2}+1)}(\frac{x}{2})^{\alpha-1}+\sum_{i=o}^{\infty}\frac{(-)^i F(\alpha-k-2i)}{(2i)!!(2k+2i+1)!!}x^{2i+k},$$

$$\alpha\in R$$

$$\alpha>k$$

where $j_k(x)$ are the spherical Bessel functions of the first kind.

Here we note in passing that the closed form expressions exist for the desired sums for $\alpha-k-\delta$ even.

Let us represent now the sum of series (2):

$$(8)\qquad S_{\mu,\nu,\alpha}=c\,\frac{\Gamma(\alpha)\Gamma(\frac{\mu+\nu-\alpha+1}{2})}{2\Gamma(\frac{\alpha+\mu+\nu+1}{2})\Gamma(\frac{\alpha+\mu-\nu+1}{2})\Gamma(\frac{\alpha-\mu+\nu+1}{2})}(\frac{x}{2})^{\alpha-1}+$$

$$+\sum_{i=o}^{\infty}\frac{(-)^i\Gamma(2i+\mu+\nu+1)F(\alpha-\mu-\nu-2i)}{i!\,\Gamma(i+\mu+1)\Gamma(i+\nu+1)\Gamma(i+\mu+\nu+1)}(\frac{x}{2})^{2i+\mu+\nu},$$

$$\alpha,\mu,\nu\in R,\qquad \alpha>0$$

$$\alpha>\mu+\nu>-\frac{1}{2},$$

$$(9)\qquad S_{\mu,\nu,\alpha}=c\,\frac{(-)^{\frac{\mu+\nu-\delta}{2}}\pi\Gamma(\alpha)}{2\Gamma(\frac{\alpha+\mu+\nu+1}{2})\Gamma(\frac{\alpha-\mu-\nu+1}{2})\Gamma(\frac{\alpha+\mu-\nu+1}{2})\Gamma(\frac{\alpha-\mu+\nu+1}{2})f(\frac{\pi\alpha}{2})}(\frac{x}{2})^{\alpha-1}+$$

$$+\sum_{i=o}^{\infty}\frac{(-)^i\binom{2i+\mu+\nu}{i}F(\alpha-\mu-\nu-2i)}{\Gamma(i+\mu+1)\Gamma(i+\nu+1)}(\frac{x}{2})^{2i+\mu+\nu},$$

$$\alpha,\mu,\nu\in R,\qquad \alpha>0,$$

$$\mu+\nu\in N_o,$$

where $\mu+\nu=\binom{2k+1}{2k}$, $f=\binom{\sin}{\cos}$, $\delta=\binom{1}{0}$, $k\in N_o$. F and c are given in the table I, where $2x$ should be taken in instead of x.

Analogously to (6) and (7), limiting or principal values of gamma functions are necessary for $\alpha-\mu-\nu=2k+1$ in (8) and for $\mu+\nu-\alpha=2k+1$ in (9), $k\in N_o$.

The shortcoming of (8), $\alpha>\mu+\nu>-\frac{1}{2}$, we due to the condition in (6). To overcome this, we additionaly give (9), but only for $\mu+\nu\in N_o$.

The reader will observe that the results just established have more general character than those discovered in [6]; besides, one of them is wrong ($\mu=1$, $\nu=0$, $\alpha=1$, $s=1$). This note one can find in [23], which is partialy incorrect, too.

For $\alpha-\mu-\nu$ even and for $s=1$, $a=1$, $b=0$ we obtain results from [7].

According to the concept of this paper, we wish to have closed form expressions and we get them from (8) and (9) for $\alpha-\mu-\nu+1+\delta$ even, where δ is given in table III. In that way the problem stated in [5] is more generaly solved. Some of these results are also given in [21].

The strange opinion of some colleagues is that some special cases of (6), (7) and (8), (9) should be pointed out. Thus, from the rich variety of closed form formulas we consider in particular (6) or (7) which for $a=2$, $b=1$, $s=1$ and $\nu=0$, $\alpha=2$ degenerate in:

$$\sum_{n=1}^{\infty} \frac{J_o((2n-1)x)}{(2n-1)^2} = \frac{\pi^2}{8} - \frac{x}{2}, \qquad 0 \le x \le \pi .$$

The formula (7) for $a=1$, $b=0$, $s=-1$ and $\alpha=m$ gives:

$$\sum_{n=1}^{\infty} \frac{(-)^{n-1} J_m(nx)}{n^m} = \frac{x^m}{2^{m+1}\Gamma(m+1)} .$$

The both formulas are in full agreement with (8), page 634. and (2), page 635. in [24].

Formulas (8) and (9) for $a=1$, $b=0$, $s=1$ lead to the known results: (13) in [7] and (10) in [23].

The fulfillments of the rest of whishes, as above, has no practical sense at present.

5. CONCLUSION

The series over Bessel functions are extremely useful for both analysis of Bessel functions and various applications. One important class of problems is obtaining closed form formulas. Although most of these formulas have been known for a long time, it seems that this important problem nevertheless has not been solved entirely.

In this paper we give some closed form expressions for two classes of series and we belive that these formulas might be useful for reducing further series, the sums of which are not known, to simpler cases or to the series the sums of which are known now. Also, we belive that this paper does contain some simple but fresh ideas.

Acknowledgments: *The authors express their sincere thanks to professors L. Gatteschi, University of Turin and B. Danković, University of Nish for fruitful discussions.*

REFERENCES

[1] *Abramowitz, M.; Stegun, A.: Handbook of Mathematical Functions, with Formulas, Graphs and Mathematical Tables, Dover Publications, N.Y., 1972.*

[2] *Askey, R.A.: Theory and Application of Special Functions, Academic Press, Inc., N.Y., 1975.*

[3] Berkesh, B.: Einige Formeln uber unendlichen Reihen Besselscher Funktionen, *Glasnik mat.fiz.astr.*,10 (1955), 161-170.

[4] Berndt, B.C.: The evaluation of character series by contour integration, *Univ.Beograd.Publ.Elektrotehn.Fak.Ser.Mat.Fiz.*, 386 (1972), 25-29.

[5] De Doelder, P.J.: On a series of product of Bessel functions of integral order, *Simon Stevin*, oct., (1960), 54-57.

[6] De Doelder, P.J.: Two infinite sums, problem 79-12, *SIAM Rev.*, 21 (1979), 395-396.

[7] Glasser, M.L.: A class of Bessel summations, *Math. comp.*, 37 (1981), 54-57.

[8] Glasser, M.L.: The evaluation of lattice sums. I. Analytic procedure, *J.Math.Phys.*, 14 (1973), 409-413.

[9] Glasser, M.L.: Private communications, Clarkson University, 1987.

[10] Gradshtein, I.S.; Ryzhik, I.M.: Tablicy Integralov, Summ, Ryadov i. Proizvedenii, Nauka, Moskva, 1971.

[11] Janković, Z.: Two recurrence formulas for the sums S_{2k}, *Glas.Mat. Ser.II*, 8 (1953), 27-29.

[12] Korenev, B.G.: Vvedenie v Teoriyu Besselevykh Funktsii, Nauka, Moskva, 1971.

[13] Lossers, O.P.: Private communications, Eindhoven University

[14] Mitrinović, D.S.; Adamović, D.D.: Nizovi i Redovi, Naučna knjiga, Beograd, 1980.

[15] Moiseev, A.I.: O razlozhenii summ $\sum n^{-\alpha} \cos 2\pi n\theta$ i $\sum n^{-\alpha} \sin 2\pi n\theta$ po stepenyam θ, *Izv.Vyssh.Uchebn.Zaved.Mat.*, 4 (1986), 75-77, RZhMat, 9 Б 1795, 1986.

[16] Petković, D.M.: Problem H-381, *Fibonacci Quart.*, feb., (1985), 89.

[17] Petković, D.M.: Infinite sums of Bessel functions, problem 85-14, solution by Glasser, M.L., *SIAM Rev.*, 28 (1986), 402-403.

[18] Prudnikov, A.P.; Brychov, Yu.A.; Marichev, O.I.: Integraly i Ryady. Elementarnye Funktsii, Nauka, Moskva, 1981.

[19] Prudnikov, A.P.; Brychov, Yu.A.; Marichev, O.I.: Integraly i Ryady. Spetsialnye Funktsii, Nauka, Moskva, 1983.

[20] Slavić, D.V.: On summation of trigonometric series, Univ.Beograd.Publ. Elektrotehn.Fak.Ser.Mat.Fiz., 263 (1969), 103-114.

[21] Stanković, M.S.; Petković, D.M.: O sumiranju nekih redova pomoću Riemannovih zeta funkcija, Informatica'81, Ljubljana, okt., (1981), 3_106.

[22] Stanković, M.S.; Petković, D.M.; Djurić, M.V.: Short table of summable series of Bessel functions, Conf.Appl.Math.5., Ljubljana, sept.,(1986), 147-152, see RZhMat. 4 Б 16, 1987.

[23] Toshić, D.Dj.: Some series of product of Bessel functions, Univ.Beograd.Publ.Elektrotehn.Fak.Ser.Mat.Fiz., 678-715 (1980), 105-110.

[24] Watson, G.N.: A Treatise on the Theory of Bessel Functions, Cambridge Univ. Press, Cambridge, 1958.

[25] Zaezdnyi, A.M.: Garmonycheskii Sintez v Radiotekhnike i Elektrosvyazi, Energiya, Leningrad, 1972.

ASYMPTOTIC BEHAVIOUR OF THE OSCILLATION OF THE SEQUENCES OF THE LINEAR TRANSFORMATIONS OF THE FOURIER SERIES

VLADIMIR N. SAVIĆ

A B S T R A C T

In this paper we consider the asymptotic behaviour of

$$(I) \qquad \varepsilon_{mn}(W^r H^\omega;\ U_n) = \sup_{f \in W^r H^\omega} \|U_m(f) - U_n(f)\|_C$$

$$(m, n, r = 1, 2, \ldots;\ m > n)$$

where $U_n(f,x)$ is the sum of Fejér, Cesaro, Rogosinski, ... of the Fourier series of the function $f \in W^r H^\omega$.

ASIMPTOTSKO PONAŠANJE OSCILACIJE NIZA LINEARNIH TRANSFORMACIJA FOURIER-OVOG REDA FUNKCIJE f. U ovom radu razmatramo asimptotsko ponašanje izraza (I), gde je $U_n(f,x)$ suma Fejér-a, Cesaro-a, Rogosinskog, ... Fourier-ovog reda funkcije $f \in W^r H^\omega$.

If n is fixed, and m sufficiently large than ε_{mn} is approximatively equal the distance between U_n and f for each $f \in W^r H^\omega$.

Definition. Let $W^r H^\omega$ $(r \in N)$ be a set 2π-periodic continuous functions f, such that $f^{(r)} \in H^\omega$, or equivalent

$$(\forall x_1, x_2 \in R)\ |f^{(r)}(x_1) - f^{(r)}(x_2)| \leq \omega(|x_1 - x_2|)$$

where ω is the modulus of continuity.

For $f\in W^r$ and $f\in W^rH^\alpha$ $(0<\alpha\leqslant 1)$ we have [1] and [2] with the corresponding results.

The fundamental results follow from the lemma 1 (see [5]) and the lemma 2 (see [4])

Lemma 1. Let $\psi\in L[a,b]$, and suppose that

(i) $\quad \Psi(x)=\int_a^x \psi(t)dt$

(ii) $\quad \Psi\uparrow(\downarrow)$ on $]a,c[$ $\quad (a<c\leqslant b)$, and

$\quad \Psi\downarrow(\uparrow)$ on $]c,b[$

(iii) $\quad \Psi(b)=0.$

Then

$$\text{(II)}\quad \sup_{f\in H^\omega[a,b]}\left|\int_a^b \psi(t)f(t)dt\right| \leqslant \int_a^c |\psi(t)|\,\omega\big(\rho(t)-t\big)dt = \int_c^b |\psi(t)|\,\omega\big(t-\rho^{-1}(t)\big)dt,$$

where the function ρ is defined with

$$\Psi(x)=\Psi\big(\rho(x)\big)\quad \big(a\leqslant x\leqslant c\leqslant \rho(x)\leqslant b\big)$$

and ρ^{-1} is the inverse function of the function ρ.

If ω is a convex modulus of continuity, then, for the function $F(x)+C$ ($C\in R$ is arbitrary constant) we have $=$ in the formula (II), and

$$F(x)=\begin{cases} -\int_x^c \omega'\big(\rho(t)-t\big)dt, & a\leqslant x\leqslant c \\ \int_c^x \omega'\big(t-\rho^{-1}(t)\big)dt, & c\leqslant x\leqslant b \end{cases}$$

Lemma 2. Let (λ_{nk}) $(n,k\in N)$ be a matrix of real numbers such that $\lambda_{nk}=0$ for $k>n$, and the sequence

$$\left\{\frac{\lambda_{n+1,k}-\lambda_{nk}}{k^2}\right\}_{k=1,+\infty} \qquad (\forall n\in N)$$

is non - increasing. If

$$U_n(f,x) = \frac{1}{\pi}\int_0^{2\pi} f(x+t)\left[\frac{1}{2}+\sum_{k=1}^{n}\lambda_{nk}\cos kt\right]dt$$

then, for a convex modulus of continuity ω, for all $r\in N$ $(r \geq 3)$ and for all $m, n\in N$ $(m > n)$

$$\varepsilon_{m,n}(W^r H^\omega;\ U_n) =$$

$$= \begin{cases} \dfrac{2}{\pi}\displaystyle\sum_{k=0}^{\left[\frac{m-1}{2}\right]} \frac{\lambda_{m,2k+1}-\lambda_{n,2k+1}}{(2k+1)^r}\int_0^{\frac{\pi}{2}}\omega(2t)\sin(2k+1)t\,dt\ ; \quad r = 2i-1 \\ \qquad\qquad (i = 2, 3, \ldots) \\ \dfrac{1}{\pi}\displaystyle\int_0^{2\pi}\psi_{r,\lambda}(t)F_{r,\lambda}(t)dt; \quad r = 2i \quad (i = 2,3,\ldots) \end{cases}$$

where

$$F_{r,\lambda}(t) = \begin{cases} F^1_{r,\lambda}(t), & 0 \leq t \leq \pi \\ -F^1_{r,\lambda}(t), & \pi \leq t \leq 2\pi \end{cases},$$

$$F^1_{r,\lambda}(t) = \begin{cases} -\displaystyle\int_x^{t_0}\omega'\left(\rho(t)-t\right)dt, & 0 \leq x \leq t_0 \\ \displaystyle\int_{t_0}^{x}\omega'\left(t-\rho^{-1}(t)\right)dt, & t_0 \leq x \leq \pi. \end{cases},$$

t_0 is a zero of the function

$$\psi_{r,\lambda}(t) = \sum_{k=1}^{m}\frac{\lambda_{m,k}-\lambda_{n,k}}{k^r}$$

on $[0,\pi]$, and the function ρ is defined with

$$\int_0^x \psi_{r,\lambda}(t)dt = \int_0^{\rho(x)}\psi_{r,\lambda}(t)dt \quad (0 \leq x \leq t_0 \leq \rho(x) \leq \pi),$$

and ρ^{-1} is the inverse function of the function ρ.

Let, now, $(\sigma_n(f,x))$ be a sequence of the sums of Fejér of the Fourier series of the function f, i.e.

$$\sigma_n(f,x) = \frac{1}{\pi}\int_0^{2\pi} f(x+t)\left[\frac{1}{2} + \sum_{k=1}^{n}\left(1 - \frac{k}{n+1}\right)\cos kt\right]dt$$

where

$$\lambda_{nk} = \begin{cases} 1 - \dfrac{k}{n+1}, & k \leq n \\ 0, & k > n \end{cases}$$

Now, we prove

Theorem 1. For all $m, n \in N$ $(m > n)$ and for a convex modulus of continuity ω we have the asymptotic equality

$$\varepsilon_{m,n}(W^r H^\omega;\ U_n) = \begin{cases} C_1 \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\omega\left(\frac{1}{n}\right)\right)\right), & r = 1 \\ C_2 \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\frac{1}{n}\right)\right), & r = 2 \\ C_r \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\dfrac{1}{n^{r-1}}\right)\omega\left(\frac{1}{n}\right)\right), & r = 2i-1 \ (i = 2,3,\ldots) \\ C_r^1 \dfrac{m-n}{(m+1)(n+1)}\left(1 + O\left(\dfrac{1}{n^{r-1}}\omega\left(\frac{1}{n}\right)\right)\right), & r = 2i \ (i = 2,3,\ldots) \end{cases}$$

where

$$C_1 = \frac{2}{\pi}\sum_{k=0}^{\infty}\int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt$$

$$C_2 = -\frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\int_0^{2\pi} F_{2,\lambda}(t)\cos kt\,dt$$

$$C_r = \frac{2}{\pi}\sum_{k=0}^{\infty}\frac{1}{(2k+1)^{r-1}}\int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt$$

$$C_r^1 = -\frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k^{r-1}}\int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt$$

Proof. For $r = 2i-1$ and $r = 2i$ from the theorem 1 (see [4]) we have

$$(1)\quad \varepsilon_{mn}(W^r H^\omega;\ U_n) = \frac{2}{\pi}\left[\frac{m-n}{(m+1)(n+1)}\sum_{k=0}^{\infty}\frac{1}{(2k+1)^{r-1}}\int_0^{\frac{\pi}{2}}\omega(2t)\sin(2k+1)t\,dt -\right.$$

$$- \frac{m-n}{(m+1)(n+1)} \sum_{k=\left[\frac{n-1}{2}\right]+1}^{\infty} \frac{1}{(2k+1)^{r-1}} \int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt +$$

$$+ \frac{1}{m+1} \sum_{k=\left[\frac{n-1}{2}\right]+1}^{\left[\frac{m-1}{2}\right]} \frac{m-2k}{(2k+1)^{r}} \int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt \Bigg], \quad r = 2i-1, \quad (i = 2,3,\ldots)$$

$$(2)\quad \varepsilon_{mn}(W^r H^{\omega};\ U_n) = \frac{1}{\pi}\left[- \frac{m-n}{(m+1)(n+1)} \sum_{k=1}^{\infty} \frac{1}{k^{r-1}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt + \right.$$

$$+ \frac{m-n}{(m+1)(n+1)} \sum_{k=n+1}^{\infty} \frac{1}{k^{r-1}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt +$$

$$\left. + \sum_{k=n+1}^{m} \frac{k-(m+1)}{(m+1)k^{r}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt \right] = \sum^{0} + \sum^{1} + \sum^{2}, \quad r = 2i, \quad (i = 2,3,\ldots)$$

Since we have

$$\int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt = 0\left(\frac{1}{2k+1}\,\omega\left(\frac{1}{2k+1}\right)\right), \qquad (k \to +\infty)$$

from (1), we obtain the theorem 1, for $r = 2i-1$ $(i = 2,3,\ldots)$.

Let, now, $f_1(t) = F_{r,\lambda}(t) - F_{r,\lambda}(0)$,

$$D_n^{(r)}(t) = \sum_{k=n+1}^{\infty} \frac{\cos kt}{k^{r-1}}, \qquad D_{mn}^{(r)}(t) = \sum_{k=n+1}^{m} \left(1 - \frac{k}{m+1}\right)\cos kt,$$

then, by [3], we get

$$(3)\quad |\sum^{1}| = \left|\frac{1}{\pi}\int_0^{2\pi} f_1(t) D_n^{(r)}(t)\,dt\right| = 0\left(\frac{m-n}{mn^{r}}\,\omega\left(\frac{1}{n}\right)\right)$$

$$(4)\quad |\sum^{2}| = \left|\frac{1}{\pi}\int_0^{2\pi} f_1(t) D_{mn}^{(r)}(t)\,dt\right| = 0\left(\frac{m-n}{mn^{r}}\,\omega\left(\frac{1}{n}\right)\right)$$

From (3), (4) and (2) it follows the theorem 1 for $r = 2i$ $(i = 2,3,\ldots)$.

If $(\tilde{\sigma}_n(f; x)$ is a sequence of the conjugate sums of Fejér of the Fourier series of the function f, then, we have, by the theorem 2 from [4]

Theorem 2. For all m, n$\in$N (m > n) and for a convex modulus of continuity ω we have the asymptotic equality

$$\varepsilon_{mn}(W^r H^\omega;\ \tilde{\sigma}_n) =$$

$$= \begin{cases} \frac{2}{\pi}\,\frac{m-n}{(m+1)(n+1)}\left\{\bar{C}_r + O\left(\frac{1}{n^{r-1}}\,\omega\left(\frac{1}{n}\right)\right)\right\} & r = 2i \\ (i = 1,2,\ldots) \\ \frac{1}{\pi}\,\frac{m-n}{(m+1)(n+1)}\left\{\bar{\bar{C}}_r + O\left(\frac{1}{n^{r-1}}\,\omega\left(\frac{1}{n}\right)\right)\right\} & r = 2i+1 \\ (i = 1,2,\ldots) \end{cases}$$

where

$$\bar{C}_r = \sum_{k=0}^{\infty} \frac{1}{(2k+1)^{r-1}} \int_0^{\frac{\pi}{2}} \omega(2t)\sin(2k+1)t\,dt$$

$$\bar{\bar{C}}_r = -\sum_{k=1}^{\infty} \frac{1}{k^{r-1}} \int_0^{2\pi} F_{r,\lambda}(t)\cos kt\,dt.$$

R E F E R E N C E S

1. Рыжанкова Г. И. О колебаниях последовательностей некотор линейных преобразований рядов Фурье, автореферат кандидатской диссертации, Киев, 1972.
2. Филипповский В. Г. О колебании последовательности полиномов, порождаемых линейными метдами суммирования рядов Фурье на классах функций Гельдера, Сборник статей: "Теория приближения функций и ее приложения", Издание Института математики АН УССР, Киев, 1974, 158-181.
3. Ефимов А. В. Приближение непрерывных периодических функци суммами Фурье, Известия АН СССР, 24 (1960), 243-296.

. Savić V.N. The oscillation of the sequences of the linear transformations of the Fourier series of the function f. Collection of scientific papers of the Faculty of Science Kragujevac, 8(1987).

5. Savić V.N. O jednom ekstremalnom problemu u prostoru neprekidnih funkcija od n promenljivih. Mat.vesnik 6 (19) (34), 1982. 165-172.

lumerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

UNIFORMLY CONVERGENT SPLINE COLLOCATION METHOD FOR A DIFFERENTIAL EQUATION WITH A SMALL PARAMETAR

K. SURLA

ABSTRACT: *For the problem:* $\varepsilon y'' + p(x)y' = f(x)$, $-\alpha y(0) + y'(0) = \alpha_0$, $y(1) = \alpha_1$, $p(x) \geq \bar{p} > 0$, *the cubic spline collocation method is derived. The uniform convergence of the first order on locally bounded mesch is achived. The method has the second order of the convergence for fixed* ε.

1. INTRODUCTION

Consider the singularly perturbed two-point boundary value problem:

$$(1) \qquad \begin{cases} Ly = \varepsilon y'' + p(x)y' = f(x), & 0 \leq x \leq 1, \quad 0 < \varepsilon \ll 1, \\ y'(0) - \alpha y(0) = \alpha_0, & y(1) = \alpha_1; \quad \alpha_0, \alpha_1 \in \mathbb{R}, \quad \alpha \geq 0, \end{cases}$$

where the functions $p, f \in C^2[0,1]$, $p(x) \geq \bar{p} > 0$. Under these assumptions problem (1) has a unique solution $y = y(x)$, which exibits a boundary layer at $x = 0$ for small ε, [2].

The ordinary cubic spline collocation methods when applied to (1) have an inherent formal cell Reynolds number limitation, i.e. $h_j p(x_j)/2\varepsilon$ must be less than or equal to 1, [3]. For "small" ε this leads to the spurious oscillations or large inaccuracies in the approximate solution, (see [1],[2]), $h_j = x_{j+1} - x_j$, $j = 0(1)n$, x_j are the points of the grid Δ.

$$\Delta: \ 0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1.$$

In order to avoid these difficulties in case of Diriclet's boundary conditions in [5] the exponential features

of the exact solution are transfered to the spline coefficients by introducing the relaxation parameter affecting the highest derivative. This parameter is determined in such a way that the truncation error of the corresponding difference scheme for the boundary layer function in case $p(x) = p = \text{const}$, vanishes. This procedure is known as the exponential fitting or the introduction of "artificial viscosity". The spline difference schemes have the same order of accuracy on a uniform and non-uniform mesh [6]. It might be expected that the exponentially fitted spline difference schemes preserve this property. However, in the case of Dirichlet's conditions the uniform convergence in [5] is obtained by putting some special conditions on the grid. The dependence of the exact solution of ε in the case of mixed boundary conditions of the type (1) is smoller then in the case of Dirichlet's one. Because of that the first order of the uniform convergence can be obtained with properly bounded local mesh ration: $h_i/h_{i\pm1} \le M$, M is a constant independent of ε and h_i. Thus, in this case the exponentially fitted spline difference scheme has the same order of the accuracy on the equidistant grid and non-equidistant one (as in [6]).

2. DERIVATION OF THE SCHEME

We seek the solution of the problem (1) in the form of the cubic spline $v(x) \in C^2[0,1]$ on the grid Δ. On each interval $[x_i, x_{i+1}]$, spline $v(x)$ has the form:

$$(2) \qquad v(x) = v_j(x) = v_j^{(0)} + (x-x_j)v_j^{(1)} + (x-x_j)^2\frac{v_j^{(2)}}{2} + \frac{(x-x_j)^3}{6}v_j^{(3)}.$$

The constants $v_j^{(k)}$ are determined from the equations (see [3]):

$$(3) \qquad \varepsilon\bar{\sigma}_j v_j^{(2)} + p_j v_j^{(1)} = f_j, \quad j = 0(1)n+1,$$

$$\bar{\sigma}_j = \rho_j \operatorname{cth}\rho_j, \quad \rho_j = h_j p_j / (2\varepsilon),$$

$$(4) \qquad v_j^{(k)}(x_j) = v_{j-1}^{(k)}(x_j), \quad k = 0,1,2; \quad j = 1(1)n,$$

$$5) \qquad v_0^{(1)} - \alpha v_0^{(0)} = \alpha_0, \quad v_{n+1}^{(0)} = \alpha_1.$$

The system (3)-(5) has 4n+4 unknowns and 4n+4 equations. The first equation presents the collocation relaxed by introducing the parameter $\bar{\sigma}_i$ (fitting factor). Equations (4) are the consequence of the continuity conditions, $v(x) \in C^2[0,1]$. By eliminating the unknowns $v_j^{(k)}$, $k = 1,2,3$; from the above equations we obtain the scheme:

$$(6) \qquad \begin{aligned} & R_h v_j = Q_h f_j, \quad j = 0(1)n \\ & v_{n+1} = \alpha_1, \quad \text{where} \\ & R_h v_j = r_j^- v_{j-1} + r_j^c v_j + r_j^+ v_{j+1}, \quad \text{for } j = 0(1)n, \\ & Q_h f_j = q_j^- f_{j-1} + q_j^c f_j + q_j^+ f_{j+1}, \quad \text{for } j=1(1)n \quad \text{and} \end{aligned}$$

$$r_j^- = \frac{3(w_{j-1}-1)}{h_{j-1}A_{j-1}}, \quad r_j^+ = \frac{3(w_{j+1}+H_j)}{h_j A_j}, \quad r_j^c = -r_j^- - r_j^+ ,$$

$$A_j = 3w_j w_{j+1} + 2H_j w_j - 2w_{j+1} - H_j, \quad w_j = c\,\mathrm{th}\,\rho_j,$$

$$q_j^+ = H_j/(p_{j+1}A_j), \quad q_j^- = 1/(p_{j-1}A_{j-1}),$$

$$q_j^c = \frac{H_{j-1}(2w_{j-1}-1)}{p_j A_{j-1} w_j} + \frac{2w_{j+1}+H_j}{p_j A_j w_j} \quad , \text{ for } j = 1(1)n.$$

Further,

$$r_0^- = 0, \quad r_0^+ = \gamma_1^{-1}, \quad r_0^c = -(1+\gamma_1\alpha)\gamma_1^{-1}$$

$$Q_h f_0 = -\alpha_0 - s_1\gamma_1^{-1}, \quad \gamma_1 = h_0\left(1 - \frac{h_0 p_0}{3\sigma_0} - h_0\,\frac{b_1 p_1}{6\sigma_1 a_1}\right),$$

$$s_1 = \frac{h_0^2}{6}\left(2\,\frac{f_0}{\sigma_0} - \frac{p_1 R_1}{\sigma_1 a_1} + \frac{f_1}{\sigma_1}\right), \quad a_j = 1 + \frac{h_{j-1}p_j}{2\sigma_j}$$

$$b_j = 1 - \frac{h_{j-1}p_{j-1}}{2\sigma_{j-1}}, \quad R_1 = \frac{h_0}{2}\left(\frac{f_0}{\sigma_0} + \frac{f_1}{\sigma_1}\right), \quad H_j = \frac{h_j}{h_{j+1}}, \quad p_j = p(x_j),$$

$$f_j = f(x_j), \qquad \sigma_j = \varepsilon\bar{\sigma}_j.$$

3. THE PROOF OF THE UNIFORM CONVERGENCE

The proof is based on the comparison function method which requires the following lemmas, [1].

LEMMA 1. ([2]). Let $f,p \in C^2[0,1]$. Then the solution of (1) satisfies the inequalities

$$|y^{(i)}(x)| \le M(1+\varepsilon^{-i+1}\exp(-2\delta x/\varepsilon)), \quad i = 0(1)4.$$

M and δ are constants independent of ε.

LEMMA 2. (maximum principle)

Let $\{v_j\}$ be a set of values at the grid points x_j satisfying $R_h v_j \ge 0$, $j = 0(1)n$. Then, $v_j \le 0$, $j = 0(1)n$.

Throughout the paper M denotes the different constants independent of ε and h_i.

LEMMA 3. There exist constants M and β independent of h_j and ε such that for $j = 1(1)n$.

a) $R_h\phi_j \ge M\min(\frac{h_j^2}{\varepsilon^2}, 1)$,

b) $R_h\psi_j \ge M\mu_j(\beta)h_j^{-1}\min(h_j^3/\varepsilon^3, 1)$,

c) $R_h\phi_o \ge M$,

d) $R_h\psi_o \ge M\mu_o(\beta)h_o^{-1}\min(h_o/\varepsilon, 1)$,

$\phi_j = -2+x_j$, $\psi_j = -\exp(-\beta t_j)$, $\mu_j(\beta) = \exp(-\beta t_j)$,

$t_j = x_j/\varepsilon$.

Functions $\phi(x)$ and $\psi(x)$ are comparison functions and we use them in order to determine how the operator R_h affects the characteristic parts of the solution $y(x)$ (β is the smallest of various positive constants appearing in the proof). From Lemma 2 and Lemma 3 we can see that

$$|v_j - y_j| \le k_1|\phi_j| + k_2|\psi_j| \tag{7}$$

if

$k_1(h_j,\varepsilon) \geq 0$ and $k_2(h_j,\varepsilon) \geq 0$ are such functions that

$$R_h(k_1\phi_j + k_2\psi_j) \geq R_h(\pm z_j) = \pm\tau_j(y) \tag{8}$$

$z_j = y_j - v_j$, $\tau_j(y)$ is a truncation error of the scheme (6) for the function y. For an arbitrary smooth function g, $\tau_j(g)$ is given by

$$\tau_j(g) = R_h g_j - Q_h(Lg)_j \ .$$

LEMMA 4. The truncation error $\tau_j(y)$ can be written in the form

$$\tau_j(y) = R_h z_j = \left(\frac{a_j\phi_{j+1,2}}{\gamma_{j+1}} - b_j\frac{\phi_{j,2}}{\gamma_j} + \phi_{j,1}\right)/(w_j+H_{j-1})$$

$j = 1(1)n$, where

$$\phi_{j,1} = \psi_{j,1} - \frac{h_{j-1}}{2}\psi_{j,2} + \frac{h_{j-1}}{2}\left(\frac{\eta_{j-1}}{\sigma_{j-1}} + \frac{\eta_j}{\sigma_j}\right)$$

$$\phi_{j,2} = \psi_{j,0} + h_{j-1}^2\left(\frac{\eta_{j-1}}{3\sigma_{j-1}} + \frac{\eta_j}{6\sigma_j} - \frac{\psi_{j,2}}{6} + p_j\frac{\phi_{j,1}}{6a_j\sigma_j}\right),$$

$\eta_j = y_j''(\sigma_j - \varepsilon)$, $\sigma_j = \varepsilon\rho_j w_j$, $\psi_{j,k} = \dfrac{h_{j-1}^{4-k}}{(4-k)!}y^{IV}(\xi_j)$, ξ_j is a fixed point belongs to $[x_{j-1},x_j]$.

For the proof see [3] or [5].

THEOREM 1. Let $f,p \in C^2[0,1]$, $p(x) \geq \bar{p} > 0$. Let v_j be defined by (6) on the grid Δ, where $h_j/h_{j\pm1} \leq M$. Then

$$|y(x_j)-v_j| \leq Mh_j^2/(\varepsilon+h_j). \tag{9}$$

Proof. From Lemma 4 and Lemma 1 we have

$$|\tau_j(y)| \leq M\frac{h_j^2}{h_j+\varepsilon}\, h_j^2\varepsilon^{-2} + \exp(-\delta x_j/\varepsilon)h_j^4\varepsilon^{-4}, \quad h_j \leq \varepsilon, \ j=1(1)n. \tag{10}$$

From Taylor's development about x_j we also have

(11) $$|\tau_0(y)| \le Mh_0^2/(h_0+\varepsilon) + h_0^2\varepsilon^{-2}\exp(-\delta x_0/\varepsilon).$$

Further, for $\varepsilon \le h_j$, after several Taylor's expansions we obtain

$$\tau_j(y) = r_j^- \frac{h_{j-1}^2}{2} y''(\xi_{1j}) + r_j^+ \frac{h_j^2}{2} y''(\xi_{2j}) + q_j^- p_{j-1} h_{j-1} y''(\xi_{3j}) -$$

$$-q_j + p_j h_j y''(\xi_{4j}) - \varepsilon(q_j^- y''_{j-1} + q_j^c y''_j + q_j^+ y''_{j+1}) ,$$

$$x_{j-1} \le \xi_{1,j},\ \xi_{3,j} \le x_j \le \xi_{2,j},\ \xi_{4j} \le x_{j+1}.$$

Since

$$|w_j| \le M,\ h_j/h_{j\pm1} \le M,\ h^k/\varepsilon^k \exp(-\delta x_j/\varepsilon) \le M\exp(-\delta x_j/2\varepsilon),$$

we have $|r_j^{\pm}| \le Mh_j^{-1}$, $|q_j^{\pm c}| \le M$ and

(12) $$|\tau_j(y)| \le M(h_j + \exp(-\delta x_{j-1}/\varepsilon)),\ j=1(1)n.$$

In a similar way, we obtain, for j=0

(13) $$|\tau_0(y)| \le M(h_0 + \exp(-\delta x_0/\varepsilon)).$$

If we take $k_1(h_j,\varepsilon) = h_j^2/(h_j+\varepsilon)$ and $k_2(h_j,\varepsilon) = h_j^2/\varepsilon$ for $h_j \le \varepsilon$, from (10), (11) and Lemma 2 we can see that (8) holds.

This leads to the estimates (7) and (9).

If $\varepsilon \le h_j$ we can take $k_1(h_j,\varepsilon) = 1$, $k_2(h_j,\varepsilon) = h_j$ and from (12),(13) and Lemma 3 we have that (8) holds, and so does Theorem 1.

THEOREM 2. Let the conditions of Theorem 1 be satisfied. Then

(14) $$|y(x)-v(x)| \le Mh^2/(\varepsilon+h), \quad h = \max_i h_i,$$

M is a constant independent of ε and h.

Proof. Since $z(x)=y(x)-v(x) \in C^4[x_j,x_{j+1}]$ we have that:

$$(15)\qquad z(x) = z_j^{(0)} + z_j^{(1)}(x-x_j) + z_j^{(2)}\frac{(x-x_j)^2}{2} + z_j^{(3)}\frac{(x-x_j)^3}{3!} +$$

$$+ y^{IV}(\xi_j)\frac{(x-x_j)^4}{4!}\,,\quad x_j \le \xi_j \le x_{j+1},\quad x \in [x_j, x_{j+1}]$$

$$z_o^{(1)} - \alpha z_o^{(0)} = 0 \quad \text{and} \quad |z_o^{(1)}| \le Mh_o^2/(h_o+\varepsilon).$$

Further,

$$a_j z_j^{(1)} = b_j z_{j-1}^{(1)} + \phi_{j,1},\ \text{and}\ |z_j^{(1)}| \le Mh_j/(h_j+\varepsilon),\ j=1(1)n.$$

$$|z_j^{(2)}| \le |(\eta_j - p_j z_j^{(1)})/\sigma_j| \le M\varepsilon^{-1}h_j/(\varepsilon+h_j),$$

$$|z_{j-1}^{(3)}| \le (z_j^{(2)} - z_{j-1}^{(2)} - \psi_{j,2})/h_{j-1} \le Mh_j^{-1}/(h_j+\varepsilon).$$

After replacing these estimates in (15) we obtain estimate (14) for $h_j \le \varepsilon$.

In the case $\varepsilon \le h_j$ we can take the form

$$(16)\qquad |z(x)| = |z_j^{(0)}| + |(x-x_j)z'(\xi_j)|,\ x \in [x_j, x_{j+1}],$$

$$x_j \le \xi_j \le x_{j+1}\ .$$

From (3) and (4) we obtain

$$(17)\qquad a_j v_j^{(1)} = b_j v_{j-1}^{(1)}\ ,\quad j = 1(1)n,$$

$$(18)\qquad v_j^{(2)} = (f_j - p_j v_j^{(1)})/\sigma_j\ ,\ j = 0(1)n+1,$$

$$(19)\qquad v_j^{(3)} = (v_{j+1}^{(2)} - v_j^{(2)})/h_j,\ j = 0(1)n.$$

Since $|z_o^{(1)}| \le Mh_o^2/(h_o+\varepsilon)$ from Lemma 1 we have

$|v_o^{(1)}| \le M$. From (17) and $|a_j| \le M$, $|b_j| \le M$ we have $|v_j^{(1)}| \le M$, $j=0(1)n$. Because of that from (18) and (19) we obtain $|v_j^{(2)}| \le M/h_j$, $|v_j^{(3)}| \le M/h_j^2$. Since $|y^{(1)}(x)| \le M$ and

$$v_j^{(1)}(x) = v_{j-1}^{(1)} + h_{j-1}v_{j-1}^{(2)} + \frac{h_{j-1}^{(3)}}{2}v_{j-1}^{(3)}\ ,\ j=1(1)n,\ \text{we have}$$

$$|v_j^{(1)}(x)| \leq M \quad \text{and} \quad |z_j^{(1)}(x)| \leq M.$$

Thus, according to (16), Theorem 2 holds.

REFERENCES

[1] A.E.Berger, J.M.Solomon, M.Ciment: An Analysis of a Uniformly Accurate Difference Method for a Singular Perturbation Problem, Math.Comput. 37(1981), 79-94.

[2] U.K.Emeljanov: O raznostnom metode rešenija tretej krajevoj zadači dlja differencijal´nogo uravnenija s malym parametrom pri staršej proizvodnoj. Žurn. vyčislit. mat. i mat.fiz. (1975),15 No. 6. 1457-1465.

[3] V.P.Il´in: O splajnovih rešenijah obyknovenyh differencijal´nyh uravnenij. žur. vyčislit, mat. i mat. fiz. (1978), 3, 621-627.

[4] K.Surla: Singularly perturbed spline collocation method for boundary value problems with mixed boundary conditions, Zb.Radova Prir.-Mat.Fak. u Novom Sadu, Ser. za Mat., 16,2(1980), 132-143.

[5] K.Surla and M.Stojanović: Singularly perturbed spline difference schemes on non-equdistant grid. Z. angew. Math. Mech.68 (1988) 3, 171-180..

[6] Ju.S.Zavijalov, B.I.Kvasov, Z.L.Mirošničenko: Metody splajn funkcii, Moskva 1980.

THE MEASURE OF APPROXIMATION FOR THE PARTICULAR SOLUTION

DJ. TAKAČI

ABSTRACT. *We observe the linear partial differential equation in the field of Mikusinski operators, F, with homogeneous conditions. For the approximate particular solution constructed in [4] we construct and estimate new measures of approximation both in a subspace F_0 of the field, F, as well as in the space L of local-integrable functions.*

1. INTRODUCTION

The nonhomogeneous differential equation with constant coefficients

$$(1)\qquad \sum_{\mu=0}^{m}\sum_{k=0}^{n} \alpha_{\mu,k} \frac{\partial^{\mu+k} x(\lambda,t)}{\partial\lambda^{\mu}\partial t^{k}} = f_1(\lambda,t); \qquad \begin{matrix} 0 \le \lambda \le \lambda_1 \\ 0 \le t \le \infty \end{matrix}$$

with conditions

$$(2)\qquad \frac{\partial^{\mu+k} x(\lambda,0)}{\partial\lambda^{\mu}\partial t^{k}} = 0 \quad \text{for} \quad \begin{matrix} \mu = 0,\dots m \\ k = 0,\dots n-1 \end{matrix}$$

$$(3)\qquad \frac{\partial^{\mu} x(0,t)}{\partial\lambda^{\mu}} = 0 \quad \text{for} \quad \mu = 0,\dots m-2 \text{ and}$$

$$\frac{\partial^{m-1} x(0,t)}{\partial\lambda^{m-1}} = \frac{t^{r-1}}{\Gamma(r)} \quad \text{for} \quad r > 0$$

($f_1(\lambda,t)$ is a continuous function) corresponds in the field F to the equation

$$(4)\qquad \sum_{\mu=0}^{m}\sum_{k=0}^{n} \alpha_{\mu,k} s^{k} x^{(\mu)}(\lambda) = f(\lambda)$$

where s is the differential operator, ℓ is the integral operator, $s = \ell^{-1}$, and $f(\lambda) = \{f_1(\lambda,t)\}$ with the conditions

$$(5)\qquad x^{(\mu)}(0) = 0 \quad \text{for } \mu = 0,\dots m-2 \text{ and } x^{(m-1)}(0) = \ell^{r}.$$

The particular solution of equation (4) can be written in the form (see [1])

$$(6)\qquad x_p(\lambda) = \frac{1}{\ell^r a_m} \int_0^\lambda f(\kappa)x_h(\lambda-\kappa)d\kappa \ ,$$

where

$$a_m = \sum_{k=0}^{n} \alpha_{m,k}s^k \quad \text{and}$$

$$x_h(\lambda) = \sum_{j=o}^{m} b_j \exp(\lambda\omega_j), \quad b_j \text{ are operators, and}$$

$$\omega_j = \sum_{i=0}^{\infty} c_{i,j}\ell^{i\alpha_j-\beta_j} \ , \quad \alpha_j > 0, \ \beta_j \le 1.$$

The approximate particular solution of equation (4) can be treated in the form (see [4])

$$(7)\qquad x_{p,n} = \frac{1}{\ell^r a_m} \int_0^\lambda f(\kappa)x_{h,n}(\lambda-\kappa)d\kappa,$$

where

$$(8)\qquad x_{h,n}(\lambda) = \sum_{j=0}^{m} b_j \exp(\lambda\omega_{j,n}) \quad \text{and}$$

$$(9)\qquad \omega_{j,n} = \sum_{i=0}^{n} c_{i,j}\ell^{i\alpha_j-\beta_j} \ ,$$

The convergence in the space of locally integrable functions L, is the convergence in all seminorms

$$(10)\qquad \|f\|_T = \int_0^T |f(t)|dt \ .$$

L_o is the subspace of L consisting of all functions f, such that $\|f\|_T > 0$, for every $T > 0$, and F_o is the algebra of all operators of the form f/g where $f \in L$ and $g \in L_o$.

The convergence type I′ in F_o is equivalent to the convergence defined by the functional $A(\cdot)$ (see [3])

$$(11)\qquad A(x) = \sum_{i=0}^{\infty} \frac{\beta_{i,1/i}(x)}{e^{ie^{i^2}(1+\beta_{i,1/i}(x))}} \ , \ x \in F_o \ ,$$

where

$$(12)\qquad \beta_{T,\varepsilon}(x) = \inf\{\|f\|_T : x = f/g, \ \|g\|_T < 1, \ \|\ell-\ell g\|_T < \varepsilon\}$$

was introduced by Burzyk ([1]).

Also, we need the following definitions.

DEFINITION 1 ([3]). Operator $\tilde{x} \in F_o$ is the approximation of the operator $x \in F_o$, according to the functional $A(x)$ with the measure of approximation $\delta > 0$ if $A(x-\tilde{x}) < \delta$.

DEFINITION 2 ([3]). The function $\tilde{f}$ from L is the approximation of the function f from L according to the functional

$$F(f) = \sum_{i=1}^{\infty} \frac{\|f\|_T}{e^{ie^{i^2}}(1+\|f\|_T)} \tag{13}$$

with the measure of approximation $\delta_L > 0$ if $F(f-\tilde{f}) < \delta_L$.

2. THE ESTIMATIONS IN THE SPACE F_o

Supposing that

$$\frac{I}{\ell^{r_{a_m}}} f(\lambda) = \{f_2(\lambda,t)\}, \quad gx_{h,n}(\lambda) \quad \text{and} \quad gx_h(\lambda)$$

for $g \in F_o$ represent functions from L, then the operators

$$z_n(\lambda) = \frac{\frac{I}{\ell^{r_{a_m}}} \int_o^{\lambda} f(\lambda) gx_{h,n}(\lambda-\kappa)d\kappa}{g} \tag{14}$$

$$z(\lambda) = \frac{\frac{I}{\ell^{r_{a_m}}} \int_o^{\lambda} f(\kappa) gx_h(\lambda-\kappa)d\kappa}{g} \tag{15}$$

belong to F_o.

Denoting by

$$\frac{\{J_g(\lambda,t)\}}{g} := \frac{\frac{I}{\ell^{r_{a_m}}} \int_o^{\lambda} f(\kappa) g(x_{h,n}(\lambda-\kappa)-x_h(\lambda-\kappa))d\kappa}{g}$$

and using relation (12), we can write

$$\beta_{T,\varepsilon}(y_n(\lambda) - y(\lambda) \le \|J_{g_1}(\lambda,t)\|_T$$

where $g_1 = \frac{\ell}{I+k} \cdot k$ satisfy for, $k > 0$, the inequalities

$$\|g_1\|_T < 1, \quad \|\ell-\ell g_1\|_T < \frac{1}{k} \quad .$$

Now, using paper [3] we obtain

$$\|Jg_1(\lambda,t)\|_T \leq \lambda M(\lambda,t) \cdot \sum_{j=1}^{m} k_{g_1}^{M}(\lambda,T,\alpha_j,\beta_j) \cdot \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j-\beta_j-1}{2}+1\right)}$$

where

$$M(\lambda,T) = \max_{0\leq\kappa\leq\lambda} \int_0^T |f_2(\kappa,t)|dt \quad \text{and} \tag{16}$$

$$k_{g_1}^{M}(\lambda,T,\alpha_j,\beta_j) = \max_{0\leq\kappa\leq\lambda} k_{g_1}((\lambda-\kappa),T,\alpha_j,\beta_j)$$

(the constants $k_{g_1}((\lambda-\kappa),T,\alpha_j,\beta_j)$ may be obtained analogous as in [3])

So, we can prove easily

LEMMA 1. The function $A(z_n(\lambda) - z(\lambda))$, where $z_n(\lambda)$ and $z(\lambda)$ are given by relations (14) and (15), can be estimated as:

$$A(z_n(\lambda)-z(\lambda)) \leq \sum_{j=1}^{m} \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j-\beta_j-1}{2}+1\right)} Q_{g_1}^{p}(\lambda,\alpha_j,\beta_j) = \delta \tag{17}$$

where

$$Q_{g_1}^{p}(\lambda,\alpha_j,\beta_j) \geq \sum_{i=1}^{\infty} \frac{\lambda M(\lambda,i) k_{g_1}^{M}(\lambda,i\alpha_j,\beta_j)}{e^{ie^{i^2}}}$$

So, we have

THEOREM 1. The sequence of approximate solutions $(x_{p,n}(\lambda))_n$ converges to the exact solution $x_p(\lambda)$ in the I' type convergence.

On using definition 1. we can say that the measure of approximation in F_o is given by (17).

3. THE ESTIMATION IN THE SPACE L

Let us suppose that the exact and the approximate particular solutions are the functions from L. Then analogously as in paper [3] we can obtain the estimation:

$$\|x_{p,n}(\lambda) - x_p(\lambda)\|_T \leq \lambda M(\lambda,T) \sum_{j=1}^{m} k^M(\lambda,T,\alpha_j,\beta_j) \cdot \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j - \beta_j - 1}{2} + 1\right)}$$

where $M(\lambda,T)$ is given by (16) and

$$k^M(\lambda,T,\alpha_j,\beta_j) = \max_{0 \leq \kappa \leq \lambda} k(\lambda-\kappa,T,\alpha_j,\beta_j) \quad \text{see [3]}$$

Now, we can prove

LEMMA 2. If $x_{p,n}(\lambda)$ and $x_p(\lambda)$ are given by relations (7) and (6) and represent functions from L, then we have:

$$(18)\quad F(x_{p,n}(\lambda) - x_p(\lambda)) \leq \sum_{j=1}^{m} \frac{1}{\Gamma\left(\frac{(n+1)\alpha_j - \beta_j - 1}{2} + 1\right)} Q^p(\lambda,\alpha_j,\beta_j) = \delta_L$$

where

$$Q^p(\lambda,\alpha_j,\beta_j) \geq \sum_{i=1}^{\infty} \frac{M(i,\lambda) k^M(\lambda,i,\alpha_j,\beta_j)}{e^{ie^{i^2}}}$$

From previous Lemma follows

THEOREM 2. If $x_{p,n}(\lambda)$ and $x_p(\lambda)$ represent functions from L, then the sequence $(x_{p,n}(\lambda))_n$ converges to $x(\lambda)$ in L.

On using definition 2 the measure of approximation in L is given by (18).

It can be remarked that the measures of approximation obtained in this paper does not depend of the lenth of the interval.

EXAMPLE. The differential equation

$$\frac{\partial^2 x(\lambda,t)}{\partial\lambda\partial t} - \frac{\partial x(\lambda,t)}{\partial t} = x(\lambda,t) + f(\lambda), \quad 0 \leq \lambda \leq \lambda_o, \quad t \geq 0$$

with conditions

$$\frac{\partial x(\lambda,0)}{\partial\lambda} = 0, \ \lambda > 0, \quad x(0,t) = 1, \ t > 0$$

corresponds to the equation

$$(s-1)x'(\lambda) - x(\lambda) = f(\lambda)\ell \ ; \quad x(0) = \ell$$

in the field F

The exact particular solution is

$$x(\lambda) = \frac{1}{s-1} \int_0^{\lambda} f(\kappa)\exp((\lambda-\kappa)w)d\kappa$$

where

$$w = \sum_{i=0}^{\infty} \ell^{i+1}$$

and the approximate one is

$$x_n(\lambda) = \frac{1}{s-1} \int_0^{\lambda} f(\kappa)\exp((\lambda-\kappa)w_n)d\kappa$$

where

$$w_n = \sum_{i=0}^{n} \ell^{i+1}$$

However $x_n(\lambda)$ and $x(\lambda)$ belong to L we can find (for $\lambda = 1$ and $f(\lambda) = e^{\lambda}$)

$$\delta_L = \frac{1}{\Gamma(\frac{n+1}{2}+1)} \cdot e\left(\frac{e^{2e^2}}{e} + \frac{e}{e-1}\right) .$$

REFERENCES

1. BURZYK J.: On convergence in the Mikusinski operational calculus, Stud. Math. 75(1983) 313-333.
2. MIKUSINSKI J.: Operational calculus, Pergamon Press, Warsawa (1959).
3. PAP E., TAKAČI Dj.: "Estimations for the solutions of operator linear differential equations, Proc.GFCA-87.(in print).
4. TAKAČI Dj.: The approximate solution of a differential equation in many steps, Zb.radova PMF u Novom Sadu knjiga (1983) 51-61.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

EXPONENTIALLY FITTED QUADRATIC SPLINE DIFFERENCE SCHEMES

Z. UZELAC and K. SURLA

ABSTRACT: *For the problem (1.1) a family of difference schemes is derived using quadratic splines* $v(x) \in C^1[0,1]$. *The schemes are uniformly convergent with first order accuracy. Numerical examples are presented.*

1. DERIVATION OF THE SCHEMES

We consider collocation spline difference schemes for singularly perturbed two point boundary value problems

$$\begin{aligned} &Ly(x) \equiv \varepsilon y''(x) + p(x)y'(x) = f(x), \qquad 0<x<1, \\ &y(0) = \alpha, \quad y(1) = \beta, \end{aligned} \tag{1.1}$$

where ε is a small parameter in $(0,1]$, $p(x)$ and $f(x)$ are sufficiently smooth functions and $p(x) \geq p > 0$. Under these assumptions (1.1) has a unique solution $y(x)$ which in general displays a boundary layer at $x=0$ for small ε. The following lemma discribes some properties of the exact solution $y(x)$.

LEMMA 1.1 ([1]) *Let* $p(x)$, $f(x) \in C^3[0,1]$. *Then the solution of (1.1) can be written in the form* $y(x) = u(x) + W(x)$ *where*

$$u(x) = \varepsilon y'(0) \exp(-p(0)x/\varepsilon)/p(0) \tag{1.2}$$

$$|W^{(i)}(x)| \leq M(1+\varepsilon^{i-1}\exp(-\delta x/\varepsilon)), \qquad i=0,1,\ldots,4,$$

and M *and* δ *are constants independent of* ε.

Let us define the uniform mesh $\{x_j\}$, $j=0(1)n+1$ by $x_j=jh$ where n is an positive integer and the mesh length $h=1/(n+1)$. We will find an approximation to the solution $y(x)$ of (1.1) in the form of a quadratic spline $v(x) \in C^1[0,1]$ wich satisfies on the each interval $I_j=[x_j,x_{j+1}]$, $j=0(1)n$:

$$v_j(x) = v_j^{(0)} + (x-x_j)v_j^{(1)} + (x-x_j)^2 v_j^{(2)}/2.$$

The approximation to $y_j = y(x_j)$ is denoted by $v_j^{(0)} = v_j(x_j)$.

Let define the fitting "comparison" problem associated with (1.1) by:

$$
\begin{aligned}
&\tilde{L}\tilde{y}(x) \equiv \tilde{\sigma}(x,\varepsilon)\tilde{y}''(x) + \tilde{p}(x)\tilde{y}'(x) = \tilde{f}(x), \qquad 0<x<1 \\
&\tilde{y}(0) = \alpha, \quad \tilde{y}(1) = \beta,
\end{aligned} \tag{1.3}
$$

where $\tilde{\sigma}(x,\varepsilon)$, $\tilde{p}(x)$ and $\tilde{f}(x)$ are piecewise polynomial approximations to $\sigma(x,\varepsilon)$, $p(x)$ and $f(x)$ respectively (the fitting factor $\sigma(x,\varepsilon)$ will be determined). It is well-known that the solution $\tilde{y}(x)$ of the "comparison" problem ($\tilde{\sigma}(x,\varepsilon)=\varepsilon$) is a good approximation to the solution $y(x)$ of (1.1) (see Berger et al. [1]).

The unknown coefficients $v_j^{(k)}$, $k=0,1,2$, $j=0(1)n$ are determined from the conditions:

- $v(x)$ satisfies equation (1.3) at the points $x_{j+1/2} \equiv (x_j+x_{j+1})/2$, $j=0(1)n$ and the boundary conditions,
- $v(x) \in C^1[0,1]$.

The above conditions give the system of $3n+3$ unknowns with the same number of equations:

$$
\begin{aligned}
&\tilde{L}v_j(x)_{x=x_{j+1/2}} = \tilde{f}_j(x)_{x=x_{j+1/2}}, \qquad j=0(1)n, \\
&v_j(x)_{x=x_{j+1}} = v_{j+1}(x)_{x=x_{j+1}}, \qquad j=0(1)n-1, \\
&v_j'(x)_{x=x_{j+1}} = v_{j+1}'(x)_{x=x_{j+1}}, \qquad j=0(1)n-1, \\
&v_0(0) = \alpha, \quad v_n(1) = \beta.
\end{aligned} \tag{1.4}
$$

When $\tilde{\sigma}(x,\varepsilon)$, $\tilde{p}(x)$ and $\tilde{f}(x)$ are piecewise constant approximations of $\sigma(x,\varepsilon)$, $p(x)$ and $f(x)$ ($\tilde{p}_j(x)=\tilde{p}_j$, $x \in I_j$, etc.), the systeme (1.4) has the following form on the interval I_{j-1}:

$$
\begin{aligned}
&\tilde{\sigma}_{j-1} v_{j-1}^{(2)} + \tilde{p}_{j-1}\left(v_{j-1}^{(1)} + \frac{h}{2} v_{j-1}^{(2)}\right) = \tilde{f}_{j-1} \\
&v_{j-1}^{(0)} + h\, v_{j-1}^{(1)} + \frac{h^2}{2} v_{j-1}^{(2)} = v_j^{(0)} \\
&v_{j-1}^{(1)} + h\, v_{j-1}^{(2)} = v_j^{(1)}.
\end{aligned} \tag{1.5}
$$

By expessing $v_{j-1}^{(2)}$ from the first equation the system (1.5) has the following form:

$$
v_j^{(0)} = v_{j-1}^{(0)} + h\, v_{j-1}^{(1)} \tilde{\gamma}_{j-1} + \tilde{f}_{j-1} \tilde{S}_{j-1} \tag{1.6}
$$

$$
v_j^{(1)} = v_{j-1}^{(1)} \tilde{A}_{j-1} + h\, \tilde{f}_{j-1} / \tilde{s}_{j-1}, \tag{1.7}
$$

where

$$\tilde{s}_{j-1} = \tilde{\sigma}_{j-1} + h\,\tilde{p}_{j-1}/2, \quad , \quad \tilde{S}_{j-1} = h^2/(2\tilde{s}_{j-1})$$

$$\tilde{\gamma}_{j-1} = h(1-h\,\tilde{p}_{j-1}/(2\tilde{s}_{j-1})), \quad \tilde{A}_{j-1} = (1-h\,\tilde{p}_{j-1}/\tilde{s}_{j-1})$$

Similarly, we have that for $x \in I_j$:

$$(1.8) \qquad v_{j+1}^{(0)} = v_j^{(0)} + \tilde{\gamma}_j\, v_j^{(1)} + \tilde{S}_j \tilde{f}_j$$

$$(1.9) \qquad v_{j+1}^{(1)} = \tilde{A}_j\, v_j^{(1)} + h\tilde{f}_j/\tilde{s}_j.$$

From (1.6) we get:

$$v_{j-1}^{(1)} = (v_j^{(0)} - v_{j-1}^{(0)} - \tilde{S}_{j-1}\,\tilde{f}_{j-1})/\tilde{\gamma}_{j-1}$$

and from (1.8):

$$v_j^{(1)} = (v_{j+1}^{(0)} - v_j^{(0)} - \tilde{S}_j\tilde{f}_j)/\tilde{\gamma}_j.$$

By substituing the above expression for $v_{j-1}^{(1)}$ and $v_j^{(1)}$ into (1.7) we get a spline difference scheme which is a member of family of implicit schemes:

$$(1.10) \quad \frac{\tilde{A}_{j-1}}{\tilde{\gamma}_{j-1}} v_{j-1}^{(0)} - \left(\frac{\tilde{A}_{j-1}}{\tilde{\gamma}_{j-1}} + \frac{1}{\tilde{\gamma}_j}\right) v_j^{(0)} + \frac{1}{\tilde{\gamma}_j} v_{j+1}^{(0)} = \frac{\tilde{S}_j}{\tilde{\gamma}_j}\tilde{f}_j + \left(\frac{h}{\tilde{s}_{j-1}} - \frac{\tilde{A}_{j-1}\tilde{S}_{j-1}}{\tilde{\gamma}_{j-1}}\right)\tilde{f}_{j-1}$$

We introduce the following notation:

$$\tilde{r}_j^- = \tilde{A}_{j-1}/\tilde{\gamma}_{j-1}, \quad \tilde{r}_j^+ = 1/\tilde{\gamma}_j, \quad \tilde{r}_j^c = -\tilde{r}_j^- - \tilde{r}_j^+,$$

$$\tilde{q}_j^- = \frac{h}{\tilde{s}_{j-1}} - \frac{\tilde{A}_{j-1}\tilde{S}_{j-1}}{\tilde{\gamma}_{j-1}}, \quad \tilde{q}_j^c = \frac{\tilde{S}_j}{\tilde{\gamma}_j}, \quad \tilde{q}_j^+ = 0.$$

Then scheme (1.10) has the abbreviated form:

$$\tilde{R}v_j^{(0)} = \tilde{Q}\tilde{f}_j$$

where $\tilde{R}v_j^{(0)} = \tilde{r}_j^-\, v_{j-1}^{(0)} + \tilde{r}_j^c\, v_j^{(0)} + \tilde{r}_j^+\, v_{j+1}^{(0)}$

$$\tilde{Q}\tilde{f}_j = \tilde{q}_j^-\tilde{f}_{j-1} + \tilde{q}_j^c\tilde{f}_j + \tilde{q}_j^+\tilde{f}_{j+1}.$$

The truncation error for the boundary layer function u(x) (1.2) for

$p(x)=p=\text{const}$ is equal to zero when

$$(1.11)\qquad \tilde{r}_j^-/\tilde{r}_j^+ = \exp(-ph/\varepsilon).$$

If $\tilde{\sigma}(x,\varepsilon)=\sigma(\varepsilon)$ when $p(x)=p=\text{const}$ then condition (1.11) gives $\sigma(\varepsilon) = \frac{hp}{2}\,\mathrm{cth}(hp/(2\varepsilon))$. When $p(x)\neq\text{const}$ we define

$$\tilde{\sigma}_j(x,\varepsilon) = \frac{h\tilde{p}_j}{2}\,\tilde{\omega}_j, \quad x \in I_j \quad \text{where} \quad \tilde{\omega}_j = \mathrm{cth}(h\tilde{p}_j/(2\varepsilon)).$$

The coefficients of the family of the spline difference schemes defined by (1.10) have the following form

$$(1.12)\qquad \begin{aligned} &\tilde{r}_j^- = (1-1/\tilde{\omega}_{j-1})/h, \quad \tilde{r}_j^+ = (1+1/\tilde{\omega}_j)/h, \quad \tilde{r}_j^c = -\tilde{r}_j^+-\tilde{r}_j^-,\\ &\tilde{q}_j^- = 1/(\tilde{p}_{j-1}\tilde{\omega}_{j-1}), \quad \tilde{q}_j^c = 1/(\tilde{p}_j\tilde{\omega}_j), \quad \tilde{q}_j^+ = 0. \end{aligned}$$

The choice of approximation to $p(x)$ and $f(x)$ determines the particular scheme.

Let $\tilde{p}_{j-1} = \tilde{p}_j = p_j$, $\tilde{f}_{j-1} = \tilde{f}_j = f_j$

then the scheme (1.12) becomes $Rv_j^{(0)} = Qf_j$ where

$$r_j^- = (\omega_j-1)p_j/(2h), \quad r_j^+ = (\omega_j+1)p_j/(2h), \quad r_j^c = -\omega_j p_j/h,$$

$$q_j^- = q_j^+ = 0, \quad q_j^c = 1, \quad \omega_j = \mathrm{cth}(hp_j/(2\varepsilon)).$$

This scheme is precisely the Allen-Southwell-Il´in scheme for which first order uniform convergence at the nodes was proved in [3] and [4]. So, the quadratic spline difference scheme has the same property.

Choosing $\tilde{p}_{j-1} = (p_{j-1}+p_j)/2$, $\tilde{f}_{j-1} = (f_{j-1}+f_j)/2$,

$$\tilde{p}_j = (p_j+p_{j+1})/2, \qquad \tilde{f}_j = (f_j+f_{j+1})/2,$$

the corresponding implicit difference scheme has the coefficients:

$$(1.13)\qquad \begin{aligned} &r_j^- = (1-1/\tilde{\omega}_{j-1})/h, \quad r_j^+ = (1+1/\tilde{\omega}_j)/h, \quad -r_j^c = r_j^+ + r_j^-,\\ &q_j^- = 1/(2\tilde{p}_{j-1}\tilde{\omega}_{j-1}), \quad q_j^+ = 1/(2\tilde{p}_j\tilde{\omega}_j), \quad q_j^c = q_j^- + q_j^+. \end{aligned}$$

This scheme will be analysed in Section 2.

2. PROOF OF THE UNIFORM CONVERGENCE

The proof is based on the comparison functions method developed by Kellogg & Tsan [4] and Berger et al. [1] .

LEMMA 2.1 *Let* $\{V_j\}$ *be a set of values at the grid points* $\{x_j\}$, $j=0(1)n+1$ *satisfying* $V_0 \leq 0$, $V_{n+1} \leq 0$ *and* $RV_j \geq 0$, $j=1(2)n$. *Then* $V_j \leq 0$ *for* $j=0(1)n+1$.

We use two comparison functions $\phi_j=-2+x_j$ and $\psi_j=-\exp(-\beta x_j/\varepsilon)=-(\mu(\beta))^j$ where $\mu(\beta)=\exp(-\beta h/\varepsilon)$, $\beta>0$ will be chosen appropriately. Lemma 2.1 implies

LEMMA 2.2 *If* $K_1(h,\varepsilon) \geq 0$ *and* $K_2(h,\varepsilon) \geq 0$ *are functions that satisfy:*

$$R(K_1(h,\varepsilon)\phi_j + K_2(h,\varepsilon)\psi_j) \geq R(\pm Z_j) = \pm\tau_j(y)$$

where $Z_j = y_j - v_j^{(0)}$, *then*

$$|Z_j| \leq K_1(h,\varepsilon)|\phi_j| + K_2(h,\varepsilon)|\psi_j|.$$

Throughout the paper $\delta, M, M_1, \ldots$ will be used to denote generic constants independent of x, h and ε.

LEMMA 2.3 *There are constants* M_1 *and* M_2 *such that for* $h \leq M_1$, $0<\beta<M_2$ *and* $j=1(2)n$ *the following holds:*

(2.1) $\quad R\phi_j \geq Mh/\varepsilon , \qquad h \leq \varepsilon,$

(2.2) $\quad R\phi_j \geq M , \qquad \varepsilon \leq h,$

(2.3) $\quad R\psi_j \geq M\mu^j(\beta)h/\varepsilon^2, \qquad h \leq \varepsilon,$

(2.4) $\quad R\psi_j \geq M\mu^j(\beta)/h, \qquad \varepsilon \leq h.$

Proof. $R\phi_j=1/\tilde{\omega}_j+1/\tilde{\omega}_{j-1}$. Hence (2.1) and (2.2) holds. Now,

(2.5) $\quad R\psi_j = \mu^{j-1}(\beta)\, r_j^+ \,(1-\mu(\beta))(\mu\beta) - r_j^-/r_j^+).$

Let $h \leq C\varepsilon$ where C is a constant independent of h and ε, then $r_j^+ \geq M/h$, $r_j^-/r_j^+ \leq \exp(-\tilde{p}_j h/\varepsilon)+Mh$ and $\mu(\beta)-r_j^-/r_j^+ \geq M\mu(\beta)h/\varepsilon$, $1-\mu(\beta)=\beta h\exp(-Q\beta h/\varepsilon)/\varepsilon$, $0<Q<1$.

From (2.5) and the above estimates we see that (2.3) holds for $h \le C_1$.

Let $\varepsilon \le C^{-1}h$ for C sufficiently large. Then $r_j^+ \ge M/h$, $r_j^-/r_j^+ \le M\exp(-\tilde{p}_{j-1}h/\varepsilon)$, $\mu(\beta) - r_j^-/r_j^+ \ge M\mu(\beta)$, for appropriately chosen C and β. Moreover (2.4) holds for $\varepsilon \le C^{-1}h$. Since (2.3) holds for $h \le C\varepsilon$ we have $R\psi_j \ge M\mu^j(\beta)/\varepsilon \ge M\mu^j(\beta)/h$.

LEMMA 2.4 *The following estimates for the truncation error of the scheme (1.13) holds:*

$$|\tau_j(y)| \le M\left(\frac{h^2}{h+\varepsilon} \cdot \frac{h}{\varepsilon} + \frac{h^3}{\varepsilon^3} \cdot \exp\left(-\delta x_j/\varepsilon\right)\right), \quad j=1(2)n, \ h \le \varepsilon \tag{2.6}$$

$$|\tau_j(y)| \le M\left(h + \exp\left(-\delta x_{j-1}/\varepsilon\right)\right), \quad j=1(2)n, \ \varepsilon \le h. \tag{2.7}$$

Proof. Let $h \le \varepsilon$, then we take

$$\begin{aligned}\tau_j(y) = {} & T_0 y_j + T_1 y_j' + T_2 y_j'' + T_3 y_j''' + r_j^- R_3(x_j, x_j - h, y) + \\ & + r_j^+ R_3(x_j, x_j + h, y) - q_j^- \varepsilon R_1(x_j, x_j - h, y'') - \\ & - q_j^- p_{j-1} R_2(x_j, x_j - h, y') - \varepsilon q_j^+ R_1(x_j, x_j + h, y'') - \\ & - q_j^+ p_{j+1} R_2(x_j, x_j + h, y')\end{aligned}$$

where

$$Rn(a,b,g) = g^{(n+1)}(\xi)(b-a)^{(n+1)}/(n+1)! = \frac{1}{n!}\int_a^b (b-s)^n g^{(n+1)}(s)ds, \quad \xi \in (a,b).$$

Since $T_0 = T_1 = 0$ we will estimate T_2, T_3 and the remainder terms.

$$T_2 = h^2(r_j^+ + r_j^-)/2 + (p_{j-1}h - 2\varepsilon)q_j^- - (p_{j+1}h + 2\varepsilon)q_j^+.$$

Since $|hp\omega - 2\varepsilon| \le Mh^2/(\varepsilon+h)$ (see [4]) we have for

$p(x) = p = \text{const}$: $|T_2^c| \le M\frac{h^2}{h+\varepsilon} \cdot \frac{h}{\varepsilon}$.

Let $\rho_{j-1} = \tilde{p}_{j-1}h/(2\varepsilon)$, $r_j^- = r^-(\rho_{j-1})$, $r_j^+ = r^+(\rho_j)$, $q_j^- = q^-(\rho_{j-1})$, $q_j^+ = 1^+(\rho_j)$.

When $p(x) \ne \text{const}$ we expand T_2 at ρ_{j-1} and using the estimation for T_2^c we get $|T_2| \le M\frac{h^3}{\varepsilon(\varepsilon+h)}$.

Consider now

$$T_3 = h^3(r_j^+ - r_j^-) + (\varepsilon h - p_{j-1}h^2/2)q_j^- - (p_{j+1}h^2/2 + \varepsilon h)q_j^+.$$

By Taylor's expansion about ρ_{j-1} we get $|T_3| \le Mh^3/\varepsilon$.

The remainder terms are bounded by

$$Mh^3(1 + \exp(-2\delta x_j/\varepsilon)/\varepsilon^3).$$

Using Lemma 1.1 we have

$$|\tau_j(W)| \le M \frac{h^3}{(\varepsilon+h)\varepsilon}(1+\exp(-2\delta x_j/\varepsilon)/\varepsilon), \quad \text{for} \quad h \le \varepsilon. \tag{2.8}$$

Since $\tau_j(y)=\tau_j(u)+\tau_j(W)$ it remains to estimate $\tau_j(u)$. Let us denote by $\tilde{\tau}_j(u)$ the truncation error for $p(x)=p(0)=p$. As $\tilde{\tau}_j \equiv 0$ after some algebra we get

$$|T_2-\tilde{T}_2|u_j'' \le M(h^3/\varepsilon+h^3x_j/\varepsilon^2)u_j/\varepsilon^2,$$

and

$$|T_3-\tilde{T}_3|u_j''' \le M\, h^3x_j u_j/\varepsilon^4.$$

The remainder terms are bounded by $Mh^3\exp(-\delta x_j/\varepsilon)/\varepsilon^3$, thus

$$|\tau_j(u)| \le Mh^3x_j \exp(-\delta x_j/\varepsilon)/\varepsilon^3 \quad \text{for} \quad h \le \varepsilon. \tag{2.9}$$

From (2.8) and (2.9) we get (2.6).

Let $\varepsilon \le h$, then we consider the truncation error in the following form

$$\begin{aligned}\tau_j(y) = {} & T_2y_j'' + r_j^- R_2(x_j,x_{j-1},y) + r_j^+ R_2(x_j,x_{j+1},y) - \\ & - q_j^- \varepsilon\, R_0(x_j,x_{j-1},y'') - q_j^- p_{j-1} R_1(x_j,x_{j-1},y') - \\ & - q_j^+ \varepsilon\, R_0(x_j,x_{j+1},y'') - q_j^+ p_{j+1}\, R_1(x_k,x_{j+1},y').\end{aligned}$$

As before, we estimate $\tau_j(u)$ and $\tau_j(W)$ separately:

$$|\tau_j(W)| \le M(h+\exp(-\delta x_{j-1}/\varepsilon)), \quad \varepsilon \le h, \tag{2.10}$$

$$|\tau_j(u)| \le M \exp(-\delta x_{j-1}/\varepsilon), \quad \varepsilon \le h. \tag{2.11}$$

From (2.10), (2.11) we get that (2.7) holds.

THEOREM 2.1 *Let $p(x)$, $f(x) \in C^3[0,1]$ in (1.1). Let $\{v_j^{(0)}\}$, $j=0(1)n+1$ be the approximation to the solution $y(x)$ of (1.1) obtained using (1.13). Then, there exist constants M and δ independent of h and ε such that for $j=0(1)n+1$*

$$|v_j^{(0)}-y_j| \le M\left(\frac{h^2}{\varepsilon+h} + \frac{h^2}{\varepsilon}\exp(-\delta x_j/\varepsilon)\right), \quad h \le \varepsilon, \tag{2.12}$$

(2.13) $\quad |v_j^{(0)} - y_j| \le Mh(1+\exp(-\delta x_{j-1}/\varepsilon)), \quad \varepsilon \le h.$

Proof. From (2.8) and (2.9) we can see that the funcions $K_1(h,\varepsilon) = h^2/(h+\varepsilon)$ and $K_2(h,\varepsilon) = h^2/\varepsilon$ satisfy Lemma 2.2, and (2.12) hold. For $\varepsilon \le h$ we use $K_1(h,\varepsilon) = h$ and $K_2(h,\varepsilon) = h\exp(h\delta/\varepsilon)$. Using Lemma 2.2, (2.10) and (2.11) we get (2.13).

3. NUMERICAL RESULTS

We present the numerical results obtained by the scheme (1.13). We consider the following simple problems:

(3.1) $\quad \varepsilon y'' + y' = x, \quad y(0) = y(1) = 0$

which the solution is

$$y(x) = (\varepsilon - 1/2)(1-\exp(-x/\varepsilon))/(1-\exp(-1/\varepsilon)) - \varepsilon x + x^2/2,$$

and

(3.2) $\quad y'' + (1+x^2)y' = -(e^x + x^2), \quad y(0) = -1, \quad y(1) = 0.$

For each problem the mesh length $h = 1/J$ was succesively halved starting with $j = 16$ and ending with $J = 1024$. The maximum error et all the mesh points $E^\infty = \max_j |y_j - v_j^{(0)}|$ is listed in Table 2. under E^∞. The numerical rate of convergence is determined as in [2]:

$$\text{rate} \equiv (\ln Z_{K,\varepsilon} - \ln Z_{K+1,\varepsilon})/\ln 2,$$

where $\quad Z_{K,\varepsilon} = \max_j \left| v_j^{h/2^K} - v_j^{h/2^{K+1}} \right|, \quad K = 0(1)4$

and $v_j^{h/2^K}$ denotes the value of $v_j^{(0)}$ et the mesh point x_j for the mesh length $h/2^K$.

Table 1: Numerical rate of convergence for (1.13) applied to (3.2)

ε / K	1/2 rate	1/4 rate	1/8 rate	1/16 rate	1/32 rate	1/64 rate	1/128 rate	1/156 rate	1/512 rate
0	2.00	1.98	1.98	1.92	1.81	1.58	1.20	.98	.95
1	2.00	2.00	2.00	1.98	1.95	1.85	1.59	1.23	1.00
2	2.00	2.00	2.00	2.00	1.99	1.96	1.85	1.59	1.23
3	2.00	2.00	2.00	2.00	2.00	1.99	1.96	1.86	1.60
4	2.00	2.00	2.00	2.00	2.00	2.00	1.99	1.96	1.86

Table 2: Numerical results for (1.13) applied to (3.1)

K	ε / J	1/2 E∞ rate	1/4 E∞ rate	1/8 E∞ rate	1/16 E∞ rate	1/32 E∞ rate	1/64 E∞ rate	1/128 E∞ rate	1/156 E∞ rate	1/512 E∞ rate
	16	1.5 E-4	5.4 E-4	1.5 E-3	3.9 E-3	8.3 E-3	1.5 E-2	2.1 E-2	2.5 E-2	2.7 E-2
0		2.00	2.00	1.99	1.98	1.92	1.72	1.32	1.00	.95
	32	3.8 E-5	1.3 E-4	3.9 E-4	9.8 E-4	2.1 E-3	4.4 E-3	7.9 E-3	1.1 E-2	1.3 E-2
1		2.00	2.00	2.00	1.99	1.97	1.92	1.72	1.34	1.03
	64	9.6 E-6	3.3 E-5	1.0 E-4	2.4 E-4	5.5 E-4	1.1 E-3	2.3 E-3	4.0 E-3	5.7 E-3
2		2.00	2.00	2.00	2.00	1.99	1.98	1.91	1.71	1.34
	128	2.4 E-6	8.4 E-6	2.5 E-5	6.2 E-5	1.3 E-4	2.9 E-4	6.1 E-4	1.1 E-3	2.0 E-3
3		2.00	2.00	2.00	2.00	2.00	1.99	1.98	1.92	1.72
	256	6.0 E-7	2.1 E-6	6.2 E-6	1.5 E-5	3.5 E-5	7.4 E-5	1.5 E-4	3.1 E-4	6.0 E-4
4		2.00	2.00	2.00	2.00	2.00	2.00	1.99	1.98	1.92
	512	1.5 E-7	5.3 E-7	1.5 E-6	3.8 E-6	8.7 E-6	1.8 E-5	3.8 E-5	7.8 E-5	1.5 E-4
	1024	3.7 E-8	1.3 E-7	3.9 E-7	9.7 E-7	2.1 E-6	4.6 E-6	9.7 E-6	1.9 E-5	3.9 E-5

REFERENCES

1. A. E. Berger & J. M. Solomon & M. Ciment, *An Analysis of a Uniformly Accurate Difference Method for a Singular Perturbation Problem.* Math. Comput. 37 (1981), 79-94.

2. E. P. Doolan & J. J. Miller & W. H. A. Schilders, *Uniform Numerical Methods for Problems with Initial and Boundary Layers*. Dublin, Boole Press (1980)

3. A. M. Il´in, *Differencing scheme for a differential equation with a small parameter affecting the highest derivative*. Mat.Zametki 6 (1969) 237-248.

4. R. B. Kellogg & A. Tsan, *Analysis of some difference approximations for a singular perturbation problem without turning points*. Math.Comp. 32 (1978) 1025-1039.

ON A NUMERICAL SOLUTION OF A POWER LAYER PROBLEM

RELJA VULANOVIĆ

ABSTRACT: *A singularly perturbed boundary value problem, whose solution has a power boundary layer, is considered. A first order numerical method, uniform in the perturbation parameter, is constructed.*

1. INTRODUCTION

The following boundary value problem is considered :

$$Lu := -(\varepsilon+x)^2 u'' + c(x)u = f(x),\ x \in I := [0, 1], \tag{1a}$$

$$Bu := (u(0), u(1)) = (U_0, U_1), \tag{1b}$$

where ε is a perturbation parameter : $0 < \varepsilon \le 1$ (usually $\varepsilon << 1$). The functions c, f and numbers U_0, U_1 are given. We assume that

$$c, f \in C^1(I), \tag{2a}$$

$$c(x) \ge 0,\ x \in I, \tag{2b}$$

$$c(0) > 0. \tag{2c}$$

Under these assumptions we shall show that the unique solution u to the problem (1) has the form :

$$u(x) = s\,v(x) + z(x), \tag{3a}$$

where

$$s \in \mathbb{R},\ |s| \le M,\ v(x) = (1+x/\varepsilon)^{-r},\ r = (\sqrt{1+4c(0)} - 1)/2,$$

$$|z^{(i)}(x)| \le M(\varepsilon+x)^{1-i},\ i=1, 2, 3,\ x \in I. \tag{3b}$$

Here and troughout the paper M denotes any positive constant independent of ε. The function v is a power boundary layer function.

The asymptotic behaviour of the solutions to problems of the power layer type was considered in many papers by S. A. Lomov, see [3]. The numerical solution of power layer problems has not been investigated to the same extent as exponential layer problems. A numerical method for another power layer problem was given in [2].

Here we shall solve (1) numerically by using standard difference schemes on a special non-equidistant mesh which is dense in the layer. The mesh is generated by a suitable function. This approach has been applied successfully to various problems of exponential layer type, cf. [4], [5], for instance. Our main result is linear convergence uniform in ε.

2. ANALYSIS OF THE CONTINUOUS PROBLEM

After giving some lemmas we shall prove (3). For the technique cf. [1], [6].

LEMMA 1. *Let* (2) *hold. Then there exists a unique solution* $u \in C^3$ *to the problem* (1) *and it satisfies*

(4a) $$|u(x)| \le M, \quad x \in I,$$

(4b) $$|u'(0)| \le M/\varepsilon, \quad |u'(1)| \le M.$$

P r o o f : The operator (L, B) is inverse monotone and the uniqueness is guaranteed. The existence and uniform boundedness follow because there exist upper and lower solutions to (1), which are bounded uniformly in ε. Indeed, let $g(x) = M(2-x^2)$, where M is a constant (independent of ε) such that

$$g(t) \ge |U_t|, \quad t = 0, 1,$$

and

$$Lg(x) = 2M(\varepsilon+x)^2 + Mc(x)(2-x^2) \ge |f(x)|.$$

Such an M exists since

$$Lg(x) \ge \gamma M(2-\delta^2) \text{ if } 0 \le x \le \delta,$$

and

$$Lg(x) \ge 2M\delta^2 \text{ if } \delta \le x \le 1,$$

where δ is a number from $(0,1]$, such that

$$c(x) \geq \gamma > 0, \quad x \in [0,\delta],$$

(δ and γ are independent of ε). Then $g(x)$ is the upper solution and $-g(x)$ is the lower solution to the problem (1). Thus, (4a) is proved.

To prove (4b) use

$$u'(b) = u'(a) + \int_a^b (\varepsilon+x)^{-2}(cu-f)(x)\,dx$$

for some $a, b \in I$. Now using (4a) and the choice $b=0$ and $a \in (0,\varepsilon)$, such that $u'(a) = (u(\varepsilon)-u(0))/\varepsilon$, we get the first inequality in (4b). Similarly, with $b=1$ and $a \in (1/2,1)$, such that $u'(a) = 2(u(1)-u(1/2))$, the second inequality follows. □

We shall need estimates for the solution to the following auxiliary problem :

(5a) $$L_1 y := -((\varepsilon+x)^2 y')' + c(x)y = f(x), \quad x \in I,$$

(5b) $$By = (U_0, U_1),$$

LEMMA 2. *Let* (2) *hold and let* $y \in C^3(I)$ *be the solution to the problem* (5). *Then :*

(6) $$|y^i(x)| \leq M(\varepsilon+x)^{-i}, \quad i=0,1,2, \quad x \in I.$$

P r o o f : The case $i=0$ can be easily proved analogously to the proof of (4a). Furthermore, analogously to the proof of the first inequality in (4b) we can get $|y'(0)| \leq M/\varepsilon$. Then we have :

$$|y'(x)| = |\varepsilon^2(\varepsilon+x)^{-2}y'(0) + (\varepsilon+x)^{-2}\int_0^x (cy-f)(t)\,dt| \leq$$

$$\leq M(\varepsilon(\varepsilon+x)^{-2} + x(\varepsilon+x)^{-2}) \leq M/(\varepsilon+x),$$

hence (6) is proved for $i=1$. Then the estimate for $i=2$ follows directly from (5a). □

Now we can prove (3):

THEOREM 1. *Let* (2) *hold. Then the unique solution to the problem* (1) *satisfies* (3).

P r o o f : Let $s = u'(0)/v'(0) = -\varepsilon u'(0)/r$. Then because of (4) we have $|s| \le M$ and

$$z'(0) = 0, \ |z'(1)| \le M,$$
$$L_1 z'(x) = F(x), \ |F(x)| \le M, \ x \in I .$$

Indeed, $F = f' - c'u - s L_1 v'$, and

$$|L_1 v'(x)| = |r((r+1)(r+2) - 2(r+1) - c(x))(\varepsilon+x)^{-1}(1+x/\varepsilon)^{-r}| =$$
$$= |r(c(0) - c(x))(\varepsilon+x)^{-1}(1+x/\varepsilon)^{-r}| \le Mx(\varepsilon+x)^{-1} \le M.$$

Thus, z' satisfies a problem of type (5) and from Lemma 2 it follows

$$|z^{(i+1)}(x)| \le M(\varepsilon+x)^{-i}, \ i = 0, 1, 2. \ \square$$

3. THE DISCRETIZATION

The discretization mesh I_h has points:

(7a) $$x_i = \lambda(t_i), \ t_i = ih, \ i = 0, 1, \ldots, n, \ h = \frac{1}{n}, \ n \in \mathbb{N},$$

where λ is a mesh generating function, cf. [4], [5], of the form:

(7b) $$\lambda(t) = \begin{cases} \omega(t) := a\varepsilon((q/(q-t))^p - 1), & t \in [0, \alpha], \\ \pi(t) := \omega'(\alpha)(t-\alpha) + \omega(\alpha), & t \in [\alpha, 1]. \end{cases}$$

Here $\alpha \in (0,1)$ is given, $q = \alpha + \varepsilon^{1/(p+1)}$, $p \ge 1/r$ (r is given in (3)) and a is determined from the condition $\pi(1) = 1$. Hence, $\lambda \in C^1(I)$.

The properties of function λ are:

(8a) $$\lambda^{(i)}(t) \geq 0, \quad i=0, 1, 2, \quad t \in I,$$

(8b) $$\lambda'(t) \leq M, \quad t \in I,$$

(8c) $$\lambda(t) \geq M\varepsilon^{1/(p+1)}, \quad t \geq \alpha,$$

(8d) $$\lambda(t) \geq M\varepsilon n, \quad t \geq q - Mh > 0.$$

In this Section constants M will be independent of h as well.

Let $w_h = [w_0, w_1, \ldots, w_n]^T \in \mathbb{R}^{n+1}$ be a mesh function on I_h. Then the discrete problem corresponding to (1) reads:

$$w_0 = U_0,$$

(9) $$L^h w_i := -(\varepsilon + x_{i-1})^2 D_h^{\cdot\cdot} w_i + c(x_i) w_i = f(x_i), \quad i=1, 2, \ldots, n-1,$$

$$w_n = U_1,$$

where

$$D_h^{\cdot\cdot} w_i = 2\left(h_{i+1} w_{i-1} - (h_i + h_{i+1}) w_i + h_i w_{i+1}\right) / \left(h_i h_{i+1} (h_i + h_{i+1})\right),$$

$$h_i = x_i - x_{i-1}, \quad i=1, 2, \ldots, n.$$

Note the shift $\varepsilon + x_{i-1}$ (instead of $\varepsilon + x_i$) which is introduced for technical reasons, (see the proof of Theorem 2 below).

Rewrite (9) in the matrix form:

$$A_h w_h = d_h,$$

where $d_h = [U_0, f(x_1), f(x_2), \ldots, f(x_{n-1}), U_1]^T \in \mathbb{R}^{n+1}$ and $A_h \in \mathbb{R}^{n+1, n+1}$ is the corresponding tridiagonal matrix. Let $\|\cdot\|$ denote the maximum norm both in $\mathbb{R}^{n+1}$ and $\mathbb{R}^{n+1, n+1}$. Then we have

(10) $$\|A_h^{-1}\| \leq M,$$

provided that h is sufficiently small, but independent of ε. Indeed, A_h is an L-matrix and for $y_h = [\,2-x_0^2,\ 2-x_1^2,\ \ldots,\ 2-x_n^2\,]^T \in \mathbb{R}$ we have:

$$A_h y_h \geq M,$$

since

$$L^h(2-x_i^2) = 2(\varepsilon+x_{i-1})^2 + c(x_i)(2-x_i^2) \geq M$$

if h is sufficiently small (compare with the proof of (4a) in Lemma 1). Thus $A_h^{-1} \geq 0$ (to be understood componentwise) and the stability (10), uniform in ε, follows.

THEOREM 2. *Let* (2) *hold and let* u *be the solution to the problem* (1). *Let* w_h *be the solution to the discrete problem* (9) *on the mesh* (7) *with sufficiently small* h *independent of* ε. *Then:*

$$\|u_h - w_h\| \leq Mh,$$

where $u_h = [\,u(x_0),\ u(x_1),\ \ldots,\ u(x_n)\,]^T \in \mathbb{R}^{n+1}$.

P r o o f : Because of (10) it is sufficient to prove

(11a) $$|R_i(v)| \leq Mh\ ,\ i = 1,\ 2,\ \ldots\ ,\ n-1,$$

and

(11b) $$|R_i(z)| \leq Mh\ ,\ i = 1,\ 2,\ \ldots\ ,\ n-1,$$

where

$$R_i(g) := L^h g(x_i) - (Lg)(x_i) = -(\varepsilon+x_{i-1})^2 D_h^{\cdot\cdot} g(x_i) + (\varepsilon+x_i)^2 g''(x_i)$$

for any $g \in C^2(I)$. Let

$$R_i(g) = R_i^1(g) + R_i^2(g)\ ,$$

$$R_i^1(g) = \left((\varepsilon+x_i)^2 - (\varepsilon+x_{i-1})^2\right) g''(x_i),$$

$$R_i^2(g) = (\varepsilon+x_{i-1})^2 \left(g''(x_i) - D_h^{\cdot\cdot} g(x_i)\right).$$

In the next steps of the proof we shall use the Taylor expansion of R_i^2 , (3b) and (8).

First it is obvious that

$$|R_i^1(z)| \le Mh_i(x_i+x_{i-1}+2\varepsilon)/(\varepsilon+x_i) \le Mh$$

because (8b) implies $h_i \le Mh$. On the other hand

$$|R_i^2(z)| \le Mh(\varepsilon+x_{i-1})^2 \max_{x_{i-1}\le x\le x_{i+1}} |z^{(3)}(x)| \le Mh$$

and (11b) is proved.

Let us now prove (11a).

1^0 Let $t_{i-1} \ge \alpha$. By using (8 a, b, c) we get for k=1, 2

$$|R_i^k(v)| \le M h (\varepsilon+x_{i-1})^{-1}(\varepsilon/(\varepsilon+x_{i-1}))^r \le M h \varepsilon^r x_{i-1}^{-(r+1)} \le M h \varepsilon^{r-(r+1)/(p+1)} \le M h .$$

2^0 Let $t_{i-1} < \alpha$ and $t_{i-1} \le q-3h$. Then $t_{i+1} < q$ and $q-t_{i+1} \ge (q-t_i)/3$. Now for k=1,2:

$$|R_i^k(v)| \le M h \lambda'(t_{i+1})\varepsilon^r(\varepsilon+x_{i-1})^{-(r+1)} \le M h (q-t_{i+1})^{-(p+1)}(\varepsilon/(\varepsilon+\lambda(t_{i-1})))^{r+1} \le$$

$$\le M h (q-t_{i-1})^{p(r+1)-(p+1)} \le M h.$$

3^0 The remaining case is: $q-3h < t_{i-1} < \alpha$. Suppose that $q-3h > 0$. Then from (8d) it follows:

$$|R_i^k(v)| \le M (\varepsilon/(\varepsilon+x_{i-1}))^r \le M h^{pr} \le M h \ , \ k=1,2.$$

Hence (11a) is proved and so is the theorem. □

4. NUMERICAL RESULTS

We shall consider the following test problem :

$$-(\varepsilon+x)^2 u'' + u = x \ , \ u(0) = 1 \ , \ u(1) = 1 + (\varepsilon/(\varepsilon+1))^r, \ r = (\sqrt{5}-1)/2.$$

Its solution is given by

$$u(x) = (1 + x/\varepsilon)^{-r} + x.$$

Let $E = \| w_h - u_h \|$, using the notation of Theorem 2. We have the following tables :

Table 1. $p = 1/r$, $\alpha = 0.8$

E	ε				
	1. -3	**1. -6**	**1. -9 , 1. -12 , 1. -18**		
n = 20	.119	.149	.153		
n = 50	.0490	.0606	.0618		

Table 2. $p = 2/r$, $\alpha = 0.8$

E	ε				
	1. -3	**1. -6**	**1. -9**	**1. -12**	**1. -18**
n = 20	6.95-3	6.14-3	7.69-3	8.66-3	8.90-3
n = 50	1.13-3	1.09-3	1.45-3	1.65-3	1.69-3

The usual notation $1.-3 = 10^{-3}$ etc. is used.

Table 1 contains the results of the method described above. The linear convergence uniform in ε is obvious. However, for all problems whose solutions have form : $u(x) = Mv(x) + b(x)$, where $|b^{(i)}(x)| \leq M$, $i = 0, 1, 2, 3$, $x \in I$, and hence for this test problem, we can prove quadratic convergence uniform in ε if we take $p \geq 2/r$ and $(\varepsilon + x_i)^2$ instead of $(\varepsilon + x_{i-1})^2$ in (9). This is illustrated by the results in Table 2.

REFERENCES

[1] R.B.Kellogg, A.Tsan : Analysis of some difference approximations for a singular perturbation problem without turning points, Math.Comput. 32 (1978), 1025-1039.

[2] В.Д.Лисейкин : О численном решении уравнений со степенным погранслоем, Ж. вычисл. матем. и матем. физ. 26 (1986), 1813-1820.

[3] В. А. Треногин : Развитие и приложения асимптотического метода Люстерника-Вышика, Успехи матем. наук 25 (1970), 123-156.

[4] R.Vulanović : On a numerical solution of a type of singularly perturbed boundary value problem by using a special discretization mesh, Zb. Rad.Prir.-Mat.Fak.Univ.Novom Sadu, Ser.Mat. 13 (1983), 187-201.

[5] R.Vulanović : Mesh construstion for discretization of singularly perturbed boundary value problems, Ph. D. Thesis, University of Novi Sad, 1986.

[6] R.Vulanović : An exponentially fitted scheme on a non-uniform mesh, Zb. Rad.Prir.-Mat.Fak.Univ.Novom Sadu, Ser. Mat. 12 (1982), 205-215.

Numerical Methods and Approximation Theory III (Niš, August 18 - 21, 1987)

A PROBLEM ON SIMULTANEOUS APPROXIMATION AND A CONJUCTURE OF HASSON

S. ZHOU

Let $C^N_{[-1,\ 1]}$ be the class of functions, which have N continuous derivatives, $P_n(f;\ x)$ be the polynomial of best approximation of degree $\leq n$ to $f \in C_{[-1,\ 1]}$ and $\Delta_n(x)=(1-x^2)^{1/2}/n+1/$

$E_n(f)= \|f - P_n(f)\| = \max\limits_{-1\leq x\leq 1} |f(x) - P_n(f,\ x)|$.

Both important and interesting question of approximation theory is: Do the derivatives of polynomials of best approximati achieve the best approximation to derivatives of function? In p riodic case, this problem had been solved long before. In algeb ic case, a classical result is that, if $f(x) \in C^N_{[-1,\ 1]}$, then there exists a $P(x) \in \Pi_n$ for $x \in (-1,\ 1)$ such that .

(1) $$|f^{(k)}(x) - P^{(k)}(x)| \leq C(N)\Delta_n^{N-k}(x)\,\omega(f^{(N)},\ \Delta_n(x)),$$

for $0\leq k\leq N$, $n\geq N$, where $\omega(f,\delta)$ is the modulus of continuity of Π_n is the set of algebraic polynomials of degree $\leq n$, $C(N)$ is a constant only depending upon N.

Considering the inequality (1), we notice that $P(x)$ is nc necessary to be the polynomial of best approximation to $f(x)$. T refore, it is natural to ask: What can one say about $P_n(f,\ x)$? M.Hasson [1] and D.Leviatan [2] have studied this problem rece tly. The result of Leviatan is that, if $f(x) \in C^N_{[-1,\ 1]}$, then

(2) $$|f_n^{(k)}(x) - P_n^{(k)}(f,x)| \leq \frac{C(N)}{n^k}\,(\Delta_n(x))^{-k}\,E_{n-k}(f^{(k)}),$$

where $x \in (-1,\ 1)$, $0\leq k\leq N$, $n\geq N$. The new question is: Can ineq lity (2) be improved? About this, M. Hasson [1] raised a conje

e as follows:

Let $N \geq 1$, then there exsists a function $f_0(x) \in C^{2N}_{[-1,\ 1]}$ for $+1 \leq k \leq 2N$ such that

$$P_n^{(k)}(f_0,\ 1) \nrightarrow f_0(1), \quad n \to \infty .$$

In this paper, we shall give a positive result to this conecture. This means, that the inequality (2) can not be improved.

Theorem. Let $N \geq 1$, then there exists a function $f_0(x) \in C^{2N}_{[-1,1]}$ or $N+1 \leq k \leq 2N$ such that

$$\overline{\lim_{n \to \infty}} \left| f_0^{(k)}(1) - P_n^{(k)}(f_0,\ 1) \right| > 0 .$$

Proof. Let n be an odd number,

$$T_n(x) = \cos(n \arccos x) = \frac{n}{2} \sum_{k=0}^{[\frac{n}{2}]} a_k\, x^{n-2k} \text{ , where } a_0 = 2^n/n ,$$

$$a_k = (-1)^k \frac{(n-k-1)!}{k!\,(n-2k)!} 2^{n-2k} \ ; \ S_{n+1}(x) = x^{2N} T_{n+1}^{(2N)}(\sqrt{1-x^2}) = \sum_{k=0}^{\frac{n+1}{2}} b_k\, x^{n-2k+1},$$

here $b_0 = (-1)^{\frac{n+1}{2}-N} 2^{n-1}(n+1)n \ldots (n-2N+2)$,

$$H_{n+2N+1}(x) = \int_0^x dx_1 \int_0^{x_1} dx_2 \ \ldots \int_0^{x_{2N-1}} S_{n+1}(x_{2N})\, dx_{2N} = \sum_{k=0}^{\frac{n+2N+1}{2}} C_k\, x^{n+2N-2k+1} ,$$

here $C_o = C_o(n) = (-1)^{\frac{n+1}{2}-N} 2^{n-1} \dfrac{(n+1)n \ldots (n-2N+2)}{(n+2)(n+3)\ldots(n+2N+1)}$.

bviously,

$$H_{n+2N+1}^{(2N)}(x) = S_{n+1}(x),$$

sing the known inequality of Bernstein type

$$\left| T_{n+1}^{(2N)}(\sqrt{1-x^2}) \right| \leq M_N\, n^{2N}\, \Delta_n^{-2N}(\sqrt{1-x^2}) \leq M_N\, x^{-2N}\, n^{2N} ,$$

hence

$$\| H_{n+2N+1}^{(2N)} \| \leq C_1(N)\, n^{2N} . \tag{3}$$

Notice that on $n+2N+2$ points $t_k = \cos(k\pi/(n+2N+1))$, $k=0,\ldots,n+2N+1$,

we have

$$T_{n+2N+1}(t_k) = (-1)^k \|T_{n+2N+1}\| ,$$

so that

$$(4) \quad H_{n+2N+1}(x) - P_{n+2N}(H_{n+2N+1}, x) = 2^{-2N-n} C_0(n) T_{n+2N+1}(x).$$

In view of the extreme properties of Chebyshev polynomials ([1])

$$(5) \quad \left| T_m^{(k)}(\pm 1) \right| = \|T_m^{(k)}(x)\| = \sup\{ \|f^{(k)}\| : f \in \Pi_m, \ \|f\| \le 1 \} = C_2(k) m^{2k}.$$

Take $\{n_j\}$ to be a sequence of odd numbers with

$$n_j^{2N+2}/n_{j+1} \longrightarrow 0, \quad j \longrightarrow \infty ,$$

for example, $n_{j+1} = (2j+1) n_j^{2N+1}$, and define

$$\frac{H_{n_\ell+2N+1}(x)}{n_\ell^{2N+2}} = h_\ell(x), \quad \sum_{\ell=j}^{\infty} h_\ell(x) = f_j(x), \quad j=0,1,\ldots .$$

Let $1 \le i \le N$. Due to (3), $f_0(x) \in C^{2N}_{[-1,1]}$ and for $0 \le k \le 2N$ from

$$\| H_{n+2N+1}^{(k)} \| \le \| S_{n+1} \| \le M_N n^{2N} ,$$

we got

$$(6) \qquad \|f_j^{(k)}(x)\| \le C_3(N) n_j^{-1} , \quad 0 \le k \le 2N.$$

From (4) ve get

$$h_j(x) - P_{n_j+2N}(h_j, x) = 2^{-2N-n_j} C_0(n_j) n_j^{-2N-2} T_{n_j+2N+1}(x) ,$$

and if combine it with (5) we obtain

$$(7) \qquad \left| h_j^{(N+i)}(1) - P_{n_j+2N}^{(N+i)}(h_j, 1) \right| \ge C(N) n_j^{2i-2} .$$

Further, from $f_0(x) - P_{n_j+2N}(f_0, x) = f_j(x) - P_{n_j+2N}(f_j, x)$ we can write