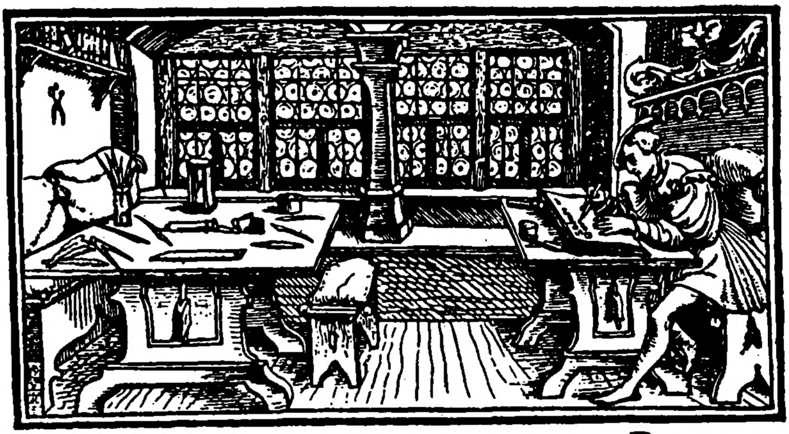
# STUDIA UNIVERSITATIS BABEŞ-BOLYAI

Mathematica

CLUJ-NAPOCA 1997



5 oct.

4

## COMITETUL DE REDACȚIE AL SERIEI MATHEMATICA:

**REDACTOR COORDONATOR:** Prof. dr. Leon ȚÂMBULEA

**MEMBRI**:

Prof. dr. Dorin ANDRICA
Prof. dr. Wolfgang BRECKNER
Prof. dr. Gheorghe COMAN
Prof. dr. Petru MOCANU
Prof. dr. Anton MUREȘAN
Prof. dr. Vasile POP
Prof. dr. Ioan PURDEA
Prof. dr. Ioan A. RUS
Prof. dr. Vasile URECHE
Conf. dr. Csaba VARGA

**SECRETAR DE REDACȚIE**: Lect. dr. Paul BLAGA

# STUDIA
# UNIVERSITATIS "BABEŞ-BOLYAI"

# MATHEMATICA
# 4

**Redacţia: 3400 Cluj-Napoca, str. M. Kogalniceanu nr. 1 • Telefon: 194315**

**SUMAR – CONTENTS – SOMMAIRE**

STUDIA UNIV. "BABEŞ-BOLYAI", MATHEMATICA, Volume XLII, Number 4, December 1997

# ON A FUNCTIONAL EQUATION

OCTAVIAN AGRATINI

**Abstract.** This is a survey paper devoted to the following functional equation

$$\sum_{k\in\mathbf{Z}} p_k u(x-k) = v(x), \quad x \in \mathbf{R},$$

which is in connection with the notion of wavelets. If $v(k)$ vanishes for $k \in \mathbf{Z}$ and if $p_k = 0$ for $k < 0$ and $k \geq m+1$, then, for $x = n$, the above equation leads us to the well-known general $m^{th}$-order linear recurrence relation. For $v(x) = u(2x)$, $x \in \mathbf{R}$, we present how this equation appears as a necessity in the field of mathematics. We also indicate three properties which must be fulfilled by the function and the sequence so that these equations admit solutions. When the sequence $(p_k)_{k\in\mathbf{Z}}$ has a compact support other properties are revealed and the technique to obtain solutions is described.

## 1. Introduction

Starting from the general $m^{th}$-order linear recurrence relation

$$\sum_{k=0}^{m} p_k u_{n-k} = 0, \quad m \geq 2,\ p_0 \neq 0,\ p_m \neq 0, \tag{1}$$

we consider a non-homogeneous equation as follows

$$\sum_{k=0}^{m} p_k u(x-k) = v(x), \quad x \in \mathbf{R}. \tag{2}$$

For $x = n$ and $v(\mathbf{Z}) = \{0\}$ we reobtain (1).

What happens if the left side of relation (2) contains an infinity of terms? In this paper we would like to study an equation of the following form

$$\sum_{k=-\infty}^{\infty} p_k u(x-k) = v(x), \quad x \in \mathbf{R}. \tag{3}$$

This equation raises new challanges such as: in which space of functions must we search the solutions and what kind of conditions must the sequence $(p_k)_{k\in\mathbf{Z}}$ fulfil

---

Received by the editors: June 12, 1997.

1991 *Mathematics Subject Classification.* 42C15, 11B37.

*Key words and phrases.* functional equations, reccurence relations.

to be compatible. Is such a research artificial or already necessary in the mathematical landscape? In the next section we will detail upon how such an equation can appear and the importance it takes. For this, we take a trip in the world of wavelets which represents a happy marriage between the results of the signal proccesing and the results in multiresolution analysis. Further on, we will list and prove some properties both of the function and the sequence which are involved in (3). In the last section we will relate the announced study under the assumption that the sequence has a finite support.

## 2. A sea of wavelets without water

We try to present the notion of wavelets. The standard references for this topic are Chui [1], Daubechies [2], Meyer [4] and the literature cited here. If we denote by $L_2(\mathbf{R})$ the space of square integrable functions defined on $\mathbf{R}$, we will refer to a function $f \in L_2(\mathbf{R})$ as being a signal with finite energy given by its norm $\|f\| = (f,f)^{1/2}$ We recall that the inner product of this space is defined by $(f,g) = \int_{\mathbf{R}} f(x)\overline{g(x)}dx$. It is well-known [3] that the Fourier transform of a function $f \in L_2(\mathbf{R})$ is given by

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{-i\xi x} f(x)dx$$

and the inverse Fourier transform is

$$f(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{i\xi x} \widehat{f}(\xi)d\xi.$$

For a given function $f$ we will use throughout the paper the notation $f_{j,k}(x) := 2^{j/2} f(2^j x - k)$. For any $j,k \in \mathbf{Z}$ we can write

$$\|f(2^j \cdot - k)\| = \left\{ \int_{\mathbf{R}} |f(2^j x - k)|^2 dx \right\}^{1/2} = 2^{-j/2}\|f\|.$$

This implies $\|f_{j,k}\| = \|f\|$, $j,k \in \mathbf{Z}$.

A multiresolution analysis (MRA) of $L_2(\mathbf{R})$ is defined as a sequence of closed subspaces $V_j$, $j \in \mathbf{Z}$, of $L_2(\mathbf{R})$ which enjoy the following properties

(i) $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots,$

(ii) $\bigcup\limits_{j\in\mathbf{Z}} V_j$ is dense in $L_2(\mathbf{R})$ and $\bigcap\limits_{j\in\mathbf{Z}} V_j = \{0\}$,

(iii) $v \in V_0 \;\Leftrightarrow\; v(\cdot + 1) \in V_0$,

$v \in V_j \;\Leftrightarrow\; v(2\cdot) \in V_{j+1}$, $j \in \mathbf{Z}$,

(iv) a function $\phi \in V_0$ exists such as the set $\{\phi_{0,k} : k \in \mathbf{Z}\}$ is a Riesz basis of $V_0$.

In accordance with this definition, if the subspace $V_0$ is generated by a single function $\phi \in L_2(\mathbf{R})$, that is $V_0 = \text{closure}_{L_2(\mathbf{R})}\langle \phi_{0,k} : k \in \mathbf{Z}\rangle$, then all the subspaces $V_j$ are also generated by the same $\phi$ namely $V_j = \text{closure}_{L_2(\mathbf{R})}\langle \phi_{j,k} : k \in \mathbf{Z}\rangle$. In fact, the set of functions $\{\phi_{j,k} : l \in \mathbf{Z}\}$ is a Riesz basis of $V_j$. We will name $\phi$ "the scaling function" or more suggestively "the father function". It is said that $\phi$ generates an MRA $\{V_j\}$ of $L_2(\mathbf{R})$. Since $\phi \in V_0 \subset V_1$ and $\{\phi_{1,k} : k \in \mathbf{Z}\}$ is a Riesz basis of $V_1$, consequently there exists a unique $l^2$-sequence $(p_k)_{k\in\mathbf{Z}}$ which describes $\phi$ that is $\phi(x) = \sum_{k\in\mathbf{Z}} p_k \phi_{1,k}(x)$. In other words, the father function satisfies the dilation equation

$$\phi(x) = \sum_{k=-\infty}^{\infty} p_k \phi(2x - k), \quad x \in \mathbf{R}. \tag{4}$$

This also called a "two-scale relation" of the function $\phi$. The sequence $(p_k)_{k\in\mathbf{Z}}$ becomes the "two-scale sequence" of $\phi$. Naturally, we consider that $\sqrt{2\pi}\hat{\phi}(0) = \int_{\mathbf{R}} \phi(x)dx$ is not zero. By integrating the relation (4) over $\mathbf{R}$, we can write

$$\int_{\mathbf{R}} \phi(x)dx = \sum_{k=-\infty}^{\infty} p_k \int_{\mathbf{R}} \phi(2x - k)dx = \frac{1}{2} \sum_{k=-\infty}^{\infty} p_k \int_{\mathbf{R}} \phi(y)dy.$$

This leads us to the following identity

$$\sum_{k=-\infty}^{\infty} p_k = 2. \tag{5}$$

At this point, we introduce $W_j$, the orthogonal complement space of $V_j$ in $V_{j+1}$, so that $V_{j+1} = V_j \oplus W_j$. We deduce that $W_j$, $j \in \mathbf{Z}$, are mutually orthogonal and

$$\bigoplus_{j=-\infty}^{\infty} W_j = L_2(\mathbf{R}).$$

Assuming that integer translates of $\phi$ generate an orthogonal basis (o.n.b.) for $V_0$, there exists a function $\psi \in W_0$ such as $\{\psi_{0,k} : k \in \mathbf{Z}\}$ forms an o.n.b. for $W_0$. At this moment the "mother wavelet" is born. Like the father $\phi$ generated an o.n.b. for $V_j$, the mother $\psi$ generates an o.n.b. for the orthocomplements $W_j$ of $V_j$, $j \in \mathbf{Z}$. It results that $\{\psi_{j,k} : (j, k) \in \mathbf{Z} \times \mathbf{Z}\}$ is an o.n.b. for $L_2(\mathbf{R})$. In fact, $\psi$ can be constructed as follows:

$$\psi(x) = \sum_{k=-\infty}^{\infty} (-1)^k \bar{c}_{1-k} \phi(2x - k) \tag{6}$$

is the general process to build wavelets bases

## 3. Other features of a father function

Applying the Fourier transform to (4), the dilation equation gives

$$\widehat{\phi}(\xi)=\sum_{k=-\infty}^{\infty}p_k\frac{1}{\sqrt{2\pi}}\int_{\mathbf{R}}\phi(2x-k)e^{-ix\xi}dx=\frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty}p_k2^{-1}e^{-ik\xi/2}\int_{\mathbf{R}}\phi(y)e^{-iy\xi/2}dy.$$

If we set $H(z):=\frac{1}{2}\sum_{k=-\infty}^{\infty}p_kz^k$, we obtain the following relation

$$\widehat{\phi}(2\xi)=H(e^{-i\xi})\widehat{\phi}(\xi). \tag{7}$$

By repeating $n$ times the relation (7) we get

$$\widehat{\phi}(2\xi)=\prod_{k=1}^{n}H(e^{-i\xi/2^{k-1}})\widehat{\phi}(\xi/2^{n-1}).$$

Since $\widehat{\phi}$ is a continuous function on $\mathbf{R}$ and assuming that $\widehat{\phi}(0)=1$, we easily deduce that

$$\widehat{\phi}(\xi)\to\prod_{k=1}^{\infty}H(e^{-i\xi/2^k}),$$

pointwise.

**Proposition 1.** *If $\phi$ defines an o.n.b. in $V_0$ the one has*

$$|h(\xi)|^2+|h(\xi+\pi)|^2=1, \tag{8}$$

*where $h(\xi)=H(e^{-i\xi})$.*

**Proof.** We can write successively:

$$\delta_{0,n}=(\phi_{0,0},\phi_{0,n})=\int_{\mathbf{R}}\phi(x)\overline{\phi}(x-n)dx=\int_{\mathbf{R}}e^{-in\xi}|\widehat{\phi}(\xi)|^2d\xi,$$

where we have used the Parseval identity. On the other hand,

$$\int_{\mathbf{R}}e^{-in\xi}|\widehat{\phi}(\xi)|^2d\xi=\sum_{k=-\infty}^{\infty}\int_{2k\pi}^{2(k+1)\pi}e^{-in\xi}|\widehat{\phi}(\xi)|^2d\xi=\int_0^{2\pi}e^{-in\xi}\left\{\sum_{k=-\infty}^{\infty}|\widehat{\phi}(\xi+2k\pi)|^2\right\}d\xi$$

Because $\frac{1}{2\pi}\int_0^{2\pi}e^{-in\xi}d\xi=\delta_{0,n}$, the above relations lead us to the following identity

$$\sum_{k=-\infty}^{\infty}|\widehat{\phi}(\xi+2k\pi)|^2=\frac{1}{2\pi}.$$

We choose $\xi:=2\xi$ and according to (7), we have

$$\frac{1}{2\pi}=\sum_{k=-\infty}^{\infty}|\widehat{\phi}(2\xi+2k\pi)|^2=\sum_{k=-\infty}^{\infty}|H(e^{-i(\xi+k\pi)})|^2|\widehat{\phi}(\xi+k\pi)|^2=\sum_{k\text{ even}}+\sum_{k\text{ odd}}=$$

$$= |H(e^{-i\xi})|^2 \sum_{k\in\mathbf{Z}} |\widehat{\phi}(\xi + 2k\pi)|^2 + |H(e^{-i(\xi+\pi)}| \sum_{k\in\mathbf{Z}} |\widehat{\phi}(\xi + (2k+1)\pi)|^2 =$$

$$= \frac{1}{2\pi}(|h(\xi)|^2 + |h(\xi+\pi)|^2).$$

We have used the fact that $h$ is $2\pi$-periodic.

Taking into account (5) we get $h(0) = H(1) = 1$. By using (8), it results $h(\pi) = H(-1) = 0$, in other words $\sum_{k=-\infty}^{\infty} (-1)^k p_k = 0$. We are now able to state another property in connection with equation (4).

**Proposition 2.** *If $\phi$ defines an o.n.b. in $V_0$ then the following identities*

$$\sum_{k\in\mathbf{Z}} p_{2k} = \sum_{k\in\mathbf{Z}} p_{2k+1} = 1$$

*hold.*

**Proposition 3.** *If $\phi$ is normalized, that is $\int_{\mathbf{R}} \phi(x)dx = 1$, then the following identities*

$$\sum_{k\in\mathbf{Z}} \phi(x-k) = \sum_{k\in\mathbf{Z}} \phi(k) = 1$$

*hold.*

**Proof.** If we put $s(x) = \sum_{k\in\mathbf{Z}} \phi(x-k)$, by using the dilation equation, we can write

$$s(x) = \sum_{k=-\infty}^{\infty} \left\{ \sum_{n\in\mathbf{Z}} p_n \phi(2x - 2k - n) \right\} = \sum_{k=-\infty}^{\infty} \left\{ \sum_{n \text{ even}} + \sum_{n \text{ odd}} \right\} =$$

$$= \sum_{k=-\infty}^{\infty} \left\{ \sum_{m\in\mathbf{Z}} p_{2m} \phi(2x - 2(k+m)) + \sum_{m\in\mathbf{Z}} p_{2m+1} \phi(2x - 2(k+m) - 1) \right\} =$$

$$= \sum_{l=-\infty}^{\infty} \phi(2x-2l) \left( \sum_{m\in\mathbf{Z}} p_{2m} \right) + \sum_{l=-\infty}^{\infty} \phi(2x-2l-1) \left( \sum_{m\in\mathbf{Z}} p_{2m+1} \right) = \sum_{l=-\infty}^{\infty} \phi(2x-l) = s(2x).$$

We have used Proposition 2. In this way, we have obtained $s(x) = s(2x)$ which implies $\widehat{s}(\xi) = 2\widehat{s}(2\xi)$. This represents a dilation equation with $p_0 = 2$ and all other coefficients are zero. The non trivial solution is $s = \delta$, the Dirac delta function ([3], Leçon $n^\circ$ 31). We deduce that $s$ is a constant. Taking $\sum_{k=-\infty}^{\infty} \phi(x-k) = \alpha$ and integrating over [0,1] we have

$$\alpha = \sum_{k=-\infty}^{\infty} \int_0^1 \phi(x-k)dx = \sum_{k=-\infty}^{\infty} \int_{-k}^{1-k} \phi(y)dy = \int_{\mathbf{R}} \phi(y)dy = 1.$$

For $x = 0$, it holds $1 = \sum_{k=-\infty}^{\infty} \phi(-k) = \sum_{k\in\mathbf{Z}} \phi(k)$ which completes the proof.

## 4. Particular approach

We are going to study a two-scale relation which is described by finite sums. We suppose that the integers $N' < N''$ exist, such as

$$(i) \quad p_{N'} \neq 0,\ p_{N''} \neq 0 \qquad (ii) \quad p_k = 0 \text{ for } k < N' \text{ and } k > N''. \tag{9}$$

We will only be concerned with scaling functions which are continuous everywhere. Because $\phi \in L_1(\mathbf{R}) \cap C(\mathbf{R})$ we are looking for the solutions $\phi$ with bounded support. Firstly, we specify that a general method for constructing the scale function $\phi$ is by using iterates and which does not involve $\widehat{\phi}$. In fact, $\phi$ solves (4) if $T(\phi) = \phi$ where $T(\phi) = \sum_{k\in\mathbf{Z}} p_k\phi(2x-k)$. We try to find this fixed point as usual: find a suitable $\phi_0$, define $\phi_n = T^n\phi_0$, and prove that $\phi_n$ has a limit. In this way, $\phi(x) = \lim_{n\to\infty}\phi_n(x)$. As a consequence of this recursive scheme we can expose

**Proposition 4.** *If the relation (9) is fulfilled then*

$$\operatorname{supp}\phi \subset [N', N''] \quad \textit{and} \quad \operatorname{supp}\psi \subset \left[\frac{1+N'-N''}{2}, \frac{1-N'+N''}{2}\right]$$

*hold, where $\phi$ and $\psi$ satisfy the equations (4) respectively (6).*

**Proof.** We use the recursive scheme and choose $\phi_0$ with compact support. Let's take $\operatorname{supp}\phi_0 = [N'_0, N''_0]$. Successive applications of $T$ define

$$\phi_{j+1}(x) = (T\phi_j)(x) = \sum_{k=N'}^{N''} p_k\phi_j(2x-k). \tag{10}$$

We have $\operatorname{supp}\phi_1 = \left[\frac{N'_0+N'}{2}, \frac{N''_0+N''}{2}\right]$ and denoting $\operatorname{supp}\phi_j = [N'_j, N''_j]$ it results $N'_{j+1} = (N'_j + N')/2$, $N''_{j+1} = (N''_j + N'')/2$. By computations, it follows

$$N'_j = \frac{N'_0}{2^j} + \left(\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^j}\right)N', \quad N''_j = \frac{N''_0}{2^j} + \left(\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^j}\right)N'',$$

and consequently $\lim_{j\to\infty} N'_j = N'$, $\lim_{j\to\infty} N''_j = N''$. This proves that $\operatorname{supp}\phi \subset [N', N'']$. In order to obtain the second inclusion, in (9) we notice that $p_{1-k}$ is only nonzero for $k \in [1-N'', 1-N'] \cap \mathbf{Z}$. On the other hand, we have $\operatorname{supp}\phi(2\cdot - k) \subset \left[\frac{N'+k}{2}, \frac{N''+k}{2}\right]$. relation (6) allows us to obtain the desired result.

Investigating (9) we remark that by a change of index in $p_k$, the relation (4) be written as follows

$$\phi(x) = \sum_{k=0}^{N} p_k\phi(2x-k), \quad p_0p_N \neq 0. \tag{11}$$

Of course, the scaling function $\phi$ must also be shifted accordingly.

Of the previous theorem, $\mathrm{supp}\phi \subset [0, N]$ and knowing that $\phi \in C(\mathbf{R})$ we deduce $\phi(0) = \phi(N) = 0$. Firstly, we need to determine $\phi(k)$, $1 \leq k \leq N-1$. Substituting $x = k$, $1 \leq k \leq N-1$, into (11) leads to $N-1$ linear equations with the $N-1$ unknowns $\phi(k)$. In matrix notation we have

$$v = Pv, \tag{12}$$

where $v$ is the column vector $(\phi(1), \phi(2), \ldots, \phi(N-1))^T$ and $P$ the $(N-1) \times (N-1)$ matrix

$$P := (p_{2j-k})_{1 \leq j,k \leq N-1} \tag{13}$$

with $j$ being the row index and $k$ the column index. Recalling that $\phi$ generates a partition of unity (see Proposition 3) we can determine the values of $\phi(k)$, $k \in \mathbf{Z}$, by finding the eigenvector $v$ corresponding to the eigenvalue 1 and imposing $\sum_{k=1}^{N-1} \phi(k) = 1$. Define $\phi_0$ to be the piecewise linear function which takes exactly the values $\phi(k)$ at the integers, that is

$$\phi_0(x) := \phi(x)(k+1-x) + \phi(k+1)(x-k), \quad x \in [k, k+1].$$

We compute $\phi_j$, by using (10) and it follows that $\phi_j$ are piecewise linear with nodes at $k/2^j \in [0, N]$, $k \in \mathbf{Z}$.

Let's make some examples. For $N = 3$, it is known the quadratic cardinal $B$-spline $N_3$ whose two-scale equation is

$$N_3(x) = \frac{1}{4}N_3(2x) + \frac{3}{4}N_3(2x-1) + \frac{3}{4}N_3(2x-2) + \frac{1}{4}N_3(2x-3).$$

However, there is another alternative [2], namely Daubechies'scaling function $\phi^D$ governed by

$$\phi^D(x) = \frac{1+\sqrt{3}}{4}\phi^D(2x) + \frac{3+\sqrt{3}}{4}\phi^D(2x-1) + \frac{3-\sqrt{3}}{4}\phi^D(2x-2) + \frac{1-\sqrt{3}}{4}\phi^D(2x-3).$$

In the following, we choose $p_0 = \mu$ and $p_1 = -\frac{1}{\mu}$ where $\mu = \left(1+\sqrt{5}\right)/2$. In concordance with Proposition 2, we must take $p_3 = \mu$ and $p_2 = 1 - \mu$. Thus the matrix $P$ defined in (13) becomes

$$P = \frac{1}{2}\begin{pmatrix} 1-\sqrt{5} & 1+\sqrt{5} \\ 1+\sqrt{5} & 1-\sqrt{5} \end{pmatrix}.$$

The solution of (12) is $v = a \begin{pmatrix} 1 & 1 \end{pmatrix}^T$ and the normalization condition implies $a = 1/2$. One obtains $\phi(1) = \phi(2) = 1/2$.

Using the golden ratio in the componence of the matrix $P$ we get a nice scale function which is defined on $\mathbf{Z}$ as follows

$$\phi(k) = \begin{cases} 1/2, & k \in \{1,2\} \\ 0, & k \in \mathbf{Z} \setminus \{1,2\}. \end{cases}$$

Having the values of $\phi(k)$ it is now easy to compute $\phi(k/2^j)$, $(k,j) \in \mathbf{Z} \times \mathbf{Z}$. In fact, the Interpolatory Graphical Display Algorithm can be applied, see [1], page 93.

## References


[1] C.K. Chui. *An Introduction to Wavelets.* Boston: Academic Press, 1992.
[2] I. Daubechies. *Ten Lecture on Wavelets.* Philadelphia: SIAM, 1992.
[3] C. Gasquet & P. Witomski. *Analyse de Fourier et Applications.* Paris: Masson, 1990.
[4] Y. Meyer. *Wavelets: Algorithms and Applications.* Philadelphia: SIAM, 1993.


"BABEŞ-BOLYAI" UNIVERSITY, FACULTY OF MATHEMATICS AND INFORMATICS,, STR. KOGĂLNICEANU, NR.1, 3400 CLUJ-NAPOCA, ROMANIA

# ON A CLASS OF FUNCTIONS OF TYPE $\alpha$ AND ORDER $\beta$ IN THE UNIT DISK

M.K. AOUF AND G.II. SĂLĂGEAN

**Abstract.** A representation formula, a distortion theorems and the radius of convexity are determined for the class $S_q(\alpha,\beta)$ of analytic functions whose power series are $f(z) = z + a_{q+1}z^{q+1} + \ldots$ and satisfy the condition

$$\left|\frac{\frac{zf'(z)}{f(z)} - 1}{\frac{zf'(z)}{f(z)} + 1 - 2\beta}\right| < \alpha,$$

for some $\alpha$ $(0 < \alpha \leq 1)$, $\beta$ $(0 \leq \beta < 1)$ and for all $z$ in $U = \{z : |z| < 1\}$. A sufficient condition for a function to belong to $S_q(\alpha,\beta)$ has also been obtained.

## 1. Introduction

Let $f(z) = z + \sum_{n=2}^{\infty} a_n z^n$ be analytic in the unit disc $U = \{z : |z| < 1\}$. The condition

$$\operatorname{Re}\left\{\frac{zf'(z)}{f(z)}\right\} > 0 \tag{1.1}$$

for $z \in U$ is a necessary and sufficient condition for $f(z)$ to be univalent and starlike in $U$. Robertson [7] was the first to introduce the notion of order for the class of starlike functions in $U$. A function $f(z)$ is said to be starlike of order $\alpha$ in $U$ if

$$\operatorname{Re}\left\{\frac{zf'(z)}{f(z)}\right\} > \alpha \tag{1.2}$$

for $|z| < 1$, for a given $\alpha$, $0 \leq \alpha < 1$.

Padmanabham [6] has introduced the "concept" of "order" of starlikeness in a different manner. Thus, according to him, a function $f(z) = z + \sum_{n=2}^{\infty} a_n z^n$ analytic in $U$

---

Received by the editors: November 10, 1996.

1991 *Mathematics Subject Classification.* 30C45.

*Key words and phrases.* analytic,starlike, distorsion theorems

and satisfying for all $z$ in $U$ the condition

$$\left|\frac{\frac{zf'(z)}{f(z)}-1}{\frac{zf'(z)}{f(z)}+1}\right|<\alpha, \tag{1.3}$$

for a given $\alpha$ $(0<\alpha\leq 1)$ is said to be starlike of "order" $\alpha$ in $U$. We denote the class of all such functions by $S(\alpha)$ and we say that $S(\alpha)$ is the class of starlike functions of type $\alpha$. For the class $S(\alpha)$ Padmanabhan [6] has obtained a representation formula, a distortion theorem and the radius of convexity. Also Mogra [3] obtained a coefficient estimates and a sufficient condition for a function to belong to $S(\alpha)$.

In [4] Mogra considered the functions $f(z)=z+\sum\limits_{k=1}^{\infty}a_{q+k}z^{q+k}$ for $q\in\{1,2,3,\ldots\}$, which are analytic in $U$ and which are starlike of type $\alpha$. The class of such functions are denoted by $S_q(\alpha)$. For the class $S_q(\alpha)$ Mogra [4] obtained a representation formula, distortion theorems and radius of convexity.

In this paper we consider the functions $f(z)=z+\sum\limits_{k=1}^{\infty}a_{q+k}z^{q+k}$ which are analytic in $U$. We say that $f(z)$ belongs to $S_q(\alpha,\beta)$, the class of starlike functions of order $\beta$ and type $\alpha$ if the condition

$$\left|\frac{\frac{zf'(z)}{f(z)}-1}{\frac{zf'(z)}{f(z)}+1-2\beta}\right|<\alpha, \tag{1.4}$$

is satisfied for some $\alpha,\beta$ $(0<\alpha,\leq 1,\ 0\leq\beta<1)$ and for all $z\in U$.

We note that by giving specific values for $\alpha,\beta$ and $q$, we obtain the following important subclasses:

(i) $S_1(\alpha,0)=S(\alpha)$,

(ii) $S_q(\alpha,0)=S_q(\alpha)$,

(iii) In [2] Juneja and Mogra introduced the class $S_1(\alpha,\beta)=S^*(\alpha,\beta)$ of starlike functions of order $\beta$ $(0\leq\beta<1)$ and type $\alpha$ $(0<\alpha\leq 1)$ and made a preliminary study of its properties.

One can easily show that: $f(z)\in S_q(\alpha,\beta)$ if and only if there is a function $f_1(z)\in S_q(\alpha)$ such that

$$f(z)=z\left[\frac{f_1(z)}{z}\right]^{(1-\beta)},\quad 0\leq\beta<1. \tag{1.5}$$

## 2. Representation formula for the class $S_q(\alpha,\beta)$

Let $B_\alpha$ $(0<\alpha\leq 1)$ denote the class of functions $\psi(z)$ which are analytic in $U$ and satisfy $|\psi(z)|\leq\alpha$ for all $z\in U$.

**Lemma 1.** *Let $H(z)=1+h_q z^q+\dots$. Then $H(z)$ is analytic and satisfies the condition*

$$\left|\frac{1-H(z)}{1-2\beta+H(z)}\right|<\alpha \quad (0<\alpha\leq 1 \text{ and } 0\leq\beta<1)$$

*for $|z|<1$ if and only if there exists a function $\psi(z)\in B_\alpha$ such that*

$$H(z)=\frac{1-(1-2\beta)z^q\psi(z)}{1+z^q\psi(z)}.$$

The proof of Lemma 1 follows exactly on the same lines as those of Padmanabhan [6, Lemma 1] and Mogra [4, Lemma 1]; so we omit the proof.

**Theorem 1.** *Let $f(z)=z+\sum_{k=1}^{\infty}a_{q+k}z^{q+k}$ be analytic in the unit disc $U$. Then $f(z)\in S_q(\alpha,\beta)$ if and only if*

$$f(z)=z\exp\left\{-2(1-\beta)\int_0^z\frac{t^{q-1}\psi(t)}{1+t^q\psi(t)}dt\right\} \tag{2.1}$$

*for some $\psi(z)\in B_\alpha$.*

*Proof.* Let $f(z)\in S_q(\alpha,\beta)$, it is easily seen that $\frac{zf'(z)}{f(z)}$ satisfies the hypothesis of Lemma 1. Hence there exists $\psi(z)\in B_\alpha$ such that

$$\frac{zf'(z)}{f(z)}=\frac{1-(1-2\beta)z^q\psi(z)}{1+z^q\psi(z)}.$$

Thus, we have

$$\frac{f'(z)}{f(z)}-\frac{1}{z}=\frac{-2(1-\beta)z^{q-1}\psi(z)}{1+z^q\psi(z)}. \tag{2.2}$$

Integration give (2.1) easily. Conversely, if $f(z)$ has the representation (2.1) for some $\psi(z)\in B_\alpha$, then, it follows that

$$\frac{zf'(z)}{f(z)}=\frac{1-(1-2\beta)z^q\psi(z)}{1+z^q\psi(z)}$$

Hence Lemma 1 gives that $f(z)\in S_q(\alpha,\beta)$ and the theorem is proved.

**Remark.** There is another proof of Theorem 1 which is an immediate consequence of (1.5) and the integral representation for $f_1(z)\in S_q(\alpha)$ given by Mogra [4, Theorem 1].

## 3. Distortion theorem for the class $S_q(\alpha,\beta)$

**Theorem 2.** *Let* $f(z)=z+\sum_{k=1}^{\infty}a_{q+k}z^{q+k}$ *be analytic in the unit disc* $U$ *and suppose* $f(z)\in S_q(\alpha,\beta)$. *Then we have, for* $z\in U$,

$$|f(z)|\leq\frac{|z|}{(1-\alpha|z|^q)^{\frac{2(1-\beta)}{q}}},\tag{3.1}$$

$$|f(z)|\geq\frac{|z|}{(1+\alpha|z|^q)^{\frac{2(1-\beta)}{q}}}.\tag{3.2}$$

*The estimates are sharp.*

*Proof.* Since $f(z)\in S_q(\alpha,\beta)$, we have, by (2.1)

$$\frac{zf'(z)}{f(z)}=\frac{1-(1-2\beta)z^q\psi(z)}{1+z^q\psi(z)}\tag{3.3}$$

for some $\psi(z)\in B_\alpha$. We can write (3.3) as

$$\frac{f'(z)}{f(z)}-\frac{1}{z}=\frac{-2(1-\beta)z^{q-1}\psi(z)}{1+z^q\psi(z)}.\tag{3.4}$$

Since $\psi(z)\in B_\alpha$, (3.4) gives

$$\log\left(\left|\frac{f(z)}{z}\right|\right)=\mathrm{Re}\,\left(\log\left(\frac{f(z)}{z}\right)\right)=$$

$$=\mathrm{Re}\,\int_0^z\left[\frac{f'(s)}{f(s)}-\frac{1}{s}\right]ds=\mathrm{Re}\,\int_0^z\frac{-2(1-\beta)s^{q-1}\psi(s)}{1+s^q\psi(s)}ds\leq$$

$$\leq\int_0^{|z|}\frac{2(1-\beta)|\psi(te^{i\theta})|t^{q-1}}{|1+t^qe^{iq\theta}\psi(te^{i\theta})|}dt\leq2(1-\beta)\int_0^{|z|}\frac{\alpha t^{q-1}}{1-\alpha t^q}dt=-\log(1-\alpha|z|^q)^{\frac{2(1-\beta)}{q}}.$$

Thus

$$\left|\frac{f(z)}{z}\right|\leq\frac{1}{(1-\alpha|z|^q)^{\frac{2(1-\beta)}{q}}}$$

which gives (3.1). To prove (3.2), we observe that the condition (1.4) coupled with an application of Schwarz's Lemma [5] implies that, for $|z|<1$, $\frac{zf'(z)}{f(z)}$ assumes values lying in the open disc $U$ on the line segment joining the points $\frac{1-\alpha(1-2\beta)|z|^q}{1+\alpha|z|^q}$ and $\frac{1+\alpha(1-2\beta)|z|^q}{1-\alpha|z|^q}$ as diameter. Hence we have

$$\frac{1-\alpha(1-2\beta)|z|^q}{1+\alpha|z|^q}\leq\mathrm{Re}\,\left\{\frac{zf'(z)}{f(z)}\right\}\leq\frac{1+\alpha(1-2\beta)|z|^q}{1+\alpha|z|^q}.\tag{3.5}$$

Let $|z|=r$, then (3.5) gives

$$r\mathrm{Re}\,\left\{\frac{\partial}{\partial r}\left(\log\frac{f(z)}{z}\right)\right\}=\mathrm{Re}\,\left\{\frac{zf'(z)}{f(z)}-1\right\}\geq\frac{-2\alpha(1-\beta)r^q}{1+\alpha r^q}.$$

Thus we have

$$\log\left|\frac{f(z)}{z}\right| = \operatorname{Re}\left\{\log\frac{f(z)}{z}\right\} \geq \int_0^r \frac{-2\alpha(1-\beta)s^{q-1}}{1+\alpha s^q}\,ds = -\log(1+\alpha r^q)^{\frac{2(1-\beta)}{q}}.$$

Hence

$$|f(z)| \geq \frac{|z|}{(1+\alpha|z|^q)^{\frac{2(1-beta)}{q}}}.$$

Equality in (3.1) and (3.2) holds for the function

$$f(z) = \frac{z}{(1-\alpha z^q)^{\frac{2(1-\beta)}{q}}}.$$

## 4. A sufficient condition for a function to belong to $S_q(\alpha,\beta)$

**Theorem 3.** *Let* $f(z) = z + \sum_{k=1}^{\infty} a_{q+k} z^{q+k}$ *be analytic in the unit disc* $U$. *If for some* $\alpha$ *(*$0 < \alpha \leq 1$*) and* $\beta$ *(*$0 \leq \beta < 1$*),*

$$\sum_{k=1}^{\infty}\{(1+\alpha)(q+k)+\alpha(1-2\beta)-1\}|a_{q+k}| \leq 2(1-\beta)\alpha, \tag{4.1}$$

*then* $f(z) \in S_q(\alpha,\beta)$.

*Proof.* We employ the technique used by Clunie and Keogh [1]. Thus suppose that (4.1) holds and that $f(z) = z + \sum_{k=1}^{\infty} a_{q+k} z^{q+k}$, then in $|z| < 1$,

$$|zf'(z) - f(z)| - \alpha|zf'(z) + (1-2\beta)f(z)| =$$

$$= \left|\sum_{k=1}^{\alpha}(q+k-1)a_{q+k}z^{q+k}\right| - \alpha\left|2(1-\beta)z + \sum_{k=1}^{\infty}(q+k+1-2\beta)a_{q+k}z^{q+k}\right| \leq$$

$$\leq \sum_{k=1}^{\infty}(q+k-1)|a_{q+k}|r^{q+k} - \alpha\left\{2(1-\beta)r - \sum_{k=1}^{\infty}(q+k+1-2\beta)|a_{q+k}|r^{q+k}\right\} <$$

$$< \left[\sum_{k=1}^{\infty}(q+k-1)|a_{q+k}| - 2(1-\beta)\alpha + \sum_{k=1}^{\infty}\alpha(q+k+1-2\beta)|a_{q+k}|\right] r =$$

$$= \left[\sum_{k=1}^{\infty}\{(1+\alpha)(q+k)+\alpha(1-2\beta)-1\}|a_{q+k}| - 2(1-\beta)\alpha\right] r \leq 0.$$

Hence it follows that $\left|\dfrac{\frac{zf'(z)}{f(z)}-1}{\frac{zf'(z)}{f(z)}+1-2\beta}\right| < \alpha$, therefore $f(z) \in S_q(\alpha,\beta)$. This completes the proof. We note that

$$f(z) = z - \frac{2(1-\beta)\alpha}{\{(1+\alpha)(q+k)+\alpha(1-2\beta)-1\}}z^{q+k}$$

is an extremal function of the theorem since $\left|\frac{\frac{zf'(z)}{f(z)}-1}{\frac{zf'(z)}{f(z)}+1-2\beta}\right| = \alpha$, for $z = 1$, $0 < \alpha \leq 1$, $0 \leq \beta < 1$ and $k = 1, 2, \ldots$. We also observe that the converse of the above theorem may not be true. For example, consider the functions $f(z)$,

$$\frac{zf'(z)}{f(z)} = \frac{1 + (1 - 2\beta)\alpha z^q}{1 - \alpha z^q}.$$

It is easily seen that $f(z) \in S_q(\alpha, \beta)$ but

$$\sum_{k=1}^{\infty} \frac{\{(1 + \alpha)(q + k) + \alpha(1 - 2\beta)\}}{2\alpha(1 - \beta)} |a_{q+k}| =$$

$$= \sum_{k=1}^{\infty} \frac{\{(1 + \alpha)(q + k) + \alpha(1 - 2\beta) - 1\}}{2\alpha(1 - \beta)} \cdot \frac{2\alpha^{\frac{(q+k-1)}{q}}(1 - \beta)}{(q + k - 1)} =$$

$$= \sum_{k=1}^{\infty} \left\{1 + \frac{\alpha(q + k + 1 - 2\beta)}{q + k - 1}\right\} \alpha^{\left(\frac{k-1}{q}\right)} > 1$$

for some $\alpha, \beta$ $(0 < \alpha \leq 1, 0 \leq \beta < 1)$, $z \in U$.

## 5. The radius of convexity for functions in the class $S_q(\alpha, \beta)$

Let $D$ denote the class of analytic functions $w(z)$ in $|z| < 1$ which satisfy the conditions (i) $|w(z)| < 1$ for $|z| < 1$. For obtaining the radius of convexity for functions in the class $f(z) \in S_q(\alpha, \beta)$, we require the following lemmas.

**Lemma 2** [8]. *If $w(z) \in D$, then for $|z| < 1$,*

$$|zw'(z) - w(z)| \leq \frac{|z|^2 - |w(z)|^2}{1 - |z|^2}. \tag{5.1}$$

**Lemma 3.** *For $w(z) \in D$, we have*

$$\operatorname{Re}\left\{\frac{z^q w'(z) + (q - 1)z^{q-1}w(z)}{[1 + \alpha z^{q-1}w(z)][1 - \alpha(1 - 2\beta)z^{q-1}w(z)]}\right\}$$

$$\leq \frac{-q}{4\alpha^2(1 - \beta)^2}\operatorname{Re}\left[\alpha p(z) - \frac{\alpha(1 - 2\beta)}{p(z)} - 2\alpha\beta\right] + \frac{r^{2q}|\alpha p(z) + \alpha(1 - 2\beta)|^2 - |1 - p(z)|^2}{4\alpha^2(1 - \beta)^2 r^{q-1}(1 - r^2)|p(z)|} \tag{5.2}$$

*where* $p(z) = \frac{1 - \alpha(1 - 2\beta)z^{q-1}w(z)}{1 + \alpha z^{q-1}w(z)}$, $r = |z|$, $0 < \alpha \leq 1$ *and* $0 \leq \beta < 1$.

*Proof.* An application of (5.1) gives (5.2) easily.

**Remark.** The transformation $p(z) = \frac{1-\alpha(1-2\beta)z^{q-1}w(z)}{1+\alpha z^{q-1}w(z)}$ maps the circle $|w(z)| \leq r$ onto the circle

$$\left|p(z) - \frac{1+\alpha^2(1-2\beta)r^{2q}}{1-\alpha^2 r^{2q}}\right| \leq \frac{2\alpha(1-\beta)r^q}{1-\alpha^2 r^{2q}}. \tag{5.3}$$

**Theorem 4.** *Let $f(z) \in S_q(\alpha,\beta)$, $0 < \alpha \leq 1$, $0 \leq \beta < \frac{1}{2}$, then for $|z| = r$, $0 \leq r < 1$,*

$$\mathrm{Re}\left\{1 + \frac{zf''(z)}{f'(z)}\right\} \geq \begin{cases} \frac{1-2\alpha[(q+1)-\beta(q+2)]+\alpha^2(1-2\beta)^2 r^{2q}}{[1+\alpha r^q][1-\alpha(1-2\beta)r^q]} \ \textit{for } R_0 \leq R_1, \\ \frac{1}{\alpha(1-\beta)r^{q-1}(1-r^2)}\Big[\sqrt{\alpha[2(1-\beta)+q]r^{q-1}(1-r^2)+1-\alpha^{2q}}\cdot \\ \cdot\sqrt{1-\alpha q(1-2\beta)r^{q-1}+\alpha q(1-2\beta)r^{q+1}-\alpha^2(1-2\beta)^2 r^{2q}} \\ -(1+\alpha^2(1-2\beta)r^{2q}\Big] - \frac{\beta q}{1-\beta} \ \textit{for } R_0 \geq R_1 \end{cases} \tag{5.4}$$

*where*

$$a = \frac{1+\alpha^2(1-2\beta)r^{2q}}{1-\alpha^2 r^{2q}}, \quad d = \frac{2\alpha(1-\beta)r^q}{1-\alpha^2 r^{2q}},$$

$$R_0 = \left\{\frac{1-\alpha q(1-2\beta)r^{q-1}+\alpha q(1-2\beta)r^{q+1}-\alpha^2(1-2\beta)^2 r^{2q}}{\alpha[2(1-\beta)+q]r^{q-1}(1-r^2)+1-\alpha^2 r^{2q}}\right\}^{\frac{1}{2}},$$

$$R_1 = a - d = \frac{1-\alpha(1-2\beta)r^q}{1+\alpha r^q}.$$

*Proof.* Since $f(z) \in S_q(\alpha,\beta)$, by (2.1) we have

$$\frac{zf'(z)}{f(z)} = \frac{1-(1-2\beta)z^q\psi(z)}{1+z^q\psi(z)}$$

where $\psi(z) \in B_\alpha$. Thus we can write $z\psi(z) = \alpha w(z)$, where $w(z) \in D$.

Consequently

$$\frac{zf'(z)}{f(z)} = \frac{1-\alpha(1-2\beta)z^{q-1}w(z)}{1+\alpha z^{q-1}q(z)}. \tag{5.5}$$

Differentiating (5.5) logarithmically we have

$$1 + \frac{zf''(z)}{f'(z)} = \frac{1-\alpha(1-2\beta)z^{q-1}w(z)}{1+\alpha z^{q-1}w(z)} -$$

$$-2\alpha(1-\beta)\left\{\frac{z^q w'(z)+(q-1)z^{q-1}w(z)}{[1+\alpha z^{q-1}w(z)][1-\alpha(1-2\beta)z^{q-1}w(z)]}\right\}. \tag{5.6}$$

An application of Lemma 3 to the above equation gives

$$\mathrm{Re}\left\{1 + \frac{zf''(z)}{f'(z)}\right\} \geq \frac{1}{2\alpha(1-\beta)}\left[\mathrm{Re}\left\{\alpha[2(1-\beta)+q]p(z) - \frac{\alpha(1-2\beta)q}{p(z)} - 2\alpha\beta q\right\} -\right.$$

$$-\frac{r^{2q}|\alpha p(z)+\alpha(1-2\beta)|^2-|1-p(z)|^2}{r^{q-1}(1-r^2)|p(z)|}\Bigg].$$

By setting $p(z)=a+\xi+i\eta$, $R^2=(a+\xi)^2+\eta^2$, where $a=\frac{1+\alpha^2(1-2\beta)r^{2q}}{1-\alpha^2r^{2q}}$ and denoting the expression of the right-hand side of (5.6) by $E(\xi,\eta)$, we get

$$E(\xi,\eta)=\frac{1}{2\alpha(1-\beta)}[\alpha[2(1-\beta)+q](a+\xi)-\alpha(1-2\beta)q(a+\xi)R^{-2}-2\alpha\beta q-$$

$$-\frac{1-\alpha^2r^{2q}}{r^{q-1}(1-r^2)}(d^2-\xi^2-\eta^2)R^{-1}] \tag{5.7}$$

where

$$d=\frac{2\alpha(1-\beta)r^q}{1-\alpha^2r^{2q}}.$$

Differentiating (5.7) partially w.r.t. $\eta$, we get

$$\frac{\partial E}{\partial\eta}=\frac{1}{2\alpha(1-\beta)}\eta R^{-4}F(\xi,\eta), \tag{5.8}$$

where

$$F(\xi,\eta)=2\alpha(1-2\beta)q(a+\xi)+\frac{(d^2-\xi^2-\eta^2)(1-\alpha^2r^{2q})}{r^{q-1}(1-r^2)}R+2\frac{1-\alpha^2r^{2q}}{r^{q-1}(1-r^2)}R^3.$$

It is easily seen that $F(\xi,\eta)>0$ for all $\alpha$, $0<\alpha\le 1$, $0\le\beta<\frac{1}{2}$, $q\ge 1$ and so (5.8) gives that the minimum of $E(\xi,\eta)$ inside the circle $\xi^2+\eta^2<d^2$ is attained on the diameter $\eta=0$. Hence putting $\eta=0$ in (5.7), we get

$$M(R)=E(\xi,o)=\frac{1}{2\alpha(1-\beta)}\Bigg[\alpha[2(1-\beta)+q](a+\xi)-\alpha(1-2\beta)q(a+\xi)R^{-2}-$$

$$-2\alpha\beta q-\frac{1-\alpha^2r^{2q}}{r^{q-1}(1-r^2)}(d^2-\xi^2)R^{-1}\Bigg]=$$

$$=\frac{1}{2\alpha(1-\beta)}\Bigg[\alpha[2(1-\beta)+q](a+\xi)-\alpha(1-2\beta)qR^{-1}-2\alpha\beta q$$

$$-\frac{1-\alpha^2r^{2q}}{r^{q-1}(1-r^2)}[d^2-(R-a)^2]R^{-1}\Bigg]=$$

$$=\frac{1}{2\alpha(1-\beta)}\Bigg[\left\{\alpha[2(1-\beta)+q]+\frac{1-\alpha^2r^{2q}}{r^{q-1}(1-r^2)}\right\}R+$$

$$+\left\{\frac{1-\alpha q(1-2\beta)r^{q-1}+\alpha q(1-2\beta)r^{q+1}-\alpha^2(1-2\beta)^2r^{2q}}{r^{q-1}(1-r^2)}\right\}R^{-1}-$$

$$-2a\frac{1-\alpha^2r^{2q}}{r^{q-1}(1-r^2)}\Bigg]-\frac{\beta q}{1-\beta},$$

where $R = a + \xi$ and $a - d \le R \le a + d$. Thus the absolute minim of $M(R)$ in $(0, \infty)$ is attained at

$$R_0 = \left\{ \frac{1 - \alpha q(1-2\beta)r^{q-1} + \alpha q(1-2\beta)r^{q+1} - \alpha^2(1-2\beta)^2 r^{2q}}{\alpha[2(1-\beta)+q]r^{q-1}(1-r^2) + 1 - \alpha^2 r^{2q}} \right\}^{\frac{1}{2}} \tag{5.9}$$

and equals

$$M(R_0) = \frac{1}{\alpha(1-\beta)r^{q-1}(1-r^2)} \Big[ \sqrt{\alpha[2(1-\beta)+q]r^{q-1}(1-r^2) + 1 - \alpha^2 r^{2q}} \cdot$$

$$\cdot \sqrt{1 - \alpha q(1-2\beta)r^{q-1} + \alpha q(1-2\beta)r^{q+1} - \alpha^2(1-2\beta)^2 r^{2q}} - (1 + \alpha^2(1-2\beta)r^{2q}) \Big] - \frac{\beta q}{1-\beta}. \tag{5.10}$$

It is easily seen that $R_0 < a + d$, but $R_0$ is not always greater than $a - d$. In such a case when $R_0 \notin [a-d, a+d]$ the minimum of $M(R)$ on the segment $[a-d, a+d]$ is attained at $R_1 = a - d$ and equals

$$M(R_1) = M(a-d) = \frac{1 - 2\alpha[(q+1) - \beta(q+2)]r^q + \alpha^2(1-2\beta)^2 r^{2q}}{[1 + \alpha r^q][1 - \alpha(1-2\beta)r^q]}. \tag{5.11}$$

The two minima given by (5.10) and (5.11) coincides when $R_0 = R_1$. The inequality (5.4) follows from (5.10) and (5.11).

The inequality signs in (5.4) are attained for the following functions:

$$f(z) = \frac{z}{(1+\alpha z^q)^{\frac{2(1-\beta)}{q}}} \text{ for } R_0 \le R_1$$

and

$$\frac{zf'(z)}{f(z)} = \frac{1 - [1 - \alpha(1-2\beta)z^{q-1}]bz - \alpha(1-2\beta)z^{q+1}}{1 - (1 + \alpha z^{q-1})bz + \alpha z^{q+1}} \text{ for } R_0 \ge R_1$$

where $b$ is determined from

$$\frac{1 - [1 - \alpha(1-2\beta)r^{q-1}]br - \alpha(1-2\beta)r^{q+1}}{1 - (1 + \alpha r^{q-1})br + \alpha r^{q+1}} = R_0.$$

## References


[1] J. Clunie and F.R. Keogh, *On starlike and convex schlicht functions*, J. London Math. Soc. 35(1970), 229-233.

[2] O.P. Juneja and M.L. Mogra, *On starlike functions of order $\alpha$ and type $\beta$*, Notices Amer. Math. Soc. 22(1975), A-384, Abstract No.75T:B80.

[3] M.L. Mogra, *On a class of starlike functions in the unit disc. I*, J. Indian Math. Soc. 40(1976), 159-161.

[4] M.L. Mogra, *On a class of starlike functions in the unit disc. II*, Indian J. Pure Appl. Math. 8(1977), no.2, 157-165.

[5] Z. Nehari, *Conformal Mapping*, Mc Graw Hill Book Co., Inc., New York, 1952.

[6] K.S. Padmanabham, *On certain classes of starlike functions in the unit disc*, J. Indian Math. Soc. (N.S.) 32(1968), 89-103.

[7] M.S. Robertson, *On the theory of univalent functions*, Ann. Math. 37(1936), 374-408.

[8] V. Singh and M.R. Goel, *On radii of convexity and starlikeness of some classes of functions*, J. Math. Soc. Japan 29(1971), 323–339.

DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE, UNIVERSITY OF MANSOURA, MANSOURA, EGYPT

"BABEŞ–BOLYAI" UNIVERSITY, FACULTY OF MATHEMATICS AND INFORMATICS, STR. KOGĂLNICEANU, NO. 1, 3400 CLUJ-NAPOCA, ROMÂNIA

STUDIA UNIV. "BABEŞ-BOLYAI", MATHEMATICA, Volume XLII, Number 4, December 1997

# APPROXIMATION OF THE SOLUTION OF A STOCHASTIC EVOLUTION EQUATION

HANNELORE BRECKNER

**Abstract.** The aim of this paper is the study of a certain stochastic evolution equation, for which one proves the existence and the (almost surely) uniqueness of the solution. One takes into consideration the sequence of the solutions of some simpler equations and one proves that the respectiv sequence is convergent almost surely to the solution of the equation in study. The hypothesis for the operators involved in equations are more general that those used so far in the literature.

## 1. Introduction

Evolution equations perturbed by white noise have important practical applications in physics, biology and engineering. Concrete problems leading to such equations occur, for example, in statistical hydromechanics [8], in genetics and neurophysiology [1], [4], in the study of random vibrations and heat conductivity with random disturbance [6].

The purpose of this paper is to investigate an abstract stochastic evolution equation (in the sense of Ito), involving a strong monotone, hemicontinuous operator and locally Lipschitz continuous nonlinearities, which satisfy a growth condition. The assumptions imposed here are more general than those in [2] and give us the possibility of approximating the solutions of a larger class of stochastic differential equations.

It is proved that the considered equation possesses a solution which is unique with probability 1. The existence of this solution is shown by constructing a sequence of stochastic processes that approximate the solution almost surely.

Received by the editors: April 1, 1997.
1991 *Mathematics Subject Classification.* 60H10, 60H15.
*Key words and phrases.* stochastic evolution equations, existence, uniqueness.

## 2. Formulation of the problem

Let $[0,T] \subset \mathbb{R}$, and let $(V, H, V^*)$ be an evolution triple with $(V, (\cdot,\cdot)_V)$ and $(H, (\cdot,\cdot))$ separable Hilbert spaces. All linear spaces that occur in our paper are considered over the field of real numbers.

In what follows we investigate the following problem which we call problem (P): seek a solution of the evolution equation

$$dx(\omega,t) = -Ax(\omega,t)dt + F(t, x(\omega,t))dt + G(t, x(\omega,t))dw(t) \tag{1}$$

(for all $t \in [0,T]$ and a.e. $\omega \in \Omega$) with the initial condition

$$x(\omega, 0) = x_0(\omega) \quad \text{a.e. } \omega \in \Omega, \tag{2}$$

where:

**(i)** $(\Omega, \mathcal{F}, P)$ is a probability space, $\{\mathcal{F}_t \mid t \in [0,T]\}$ is a filtration (contained in $\mathcal{F}$) with respect to a given real Wiener process $\left(w(t)\right)_{t\in[0,T]}$;

**(ii)** $A : V \to V^*$ is a strong monotone, hemicontinous operator satisfying the following growth condition: there exists a constant $c_A > 0$ such that

$$\|A(v)\|_{V^*}^2 < c_A(1 + \|v\|_V^2)$$

for all $v \in V$;

**(iii)** $F, G : [0,T] \times H \to H$ are functions satisfying the following conditions:

**(a)** there exists a constant $\alpha > 0$ and a sequence $(\beta_N)$ of positive constants such that, for all natural numbers $N$, all $t_1, t_2 \in [0,T]$ and all $v_1, v_2 \in H$ with $\|v_1\| \leq N$, $\|v_2\| \leq N$, the inequalities

$$\|F(t_1, v_1) - F(t_1, v_2)\|^2 \leq \alpha|t_1 - t_2|^2 + \beta_N\|v_1 - v_2\|^2,$$

$$\|G(t_1, v_1) - G(t_1, v_2)\|^2 \leq \alpha|t_1 - t_2|^2 + \beta_N\|v_1 - v_2\|^2$$

hold;

**(b)** there exists a constant $\gamma > 0$ such that for all $t \in [0,T]$ and all $v \in H$ the inequalities

$$\|F(t,v)\|^2 \leq \gamma(1 + \|v\|^2), \quad \|G(t,v)\|^2 \leq \gamma(1 + \|v\|^2)$$

hold;

**(iv)** $x_0 \in \mathcal{L}_H^2(\Omega)$ is $\mathcal{F}_0$-measurable.

If $(S, \|\cdot\|_S)$ is a Banach space, then we denote by $\mathcal{L}_S^2(\Omega \times [0,T])$ the linear space of all functions $v : \Omega \times [0,T] \to S$ which are $\mathcal{F} \times \mathcal{B}^1_{[0,T]}$-measurable and for which

$$E\int_0^T \|v(\omega,t)\|_S^2 < \infty.$$

Further we denote by $\mathcal{L}_S^2(\Omega)$ the linear space of all functions $v : \Omega \to S$ which are $\mathcal{F}$-measurable and for which $E\|v(\omega,t)\|_S^2 < \infty$. As usual, the values of a function $v \in \mathcal{L}_S^2(\Omega \times [0,T])$ or $v \in \mathcal{L}_S^2(\Omega)$ will not be denoted by $v(\omega,t)$ and $v(\omega)$, but simply by $v(t)$ and $v$, respectively.

An adapted process $\left(x(t)\right)_{t\in[0,T]}$ from the space $\mathcal{L}_V^2(\Omega \times [0,T])$ is said to be a *solution of problem* (P) if it satisfies equation (1) and condition (2) in the following sense:

$$(x(t) - x_0, v) = -\int_0^t \langle Ax(s), v\rangle ds + \int_0^t (F(s,x(s)), v) ds + \int_0^t (G(s,x(s)), v) dw(s)$$

for a.e. $\omega \in \Omega$ and for all $v \in V, t \in [0,T]$. Here $\langle\cdot,\cdot\rangle$ denotes the duality between $V$ and $V^*$.

## 3. Main Result

For each natural number $N$ we define the functions $F_N, G_N : [0,T] \times H \to H$ by

$$F_N(t,x) = \begin{cases} F(t,x) & , \quad \|x\| \le N \\ F(t, \frac{Nx}{\|x\|}) & , \quad \|x\| > N, \end{cases}$$

$$G_N(t,x) = \begin{cases} G(t,x) & , \quad \|x\| \le N \\ G(t, \frac{Nx}{\|x\|}) & , \quad \|x\| > N. \end{cases}$$

By using hypothesis (iii) and the properties of a norm in a Hilbert space, it can be proved that $F_N$ and $G_N$ are Lipschitz continuous, i.e., they satisfy

$$\|F_N(t_1,v_1) - F_N(t_1,v_2)\|^2 \le \alpha|t_1 - t_2|^2 + \beta_N\|v_1 - v_2\|^2,$$

$$\|G_N(t_1,v_1) - G_N(t_1,v_2)\|^2 \le \alpha|t_1 - t_2|^2 + \beta_N\|v_1 - v_2\|^2$$

for all $v_1, v_2 \in H$ and all $t_1, t_2 \in [0,T]$.

The basic idea which will be used in our investigations is to approximate problem (P) by a sequence of problems $(\mathrm{P}_N)$ $(N \in \{1,2,\dots\})$. The problem $(\mathrm{P}_N)$ requires to find an adapted process $\left(x_N(t)\right)_{t\in[0,T]}$ from the space $\mathcal{L}^2_V(\Omega\times[0,T])$ that satisfies the evolution equation

$$\begin{aligned}(x_N(t)-x_0,v) = & -\int_0^t\langle Ax_N(s),v\rangle ds+\int_0^t(F_N(s,x_N(s)),v)ds+\\ & +\int_0^t(G_N(s,x_N(s)),v)dw(s)\end{aligned}$$

for all $v\in V$, $t\in[0,T]$ and a.e. $\omega\in\Omega$.

Concerning problem $(\mathrm{P}_N)$ the following theorem was established in [2], p. 133.

**Theorem 3.1.** *For each natural number $N$ there exists an almost surely unique process $\left(x_N(t)\right)_{t\in[0,T]}$, which is a solution of problem $(\mathrm{P}_N)$ and which has continuous trajectories in $H$.*

Let $a$ be a positive number, fixed for the moment. To this number we assign the following problem $(\mathrm{P}^a_N)$: find an adapted process $\left(x_N(t)\right)_{t\in[0,T]}$ from the space $\mathcal{L}^2_V(\Omega\times[0,T])$ that satisfies

$$\begin{aligned}(e^{-at}x_N(t) \quad - \quad x_0,v) = & -\int_0^t\langle e^{-as}(A+aJ)(x_N(s)),v\rangle ds+ \\ & +\int_0^t(e^{-as}F_N(s,x_N(s)),v)ds+\int_0^t(e^{-as}G_N(s,x_N(s)),v)dw(s),\end{aligned}\tag{3}$$

for a.e. $\omega\in\Omega$ and for all $v\in V$, $t\in[0,T]$ ($J:V\to V^*$ is the duality map between $V$ and $V^*$).

Using the Ito formula, it can be shown that $\left(x_N(t)\right)_{t\in[0,T]}$ is a solution of problem $(\mathrm{P}_N)$ if and only if it is a solution of problem $(\mathrm{P}^a_N)$. The advantage of problem $(\mathrm{P}^a_N)$ is the possibility of a favorable choice of $a$ and so, as it will be revealed in the sequel, some useful properties can be obtained.

**Lemma 3.2.** *There exists a constant $c_1>0$ such that for all natural numbers $N$ the following inequality holds:*

$$E\sup_{0\le t\le T}\|x_N(t)\|^2+E\int_0^T\|x_N(t)\|^2_V dt\le c_1,$$

*where $\left(x_N(t)\right)_{t\in[0,T]}$ is the solution of problem $(\mathrm{P}_N)$.*

*Proof.* Since the process $\left(x_N(t)\right)_{t\in[0,T]}$ is also a solution of ($P_N^a$), we can apply the Ito formula to (3) and obtain

$$e^{-2at}\|x_N(t)\|^2 = \|x_0\|^2 - 2\int_0^t e^{-2as}\langle (A+aJ)(x_N(s)), x_N(s)\rangle ds + \tag{4}$$

$$+ \quad 2\int_0^t e^{-2as}(F_N(s,x_N(s)),x_N(s))ds + \int_0^t e^{-2as}\|G_N(s,x_N(s))\|^2 ds +$$

$$+ \quad 2\int_0^t e^{-2as}(G_N(s,x_N(s)),x_N(s))dw(s).$$

By the monotonicity of $A$ and elementary computations we obtain

$$2\int_0^t e^{-2as}\langle (A+aJ)(x_N(s)),x_N(s)\rangle ds \geq (a-1)\int_0^t e^{-2as}\|x_N(t)\|_V^2 ds - T\|A(0)\|_{V^*}^2.$$

Consequently (4) implies

$$\sup_{0\leq s\leq t} e^{-2as}\|x_N(s)\|^2 + (a-1)\int_0^t e^{-2as}\|x_N(s)\|_V^2 ds \leq k_1 + \tag{5}$$

$$+ \quad \int_0^t e^{-2as}\left(\|F_N(s,x_N(s))\|^2 + \|G_N(s,x_N(s))\|^2\right)ds + \int_0^t e^{-2as}\|x_N(s)\|^2 ds +$$

$$+ \quad 2\sup_{0\leq s\leq t}\int_0^s e^{-2ar}(G_N(r,x_N(r)),x_N(r))dw(r),$$

where $k_1$ is a positive constant, which does not depend on $N$. By applying hypothesis (iii) and the mathematical expectation to (5) we get

$$E\sup_{0\leq s\leq t} e^{-2as}\|x_N(s)\|^2 + (a-1)E\int_0^t e^{-2as}\|x_N(s)\|_V^2 ds \leq k_1 + \tag{6}$$

$$+ \quad (1+2\gamma)E\int_0^t e^{-2as}\|x_N(s)\|^2 ds + 2E\sup_{0\leq s\leq t}\int_0^s e^{-2ar}(G_N(r,x_N(r)),x_N(r))dw(r).$$

Now Burkholder's inequality (see [3]) yields

$$2E \sup_{0\le s\le t} \int_0^s e^{-2ar}(G_N(r,x_N(r)),x_N(r))dw(r) \le$$

$$\le \quad 6E\Big(\int_0^t e^{-4as}(G_N(s,x_N(s)),x_N(s))^2 ds\Big)^{\frac{1}{2}} \le$$

$$\le \quad \frac{1}{2}E \sup_{0\le s\le t} e^{-2as}\|x_N(s)\|^2 + 18\gamma E\int_0^t e^{-2as}(1+\|x_N(s)\|^2)ds.$$

Using these inequalities and choosing $a > 2$, it follows from (6) that

$$E \sup_{0\le s\le t} e^{-2as}\|x_N(s)\|^2 + E\int_0^t e^{-2as}\|x_N(s)\|_V^2 ds \le k_2 + k_3 E\int_0^t \sup_{0\le r\le s} e^{-2ar}\|x_N(r)\|^2 dr,$$

where $k_2$ and $k_3$ are positive constants which do not depend on $N$. By applying Gronwall's Lemma (see [7], Lemma 3, p. 311) we get

$$E \sup_{0\le s\le t} e^{-2as}\|x_N(s)\|^2 + E\int_0^t e^{-2as}\|x_N(s)\|_V^2 ds \le k_2 e^{k_3 T} \tag{7}$$

for all $t \in [0,T]$. In particular, if we take $t = T$, then we obtain

$$E\Big(\sup_{0\le s\le T}\|x_N(s)\|^2 + \int_0^T \|x_N(s)\|_V^2 ds\Big) \le k_2 e^{(k_3+2a)T}. \tag{8}$$

Next we set $c_1 = k_2 e^{(k_3+2a)T}$ and hence the conclusion of the lemma holds. □

For each natural number $N$ we define the function $\mathcal{T}_N : \Omega \to \mathbb{R}$, called *stopping time*, by

$$\mathcal{T}_N(\omega) = \sup\Big\{t\in[0,T] \;\mid\; \sup_{0\le s\le t}\|x_N(s)\|^2 + \int_0^t \|x_N(s)\|_V^2 ds \le N^2\Big\} \quad \text{for a.e. } \omega\in\Omega.$$

**Lemma 3.3.** *Let $M$ and $N$ be two natural numbers such that $M > N$. For all $\epsilon > 0$ the following relation holds:*

$$P\Big\{\sup_{0\le t\le \mathcal{T}_N}\|x_N(t)-x_M(t)\|^2 + \int_0^{\mathcal{T}_N}\|x_N(t)-x_M(t)\|_V^2 dt > \epsilon\Big\} = 0. \tag{9}$$

*Proof.* For all $t \in [0, \mathcal{T}_N]$ we have

$$\sup_{0 \le s \le t} \|x_N(s)\|^2 + \int_0^t \|x_N(s)\|_V^2 ds < N^2 \tag{10}$$

and for all $s \in [0, t]$

$$F_N(s, x_N(s)) \quad F_M(s, x_N(s)), \quad G_N(s, x_N(s)) = G_M(s, x_N(s)). \tag{11}$$

Since $\left(x_N(t)\right)_{t \in [0,T]}$ and $\left(x_M(t)\right)_{t \in [0,T]}$ are the solutions of $(\mathrm{P}_\mathrm{N}^\mathrm{s})$ and $(\mathrm{P}_\mathrm{M}^\mathrm{s})$, respectively, we can apply the Ito formula (with $\|\cdot\|^2$) to the difference of this two processes and obtain

$$e^{-2at}\|x_N(t) - x_M(t)\|^2 = \int_0^t e^{-2as}\|G_N(s, x_N(s)) - G_M(s, x_M(s))\|^2 ds - \tag{12}$$

$$- \quad 2\int_0^t e^{-2as}\langle (A + aJ)(x_N(s)) - (A + aJ)(x_M(s)), x_N(s) - x_M(s)\rangle ds +$$

$$+ \quad 2\int_0^t e^{-2as}(F_N(s, x_N(s)) - F_M(s, x_M(s)), x_N(s) - x_M(s)) ds +$$

$$+ \quad 2\int_0^t e^{-2as}(G_N(s, x_N(s)) - G_M(s, x_M(s)), x_N(s) - x_M(s)) dw(s),$$

for all $t \in [0, T]$ and a.e. $\omega \in \Omega$.

Let $t \in [0, \mathcal{T}_N]$ be arbitrarily choosen. In virtue of (10), (11), hypothesis (iii) and the monotonicity of $A$ we obtain from (12) that

$$e^{-2at}\|x_N(t) - x_M(t)\|^2 + 2a\int_0^t \|x_N(s) - x_M(s)\|_V^2 ds \le \tag{13}$$

$$\le \quad 2\int_0^t e^{-2as}(G_N(s, x_N(s)) - G_M(s, x_M(s)), x_N(s) - x_M(s)) dw(s) +$$

$$+ \quad (2\beta_M + 1)\int_0^t e^{-2as}\|x_N(s) - x_M(s)\|^2 ds$$

for all $t \in [0, \mathcal{T}_N]$. Therefore

$$E \sup_{0 \le s \le t \wedge \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|^2 + 2aE \int_0^{t \wedge \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|_V^2 ds \le \quad (14)$$

$$\le \quad 4E \sup_{0 \le s \le t \wedge \mathcal{T}_N} \int_0^s e^{-2as} (G_N(r, x_N(r)) - G_M(r, x_M(r)), x_N(r) - x_M(r)) dw(r) +$$

$$+ \quad 2(2\beta_M + 1)E \int_0^{t \wedge \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|^2 ds.$$

Now Burkholder's inequality yields

$$2E \sup_{0 \le s \le t \wedge \mathcal{T}_N} \int_0^s e^{-2ar} (G_N(r, x_N(r)) - G_M(r, x_M(r)), x_N(r) - x_M(r)) dw(r) \le$$

$$\le \quad 6E\Big( \int_0^{t \wedge \mathcal{T}_N} e^{-4as} (G_N(s, x_N(s)) - G_M(r, x_M(r)), x_N(r) - x_M(r))^2 ds\Big)^{\frac{1}{2}} \le$$

$$\le \quad \frac{1}{2} E \sup_{0 \le s \le t \wedge \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|^2 + 18\beta_M E \int_0^{t \wedge \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|^2 ds.$$

Consequently (14) implies

$$E \sup_{0 \le s \le t \wedge \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|^2 + 4aE \int_0^{t \wedge \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|_V^2 dt \le$$

$$\le \quad 2(k_3\beta_M + 1)E \int_0^t \sup_{0 \le r \le s \wedge \mathcal{T}_N} e^{-2ar} \|x_N(r) - x_M(r)\|^2 ds,$$

where $k_3$ is a positive constant which does not depend on $M$ and $N$. By Gronwall's Lemma it follows that

$$E \sup_{0 \le s \le \mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|^2 + 4aE \int_0^{\mathcal{T}_N} e^{-2as} \|x_N(s) - x_M(s)\|_V^2 ds = 0.$$

Since $e^{-2as} \ge e^{-2aT}$ for all $s \in [0, T]$, we conclude that

$$E \sup_{0 \le s \le \mathcal{T}_N} \|x_N(s) - x_M(s)\|^2 + 4aE \int_0^{\mathcal{T}_N} \|x_N(s) - x_M(s)\|_V^2 ds = 0.$$

Hence by Markov's inequality it follows that (9) holds. $\square$

**Lemma 3.4.** *There exists an adapted process* $\left(x(t)\right)_{t\in[0,T]}$ *satisfying*

$$\lim_{N\to\infty}\left(\sup_{0\le t\le T}\|x_N(t)-x(t)\|^2+\int_0^T\|x_N(t)-x(t)\|_V^2\,dt\right)=0 \qquad \textit{for a.e.}\quad \omega\in\Omega$$

*and which has continuous trajectories in* $H$.

*Proof.* Let $N$ be any natural number. By the definition of $\mathcal{T}_N$ we have

$$P\{\mathcal{T}_N<T\}=P\Big\{\sup_{0\le t\le T}\|x_N(t)\|^2+\int_0^T\|x_N(t)\|_V^2\,dt>N^2\Big\}.$$

Furthermore, from Markov's inequality and from Lemma 3,2 we get

$$P\Big\{\sup_{0\le t\le T}\|x_N(t)\|^2 \;+\; \int_0^T\|x_N(t)\|_V^2\,dt>N^2\Big\}\le$$

$$\le\;\frac{1}{N^2}E\Big(\sup_{0\le t\le T}\|x_N(t)\|^2+\int_0^T\|x_N(t)\|_V^2\,dt\Big)\le\frac{c_1}{N^2}.$$

Consequently we conclude that

$$P\{\mathcal{T}_N<T\}\le\frac{c_1}{N^2}. \tag{15}$$

Next we notice that

$$P\Big\{\sup_{0\le t\le T}\|x_N(t)-x_{N+1}(t)\|^2+\int_0^T\|x_N(t)-x_{N+1}(t)\|_V^2\,dt>\epsilon\Big\}\le P\{\mathcal{T}_N<T\}+$$

$$+P\Big\{\sup_{0\le t\le \mathcal{T}_N}\|x_N(t)-x_{N+1}(t)\|^2+\int_0^{\mathcal{T}_N}\|x_N(t)-x_{N+1}(t)\|_V^2\,dt>\epsilon\Big\}$$

for any $\epsilon>0$. By applying Lemma 3.3 and (15) it follows that

$$P\Big\{\sup_{0\le t\le T}\|x_N(t)-x_{N+1}(t)\|^2+\int_0^T\|x_N(t)-x_{N+1}(t)\|_V^2\,dt>\epsilon\Big\}\le\frac{c_1}{N^2}.$$

Hence, by Borel-Cantelli's Lemma, there exists an adapted process $\left(x(t)\right)_{t\in[0,T]}$ such that

$$\sup_{0\le t\le T}\|x_N(t)-x(t)\|^2+\int_0^T\|x_N(t)-x(t)\|_V^2\,dt\to 0 \qquad \text{for a.e. } \omega\in\Omega.$$

Since $\left(x_N(t)\right)_{t\in[0,T]}$ has continuous trajectories in $H$ and the sequence $(x_N)$ of processes converges uniformly to $x$ in the norm of $H$, we conclude that $\left(x(t)\right)_{t\in[0,T]}$ has also continuous trajectories in $H$. □

**Lemma 3.5.** *For all $t \in [0,T]$, all $v \in V$ and a.e. $\omega \in \Omega$ the following convergence holds:*

$$\langle \int_0^t Ax_N(s)ds, v\rangle \to \langle \int_0^t Ax(s)ds, v\rangle.$$

*Proof.* We define $\hat{A} : \mathcal{L}_V^2[0,T] \to \left(\mathcal{L}_V^2[0,T]\right)^*$ by

$$\left(\hat{A}(u)\right)(v) = \int_0^T \langle Au(s), v(s)\rangle ds.$$

Since $A$ is a monotone and hemicontinuous operator, it is maximal monotone (see [10], p. 474). It can be proved that $\hat{A}$ is also maximal monotone.

From Lemma 3.4 it follows that there exists an $\Omega' \subseteq \Omega$ with $P(\Omega') = 1$ such that

$$\lim_{N\to\infty} \Big( \sup_{0\le t\le T} \|x_N(t) - x(t)\|^2 + \int_0^T \|x_N(t) - x(t)\|_V^2 dt\Big) = 0 \quad \text{for all } \omega \in \Omega'.$$

Let $\omega \in \Omega'$ be arbitrarily choosen. The sequence $\left(x_N(\omega)\right)$ converges to $x(\omega)$ in the norm of the space $\mathcal{L}_V^2[0,T]$ and $A$ satisfies the growth condition given in (ii). Consequently $\left(\hat{A}x_N(\omega)\right)$ is a bounded sequence in $\mathcal{L}_{V^*}^2[0,T]$. By applying properties of the weak convergence (see [9], Theorem 21.D, p. 255) we conclude that there exist a function $B(\omega) \in \left(\mathcal{L}_V^2[0,T]\right)^*$ and a subsequence $\left(\hat{A}x_{N'}(\omega)\right)$ of $\left(\hat{A}x_N(\omega)\right)$ which converges weakly to $B(\omega)$ in the space $\left(\mathcal{L}_V^2[0,T]\right)^*$. From the maximal monotonicity of $\hat{A}$ it follows that $\hat{A}x(\omega) = B(\omega)$ (see [10], Proposition 31.6, p. 821). Hence

$$\hat{A}x_{N'}(\omega) \rightharpoonup \hat{A}x(\omega) \quad \text{in} \quad \left(\mathcal{L}_V^2[0,T]\right)^*.$$

On the other hand, the maximal monotonicity of $\hat{A}$ implies that each weakly convergent subsequence of $\left(\hat{A}x_N(\omega)\right)$ has the same limit $\hat{A}x(\omega)$. Consequently the whole sequence $\left(\hat{A}x_N(\omega)\right)$ must converge weakly to $\hat{A}x(\omega)$ (see [9], Proposition 21.23, p. 258). This means that

$$\int_0^T \langle Ax_N(\omega, s) - Ax(\omega, s), v(s)\rangle ds \to 0$$

for all $v \in \mathcal{L}^2_V[0,T]$ and all $\omega \in \Omega'$, and hence

$$\langle \int_0^t \Big(Ax_N(s) - Ax(s)\Big)ds, v\rangle \to 0 \quad \text{for all } t \in [0,T],\ v \in V,\ \text{a.e. } \omega \in \Omega. \qquad \square$$

**Lemma 3.6.** *For all $t \in [0,T]$ the following convergences hold:*

$$\int_0^t F_N(s, x_N(s))ds \to \int_0^t F(s, x(s))ds \quad a.e. \quad \omega \in \Omega,$$

$$\int_0^t G_N(s, x_N(s))dw(s) \xrightarrow{P} \int_0^t G(s, x(s))dw(s).$$

*Proof.* From Lemma 3.4 it follows that there exists an $\Omega' \subseteq \Omega$ with $P(\Omega') = 1$ such that

$$\lim_{N\to\infty} \Big( \sup_{0\le t\le T} \|x_N(t) - x(t)\|^2 + \int_0^T \|x_N(t) - x(t)\|_V^2 dt\Big) = 0 \quad \text{for all } \omega \in \Omega'.$$

Let $\omega \in \Omega'$ be arbitrarily choosen. Since $\Big(x(t)\Big)_{t\in[0,T]}$ has continuous trajectories in $H$, it follows that there exists a constant $c_2(\omega) > 0$ such that

$$\|x(\omega, t)\|^2 \le c_2(\omega) \qquad \text{for all} \quad t \in [0,T].$$

The sequence $(x_N(\omega))$ of processes converges uniformly (with respect to $t$) to $x(\omega)$ (see Lemma 3.4) and hence there exists an index $N_0(\omega)$ such that $N_0(\omega) \ge c_2(\omega)$ and

$$\|x_N(\omega, t)\|^2 \le N_0(\omega) \qquad \text{for all} \quad t \in [0,T] \quad \text{and all} \quad N \ge N_0(\omega).$$

Hence for all $t \in [0,T]$ and all $N \ge N_0(\omega)$ we have

$$F_N(t, x_N(\omega,t)) = F(t, x_N(\omega,t)), \qquad G_N(t, x_N(\omega,t)) = G(t, x_N(\omega,t)),$$

and

$$F_N(t, x(\omega,t)) = F(t, x(\omega,t)), \qquad G_N(t, x(\omega,t)) = G(t, x(\omega,t)).$$

Since the functions $F, G$ are locally Lipschitz continuous, we conclude by Lemma 3.4 that for all $t \in [0,T]$ and all $\omega \in \Omega'$ we have

$$F_N(t, x_N(\omega,t)) \to F(t, x(\omega,t)), \qquad G_N(t, x_N(\omega,t)) \to G(t, x(\omega,t))$$

and

$$\int_0^t F_N(s, x_N(\omega,s))ds \to \int_0^t F(s, x(\omega,s))ds,$$

$$\int_0^t \|G_N(s, x_N(\omega, s)) - G(s, x(\omega, s))\|^2 ds \to 0.$$

From the definition of the stochastic integral (see [5]) it follows that

$$\int_0^t G_N(s, x_N(s))dw(s) \xrightarrow{P} \int_0^t G(s, x(s))dw(s). \qquad \square$$

Now we can state the main result of this paper.

**Theorem 3.7.** *The following assertions are true:*

1° *Problem* (P) *admits a solution.*

2° *Any solution of* (P) *has continuous trajectories in $H$.*

3° *The solution of problem* (P) *is unique with probability 1.*

*Proof.* 1° Since $\left(x_N(t)\right)_{t\in[0,T]}$ is a solution of problem ($P_N$) we have

$$\begin{aligned}(x_N(t) - x_0, v) = \; & - \int_0^t \langle Ax_N(s), v\rangle ds + \int_0^t (F_N(s, x_N(s)), v)ds + \\ & + \int_0^t (G_N(s, x_N(s)), v)dw(s)\end{aligned}$$

for a.e. $\omega \in \Omega$ and for all $v \in V, t \in [0,T]$. Passing to the limit, when $N \to \infty$, and applying the Lemmas 3.5 and 3.6 (the convergence in probability implies the convergence a.e. for a subsequence) we obtain

$$(x(t) - x_0, v) = -\int_0^t \langle Ax(s), v\rangle ds + \int_0^t (F(s, x(s)), v)ds + \int_0^t (G(s, x(s)), v)dw(s)$$

for a.e. $\omega \in \Omega$ and for all $v \in V, t \in [0,T]$. Thus $\left(x(t)\right)_{t\in[0,T]}$ is a solution of problem (P).

2° In 1° we showed that the process $\left(x(t)\right)_{t\in[0,T]}$ has a stochastic differential over the evolution triplet $(V, H, V^*)$. Then it follows from the Ito formula for $\|\cdot\|^2$ that there exists a continuous modification (in $H$) of $\left(x(t)\right)_{t\in[0,T]}$. Subsequently we can identify the solution $\left(x(t)\right)_{t\in[0,T]}$ of (P) with a process which has continuous trajectories in $H$ (see [2], Theorem 3.4, p. 42).

3° Let $\left(u_1(t)\right)_{t\in[0,T]}$ and $\left(u_2(t)\right)_{t\in[0,T]}$ be two solutions of problem (P). By applying the Ito formula (with $\|\cdot\|^2$) to the difference of this two processes we obtain

$$\|u_1(t)-u_2(t)\|^2 = \int_0^t \|G(s,u_1(s))-G(s,u_2(s))\|^2 ds - \tag{16}$$
$$- \quad 2\int_0^t \langle Au_1(s)-Au_2(s), u_1(s)-u_2(s)\rangle ds +$$
$$+ \quad 2\int_0^t \left(F(s,u_1(s))-F(s,u_2(s)), u_1(s)-u_2(s)\right) ds +$$
$$+ \quad 2\int_0^t \left(G(s,u_1(s))-G(s,u_2(s)), u_1(s)-u_2(s)\right) dw(s)$$

for all $t\in[0,T]$ and a.e. $\omega\in\Omega$.

Let $\mathcal{T}:\Omega\to\mathbb{R}$ be defined by

$$\mathcal{T}(\omega) = \sup\left\{t\in[0,T] \mid \sup_{0\le s\le t}\|u_1(s)\|^2 \le N^2 \quad \text{and} \quad \sup_{0\le s\le t}\|u_2(s)\|^2 \le N^2\right\}.$$

For all $s\in[0,\mathcal{T}]$ we have

$$\|F(s,u_1(s))-F(s,u_2(s))\|^2 \le \beta_N\|u_1(s)-u_2(s)\|^2,$$

$$\|G(s,u_1(s))-G(s,u_2(s))\|^2 \le \beta_N\|u_1(s)-u_2(s)\|^2.$$

By using the monotonicity of $A$ and the inequalities from above, it follows from (16) that

$$\|u_1(t\wedge\mathcal{T})-u_2(t\wedge\mathcal{T})\|^2 \le (2\beta_N+1)\int_0^{t\wedge\mathcal{T}} \|u_1(s)-u_2(s)\|^2 ds +$$
$$+ \quad 2\int_0^{t\wedge\mathcal{T}} \left(G(s,u_1(s))-G(s,u_2(s)), u_1(s)-u_2(s)\right) dw(s)$$

for all $t\in[0,T]$. Since $\mathcal{T}$ is a stopping time, we have

$$E\int_0^{t\wedge\mathcal{T}} \left(G(s,u_1(s))-G(s,u_2(s)), u_1(s)-u_2(s)\right) dw(s) = 0 \quad \text{for all} \quad t\in[0,T].$$

Therefore

$$E\|u_1(t\wedge\mathcal{T})-u_2(t\wedge\mathcal{T})\|^2 \le (2\beta_N+1)\int_0^t E\|u_1(s\wedge\mathcal{T})-u_2(s)\|^2 ds.$$

By applying Gronwall's Lemma we obtain

$$E\|u_1(t \wedge T) - u_2(t \wedge T)\|^2 = 0 \quad \text{for all} \quad t \in [0,T].$$

Hence

$$E\|u_1(t) - u_2(t)\|^2 = 0 \quad \text{for all} \quad t \in [0,T].$$

We notice that

$$P\{u_1(t) \neq u_2(t)\} \leq P\{\sup_{0 \leq s \leq T} \|u_1(s)\|^2 \geq N^2\} + P\{\sup_{0 \leq s \leq T} \|u_2(s)\|^2 \geq N^2\}. \quad (1$$

But $\left(u_1(t)\right)_{t \in [0,T]}$ and $\left(u_2(t)\right)_{t \in [0,T]}$ being solutions of (P), they have continuous tra jectories in $H$ and hence they are a.e. bounded. Consequently

$$\lim_{N \to \infty} P\{\sup_{0 \leq s \leq T} \|u_1(s)\|^2 \geq N^2\} = 0, \quad \lim_{N \to \infty} P\{\sup_{0 \leq s \leq T} \|u_2(s)\|^2 \geq N^2\} = 0.$$

Therefore (17) implies that

$$P\{u_1(t) = u_2(t)\} = 1 \qquad \text{for all} \quad t \in [0,T].$$

By applying again the continuity of the two considered processes we get

$$P\left\{\sup_{0 \leq s \leq T} \|u_1(s) - u_2(s)\|^2 = 0\right\} = 1.$$

This means that the solution of problem (P) is unique with probability 1. □

*Remark.* From Lemma 3.4 and Theorem 3.7 we conclude that the process $\left(x_N(t)\right)_{t \in [0,T]}$ approximates (almost surely) the solution of problem (P). Thus the results of this pape represent also a constructive method for proving the existence of the solution of th considered problem.

## References


[1] P. L. Chow: ***Stability of nonlinear stochastic evolution equations.*** J. Math. Anal. Appl 89(1982), 400-419.

[2] W. Grecksch, C. Tudor: *Stochastic Evolution Equations.* Akademie Verlag, Berlin, 1995.

[3] X. Mao: *Exponential Stability of Stochastic Differential Equations.* Marcel Dekker, Inc. New York, 1994.

[4] H. P. Mc Kean Jr.: *Nagumo's equation.* Adv. in Math. 4(1970), 209-223.

[5] B. Oksendal: *Stochastic Differential Equations.* Springer-Verlag, Berlin, 1985.

[6] J. vom Scheidt: *Stochastic Equations of Mathematical Physics.* Akademie Verlag, Berlin, 1990.

[7] V. M. Tichomirov: *Theorie der Extremalaufgaben.* VEB Deutscher Verlag der Wis senschaften, Berlin, 1979.

[8] M. I. Vishik, A. O. Fursikov: *Mathematical Problems of Statistical Hydromechanics.* Kluwer Academic Publishers, Dordrecht, 1988.

[9] E. Zeidler: *Nonlinear Functional Analysis and it's Applications.* Vol. II/A: *Linear Monotone Operators.* Springer-Verlag, New York, 1990.

[10] E. Zeidler: *Nonlinear Functional Analysis and it's Applications*. Vol. II/B: *Nonlinear Monotone Operators*. Springer-Verlag, New York, 1990.

BABEŞ-BOLYAI UNIVERSITY, FACULTY OF MATHEMATICS AND COMPUTER SCIENCE, STR. KOGĂLNICEANU NR. 1, RO-3400 CLUJ-NAPOCA, ROMANIA

# FREE CONVECTION IN AN INCLINED SQUARE ENCLOSURE FILLED WITH A HEAT-GENERATING POROUS MEDIUM

I. CHIOREAN AND I. POP

**Abstract.** The numerical solution of a steady state problem of free convection in an inclined enclosure bounded by four rigid walls of constant temperature, filled with porous medium, is studied. The multigrid iterative method is used. Several graphics showing the solution and the heat-transfer are given.

## Introduction

Free convection in an enclosure filled with a fluid-saturated porous medium has occupied the center stage in many fundamental heat transfer analysis. A great deal of research both theoretical and experimental has been accumulated on this topic during the last two decades, see the recent monograph by Nield and Bejan [1]. However, there is relatively little work published on free convection in enclosures filled with heat-generating porous medium. To the authors' knowledge, studies reported for this case are only those by Vasseur and al. [2] and Prasad [3]. Thus, Vasseur and al. [2] have presented numerical solutions for the problem of free convection in a porous layer bounded by two horizontal concentric cylinders with uniformly distributed energy sources. Prasad [3], on the other hand, has studied numerically the steady free convection in a vertical rectangular cavity filled with a heat generating saturate porous medium where the vertical walls of the cavity are isothermally cooled and the horizontal walls are adiabatic.

The purpose of the present paper is to analyze the steady two dimensional free convection in an inclined square enclosure bounded by four rigid walls of constant temperature and containing a heat-generating fluid-saturated porous medium. The problem is studied numerically using the finite difference scheme and a multigrid method for a wide range of the Rayleigh numbers, Ra, and inclination and $\phi$. Solutions for the flow and temperature fields are presented in the form of streamlines and isotherms.

Received by the editors: October 24, 1996.

1991 *Mathematics Subject Classification.* 76S05.

*Key words and phrases.* convection, porous medium, rigid walls, multigrid methods.

It is worth mentioning to this end that free convection in enclosed spaces with internal heat generation is of possible importance in the problems of radioactive waste heat removal, electrolytic processes, geothermal energy, nuclear fusion and exothermic chemical reactions among other practical applications.

## Basic equations

The schematic diagram of a two-dimensional square cavity inclined at an angle $\phi$ to the horizontal is shown in Fig.1.

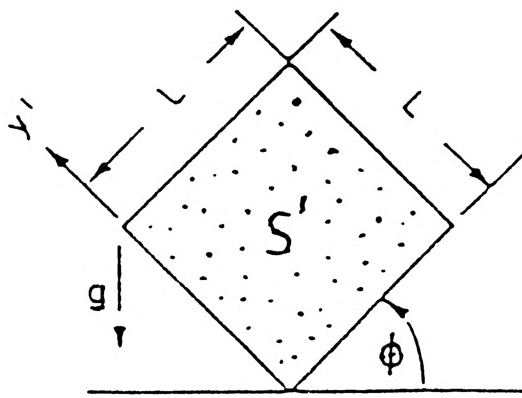



FIGURE 1. Schematic diagram of the enclosure

The cavity is filled with a fluid saturated porous medium which generates heat at a uniform rate $S$. All the walls of cavity are assumed to be impermeable and are maintained at the constant temperature $T$. In the porous medium, the Darcy law is assumed to hold, the fluid is assumed to be Bussineque fluid, and the viscous drag and inertia terms of the momentum equations are neglected because their magnitude are of small order compared to other terms of low Darcy number.

With these assumptions, the conservative equations of mass, momentum and energy for steady, two-dimensional flow in an isotropic porous medium can be written as:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{1}$$

$$\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x} = \frac{g\beta K}{v}\left(\frac{\partial T}{\partial y}\sin\phi - \frac{\partial T}{\partial x}\cos\phi\right) \tag{2}$$

$$u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} = \alpha\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + \frac{S}{(\rho c)_f} \tag{3}$$

where $(u, v)$ are the velocity components along $(x, y)$-axes, $T$ is the temperature and $S$ is the rate of volumetric heat generation. Equations (2) and (3) are subject to the boundary conditions

$$u = v = 0,\ T = T_0 \text{ on } x = 0, 1 \text{ and } y = 0, 1. \tag{4}$$

Next, we introduce the non-dimensional variables defined as

$$X = x/l, \quad Y = y/l, \quad U = (l/\alpha)u, \quad V = (l/\alpha)v, \quad \theta = (k/Sl^2)(T - T_0).$$

Equations (1) to (3) then become

$$\nabla^2\psi = \mathrm{Ra}\left(\frac{\partial\theta}{\partial Y}\sin\phi - \frac{\partial\theta}{\partial X}\cos\phi\right) \tag{5}$$

$$\nabla^2\theta + 1 = \frac{\partial\psi}{\partial Y}\frac{\partial\theta}{\partial X} - \frac{\partial\psi}{\partial X}\frac{\partial\theta}{\partial Y} \tag{6}$$

subject to the boundary conditions

$$\psi = \phi = 0 \text{ for } X = 0, 1 \text{ and } Y = 0, 1. \tag{7}$$

Here $\nabla^2$ is the two-dimensional Laplacian and $\psi$ is the stream function defined in th usual way

$$U = \frac{\partial\psi}{\partial Y}, \quad V = -\frac{\partial\psi}{\partial X} \tag{8}$$

Also, Ra is the modified Rayleigh number for a porous medium defined as

$$\mathrm{Ra} = b\beta K(sl^2/k)l/\alpha v. \tag{9}$$

## Results and discussions

Equations (5) and (6) along with the boundary conditions (7) were solved numerically using the multigrid method with a Gauss-Seidel smoother. The solution technique is well described in the literature [5,6]. A convergence criterion as

$$\left\| \frac{\psi^n - \psi^{n-1}}{\psi^n} \right\| < 10^{-5} \tag{10}$$

was used for the present computations, both for $\psi$ and for $\theta$ functions, where $n$ is the number of iterations and the Euclidean norm is considered.

Numerical results were obtained for a wide range values of the Rayleigh number, Ra, and inclination angle, $\phi$. However, we present results here only for Ra=700 and 4500 when $\phi = 0°$ (vertical square enclosure) and $\phi = 60°$ (inclined enclosure). Representative streamlines and isotherms are shown in Figs.2 to 5. These figures clearly reveal the influence of natural convection for all Rayleigh numbers considered. It is also seen that there is a strong pair of counter-rotating rolls for the inclined enclosure. The hot interior fluid moves upward in the middle along the line nearly parallel to the direction of gravity, then turns to the direction of the upper edge and divides the whole cross-section in two halves. The flow divides at the top and moves downward separately along the cold side walls.

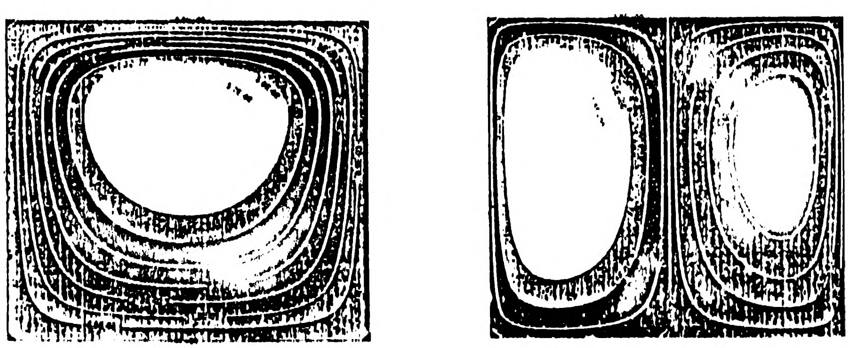


FIGURE 2. Temperature and stream function for Ra=700 and $\phi = 0°$

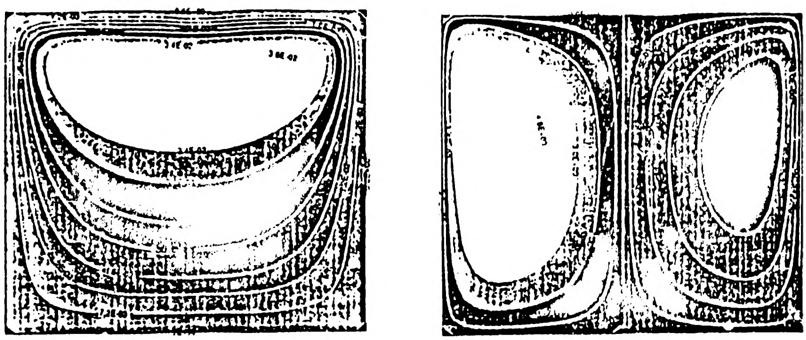


FIGURE 3. Temperature and stream function for Ra=4500 and $\phi = 0°$

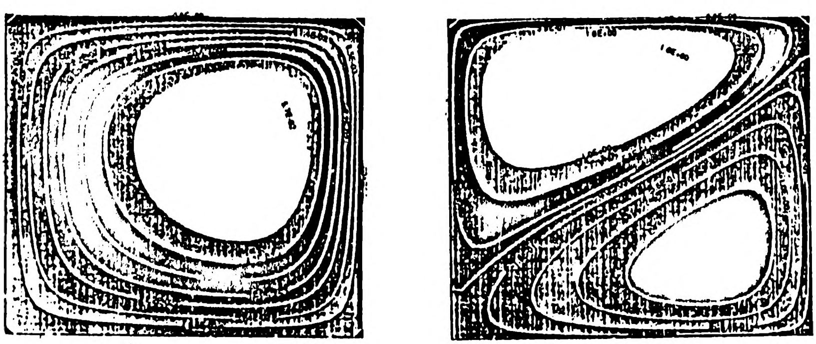


FIGURE 4. Temperature and stream function for Ra=700 and $\phi = 60°$

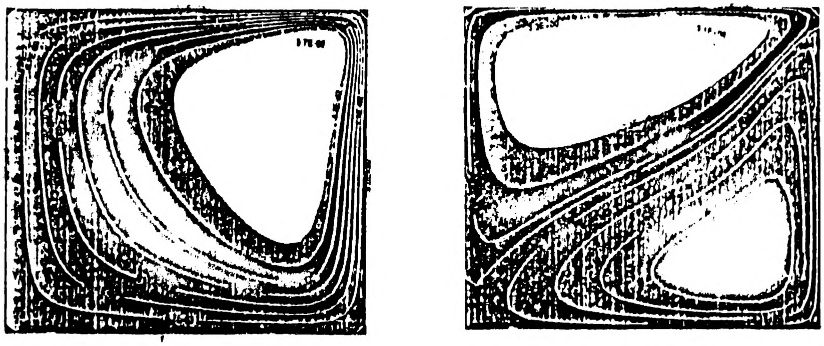


FIGURE 5. Temperature and stream function for Ra=4500 and $\phi = 60°$

The average Nusslet number is defined as in [7]:

$$\overline{\mathrm{Nu}} = \frac{1}{2}\int_0^1 \left(\frac{\partial\theta}{\partial n}\right)_{wall} dn \tag{11}$$

where $n$ denotes the $x-$ or $y-$ direction. The variation of $\overline{\mathrm{Nu}}$ with $\phi = 0°$ and $\phi = 60°$ is shown in Figs.6 and 7 for Ra=10, 100 and 1500, respectively.

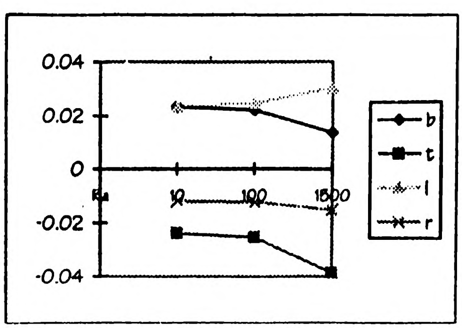



FIGURE 6. The variation of Nusselt number for $\phi = 0°$

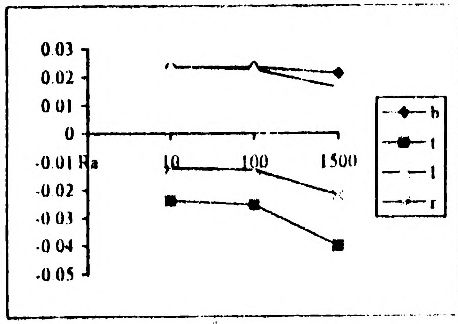



FIGURE 7. The variation of Nusselt number for $\phi = 60°$

**Nomenclature**

$g$ - gravitational acceleration

$V$ - velocity of the fluid

$p$ - pressure of the fluid

$T$ - temperature of the fluid

$K$ - permeability of the saturated porous medium

$k$ - thermal conductivity of porous medium

$S$ - rate of internal heat generation of porous medium

Ra - internal Rayleigh number

$L$ - characteristic length of the porous medium

$t$ - time

$u, v$ - velocity components

$x, y$ - coordinates

**Greek symbols**

$\alpha$ - thermal diffusivity

$\rho$ - density of fluid

$\mu$ - viscosity of fluid

$\theta$ - nondimensional temperature

$(\rho c)_f$ - heat capacity of fluid

$(\rho c)_p$ - heat capacity of porous medium

$\beta$ - thermal expansion coefficient

$\psi$ - dimensionless stream function

$\phi$ - angular coordinate

**Superscripts**

$n$ - number of iterations

**Subscripts** 0 - value at reference temperature and density

## References


[1] Nield, D.A. and Bejan, A., *Convection in Porous Media*, Springer-Verlag, Berlin, 1992.

[2] Vasseur, P., Nguyen, I.H., Robillard, L., Thi, V.K.I., *Natural convection between horizontal concentric cylinders filled with a porous layer with internal heat generation*, Int. J. Heat Mass Transfer 27, 1984, 337-349.

[3] Prasad, V., *Thermal Convection in a Rectangular Cavity filled with a Heat-Generating, Darcy Porous Medium*, J. heat Transfer, 109, 1987, 697-703.

[4] May, H.-O., *A numerical study on natural convection in an inclined square enclosure containing internal heat sources*, Int. J. Heat Mass Transfer, 34, 1991, 919-928.

[5] Roache, J.P., *Computational Fluid Dynamics*, Hermosa, Albuquerque, New Mexico, 1985.

[6] Hackbusch, W., *Multigrid Methods and Applications*, Springer Verlag, Berlin, 1985.

[7] Lee, I.-H., Goldstein, R.I., *An experimental study on natural convection heat transfer in an inclined square enclosure containing internal energy sources*, J. Heat Transfer, 110, 1988, 345-349.


"BABEŞ-BOLYAI" UNIVERSITY, FACULTY OF MATHEMATICS AND INFORMATICS, STR. KOGĂLNICEANU, NO. 1, 3400 CLUJ-NAPOCA, ROMÂNIA

STUDIA UNIV. "BABEŞ-BOLYAI", MATHEMATICA, Volume XLII, Number 4, December 199

# A NEW PROOF FOR A BASIC THEOREM CONCERNING ITERATIVE METHODS WITH $N$ STEPS

BÉLA FINTA

**Abstract.** In this paper one gives a new proof of a theorem concerning the iterative methods with $n$ steps, by using the Banach fixed point principle. Other proofs can be found in [1] and [3].

Let $(X, \rho)$ be a metric space, and $B \subset X$, $B \neq \emptyset$ a sphere. Let's consider the equation $\varphi(x) = x$, where $\varphi : B \to X$ is a given function. We can observe that the solutions of this equation are fixed points for $\varphi$. To solve this equation a possibility is to build a new function $F : B^n \to X$, where $n \geq 1$ is a fixed natural number, so that the restriction of $F$ on the diagonal set $B^n$ coincides with $\varphi$, i.e. $F(x, x, \ldots, x) = \varphi(x)$, for every $x \in B$.

Now we consider the sequence $\{x^k\}_{k \in \mathbb{N}}$ given by the following iterative method with $n$ steps:

$$x^n = F(x^{n-1}, x^{n-2}, \ldots, x^1, x^0)$$

and

$$x^{k+n} = F(x^{k+n-1}, x^{k+n-2}, \ldots, x^{k+1}, x^k)$$

for every $k = 1, 2, \ldots$ and $x^0, x^1, \ldots, x^{n-2}, x^{n-1} \in B$. The following well known theorem gives us some necessary conditions which assure us the existence and the convergence of the sequence $\{x^k\}_{k \in \mathbb{N}}$ to the fixed point of $\varphi$.

**Theorem 1.** *Let $(X, \rho)$ be a complete metric space and $B \subset X$, $B \neq \emptyset$ a closed sphere and we suppose that the function $F$ satisfies the following conditions:*

*i) transforms the set $B^n$ into $B$;*

*ii) verifies the equality $F(x, x, \ldots, x) = \varphi(x)$ for every $x \in B$;*

*iii) satisfies the Lipschitz condition: for every $y^1, y^2, \ldots, y^n, z^1, z^2, \ldots z^n \in B$ we have $\rho\left(F(y^1, y^2, \ldots, y^n), F(z^1, z^2, \ldots, z^n)\right) \leq \sum_{i=1}^{n} \alpha_i \rho(y^i, z^i)$, where $\alpha_i \geq 0$, $i =$*

Received by the editors: December 10,1996.

1991 *Mathematics Subject Classification.* 44-00, 65-00.

*Key words and phrases.* iterative methods, Banach fixed point principle.

$\overline{1,n}$ *are real numbers so that* $\sum_{i=1}^{n} \alpha_i < 1$. *It results that the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *is well defined, it is convergent for every* $x^0, x^1, \ldots, x^{n-1} \in B$ *and if we denote by* $x^* = \lim_{k\to\infty} x^k$, *then* $x^k$ *is the unique fixed point of the* $\varphi$ *in* $B$.

This theorem has several known different proofs like in [1], [2]. The reason of this work is to give a new short proof which is based on the Banach fixed point theorem.

*Proof.* We build the following function $\Phi : B^n \to B^n$ given by formulae

$$\Phi(y^n, y^{n-1}, \ldots, y^2, y^1) = (u^n, u^{n-1}, \ldots, u^2, u^1),$$

where $u^1 = F(y^n, y^{n-1}, \ldots, y^1)$ and $u^i = F(u^{i-1}, \ldots, u^1, y^n, \ldots, y^i)$ for every $i = \overline{2,n}$. We can observe that the function $\Phi$ is well defined using condition i). On the space $X^n$ we consider the distance function $\rho_\infty : X^n \times X^n \to [0, +\infty)$,

$$\rho_\infty\left((y^1, y^2, \ldots, y^n), (z^1, z^2, \ldots, z^n)\right) = \max_{1\le i\le n}\{\rho(y^i, z^i)\}.$$

It is not difficult exercise to show that $\rho_\infty$ is a metric on the space $X^n$, and because $(X, \rho)$ is a complete metric space results that $(X^n, \rho_\infty)$ is a complete metric space, too. Now, as $B$ is a closed sphere in $X$, we obtain that $B^n$ is a closed set in $X^n$, so $(B^n, \rho_\infty)$ is a complete metric space, too. It remains to show that $\Phi$ is a contraction on $B^n$. Indeed:

$$\begin{aligned}
&\rho_\infty\left(\Phi(y^n, y^{n-1}, \ldots, y^2, y^1),\ \Phi(z^n, z^{n-1}, \ldots, z^2, z^1)\right) = \\
&= \rho_\infty\left((u^n, u^{n-1}, \ldots, u^2, u^1), (v^n, v^{n-1}, \ldots, v^2, v^1)\right) = \max_{1\le i\le n}\{\rho(u^i, v^i)\}.
\end{aligned}$$

We use the Lipschitz condition:

$$\begin{aligned}
\rho(u^1, v^1) &= \rho\left(F(y^n, y^{n-1}, \ldots, y^2, y^1), F(z^n, z^{n-1}, \ldots, z^2, z^1)\right) \le \\
&\le \sum_{i=1}^{n} \alpha_i \cdot \rho(y^{n-i+1}, z^{n-i+1}) \le \sum_{i=1}^{n} \alpha_i \cdot \max_{1\le i\le n}\{\rho(y^{n-i+1}, z^{n-i+1})\} = \\
&= \left(\sum_{i=1}^{n} \alpha_i\right) \cdot \max_{1\le i\le n}\{\rho(y^i, z^i)\} = k_1 \cdot \max_{1\le i\le n}\{\rho(y^i, z^i)\},
\end{aligned}$$

where we denote by $k_1 = \sum_{i=1}^{n} \alpha_i < 1$.

$$\begin{aligned}
\rho(u^2, v^2) &= \rho\left(F(u^1, y^n, y^{n-1}, \ldots, y^2), F(v^1, z^n, z^{n-1}, \ldots, z^2)\right) \le \\
&\le \alpha_1 \cdot \rho(u^1, v^1) + \sum_{i=2}^{n} \alpha_i \cdot \rho(y^{n-i+2}, z^{n-i+2}) \le \\
&\le \alpha_1 \cdot k_1 \cdot \max_{1 \le i \le n} \{\rho(y^i, z^i)\} + \sum_{i=2}^{n} \alpha_i \cdot \max_{2 \le i \le n} \{\rho(y^i, z^i)\} \le \\
&\le \left(\alpha_1 k_1 + \sum_{i=2}^{n} \alpha_i\right) \cdot \max_{1 \le i \le n} \{\rho(y^i, z^i)\} = k_2 \cdot \max_{1 \le i \le n} \{\rho(y^i, z^i)\},
\end{aligned}$$

where we denote by $k_2 = \alpha_1 k_1 + \sum_{i=2}^{n} \alpha_i < \alpha_1 + \sum_{i=2}^{n} \alpha_i = k_1 < 1$.

By induction we show that for $i = \overline{3, n}$ we have

$$\rho(u^i, v^i) \le k_i \cdot \max_{1 \le i \le n} \{\rho(y^i, z^i)\},$$

where $k_i < k_1$. Indeed:

$$\begin{aligned}
\rho(u^i, v^i) &= \rho\left(F(u^{i-1}, \ldots, u^1, y^n, \ldots, y^i), F(v^{i-1}, \ldots, v^1, z^n, \ldots, z^i)\right) \le \\
&\le \sum_{j=1}^{i-1} \alpha_j \cdot \rho(u^{i-j}, v^{i-j}) + \sum_{j=i}^{n} \alpha_j \cdot \rho(y^{n-j+i}, z^{n-j+i}) \le \\
&\le \sum_{j=1}^{i-1} \alpha_j k_{i-j} \cdot \max_{1 \le i \le n} \{\rho(y^i, z^i)\} + \sum_{j=i}^{n} \alpha_j \cdot \max_{i \le j \le n} \{\rho(y^{n-j+i}, z^{n-j+1})\} \le \\
&\le \sum_{j=1}^{i-1} \alpha_j k_{i-j} \cdot \max_{1 \le i \le n} \{\rho(y^i, z^i)\} + \left(\sum_{j=i}^{n} \alpha_j\right) \cdot \max_{i \le j \le n} \{\rho(y^{n-j+i}, z^{n-j+1})\} \le \\
&\le \left(\sum_{j=1}^{i-1} \alpha_j k_{i-j} + \sum_{j=i}^{n} \alpha_j\right) \cdot \max_{1 \le i \le n} \{\rho(y^i, z^i)\},
\end{aligned}$$

where we denote by $k_i = \sum_{j=1}^{i-1} \alpha_j k_{i-j} + \sum_{j=i}^{n} \alpha_j$. We obtain immediately that

$$k_i < \sum_{j=1}^{i-1} \alpha_j \cdot 1 + \sum_{j=i}^{n} \alpha_j = k_1$$

, because $k_{i-j} < k_1 < 1$ for $j = \overline{1,i}$ ly:

$$\rho_\infty\left(\Phi(y^n, y^{n-1}, \ldots, y^2, y^1), (z^n, z^{n-1}, \ldots, z^2, z^1)\right\} =$$

$$= \max_{1\le i\le n}\{\rho(u^i, v^i)\} \le \max_{1\le i\le n}\left\{ \{k_i\} \cdot \max_{1\le j\le n}\{\rho(y^j, z^j)\}\right\} =$$

$$= \max_{1\le i\le n}\{k_i\} \cdot \max_{1\le j\le n}\{\rho(y^j, z^j)\} = k_1 \cdot \max_{1\le j\le n}\{\rho(y^j, z^j)\} =$$

$$= k_1 \cdot \rho_\infty\left((y^n, y^{n-1}, \ldots, y^2, y^1), (z^n, z^{n-1}, \ldots, z^2, z^1)\right),$$

where $k_1 = \max_{1\le i\le n}\{k_i\} < 1$ is the constant of contraction.

Now we choose arbitrary $x^0, x^1, \ldots, x^{n-1} \in B$. The Banach fixed point theorem assures us the existence and uniqueness of the fixed point of $\varphi : (x_n^*, \ldots, x_1^*) \in B^n$ such that the sequence $\left\{(x^{kn+n-1}, x^{kn+n-2}, \ldots, x^{kn+1}, x^{kn})\right\}_{k\in\mathbf{N}}$ generated by

$$(x^{kn+n-1}, x^{kn+n-2}, \ldots, x^{kn+1}, x^{kn}) =$$

$$= \Phi\left(x^{(k-1)\cdot n+n-1}, x^{(k-1)\cdot n+n-2}, \ldots, x^{(k-1)\cdot n+1}, x^{(k-1)\cdot n}\right)$$

for $k \in \mathbf{N}^* = \mathbf{N}\setminus\{0\}$ converges to this point. This implies for components that

$$\lim_{k\to\infty} x^{kn+i-1} = x_i^*$$

for every $i = \overline{1,n}$. If we choose as initial points

$$(x^n, x^{n-1}, \ldots, x^1) \in B^n,$$

$$(x^{n+1}, x^n, \ldots, x^2) \in B^n, \ldots, \quad \text{and}$$

$$(x^{2n-2}, x^{2n-3}, \ldots, x^{n-1}) \in B^n,$$

respectively and we generate the corresponding sequence by $\Phi$, then the first terms give us the following limits: $\lim_{k\to\infty} x^{kn+n} = \lim_{k\to\infty} x^{kn+n+1} = \cdots = \lim_{k\to\infty} x^{kn+2n-1} = x_n^*$ i.e. the sequence $\{x^k\}_{k\in\mathbf{N}}$ is convergent and exists the $\lim_{k\to\infty} x^k = x_n^* = x^* \in B$. Using the fact that for the subsequences $lim_{k\to\infty} x^{kn+i-1} = x_i^*$ for every $i = \overline{1,n}$ we obtain that $x_1^* = x_2^* = \cdots = x_n^* = x^*$. The Lipschitz condition for $F$ implies the continuity of $F$ so we can take the limit in the recurrence relation $x^{k+n} = F\left(x^{k+n-1}, x^{k+n-2}, \ldots, x^{k+1}, x^k\right)$. Consequently $x^* = F(x^*, x^*, \ldots, x^*, x^*) = \varphi(x^*)$ using ii). So we obtained that $x^*$ is a fixed point for $\varphi$. If we suppose that $y^*$ is another fixed point of $\varphi$ then the following relations: $\rho(x^*, y^*) = \rho(\varphi(x^*), \varphi(y^*)) = \rho(F(x^*, x^*, \ldots, x^*), F(y^*, y^*, \ldots, y^*)) \le$ $\le \sum_{i=1}^n \alpha_i \rho(x^*, y^*)$ imply that $1 \le \sum_{i=1}^n \alpha_i$, which means a contradiction with the assumption iii) q.e.d. □

A similar proof appears in [3] for the case of the space $X \times X$, which was the starting point for this proof on the space $X^n$. If we work on the whole space $X^n$ instead of the set $B^n$, then we obtain the shown theorem without condition i).

## References


[1] Gh. Coman, G. Pavel, I. Rus, I.A. Rus, Introducere în teoria ecuaţiilor operatoriale, Editura Dacia, Cluj-Napoca, 1976.

[2] I. Păvăloiu, Rezolvarea ecuaţiilor prin interpolare, editura Dacia, Cluj-Napoca, 1981.

[3] B. Finta, Some remarks about iterative methods with two steps, Octogon, vol. 5, nr. 2, October, Braşov, (in appear).


Târgu-Mureş Technical University, 4300 Târgu-Mureş, Romania

# FERMAT'S EQUATION IN THE SET OF MATRICES AND SPECIAL FUNCTIONS

ALEKSANDER GRYTCZUK

**Abstract.** In this paper we give an extension and stronger form of our result presented in Theorem 1 of [3]. Moreover we consider some connections between Fermat's equation in the set of matrices and the set of special functions.

## 1. Introduction

Following recently result given by A. Wiles [8] and R. Taylor and A. Wiles [7] we know that the Fermat's equation

$$X^n + Y^n = Z^n \tag{1}$$

has no solutions in positive integers $X, Y, Z$ if $n > 2$.

In contrast to this situation the equation (1) has infinitely many solutions in $2 \times 2$ integral matrices $X, Y, Z$ for the exponent $n = 4$. This fact has been discovered by R.Z. Domiaty [2] in 1966. Namely, he remarked that if

$$X = \begin{pmatrix} 0 & 1 \\ a & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & 1 \\ b & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 0 & 1 \\ c & 0 \end{pmatrix}$$

where $a, b, c$ are integer solutions of the Pythagorean equation $a^2 + b^2 = c^2$ then

$$X^4 + Y^4 = Z^4.$$

Another results connected with solvability of (1) are described by P. Ribenboim in [6]. Important problem in these investigations is to give a necessary and sufficient condition for solvability (1) in the set of matrices. For the case when the matrices $X, Y, Z$ belongs to $SL_2(Z)$, $SL_3(Z)$, $GL_3(Z)$ such conditions has been find recently by A. Khazanov [4]. In our paper [3], Thm.1 has given a necessary condition for solvability (1) in the set of $2 \times 2$ integral matrices. In the present paper we give an extension of our

Received by the editors: December 10, 1996.
1991 *Mathematics Subject Classification.* 11C20, 11D41
*Key words and phrases.* Fermat's equation, special functions

result to the set of arbitrary $2\times 2$ matrices under some weaker assumption. Namely following theorem is true:

**Theorem 1.** *Let* $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $Y = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ *and* $Z = \begin{pmatrix} p & q \\ u & v \end{pmatrix}$ *be a given matrices such that one of* $X, Y, Z$*; say* $X$ *satisfies* $X^k \neq \gamma I$ *for every positive integer* $k$*, where* $\gamma \neq 0$ *and* $I$ *is identity matrix. If the equation (1) is satisfied by the matrices* $X, Y, Z$ *then*

$$\det \begin{pmatrix} a-d & e-h & p-v \\ b & f & q \\ c & g & u \end{pmatrix} = 0. \tag{2}$$

Moreover, we prove the following:

**Theorem 2.** *If the matrix* $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \neq 0$ *has eigenvalues* $\alpha, \beta$ *such that* $\alpha \neq \beta$ *and* $\frac{\alpha}{\beta}$ *is not a root of unity then for every positive integer* $k$ *we have* $X^k \neq \gamma I$.

It is easy to see that if $X, Y, Z$ are integral matrices then Theorem 1 and Theorem 2 implies our Theorem 1 of [3]. Moreover we observe that if the matrix $X = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}$, where $a \neq 0$, $b \neq 0$ then by induction follows that for every natural number $k$ we have

$$X^k = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}^k = \begin{pmatrix} a^k & ka^{k-1}b \\ 0 & a^k \end{pmatrix} \neq \gamma I.$$

But the matrix $X$ has eigenvalues $\alpha = \beta = a$, so the converse is not true in general case. Hence, the result contained in the Theorem 1 is stronger than corresponding result in [3]. Further we establish some connections between the solutions of (1) in the set of matrices and the solutions of this equation in the set of special functions of one variables satisfying the condition:

$$f(x+w) = \frac{af(x)+b}{cf(x)+d}, \tag{$*$}$$

where $w \neq 0$ and $a, b, c, d$ are the elements of a fixed number field $K$.

Such type functions when $K = R$ has been considered by A.W. Kuzel in [5].

## 2. Proof of the Theorem 1

In the proof of the Theorem 1 we use the following

**Lemma.** *Let* $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *be a given matrix. Then for every natural number* $n \geq 2$

$$X^n = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^n = \begin{pmatrix} F_1(a) & b\Psi_1 \\ c\Psi_1 & F_1(d) \end{pmatrix}, \tag{3}$$

*where* $F_1(a) = F_1(a,b,c,d)$, $F_1(d) = F_1(d,a,b,c)$ *and* $\Psi_1 = \Psi_1(a,b,c,d)$ *are polynomials such that*

$$F_1(a) - F_1(d) = (a-d)\Psi_1. \tag{4}$$

The proof of this Lemma follows by induction and is similar to the proof of the corresponding Lemma of [1].

Now, suppose that (1) is satisfied by the matrices $X, Y, Z$ for some $n \geq 2$. Then by Lemma it follows that

$$\begin{pmatrix} F_1(a) & b\Psi_1 \\ c\Psi_1 & F_1(d) \end{pmatrix} + \begin{pmatrix} F_2(e) & f\Psi_2 \\ g\Psi_2 & F_2(h) \end{pmatrix} = \begin{pmatrix} F_3(p) & q\Psi_3 \\ u\Psi_3 & F_3(v) \end{pmatrix}. \tag{5}$$

From (5) we obtain

$$\begin{cases} F_1(a) + F_2(e) = F_3(p) \\ F_1(d) + F_2(h) = F_3(v) \\ b\Psi_1 + f\Psi_2 = q\Psi_3 \\ c\Psi_1 + g\Psi_2 = u\Psi_3 \end{cases} \tag{6}$$

By (4) of Lemma follows

$$F_1(a) - F_1(d) = (a-d)\Psi_1, \; F_2(e) - F_2(h) = (e-h)\Psi_2, \; F_3(p) - F_3(v) = (p-v)\Psi_3. \tag{7}$$

From the first two equations of (6) we obtain

$$F_1(a) - F_1(d) + F_2(e) - F_2(h) = F_3(p) - F_3(v) \tag{8}$$

Substituting (7) to (8) we get

$$(a-d)\Psi_1 + (e-h)\Psi_2 = (p-v)\Psi_3. \tag{9}$$

Hence, by (9) and (6) it follows that the system of the following equations

$$\begin{cases} (a-d)\Psi_1 + (e-h)\Psi_2 - (p-v)\Psi)3 = 0 \\ b\Psi_1 + f\Psi_2 - q\Psi_3 = 0 \\ c\Psi_1 + g\Psi_2 - u\Psi_3 = 0 \end{cases} \tag{10}$$

has a solution with respect to $\Psi_1, \Psi_2, \Psi_3$.

By the assumption of the Theorem 1 follows that one of $\Psi_1, \Psi_2, \Psi_3$ is different from zero. Really, suppose that for example $\Psi_1 = 0$. Then using Lemma for the matrix $X$ we obtain

$$X^n = \begin{pmatrix} F_1(a) & b\Psi_1 \\ c\Psi_1 & F_1(d) \end{pmatrix} = \begin{pmatrix} F_1(a) & 0 \\ 0 & F_1(d) \end{pmatrix}$$

and $F_1(a) - F_1(d) = (a-d)\Psi_1 = 0$. Hence $F_1(a) = F_1(d) = \gamma \neq 0$ and $X^n = \gamma I$, so contrary to the assumption of the theorem. Therefore we have $\Psi_1 \neq 0$ and the system of homogeneous equations (10) has non-trivial solution with respect to $\Psi_1, \Psi_2, \Psi_3$. Thus by well-known result about such systems we obtain that

$$\det \begin{pmatrix} a-d & e-h & p-v \\ b & f & q \\ c & g & u \end{pmatrix} = 0.$$

The proof of the Theorem 1 is complete.

## 3. Proof of Theorem 2

We prove the Theorem 2 by contraposition. Suppose that for some positive integer $k$

$$X^k = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^k = \gamma I, \quad \gamma \neq 0. \tag{11}$$

Let $\alpha, \beta$ be the eigenvalues of the matrix $X$. Then it is well-known that the matrix $X^k$ has the eigenvalues $\alpha^n$, $\beta^k$ such that

$$Tr X^k = \alpha^k + \beta^k, \quad \det X^k = \alpha^k \beta^k. \tag{12}$$

From (11) and (12) we obtain

$$2\gamma = \alpha^k + \beta^k, \quad \gamma^2 = \alpha^k \beta^k. \tag{13}$$

It is easy to see that (13) implies $\alpha^k - \beta^k = 0$, so $\alpha = \beta$ or $\frac{\alpha}{\beta}$ is a root of unity of degree $k$. The proof is complete.

## 4. Connections with the set of special functions

Denote by $\Phi(x, w)$ the set of all functions $f$ satisfying the condition (*). In this set we introduce the following operation:

$$f^2(x+w) = f(x+w) * f(x+w) = f(x+2w). \tag{14}$$

By easy calculation from (14) and (*) follows that

$$f^2(x+w) = \frac{a' f(x) + b'}{c' f(x) + d'}, \tag{15}$$

where $a' = a^2 + bc$, $b' = b(a+d)$, $c' = c(a+d)$, $d' = d^2 + bc$.

By induction we obtain

$$f^n(x+w) = f(x+nw) = \frac{a_n f(x) + b_n}{c_n f(x) + d_n}. \tag{16}$$

From (16) follows that for every natural number $n$ we have $f^n(x+w) \in \Phi(x, w)$. Now suppose that $f, g \in \Phi(x, w)$ and let

$$f(x+w) = \frac{af(x)+b}{cf(x)+d}, \quad g(x+w) = \frac{rg(x)+s}{tg(x)+u}. \tag{17}$$

Then we define the addition operation as follows

$$f(x+w) \oplus g(x+w) = \frac{(a+r)(f+g)(x)+b+s}{(c+t)(f+g)(x)+d+u} = \frac{\widehat{a}(f+g)(x)+\widehat{b}}{\widehat{c}(f+g)(x)+\widehat{d}}. \tag{18}$$

Consider the set $\Phi(x, w)$ with the operations " $*$ "," $\oplus$" defined by (14) and (18). Moreover let $M_2(K)$ denote the set of all $2 \times 2$ matrices with entries belonging to $K$. Then we observe that the mapping $\Phi : \Phi(x, w) \to M_2(K)$ defined by

$$\Phi(f(x+w)) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A\hat{\imath}M_2(K) \tag{19}$$

is an isomorphism. Therefore as consequence we obtain the following:

**Corollary 1.** *Fermat's equation (1) in the set of $M_2(K)$ is equivalent to the following equations:*

$$f^n(x+w) \oplus g^n(x+w) = h^n(x+w)$$

*in the set $\Phi(x, w)$.*

Now we can observe that if for some natural number $k$ is satisfied the condition

$$A^k = \begin{pmatrix} a^b & \\ c & d \end{pmatrix}^k = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix} = \gamma I,$$

where $\gamma \neq 0$ then the function $f^k \in \Phi(x, w)$ is a periodic function.

Really, the condition $A^k = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix}$, $\gamma \neq 0$ by isomorphism (19) is equivalent to

$$f^k(x+w) = \frac{a_k f(x) + b_k}{c_k f(x) + d_k} = \frac{\gamma f(x) + 0}{0 f(x) + \gamma} = \frac{\gamma f(x)}{\gamma} = f(x). \tag{20}$$

Hence the function $f^k(x+w) = f(x+kw) = f(x)$ is a periodic function.

Finally we observe that if $\Phi_3(x, w)$ denote the set of all functions $f \in \Phi(x, w)$ with at least three distinct zeros then the function $f^k \in \Phi_3(x, w)$ is a periodic if and only if $A^k = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix} = \gamma I$, $\gamma \neq 0$. Indeed it suffices to show that if $f^k \in \Phi(x, w)$ is a periodic then $A^k = \gamma I$, $\gamma \neq 0$. Using (20) and denoting by $y = f(x)$ we obtain

$$c_k y^2 + (d_k - a_k)y - b_k = 0. \tag{21}$$

The condition that the function $f$ has at least three distinct zeros implies that in (21) all coefficients must be equal to zero. Hence, $c_k = d_k - a_k = b_k = 0$ and by isomorphism we obtain

$$A^k = \begin{pmatrix} a_k & b_k \\ c_k & d_k \end{pmatrix} = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix}, \quad \gamma = a_k = d_k \neq 0.$$

From these considerations follows the following:

**Corollary 2.** *The functions $f^k, g^r, h^s \in \Phi(x, w)$ if and only if there exists $\alpha$, $\beta$, $\gamma$, $\alpha\beta\gamma \neq 0$ such that $A^k = \alpha I$, $B^r = \beta I$, $C^s = \gamma I$.*

In this case the Fermat's equation $f^{kn}(x+w) \oplus g^{rn}(x+w) = h^{sn}(x+w)$ is equivalent to the equation $A^{kn} + B^{rn} = C^{sn}$ and by Corollary 2 the last equation is equivalent to $\alpha^n + \beta^n = \gamma^n$.

## References


[1] K. Bialek and A. Grytczuk, *The equation of Fermat in $G_2(k)$ and $Q\left(\sqrt{k}\right)$*, Acta Acad. Pa. Agriensis, Sectio Mat. Eger XIII/11, (1987), 81-90.

[2] R.Z. Domiaty, *Solutions of $x^4 + y^4 = z^4$ in $2 \times 2$ integral matrices*, Amer. Math. Monthly, 73(1966), 631.

[3] A. Grytczuk, *On Fermat's equation in the set of integral $2 \times 2$ matrices*, Period. Math. Hungar., vol.30, 1(1995), 67-72.

[4] A. Khazanov, *Fermat's equation in matrices*, Serdica Math. J., 21(1995), 19-40.
[5] A.W. Kuzel, *Matchmatical improvisations*, Kiev, 1983 (in Russian).
[6] P. Ribenboim, *13 Lectures on Fermat Last Theorem*, Springer-Verlag, 1979.
[7] R.Taylor and A. Wiles, *Ring theoretic properties of certain Hecke algebras*, Annals of Math., 141(1995), 553-572.
[8] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem*, Annals of Math., 141(1995), 443-551.

INSTITUTE OF MATHEMATICS, T. KOTARBIŃSKI PEDAGOGICAL UNIVERSITY, 65-069 ZIELONA GÓRA, POLAND

# LAGRANGE-JACOBI AND SUNDMAN RELATIONS FOR A SUM OF HOMOGENEOUS POTENTIALS

VASILE MIOC AND CRISTINA STOICA

**Abstract.** One considers the $n$-body problem in an attractive field defined by a sume of homogeneous force functions. Starting from the equations of motion and the prime integrals, one obtains a relation between the momentum of inertia, the potential and the energy constant, the analogous of the Lagrange-Jacobi relation from the Newtonian case. One also obtains two relations between the momentum of inertia, the potential and the angular momentum, corresponding to the Sundman relations for the Newtonian case.

## 1. Introduction

Consider a system of $n$ interacting particles of masses $m_i > 0$, $i = \overline{1,n}$, in the Euclidean space $\mathbf{R}^3$, let $r_i = (x_i, y_i, z_i) \in \mathbf{R}^3$ be their position vectors, and let $r = (r_1, r_2, \ldots, r_n) \in \mathbf{R}^{3n}$ be the configuration of the system. Let the force field be defined by the superposition of a number $N$ (finite or infinite) of homogeneous potential functions

$$U : \mathbf{R}^{3n} \setminus \Delta \to [0, \infty), \quad U(r) = \Sigma_k U_k(r), \quad U_k(r) = \Sigma^* A_{k,ij} r_{ij}^{-\alpha_k}, \tag{1}$$

where $\alpha_k \geq 0$, $r_{ij} = |r_i - r_j|$ is the Euclidean distance between $i$-th and $j$-th particles, $\Delta$ stands for the collision set

$$\Delta = \bigcup_{1 \leq i < j \leq n} \{r | \, r_i = r_j\}$$

$A_{k,ij} : \mathbf{R}^2 \to [0, \infty)$ are symmetric positive functions of masses

$$A_{k,ij} = A_k(m_i, m_j) = A_k(m_j, m_i) = A_{k,ji}.$$

Received by the editors: November 26, 1996.

1991 *Mathematics Subject Classification.* 70F05

*Key words and phrases.* homogeneous potentials, Lagrange-Jacobi relations

and we use (throughout this paper) the abbreviating notation

$$\Sigma_k = \Sigma_{k=1}^{N}, \quad \Sigma_i = \Sigma_{i=1}^{n}, \quad \Sigma^* = \Sigma_{1\le i<j\le n}.$$

To be retained that the index $k$ refers to the potential in the sum (1), while the indices $i$ and $j$ refer to particles.

Such sums of potentials constitute a natural extension of homogeneous ($N = 1$) and quasihomogeneous ($N = 2$) potentials. The homogeneous models (from $\alpha_1 = 1$, the Newtonian case, up to $\alpha_1 = 6$ or 8, the Van der Waals case) are well known in physics, and their study by the methods of mechanics led to many interesting results (see [2, 5, 10]). The attraction described by central potentials with $N > 1$ also models various situations. For $N = 2$, $\alpha_1 = 1$, $\alpha_2 = 2$, we have Maneff's field [3, 4, 6, 8, 9] or Maneff-type fields [1]. For $N = 2$, $\alpha_1 = 1$, $\alpha_2 = 3$, we recover Schwarzschild's field. Fock's field (e.g. [7]) is recovered for $N = 4$, $\alpha_k = k$. The case $\alpha_{2s+1} \neq 0$, $\alpha_{2s} = 0$ models the motions in the equatorial plane of a body which generates a field featured by zonal harmonics; and so forth.

In this paper we shall extend some results (known for the homogeneous case; see [2, 11]) concerning the $n$-body problem to such sums of potentials. We shall establish a relation analogous to that of Lagrange-Jacobi, and an inequality (in two variants) corresponding to Sundman's one.

## 2. Equations of motion and first integrals

The equations of motion have the form

$$m_i \ddot{r}_i = \partial U/\partial r_i = -\Sigma^*(r_i - r_j)\Sigma_k \alpha_k A_{k,ij} r_{ij}^{-\alpha_k - 2} \tag{2}$$

Standard results of the theory of differential equations ensure, for given initial conditions $(r, \dot{r})(0)$, the existence and uniqueness of an analytic solution of equations (2), defined on an interval $(t^-, t^+)$, $t^- < 0 < t^+$. Equations (2) being time-reversible, we may confine our study to $[0, t^+)$. The solution can be analytically extended to a maximal interval $[0, t^*)$, $0, t^+ \le t^* \le \infty$; it is regular for $t^* = \infty$, and encounters a singularity else.

It is easy to establish that equations (2) admit ten first integrals: those of mass centre, those of angular momentum, and that of energy, respectively

$$\Sigma_i m_i \dot{r}_i = a, \quad \Sigma_i m_i r_i - t\Sigma_i m_i \dot{r}_i = b, \quad a, b \in \mathbf{R}^3.$$

$$\Sigma_i m_i r_i \times \dot{r}_i = C = constant, \quad C \in \mathbf{R}^3; \tag{4}$$

$$T - U = h = constant, \quad h \in \mathbf{R}, \tag{5}$$

where the kinetic energy of the system, $T : \mathbf{R}^{3n} \to [0, \infty)$, has the expression

$$T(\dot{r}) = \Sigma_i m_i \dot{r}_i^2/2. \tag{6}$$

(We used the notation $u^2 = |u|^2$, $u \in \mathbf{R}^3$.) Fixing the origin of the coordinates in the mass centre, equations (2), (4), (5) keep their form, but (3) become

$$\Sigma_i m_i r_i = 0, \quad \Sigma_i m_i \dot{r}_i = 0. \tag{7}$$

## 3. Lagrange-Jacobi relation

Let $J : \mathbf{R}^{3n} \to [0, \infty)$, defined by

$$2J(r) = \Sigma_i m_i r_i^2, \tag{8}$$

be the moment of inertia. We can state

**Proposition 1.** *For potentials of the form (1) the following equality holds:*

$$\ddot{J} = \Sigma_k (2 - \alpha_k) U_k + 2h. \tag{9}$$

*Proof.* Differentiating (8) twice with respect to time, we get successively

$$\dot{J} = \Sigma_i m_i r_i \cdot \dot{r}_i, \tag{10}$$

$$\ddot{J} = \Sigma_i m_i (\dot{r}_i^2 + r_i \cdot \ddot{r}_i). \tag{11}$$

By (1), it is easy to show that

$$\Sigma_i r_i (\partial U . \partial r_i) = -\Sigma_k \alpha_k U_k$$

or, taking into account (2),

$$\Sigma_i m_i r_i \cdot \ddot{r}_i = -\Sigma_k \alpha_k U_k. \tag{12}$$

On the other hand, (5) and (6) lead to

$$\Sigma_i m_i \dot{r}_i^2 = 2\Sigma_k U_k + 2h. \tag{13}$$

Finally, adding together (12) and (13), and replacing the resulting expression in (11), we obtain the relation (9). □

This equality constitutes the analogous of the Lagrange-Jacobi relation (known for the Newtonian potential).

## 4. Sundman's inequalities

These inequalities connect the potential, the moment of inertia, and the angular momentum. We can state

**Proposition 2.** *For potentials of the form (1), the following inequality holds:*

$$C^2 \leq 2J(\ddot{J} + \Sigma_k \alpha_k U_k). \tag{14}$$

*Proof.* By (4), using $|\Sigma p| \leq \Sigma |p|$, we have

$$|C| \leq \Sigma_i m_i |r_i \times \dot{r}_i|, \tag{15}$$

which, using $|u \times v| \leq |u||v|$ and squaring, leads to

$$C^2 \leq \left[\Sigma_i \left(\sqrt{m_i}|r_i|\right)\left(\sqrt{m_i}|\dot{r}_i|\right)\right]^2.$$

But

$$(\Sigma pq)^2 \leq (\Sigma p^2)(\Sigma q^2), \tag{16}$$

hence we can write

$$C^2 \leq (\Sigma_i m_i r_i^2)(\Sigma_i m_i \dot{r}_i^2),$$

which, taking into account (6) and (8), becomes

$$C^2 \leq 4JT. \tag{17}$$

By (5) and (9), one obtains easily

$$2T = \ddot{J} + \Sigma_k \alpha_k U_k, \tag{18}$$

which, replaced in (17), leads to the inequality (14). □

This result can be refined, in the form of

**Proposition 3.** *For potentials of the form (2), an inequality stronger than (14) holds:*

$$C^2 \leq 2J(\ddot{J} + \Sigma_k \alpha_k U_k) - \dot{J}^2. \tag{19}$$

*Proof.* By (10), using $|\Sigma p| \leq \Sigma |p|$, we have

$$|\dot{J}| \leq \Sigma_i m_i |r_i||\dot{r}_i|.$$

Squaring and using (16), we obtain

$$\dot{J}^2 \leq (\Sigma_i m_i r_i^2)(\Sigma_i m_i \dot{r}_i^2),$$

which, taking into account (8), can be written in the form

$$\dot{J}^2 \leq 2J\Sigma_i m_i (r_i \cdot \dot{r}_i)^2 / r_i^2. \tag{20}$$

Now, writing (15) under the form

$$|C| \leq \Sigma_i \left(\sqrt{m_i}|r_i|\right)\left(\sqrt{m_i}|r_i \times \dot{r}_i|/|r_i|\right),$$

using again (16), then squaring and using (8), we get

$$C^2 \leq 2J\Sigma_i m_i |r_i \times \dot{r}_i|^2 / r_i^2. \tag{21}$$

Adding together (20) and (21), then taking into consideration the well-known relation $|u \times v|^2 + |u \cdot v|^2 = |u|^2|v|^2$, and finally using (6), we obtain easily

$$C^2 + \dot{J}^2 \leq 4JT. \tag{22}$$

The substitution of $T$ from (18) in (22) yields the inequality (19). □

The inequalities (14) and (19) are analogous to those established by Sundman for the Newtonian potential.

## References


[1] Delgado, J., Diacu, F.N., Lacomba, E.A., Mingarelli, A., Mioc, V., Perez, E., Stoica, C., *The Global Flow of the Manev Problem*, J. Math. Phys., 37(1996) 2748-2761.

[2] Diacu, F.N., *Total Collapse Dynamics for Particle Systems*, Libertas Math., 10(1990), 161-170.

[3] Diacu, F.N., *The Planar Isosceles Problem for Maneff's Gravitational Law*, J. Math. Phys., 34(1993), 5671-5690.

[4] Diacu, F.N., Mingarelli, A., Mioc, V., Stoica, C., *The Manev Two-Body Problem: Quantitative and Qualitative Theory*, in R.P. Agarwal (ed.), *Dynamical Systems and Applications*, World Sci. Ser. Appl. Anal., 4, World Scientific Publ. Co., Singapore, 1995, 213-227.

[5] McGehee, R., *Double Collisions for a Classical Particle System with Nongravitational Interactions*, Comment. Math. Helvetici, 56(1981), 524-557.

[6] Maneff, G., *La gravitation et l'énergie au zéro*, C.R. Acad. Sci. Paris, 190(1930), 1374-1377.

[7] Mioc, V., *Elliptic-Type Motion in Fock's Gravitational Field*, Astron. Nachr., 315(1994), 175-180.

[8] Mioc, V., Stoica, C., *Discussion et résolution complète du problème des deux corps dans le champs gravitationnel de Maneff*, C.R. Acad. Sci. Paris, 320(1995), ser.I, 645-648.

[9] Mioc, V., Stoica, C., *Discussion et résolution complète du problème des deux corps dans le champs gravitationnel de Maneff* (II), C.R. Acad. Sci. Paris, 321(1995), ser.I, 961-964.

[10] Moser, J., *Three Integrable Hamiltonian Systems Connected with Isospectral Deformation*, Adv. Math., 16(1975), 197-220.

[11] Wintner, A., *The Analytical Foundations of Celestial Mechanics*, Princeton Univ. Press, Princeton, 1941.

ASTRONOMICAL INSTITUTE OF THE ROMANIAN ACADEMY, ASTRONOMICAL OBSERVATORY CLUJ-NAPOCA, 3400 CLUJ-NAPOCA, ROMANIA

INSTITUTE FOR GRAVITATION AND SPACE SCIENCES, LABORATORY FOR GRAVITATION, 71111 BUCHAREST, ROMANIA

# ABOUT AN INTEGRAL OPERATOR PRESERVING THE UNIVALENCE

VIRGIL PESCAR

**Abstract.** In this work an integral operator is studied and the author determines conditions for the univalence of this integral operator.

## 1. Introduction

Let A be the class of the functions f which are analytic in the unit disc $U = \{z \in C; |z| < 1\}$ and $f(0) = f'(0) - 1 = 0$.

We denote by S the class of the function $f \in A$ which are analytic in U.

Many authors studied the problem of integral operators which preserve the class S. In this sense an important result is due to J. Pfaltzgraff [4].

**Theorem A.** *[4] If $f(z)$ is univalent in $U$, $\alpha$ a complex number and $|\alpha| \leq \frac{1}{4}$, then the function*

$$G_\alpha(z) = \int_0^z [f'(\xi)]^\alpha d\xi \tag{1}$$

*is univalent in $U$.*

**Theorem B.** *[3] If the function $g \in S$ and $\alpha$ is a complex number, $|\alpha| \leq \frac{1}{4n}$, then the function defined by*

$$G_{\alpha,n}(z) = \int_0^z [g'(u^n)]^\alpha du \tag{2}$$

*is univalent in $U$ for all positive integer $n$.*

## 2. Preliminaries

For proving our main result we will need the following theorem and lemma.

---

1991 *Mathematics Subject Classification.* 30E20.

*Key words and phrases.* univalence, integral operators.

**Theorem C.** *[1] If the function $f$ is regular in the unit disc $U$, $f(z) = z + a_2z^2 + \dots$ and*

$$(1 - |z|^2)\left|\frac{zf''(z)}{f'(z)}\right| \le 1 \tag{3}$$

*for all $z \in U$, then the function $f$ is univalent in $U$.*

**Lemma Schwarz.** *[2] If the function $g$ is regular in $U$, $g(0) = 0$ and $|g(z)| \le 1$ for all $z \in U$, then the following inequalities hold*

$$|g(z)| \le |z| \tag{4}$$

*for all $z \in U$, and $|g'(0)| \le 1$, the equalities (in inequality (4) for $z \neq 0$) hold only in the case $g(z) = \epsilon z$, where $|\epsilon| = 1$.*

## 3. Main result

**Theorem 1.** *Let $\gamma$ be a complex number and the function $g \in A$, $g(z) = z + a_2z^2 + \dots$. If*

$$\left|\frac{h''(z)}{g'(z)}\right| \le \frac{1}{n} \tag{5}$$

*for all $z \in U$ and*

$$|\gamma| \le \frac{1}{\left(\frac{n}{n+2}\right)^{\frac{n}{2}} \frac{2}{n+2}} \tag{6}$$

*then the function*

$$G_{\gamma,n}(z) = \int_0^z [g'(u^n)]^\gamma \, du \tag{7}$$

*is univalent in $U$ for all $n \in N^* - \{1\}$.*

*Proof.* Let us consider the function

$$f(z) = \int_0^z [g'(u^n)]^\gamma \, du. \tag{8}$$

The function

$$h(z) = \frac{1}{\gamma}\frac{f''(z)}{f'(z)}, \tag{9}$$

where the constant $\gamma$ satisfies the inequality (6) is regular in U. From (9) and (8) it follows that

$$h(z) = \frac{\gamma}{|\gamma|}\left[\frac{nz^{n-1}g''(z^n)}{g'(z^n)}\right]. \tag{10}$$

Using (10) and (5) we have

$$|h(z)| \leq 1, \tag{11}$$

for all $z \in U$. From (10) we obtain $h(0) = 0$ and applying Schwarz-Lemma we have

$$\frac{1}{|\gamma|}\left|\frac{f''(z)}{f'(z)}\right| \leq |z|^{n-1} \leq |z| \tag{12}$$

for all $z \in U$, and hence, we obtain

$$\left(1-|z|^2\right)\left|\frac{zf''(z)}{f'(z)}\right| \leq |\gamma|\left(1-|z|^2\right)|z|^n. \tag{13}$$

Let us consider the function $Q:[0,1] \to R$, $Q(x) = \left(1-x^2\right)x^n$; $x = |z|, z \in U$, which has a maximum at a point $x = \sqrt{\frac{n}{n+2}}$, and hence

$$Q(x) < \left(\frac{n}{(n+2)^{\frac{n}{2}}}\right)\frac{2}{n+2} \tag{14}$$

for all $x \in (0,1)$. Using this result and (13) we have

$$\left(1-|z|^2\right)\left|\frac{zf''(z)}{f'(z)}\right| \leq |\gamma|\left(\frac{n}{(n+2)}\right)^{\frac{n}{2}}\frac{2}{n+2}. \tag{15}$$

From (15) and (6) we obtain

$$\left(1-|z|^2\right)\left|\frac{zf''(z)}{f'(z)}\right| \leq 1 \tag{16}$$

for all $z \in U$. From (16) and (8) and Theorem C it follows that $G_{\gamma,n}$, is in the class S. □

*Remark.* For $n = 2$, we obtain $|\gamma| \leq 4$ and the function $G_{\gamma,2}$ is in the class S.

## References


[1] J. Becker, Löwnersche Differentialgleichung und quasikonform fortsetzbare schlichte Funcktionen, J.Reine Angew. Math., 255(1972),23-43.

[2] G.M. Goluzin, Gheometriceskaia teoria funktii Kompleksnogo peremennogo, ed. a II-a, Nauka, Moscova, 1966.

[3] N.N. Pascu, V. Pescar, On the integral operators of Kim-Merkes and Pfaltzgraff, Studia (Mathematica), Univ. Babeş-Bolyai, Cluj-Napoca, 32, 2(1990), 185-192.

[4] J. Pfaltzgraff, Univalence of the integral $\int_0^z [f'(t)]^c\,dt$, Bull. London Math. Soc. 7(1975), No. 3, 254-256.


"Transilvania" University of Braşov Faculty of Science Department of Mathematics 2200 Braşov ROMANIA

STUDIA UNIV. "BABEŞ-BOLYAI", MATHEMATICA, Volume XLII, Number 4, December 1997

# A STABILIZED APPROACH FOR THE CHEBYSHEV-TAU METHOD

IULIU SORIN POP

**Abstract.** We consider a different approach for the Chebyshev-tau spectral method by a modification of the basis for the test function space. This leads to sparse matrices, which are better conditioned than those generated by the usual method, as being pointed out by some numerical examples.

## 1. Introduction

Spectral methods have been studied intensively in the last two decades because of their good approximation properties. This advantage was shadowed by some difficulties generated by this discretization. Thus, the matrices which arise in the spectral discretization of differential equations are generally full and their condition number increases strongly with the number of shape-functions. Therefore, it is quite difficult to get efficient iterative solvers, mainly for the Galerkin or the tau variant of these methods. Moreover, especially for fourth order problems, stability and numerical accuracy of the computation can be strongly affected when a discretization using a large number of shape-functions is applied, and the theoretical accuracy of these methods can be lost. There are several works concerned with the problems mentioned above in any of the three existing types of spectral methods (see, for example [4], [8] for the tau method, [7], [9], [10], [11], [12] for the collocation variant or [15], [16] for the Galerkin approach – all of these cited only in conjunction with Chebyshev polynomials).

The type of the spectral method is dictated by the application. For example, collocation methods are suited to nonlinear problems or complicated coefficients, while Galerkin ones have the advantage of a more convenient analysis and optimal error estimates. The tau method can be appropriate in the case of complicated (nonlinear)

Received by the editors: September 3,1996.

1991 *Mathematics Subject Classification.* 65F10.

*Key words and phrases.* spectral methods, sparse matrices, Chebyshev polynomials.

boundary conditions, where a Galerkin approach would be impossible and the collocation extremely tedious.

Our work is focused on the tau spectral method using Chebyshev polynomials. We try to present a slightly different approach by a modification of the test-function basis. This leads to better conditioned matrices which, in case of linear equations having constant coefficients, are also sparse (banded). These features are exemplified on some model problems, where the applicability of the Bi-CGSTAB [17] algorithm is studied. The paper is organized as follows: In section 2, some basic properties of Chebyshev polynomials are provided. The following section deals with the convergence of the Chebyshev-tau method. Next, the new approach for the tau method is presented, together with some details regarding the discretization matrices. Finally, we give some numerical examples.

## 2. Chebyshev polynomials

In the following we will denote by $\mathcal{L}^2_\omega(-1,1)$, $\mathcal{H}^k_\omega(-1,1)$, $\mathcal{H}^k_{\omega,0}(-1,1)$, $(\cdot,\cdot)_{k,\omega}$, $\|\cdot\|_{k,\omega}$, $|\cdot|_{k,\omega}$ the corresponding weighted Sobolev spaces, scalar products, norms and seminorms on $(-1,1)$, where $\omega(x) = \frac{1}{\sqrt{(1-x^2)}}$ is the Chebyshev weight. Let $\mathbb{P}_N$ be the space of (real) polynomials of maximal order $N$ and

$$T_k(x) = \cos(k \arccos(x)),\ k \in \mathbb{N}$$

be the $k^{th}$ order Chebyshev polynomial of the first kind. The following properties can be found for example in [5] or [6]

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x),\ k > 0, \tag{2.1}$$

$$T_k(x)T_p(x) = \frac{1}{2}(T_{k+p} + T_{|k-p|}), \tag{2.2}$$

$$T_k(\pm 1) = (\pm 1)^k,\ T_k'(\pm 1) = (\pm 1)^k k^2, \tag{2.3}$$

$$(T_n, T_m)_{0,\omega} = \frac{\pi}{2} c_n \delta_{n,m}, \tag{2.4}$$

where $\delta_{n,m}$ represents the Kronecker symbol and $c_i := \begin{cases} 2, & \text{if } i = 0 \\ 1, & \text{if } i > 0 \end{cases}$. The properties above are the starting point for the development of Chebyshev spectral methods for differential equations. The idea is to approximate the unknown function by a Chebyshev series and to make use of the differentiation rule which can be deduced from (2.1). Thus, if

$$u(x)=\sum_{k=0}^{\infty} a_k T_k(x), \tag{2.5}$$

its derivative can be expressed in the form ([5])

$$u'(x)=\sum_{k=0}^{\infty} a_k^{(1)} T_k(x), \tag{2.6}$$

where

$$a_k^{(1)}=\frac{2}{c_k}\sum_{p=k+1,\, p+k-odd}^{\infty} p a_p. \tag{2.7}$$

Similar relations can be deduced for other operators. After a spectral discretization is done, one has to project the initial spaces onto finite dimensional ones. Therefore, one would consider a truncated series up to an order $N$ in (2.5) and (2.6) in the Galerkin or tau case, or an interpolant of the same order for the collocative methods. Then the result will be projected on a finite dimensional space in order to get a finite system, where the in the collocative approach Dirac distributions can be considered as the "test functions".

## 3. The Chebyshev-tau method

In this section the tau variant is described generally and the convergence of the Chebyshev-tau method for a $4^{th}$ order Dirichlet boundary value model-problem is given. There are several methods for proving the convergence of the Chebyshev-tau method (see, for example [4] or [14]) but we will respect the approach in [5], chapter 10.

Let us consider the following examples

$$\begin{cases} L_1 u \equiv -u''(x)+\lambda^2 u(x)=f_1(x),\ x\in[-1,1], \\ u(\pm 1)=0, \end{cases} \tag{3.1}$$

respectively

$$\begin{cases} L_2 u \equiv u^{(IV)}(x)+\lambda^2 u(x)=f_2(x),\ x\in[-1,1], \\ u(\pm 1)=u'(\pm 1)=0. \end{cases} \tag{3.2}$$

Both problems can be treated in a similar manner, therefore we will consider a unified approach. Let us denote them by $(\mathcal{P}_k)$, where $k=1$ in the first case and $k=2$ in the second one. In the following the index $k$ appearing in any of the relation below should be replaced with the corresponding value for the problems above. In both case the existence and uniqueness of a (variational) solution can be obtained for any $f$ in $\mathcal{H}^{-k}(-1,1)$ (the dual space of $\mathcal{H}^k_{\omega,0}(-1,1)$). For working in weighted Sobolev spaces the following bilinear forms should be considered

$$a_k : \mathcal{H}^k_\omega(-1,1) \times \mathcal{H}^k_{\omega,0}(-1,1) \longrightarrow \mathbb{R}, \ a_k(u,v) = \int_{-1}^{1} u^{(k)}(v\omega)^{(k)}\, dx + \int_{-1}^{1} \lambda^2 uv\omega\, dx, \ k = \tag{3.3}$$

Then, for any $f$ in $\mathcal{H}^{-k}_\omega(-1,1)$ the problem $(\mathcal{P}_k)$ is equivalent to the following variational one

$$\begin{cases} u_k \in \mathcal{H}^k_{\omega,0}(-1,1), \\ \forall v \in \mathcal{H}^k_{\omega,0}(-1,1),\ a_k(u_k,v) = (f,v)_{0,\omega}. \end{cases} \tag{3.4}$$

Continuity and ellipticity for $a_k$ are proven, e.g., in [5], chapter 11 ($k=1$) and [3], lemma III.1 ($k=2$). Therefore, existence and uniqueness for the solution $u_k$ is assured in both cases and one gets

$$\|u_k\|_{k,\omega} \leq \|f\|_{-k,\omega}. \tag{3.5}$$

The corresponding discrete problems obtained by a Chebyshev-tau discretization, assuming $f \in \mathcal{L}^2_\omega(-1,1)$, are

$$(\mathcal{P}_{k,N}) \quad \begin{cases} u_{k,N} \in \mathbb{P}_N \cap \mathcal{H}^k_{\omega,0}(-1,1), \\ \forall v \in \mathbb{P}_{N-2k},\ a_k(u_k,v) = (f,v)_{0,\omega}, \end{cases} \tag{3.5.a}$$

or their equivalent strong form

$$(\mathcal{P}_{k,N}) \quad \begin{cases} u_{k,N} \in \mathbb{P}_N, \\ (L_k u_k, v)_{0,\omega} = (f,v)_{0,\omega},\ \forall v \in \mathbb{P}_{N-2k}, \\ u^{(l)}_{k,N}(\pm 1) = 0,\ l = \overline{0,k-1}. \end{cases} \tag{3.5.b}$$

**Theorem 3.1.** *For any $f \in \mathcal{L}^2_\omega(-1,1)$, the problem $(\mathcal{P}_{k,N})$ has a unique solution $u_{k,N} \in \mathbb{P}_N \cap \mathcal{H}^{2k}_{\omega,0}(-1,1)$ which converges to the solution $u_k$ of the problem $(\mathcal{P}_k)$ as $N$ tends to infinity. Moreover, if $u_k \in \mathcal{H}^\alpha_\omega(-1,1)$ with $\alpha \geq 2k$, then the error is bounded by*

$$\|u_k - u_{k,N}\|_{2k,\omega} \leq CN^{2k-\alpha}\|u\|_{\alpha,\omega}. \tag{3.6}$$

*Proof.* This result is obtained for $(\mathcal{P}_1)$ in [5], chapter 10. A similar approach can be considered for $(\mathcal{P}_2)$. The continuous "inf-sup" condition ([1], theorem 5.2.1) is fulfilled for $L_2$ because of the ellipticity and boundedness of $a_2$. In the discrete case, for any $u_N \in \mathbb{P}_N$, $u_N^{(IV)} \in \mathbb{P}_{N-4}$ can be taken as test function. Notice that, if $u_N \in \mathbb{P}_N \cap \mathcal{H}^2_{\omega,0}(-1,1)$ and

$u_N^{(IV)} \equiv 0$, then $u_N \equiv 0$, hence in this case the condition is trivially satisfied. Therefore we obtain

$$\sup_{v\neq 0, v\in \mathbb{P}_{N-4}} \frac{(L_2 u_N, v)_{0,\omega}}{\|v\|_{0,\omega}} \geq \frac{(L_2 u_N, u_N^{(IV)})_{0,\omega}}{\|u_N^{(IV)}\|_{0,\omega}} = \frac{|u_N|_{4,\omega}^2 + \lambda^2 \int_{-1}^{1} (u_N \omega)'' u_N''}{|u_N|_{4,\omega}} \geq$$

$$\geq \frac{|u_N|_{4,\omega}^2 + C\lambda^2 |u_N|_{2,\omega}}{|u_N|_{4,\omega}} \geq |u_N|_{4,\omega}.$$

Now, applying Poincaré's inequality (Appendix of [5]) successively, we get

$$|u_N|_{4,\omega} \geq C\|u_N\|_{4,\omega}$$

and by theorem 6.2.1 in [1] the first part of the proof is shown. In order to get the error estimate we have to remind that the discrete stability condition yields

$$\|u_2 - u_{2,N}\|_{4,\omega} \leq C \inf_{p\in \mathbb{P}_N \cap \mathcal{H}^2_{\omega,0}(-1,1)} \|u - p\|_{4,\omega}. \tag{3.7}$$

Theorem 4.1 in [13] provides the existence of an operator $\Pi_{4,N}^{2,0} : \mathcal{H}^\alpha_\omega(-1,1) \cap \mathcal{H}^2_{\omega,0}(-1,1) \to \mathbb{P}_N \cap \mathcal{H}^2_{\omega,0}(-1,1)$ such that for any $u \in \mathcal{H}^\alpha_\omega(-1,1) \cap \mathcal{H}^2_{\omega,0}(-1,1)$ the following holds

$$\forall \beta,\ 0 \leq \beta \leq 4,\ \|u - \Pi_{4,N}^{2,0} u\|_{\beta,\omega} \leq C N^{\beta-\alpha} \|u\|_{\alpha,\omega}. \tag{3.8}$$

The last two relations with $\beta = 4$ in (3.8) leads to the desired estimate. □

## 4. A different approach

As pointed out in the introduction, we will suggest a different approach for the tau method. The basic idea is to preserve the spaces in which this method was formulated originally, while considering a different basis for the projection space (test functions). We will restrict ourselves to ordinary differential equations, but these ideas can be applied also in the multidimensional case. Let us define the functions ($k \in \mathbb{N}$)

$$\Phi_k^{(0)}(x) := \frac{2}{\pi} T_k(x),\ \ \Phi_{k+i}^{(i)}(x) := \frac{\Phi_{k+i-1}^{(i-1)}(x) - \Phi_{k+i+1}^{(i-1)}(x)}{2(k+i)},\ i > 0 \tag{4.1}$$

and

$$\tilde{\Phi}_k^{(0)}(x) := d_k \Phi_k^{(0)}(x),\ \ \tilde{\Phi}_{k+i}^{(i)}(x) := \frac{\tilde{\Phi}_{k+i-1}^{(i-1)}(x) - \tilde{\Phi}_{k+i+1}^{(i-1)}(x)}{2(k+i)},\ i > 0, \tag{4.2}$$

where $c_k$ was given in (2.4), $d_k := d_k(N,\beta) := \begin{cases} 1, & \text{if } k = \overline{0, N-\beta} \\ 0, & \text{otherwise} \end{cases}$ and $\beta$ stands for the order of the differential operator. The following lemma justifies the choice of $\left\{\tilde{\Phi}_{k+i}^{(i)}, k = \overline{0, N-\beta}\right\}$ as test function basis for the Chebyshev-tau method.

**Lemma 4.1.** *For any $i \in \mathbb{N}$ the following relations hold*

*a)* $\tilde{\Phi}_{k+i}^{(i)} \equiv 0 \ \forall k > N - \beta$;

*b)* $span \left\{ \tilde{\Phi}_{k+i}^{(i)}, k = \overline{0, N-\beta} \right\} = \mathbb{P}_{N-\beta}$.

*Proof.* The case $i = 0$ is obvious. Then, both $a)$ and $b)$ can be proven by mathematical induction after $i$. □

*Remark 4.1.* If the Chebyshev-tau discretization matrices have already been constructed, then testing with the functions described above is similar to an algebraic transformation of the resulting system. This transformation can be described in an iterative way and it refers only to the part of the system corresponding to the differential operator. Let us assume that we have written first the equations for the boundary conditions, and then those resulting form the test with $T_k$, $k = \overline{0, N-\beta}$. At step $j$, $1 \leq j \leq i$, one has to subtract the equation number $k + \beta + 2$ from $k + \beta$ and divide the result with $2(k+j)$. When $k + 2 \geq N - \beta$ only the division should be performed. The resulting system will be identical to the one which arises when $\left\{ \tilde{\Phi}_{k+i}^{(i)}, k = \overline{0, N-\beta} \right\}$ are considered as test functions. However, this does not take advantage of the "sparsity potential" of this approach and the reasons will be seen below.

**Lemma 4.2.** *Let $u(x) = \sum_{k=0}^{\infty} a_k T_k(x)$. Then the following relations are satisfied ($\forall k \in \mathbb{N}$)*

$$(u, \Phi_k^{(0)})_{0,\omega} = c_k a_k, \ (f, \Phi_{k+1}^{(1)})_{0,\omega} = \frac{1}{(k+1)}[c_k a_k - a_{k+2}]. \tag{4.3a}$$

$$(u, \Phi_{k+2}^{(2)})_{0,\omega} = \frac{c_k a_k}{4(k+1)(k+2)} - \frac{a_{k+2}}{2(k+1)(k+3)} + \frac{a_{k+4}}{4(k+2)(k+3)}, \tag{4.3b}$$

$$\begin{aligned}(u, \Phi_{k+3}^{(3)})_{0,\omega} = &\frac{c_k a_k}{8(k+1)(k+2(k+3)} - \frac{3a_{k+2}}{8(k+1)(k+3)(k+4)} + \\ &+ \frac{3a_{k+4}}{8(k+2)(k+3)(k+5)} - \frac{a_{k+6}}{8(k+3)(k+4)(k+5)},\end{aligned} \tag{4.3c}$$

$$\begin{aligned}(u, \Phi_{k+4}^{(4)})_{0,\omega} = &\frac{c_k a_k}{16(k+1)(k+2(k+3)(k+4)} - \frac{4a_{k+2}}{16(k+1)(k+3)(k+4)(k+5)} + \\ &+ \frac{6a_{k+4}}{16(k+2)(k+3)(k+5)(k+6)} - \frac{4a_{k+6}}{16(k+3)(k+4)(k+5)(k+7)} + \\ &+ \frac{a_{k+8}}{16(k+4)(k+5)(k+6)(k+7)}.\end{aligned} \tag{4.3d}$$

*Proof.* The above relations can be obtained by a direct computation using the definition of $\Phi_{k+i}^{(i)}$, $i = \overline{0,4}$. □

*Remark 4.2.* The relations in (4.3a-d) also hold if $\left\{\Phi_{k+i}^{(i)}, k=\overline{0, N-\beta}\right\}$ are considered as test functions. In this case one has to replace $a_{k+j}$ by $d_{k+j}a_{k+j}$, but only when $j > 0$.

**Lemma 4.3.** *Let $u(x)=\sum_{k=0}^{\infty} a_k T_k(x)$. Then, for any $i > 0$, $k \in$ N*

$$(u', \Phi_{k+i}^{(i)})_{0,\omega} = (u, \Phi_{k+i}^{(i-1)})_{0,\omega}. \tag{4.4}$$

*Proo* We use again the mathematical induction after $i$. The case $i=0$ can be obtained directly from the relations in (2.5) – (2.7) and (4.1). Assuming that (4.4) holds for $j=\overline{0, i-1}$, (4.1) gives us

$$(u', \Phi_{k+i}^{(i)})_{0,\omega} = \frac{1}{2(k+i)}\left[(u', \Phi_{k+i-1}^{(i-1)})_{0,\omega} - (u', \Phi_{k+i+1}^{(i-1)})_{0,\omega}\right]. \tag{4.5}$$

Now, because of the assumption we have made, we get

$$(u', \Phi_{k+i-1}^{(i-1)})_{0,\omega} = (u, \Phi_{k+i-1}^{(i-2)})_{0,\omega},\ (u', \Phi_{k+i+1}^{(i-1)})_{0,\omega} = (u, \Phi_{k+i+1}^{(i-2)})_{0,\omega}, \tag{4.6}$$

and by (4.1)

$$\Phi_{k+i-1}^{(i-1)} = \frac{\Phi_{k+i-1}^{(i-2)} - \Phi_{k+i+1}^{(i-2)}}{2(k+i)}. \tag{4.7}$$

Putting together the relations in (4.5) – (4.7) we get the desired result. □

*Remark 4.3.* Lemma 4.2 remains true also in the case defined in (4.2).

*Remark 4.4.* Lemmas 4.1 and 4.2 give an iterative way to build the differentiation matrices for the desired differential operator. Lemma 4.2 also suggests that, in the case of a differential operator of order $\beta$, a reasonable choice for the test functions would be $\left\{\tilde{\Phi}_{k+i}^{(i)}, k=\overline{0, N-\beta}\right\}$. This reduces all the differentiation matrices to banded ones. The numerical examples are performed in this manner.

*Remark 4.5.* Similar differentiation matrices are obtained in the integral formulation of the Chebyshev-tau method ([4], [8]). The difference could appear in the case of nonconstant coefficient problems or nonlinear ones. But the approach proposed above has also a stabilization effect, in the sense that the discretization matrices have elements with a reduced order of magnitude. More, in comparison to the classical approach, the complexity of the computations involved in this discretization is decreased. The following lemma, which can be useful in the case of nonlinear problems (e.g. Burger's or the Navier Stokes equations), sustains the former statement.

**Lemma 4.4.** *Let* $v(x) = \sum_{k=0}^{\infty} a_k T_k(x)$ *and* $v(x) = \sum_{k=0}^{\infty} b_k T_k(x)$. *Then, for any* $k \in \mathbb{R}$ *we have*

$$(v'u, \Phi_{k+1}^{(1)})_{0,\omega} = \frac{1}{4(k+1)} \left\{ A_k a_0 + \sum_{j=1}^{\infty} \left[ 2(k+1+j) b_{|k+1-j|} + 2(k+1+j) b_{k+j+1} \right] a_j \right\}. \tag{4.8a}$$

$$(u'u, \Phi_{k+1}^{(1)})_{0,\omega} = \frac{1}{4(k+1)} \begin{cases} \sum_{p=1,\, p-odd}^{\infty} p a_0 a_p + 2\sum_{j=0}^{\infty} a_j a_{j+1}, & \textit{if } k=0, \\ a_0 a_1 + \sum_{j=1}^{k} 2(k+1-j) a_j a_{k+1-j} + \\ +2(k+1) \sum_{j=1}^{\infty} a_j a_{k+1+j}, & \textit{otherwise}, \end{cases} \tag{4.8b}$$

*where* $A_k := \begin{cases} 3b_1 + \sum_{p=3,\, p+k-odd}^{\infty} p b_p, & \textit{if } k=0, \\ b_1, & \textit{otherwise} \end{cases}$

*Proof.* Both relations can be obtained from (2.2), (2.6) and (4.4), but the calculus is quite tedious and therefore it is skipped. □

*Remark 4.6.* Similar features are obtained for other bases defined in (4.1) or (4.2).

It is worth to complete the approach with a treatment of the algebraic equations corresponding to the boundary conditions. By the method described above, the elements of the part of the discretization matrix corresponding to the differential operator are scaled to $O(1)$ (in fact the order of magnitude for the coefficients of the equation); this results from the discretization of the highest order derivative. Therefore, it is natural to modify the equations arising from the boundary conditions similarity. The simplest way is to divide any of the equations mentioned above by a number of order $O(N^{2k})$, where $k$ is the highest order of the derivatives appearing in the corresponding boundary condition. Although it seems trivial, using this trick a sensible improvement of the condition number of the discretization matrix can be obtained.

## 5. Numerical examples

In this part some results obtained with the Chebyshev-tau method in both approaches are given. All the computation are performed in double precision on an IBM-RS/6000 computer, and NAG routines are used to compute the necessary eigenvalues

At first, we compare the condition numbers of the resulting Chebyshev tau discretization matrices for the problems in (3.1) and (3.2) in the classical respectively the modified approach. This number plays a determinant role in the convergence behavior of iterative methods and represents an important source of roundoff errors. The results are presented in tables 5.1a and b for the problem in (3.1) and in tables 5.2a and b for the one in (3.2). For the first problem, the condition number is of order $O(N^4)$ in the classical tau method, while in the modified variant decreases to $O(N)$. Similar features are obtained for the second problem. In this case, the matrices generated by the classical method have a condition number proportional to $O(N^8)$, while in the modified approach the same characteristic has been reduced to $O(N^3)$.

The increased stability of the modified approach is shown through the results obtained for the examples mentioned before. Here $\mathbf{f}$ was taken such that $u(x) = \sin^2 \pi x$ is the exact solution. The discrete problems are solved using the (unpreconditioned) Bi-CGSTAB algorithm [17]. The stopping criterion is set to $\|\mathbf{r}^{(\mathbf{k})}\|_2/\|\mathbf{b}\|_2 \leq \epsilon$, where $\epsilon = 10^{-8}$ for the problem in (3.1) and $\epsilon = 10^{-6}$ for the one in (3.2) ($\mathbf{r}^{(\mathbf{k})}$ stands for the residual of the $k$-th iterate of the linear problem $\mathbf{A}\mathbf{u} = \mathbf{b}$; $\mathbf{r}^{(\mathbf{k})} = \mathbf{b} - \mathbf{A}\mathbf{u}^{(\mathbf{k})}$). From Tables 5.3a and b, a gain in accuracy can be observed in the modified version. We had no problems in achieving the stopping criterion in a moderate number of iterations, even in the cases when the classical tau method failed. In the tables '*' indicates divergence, or non-convergence in 250 steps. The failure of the method in the classical approach is due to roundoff errors, the algorithm being finite. In the modified approach, the number of iterations tends to remain stable with respect to the discretization order, while a significant increasal with $N$ can be seen in the classical method. Therefore, we can affirm that the proposed variant is more robust.

**Concluding remarks.** We have proposed a different approach for the Chebyshev tau method, which removes partially some of the inconvenients of the classical one. More sparsity and better conditioned matrices are provided, and therefore an improved stability and converging properties are obtained. The modification is easy to implement. However, a good preconditioner for this method is still to be found.

**Acknowledgments** This work was done at the Interdisciplinary Center for Scientific Computing (IWR) of the University of Heidelberg and was supported partially by the DAAD organization. The guidance of Prof.Dr.Dr.h.c. Willy Jäger is gratefully acknowledged. The author also expresses his gratitude to Prof. Dr. C. I. Gheorghiu for his suggestions. Many thanks to Dr. M. Hiegemann for the careful reading the manuscript and for his remarks.

Table 5.1a Problem (3.1), classical

| $N$ | $\lambda$ | $cond$ | $cond \cdot N$ |
|---|---|---|---|
| 8 | 0.0 | 575.17 | 0.876E-01 |
| | 10.0 | 310.21 | 0.472E-01 |
| 32 | 0.0 | 0.115E+06 | 0.971E-01 |
| | 10.0 | 0.630E+05 | 0.531E-01 |
| 128 | 0.0 | 0.273E+08 | 0.987E-01 |
| | 10.0 | 0.149E+08 | 0.540E-01 |
| 256 | 0.0 | 0.432E+09 | 0.990E-01 |
| | 10.0 | 0.236E+09 | 0.542E-01 |
| 512 | 0.0 | 0.686E+10 | 0.991E-01 |
| | 10.0 | 0.376E+10 | 0.542E-01 |

Table 5.1b The same, modified-tau

| $N$ | $\lambda$ | $cond$ | $cond \cdot N^{-1}$ |
|---|---|---|---|
| 8 | 0.0 | 7.64 | 0.849 |
| | 10.0 | 36.40 | 4.045 |
| 32 | 0.0 | 24.70 | 0.748 |
| | 10.0 | 73.77 | 2.235 |
| 128 | 0.0 | 92.62 | 0.717 |
| | 10.0 | 148.32 | 1.149 |
| 256 | 0.0 | 183.13 | 0.712 |
| | 10.0 | 209.82 | 0.816 |
| 512 | 0.0 | 364.15 | 0.709 |
| | 10.0 | 296.80 | 0.578 |

Table 5.2a Problem (3.2), classical-tau

| $N$ | $\lambda$ | $cond$ | $cond \cdot N^{-8}$ |
|---|---|---|---|
| 8 | 0.0 | 0.625E+05 | 0.145E-02 |
| | 10.0 | 0.564E+05 | 0.131E-02 |
| 32 | 0.0 | 0.326E+10 | 0.232E-02 |
| | 10.0 | 0.295E+10 | 0.209E-02 |
| 128 | 0.0 | 0.186E+15 | 0.242E-02 |
| | 10.0 | 0.168E+15 | 0.219E-02 |
| 256 | 0.0 | 0.464E+17 | 0.244E-02 |
| | 10.0 | 0.420E+17 | 0.220E-02 |
| 512 | 0.0 | 0.117E+20 | 0.244E-02 |
| | 10.0 | 0.106E+20 | 0.221E-02 |

Table 5.2b The same, modified-tau

| $N$ | $\lambda$ | $cond$ | $cond \cdot N^{-3}$ |
|---|---|---|---|
| 8 | 0.0 | 105.68 | 0.144 |
| | 10.0 | 31.96 | 0.438E-01 |
| 32 | 0.0 | 0.465E+03 | 0.129 |
| | 10.0 | 0.423E+03 | 0.456E-01 |
| 128 | 0.0 | 0.261E+06 | 0.121 |
| | 10.0 | 0.970E+05 | 0.451E-01 |
| 256 | 0.0 | 0.204E+07 | 0.119 |
| | 10.0 | 0.764E+06 | 0.450E-01 |
| 512 | 0.0 | 0.161E+08 | 0.119 |
| | 10.0 | 0.607E+07 | 0.449E-01 |

Table 5.3a : Number of iterations and absolute error for (3.1)

| | classical-tau | | | modified-tau | | |
|---|---|---|---|---|---|---|
| | $\lambda$ | | | $\lambda$ | | |
| $N$ | 0.0 | 10.0 | 100.0 | 0.0 | 10.0 | 100.0 |
| 8 | 9 $(0.381 \cdot 10^{-1})$ | 14 $(0.329 \cdot 10^{-1})$ | 18 $(0.200 \cdot 10^{-1})$ | 5 $(0.381 \cdot 10^{-1})$ | 10 $(0.329 \cdot 10^{-1})$ | 12 $(0.200 \cdot 10^{-1})$ |
| 16 | 20 $(0.347 \cdot 10^{-6})$ | 32 $(0.320 \cdot 10^{-6})$ | 100 $(0.761 \cdot 10^{-6})$ | 5 $(0.347 \cdot 10^{-6})$ | 11 $(0.322 \cdot 10^{-6})$ | 28 $(0.195 \cdot 10^{-6})$ |
| 32 | 63 $(0.106 \cdot 10^{-6})$ | 51 $(0.172 \cdot 10^{-6})$ | * (*) | 5 $(0.999 \cdot 10^{-15})$ | 11 $(0.166 \cdot 10^{-7})$ | 51 $(0.164 \cdot 10^{-4})$ |
| 64 | 196 $(0.409 \cdot 10^{-7})$ | 225 $(0.451 \cdot 10^{-6})$ | * (*) | 5 $(0.166 \cdot 10^{-14})$ | 11 $(0.237 \cdot 10^{-7})$ | 44 $(0.131 \cdot 10^{-4})$ |
| 128 | * (*) | * (*) | * (*) | 5 $(0.255 \cdot 10^{-14})$ | 11 $(0.301 \cdot 10^{-7})$ | 42 $(0.744 \cdot 10^{-5})$ |
| 256 | * (*) | * (*) | * (*) | 5 $(0.999 \cdot 10^{-15})$ | 11 $(0.308 \cdot 10^{-7})$ | 44 $(0.727 \cdot 10^{-5})$ |
| 512 | * (*) | * (*) | * (*) | 5 $(0.878 \cdot 10^{-14})$ | 11 $(0.333 \cdot 10^{-7})$ | 46 $(0.296 \cdot 10^{-4})$ |

Table 5.3b : Number of iterations and absolut error for (3.2)

| | classical-tau | | | modified-tau | | |
|---|---|---|---|---|---|---|
| | $\lambda$ | | | $\lambda$ | | |
| $N$ | 0.0 | 10.0 | 100.0 | 0.0 | 10.0 | 100.0 |
| 8 | 16 (0.801) | 18 (0.400) | 17 (0.103) | 5 (0.801) | 8 (0.400) | 12 (0.103) |
| 16 | 62 $(0.142 \cdot 10^{-3})$ | 71 $(0.494 \cdot 10^{-3})$ | 58 $(0.360 \cdot 10^{-4})$ | 6 $(0.183 \cdot 10^{-5})$ | 9 $(0.175 \cdot 10^{-5})$ | 13 $(0.484 \cdot 10^{-5})$ |
| 32 | * (*) | * (*) | * (*) | 6 $(0.322 \cdot 10^{-10})$ | 9 $(0.394 \cdot 10^{-7})$ | 17 $(0.851 \cdot 10^{-6})$ |
| 64 | * (*) | * (*) | * (*) | 6 $(0.189 \cdot 10^{-6})$ | 11 $(0.274 \cdot 10^{-8})$ | 16 $(0.332 \cdot 10^{-3})$ |
| 128 | * (*) | * (*) | * (*) | 6 $(0.189 \cdot 10^{-6})$ | 10 $(0.668 \cdot 10^{-5})$ | 12 $(0.442 \cdot 10^{-2})$ |
| 256 | * (*) | * (*) | * (*) | 8 $(0.384 \cdot 10^{-4})$ | 11 $(0.188 \cdot 10^{-5})$ | 13 $(0.952 \cdot 10^{-3})$ |
| 512 | * (*) | * (*) | * (*) | 10 $(0.205 \cdot 10^{-7})$ | 18 $(0.360 \cdot 10^{-7})$ | 13 $(0.316 \cdot 10^{-3})$ |

**References**


[1] Babuška, I.; Aziz, A.K.: Survey lectures on the mathematical foundations of the finite element method, in *The Mathematical Foundation of the Finite Element Method*, A. K Aziz (ed.), Academic Press, London, New York, 1972, 3–359.

[2] Bernardi, C.; Maday, Y.: Properties of some weighted Sobolev spaces and application to spectral approximations SIAM J. Numer. Anal. **26** (1989), 769–829.

[3] Bernardi, C.; Maday, Y.: Some spectral approximations of one dimensional fourth-order problems, in *Progress in*

*Approximation Theory*, P. Nevai and A. Pinkus (eds.), Academic Press, Boston, MA, 43-116.

[4] CABOS, CH.: A preconditioning of the tau operator for ordinary differential equations, ZAMM **74** (1994), 521-532.

[5] CANUTO, C.; HUSSAINI, M.Y.; QARTERONI, A.; ZANG, T.A.: *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, Heidelberg, New-York, 1988.

[6] FOX, L.; PARKER, I.B.: *Chebyshev Polynomials in Numerical Analysis*, Oxford University Press, London, 1968.

[7] FUNARO, D.; HEINRICHS, W.: Some results about the pseudospectral approximation of one-dimensional fourth-order problems, Numer. Math. **58** (1990), 399-418.

[8] HIEGEMANN, M.: Chebyshev matrix operator method for the solution of integrated forms of linear ordinary differential equations, Acta Mech. (to appear).

[9] HEINRICHS, W.: Improved condition number for spectral methods, Math. Comput. **53** (1989), 103-109.

[10] HEINRICHS, W.: Stabilization techniques for spectral methods, J. Sci. Comp. **6** (1991), 1-19.

[11] HEINRICHS, W.: A stabilized treatment of the biharmonic operator with spectral methods, SIAM J. Sci. Stat. Comput. **12** (1991), 1162-1172.

[12] HUANG, W.; SLOAN, D.M.: The pseudospectral method for solving differential eigenvalue problems, J. Comput. Phys. **111** (1994), 399-409.

[13] MADAY, Y: Analysis of spectral projectors in one-dimensional domains, Math. Comput. **55** (1990), 537-562.

[14] ROOS, H.G.; PFEIFFER, E.: A convergence result for the tau method, Computing **42** (1989), 81-84.

[15] SHEN, J.: Efficient spectral-Galerkin method II. Direct solvers of second and fourth order equations by using Chebyshev polynomials, SIAM J. Sci. Stat. Comput. **16** (1995), 74-87.

[16] SHEN, J.: Effcient Chebyshev-Legendre Galerkin methods for elliptic problems, Houston J. Math. (to appear).

[17] VAN DER VORST, H.A.: Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems, SIAM J. Sci. Stat. Comput. **13** (1992), 631-644.

"BABEŞ-BOLYAI" UNIVERSITY, FACULTY OF MATHEMATICS, M. KOGALNICEANU 1, 3400 CLUJ NAPOCA, ROMANIA

*E-mail address*: `ispop@cs.ubbcluj.ro`

STUDIA UNIV. "BABEŞ-BOLYAI", MATHEMATICA, Volume XLII, Number 4, December 1997

# HILBERT NUMBER OF AN ALGEBRAIC SURFACE

PINGXING SHENG

**Abstract.** In this article, we find t he maximum number of sheets for an algebraic equation of degree $n$ in space is bounded above by $\sum_{i=1}^{n} i^2 - n + 1$. Some impossibility of specific configurations is discussed. It solves the first part of Hilbert's 16th problem.

What is the maximum number of sheets for an algebraic equation of degree $n$ in space? What is the maximum number of the connected components of the complement of an algebraic hypersurface of degree $n$? How do self-intersection curves of an algebraic hypersurface behave? Can one classify all sheets of an algebraic hypersurface? What are the possible configurations of sheets? It turns out that above questions are interesting and very classic in algebraic geometry.

We consider the following algebraic equation $P(x, y, z) = 0$ with $deg(P) = n$. In first part of Hilbert's 16th problem, one discusses the maximum number of sheets for above equation. The definition and classification of sheets are important in order to understand the problem. Usually a sheet means a piece of regular surface of the algebraic surface $P(x, y, z) = 0$. The maximum number of connected components of the complement of an algebraic hypersurface has been discussed often to avoid the confusion in definition. A lot of beautiful results have been obtained by different mathematicians. I apologize first for not going to summarize all different results in this direction. Along this line, some results have been obtained by Gudkov, Rassias and others. One may consult the references for literature. In [6], Rassias discussed same problem for planar algebraic curves, and made a progress for first part of Hilbert's 16th problem.

For a given algebraic hypersurface $P(x, y, z) = 0$, there are sometimes many isolated surfaces. We are first interested in considering some very specific isolated surfaces which are submanifolds of $R^2$ with only infinity as boundary or without boundary. For

Received by the editors: October 15, 1996.
1991 *Mathematics Subject Classification.* 14A10, 58F22.
*Key words and phrases.* Hilbert 16th problem, algebraic surface.

example, $(x^2 + y^2 + z^2 - 1)(x + y + z - 100) = 0$ consists of a unit sphere and a plane. The unit sphere $x^2 + y^2 + z^2 = 1$ and the plane $x + y + z = 100$ are two isolated surfaces of the algebraic equation $(x^2 + y^2 + z^2 - 1)(x + y + z - 100) = 0$. These two surfaces do not intersect each other, so they are called isolated surface. The following results is trivial according to the factorization of polynomials and the fact that each surface of a given algebraic hypersurface has at least degree one in its algebraic representation.

**Proposition 1.** *The maximum number of isolated sheets (surface) for an algebraic equation of degree $n$ is bounded above by $n$.*

If considering regular sheets, one is interested in knowing configurations of sheets when the maximum number reaches. Even in the plane, another still open problem is to find all possible configurations of closed branches of a plane curve of degree $n$ with the maximum number $1 + \frac{(n-1)(n-2)}{2}$. Since many configurations can occur as the maximum number reaches, a special interest is to know if two extreme can happen, namely if no closed branch of curves can sit in interior of another and if each closed branch sits in interior of another closed branch (namely they are all nested).

In [6], Rassias considered the maximum number of connected regions of that possible straight lines can induce. From statement of Hilbert's 16th problem, "as to the curves of 6th order, via a complicated process, it is true that of eleven branches which they can have according to Harnack, by no means all can lie external to one another", it seems that the first extreme case will not happen for any algebraic plane curve of degree 6. However, a proof does not occur for any degree $n$. [1] For the second extreme case, it seems not difficult to prove.

For algebraic hypersurface, such question is much difficult because too many cases (too many configurations) can occur when the maximum number of sheets of an algebraic hypersurface reaches. The most difficulty is to know what character of a polynomial determines the configurations of sheets. It seems that the degree of a given polynomial determines the maximum number of sheets. Then, what is the maximum number of sheets for a given algebraic hypersurface of degree $n$? The following questions may be helpful in solving or answering such open problem. For any curve on the algebraic hypersurface $P(x, y, z) = 0$, how many sheets can have this curve as their common boundary? Furthermore, for any point on the surface, how many common boundaries

[1] Petrovskii showed that the number of ovals of a curve of degree $2n$ not containing each other is less than or equal to $\frac{3}{2}n(n-1) + 1$.

(self-intersection curves) of sheets can have this point as their common vertex. These two questions play a central role in proving main theorems. It seems that we can only have parallel $n$ isolated sheets for configurations when maximum number of isolated sheets reaches. It makes us conjecture that the maximum number of sheets for an algebraic equation of degree $n$ is less than or equal to the maximum number of sheets for the algebraic equation of degree $n$, $P_1P_2\dots P_n = 0$, with degree one for all $P_i$. Now we expect to know the maximum number of sheets of $P_1P_2\dots P_n = 0$.

**Theorem 1.** *For the algebraic equation $P_1P_2\dots P_n = 0$, where $deg(P_i) = 1$ for all $i$, the maximum number of sheets is bounded above by*

$$\sum_{i=1}^{n} i^2 - n + 1.$$

*Proof.* We are going to show by induction on $n$. When $n = 1$, $P_1 = 0$ represents a plane in space which is a sheet. Hence the result is trivial. Now we assume that if $n = k - 1$, it is true that $P_1P_2\dots P_{k-1} = 0$ has no more than

$$\sum_{i=1}^{k-1} i^2 - k + 2$$

sheets. For $n = k$, consider the system $P_1P_2\dots P_{k-1} = 0$, $P_k = 0$. If the plane $P_k = 0$ intersects with all other planes $P_1 = 0$, $P_2 = 0, \dots,$ $P_{k-1} = 0$, we want to know how many extra sheets can be created after the plane $P_k = 0$ intersects with others. It is easy to see that any two planes in space can intersect with each other once, thus the plane $P_k = 0$ can intersect with other planes totally no more than $k - 1$ times. Look at each intersection, to find how many extra sheets can be created. Let $L$ be an intersection straight line of two planes, parametrized by $x = at + b$, $y = ct + d$, $z = et + f$. Since $P_1P_2\dots P_k(at + b, ct + d, et + f) = 0$ cam only have $k$ isolated roots in $t$, it indicates that surface $P_1P_2\dots P_k = 0$ can have no more than $k + 1$ pieces on the straight line $L$. Therefore, each intersection can not create more than $k + 1$ sheets. So extra $(k - 1)(k + 1) = k^2 - 1$ sheets can be created after the plane $P_k = 0$ intersects with others. It has

$$\sum_{i=1}^{k-1} i^2 - k + 2 + k^2 - 1 = \sum_{i=1}^{k} i^2 - k + 1.$$

By induction assumption, we have proved that the maximum number of for $P_1 P_2 \dots P_k = 0$ is bounded above by

$$\sum_{i=1}^{k} i^2 - k + 1.$$

$\square$

Next, we want to show that the maximum number of sheets for an arbitrary algebraic equation of degree $n$ is also bounded above by

$$\sum_{i=1}^{n} i^2 - n + 1$$

and want to know if the maximum number reaches only if $P(x, y, z)$ can be decomposed into $P_1 P_2 \dots P_n$. Does it imply that if $P$ is irreducible, the maximum number of sheets can be reduced?

**Theorem 2.** *The maximum number of sheets for an algebraic equation of degree $n$ is bounded above by*

$$\sum_{i=1}^{n} i^2 - n + 1.$$

*Proof.* We try to fulfill a proof via the following two lemmas and the theorem 1. First we have to show that the maximum number of sheets of $P(x, y, z) = 0$ having a curve as a common boundary is less than or equal to the maximum number of sheets of $P_1 P_2 \dots P_n = 0$ having a straight line as a common boundary. Second, one needs to show that the maximum number of paths on $P(x, y, z) = 0$ having a point as a common vertex is less than or equal to the maximum number of half rays on $P_1 P_2 \dots P_n = 0$ having a point as a common vertex. Then by the theorem 1, proposition 1 and the fact that each sheet has a curve as a boundary or a point as a common vertex (nonisolated sheet only), a proof of the theorem can be obtained.

We may first fix some terminology that a common boundary of sheets is a self-intersection curve of algebraic hypersurface $P(x, y, z) = 0$. A common vertex is an intersection point of common boundaries of sheets.

**Lemma 1.** *The maximum number of sheets having a curve as a common boundary is less than or equal to $2n$, which is a maximum number of sheets having a straight line as a common boundary.*

*Proof.* Consider any curve $L$ on the hypersurface $P = 0$, and assume there are $\alpha$ sheets having this curve as their common boundary. If $L$ is a straight line, $\alpha \leq 2n$ trivially because maximum $n$ planes can occur and each plane turns out to be two sheets via the separation line $L$ and because any surface other than plane containing such line has at least degree two and induces less sheets via the line $L$. If $L$ is not a straight line, then each algebraic sheet with partial boundary $L$ has at least degree 2, thus $\alpha \leq 2n$ too. It implies that the maximum number of sheets having a curve $L$ as their common boundary for an algebraic equation of degree $n$ is less than or equal to the maximum number of sheets having a straight line as their common boundary for the algebraic equation $P_1 P_2 \ldots P_n = 0$ with $deg(P_i) = 1$ for all $i$. □

**Lemma 2.** *The maximum number of paths (boundaries of sheets) on $P(x, y, z) = 0$ having a point as a common vertex is less than or equal to the maximum number of half rays (boundaries of sheets) on $P_1 P_2 \ldots P_n = 0$ having a point as a common vertex.*

*Proof.* It is trivial. □

Since a common vertex is an intersection point of common boundaries of sheets and can be thought a special point on some curve which is a boundary of sheets, via the lemma 1, that the maximum number of half rays (boundaries of sheets) on $P_1 P_2 \ldots P_n = 0$ having a point as common vertex is equal to $2n$ implies the lemma 2. □

Let us review Bezout's theorem here for later use.

**Bezout's Theorem.** *Let $f_1, f_2, \ldots, f_m$ be hypersurface of $KP^m$ which only intersect in a finite set $\{m_j\}$ of points and let $d_i$ be the degree of $f_i$. There may then be assigned multiplicities to the $m_j$ independent of the coordinate system such that counted with these multiplicities the number of intersections is $d = d_1 d_2 \ldots d_m$.*

**Generalized Bezout's Theorem.** *If two algebraic manifolds intersect normally and have orders $m$ and $n$ respectively, then their intersection is an algebraic manifold of order $mn$.*

We first consider some open problem for plane algebraic curve. Harnack showed that for a plane algebraic equation of degree $n$, the maximum number of closed branches is bounded above by $1 + \frac{(n-1)(n-2)}{2}$. It seems that this number is somehow related to the number of double points. The following theorem is classic algebraic geometry indicates that.

**Theorem.** *An irreducible curve $C^n$ cannot have more than $\frac{(n-1)(n-2)}{2}$ double points.*

It is natural to ask what intrinsic property of the algebraic equation $P(x,y)=0$ determines the configurations of closed branches when the maximum number reaches. Especially, two extreme cases are that all branches lie external to one another and all branches are nested. As to the curves of the 6th order, Hilbert found that the first case cannot occur. It is easy to have the following proposition for second case. When $n=2$, it is trivial.

**Proposition 2.** *For $n \geq 3$, when the maximum number of closed branches reaches, all closed branches cannot be nested.*

*Proof.* This is very trivial if one considers a straight line intersecting with each closed branch twice and Bezout's theorem on the maximum number of intersection points of a curve of order $n$ and a straight line of order 1, because $2\left(1+\frac{(n-1)(n-2)}{2}\right) > n$ if $n \geq 3$. □

It seems much difficult to show the first case in general. However, above classic theorem and Harnack's theorem can be used to carry out a proof.

**Proposition 3.** *For an irreducible algebraic curve of degree $n$ ($n \geq 5$), all closed branches cannot lie external to one another, when the maximum number of closed branches $1+\frac{(n-1)(n-2)}{2}$ reaches.*

*Proof.* We mainly classify some nodes and conclude the impossibility of above configurations, by using some induction on the number of nodes (double points, multi-nodes) and the degree $n$. Some closed branch is an oval without any node, for example,

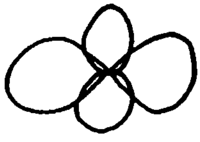

Without discussion on the possibility of such oval for an irreducible algebraic curve, an intuitive argument leads to the fact that each oval has at least degree two in its algebraic representation. We assume that among $1+\frac{(n-1)(n-2)}{2}$ closed branches, only one oval and all others induced by some nodes. By the assumption, we have $\frac{(n-1)(n-2)}{2}$

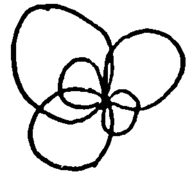
double points, and one oval with at least degree two. It implies that an irreducible algebraic curve of degree $n-2$ has $\frac{(n-1)(n-2)}{2}$ double points. It contradicts to above classic theorem. Similarly if we have $k$ ovals without double points ($k \geq [\frac{n}{2}]$) and other closed branches induced by simple double points, then an irreducible algebraic curve of degree $n-2k$ can have $1-k+\frac{(n-1)(n-2)}{2}$ double points. It also contradicts to the above theorem because $1-k+\frac{(n-1)(n-2)}{2} > \frac{(n-2k-1)(n-2k-2)}{2}$ for $n \geq 3$, $1 \leq k \leq [\frac{n}{2}]$. If a node induces more than one closed branches, for examples, two or three closed branches,

(a)

(b)

then we here call such node a multi-nodes (different from usual definition of multi-nodes) and use similar argument for each compound closed branches induced by one multi-nodes. As a simple fact, the degree of each compound closed branches induced by one multiple point is at least the number of closed branches, namely (a) has at least degree two and (b) has at least degree three, etc. In general, if a multi-nodes induces $k$ closed branches, then the algebraic curve generating that compound closed branches has at least degree $k$. Here we have to clarify two different questions on closed branches of an irreducible algebraic curve and on connected regions induced by an irreducible algebraic curve. The following example are not the situations that each closed branch lie external to one another.





However, the connected regions induced by an irreducible algebraic curve on above cases are nine and eleven. The closed branches of an irreducible algebraic curve are four and five although they do not lie external to one another. Therefore, such cases are not included in our assumption, otherwise, there is nothing to prove because we already know when the maximum number reaches, all closed branches do not lie external to one another. Now if $k$ closed branches among $1+\frac{(n-1)(n-2)}{2}$ circuits is induced by a multi-nodes, then algebraic curve of degree $n-k$ has $1-k+\frac{(n-1)(n-2)}{2}$ closed branches, each with one double point. According to above classic theorem, an irreducible algebraic curve of $n-k$ cannot have more than $\frac{(n-k-1)(n-k-2)}{2}$ double points. Since

$$\frac{(n-k-1)(n-k-2)}{2}=\frac{(n-1)(n-2)}{2}-\frac{k(n-2)}{2}-\frac{k(n-1)}{2}+\frac{k^2}{2}=$$

$$=\frac{(n-1)(n-2)}{2}-\frac{k(2n-3)}{2}+\frac{k^2}{2}=\frac{(n-1)(n-2)}{2}-k+1-1+\frac{5k}{2}-kn+\frac{k^2}{2}.$$

If $-1+\frac{5k}{2}-kn+\frac{k^2}{2}<0$, then $\frac{(n-1)(n-2)}{2}-k+1>\frac{(n-k-1)(n-k-2)}{2}$. It is a contradiction. Now we are going to show $-1+\frac{5k}{2}-kn+\frac{k^2}{2}<0$ for $n\geq 5$ and $k\leq n$. In fact, $-1+\frac{5k}{2}-kn+\frac{k^2}{2}=k\left(\frac{5}{2}-n+\frac{k}{2}\right)-1=\frac{k(5-2n_k)}{2}-1<0$ if $5-2n+k\leq 0$, namely $\frac{5+k}{2}\leq n$. Clearly $n\geq\frac{5+k}{2}$ if $n\geq 5$ and $k\leq n$. For any combinations of ovals without nodes, multi-nodes inducing several closed branches and each double point having only one closed branch, it is similar. A contradiction follows. Therefore, the proof is completed. □

As we mentioned before, for an irreducible algebraic surface of degree $n$, is it possible to reach to the maximum number of sheets, $\sum_{i=1}^{n} i^2-n+1$? In order to answer such question, one needs some classification of nodes and intersection curves of algebraic hypersurface. It seems that the possibility of such positive answer is rare except for $n=1$. How to determine the degree of intersection curve is a key point. We claim that for an irreducible algebraic hypersurface of degree $n$ ($n\geq 2$), it is impossible to reach the maximum number of sheets, $\sum_{i=1}^{n} i^2-n+1$. However, a lower upper bound for such case cannot be provided here. It is still open. For configurations of sheets as the maximum number reaches, it seems much difficult to discuss. We may discuss the configurations of connected components of the complement of an algebraic hypersurface instead. Two extreme cases that all connected components are disjoint and all connected components are nested are the most interesting. It may easily provide an exact answer for such two situations.

As the problem mentioned in [8], the extension of Hilbert's 16th problem to arbitrary dimensional algebraic hypersurface is certainly an interesting question, especially for 4-dimensional case $P(x_1, x_2, x_3, x_4) = 0$. Discussion of sheets of such algebraic hypersurface may lead to some idea for classification of submanifolds of $R^4$. This is a very popular question in differential topology, differential manifolds, geometric topology, and so on. Poincaré's conjecture for $S^2$ in differential topology is very elegant problem which is still open at this moment. It leads to the impossibility of the classification of compact 3-manifolds in $R^4$. We hope to see some progress along this line and some new discussion on 4-dimensional hypersurface.

**Acknowledgment.**

Many thanks go to professors S. Smale and C. Pugh for their benefit suggestions on the original manuscript. I take this opportunity to thank professor Th. M. Rassias for communicating his ideas on solving Hilbert's 16th problem, for providing many profitable comments, and for reading the manuscript.

## References

[1] AMS, *Proceedings of symposia in pure mathematics*, Vol. XXV, III, 1974.
[2] J.L. Coolidge, *A treatise on algebraic plane curves*, Dover Publication, Inc., 1959.
[3] D.A. Gudkov, *The topology of real projective algebraic varieties*, Russian Math. Surveys, 29(4), 1974, 1-79.
[4] M.W. Hirsch, *Differential topology*, GTM, 1976.
[5] S. Lang, *Introduction to algebraic geometry*, 1958.
[6] Th. M. Rassias *Remarks on the Hilbert's 16th problem*, Proceedings of the Academy of Athens, Vol.54, 1979.
[7] Th. M. Rassias, *On the topology of algebraic curves*, Bull. of the Inst. of Math., Academia Sinica, V13, No.3, 1985.
[8] Th. M. Rassias, *A remark and problem for polynomials of several variables*, General Inequalities 6, 6th International Conference on General Inequalities, Oberwolfach, Dec. 9-15, 1990, 487-488.
[9] J.G. Semple and L. Roth, *Introduction to algebraic geometry*, Oxford Univ. Press, 1985.
[10] P.X. Sheng, *Hilbert's 16th problem*, Analysis and Topology (edited by C. Andreian-Cazacu, O. Lehto, Th. M. Rassias), 1996.

DEPARTMENT OF MATHEMATICS, SHANGHAI UNIVERSITY, SHANGHAI 201800, P.R.C.

STUDIA UNIV. "BABEŞ-BOLYAI", MATHEMATICA, Volume XLII, Number 4, December 199?

# THE MANEFF-TYPE TWO-BODY PROBLEM IN VELOCITY PLANE

CRISTINA STOICA AND VASILE MIOC

**Abstract.** One studies the qualitative evolution of the relative motion in the two-body problem for Maneff-type potentials. Using the prime integrals of the angular momentum and energy, one gets the trajectories in the plane of the polar components of the velocities in the case of nonradial motions for all the combinations of the parameters $(A, B)$ of the field, the angular momentum $(C)$ and energy constant $(h)$, corresponding to the real motion.

## 1. Introduction

Consider a central force field of centre $M$ characterized by a quasihomogeneous potential function of the form $A/r+B/r^2$, where $r=|\mathbf{r}|$ ($\mathbf{r}$ = position vector of a particle with respect to $M$), and $A, B$ = constants. We have called it (e.g. [13]) Maneff-type field, because such a potential function generalizes that proposed by G. Maneff [6-9] and reconsidered recently in a series of studies having as departure point F.N. Diacu's [2] researches.

Let a particle of unit mass be moving in a Maneff-type field. Its relative motion with respect to $M$ will be planar and described by the equation

$$\ddot{\mathbf{r}} = -A\mathbf{r}/r^3 - 2B\mathbf{r}/r^4 \tag{1}$$

where dots mark time-differentiation.

Concrete expressions assigned to the parameters $A$ and $B$ can model various physical situations (for example, see [13]) belonging to (celestial) mechanics [3, 4, 11-13], theoretical physics [6-9], relativity (cf. [10]), astrophysics [16], atomic physics [2], and so forth.

Received by the editors: November 7, 1996.

1991 *Mathematics Subject Classification.* 70F05.

*Key words and phrases.* Maneff field, two-body problem, velocity plane

The Maneff-type two-body problem was tackled by us in some previous papers (see [13]) from different standpoints. The same problem will be approached here through the study of the trajectories in velocity plane.

As proved in [5], to complete a well-known result (e.g. [15]), the Keplerian trajectories of the Newtonian two-body problem are represented by circles (or arcs of circles) in the plane defined by the polar components of the relative velocity. In [14] we have generalized this result, showing that in the same velocity plane the trajectories of the Maneff two-body problem are mainly ellipses (or arcs of ellipses). In this paper we shall prove that the corresponding trajectories for the Maneff-type two-body problem are conic sections (degenerate or not) or portions of them. These trajectories will be identified for every situation belonging to the allowed interplay among field parameters, angular momentum and total energy. For each such situation, the qualitative behaviour of the particle will be pointed out.

## 2. Equations of motion and first integrals

Using polar coordinates $(r, u)$, equation (1) transform into

$$\ddot{r} - r\dot{u}^2 = -A/r^2 - 2B/r^3, \tag{2}$$

$$r\ddot{u} + 2\dot{r}\dot{u} = 0, \tag{3}$$

system to which we attach the initial conditions

$$(r, u, \dot{r}, \dot{u})(t_0) = (r_0, u_0, \dot{r}_0 = V_0 \cos\alpha, \dot{u}_0 = V_0 \sin\alpha / r_0), \tag{4}$$

where $V_0 = V(t_0)$, $V = |\dot{r}|$ = velocity, $\alpha$ = angle between initial radius vector and initial velocity.

The field being central, the angular momentum is conserved, and (3) provides the first integral

$$r^2\dot{u} = C, \tag{5}$$

where $C = r_0 V_0 \sin\alpha$ is the constant angular momemtum. The first integral of energy can also be easily obtained by the usual technique

$$V^2 \equiv \dot{r}^2 + r^2\dot{u}^2 = 2A/r + 2B/r^2 + h, \tag{6}$$

where $h = V_0^2 - 2A/r_0 - 2B/r_0^2$ is the constant of energy.

## 3. Trajectories in velocity plane

It is clear that the conservation of the angular momemtum restricts the velocity space to a plane. So, let $V_r = \dot{r}$, $V_u = r\dot{u}$ be the polar components of the velocity. By (5) we have $rV_u = C$, and replacing this in (6) we get

$$\frac{C^2 - 2B}{C^2} V_u^2 - 2\frac{A}{C} V_u + V_r^2 = h. \tag{7}$$

Observe that we have tacitly that $C \neq 0$. Indeed, for $C = 0$ (radial motion) we have $V_u = 0$ and the study of the trajectories in velocity plane becomes meaningless. It goes without saying that in such a case it is natural to resort to the phase plane $(r, \dot{r})$. Actually such a study was already performed [13] for both zero and nonzero $C$; however our present study is not useless because the curves in the $(V_u, V_r)$ plane are conic sections and this makes easier the investigation of the qualitative behaviour of the particle.

Indeed, putting $x = V_u$, $y = V_r$, $a_{11} = (C^2 - 2B)/C^2$, $a_{22} = 1$, $a_{13} = -A/C$, $a_{33} = -h$, $a_{12} = a_{23} = 0$, (7) reads

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33} = 0,$$

namely the equation of a conic section. The sign of the expression

$$\delta = a_{11}a_{22} - a_{12}^2 = (C^2 - 2B)/C^2$$

gives the kind of the conic section, while the nature of this one is given by the sign of

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} = \frac{h(2B - C^2) - A^2}{C^2}.$$

Observe that there exists a critical value $h_c = A^2/(2B - C^2)$ for which the conic sections are degenerate ($\Delta = 0$).

Now, if $C^2 < 2B$ we have $\delta < 0$, and (7) represents a family of hyperbolas of the form

$$(V_u - w)^2/a^2 - V_r^2/b^2 = 1, \tag{8}$$

with:

- centre: $P(w, 0) = (-AC/(2B - C^2), 0)$;
- semiaxes: $a = \sqrt{C^2[A^2 - h(2B - C^2)]}/(2B - C^2)$,
  $b = \sqrt{A^2/(2B - C^2) - h}$;

- foci: $\left(\left[-AC \pm \sqrt{2B[A^2 - h(2B - C^2)]}\right]/(2B - C^2), 0\right)$;
- asymptotes: $V_r = \pm\left(\sqrt{2B - C^2}/C^2\right)[V_u + AC/(2B - C^2)]$;
- intersections with $V_u$ axis: $\left(C\left[-A \pm \sqrt{A^2 - h(2B - C^2)}\right]/(2B - C^2), 0\right)$.

It is clear that for the family of hyperbolas with the above characteristics we must have $h < h_c$ ($\Delta < 0$). If $h = h_c$, (7) represents the asymptotes of the family (8). If $h > h_c$ ($\Delta > 0$), (7) represents the family of conjugate hyperbolas with respect to (8).

If $C^2 = 2B$ then $\delta = 0$, and (7) represents a family of parabolas of the form

$$V_r^2 = 2p(V_u - q), \tag{9}$$

with $p = A/C$, $q = -hC/(2A)$, focus $((A^2 - hC^2)/(2AC), 0)$, and intersection with $V_u$ axis $(-hC/(2A), 0)$. The parabolas are nondegenerate for $A \neq 0$. For $A = 0$ we have $\Delta = 0$, and any parabola reduces to a couple of straight lines (distinct or not) parallel to $V_u$ axis.

Finally, $C^2 > 2B$ leads to $\delta > 0$, and (7) represents a family of ellipses of the form

$$(V_u - w)^2/a^2 + V_r^2/b^2 = 1, \tag{10}$$

with:

- centre: $P(w, 0) = (AC/(C^2 - 2B), 0)$;
- semiaxes: $a = \sqrt{C^2[A^2 + h(C^2 - 2B)]}/(C^2 - 2B)$,
  $b = \sqrt{A^2/(C^2 - 2B) + h}$;
- intersection with $V_u$ axis: $\left(C\left[-A \pm \sqrt{A^2 + h(C^2 - 2B)}\right]/(C^2 - 2B), 0\right)$;
- intersection with $V_r$ axis: $\left(0, \pm\sqrt{A^2/(C^2 - 2B) + h}\right)$;
- foci: $\left(\left[AC \pm \sqrt{2B[A^2 + h(C^2 - 2B)]}\right]/(C^2 - 2B), 0\right)$ for $B > 0$,
  $\left(AC/(C^2 - 2B), \pm\sqrt{-2B[A^2 + h(C^2 - 2B)]}/(C^2 - 2B)\right)$ for $B < 0$;

(for $b = 0$, (10) reduces to a family of circles centered in $P(A/C, 0)$ of radii $\sqrt{A^2/C^2 + h}$).

Observe that for the family of ellipses with the above characteristics we must have $h > h_c$ ($\Delta < 0$). If $h = h_c$ the whole family reduces to the point $P$. If $h < h_c$ ($\Delta > 0$), (7) represents a family of imaginary ellipses.

Till now we did not use the somewhat natural condition $C$ nonnegative. Together with $C \neq 0$, this leads immediately to $V_u > 0$, hence the motion in velocity plane is possible only in the halfplane $V_u > 0$ (at limits $V_u \to 0$ when $r \to \infty$, and conversely).

As a consequence, more restrictive conditions will be imposed to $h$. Examining (7) and taking also into account the features of the above conic sections, we find that the motion is not possible for:

$\{C^2 = 2B,\ A < 0,\ h \leq 0\}$: parabolas lying wholly in the forbidden plane $V_u \leq 0$;

$\{C^2 = 2B,\ A = 0,\ h < 0\}$: imaginary parallel straight lines;

$\{C^2 > 2B,\ A \leq 0,\ h \leq 0\}$: real ellipses lying wholly in the forbidden plane $V_u \leq 0$;

$\{C^2 > 2B,\ A > 0,\ h < h_c\}$: imaginary ellipses.

## 4. Interpretation of motion in velocity plane

We have shown that in the velocity plane $(V_u, V_r)$ the trajectories are only conic sections (degenerate or not). The motion on these curves may only the following characteristics (see also Figures 1-3 below):

- monotonic increase/decrease of $V_u$, tending to $\infty/0$, or to an equilibrium;

- monotonic increase/decrease of $V_u$ up to a maximum/minimum value, then monotonic decrease/increase, tending to $0/\infty$'

- oscillation of $V_u$ between two finite and positive limit values;

- constancy of $V_u$.

Let us now interpret these scenarios in terms of real motion. To do this, remind that $V_u = C/r$, so increase/decrease of $V_u$ means decrease/increase of $r$; $V_u \to \infty$ $(r \to 0)$ means collision (if $V_r < 0$) or ejection (if $V_r > 0$); $V_u \to 0$ $(r \to \infty)$ means escape.

Also observe that, since $\dot{u} > 0$ (see (5)) during the whole motion, to every segment of monotonic increase/decrease of $V_u$ on the curves in velocity plane corresponds a spiral motion of the particle (performed inwards/outwards). Accordingly, the oscillation of $V_u$ between two finite and positive limit values means that the particle moves on a noncollisional bounded trajectory, quasiperiodic or periodic (i.e. filling densely or not the annulus determined by the two limits of $r$ corresponding to those of $V_u$; see also [1]). Lastly, the constancy of $V_u$ means circular motion of the particle.

With this interpretation, the qualitative behaviour of the particle for the whole allowed interplay among field parameters, angular momentum and total energy can be described by the following

**Theorem.** *Consider the two-body problem in a Maneff-type field. The only possible scenarios for the relative motion with nonzero angular momentum are:*

$S_1$: *spiral motion inwards ending in collision;*

$S_2$: *spiral motion outwards up to a finite maximum distance, then spiral motion inwards ending in collision;*

$S_3$: *spiral motion outwards tending asymptotically to the equilibrium circular orbit;*

$S_4$: *periodic (rosette) or quasiperiodic motion, starting inwards;*

$S_5$: *circular motion;*

$S_6$: *periodic (rosette) or quasiperiodic motion, starting outwards;*

$S_7$: *spiral motion inwards tending asymptotically to the equilibrium circular orbit;*

$S_8$: *spiral motion inwards up to a nonzero minimum distance, then spiral motion outwards leading to escape;*

$S_9$: *spiral motion outwards leading to escape.*

The qualitative characteristics of these motions (maximum and minimum extenses, radius of equilibrium circular orbit) can be easily found from the qualitative features of the conic sections pointed out in Section 3.

In the next sections we shall examine in detail those trajectories in the velocity plane which represent real motion, pointing out the corresponding scenarios case by case. The initial conditions for the velocity plane motion will be denoted by $(V_u^0, V_r^0)$.

## 5. Hyperbolas in velocity plane

The condition for such trajectories is $C^2 < 2B$. The motion of these curves is represented in Figure 1, where the $V_r$ axis changes its position according to the considered case (remind that the allowed halfplane for real motion is $V_u > 0$). The corresponding curves (or arcs of curves) for each allowed combination $\{A, B, C, h\}$ are easy to identify.

Fig.1

**5.1.** Case $A < 0$. In this case $w > 0$. If $h \leq 0$ we have

$$V_r^0 \leq 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_2.$$

If $0 < h < h_c$ we have

$$V_u^0 < w: \quad V_r^0 < 0 \Rightarrow S_8, \quad V_r^0 \geq 0 \Rightarrow S_9;$$

$$V_u^0 > w: \quad V_r^0 \leq 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_2.$$

If $h = h_c$ we have

$$V_u^0 < w: \quad V_r^0 < 0 \Rightarrow S_7, \quad V_r^0 > 0 \Rightarrow S_9;$$

$$V_u^0 = w: \quad (V_r^0 = 0) \Rightarrow S_5 \text{ (unstable orbit)};$$

$$V_u^0 > w: \quad V_r^0 < 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_3.$$

If $h > h_c$ we have

$$V_r^0 < 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_9.$$

**5.2.** Case $A = 0$. In this case $h_c = 0$ and $w = 0$. If $h < 0$ we have

$$V_r^0 \leq 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_2.$$

If $h \geq 0$ we have

$$V_r^0 < 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_9.$$

**5.3.** Case $A > 0$. In this case $w < 0$. If $h < 0$ we have

$$V_r^0 \leq 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_2.$$

If $h \geq 0$ we have

$$V_r^0 < 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_9.$$

## 6. Parabolas in velocity plane

The condition for such trajectories is $C^2 = 2B$. The motion on these curves is represented in Figure 2(a,b,c); the $V_r$ axis also changes its position according to the considered case.

Fig.2a

Fig.2b

Fig.2c

**6.1.** Case $A < 0$ (Figure 2a). Real motion is possible only for $h > 0$, and we have

$$V_r^0 < 0 \Rightarrow S_8, \quad V_r^0 \geq 0 \Rightarrow S_9.$$

**6.2.** Case $A = 0$ (Figure 2b). Real motion is possible only for $h \geq 0$. If $h = 0$ every trajectory reduces to a point $(V_u^0, 0)$ on the positive $V_u$ semiaxis; this means $S_5$ and the circular motion is stable.

If $h > 0$ we have

$$V_r^0 < 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_9.$$

**6.3.** Case $A > 0$ (Figure 2c). If $h < 0$ we have

$$V_r^0 < 0 \Rightarrow S_1, \quad V_r^0 \geq 0 \Rightarrow S_2.$$

If $h \geq 0$ we have

$$V_r^0 < 0 \Rightarrow S_1, \quad V_r^0 > 0 \Rightarrow S_9.$$

## 7. Ellipses in velocity plane

The condition for such trajectories is $C^2 > 2B$. The motion of these curves or arcs of curves (easy to identify for each allowed combination $\{A, B, C, h\}$) is represented in Figure 3; the $V_r$ axis changes its position according to the considered case.

Fig.3

**7.1.** Case $A < 0$. In this case $w < 0$ and real motion is possible only for $h > 0$. We have

$$V_r^0 < 0 \Rightarrow S_8, \quad V_r^0 \geq 0 \Rightarrow S_9.$$

**7.2.** Case $A = 0$. Now $w = 0$ and real motion is possible only for $h > 0$. The same initial conditions lead to the same scenarios as above.

**7.3.** Case $A > 0$. In this case $w > 0$ and real motion is possible only for $h \geq h_c$.

If $h = h_c$ all trajectories reduce to point $P$; this means $S_5$ and the circular motion is stable.

If $h_c < h < 0$ we have

$$\{(V_u^0 = w - a) \vee (V_r^0 < 0)\} \Rightarrow S_4, \quad \{(V_u^0 = w + a) \vee (V_r^0 > 0)\} \Rightarrow S_6.$$

If $h \geq 0$ we have

$$V_r^0 < 0 \Rightarrow S_8, \quad V_r^0 \geq 0 \Rightarrow S_9.$$

## 8. Concluding remarks

Tackling a qualitative study of the two-body problem in Maneff-type fields we replaced the usual phase plane $(r, \dot{r})$ by the velocity plane $(V_u, V_r)$. The trajectories which represent the real motion in this plane were found to be conic sections (nondegenerate or degenerate) or portions of them.

Compared with the use of the $(r, \dot{r})$ plane, that of the velocity plane has both advantages and disadvantages. The main advantage consists of the fact that the trajectories in velocity plane are conic sections, and their behaviour is very well known (providing immediately the image of the qualitative behaviour of the particle), while the usual phase curves are more complicated (see [13]). This makes fairly immediate the proof of the theorem stated in Section 4 (theorem also stated in [13], but with a considerably longer proof based on the study of usual trajectories). The disadvantage is the fact that one cannot study in this way the rectilinear motion (case of zero angular momentum).

Surveying the whole allowed interplay among field parameters, angular momentum and total energy, nine possible scenarios for the real motion were found. These scenarios illustrate four essential trends: collision path, escape path, trajectory tending to equilibrium, and quasiperiodic or periodic motion.

As a final remark, if our field is just Maneff's one (see e.g. [2, 11, 14]), in realistic astronomical situations we shall have $C^2 > 2B$ (see [14]) and $A > 0$, finding the ellipses discussed in subsection 7.3, and recovering in this way the results obtained in [14]. Putting further $B = 0$, the circles obtained in [5] are recovered.

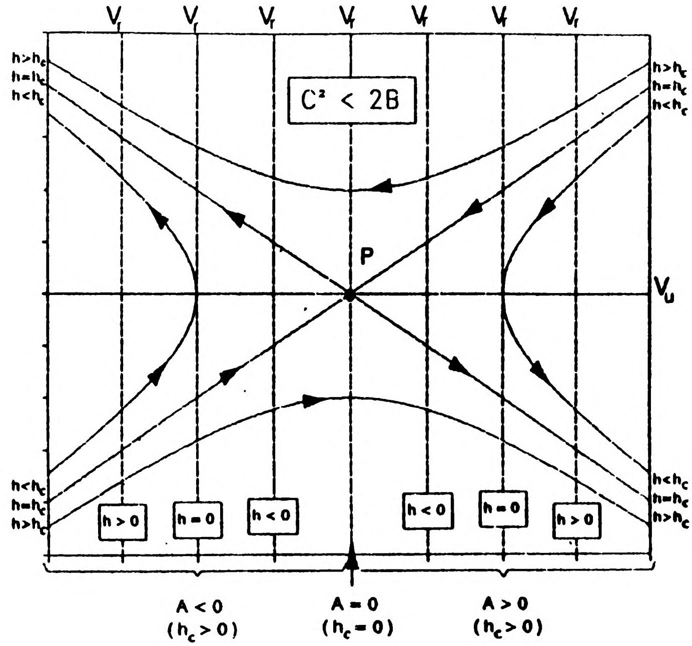



FIGURE 1

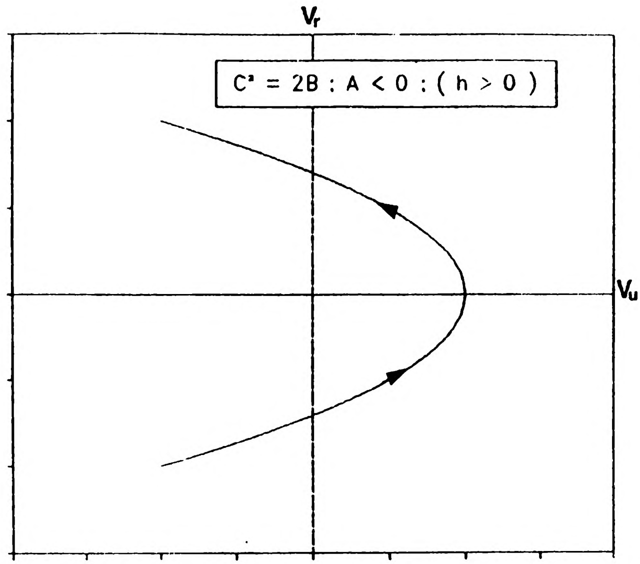



FIGURE 2A

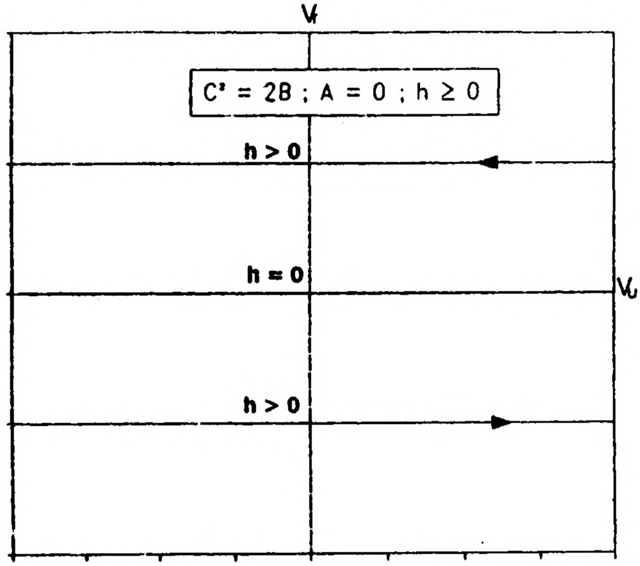



FIGURE 2B

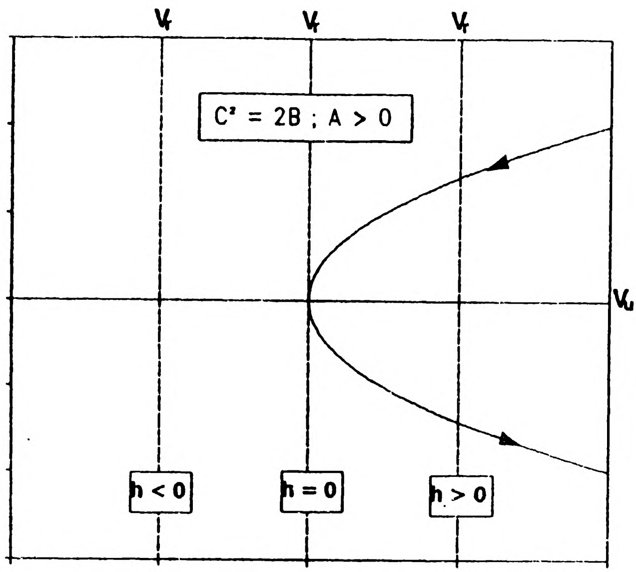



FIGURE 2C

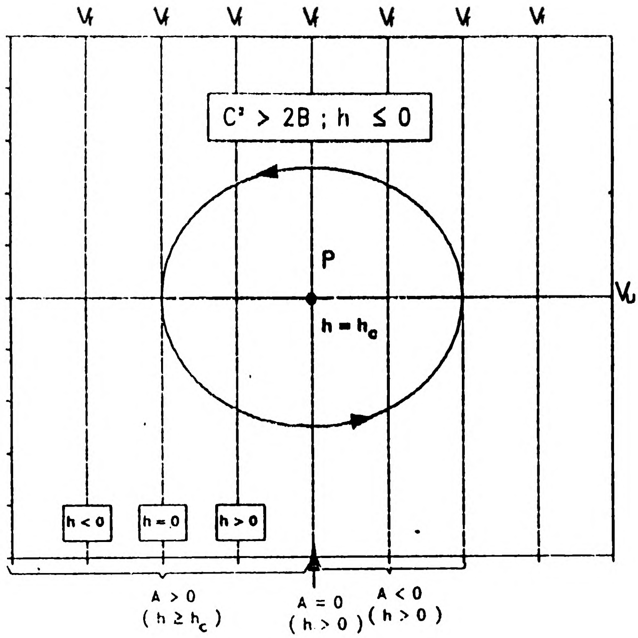



FIGURE 3

## References


[1] Arnold, V., *Les méthodes mathématiques de la mécanique classique*, Ed. Mir, Moscou, 1976.

[2] Diacu, F.N., *The Planar Isosceles Problem for Maneff's Gravitational Law*, J. Math. Phys., 34(1993), 5671-5690.

[3] Diacu, F.N., *Near-Collision Dynamics for Particle Systems with Quasihomogeneous Potentials*, J. Diff. Eq. (1996) 128(1996) 58-77.

[4] Diacu, F.N., Mingarelli, A., Mioc, V., Stoica, C., *The Manev Two-Body Problem: Quantitative and Qualitative Theory*, in R.P. Agarwal (ed.), *Dynamical Systems and Applications*, World Sci. Ser. Appl. Anal., Vol.4, world Scientific Publ. Co., Singapore, 1995, 213-227.

[5] Krpić, D., Anicin, I., *In Velocity Spaces the (Unperturbed) Planets Move in Circles Only*, Publ. Obs. Astron. Belgrade, No.44(1993), 95-98.

[6] Maneff, G., *La gravitation et le principe de l'égalité de l'action et de la réaction*, C.R. Acad. Sci. Paris, 178(1924), 2159-2161.

[7] Maneff, G., *Die Gravitation und das Prinzip von Wirkung und Gegenwirkung*, Z. Phys., 31(1925), 786-802.

[8] Maneff, G., *Le principe de la moindre action et la gravitation*, C.R. Acad. Sci. Paris, 190(1930), 963-965.

[9] Maneff, G., *La gravitation et l'énergie au zéro*, C.R. Acad. Sci. Paris, 190(1930), 1374-1377.

[10] Mioc, V., *Elliptic-Type Motion in Fock's Gravitational Field*, Astron. Nachr., 315(1994), 175-180.

[11] Mioc, V., Stoica, C., *Discussion et résolution complète du problème des deux corps dans le champ gravitationnel de Maneff*, C.R. Acad. Sci. Paris, 320(1995), 645-648.

[12] Mioc, V., Stoica, C., *Discussion et résolution complète du problème des deux corps dans le champ gravitationnel de Maneff* (II), C.R. Acad. Sci. Paris, 321(1995), 961-964.

[13] Mioc, V., Stoica, C., *A Qualitative Study of the Two-Body Problem in Maneff-Type Fields*, Rom. Astron. J., 5(1995) (to appear).

[14] Mioc, V., Stoica, C., *Unperturbed Trajectories in Maneff's Gravitational Field are Ellipses in Velocity Plane*, Bull. Astron. Belgrade, No.152(1995), 43-47.

[15] Sommerfeld, A., *Mechanics*, Academic Press, New York, 1964.

[16] Ureche, V., *Free-Fall Collapse of a Homogeneous Sphere in Maneff's Gravitational Field*, Rom. Astron. J., 5(1995), 145-148.

INSTITUTE FOR GRAVITATION AND SPACE SCIENCES, LABORATORY FOR GRAVITATION, 71111 BUCHAREST, ROMANIA

ASTRONOMICAL INSTITUTE OF THE ROMANIAN ACADEMY, ASTRONOMICAL OBSERVATORY CLUJ-NAPOCA, 3400 CLUJ-NAPOCA, ROMANIA

STUDIA UNIV. "BABEŞ-BOLYAI", MATHEMATICA, Volume XLII, Number 4, December 1997

# ASYMPTOTIC FORMULAE CONCERNING ARITHMETICAL FUNCTIONS DEFINED BY CROSS-CONVOLUTIONS, II. THE DIVIZOR FUNCTION

LÁSZLÓ TÓTH

**Abstract.** Let $A$ be a regular convolution of Narkiewicz and let $\tau_A(n)$ denote the number of $A$-divisors of $n$. We establish an asymptotic formula for the summatory function of $\tau_A$ if $A$ is a cross-convolution investigated in the first part of the present paper.

## 1. Introduction

In the first part [T97] of this paper and in [TH96] we introduced the notion of cross-convolution of arithmetical functions as a special case of Narkiewicz's [Nar63] regular convolution as follows. Let $\mathbb{N}$ denote the set of positive integers and let $A$ be a regular convolution given by

$$(f *_A g)(n) = \sum_{d \in A(n)} f(d)g(n/d),$$

see [Nar63], [McC86], [Sit78], [T97]. The elements of the set $A(n)$ are called the $A$-divisors of $n$. We say that $A$ is a *cross-convolution* if for every prime $p$ we have either $A(p^a) = \{1, p, p^2, ..., p^a\} \equiv D(p^a)$ or $A(p^a) = \{1, p^a\} \equiv U(p^a)$ for every $a \in \mathbb{N}$. Let $P$ and $Q$ be the sets of the primes of the first and second kind of above, respectively, where $P \cup Q = \mathbb{P}$ is the set of all primes. For $P = \mathbb{P}$ and $Q = \emptyset$ we have the Dirichlet convolution $D$ and for $P = \emptyset$ and $Q = \mathbb{P}$ we obtain the unitary convolution $U$.

Furthermore, let $(P) = \{1\} \cup \{n \in \mathbb{N} : \text{each prime factor of } n \text{ belongs to } P\}, (Q) = \{1\} \cup \{n \in \mathbf{N} : \text{each prime factor of } n \text{ belongs to } Q\}$. Every $n \in \mathbb{N}$ can be written uniquely in the form $n = n_P n_Q$, where $n_P \in (P), n_Q \in (Q)$ and $(n_P, n_Q) = 1$. If $A$ is a cross-convolution, then $A(n) = \{d \in \mathbb{N} : d|n, (d, n/d) \in (P)\}$ for every $n \in \mathbb{N}$.

Received by the editors: October 23, 1996.
1991 *Mathematics Subject Classification.* 11A25, 11N37.
*Key words and phrases.* Narkiewicz's regular convolution, divisor function, asymptotic formula.

Let $\tau_A(n)$ denote the number of $A$-divisors of $n$. Observe that if $A$ is a cross-convolution, then $\tau_A(n) = \tau(n_P)\tau^*(n_Q) = \tau(n_P)2^{\omega(n_Q)}$, where $\omega(n_Q)$ represents the number of distinct prime factors of $n_Q$.

The aim of the present paper is to establish an asymptotic formula for the summatory function of $\tau_A$ if $A$ is a cross-convolution, which generalizes and unifies the corresponding known results concerning the divisor function $\tau$ and its unitary analogue $\tau^*$.

Our method is elementary and it applies an asymptotic estimate of B. GORDON and K. ROGERS [GR64] concerning the divisor function $\tau$. The results of this paper are parts of our thesis [T95].

Preliminaries

We need the following lemmas.

**Lemma 1.** *Let $A$ be a cross-convolution and let $h$ be the multiplicative function defined by $h(1) = 1$ and*

$$h(p^a) = \begin{cases} -1, & \text{if } p \in Q \text{ and } a = 1, \\ 0, & \text{otherwise,} \end{cases}$$

*for every prime power $p^a$. Then*

$$\tau_A(n) = \sum_{d^2e=n} h(d)\tau(e),$$

*for every $n \in \mathbb{N}$.*

*Proof.* Taking into account the multiplicativity of the functions $\tau_A, h$ and $\tau$ it is sufficient to verify the above identity for $n = p^a$, a prime power. Let $F(n) = \sum_{d^2e=n} h(d)\tau(e)$. Then for $p \in P$ we have $F(p^a) = h(1)\tau(p^a) = \tau(p^a) = \tau_A(p^a)$, and for $p \in Q$ we obtain $F(p^a) = h(1)\tau(p^a) + h(p)\tau(p^{a-2}) = (a+1) - (a-1) = 2 = \tau^*(p^a) = \tau_A(p^a)$ if $a \geq 2$ and $F(p) = h(1)\tau(p) = 2 = \tau_A(p)$, which completes the proof. □

If $Q = \mathbb{P}$, then $h$ is the Möbius function $\mu$ and we reobtain the identity given by M. V. SUBBARAO and D. SURYANARAYANA [SSur78], page 5.

The following properties were suggested by the paper of D. SURYANARAYANA [Sur69] and they represent generalizations of the results of that paper.

**Lemma 2.** *If $h$ is the function defined in Lemma 1, $u \in \mathbb{N}$ and $s > 1$, then*

$$\sum_{\substack{n=1\\(n,u)=1}}^{\infty} \frac{h(n)}{n^s} = \frac{u_Q^s}{\zeta_Q(s)\phi_s(u_Q)},$$

*the series being absolutely convergent, where*

$$\zeta_Q(s) = \sum_{\substack{n=1\\n\in(Q)}}^{\infty} \frac{1}{n^s} \qquad \text{and} \qquad \phi_s(n) = \sum_{d|n} d^s \mu(n/d).$$

*Proof.* The absolute convergence of the series follows at once by $|h(n)| \leq 1$ for every $n \in \mathbb{N}$ and by $s > 1$. Note that $\zeta_Q(s) = \prod_{p\in Q}(1 - \frac{1}{p^s})^{-1}$, where $s > 1$, see [T97], lemma 3. The function $h$ is multiplicative, hence we can use the Euler product formula:

$$\sum_{\substack{n=1\\(n,u)=1}}^{\infty} \frac{h(n)}{n^s} = \prod_{\substack{p\nmid u\\p\in Q}}(1 - \frac{1}{p^s}) = \prod_{p\in Q}(1 - \frac{1}{p^s}) \prod_{\substack{p|u\\p\in Q}}(1 - \frac{1}{p^s})^{-1} = \frac{u_Q^s}{\zeta_Q(s)\phi_s(u_Q)}.$$

□

**Lemma 3.** *If $h$ is the function defined in Lemma 1, $u \in \mathbb{N}$ and $s > 1$, then*

$$\sum_{\substack{n=1\\(n,u)=1}}^{\infty} \frac{h(n)\log n}{n^s} = \frac{u_Q^s}{\zeta_Q(s)\phi_s(u_Q)}(\alpha_s(u_Q) + \frac{\zeta_Q'(s)}{\zeta_Q(s)}),$$

*where $\zeta_Q'$ is the derivative of $\zeta_Q$ and*

$$\alpha_s(u) = \sum_{p|u} \frac{\log p}{p^s - 1};$$

*Proof.* The series is uniformly convergent for $s \geq 1 + \varepsilon > 1$, with $\varepsilon > 0$ and the formula is obtained by termwise differentiation with respect to $s$ of the series given in Lemma 2, see [Sur69], lemma 2.5 for $h = \mu$. □

We need the following result due by B. Gordon and K. Rogers [GR64], lemma 3, see also [SSur78], lemma 2.

**Lemma 4.** *If $u \in \mathbb{N}$, then*

$$\sum_{\substack{n\leq x\\(n,u)=1}} \tau(n) = (\frac{\phi(u)}{u})^2 x(\log x + 2C - 1 + 2\alpha(u)) + O(\sqrt{x}S(u)),$$

*where $\varphi \equiv \phi_1$ is the Euler totient function, $C$ is Euler's constant and*

$$\alpha(u) \equiv \alpha_1(u) = \sum_{p|u} \frac{\log p}{p-1}, \qquad S(u) = \sum_{d|u} \frac{3^{\omega(d)}}{\sqrt{d}}.$$

**Lemma 5.** *If $A$ is a cross-convolution and the set $Q$ is finite, then for $s > 0$ we have*

$$\sum_{\substack{n<x \\ n\in\overline{(Q)}}} \frac{1}{n^s} = O(1), \tag{i}$$

$$\sum_{\substack{n>x \\ n\in(Q)}} \frac{1}{n^s} = O(\frac{1}{x^s}), i \tag{i}$$

$$\sum_{\substack{n>x \\ n\in(Q)}} \frac{\log n}{n^s} = O(\frac{\log x}{x^s}). ii \tag{i}$$

*Proof.* For every $p \in Q$ define $\alpha_p \in \mathbb{N}$ such that $p^{\alpha_p} \leq x < p^{\alpha_p+1}$. Then

$$\sum_{\substack{n<x \\ n\in\overline{(Q)}}} \frac{1}{n^s} \leq \prod_{p\in Q}(1+\frac{1}{p^s}+\frac{1}{p^{2s}}+...+\frac{1}{p^{\alpha_p s}}) < \prod_{p\in Q}(1-\frac{1}{p^s})^{-1},$$

which proves formula (i). Furthermore, let $Q = \{p_1, p_2, ..., p_t\}$, where $p_1 < p_2 < ... < p_t$. Then

$$\sum_{\substack{n>x \\ n\in(Q)}} \frac{1}{n^s} = \sum_{n=p_1^{a_1}...p_t^{a_t}>x} \frac{1}{(p_1^{a_1}...p_t^{a_t})^s} \leq \sum_{2^{a_1+...+a_t}>x} \frac{1}{2^{s(a_1+...+a_t)}},$$

where $p_1^{a_1}...p_t^{a_t} \geq p_1^{a_1+...+a_t} \geq 2^{a_1+...+a_t}$ and $2^{a_1+...+a_t} > x$ implies $p_1^{a_1}...p_t^{a_t} > x$.

Denoting $a_1 + ... + a_t = a \in \mathbb{N}$ and $a(x) = \log x/\log 2$ we have

$$\sum_{\substack{n>x \\ n\in(Q)}} \frac{1}{n^s} \leq \sum_{2^a>x} \frac{1}{2^{sa}} = \sum_{a>a(x)} \frac{1}{2^{sa}} = \frac{1}{(2^s-1)2^{s[a(x)]}} = O(\frac{1}{x^s}),$$

where $[a(x)]$ stands for the integer part of $a(x)$, and it follows (ii). Formula (iii) can be proved in the same way. □

## 2. Main results

Now we are ready to prove the following asymptotic formula.

**Theorem 1.** *If $A$ is a cross-convolution and $u \in \mathbb{N}$, then*

$$\sum_{\substack{n<x \\ (n,u)=1}} \tau_A(n) = (\frac{\phi(u)}{u})^2 \frac{u_Q^2 x}{\zeta_Q(2)\phi_2(u_Q)}(\log x + 2C - 1 + 2\alpha(u) - 2\beta(u_Q) - 2\frac{\zeta_Q'(2)}{\zeta_Q(2)})$$

$$+O(H(x,Q)S(u)),$$

*where*

$$\beta(u) \equiv \alpha_2(u) = \sum_{p|u} \frac{\log p}{p^2-1},$$

*and* $H(x,Q)=\sqrt{x}$ *(Q finite),* $\sqrt{x}\log x$ *(Q infinite).*

*Proof.* Using Lemmas 1 and 4 we deduce

$$\sum_{\substack{n\le x\\(n,u)=1}} \tau_A(n) = \sum_{\substack{d^2e=n\le x\\(d^2e,u)=1}} h(d)\tau(e) = \sum_{\substack{d\le\sqrt{x}\\(d,u)=1}} h(d) \sum_{\substack{e\le x/d^2\\(e,u)=1}} \tau(e)$$

$$= \sum_{\substack{d\le\sqrt{x}\\(d,u)=1}} h(d)\left( (\frac{\phi(u)}{u})^2\frac{x}{d^2}(\log\frac{x}{d^2}+2C-1+2\alpha(u)) + O(S(u)\sqrt{\frac{x}{d^2}})\right)$$

$$= (\frac{\phi(u)}{u})^2 x(\log x+2C-1+2\alpha(u)) \sum_{\substack{d\le\sqrt{x}\\(d,u)=1}} \frac{h(d)}{d^2}$$

$$-2x(\frac{\phi(u)}{u})^2 \sum_{\substack{d\le\sqrt{x}\\(d,u)=1}} \frac{h(d)\log d}{d^2} + O(S(u)\sqrt{x}\sum_{\substack{d\le\sqrt{x}\\d\in(Q)}}\frac{1}{d}).$$

Applying Lemmas 2 and 3 for $s=2$ and the well-known estimates

$$\sum_{n>x}\frac{1}{n^s} = O(x^{1-s}),\quad s>1, \tag{1}$$

$$\sum_{n>x}\frac{\log n}{n^s} = O(x^{1-s}\log x),\quad s>1. \tag{2}$$

we have

$$\sum_{\substack{d\le\sqrt{x}\\(d,u)=1}} \frac{h(d)}{d^2} = \sum_{\substack{d=1\\(d,u)=1}}^{\infty} \frac{h(d)}{d^2} + O(\sum_{\substack{d>\sqrt{x}\\d\in(Q)}}\frac{1}{d^2})$$

$$= \frac{u_Q^2}{\zeta_Q(2)\phi_2(u_Q)} + O(\frac{1}{(\sqrt{x})^m}),$$

$$\sum_{\substack{d\le\sqrt{x}\\(d,u)=1}} \frac{h(d)\log d}{d^2} = \sum_{\substack{d=1\\(d,u)=1}}^{\infty} \frac{h(d)\log d}{d^2} + O(\sum_{\substack{d>\sqrt{x}\\d\in(Q)}}\frac{\log d}{d^2})$$

$$= \frac{u_Q^2}{\zeta_Q(2)\phi_2(u_Q)}(\beta(u_Q)+\frac{\zeta_Q'(2)}{\zeta_Q(2)}) + O(\frac{\log x}{(\sqrt{x})^m}),$$

where $m=2$ if $Q$ is finite, see Lemma 5 and $m=1$ if $Q$ is infinite. We also have

$$\sum_{\substack{n\le x\\n\in(Q)}}\frac{1}{n} = \begin{cases} O(1), & \text{if } Q \text{ is finite},\\ O(\log x), & \text{otherwise},\end{cases}$$

see Lemma 5/(i), which completes the proof. □

In the unitary case we reobtain the formula proved by M. V. SUBBARAO and D. SURYANARAYANA [SSur78], lemma 3 for the function $\tau^*$.

For $u = 1$ we obtain the following result.

**Theorem 2.** *If $A$ is a cross-convolution, then*

$$\sum_{n\leq x} \tau_A(n) = \frac{x}{\zeta_Q(2)}(\log x + 2C - 1 - 2\frac{\zeta'_Q(2)}{\zeta_Q(2)}) + O(H(x,Q)),$$

*where $H(x,Q)$ is defined by Theorem 1.*

For $A = D$ we reobtain Dirichlet's well-known formula and for $A = U$ we have the formula established by Mertens, see E. COHEN [Co60].

The remainder term the above formula can be improved for $A = U$ into $O(\sqrt{x})$ using analytical methods, see A. A. GIOIA and A. M. VAIDYA [GV66].

## References


[1] E.Cohen, *The number of unitary divisors of an integer*, Amer. Math. Monthly **67**(1960), 879-880.

[2] A. A. Gioia, A. M. Vaidya, *The number of squarefree divisors of an integer*, Duke Math. J. **33**(1966), 797-799.

[3] B. Gordon, K. Rogers, *Sums of the divisor function*, Canadian J. Math. **16**(1964), 151-158.

[4] P. J. McCarthy, *Introduction to arithmetical functions*, Springer- Verlag, New York, Berlin, Heidelberg, Tokyo, 1986.

[5] W. Narkiewicz, *On a class of arithmetical convolutions*, Colloq. Math. **10**(1963), 81-94.

[6] V. Sita Ramaiah, *Arithmetical sums in regular convolutions*, J. Reine Angew. Math. **303/304**(1978), 265-283.

[7] M. V. Subbarao, D. Suryanarayana, *Sums of the divisor and unitary divisor functions*, J. Reine Angew. Math. **302**(1978), 1-15.

[8] D. Suryanarayana, *The greatest divisor of n which is prime to k*, Math. Student **37**(1969), 147-157.

[9] L. Tóth, *Contributions to the theory of arithmetical functions defined by regular convolutions* (Romanian), thesis, "Babeş-Bolyai" University, Cluj-Napoca, 1995.

[10] L. Tóth, *Asymptotic formulae concerning arithmetical functions defined by cross-convolutions, I. Divisor-sum functions and Euler-type functions*, Publ. Math. Debrecen **50**(1997), 159-176.

[11] L. Tóth, P. Haukkanen, *A generalization of Euler's $\phi$-function with respect to a set of polynomials*, Ann. Univ. Sci. Budap. Rolando Eötvös, Sect. Math. **39**(1996), 69-83.



FACULTY OF MATHEMATICS AND COMPUTER SCIENCE "BABEŞ-BOLYAI" UNIVERSITY STR. M. KOGĂLNICEANU 1 RO-3400 CLUJ-NAPOCA ROMANIA

*E-mail address*: ltoth@math.ubbcluj.ro



BIBLIOTECA FACULTĂŢII
CLUJ-NAPOCA
DE MATEMATICĂ

În cel de al XLII-lea an (1997) *STUDIA UNIVERSITATIS BABEȘ-BOLYAI* apare în următoarele serii:

matematică (trimestrial)
informatică (semestrial)
fizică (semestrial)
chimie (semestrial)
geologie (semestrial)
geografie (semestrial)
biologie (semestrial)
filosofie (semestrial)
sociologie (semestrial)
politică (anual)
efemeride (anual)
studii europene (semestrial)
business (semestrial)
psihologie-pedagogie (semestrial)
științe economice (semestrial)
științe juridice (semestrial)
istorie (trei apariții pe an)
filologie (trimestrial)
teologie ortodoxă (semestrial)
teologie catolică (anual)
educație fizică (anual)

In the XLII-th year of its publication (1997) *STUDIA UNIVERSITATIS BABEȘ-BOLYAI* is issued in the following series:

mathematics (quarterly)
computer science (semesterily)
physics (semesterily)
chemistry (semesterily)
geology (semesterily)
geography (semesterily)
biology (semesterily)
philosophy (semesterily)
sociology (semesterily)
politics (yearly)
ephemerides (yearly)
european studies (semesterily)
business (semesterily)
psychology - pedagogy (semesterily)
economic sciences (semesterily)
juridical sciences (semesterily)
history (three issues per year)
philology (quarterly)
orthodox theology (semesterily)
catholic theology (yearly)
physical training (yearly)

Dans sa XLII-e année (1997) *STUDIA UNIVERSITATIS BABEȘ-BOLYAI* paraît dans les séries suivantes:

mathématiques (trimestriellement)
informatiques (semestriellement)
physique (semestriellement)
chimie (semestriellement)
géologie (semestriellement)
géographie (semestriellement)
biologie (semestriellement)
philosophie (semestriellement)
sociologie (semestriellement)
politique (annuel)
ephemerides (annuel)
études européennes (semestriellement)
affaires (semestriellement)
psychologie - pédagogie (semestriellement)
études économiques (semestriellement)
études juridiques (semestriellement)
histoire (trois apparitions per année)
philologie (trimestriellement)
théologie orthodoxe (semestriellement)
théologie catholique (annuel)
éducation physique (annuel)