NEW TAYLOR-LIKE EXPANSIONS FOR FUNCTIONS OF TWO VARIABLES AND ESTIMATES OF THEIR REMAINDERS

S.S. DRAGOMIR, FENG QI, G. HANNA AND P. CERONE

ABSTRACT. In this article, a generalisation of Sard's inequality for Appell polynomials is obtained. Estimates for the remainder are also provided.

1. INTRODUCTION

Let \( x \in [a, b] \) and \( y \in [c, d] \). If \( f(x, y) \) is a function of two variables we shall adopt the following notation for partial derivatives of \( f(x, y) \):

\[
\begin{align*}
f^{(i,j)}(x, y) & \triangleq \frac{\partial^{i+j} f(x, y)}{\partial x^i \partial y^j}, \\
f^{(0,0)}(x, y) & \triangleq f(x, y), \\
f^{(i,j)}(\alpha, \beta) & \triangleq f^{(i,j)}(x, y)|_{(x,y)=(\alpha,\beta)}
\end{align*}
\]

for \( 0 \leq i, j \in \mathbb{N} \) and \( (\alpha, \beta) \in [a, b] \times [c, d] \).

A. H. Stroud has pointed out in [6] that one of the most important tools in the numerical integration of double integrals is the following Taylor's formula [6, p. 138 and p. 157] due to A. Sard [5]:

---

2000 Mathematics Subject Classification: Primary 26D10, 26D15; Secondary 41A55.

Key words and phrases: Appell polynomials, Bernoulli polynomials, Euler polynomials, generalized Taylor's formula, Taylor-like expansion, double integral, remainder.

The second author was supported in part by NNSF (\#10001016) of China, SF for the Prominent Youth of Henan Province, SF of Henan Innovation Talents at Universities, NSF of Henan Province (\#004051800), SF for Pure Research of Natural Science of the Education Department of Henan Province (\#1999110004), Doctor Fund of Jiaozuo Institute of Technology, China.

© The Korean Society for Industrial and Applied Mathematics, 2005
Theorem A. If \( f(x, y) \) satisfies the condition that all the derivatives \( f^{(i,j)}(x, y) \) for \( i + j \leq m \) are defined and continuous on \([a, b] \times [c, d] \), then \( f(x, y) \) has the expansion

\[
f(x, y) = \sum_{i+j \leq m} \frac{(x-a)^i}{i!} \frac{(y-c)^j}{j!} f^{(i,j)}(a, c) \\
+ \sum_{j < q} \frac{(y-c)^j}{j!} \int_a^x \frac{(x-u)^{m-j-1}}{(m-j-1)!} f^{(m-j,j)}(u, c) \, du \\
+ \sum_{i < p} \frac{(x-a)^i}{i!} \int_c^y \frac{(y-v)^{m-i-1}}{(m-i-1)!} f^{(i,m-i)}(a, v) \, dv \\
+ \int_a^x \int_c^y \frac{(x-u)^{p-1}}{(p-1)!} \frac{(y-v)^{q-1}}{(q-1)!} f^{(p,q)}(u, v) \, dv \, du,
\]

where \( i, j \) are nonnegative integers; \( p, q \) are positive integers; and \( m \triangleq p + q \geq 2 \).

Essentially, the representation (2) is used for obtaining the fundamental Kernel Theorems and Error Estimates in numerical integration of double integrals [6, p. 142, p. 145 and p. 158] and has both theoretical and practical importance in the domain as a whole.

Definition 1. A sequence of polynomials \( \{P_i(x)\}_{i=0}^\infty \) is called harmonic [4] if it satisfies the recursive formula

\[
P_i'(x) = P_{i-1}(x)
\]

for \( i \in \mathbb{N} \) and \( P_0(x) = 1 \).

A slightly different concept that specifies the connection between the variables is the following one.

Definition 2. We say that a sequence of polynomials \( \{P_i(t, x)\}_{i=0}^\infty \) satisfies the Appell condition [2] if

\[
\frac{\partial P_i(t, x)}{\partial t} = P_{i-1}(t, x)
\]

and \( P_0(t, x) = 1 \) for all defined \((t, x)\) and \( n \in \mathbb{N} \).

It is wellknown that the Bernoulli polynomials \( B_i(t) \) can be defined by the following expansion

\[
\frac{x e^{tx}}{e^x - 1} = \sum_{i=0}^\infty \frac{B_i(t)}{i!} x^i, \quad |x| < 2\pi, \quad t \in \mathbb{R}.
\]

It can be shown that the polynomials \( B_i(t) \), \( i \in \mathbb{N} \), are uniquely determined by the two formulae

\[
B_i'(t) = iB_{i-1}(t), \quad B_0(t) = 1;
\]

and

\[
B_i(t + 1) - B_i(t) = it^{i-1}.
\]
The Euler polynomials can be defined by the expansion

\begin{equation}
\frac{2e^{tx}}{e^x + 1} = \sum_{i=0}^{\infty} \frac{E_i(t)}{i!} x^i, \quad |x| < \pi, \quad t \in \mathbb{R}.
\end{equation}

It can also be shown that the polynomials $E_i(t)$, $i \in \mathbb{N}$, are uniquely determined by the two properties

\begin{align}
E'_i(t) &= i E_{i-1}(t), \quad E_0(t) = 1; \\
\text{and} \quad E_i(t + 1) + E_i(t) &= 2t^i.
\end{align}

For further details about Bernoulli polynomials and Euler polynomials, please refer to [1, 23.1.5 and 23.1.6].

There are many examples of Appell polynomials. For instance, for $\lambda$ a nonnegative integer, $\theta \in \mathbb{R}$ and $\lambda \in [0, 1],$

\begin{align}
P_{i,\lambda}(t) &\triangleq P_{i,\lambda}(t; x; \theta) = \frac{[t - (\lambda \theta + (1 - \lambda)x)]^i}{i!}, \\
P_{i,B}(t) &\triangleq P_{i,B}(t; x; \theta) = \frac{(x - \theta)^i}{i!} B_i \left( \frac{t - \theta}{x - \theta} \right) \quad ([4]), \\
P_{i,E}(t) &\triangleq P_{i,E}(t; x; \theta) = \frac{(x - \theta)^i}{i!} E_i \left( \frac{t - \theta}{x - \theta} \right) \quad ([4]).
\end{align}

In [4], the following generalized Taylor’s formula was established.

**Theorem B.** Let $\{P_i(x)\}_{i=0}^{\infty}$ be a harmonic sequence of polynomials. Further, let $I \subset \mathbb{R}$ be a closed interval and $a \in I$. If $f : I \to \mathbb{R}$ is any function such that $f^{(n)}(x)$ is absolutely continuous for some $n \in \mathbb{N}$, then, for any $x \in I$, we have

\begin{equation}
f(x) = f(a) + \sum_{k=1}^{m} (-1)^{k+1} \left[ P_k(x)f^{(k)}(x) - P_k(a)f^{(k)}(a) \right] + R_n(f; a, x),
\end{equation}

where

\begin{equation}
R_n(f; a, x) = (-1)^n \int_a^x P_n(t)f^{(n+1)}(t) \, dt.
\end{equation}

The fundamental aim of this article is to obtain a generalisation of the Taylor-like formula (2) for Appell polynomials and to study its impact on the numerical integration of double integrals.

### 2. Two New Taylor-like Expansions

Following a similar argument to the proof of Theorem 2 in [4], we obtain the following result.
Theorem 1. If \( g : [a, b] \to \mathbb{R} \) is such that \( g^{(n-1)} \) is absolutely continuous on \([a, b]\), then we have the generalised integration by parts formula for \( x \in [a, b] \)

\[
\int_a^b g(t) \, dt = \sum_{k=1}^n (-1)^{k+1} \left[ P_k(b, x)g^{(k-1)}(b) - P_k(a, x)g^{(k-1)}(a) \right] 
+ (-1)^n \int_a^b P_n(t, x)g^{(n)}(t) \, dt.
\]

(16)

Proof. By integration by parts we obtain, on using the Appell condition (4),

\[
(-1)^n \int_a^b P_n(t, x)g^{(n)}(t) \, dt 
= (-1)^n P_n(t, x)g^{(n-1)}(t) \bigg|_a^b 
+ (-1)^{n-1} \int_a^b P_{n-1}(t, x)g^{(n-1)}(t) \, dt 
= (-1)^n \left[ P_n(b, x)g^{(n-1)}(b) - P_n(a, x)g^{(n-1)}(a) 
- \int_a^b P_{n-1}(t, x)g^{(n-1)}(t) \, dt \right].
\]

Clearly, the same procedure can be used for the term \( \int_a^b P_{n-1}(t, x)g^{(n-1)}(t) \, dt \). Therefore, formula (16) follows from successive integration by parts. \( \blacksquare \)

Theorem 2. Let \( D \) be a domain in \( \mathbb{R}^2 \) and the point \((a, c) \in D\). Also, let \( \{P_i(t, x)\}_{i=0}^\infty \) and \( \{Q_j(s, y)\}_{j=0}^\infty \) be two Appell polynomials. If \( f : D \to \mathbb{R} \) is such that \( f^{(i,j)}(x, y) \) are continuous on \( D \) for all \( 0 \leq i \leq m \) and \( 0 \leq j \leq n \), then

\[
f(x, y) = f(a, c) + C(f, P_m, Q_n) + D(f, P_m, Q_n) + S(f, P_m, Q_n) + T(f, P_m, Q_n),
\]

where

\[
C(f, P_m, Q_n) = \sum_{i=1}^m (-1)^{i+1} \left[ P_i(x, x)f^{(i,0)}(x, c) - P_i(a, x)f^{(i,0)}(a, c) \right] 
+ \sum_{j=1}^n (-1)^{j+1} \left[ Q_j(y, y)f^{(0,j)}(a, y) - Q_j(c, y)f^{(0,j)}(a, c) \right],
\]

(19)

\[
D(f, P_m, Q_n)
= \sum_{i=1}^m \sum_{j=1}^n (-1)^{i+j} P_i(x, x) \left[ Q_j(y, y)f^{(i,j)}(x, y) - Q_j(c, y)f^{(i,j)}(x, c) \right] 
- \sum_{i=1}^m \sum_{j=1}^n (-1)^{i+j} P_i(a, x) \left[ Q_j(y, y)f^{(i,j)}(a, y) - Q_j(c, y)f^{(i,j)}(a, c) \right],
\]

(20)

\[
S(f, P_m, Q_n)
\]

(21)
\[ = (-1)^m \int_a^x P_m(t, x) f^{(m+1,0)}(t, c) \, dt + (-1)^n \int_c^y Q_n(s, y) f^{(0,n+1)}(a, s) \, ds \]

\[ + \sum_{i=1}^m (-1)^{n+i+1} \int_c^y Q_n(s, y) \left[ P_i(x, x) f^{(i,n+1)}(x, s) - P_i(a, x) f^{(i,n+1)}(a, s) \right] \, ds \]

\[ + \sum_{j=1}^n (-1)^{m+j+1} \int_a^x P_m(t, x) \left[ Q_j(y, y) f^{(m+1,j)}(t, y) - Q_j(c, y) f^{(m+1,j)}(t, c) \right] \, dt \]

and

\[ (22) \quad T(f, P_m, Q_n) = (-1)^{m+n} \int_a^x \int_c^y P_m(t, x) Q_n(s, y) f^{(m+1,n+1)}(t, s) \, ds \, dt. \]

**Proof.** Let \( P_m(t, x) \) be an Appell polynomial. Applying formula (14) to the function \( f(x, y) \) with respect to variable \( x \) yields

\[ f(x, y) = f(a, y) + \sum_{i=1}^m (-1)^{i+1} \left[ P_i(x, x) f^{(i,0)}(x, y) - P_i(a, x) f^{(i,0)}(a, y) \right] \]

\[ + (-1)^m \int_a^x P_m(t, x) f^{(m+1,0)}(t, y) \, dt. \]

Similarly, for the functions \( f^{(i,0)}(x, y) \), \( f^{(i,0)}(a, y) \), \( f^{(m+1,0)}(t, y) \) and \( f(a, y) \), we have

\[ f^{(i,0)}(x, y) = f^{(i,0)}(x, c) + (-1)^n \int_c^y Q_n(s, y) f^{(i,n+1)}(x, s) \, ds \]

\[ + \sum_{j=1}^n (-1)^{j+1} \left[ Q_j(y, y) f^{(i,j)}(x, y) - Q_j(c, y) f^{(i,j)}(x, c) \right], \]

\[ f^{(i,0)}(a, y) = f^{(i,0)}(a, c) + (-1)^n \int_c^y Q_n(s, y) f^{(i,n+1)}(a, s) \, ds \]

\[ + \sum_{j=1}^n (-1)^{j+1} \left[ Q_j(y, y) f^{(i,j)}(a, y) - Q_j(c, y) f^{(i,j)}(a, c) \right], \]

\[ f^{(m+1,0)}(t, y) = f^{(m+1,0)}(t, c) + (-1)^n \int_c^y Q_n(s, y) f^{(m+1,n+1)}(a, s) \, ds \]

\[ + \sum_{j=1}^n (-1)^{j+1} \left[ Q_j(y, y) f^{(m+1,j)}(t, y) - Q_j(c, y) f^{(m+1,j)}(t, c) \right], \]

\[ f(a, y) = f(a, c) + (-1)^n \int_c^y Q_n(s, y) f^{(0,n+1)}(a, s) \, ds \]

\[ + \sum_{j=1}^n (-1)^{j+1} \left[ Q_j(y, y) f^{(0,j)}(a, y) - Q_j(c, y) f^{(0,j)}(a, c) \right]. \]
Substituting formulae (24)–(27) into (23) produces

\[
f(x, y) = f(a, c) + \sum_{i=1}^{m} (-1)^{i+1} [P_i(x, x)f^{(i,0)}(x, c) - P_i(a, x)f^{(i,0)}(a, c)]
\]

\[+ \sum_{j=1}^{n} (-1)^{j+1} [Q_j(y, y)f^{(0,j)}(a, y) - Q_j(c, y)f^{(0,j)}(a, c)]
\]

\[+ \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{i+j} P_i(x, x)[Q_j(y, y)f^{(i,j)}(x, y) - Q_j(c, y)f^{(i,j)}(x, c)]
\]

\[- \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{i+j} P_i(a, x)[Q_j(y, y)f^{(i,j)}(a, y) - Q_j(c, y)f^{(i,j)}(a, c)]
\]

(28) \[+ (-1)^m \int_{a}^{x} P_m(t, x)f^{(m+1,0)}(t, c) dt + (-1)^n \int_{c}^{y} Q_n(s, y)f^{(0,n+1)}(a, s) ds
\]

\[+ \sum_{i=1}^{m} (-1)^{n+i+1} \int_{c}^{y} Q_n(s, y)[P_i(x, x)f^{(i,n+1)}(x, s) - P_i(a, x)f^{(i,n+1)}(a, s)] ds
\]

\[+ \sum_{j=1}^{n} (-1)^{m+j+1} \int_{a}^{x} P_m(t, x)[Q_j(y, y)f^{(m+1,j)}(t, y) - Q_j(c, y)f^{(m+1,j)}(t, c)] dt
\]

\[+ (-1)^{m+n} \int_{a}^{x} \int_{c}^{y} P_m(t, x)Q_n(s, y)f^{(m+1,n+1)}(t, s) ds. dt.
\]

The proof of Theorem 2 is complete. 

**Remark 1.** If we take

(29) \[P_i(t, x) = P_{m, \lambda}(t, x; a), \quad Q_j(s, y) = Q_{j, \mu}(s, y; c)
\]

for 0 \(\leq i \leq m\), 0 \(\leq j \leq n\) and \(\lambda, \mu \in [0, 1]\) in Theorem 2, then the expressions simplify to give, on using (11),

(30) \[C(f, P_m, Q_n) = \sum_{i=1}^{m} \frac{(x - a)^i}{i!} [ (1 - \lambda)^i f^{(i,0)}(a, c) + \lambda^i f^{(i,0)}(x, c)]
\]

\[+ \sum_{j=1}^{n} \frac{(y - c)^j}{j!} [ (1 - \mu)^j f^{(0,j)}(a, c) + \mu^j f^{(0,j)}(a, y)]
\],

(31) \[D(f, P_m, Q_n) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\lambda^i(x - a)^j(y - c)^j}{i! \cdot j!} [\mu^j f^{(i,j)}(x, y) + (1 - \mu)^j f^{(i,j)}(x, c)]
\]

\[- \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(1 - \lambda)^i(x - a)^j(y - c)^j}{i! \cdot j!} [\mu^j f^{(i,j)}(a, y) + (1 - \mu)^j f^{(i,j)}(a, c)],
\]
\[ S(f, P_m, Q_n) = (-1)^m \int_a^x \frac{[t - (\lambda a + (1 - \lambda)x)]^m}{m!} f^{(m+1,0)}(t,c) \, dt \]
+ \[ (-1)^n \int_c^y \frac{[s - (\mu c + (1 - \mu)y)]^n}{n!} f^{(0,n+1)}(a,s) \, ds \]
+ \[ \sum_{i=1}^m \int_c^y \frac{[\mu c + (1 - \mu)y - s]^n(x - a)^i}{n! \cdot i!} [(\lambda - 1)^i f^{(i,n+1)}(a,s) - \lambda^i f^{(i,n+1)}(x,s)] \, ds \]
+ \[ \sum_{j=1}^n \int_a^x \frac{[\lambda a + (1 - \lambda)x - t]^m(y - c)^j}{m! \cdot j!} [(\mu - 1)^j f^{(m+1,j)}(t,c) - \mu^j f^{(m+1,j)}(t,y)] \, dt, \]

and

\[ T(f, P_m, Q_n) = \]
\[ \int_a^x \int_c^y \frac{[(\lambda a + (1 - \lambda)x - t)^m[(\mu c + (1 - \mu)y - s)^n]}{m! \cdot n!} f^{(m+1,n+1)}(t,s) \, ds \, dt. \]

Notice that, taking \( \lambda = 0 \) and \( \mu = 0 \) in (29), then we can deduce Theorem A from Theorem 2.

Other choices of Appell type polynomials will provide generalizations of Theorem A.

The following approximation of double integrals in terms of Appell polynomials holds.

**Theorem 3.** Let \( \{P_i(t,x)\}_{i=0}^\infty \) and \( \{Q_j(s,y)\}_{j=0}^\infty \) be two Appell polynomials and \( f : [a,b] \times [c,d] \subset \mathbb{R}^2 \to \mathbb{R} \) such that \( f^{(i,j)}(x,y) \) are continuous on \( [a,b] \times [c,d] \) for all \( 0 \leq i \leq m \) and \( 0 \leq j \leq n \). We then have

\[ \int_a^b \int_c^d f(t,s) \, ds \, dt = A(f, P_m, Q_n) + B(f, P_m, Q_n) + R(f, P_m, Q_n), \]

where

\[ A(f, P_m, Q_n) = \]
\[ \sum_{i=1}^m \sum_{j=1}^n (-1)^{i+j} P_i(a,b) [Q_j(d,d) f^{(i-1,j-1)}(a,d) - Q_j(c,d) f^{(i-1,j-1)}(a,c)] \]
\[ - \sum_{i=1}^m \sum_{j=1}^n (-1)^{i+j} P_i(b,b) [Q_j(d,d) f^{(i-1,j-1)}(b,d) - Q_j(c,d) f^{(i-1,j-1)}(b,c)], \]

and

\[ B(f, P_m, Q_n) = \]
\[ \int_a^b \int_c^d f(t,s) \, ds \, dt. \]
\[ B(f, P_m, Q_n) = \sum_{j=1}^{n} (-1)^j Q_j(c, d) \int_a^b f^{(0,j-1)}(t, c) \, dt \]
\[ \quad - \sum_{j=1}^{n} (-1)^j Q_j(d, d) \int_a^b f^{(0,j-1)}(t, d) \, dt \]
\[ \quad + \sum_{i=1}^{m} (-1)^i P_i(a, b) \int_c^d f^{(i-1,0)}(a, s) \, ds \]
\[ \quad - \sum_{i=1}^{m} (-1)^i P_i(b, b) \int_c^d f^{(i-1,0)}(b, s) \, ds \]

and

\[ R(f, P_m, Q_n) = (-1)^{m+n} \int_a^b \int_c^d P_m(t, b)Q_n(s, d) f^{(m,n)}(t, s) \, ds \, dt. \]

**Proof.** Using the generalized integration by parts formula consecutively yields
\[
\int_a^b \int_c^d P_m(t, b)Q_n(s, d) f^{(m,n)}(t, s) \, ds \, dt \\
= \int_a^b P_m(t, b) \left[ \int_c^d Q_n(s, d) f^{(m,n)}(t, s) \, ds \right] \, dt \\
= (-1)^m \int_a^b P_m(t, b) \left\{ \int_c^d f^{(m,0)}(t, s) \, ds \\
\quad + \sum_{j=1}^{n} (-1)^j \left[ Q_j(d, d) f^{(m,j-1)}(t, d) - Q_j(c, d) f^{(m,j-1)}(t, c) \right] \right\} \, dt \\
= (-1)^m \int_a^b \int_c^d P_m(t, b) f^{(m,0)}(t, s) \, ds \, dt \\
\quad + \sum_{j=1}^{n} (-1)^m+1 Q_j(d, d) \int_a^b P_m(t, b) f^{(m,j-1)}(t, d) \, dt \\
\quad - \sum_{j=1}^{n} (-1)^m+1 Q_j(c, d) \int_a^b P_m(t, b) f^{(m,j-1)}(t, c) \, dt \\
= (-1)^m \int_c^d (-1)^n \left\{ \int_a^b f(t, s) \, dt \\
\quad + \sum_{i=1}^{m} (-1)^i \left[ P_i(b, b) f^{(i-1,0)}(b, s) - P_i(a, b) f^{(i-1,0)}(a, s) \right] \right\} \, ds \\
\quad + \sum_{j=1}^{n} (-1)^{m+j} Q_j(d, d) \left\{ (-1)^m \left[ \int_a^b f^{(0,j-1)}(t, d) \, dt \right] \right\}
\]
\[
+ \sum_{i=1}^{m} (-1)^i \left( P_i(b, b) f^{(i-1,j-1)}(b, d) - P_i(a, b) f^{(i-1,j-1)}(a, d) \right) \right] \\
- \sum_{j=1}^{n} (-1)^{n+j} Q_j(c, d) \left\{ (-1)^m \left[ \int_a^b f^{(0,j-1)}(t, c) \, dt \right. \\
+ \sum_{i=1}^{m} (-1)^i \left( P_i(b, b) f^{(i-1,j-1)}(b, c) - P_i(a, b) f^{(i-1,j-1)}(a, c) \right) \right\} \\
= (-1)^{m+n} \int_a^b \int_c^d f(t, s) \, ds \, dt \\
+ \sum_{i=1}^{m} (-1)^{m+n+i} \int_c^d \left[ P_i(b, b) f^{(i-1,0)}(b, s) - P_i(a, b) f^{(i-1,0)}(a, s) \right] \, ds \\
+ \sum_{j=1}^{n} (-1)^{m+n+j} Q_j(d, d) \int_a^b f^{(0,j-1)}(t, d) \, dt \\
+ \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{m+n+i+j} P_i(b, b) Q_j(d, d) f^{(i-1,j-1)}(b, d) \\
- \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{m+n+i+j} P_i(a, b) Q_j(d, d) f^{(i-1,j-1)}(a, d) \\
- \sum_{j=1}^{n} (-1)^{m+n+j} Q_j(c, d) \int_a^b f^{(0,j-1)}(t, c) \, dt \\
+ \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{m+n+i+j} P_i(a, b) Q_j(c, d) f^{(i-1,j-1)}(a, c) \\
- \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{m+n+i+j} P_i(b, b) Q_j(c, d) f^{(i-1,j-1)}(b, c) \\
= (-1)^{m+n} \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{i+j} P_i(b, b) \left[ Q_j(d, d) f^{(i-1,j-1)}(b, d) \\
- Q_j(c, d) f^{(i-1,j-1)}(b, c) \right] \\
+ (-1)^{m+n} \sum_{i=1}^{m} \sum_{j=1}^{n} (-1)^{i+j} P_i(a, b) \left[ Q_j(c, d) f^{(i-1,j-1)}(a, c) \\
- Q_j(d, d) f^{(i-1,j-1)}(a, d) \right] \\
+ (-1)^{m+n} \sum_{i=1}^{m} \left( -1)^i P_i(b, b) \int_c^d f^{(i-1,0)}(b, s) \, ds \right]
\]
\[-(-1)^{m+n} \sum_{i=1}^{m} (-1)^j P_i(a, b) \int_c^d f^{(i-1,0)}(a, s) \, ds \]
\[+ (-1)^{m+n} \sum_{j=1}^{n} (-1)^j Q_j(d, d) \int_a^b f^{(0,j-1)}(t, d) \, dt \]
\[- (-1)^{m+n} \sum_{j=1}^{n} (-1)^j Q_j(c, d) \int_a^b f^{(0,j-1)}(t, c) \, dt \]
\[+ (-1)^{m+n} \int_a^b \int_c^d f(t, s) \, ds \, dt. \]

The proof of Theorem 3 is complete.

**Remark 2.** As usual, let \( B_i, i \in \mathbb{N} \), denote Bernoulli numbers.

From properties (6) and (7), (9) and (10) of the Bernoulli and Euler polynomials respectively, we can easily obtain that, for \( i \geq 1 \),

\[(38) \quad B_{i+1}(0) = B_{i+1}(1) = B_{i+1}, \quad B_1(0) = -B_1(1) = -\frac{1}{2}, \]

and, for \( j \in \mathbb{N} \),

\[(39) \quad E_j(0) = -E_j(1) = -\frac{2}{j+1}(2^{j+1} - 1)B_{j+1}. \]

It is also a well known fact that \( B_{2i+1} = 0 \) for all \( i \in \mathbb{N} \).

As an example, taking \( P_i(t, x) = P_{i,B}(t, x; a) \) and \( Q_j(s, y) = P_{j,E}(s, y; c) \) from (12) and (13) for \( 0 \leq i \leq m \) and \( 0 \leq j \leq n \) in Theorem 3 and using (38) and (39) yields

\[(40) \quad A(f, P_m, Q_n) = \sum_{i=1}^{m} \sum_{j=2}^{n} \frac{(a - b)^i(c - d)^j}{i! \cdot j!} \cdot \frac{2(2^{j+1} - 1)}{j + 1} B_iB_{j+1} \]
\[\times [f^{(i-1,j-1)}(a, d) + f^{(i-1,j-1)}(a, c) - f^{(i-1,j-1)}(b, d) - f^{(i-1,j-1)}(b, c)] \]
\[+ (b - a) \sum_{i=1}^{m} \frac{(2^{i+1} - 1)(c - d)^i}{(i + 1)!} B_{j+1} \]
\[\times [f^{(i-1,0)}(a, d) + f^{(i-1,0)}(a, c) + f^{(i-1,0)}(b, d) + f^{(i-1,0)}(b, c)]. \]
\[ B(f, P_m, Q_n) = \]
\[ 2 \sum_{j=1}^{n} \frac{(1 - 2^{j-1})(c - d)^j}{(j + 1)!} B_{j+1} \int_{a}^{b} \left[ f^{(0,j-1)}(t, c) + f^{(0,j-1)}(t, d) \right] dt \]
\[ + \sum_{j=2}^{n} \frac{(a - b)^j}{j!} B_j \int_{c}^{d} \left[ f^{(i-1,0)}(a, s) - f^{(i-1,0)}(b, s) \right] ds \]
\[ + \frac{b - a}{2} \int_{c}^{d} \left[ f(a, s) + f(b, s) \right] ds, \]

and

\[ R(f, P_m, Q_n) = \]
\[ \frac{(a - b)^m(c - d)^n}{m! \cdot n!} \int_{a}^{b} \int_{c}^{d} B_m \left( \frac{t - a}{b - a} \right) E_n \left( \frac{s - c}{d - c} \right) f^{(m,n)}(t, s) ds dt. \]

3. ESTIMATES OF THE REMAINDERS

In this section, we will give some estimates for the remainders of expansions in Theorem 2 and Theorem 3.

We firstly need to introduce some notation.

For a function \( \ell : [a, b] \times [c, d] \to \mathbb{R} \), then for any \( x, y \in [a, b], u, v \in [c, d] \) we define

\[ \| \ell \|_{\ell, \infty} := \text{ess sup} \{ |\ell(t, s)| \}, \]

\[ t \in [x, y] \text{ or } [y, x] \text{ and } s \in [u, v] \text{ or } [v, u] \]

and

\[ \| \ell \|_{\ell, p} := \left( \int_{x}^{y} \int_{u}^{v} |h(t, s)|^p ds dt \right)^{\frac{1}{p}}, \quad p \geq 1. \]

The following result establishing bounds for the remainder in the Taylor-like formula (18) holds.

**Theorem 4.** Assume that \( \{ P_i(t, x) \}_{i=0}^{\infty}, \{ Q_j(s, y) \}_{j=0}^{\infty} \) and \( f \) satisfy the assumptions of Theorem 2. Then we have the representation (18) and the remainder satisfies the estimate

\[ |T(f, P_m, Q_n)| \leq \left\{ \begin{array}{l}
\|P_m(\cdot, x)\|_{[a,x], \infty} \|Q_n(\cdot, y)\|_{[c,y], \infty} \|f^{(m+1,n+1)}\|_{[a,x] \times [c,y], 1}, \\
\|P_m(\cdot, x)\|_{[a,x], p} \|Q_n(\cdot, y)\|_{[c,y], p} \|f^{(m+1,n+1)}\|_{[a,x] \times [c,y], p}, \\
\quad \text{where } p > 1, \quad \frac{1}{p} + \frac{1}{q} = 1; \\
\|P_m(\cdot, x)\|_{[a,x], 1} \|Q_n(\cdot, y)\|_{[c,y], 1} \|f^{(m+1,n+1)}\|_{[a,x] \times [c,y], \infty}. 
\end{array} \right. \]
The proof follows in a straightforward fashion on using Hölder’s inequality applied for the integral representation of the remainder $T(f, P_m, Q_n)$ provided by equation (22). We omit the details.

The integral remainder in the cubature formula (34) may be estimated in the following manner.

**Theorem 5.** Assume that $\{P_i(t, x)\}_{i=0}^{\infty}$, $\{Q_j(s, y)\}_{j=0}^{\infty}$ and $f$ satisfy the assumptions in Theorem 3. Then one has the cubature formula (34) and, the remainder $R(f, P_m, Q_n)$ satisfies the estimate:

$$
|R(f, P_m, Q_n)| \leq \begin{cases}
||P_m(\cdot, b)||_{[a,b],\infty} ||Q_n(\cdot, d)||_{[c,d],\infty} ||f^{(m,n)}||_{[a,b] \times [c,d], 1}, \\
||P_m(\cdot, b)||_{[a,b],p} ||Q_n(\cdot, d)||_{[c,d],p} ||f^{(m,n)}||_{[a,b] \times [c,d], p}, \\
\text{where } p > 1, \frac{1}{p} + \frac{1}{q} = 1; \\
||P_m(\cdot, b)||_{[a,b],1} ||Q_n(\cdot, d)||_{[c,d],1} ||f^{(m,n)}||_{[a,b] \times [c,d],\infty}.
\end{cases}
$$

(44)

**Remark 1.** If we consider the particular instances of Appell polynomials provided by (11), (12) and (13), then a number of particular formulae may be obtained. Their remainder may be estimated by the use of Theorems 4 and 5, providing a 2-dimensional version of the results in [4].

For instance, if we consider from (11),

$$
P_m(\lambda)(t, x; a) = \frac{[t - (\lambda a + (1 - \lambda) x)]^m}{m!}
$$

(45)

$$
Q_n(\mu)(s, y; c) = \frac{[s - (\mu c + (1 - \mu) y)]^n}{n!}
$$

(46)

then we obtain the following result.

**Theorem 6.** Let $\{P_{m,\lambda}(t, x; a)\}_{m=0}^{\infty}$, $\{Q_{n,\mu}(s, y; c)\}_{n=0}^{\infty}$ and $f$ satisfy the assumptions of Theorem 2. Then we have the representation (18) and the remainder satisfies for $a \leq x$, $c \leq y$, the estimate

$$
|T(f, P_{m,\lambda}, Q_{n,\mu})| \leq \begin{cases}
\frac{(x-a)^m(y-c)^n}{m!n!} \lambda \mu \|f^{(m+1,n+1)}\|_{[a,x] \times [c,y], 1}, \\
\frac{1}{m!n!} \left[ \frac{(x-a)^{m+1}(y-c)^{n+1}}{(m+1)(n+1)} \right]^{\frac{1}{q}} \lambda \mu \|f^{(m+1,n+1)}\|_{[a,x] \times [c,y], q}, \\
\text{where } p > 1, \frac{1}{p} + \frac{1}{q} = 1; \\
\frac{(x-a)^{m+1}(y-c)^{n+1}}{(m+1)(n+1)} \lambda \mu \|f^{(m+1,n+1)}\|_{[a,x] \times [c,y],\infty}.
\end{cases}
$$

(47)
where
\[
\lambda_1 = \left[ \lambda^{m+1} + (1 - \lambda)^{m+1} \right],
\]
\[
\lambda_p = \left[ \lambda^{mq+1} + (1 - \lambda)^{mq+1} \right]^\frac{1}{p} \quad \text{and}
\]
\[
\lambda_\infty = \left[ \frac{1}{2} + \left| \lambda - \frac{1}{2} \right| \right]^m.
\]

and similar for \( \mu_1, \mu_p \) and \( \mu_\infty \).

Proof. Utilizing equations (45) and (46) and using Hölder’s inequality for double integrals and the properties of the modulus on equation (22), then we have that

\[
\left| \int_a^x \int_c^y T(f, P_{m,\lambda}, Q_{n,\mu}) \right| = \left| \int_a^x \int_c^y P_{m,\lambda}(t, x; a) \ Q_{n,\mu}(s, y; c) f^{(m+1,n+1)} \ ds \ dt \right|
\leq \int_a^x \int_c^y \left| P_{m,\lambda}(t, x; a) \ Q_{n,\mu}(s, y; c) \right| f^{(m+1,n+1)} \ ds \ dt
\leq \sup_{(t,s)\in[a,x] \times [c,y]} \left| P_{m,\lambda}(t, x; a) \ Q_{n,\mu}(s, y; c) \right| \left\| f^{(m+1,n+1)} \right\|_{[a,x] \times [c,y], 1}
\leq \left\{ \left( \int_a^x \int_c^y \left| P_{m,\lambda}(t, x; a) \ Q_{n,\mu}(s, y; c) \right|^q \ dt \ ds \right)^{\frac{1}{q}} \left\| f^{(m+1,n+1)} \right\|_{[a,x] \times [c,y], p} \right. \quad p > 1, \ \frac{1}{p} + \frac{1}{q} = 1;
\sup_{(t,s)\in[a,x] \times [c,y]} \left| P_{m,\lambda}(t, x; a) \ Q_{n,\mu}(s, y; c) \right| \left\| f^{(m+1,n+1)} \right\|_{[a,x] \times [c,y], \infty}.\]

(48)

Now, the result in equation (48) can be further simplified by the application of equations (45) and (46), given that,
\[
\alpha = (1 - \lambda) \ x + \lambda \ a \quad \text{and} \quad \beta = (1 - \mu) \ y + \mu \ c.
\]

It then follows
\[
\sup_{(t,s)\in[a,x] \times [c,y]} \left| P_{m,\lambda}(t, x; a) \ Q_{n,\mu}(s, y; c) \right|
= \sup_{t\in[a,c]} \left| P_{m,\lambda}(t, x; a) \right| \sup_{s\in[c,y]} \left| Q_{n,\mu}(s, y; c) \right|
= \max \left\{ \frac{(\alpha - a)^m}{m!}, \frac{(x - a)^m}{m!} \right\} \times \max \left\{ \frac{(\beta - c)^n}{n!}, \frac{(y - \beta)^n}{n!} \right\}
= \frac{(x - a)^m (y - c)^n}{m! n!} \left[ \max\{1 - \lambda, \lambda\} \right]^m \times \left[ \max\{1 - \mu, \mu\} \right]^n
= \frac{(x - a)^m (y - c)^n}{m! n!} \left[ \frac{1}{2} + \left| \lambda - \frac{1}{2} \right| \right]^m \times \left[ \frac{1}{2} + \left| \mu - \frac{1}{2} \right| \right]^n.\]
giving the first inequality in (47) where we have used the fact that
\[
\max \{X, Y\} = \frac{X + Y}{2} + \left|\frac{Y - X}{2}\right|
\]
Further, we have
\[
\left( \int_a^x \int_c^y |P_{m,\lambda}(t, x; a)Q_{n,\mu}(s, y; c)|^q \, ds \, dt \right)^{\frac{1}{q}}
\]
\[
= \left( \int_a^x |P_{m,\lambda}(t, x; a)|^q \, dt \right)^{\frac{1}{q}} \left( \int_c^y |Q_{n,\mu}(s, y; c)|^q \, ds \, dt \right)^{\frac{1}{q}}
\]
\[
= \frac{1}{m!n!} \left[ \int_a^\alpha (\alpha - t)^{mq} \, dt + \int_\alpha^x (t - \alpha)^{mq} \, dt \right]^\frac{1}{q}
\times \left[ \int_c^\beta (\beta - s)^{nq} \, ds + \int_\beta^y (s - \beta)^{nq} \, ds \right]^\frac{1}{q}
\]
\[
= \frac{1}{m!n!} \left[ \frac{(x - a)^{mq+1}(y - c)^{nq+1}}{(mq + 1)(nq + 1)} \right]^\frac{1}{q} \lambda_p \mu_p
\]
producing the second inequality in (47).
Finally,
\[
\int_a^x \int_c^y |P_{m,\lambda}(t, x; a)Q_{n,\mu}(s, y; c)| \, dt \, ds
\]
\[
= \int_a^x \frac{(t - \alpha)^m}{m!} \, dt \int_c^y \frac{(s - \beta)^n}{n!} \, ds
\]
\[
= \left[ \int_\alpha^x \frac{(\alpha - t)^m}{m!} \, dt + \int_\alpha^x \frac{(t - \alpha)^m}{m!} \, dt \right] \times \left[ \int_\beta^y \frac{(\beta - s)^n}{n!} \, ds + \int_\beta^y \frac{(s - \beta)^n}{n!} \, ds \right]
\]
\[
= \frac{(x - a)^{m+1}(y - c)^{n+1}}{(m + 1)!(n + 1)!} \left[ (1 - \lambda)^{m+1} + \lambda^{m+1} \right] \times \left[ (1 - \mu)^{n+1} + \mu^{n+1} \right]
\]
gives the last inequality in (47). Thus the theorem is completely proved.

Remark 2. By taking \(\lambda = \mu = 0\) or 1, we recapture the result obtained by G. Hanna et al. in [3].

In a similar fashion, we can state the remainder \(R(f, P_{m,\lambda}, Q_{n,\mu})\) estimate in the cubature formula (34) as in the following
Theorem 7. Let \( \{P_m(t, x; a)\}_{m=0}^{\infty}, \{Q_n(t, x; a)\}_{n=0}^{\infty} \) and \( f \) satisfy the assumptions of Theorem 3, then the remainder \( R(f, P_m, Q_n, \mu) \) estimate in the cubature formula (34) satisfies the following

\[
| R(f, P_m, Q_n, \mu) | \leq \left\{ \begin{array}{l}
\frac{(b-a)^{m(d-c)}}{m!n!} \lambda_{\infty} \mu_{\infty} \| f^{(m,n)} \|_{[a,b] \times [c,d], 1}, \\
\frac{1}{m!n!} \left[ \frac{(b-a)^{m+1}(d-c)^{n+1}}{(m+1)(n+1)} \right]^{\frac{1}{q}} \lambda_p \mu_p \| f^{(m,n)} \|_{[a,b] \times [c,d], q}, \\
\quad \text{where } p > 1, \frac{1}{p} + \frac{1}{q} = 1; \\
\frac{(b-a)^{m+1}(d-c)^{n+1}}{(m+1)(n+1)} \lambda_1 \mu_1 \| f^{(m,n)} \|_{[a,b] \times [c,d], \infty}.
\end{array} \right.
\]

The proof is similar to the one in Theorem 6 applied on the interval \([a, b] \times [c, d]\), and we omit the details.

Acknowledgements. This paper was completed during the second author's visit to the RGMIA between November 1, 2001 and January 31, 2002, as a Visiting Professor with grants from the Victoria University of Technology and Jiaozuo Institute of Technology.

References


S.S. Dragomir, P. Cerone and G. Hanna:
SCHOOL OF COMPUTER SCIENCE AND MATHEMATICS, VICTORIA UNIVERSITY OF TECHNOLOGY, PO BOX 14428, MELBOURNE CITY MC, VICTORIA 8001, AUSTRALIA.
E-mail: sever.dragomir@vu.edu.au (S.S. Dragomir)
E-mail: george@csu.wu.edu.au (G. Hanna)
E-mail: pc@csu.vu.edu.au (P. Cerone)

Feng Qi:
DEPT. OF MATHEMATICS, JIAOZUO INSTITUTE OF TECHNOLOGY, JIAOZUO CITY, HENAN 454000, CHINA
E-mail: qifeng@jzit.edu.cn
THE STABILITY IN AN INCLINED LAYER OF VISCOELASTIC FLUID FLOW OF HYDROELECTRIC NATURAL CONVECTION

A.A. El-Bary

ABSTRACT The problem of the onset stability in an inclined layer of dielectric viscoelastic fluid (Walter's liquid B') is studied. The analysis is made under the simultaneous action of a normal a.c. electric field and the natural convection flow due to uniformly distributed internal heat sources. The power series method used to obtain the eigen value equation which is then solved numerically to obtain the stable and unstable solutions. Numerical results are given and illustrated graphically.

1. INTRODUCTION

The phenomenal growth of energy requirements in recent years has been attracting considerable attention all over the world. This has resulted in a continuous exploration of new ideas and avenues in harnessing various conventional energy sources, such as tidal waves, wind power, geo-thermal energy, etc. It is obvious that in order to utilize geo-thermal energy to a maximum, one should have a complete and precise knowledge of the amount of perturbations needed to generate convection currents in geo-thermal fluid. Also, knowledge of the quantity of perturbations that are essential to initiate convection currents in mineral fluids found in the earth's crust helps one to utilize the minimal energy to extract the minerals. For example, in the recovery of hydro-carbons from underground petroleum deposits, the use of thermal processes is increasingly gaining importance as it enhances recovery. Heat is being injected into the reservoir in the form of hot water or steam or heat can be generated by burning part of the crude in the reservoir. In all such thermal recovery processes, fluid flow takes place through a dielectric medium and convection currents are detrimental.

In technological fields there exist important class of fluid, called non-Newtonian fluid, are also being studied extensively because of their practical applications, such as fluid film lubrication, analysis of polymers in chemical engineering etc. the micropolar fluid is famous case for non-Newtonian fluid as El-Bary [1]. Also, another example for non-Newtonian fluid is viscoelastic fluid. A detailed theoretical investigation has recently begun for the viscoelastic prototype designated liquid B'Walters [2] and Beard and Walters [3]. Many other authors have contributed to the subject. Sen [4] studied the behavior of unsteady free convection flow of a viscoelastic fluid past an infinite porous plate with constant suction. The effects of suction, free oscillations and free convection currents on flow have been studied by Soundalgeker and Patil [5]. Singh and Singh [6] have studied the magnetohydrodynamic flow of viscoelastic fluid past an accelerated plate. The flow of viscoelastic and electrically conducting fluid past an infinite plate has been studied by Sherief and Ezzat [7]. In most of the above applications, the method of solution due to Lighthil [8] and Stuart [9] is utilized.

The method of the matrix exponential, proposed by Ezzat [10-13], which constitutes the basis of the
state space approach of modern control theory is applied to the non-dimensional equations of a viscoelastic fluid flow of hydromagnetic free convection flows.

A temperature gradient applied to a dielectric fluid produces a gradient in the dielectric constant and electrical conductivity. The application of a dc electric field to the results in the accumulation of free charge in the fluid. The free charge build up occurs exponentially in tie with a time constant. This constant is known as the electrical relaxation time. If an ac electric field is applied at frequency much higher than the reciprocal of the electrical relaxation, the free charge does not have time to accumulate. The electrical relaxation times of most dielectric fluids appear to be sufficiently long to make free charge effects negligible at standard power line frequencies is so low that it makes no significant contribution to the temperature field. Furthermore, variations in the body force are so rapid that its mean value can be assumed as the effective value in determining fluid motions, except in the case of fluids of extremely low viscosity. Thus, the case of an ac electric field is more tractable than that of a dc electric field. Turnbull and Melcher [14] and Turnbull [15] have examined the ac case.

An important stability problem is the thermal convection in a thin layer of fluid heated from below. A detailed account of thermal convection in a thin layer of Newtonian fluid heated from below, under varying assumptions, has been given by Chandrasekhar [16]. The problem of the onset of convective instability in an inclined fluid layer including heat sources in the presence of a temperature gradient and an a.c. electric field was studied by Mohamed and et al. [17]. They used the power-series method to obtain the eigenvalue equation that is then solved numerically to obtain the stable and unstable solutions. The stability of viscoelastic conducting liquid heated from below in the presence of a magnetic field is studied by Othman and Ezzat [18]. Ezzat and Othman [19] are studied the effect of a vertical ac electric field on the onset of convective instability in a dielectric micropolar fluid layer form below confined between two horizontal planes under the simultaneous action of the rotation of the system and the vertical temperature gradient.

The purpose of this work is to study the stability of natural convection in an inclined fluid layer (Walter’s liquid B’) with internal heat generation in the presence of an ac electric field.

2. FORMULATION OF THE PROBLEM

We consider an infinite incompressible and dielectric viscoelastic fluid layer confined between two parallel plates which are separated by a distance and inclined from the vertical by an angle $\theta$. It is assumed that the fluid layer is heated internally by a uniform distribution of heat sources and that the two plates are maintained at constant and equal temperatures $T_0$. The plate at $x = -\frac{L}{2}$ is maintained at electric potential ($\phi_1 = 0$), whereas the plate at $x = \frac{L}{2}$ is kept at constant and high alternating potential whose root-mean-square value is $\phi_2$.

Under the foregoing assumptions, the basic equations can be written as [2]

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$$  \hspace{1cm} (1)

$$\rho \left[ \frac{\partial v_i}{\partial t} + v_k \frac{\partial v_i}{\partial x_k} \right] = \rho g - \frac{\partial P}{\partial x_i} + \eta_0 \frac{\partial^2 v_i}{\partial x_k \partial x_k} + f_e - k_o \left( \frac{\partial}{\partial t} \left( \frac{\partial^2 v_i}{\partial x_k \partial x_k} \right) \right)$$

$$+ \nu_m \left( \frac{\partial^3 v_i}{\partial x_m \partial x_k \partial x_k} \right) - \left( \frac{\partial v_i}{\partial x_m} \right) \left( \frac{\partial^2 v_m}{\partial x_k \partial x_k} \right) - 2 \left( \frac{\partial v_i}{\partial x_m} \right) \left( \frac{\partial^2 v_m}{\partial x_k \partial x_k} \right),$$  \hspace{1cm} (2)
\[ \rho c_v \left[ \frac{\partial T}{\partial t} + (\mathbf{v} \cdot \nabla) T \right] = k \nabla^2 T + Q \]  
(3)

\[ \text{div} (\varepsilon \mathbf{E}) = 0 \]  
(4)

and

\[ \text{curl} \mathbf{E} = 0 \quad \text{or} \quad \mathbf{E} = -\nabla \phi \]  
(5)

where, \( \mathbf{v} = (u, v, w) \) is the velocity of the fluid, \( \mathbf{g} = (-g \sin \theta, 0, -g \cos \theta) \) is the gravitational acceleration, \( \rho \) is the mass density, \( P \) is the pressure, \( \eta_0 \) is the limiting viscosity at small rate of shear, \( K_o \) is the elastic constant of Walters’ liquid B’, \( c_v \) is the specific heat at constant volume, \( k \) is the thermal conductivity, \( T \) is the temperature of the fluid, \( Q \) is the heat generation within the fluid per unit volume per unit time, \( \varepsilon \) is the dielectric constant, \( \mathbf{E} = [E_x, 0, 0] \) is the electric field, \( \phi \) is the root-mean-square value of the electric potential and \( f_e \) is the force of electrical origin which may be expressed as Landau [20] in the form,

\[ f_e = \rho_e E - \frac{1}{2} E^2 \nabla \varepsilon + \frac{1}{2} \nabla (\rho \frac{\partial \varepsilon}{\partial \rho} E^2). \]  
(6)

taking into account the fact the free charge density \( \rho_e \) is zero.

If we replace the pressure by

\[ P^* = P - \frac{1}{2} \rho \frac{\partial \varepsilon}{\partial \rho} E^2 \]  
(7)

The electrostriction term disappear from the equation (2), which can be rewritten in the form

\[
\rho \left[ \frac{\partial v_i}{\partial t} + v_k \frac{\partial v_i}{\partial x_k} \right] = \rho g - \frac{\partial P^*}{\partial x_i} + \eta_0 \frac{\partial^2 v_i}{\partial x_k \partial x_k} - \frac{1}{2} E^2 \frac{\partial \varepsilon}{\partial x_i} - K_o \left[ \frac{\partial}{\partial t} \left( \frac{\partial^2 v_i}{\partial x_k \partial x_k} \right) \right] + \nu_m \left( \frac{\partial^3 v_i}{\partial x_m \partial x_k \partial x_k} - \left( \frac{\partial v_i}{\partial x_m} \right) \left( \frac{\partial^2 v_m}{\partial x_k \partial x_k} \right) - 2 \left( \frac{\partial v_m}{\partial x_k} \right) \left( \frac{\partial^2 v_i}{\partial x_m \partial x_k} \right) \right].
\]  
(8)

The boundary conditions

\[ u = v = w = 0 \quad \text{at} \quad x = \pm \frac{\lambda}{2}, \]  
(9)

\[ T = T_0 \quad \text{at} \quad x = \pm \frac{\lambda}{2}, \]  
(10)

\[ \phi = 0 \quad \text{at} \quad x = -\frac{\lambda}{2}, \]  
(11)

\[ \phi = \phi_2 \quad \text{at} \quad x = -\frac{\lambda}{2}, \]  
(12)

The mass density \( \rho \) and the dielectric constant \( \varepsilon \) are assumed to be linearly dependent on temperature as [18]:

\[ \rho = \rho_0 [1 - \alpha (T - T_o)], \quad \alpha > 0 \]  
(13)

\[ \varepsilon = \varepsilon_0 [1 - e (T - T_o)], \quad e > 0 \]  
(14)

where the subscript “0” refers to values at the midplane \( x = 0 \), \( \alpha \) is the coefficient of volume expansion and \( e \) is the coefficient of relative variation of the dielectric constant with temperature.

We first obtained the following steady solutions (denoted by an overbar).

\[ \overline{T} = T_0 + \frac{Q}{2k} \left( \frac{\lambda}{4} - x^2 \right) \]  
(15)
\[ \bar{\rho} = \rho_o \left[ 1 - \frac{\alpha Q \lambda}{2k} \left( \frac{\lambda}{4} - x^2 \right) \right], \]  
\[ \bar{\varepsilon} = \varepsilon_o \left[ 1 - \frac{eQ \lambda}{2k} \left( \frac{\lambda}{4} - x^2 \right) \right], \]  
\[ \bar{P} = P_o^* - \rho_o g \cos \theta \left[ 1 - \frac{\alpha Q \lambda^2}{10k} \right] x + \rho_o g \sin \theta \left[ \left( 1 - \frac{\alpha Q \lambda^2}{8k} \right) x - \frac{\alpha Q}{6k} x^3 \right] \]  
\[ - \frac{1}{2} \bar{E}_x \frac{\partial \bar{\varepsilon}}{\partial x} dx \]  
(18)

Since the flow is assumed to be along the z-axis, \( \bar{w} \) is obtained in the form
\[ \bar{w} = \frac{\alpha g Q}{k \nu} \left[ \frac{\lambda^4}{1920} - \frac{\lambda^2}{80} x^2 + \frac{1}{24} x^4 \right] \cos \theta, \quad \bar{u} = \bar{v} = 0, \]  
(19)
\[ \bar{E}_x = \frac{E_o}{\left[ 1 - \frac{eQ \lambda}{2k} \left( \frac{\lambda}{4} - x^2 \right) \right]}, \quad \bar{E}_x = 0, \quad \bar{E}_z = 0, \]  
(20)
\[ \bar{\phi} = -i \bar{E}_x dx \]  
(21)

where \( P_o^* \) is the pressure at \( z = 0 \) and \( x = 0 \), \( \nu = -\frac{\eta}{\rho_o} \) is the kinematic viscosity, and \( \bar{w} \) and \( \bar{P} \) have been determined under the condition that the total flux of flow across a plane \( z = \) constant is zero \[17\].

Let this initial steady state be slightly perturbed where any physical quantities \( \psi \) can be expressed after perturbation by the simple relation \( \psi = \psi + \psi' \), and prime refers to perturbed quantities. Following the usual steps of linear stability theory we can obtain the following main equations:
\[ \left( \frac{\partial}{\partial t} - \nu \nabla^2 \right) \nabla^2 u' + \bar{w} \frac{\partial}{\partial z} \nabla^2 u' - \frac{d^2 \bar{w}}{dx^2} \frac{\partial u'}{\partial z} + \alpha g \cos \theta \frac{\partial}{\partial z} \left( \frac{\partial T'}{\partial x} \right) - \alpha g \sin \theta \nabla_1^2 T' \]  
\[ - \frac{\bar{E}_x}{\rho_o} \frac{d \bar{\varepsilon}}{dx} \frac{\partial}{\partial x} \nabla_1^2 \phi' + \frac{\varepsilon_o}{\rho_o} \frac{\partial}{\partial x} \bar{E}_x \nabla_1^2 T' + K_o \left[ \frac{\partial}{\partial t} + \bar{w} \frac{\partial}{\partial z} \right] \nabla^4 u' = 0, \]  
(22)
\[ \left( \frac{\partial}{\partial t} - K \nabla^2 \right) T' + \frac{\partial T'}{\partial z} + \frac{d T}{dx} u' = 0, \]  
(23)
\[ \nabla^2 \phi' + e E_o \frac{\partial T'}{\partial x} = 0, \]  
(24)

where \( K = \frac{k}{\rho_o c_v} \) is the thermal diffusivity and \( \nabla_1^2 = \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \) is the two-dimensional Laplacian.

The associated boundary conditions are given by
\[ u' = v' = w' = T' = \phi' = 0 \quad \text{at} \quad x = \pm \frac{\lambda}{2} \]  

(25)

We first rendered Eqs. (22) – (24) and the boundary conditions (25) in a dimensionless form by choosing \( \frac{\lambda}{v}, \frac{\lambda}{\alpha}, \frac{\alpha^2}{\epsilon}, \lambda \) as the units of length, time, velocity, temperature and electrostatic potential respectively and the equations are then simplified in the usual manner by decomposing the solution in terms of normal modes, so that

\[ [u', T', \phi'] = [U(x), \Theta(x), \Phi(x)] \exp[i(\sigma t + i(a_y y + a_z z))] \]

(26)

where "a_y" and "a_z" are the (real) wave numbers in the y and z directions and \( \sigma \) is the complex time constant.

This makes it possible to obtain the following system of equations:

\[
\begin{align*}
&[\sigma - (D^2 - \lambda^2)](D^2 - \lambda^2) U + i a_y R_H \cos \theta [\overline{w_1}(D^2 - \lambda^2)U - D^2 \overline{w_1} U + D \Theta] \\
&+ R_H \lambda^2 \sin \theta \Theta + R_A \lambda^2 \Theta + [\sigma + R_H \cos \theta \overline{w_1}] (D^2 - \lambda^2)^2 U = 0, \\
&[P, \sigma - (D^2 - \lambda^2)] \Theta + P_i [a_z R_H \cos \theta \overline{w_1} \Theta - x U] = 0,
\end{align*}
\]

(27)

and

\[ (D^2 - \lambda^2) \Phi + D \Theta = 0. \]

(29)

The boundary conditions are

\[ U = DU = \Theta = \Phi = 0 \quad \text{at} \quad x = \pm \frac{1}{2} \]

(30)

where,

\[ P_r = \frac{\nu}{K} \quad \text{Prandtl number}, \]

(31)

\[ R_H = \frac{\alpha g \mu \lambda}{\nu^2 k} \quad \text{heat Rayleigh number}, \]

(32)

\[ R_A = \frac{\epsilon_0 \epsilon^2 E^2 Q^2 \lambda^4}{\rho_o k^2 \nu^2} \quad \text{electric Rayleigh number}, \]

(33)

\[ K_o = \frac{K_o}{\lambda^2} \quad \text{the dimensionless elastic constant} \]

(34)

\[ \overline{w_1} = \frac{1}{1920} - \frac{1}{80} x^2 + \frac{1}{24} x^4 \quad \text{the dimensionless velocity}, \]

(35)

and \( D \) denotes differentiation with respect to \( x \).

To describe three-dimensional disturbances, it is convenient to introduce the parameters

\[ \lambda^2 = a_y^2 + a_z^2 \]

(36)

and,

\[ a = \frac{a_z}{\lambda} \]

(37)

then, equations (27) and (28) become

\[
\begin{align*}
&D^2 - \lambda^2 - \sigma \left[D^2 - \lambda^2\right] U - i a \lambda R_H \cos \theta [\overline{w_1}(D^2 - \lambda^2) - D^2 \overline{w_1}] U - i a \lambda R_H \cos \theta D \Theta \\
&- R_H \lambda^2 \sin \theta \Theta - \lambda^2 x R_A [D \Phi + \Theta] - K_o^* [\sigma + R_H \overline{w_1}] (D^2 - \lambda^2)^2 U = 0,
\end{align*}
\]

(38)
\[ [D^2 - \lambda^2 - P, \sigma] \Theta - P, [i a \lambda R_H \cos \theta \bar{w}_1, \Theta - x U] = 0 \]  
(39)

It should be noted that the value of the parameter "a" is within the range \(0 \leq a \leq 1\). The case \(a = 0\) corresponds to the case \(a_x = 0\) (i.e. longitudinal rolls); the case \(a = 1\) corresponds to the case \(a_y = 0\) (i.e. transverse rolls).

3. Solution

Equations (38) and (39) can be rewritten in the form:

\[
\begin{align*}
(b_1 + b_2 x^2 + b_3 x^4) D^4 U -(c_1 - c_2 x^2 + c_3 x^4) D^2 U + (f_1 - f_2 x^2 + f_3 x^4) U \\
- S D \Theta - F \Theta - \Gamma x (\Theta + D \Phi) &= 0 , \\
D^2 \Theta - (h_1 + h_2 x^2 + h_3 x^4) \Theta + P, x U &= 0 ,
\end{align*}
\]

where,

\[
\begin{align*}
S &= i a \lambda R_H \cos \theta \\
F &= \lambda^2 R_A \\
\Gamma &= \lambda^2 R_H \\
b_1 &= [1 - K_o^*(\sigma + \frac{1}{1920} R_H \cos \theta)] , \\
b_2 &= \frac{1}{80} K_o^* R_H \cos \theta , \\
b_3 &= -\frac{1}{24} K_o^* R_H \cos \theta , \\
c_1 &= 2 \lambda^2 + \sigma + \frac{1}{1920} S - 2 \lambda^2 K_o^* (\sigma - \frac{1}{1920} R_H \cos \theta) , \\
c_2 &= \frac{1}{80} S + \frac{1}{40} \lambda^2 K_o^* R_H \cos \theta , \\
c_3 &= \frac{1}{24} S - \frac{1}{12} \lambda^2 K_o^* R_H \cos \theta , \\
f_1 &= \lambda^2 (\lambda^2 + \sigma) + S (\frac{\lambda^2}{1920} - \frac{1}{40}) - \lambda^4 K_o^* (\sigma + \frac{1}{1920} R_H \cos \theta) \\
f_2 &= S (\frac{\lambda^2}{80} - \frac{1}{2}) - \frac{\lambda^4}{80} K_o^* R_H \cos \theta , \\
f_3 &= \frac{\lambda^2}{24} S - \frac{\lambda^4}{24} K_o^* R_H \cos \theta , \\
h_1 &= \lambda^2 + P, (\sigma + \frac{1}{1920} S) , \\
h_2 &= \frac{1}{80} P, S , \\
h_3 &= \frac{1}{24} P, S ,
\end{align*}
\]

The power series method is adopted to solve equations (40), (41) and (29), since this method is much less laborious than other various approximate methods and moreover, it enables one to obtain essentially exact values of the stability condition, unless the product \(\lambda R_H\) is exceedingly large.

Applying this power series method, the general solutions of equations (40), (41) and (29) can be constructed in the form

\[
\begin{align*}
U &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} C_m C_n A(m,n) x^{n-1} , \\
\Theta &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} C_m C_n B(m,n) x^{n-1} , \\
\Phi &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} C_m C_n H(m,n) x^{n-1} ,
\end{align*}
\]
where $C_1$ to $C_8$ are arbitrary constants. The series coefficients $A(m, n)$, $B(m, n)$, and $H(m, n)$ are found from equations (40), (41) and (29) to obey the following recurrence relations:

$$A(m, n) = \delta_{m,n} \text{ for } m = 1 \rightarrow 8, \quad n = 1, 2, 3, 4 \quad (52)$$

$$A(m, n) = \frac{1}{(n-1)(n-2)(n-3)(n-4)b_1} \left\{ [(n-3)(n-4)c_1 A(m, n-2) - f_1 A(m, n-4)] \Delta_{x,n} + \Delta_{y,n} \right\}$$

$$+ \Gamma B(m, n-5) \Delta_{y,n} - [(n-3)(n-4)(n-5)(n-6)b_2 A(m, n-2)$$

$$+ (n-5)(n-6)c_2 A(m, n-4) - f_2 A(m, n-6) - \Gamma B(m, n-6)] \Delta_{y,n} \right\}$$

$$+ (n-7) \Gamma H(m, n-6) \Delta_{y,n} + [(n-7)(n-8)c_3 A(m, n-6) - f_3 A(m, n-8)$$

$$+ (n-5)(n-6)(n-7)(n-8)b_3 A(m, n-4) - (n-8)S B(m, n-7)] \Delta_{y,n} \right\}$$

$$\text{for } m = 1 \rightarrow 8, \quad n = 5, 6, 7, \ldots \quad (53)$$

$$B(m, n) = \delta_{m,n+4} \text{ for } m = 1 \rightarrow 8, \quad n = 1, 2 \quad (54)$$

$$B(m, n) = \frac{1}{(n-1)(n-2)} \left\{ h_1 B(m, n-2) \Delta_{x,n} - [h_2 B(m, n-4) + P, A(m, n-3)] \Delta_{y,n} \right\}$$

$$+ h_3 B(m, n-6) \Delta_{y,n} \right\}$$

$$\text{for } m = 1 \rightarrow 8, \quad n = 3, 4, 5 \quad (55)$$

$$H(m, n) = \delta_{m,n+6} \text{ for } m = 1 \rightarrow 8, \quad n = 1, 2 \quad (56)$$

$$H(m, n) = \frac{1}{(n-1)(n-2)} \left\{ i^2 H(m, n-2) - (n-2) B(m, n-1) \right\} \Delta_{y,n}$$

$$\text{for } m = 1 \rightarrow 8, \quad n = 1, 2 \quad (57)$$

where

$$\delta_{i,j} = 0 \text{ for } i \neq j \quad (58)$$

$$\delta_{i,j} = 1 \text{ for } i = j \quad (59)$$

$$\Delta_{i,j} = 0 \text{ for } i > j \quad$$

$$\Delta_{i,j} = 1 \text{ for } i \leq j \quad$$

Let us now impose the boundary conditions (30) to obtain eight homogeneous algebraic equations for eight unknown constants $C_1$ to $C_8$. The requirement that the determinant of coefficients of $C_1$ to $C_8$ must vanish in order to ensure a nontrivial solution of the form

$$|X(\lambda, m)| = 0 \quad (60)$$

where,

$$X(1, m) = \sum_{n=1}^\infty \lambda^2 A(m, n) \left( \frac{1}{2} \right)^{n-1}, \quad X(2, m) = \sum_{n=1}^\infty \lambda(n-1) A(m, n) \left( \frac{1}{2} \right)^{n-2}$$

$$X(3, m) = \sum_{n=1}^\infty \lambda^2 B(m, n) \left( \frac{1}{2} \right)^{n-1}, \quad X(4, m) = \sum_{n=1}^\infty \lambda H(m, n) \left( \frac{1}{2} \right)^{n-1}$$

$$X(5, m) = \sum_{n=1}^\infty \lambda A(m, n) \left( -\frac{1}{2} \right)^{n-1}, \quad X(6, m) = \sum_{n=1}^\infty \lambda(n-1) A(m, n) \left( -\frac{1}{2} \right)^{n-2}$$

$$X(7, m) = \sum_{n=1}^\infty \lambda B(m, n) \left( -\frac{1}{2} \right)^{n-1}, \quad X(8, m) = \sum_{n=1}^\infty \lambda H(m, n) \left( -\frac{1}{2} \right)^{n-1} \quad (61)$$
4. NUMERICAL RESULTS

The case when $0^\circ < \theta \leq 90^\circ$, $E \neq 0$, $K_o^* = 0$ and $a = 0$ has been studied in [17].

In the present paper the case $0^\circ \leq \theta \leq 90^\circ$, $E \neq 0$, $a = 0$ (i.e. $a_z = 0$ or longitudinal rolls) and $K_o^* \neq 0$ is treated. It is shown that if we fix the values of Prandtl number $P_T$, wave number $\lambda$, heat Rayleigh number $R_H$, the inclination angle $\theta$, the time constant ($\sigma = 1$) and the elastic constant $K_o^*$, one obtains a quadratic equation of $R_A$ with coefficients that depends on $\lambda$ as a parameter. This equation can be solved numerically to obtain the critical values of $R_A$. The function $R_A(\lambda)$ is illustrated graphically for various fixed values of $K_o^*$, $R_A$ and $\theta$ as parameters. It can be seen that the curves obtained pass through a critical minimum values $R_{Ac}$ corresponds to a critical wave number $\lambda_c$.

Figures 1, 2 and 3 show the plots of $R_A$, when $P_T = 5$, $R_H = 1000$ and $0^\circ \leq \theta \leq 90^\circ$, for the values of $K_o^* = 0$, 0.1, 0.4 and $0 < \lambda < 3$ at the different values of $\theta$. We notice from these figures that as $\theta$ increases the critical values of the electric Rayleigh number increases, which indicates that the angle of the inclination has destabilizing effect but the elastic coefficient $K_o^*$ has stabilizing effect.

![Graph 1](image1.png)  
![Graph 2](image2.png)
Figures 4, 5, and 6 show the plots of $R_{A\theta}$ when $P_r = 5$, $\lambda = 1.8$ and $0^\circ \leq \theta \leq 90^\circ$ for the values of $K^*_o = 0, 0.1, 0.4$ and $100 \leq R_H \leq 1000$. The critical value of electric Rayleigh number decreases as $K^*_o$ increases.
Figures 7 and 8 show that for fixed values of $K_0^*$ and $P_r$, as the heat Rayleigh number increases the critical values of electric Rayleigh number decreases.
5. CONCLUSION

In general one can conclude that:
1. At fixed values of $P_n$, $R_H$ and for various values of $K^*_v$ it is noticed that the angle of inclination about the vertical axis has instability effect.
2. For fixed $P_n$, $R_H$ and $\theta$ it is shows that the elastic coefficient $K^*_v$ has stabilizing effect.

REFERENCES


A.A. El-Bary
Basic and applied science dept.
Arab Academy for Science and Technology
P.O.Box 1029 Alexandria Egypt.
E-mail: aaelbary@aast.edu
A SURFACE RECONSTRUCTION METHOD FOR SCATTERED POINTS ON PARALLEL CROSS SECTIONS

PHILSU KIM

ABSTRACT. We consider a surface reconstruction problem from geometrical points (i.e., points given without any order) distributed on a series of smooth parallel cross sections in $\mathbb{R}^3$. To solve the problem, we utilize the natural points ordering method in $\mathbb{R}^2$, described in [18], which is a method of reconstructing a curve from a set of sample points and is based on the concept of diffusion motions of a small object from one point to the other point. With only the information of the positions of these geometrical points, we construct an acceptable surface consisting of triangular facets using a heuristic algorithm to link a pair of parallel cross sections constructed via the natural points ordering method. We show numerical simulations for the proposed algorithm with some sets of sample points.

1. INTRODUCTION

The problem that we treat is to reconstruct a 3D surface from a set of sample points on a series of parallel planar cross-sections (or called slices) corresponding to different levels. It occurs from various fields, for instance, medical imaging, digitization of objects, and GIS systems and so on, and a lot of algorithms with the work of Keppel [15] as its starting point have been proposed. The problem has been explored in four major directions: (i) Delaunay triangulation method [2, 3, 8], (ii) partial differential equation (PDE) method [4, 5, 19], (iii) optimal method [1, 12, 14, 15], and (iv) heuristic method [6, 13, 16]. The most general method of them is the Delaunay triangulation one, which solves the problem in a $n$-dimensional space for reconstructing a surface from a set of sample points arbitrarily distributed in space. This does not require any particular structures such as that the sample points lie on a series of parallel planar cross-sections, however it is somewhat expensive compared with other methods. The optimal method and heuristic method are more simple and efficient algorithms than the Delaunay triangulation one. According to the number of contours in each cross section and the position of them, these two methodologies have a problem requiring some meticulous care, so called branching problem. The PDE method is very effective one to solve the branching problem.

2000 Mathematics Subject Classification: 35Q30, 76N10.

Key words and phrases: sample points, curve reconstruction, surface reconstruction, ordering, Brownian motion.

© The Korean Society for Industrial and Applied Mathematics, 2005
A common assumption in the methodologies (ii)-(iv) is that the set of sample points on each cross section are well-ordered, that is, a parametrization on each cross section is possible. However, it is far from realistic situations, and hence the problem of reconstructing a curve from a set of scattered sample points on a planar curve is important to reconstruction methods of a surface. A number of related numerical results can be found in [7, 9, 10, 11, 17]. Most of these methods are based on the use of Delaunay triangulation, and give a reasonable curve under a given criterion such as local feature size.

At most recent, the author [18] developed an efficient method, so called a natural points ordering method, for solving the problem of reconstructing a curve from a set of sample points arbitrarily distributed in $\mathbb{R}^2$. It consists of defining the order of the sample points and piecewise edges one by one using only a local decision criterion, so called a natural distance, which is based on a property of the stochastic process (Brownian motion) which models diffusing motions of a small object from one point to the other point. The purpose of this paper is to utilize this algorithm to develop a heuristic method giving an acceptable surface from a set of sample points on a series of parallel planar cross-sections.

The rest of this paper is organized as follows. In the first part of section 2, we precisely describe the problem of interest. We then review the concept of the natural distance [18] and utilize it to reconstruct each cross-section. The second part of section 2 is devoted to solve the tiling problem between two consecutive cross sections. We develop a heuristic method which creates triangular patches sequentially by satisfying a local criterion, so called a smallest inner angle criterion, and then "stitching" all triangulations together.

In section 3, we execute three numerical simulations for the algorithm and shows the potentiality of the present algorithm. Finally, we finish the paper with some conclusions and comments.

2. A HEURISTIC METHOD FOR SOLVING TILING PROBLEM

We begin with this section precisely setting up the problem of interest and then briefly recall the concept of the natural distance and the natural points ordering method to solve the described problem.

2.1. Setting of the problem. Suppose that there are arbitrarily scattered sample points $p_{ij} = (x_{ij}, y_{ij}, z_i)$ on a smooth surface $\Sigma$ in $\mathbb{R}^3$, where $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, N_i$. Denote the set of all sample points by $\mathcal{P}$ and all sample points with level $z = z_i$ as $\mathcal{P}_i$, $i = 1, 2, \ldots, M$, where we assume $z_1 < z_2 < \cdots < z_M$. We call each set $\mathcal{P}_i$ as a cross section or a contour. The problem to solve can then be summarized as the following questions. For given the set $\mathcal{P} = \bigcup_{i=1}^{M} \mathcal{P}_i$ of sample points,

- how can we reconstruct a piecewise interpolating curve from the set $\mathcal{P}_i$ of sample points with only knowledge of the coordinates $p_{ij} = (x_{ij}, y_{ij}, z_j)$?
how can we connect the $p_{ij}$'s and $p_{i+1k}$'s with straight lines in such a way as to form a triangular facet (or called an elementary tile) surface spanning two nested cross sections $P_i$ and $P_{i+1}$? This last question is called a tiling problem.

More concisely, we settle the problem to solve following [13]. A contour segment is a linear approximation of the curve connecting consecutive points in a single cross section. An elementary tile (or triangular patches) is a triangular face composed of a single contour segment and two spans connecting the endpoints of a contour segment with a common point on the adjacent contour. The spans will be designated as "left" and "right" for obvious reasons (see Figure 1). Then the problem of reconstructing a 3D surface from a set of sample points on a series of parallel planar cross-sections is to find a set of elementary tiles which defines a surface satisfying two constraints:

(C1) Each contour segment will appear in exactly one elementary tile.
(C2) If a span appears as the left (right) span of some tile in the set, it will appear as the right (left) span exactly one other tile in the set.

A set of tiles which satisfies these two conditions is called an acceptable surface.

As described in the introduction, the main focus of this paper is to utilize the natural points ordering method reviewed in subsequent subsection and introduce a new heuristic method for solving the described problem above. Therefore, we consider only the simple case that the set $P$ of sample points satisfies the following restrictions:

- Sample points on each set $P_i$ belong to a simple smooth closed curve, which satisfies a local convexity described in Definition 2.1.
- Each cross section $P_i$ consists of only one contour. That is, there are no branching problem.
- For each index $i$, two consecutive cross section $P_i$ and $P_{i+1}$ are so close and similar-shaped that the situations such as Figure 2 do not occur.

2.2. The natural points ordering method. For given two points $p$ and $q$ in $\mathbb{R}^2$ which are already ordered in the direction $\overrightarrow{pq}$, the natural distance is an answer to
the question: for each $r$ which is lying on the opposite side of $p$, centered around the line perpendicular to $\overrightarrow{pq}$ containing $q$ (see Figure 3), how can we construct a distance between two points $q$ and $r$ which reflects the 'smoothness' of the piecewise linear arc $\overrightarrow{qr}$ and the 'closeness' between $q$ and $r$ simultaneously? In order to give a reasonable answer, we considered a small object moving from $q$ in the direction of $\overrightarrow{pq}$ with constant velocity 1 and simultaneously diffusing randomly on the line which is perpendicular to $\overrightarrow{pq}$. A trace of such object corresponds to a graph of sample path of Brownian motion starting from $q$ when we think of $\overrightarrow{pq}$ as a directional vector of time axis. The authors[18] then defined the natural distance $f(r)$ from $q$ to $r$ for directional axis $\overrightarrow{pq}$ as

$$f(r) = t(r) + K \frac{s(r)}{\sqrt{t(r)}},$$

where

$$t(r) = |\overrightarrow{qr}| \cos \theta, \quad s(r) = |\overrightarrow{qr}| \sin \theta,$$

where $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$ is the signed angle between $\overrightarrow{pq}$ and $\overrightarrow{qg}$ defined by $\cos \theta = \frac{\overrightarrow{pq} \cdot \overrightarrow{qg}}{|\overrightarrow{pq}| |\overrightarrow{qg}|}$ (see Figure 3).

Then the first factor $t(r)$ was construed as the diffusion time of the small object from $q$ to $r$ for directional vector $\overrightarrow{pq}$ and call it the time distance, while the second factor $\frac{s(r)}{\sqrt{t(r)}}$ as the transition density measuring how smoothly moves the small object from $q$ to $r$, and the probability that the small object moves from $q$ to $r$ with reaching time $t(r)$, and call it the standardized probability distance. For a detailed illustration, one refers to the paper [18]. We call $K$ the subjective weight, since it represents how much weight is given to the second factor. That is, the larger the subjective weight $K$ is, the more sensitive than the natural distance it is to the magnitude of the change of the probability distance $\frac{s(\cdot)}{\sqrt{t(\cdot)}}$ than that of the time distance $t(\cdot)$. 

\[ \text{Figure 2. (a) Far away situation (b) Dissimilar situation} \]
A SURFACE RECONSTRUCTION METHOD FOR SCATTERED POINTS

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure3.png}
\caption{New coordinates for a scattered point $r$ with $\theta > 0$ based on two initial points $p$ and $q$.}
\end{figure}

\begin{table}
\centering
\caption{An algorithm for the natural points ordering method for given sample points}
\begin{tabular}{|l|}
\hline
Step 0. Assumption: A set of sample points $S$ and the first and second ordered points $p_1$ and $p_2$ are already given. \\
Step 1. Let $i = 2$ and let $C_{i+1}$ be the set of candidate points of $p_{i+1}$ consisting of the points located on the opposite side of $p_{i-1}$ centering around the line perpendicular to $\overrightarrow{p_i-1p_i}$ containing $p_i$. \\
Step 2. Set $\tilde{S}$ to be an ordered set with $p_1$, $p_2$. \\
Step 3. Initialize the starting direction vector $\overrightarrow{p_i-1p_i}$ and calculate it and $|\overrightarrow{p_i-1p_i}|$. \\
Step 4. For each candidate point $r \in C_{i+1}$, calculate $\overrightarrow{qr}$, $t(r)$ and $s(r)$. \\
Step 5. Find the next point $p_{i+1}$ such that $p_{i+1} = \text{argmin}_{r \in C_{i+1}} f(r)$. \\
Step 6. Append the point $p_{i+1}$ to $\tilde{S}$. \\
Step 7. Set $i$ to be $i + 1$ and find the set $C_{i+1}$ of candidate points. \\
Step 8. If $C_{i+1}$ is nonempty then go to Step 3, otherwise stop. \\
\hline
\end{tabular}
\end{table}

Using the natural distance described above, the authors [18] developed a method to solve the curve reconstruction problem in $\mathbb{R}^2$ so called the natural points ordering method. It provides an algorithm to us for consecutive myopic choices of sample points in a natural way. In fact, it is designed to choose the points where the natural distance from the starting point (chosen one step before) is minimized. The algorithm can be summarized as in Table 1.

In Table 1, for a unique set $\tilde{S}$ of ordered points, it is sufficient to fulfill the assumption that in each steps, the new coordinate representations $(t(r), s(r))$ for each candidate points in $C_*$ are different one another.

\subsection{2.3. Tiling method.}\ To describe the heuristic method we use, suppose we are given sample points distributed on two consecutive cross sections $\mathcal{P}_i$ and $\mathcal{P}_{i+1}$ as in section 1
and these set are well-ordered by the algorithm given in Table 1. Since both cross-sections are closed, we may think of the points $p_{ij}$ as being extended periodically; i.e. $p_{iN_j + j} = p_{ij}, j = 1, 2, \cdots, N_1$. Hereafter, we assume that two consecutive cross sections $\mathcal{P}_i$ and $\mathcal{P}_{i+1}$ are on a same plane without any special mention.

**Definition 2.1.** For given two consecutive cross sections $\mathcal{P}_i$ and $\mathcal{P}_{i+1}$, we say that a discrete set $\mathcal{P}_{i+1}$ has a local convexity with respect to a discrete set $\mathcal{P}_i$ provided the following conditions are satisfied: For each contour segment $L_k$ connecting consecutive points $p_{ik}$ and $p_{ik+1}$ in $\mathcal{P}_i$ with direction $\overrightarrow{p_{ik}p_{ik+1}}$, let $c_k$ be the mid point of $p_{ik}$ and $p_{ik+1}$, and $D_r(c_k)$ the closed disk with center $c_k$ and radius $r$. Further we denote $D^+_r(c_k)$ as the left hand side of $L_k$ and $D^-_r(c_k)$ is as the remaining portion of $D_r(c_k)$ (see Figure 4 (a)). Then there exists a positive number $\delta$ such that for any index $k$,

(i) either $D^+_\delta(c_k) \cap \mathcal{P}_{i+1} \neq \emptyset$ or $D^-_\delta(c_k) \cap \mathcal{P}_2 \neq \emptyset$; and

(ii) furthermore, the piecewise linear approximation of the nonempty set is convex (see Figure 4 (b)).

Here, we say the positive number $\delta$ as a local convexity radius.

Hereafter, for each $i$, we assume that the cross-section $\mathcal{P}_{i+1}$ has a local convexity with respect to $\mathcal{P}_i$ with sufficiently small local convexity radius $\delta$.

For simplicity of descriptions, let $i = 1$ hereafter, and $k$ be a fixed index. For each $p_{2j} \in \mathcal{P}_2$, let $p'_{2j}$ be the end point of the projection vector of $\overrightarrow{p_{1k}p_{2j}}$ on the ray containing the contour segment $L_k$ (See Figure 5). Let $E_k$ be a subset of $\mathcal{P}_2$ defined by

$$E_k = \arg\max_{p_{2j} \in D^+_\delta(c_k) \cap \mathcal{P}_2} \frac{\overrightarrow{r_kp_{2j}} \cdot \overrightarrow{p_{2j}p'_{2j}}}{\|p_{2j}p'_{2j}\|}, \quad k = 1, 2, \cdots, N_i.$$ 

Then the set $E_k$ has the following property;
Theorem 2.2. Assume that $\delta$ is the smallest radius such that $D_3^+(c_k) \cap \mathcal{P}_2$ is nonempty and the linear approximation of its elements if exists is convex. Then the set $E_k$ is one of the following cases:

(i) $E_k$ is a set with single element;
(ii) if not the case (i), either the linear approximation of the set $E_k$ is a part of a straight line passing through the point $r_k$ or there are only two points, say $q_{j_1}$ and $q_{j_2}$, such that $q_{j_1}$ and $q_{j_2}$ are symmetric with respect to the line $R_k$.

Proof. If the elements of $E_k$ are one or two, then the assertion holds clearly. So we assume that the number of element of $E_k$ is larger than three and the linear approximation of the set $E_i$ is not a part of a straight line. Then there exist three points $q_1, q_2, q_3 \in \mathcal{P}_2$ such that $q_1, q_2 \in E_k$ and $q_3 \in D_3^+ \cap \mathcal{P}_2 \setminus E_k$. If $|c_kq_3| > |c_kq_j|$, $j = 1, 2$, or $|c_kq_3|$ is larger than one of $|c_kq_j|$, $j = 1, 2$, then either there exist a positive number $\gamma$ such that $D_3^+ \cap \mathcal{P}_2$ is nonempty and the linear approximation of its elements is convex, which is a contradiction to the assumption, or the linear approximation of $q_1, q_2$, and $q_3$ is either concave or a wedge, which is also a contradiction to the assumption. If $|c_kq_3| < |c_kq_j|$, $j = 1, 2$, then the angle between two vectors $\overrightarrow{r_kc_k}$ and $\overrightarrow{r_kq_3}$ is less than that between two vectors $\overrightarrow{r_kc_k}$ and $\overrightarrow{r_kq_1}$. So

$$\frac{\overrightarrow{r_kq_1} \cdot \overrightarrow{q_1q_3}}{|\overrightarrow{r_kq_1}||\overrightarrow{q_1q_3}|} < \frac{\overrightarrow{r_kq_3} \cdot \overrightarrow{q_3q_1}}{|\overrightarrow{r_kq_3}||\overrightarrow{q_3q_1}|},$$

which is a contradiction to the fact $q_1 \in E_k$. \qed

Using this theorem, we now introduce our heuristic method for solving the tiling problem. For the simplicity of the notations, we let $\mathcal{T}_{ij}$ be the $(i+j)$th triangular facet. Then if the triangle $\mathcal{T}_{ij}$ has two vertices in $\mathcal{P}_1$, say $p_{1l}$, $p_{1l+1}$, consecutively, and one vertex in $\mathcal{P}_2$, say $p_{2m}$, (it calls a triangle of Type I), then $i$ and $j$ mean $i = l - 1$ and $j = m$, while if $\mathcal{T}_{ij}$ has two vertices in $\mathcal{P}_2$, say $p_{2l}$, $p_{2l+1}$, consecutively, and one vertex
in $P_1$, say $p_{1m}$, (it calls a triangle of Type II), then $i = m - 1$ and $j = l$. Further, let $T_{kl}$ be the $l$th vertex of the triangle $T_{ij}$ with $k = i + j$. As described in the following, we will directly construct the triangles of Type II from nested two triangles of Type I without any computation. We are now ready to introduce our algorithm.

*Initial step.* To construct first triangular facet $T_{01}$, start with $T_{11} = p_{11}$ and $T_{12} = p_{12}$. In order to choose the third vertex of $T_{01}$, consider the set $E_1$. If $E_1$ is a set with single element, we choose the third vertex $T_{13}$ of the first triangle as the point in $E_1$. If the number of element of $E_1$ is more than one, the above theorem shows that either the linear approximation of the set $E_1$ is a part of a straight line passing through the point $r_1$ or there are only two points, say $p_{2j_1}$ and $p_{2j_2}$, which are symmetric with respect to the line $R_1$.

If the first case occurs, we let $T_{13} = \arg\min_{q \in E_1} |\overrightarrow{q_c}|$, while if the second case occurs, we let $T_{13} = \arg\max_{q \in E_1} \cos^{-1} \left( \frac{\overrightarrow{T_{11}} \cdot \overrightarrow{T_{13}}}{|\overrightarrow{T_{11}}||\overrightarrow{T_{13}}|} \right)$. Once the vertex $T_{13}$ is chosen, then we reorder the ordered set $P_2$ starting with $T_{13}$.

*Middle steps.* Assume that $T_{01}, T_{11}, \ldots, T_{lm_1}, \ldots, T_{ln}, \ldots, T_{ln_1}, \ldots, T_{k-1j}, k \geq 2$, $j \geq 1$ have already been constructed, where $T_{k-1j}$ is a triangle of Type I. Then three vertices of $T_{k-1j}$ are $p_{1k}, p_{1k+1}, p_{2j}$. In order to find a next triangle facet, we let $A = p_{1k+1}, B = p_{1k+2}$ and consider the set $E_{k+1}$. To find an acceptable surface, we reset $E_{k+1}$ by eliminating the points $p_{2c}$ with $c < j$ if it exist. Now let $C = p_{2e}$ be the points in $E_{k+1}$ such that:

- If $E_{k+1}$ is a set with single element, we let $C$ be the point of $E_{k+1}$;
- If the number of element of $E_{k+1}$ is more than one, then either the linear approximation of the set $E_{k+1}$ is a part of a straight line passing through the point $r_{k+1}$ or there are only two points, which are symmetric with respect to the line $R_{k+1}$. If the first case occurs, we let $C$ be the point such that $C = \arg\min_{q \in E_{k+1}} |\overrightarrow{q_{ck+1}}|$, while if the second case occurs, we let $C$ be the point such that $C = \arg\max_{q \in E_{k+1}} \cos^{-1} \left( \frac{\overrightarrow{Aq} \cdot \overrightarrow{AB}}{|\overrightarrow{Aq}||\overrightarrow{AB}|} \right)$.

If $e = j$, we construct the next triangle as $T_{ke}$ with vertices $T_{k+e1} = A, T_{k+e2} = B$ and $T_{k+e3} = C$. When $e > j$, first construct triangles of Type II, $T_{kj}, T_{kj+1}, \ldots, T_{ke-2}, T_{ke-1}$ and then construct a triangle of Type I, $T_{ke}$. Repeat these procedures until the point $C = p_{2N_2+1} = p_{21}$ is chosen. We let the last chosen triangle is $T_{fN_2+1}$. (See Figure 6 (a)).

*Final step.* In the last of the middle steps, if the index $f + 1$ equals to $N_1$, stop the algorithm. Otherwise, construct triangles of Type I, $T_{f+1N_2+1}, \ldots, T_{N_1N_2+1}$. (See Figure 6 (b)).

This algorithm can be summarized as in Table 2.
Remark 2.3. For the fixed $k$, if we let
\[ A(p_{2j}) = \cos^{-1} \left( \frac{\overrightarrow{r_k c_k} \cdot \overrightarrow{r_k p_{2j}}}{|\overrightarrow{r_k c_k}| |\overrightarrow{r_k p_{2j}}|} \right), \quad p_{2j} \in V_0(r_k) \cap \mathcal{P}_2, \]
then the function $A(p_{2j})$ is the angle between two vectors $\overrightarrow{r_k c_k}$ and $\overrightarrow{r_k p_{2j}}$. Further, we can see that
\[ E_k = \arg\min_{p_2 \in V_0(r_k) \cap \mathcal{P}_2} |A(p_2)|. \]
In this sense, we call the above algorithm as a smallest inner angle criterion.

Note that the calculation of the set $E_k$ is based on the problem of finding the location of the point $r_k$. We thus close this section after some discussion for it.

**Theorem 2.4.** For each $k$, if we let $r_k$ be the intersection point with the boundary of $D_{\delta}^+(c_k)$ and the ray $R_k$ such as in see Figure 5, then it can be expressed as
\[ r_k = c_k - \delta \frac{d_j}{|d_j|}, \quad d_j = \frac{\overrightarrow{p_{1k} p_{2j}} \cdot \overrightarrow{p_{1k} p_{1k+1}}}{|\overrightarrow{p_{1k} p_{1k+1}}|^2} \overrightarrow{p_{1k} p_{1k+1}} + \overrightarrow{p_{2j} p_{1k}}, \]
where $p_{2j}$ is any point in the set $D_{\delta}^+(c_k) \cap \mathcal{P}_2$.

**Proof.** Let $p_{2j}$ be a point in $D_{\delta}^+(c_k) \cap \mathcal{P}_2$ and $p'_{2j}$ be its projection onto the line containing the vector $\overrightarrow{p_{1k} p_{1k+1}}$. Then the vector $\overrightarrow{p_{1k} p_{2j}}$ can be expressed as
\[ \overrightarrow{p_{1k} p_{2j}} = \frac{\overrightarrow{p_{1k} p'_{2j}} \cdot \overrightarrow{p_{1k} p_{1k+1}}}{|\overrightarrow{p_{1k} p_{1k+1}}|^2} \overrightarrow{p_{1k} p_{1k+1}}. \]
TABLE 2. An algorithm for the tiling problem with given two well-ordered consecutive cross-sections

Step 0. Assumption: A set of sample points \( P = P_1 \cup P_2 \), where the points in \( P_i \) are well-ordered as \( p_{i1}, \ldots, p_{in_i}, i = 1, 2 \). Let \( \delta \) and \( \theta \) be given real numbers.

Step 1. With the starting points \( p_{11} \) and \( p_{12} \), calculate \( r_1, c_1 \) and \( E_1 \).

Step 2. Find the third vertex \( T_{13} \) in \( E_1 \) for the first triangle \( T_{01} \).

Step 3. Reorder the set \( P_2 \) starting at \( T_{13} \) and let \( j = 1, k = 1 \).

Step 4. Calculate \( r_{k+1}, c_{k+1} \) and \( E_{k+1} \).

Step 5. Reset \( E_{k+1} \) by eliminating the points \( p_{ac} \in E_{k+1} \) such that \( c < j \) if exists.

Step 6. Choose a point \( p_{2c} \in E_{k+1} \) such that:

- if \( E_{k+1} \) is a set with single element, we let \( p_{2c} \) be the point of \( E_{k+1} \);
- if the piecewise linear approximation of the set \( E_{k+1} \) is a part of a straight line passing through the point \( r_{k+1} \), then \( p_{2c} = \arg\min_{q \in E_{k+1}} |q - c_{k+1}| \);
- otherwise \( p_{2c} = \arg\max_{q \in E_{k+1}} \cos^{-1} \left( \frac{\overrightarrow{P_{1k+1}P_{1k+2}} \cdot \overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}}}{|\overrightarrow{P_{1k+1}P_{1k+2}}||\overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}}|} \right) \).

Step 7. If \( j = e \), construct a triangle \( T_{kj} \) of Type I, while if \( e > j \), then first construct triangles \( T_{kj}, T_{kj+1}, \ldots, T_{ke-1} \) of Type II using the points \( p_{2j}, p_{2j+1}, \ldots, p_{2e}, p_{1k+1} \) and then the triangle \( T_{ke} \) of Type I with vertices \( p_{1k+1}, p_{1k+2}, p_{2e} \).

Step 8. Put \( j = e \) and \( k = k + 1 \).

Step 9. If \( e < N_2 + 1 \), then go to Step 4, otherwise go to next step.

Step 10. If \( k + 1 < N_1 + 1 \), construct triangles \( T_{ke}, T_{k+1e}, \ldots, T_{N_1e} \) using the points \( p_{1k+1}, p_{1k+2}, \ldots, p_{1N_1+1}, p_{2j} \) and then stop. Otherwise, stop.

and hence the point \( p'_{2j} \) is given by

\[
p'_{2j} = \frac{\overrightarrow{P_{1k+1}P_{2j}} \cdot \overrightarrow{P_{1k+1}P_{1k+2}}}{|\overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}}|^2} \overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}} + p_{1k}.
\]

Hence if we let \( \overrightarrow{d_j} = p_{2j}p'_{2j} \), we get

\[
\overrightarrow{d_j} = p'_{2j} - p_{2j} = \frac{\overrightarrow{P_{1k+1}P_{2j}} \cdot \overrightarrow{P_{1k+1}P_{1k+2}}}{|\overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}}|^2} \overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}} + p_{1k} - p_{2j}
\]

\[
= \frac{\overrightarrow{P_{1k+1}P_{2j}} \cdot \overrightarrow{P_{1k+1}P_{1k+2}}}{|\overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}}|^2} \overrightarrow{P_{1k+1}P_{1k+1}P_{1k+2}} + p_{2j}p_{1k}.
\]

Since \( r_kc_k = \delta d_j/|d_j| \), we can complete the assertion. \( \square \)

3. SIMULATION RESULTS

As numerical simulations for the proposed heuristic method giving an acceptable surface, in this section, we consider three examples. First, to illustrate our algorithm,
we treat a set of sample points on two simple cross-sections, which are perfectly convex curve in $\mathbb{R}^3$. The second example is a set of sample points on two cross-sections which are not convex, but satisfy the local convexity. Finally, we consider a set of sample points arbitrary distributed on a series of parallel cross sections in the unit half-sphere.

**Example 1.** As shown in Figure 7 (a), the considered set of sample points are scattered on two convex curves. Figure 7 (b) displays the ordered set of sample points on each cross section obtained by the natural points ordering method. Figure 7 (c) shows that the presented algorithm gives an acceptable surface. The numbers in each triangle denote the order of the triangle facets constructed from the algorithm.
Example 2. The sample points to be considered in this example are on two non-convex cross sections as shown in Figure 8 (a). As shown in the contour graph of the cross sections, these two cross sections have similar shape with elbows and hence satisfy the local convexity condition. The proposed algorithm well behavior in this case also and gives an acceptable surface as shown in Figure 8 (b).

Example 3. We consider 20 parallel planner cross sections in the upper hemisphere with radius 1, where we assume that the cross sections are randomly distributed with starting cross section in the plane $z = 0$. If we let the level of each cross section $\mathcal{P}_i$ as $z_i$, the sample points in $\mathcal{P}_i$ are distributed on the level curve $x^2 + y^2 = \sqrt{1 - z_i^2}$ with step size $\pi/(15(1 - z_i)^2 + 10)$ provided $z_i \neq 1$. If $z_i = 1$, the set $\mathcal{P}_i$ is a single set of point $(0, 0, 1)$. Of course, we assume that the sample points in each cross section $\mathcal{P}_i$ are randomly distributed. In Figure 9 (a), we show the set of sample points with its projection onto the plane $z = -1$. In particular, the graph in the projection of the
second figure of Figure 9 (a) is constructed by connecting the sample points with given order, and it shows that the sample points are randomly distributed. The proposed algorithm well behavior in this example also and gives an almost perfect hemisphere surface as shown in Figure 9 (b). Figure 9 (c) is the graph viewed in the bottom direction of the hemisphere.

4. Concluding Remark

We develop a heuristic method for reconstructing 3D surface from sample points arbitrary scattered on a series of parallel planer cross sections, where we assume that consecutive cross sections are very similar. The method is established by using the natural points ordering method developed in [18] and a smallest inner angle criterion.
This provides us a way to reconstruct and the order of the triangle facets simultaneously. By stitching all triangle facets together, we can obtain desired or acceptable surfaces. Numerical simulations say that it can be very efficient to reconstruct an acceptable surface from unorganized sample points distributed on a series of parallel planner cross sections.

Although our focus is on reconstructing acceptable surfaces from sample points distributed on parallel cross sections which are very similar, the method may be extendable to other situations, for instance, parallel cross sections with dissimilar portion, parallel cross sections with branching portion and so on. We conjecture that one efficient way to solve these problems is combining the proposed method and a partial differential equation method technique. That is, one adopts the present algorithm for the similar portion of the cross sections, while for the dissimilar portion or the branching portion, one first models a suitable partial differential equation describing these areas and then solve it.

Acknowledgement This research was supported by Kyungpook National University Research Fund, 2005.

References


Philsu Kim:
Department of Mathematics, Kyungpook National University, Daegu 702-701, Korea
E-mail: kimps@mail.knu.ac.kr
INFLUENCE OF THERMAL CONDUCTIVITY AND VARIABLE VISCOSITY ON THE FLOW OF A MICROPOLAR FLUID PAST A CONTINUOUSLY MOVING PLATE WITH SUCTION OR INJECTION

A. M. Salem* and S. N. Odda**

ABSTRACT This paper investigates the influence of thermal conductivity and variable viscosity on the problem of micropolar fluid in the presence of suction or injection. The fluid viscosity is assumed to vary as an exponential function of temperature and the thermal conductivity is assumed to vary as a linear function of temperature. The governing fundamental equations are approximated by a system of nonlinear ordinary differential equations and are solved numerically by using shooting method. Numerical results are presented for the distribution of velocity, microrotation and temperature profiles within the boundary layer. Results for the details of the velocity, angular velocity and temperature fields as well as the friction coefficient, couple stress and heat transfer rate have been presented.

1. INTRODUCTION

Recently, the study of the dynamics of micropolar fluids has received considerable interest, because of its wide applicability in energy, such as geothermal energy technology, petroleum recovery, glass fiber production, metal extrusion, hot rolling, the cooling and/or drying of paper and textiles, and wire drawing.

Most of the existing analytical studies for this problem are based on the constant physical properties of the ambient fluid [1,2,3]. However, it is known that these properties may change with temperature [4]. To accurately predict the flow and heat transfer rates it is necessary to take into account this variation of viscosity and thermal conductivity. The study of heat transfer and the flow field is necessary for determining the quality of the final products of these processes as explained by Karwe and Jaluria [5,6].

In studying the motion of such a fluid, the non-linearity of the basic equation and additional mathematical difficulties associated with it has led several investigators to explore the perturbation and numerical methods. Hydrodynamic flows of a viscous and incompressible fluid have been studied under different physical conditions with variable fluid properties by Hassani [4] and Seddeek [7]. In many particle engineering system, both the plane surface and the ambient fluid are moving in parallel.

Hence, the aim of the present work is to study the effects of variable viscosity and variable thermal conductivity on heat transfer from moving plate in a steady, incompressible, micropolar fluid in the presence of suction or injection.

© THE KOREAN SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS, 2005
2. Mathematical Formulation:

We consider a steady two-dimensional flow of a micropolar incompressible fluid past a continuously moving plate with suction or injection. The origin is located at the spot through which the plate is drawn in the fluid medium, the x-axis is chosen along the plate and y-axis is taken normal to it. We assume that the fluid properties are isotropic and constant, except for the fluid viscosity \( \mu \), which is assumed to vary as an exponential function of temperature \( T \), in the form

\[
\mu = \mu_0 e^{-\beta_1 \Theta}
\]

(1)

Also, we assume that, the fluid thermal conductivity \( k \) is assumed to vary as a linear function of temperature in the form [8]

\[
k = k_0 (1 + \beta_2 \Theta)
\]

(2)

Where \( \beta_1 \) and \( \beta_2 \) are parameters depending on the nature of the fluid and \( \mu_0, k_0 \) are the thermal diffusivity and viscosity at temperature \( T_w \), respectively.

Under the usual boundary layer approximation, the governing equation for this problem can be written as follows

\[
\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0
\]

(3)

\[
u \frac{\partial u}{\partial x} + \nu \frac{\partial u}{\partial y} = \frac{1}{\rho} \frac{\partial}{\partial y} \left( \mu \frac{\partial u}{\partial y} \right) + k_1 \frac{\partial \Theta}{\partial y}
\]

(4)

\[
G \frac{\partial^2 \Theta}{\partial y^2} - 2 \Theta - \frac{\partial u}{\partial y} = 0
\]

(5)

\[
u \frac{\partial T}{\partial x} + \nu \frac{\partial T}{\partial y} = \frac{1}{\rho c_p} \frac{\partial}{\partial y} \left( k \frac{\partial u}{\partial y} \right) + \frac{v}{c_p} \left( \frac{\partial u}{\partial y} \right)^2
\]

(6)

Subject to the boundary conditions

\[
\begin{align*}
\begin{cases}
\begin{align*}
u &= U_0, & v = V_w, & T = T_w, & \sigma = 0 & \text{at } y = 0, \\
u &\to 0, & T &\to T_\infty, & \sigma &\to 0 & \text{as } y \to \infty
\end{align*}
\end{cases}
\end{align*}
\]

(7)

where \( x \) and \( y \) are the coordinate direction, \( u, v, \sigma \) and \( T \) are the fluid velocity components in the \( x \) and \( y \) directions, the component of microrotation and temperature, respectively. \( k_1, G_1 \) and \( \rho \) are the coupling constant, the microrotation constant and density of the fluid, respectively. \( c_p, U_0, V_w, T_w \) and \( T_\infty \) are the specific heat of the fluid at constant pressure, the uniform velocity of the plate, a non-zero velocity component at the wall, the temperature of the plate and the temperature of the fluid far away from the plate.

The governing equations (3)-(6) can be expressed in a simpler form by introducing the following transformation:
\begin{align*}
\eta = y \sqrt[4]{\frac{U_0}{2v x}}, \quad \psi = \sqrt{2v U_0} x f(\eta), \quad T = (T_w - T_\infty) \Theta(\eta) + T_\infty, \\
\sigma = \sqrt{\frac{U_0^3}{2v x}} g(\eta), \quad u = \frac{\partial \psi}{\partial y}, v = -\frac{\partial \psi}{\partial x}
\end{align*}

(8)

to obtain the ordinary differential equations for the function \( f(\eta), \ g(\eta) \) and \( \Theta(\mu) \)

\begin{align*}
f''' + e^{-1} \left( f f'' + Kg' \right) - \beta_1 f'' \Theta = 0 \\
Gg'' - 2(2g + f''') = 0 \\
\Theta''(1 + \beta_2 \Theta) + pr(Ec f'' + f' \Theta + \beta_2 \Theta'') = 0
\end{align*}

(9) (10) (11)

and the boundary conditions (7) become

at \( \eta = 0 \): \( f = F_w, \ f' = 1, \ \Theta = 1, \ g = 0 \)

as \( \eta \to \infty \): \( f' = \Theta = g = 0 \)

(12)

In the above equations, a prime denotes differentiation with respect to \( \eta \), and

\begin{align*}
K = \frac{k_l}{v}, \quad G = \frac{G_1 U_0}{v x}, \quad pr = \frac{\rho v c_p}{k_o}, \\
Ec = \frac{U_0^2}{c_p(T_w - T_\infty)}, \quad F_w = -V_m \sqrt{\frac{2x}{v U_0}}
\end{align*}

(13)

are the Coupling constant parameter, Microrotation parameter, prandtl number, Eckert number and mass transfer parameter, respectively. Here \( F_w \) is positive for suction and negative for injection.

For micropolar boundary layer flow, the wall skin friction \( \tau_w \) is given by

\[ \tau_w = [(\mu + k) \frac{\partial u}{\partial y} + k \sigma]_{y=0} \]

(14)

The skin friction coefficient \( C_f \) can be defined as

\[ C_f = \frac{\tau_w}{\frac{1}{2} \rho U_0^2} = -2Re_x \frac{1}{2} f''(0), \]

(15)

where \( Re_x = \frac{U_0 x}{v} \) the local Reynolds number.

The local heat flux coefficient (or local Nusselt number) may be written as

\[ Nu_x = \frac{q_{w x}}{k(T_w - T_\infty)} = -\frac{1}{2} Re_x \frac{1}{2} \Theta'(0), q_w = -k \frac{\partial T}{\partial y} \bigg|_{y=0} \]

(16)
3. Results and Discussion

Equations (9), (10) and (11) with boundary conditions (12) can be integrated numerically by the Runge-Kutta method with a systematic guessing of $f''(0)$, $g'(0)$ and $\Theta'(0)$ by the shooting technique. To assess the accuracy of the present method, comparisons between the present results and previously published data[9]. Table I presents the comparison of $f''(0)$, also Table II presents the comparison of the heat transfer rates $-\Theta'(0)$. In fact, this results show a close agreement, hence an encouragement for further study of the effects of other varies of parameters on the continuous moving surface.

<table>
<thead>
<tr>
<th>k</th>
<th>Soundalgekar et al.[9]</th>
<th>Present result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.6291</td>
<td>0.622588</td>
</tr>
<tr>
<td>0.2</td>
<td>0.6225</td>
<td>0.616853</td>
</tr>
</tbody>
</table>

Table II Comparison of the heat transfer rate $-\Theta'(0)$

<table>
<thead>
<tr>
<th>k</th>
<th>Soundalgekar et al.[9]</th>
<th>Present result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G=2</td>
<td>G=4</td>
</tr>
<tr>
<td>0.1</td>
<td>1.944</td>
<td>1.946</td>
</tr>
<tr>
<td>0.5</td>
<td>1.801</td>
<td>-</td>
</tr>
</tbody>
</table>

To study the behavior of the velocity, the angular velocity and the temperature profiles, curves are drawn for various values of the parameters that describe the flow in the case of air ($\text{Pr}=0.733$) and water ($\text{Pr}=3$) at $n_\infty = 7$, $k=0.1$, $G=2$ and $Ec=0.02$.

Figs. 1-3 show the effect of the variable viscosity parameter $\beta_1$ on the velocity, the angular velocity and the temperature distribution, respectively. As shown, the velocity and the angular velocity are decreasing with increasing $\beta_1$, but the temperature increases as $\beta_1$ increases. So, we conclude that for both air and water, the consequence of having a significant temperature dependent viscosity is to produce a marked effect on the temperature field in these convection flows. The temperature variation for both air and water is shown in Fig. 4, for various values of the thermal conductivity $\beta_2$ parameter. Clearly the temperature profiles increases as the thermal conductivity parameter $\beta_2$ increases in two cases.
Fig. 1. Effects of viscosity parameter $\beta_1$ on velocity distribution $f'$.

Fig. 2. Effects of viscosity parameter $\beta_1$ on angular velocity $g$. 
Fig. 3. Effects of viscosity parameter $\beta_1$ on temperature distribution $\Theta$

Fig. 4. Effects of the thermal conductivity parameter $\beta_2$ on temperature distribution $\Theta$
Figs. 5-7 illustrate the velocity, angular velocity and temperature fields, respectively, for different values of porosity parameter $F_w$. This figures indicate that, all of this quantities decreases as $F_w$ increases. It can be seen that the velocity increases monotonically with injection ($F_w < 0$) and decreases with increases in suction ($F_w > 0$). Also, increasing values of the injection parameter move the location of the maximum value of the microrotation away from the surface.

![Graph of velocity distribution](image1)

**Fig. 5.** Effects of porosity parameter $F_w$ on velocity distribution $f'$

![Graph of angular velocity](image2)

**Fig. 6.** Effects of porosity parameter $F_w$ on angular velocity $g$
Fig. 7. Effects of porosity parameter $F_w$ on the temperature distribution $\Theta$

Table III represents values of the skin friction coefficient $f''(0)$, plate couple stress $g'(0)$ and the heat transfer rates $-\Theta'(0)$ for various values of the variable viscosity parameter $\beta_1$, the variable thermal conductivity parameter $\beta_2$ and the porosity parameter $F_w$. It is clear that, with increasing $\beta_1$, $f''(0)$ and $-\Theta'(0)$ decrease and $g'(0)$ increases in two cases (air and water), whereas with increasing $\beta_2$, $f''(0)$, $g'(0)$ and $-\Theta'(0)$ decrease the case of air, but $f''(0)$ and $g'(0)$ increase and $-\Theta'(0)$ decreases as $\beta_2$ increases in the case of water. $g'(0)$ and $-\Theta'(0)$ increase as $F_w$ increases whereas $f''(0)$ decreases. We conclude that for the case of injection ($F_w < 0$), the heat transfer rate was reported to decrease with increased injection.

<table>
<thead>
<tr>
<th>pr</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$F_w$</th>
<th>$f''(0)$</th>
<th>$g'(0)$</th>
<th>$-\Theta'(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.733</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.1</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.485592</td>
<td>0.312835</td>
<td>0.464397</td>
<td></td>
</tr>
<tr>
<td>-0.4</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.372869</td>
<td>0.275746</td>
<td>0.485617</td>
<td></td>
</tr>
<tr>
<td>-0.7</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.262136</td>
<td>0.234901</td>
<td>0.507664</td>
<td></td>
</tr>
<tr>
<td>-0.2</td>
<td>0.5</td>
<td>0.0</td>
<td>-0.552239</td>
<td>0.332948</td>
<td>0.383020</td>
<td></td>
</tr>
<tr>
<td>-0.2</td>
<td>2.0</td>
<td>0.0</td>
<td>-0.556309</td>
<td>0.331002</td>
<td>0.228236</td>
<td></td>
</tr>
<tr>
<td>-0.2</td>
<td>4.0</td>
<td>0.0</td>
<td>-0.558481</td>
<td>0.330390</td>
<td>0.158079</td>
<td></td>
</tr>
<tr>
<td>-0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>-0.241789</td>
<td>0.211229</td>
<td>0.221376</td>
<td></td>
</tr>
<tr>
<td>-0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.551329</td>
<td>0.332630</td>
<td>0.426182</td>
<td></td>
</tr>
<tr>
<td>-0.2</td>
<td>0.2</td>
<td>-0.8</td>
<td>-1.003199</td>
<td>0.452500</td>
<td>0.721355</td>
<td></td>
</tr>
<tr>
<td>3.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.0</td>
<td>-0.666476</td>
<td>0.362577</td>
<td>0.903471</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.5</td>
<td>0.0</td>
<td>-0.813429</td>
<td>0.390704</td>
<td>0.879574</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.5</td>
<td>0.0</td>
<td>-0.982137</td>
<td>0.420909</td>
<td>0.852131</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.717061</td>
<td>0.370385</td>
<td>1.191196</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.713081</td>
<td>0.371713</td>
<td>0.895878</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.710075</td>
<td>0.372670</td>
<td>0.735516</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>-0.8</td>
<td>-0.259679</td>
<td>0.230134</td>
<td>0.203541</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.8</td>
<td>-1.409252</td>
<td>0.499884</td>
<td>2.743576</td>
<td></td>
</tr>
</tbody>
</table>
REFERENCES


A. M. Salem
Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt
E-mail :

S. N. Odda
Department of Mathematics, Faculty of Women, Ain-Shams University, Cairo, Egypt
E-mail :
GENERALIZED DIFFERENCE METHODS FOR ONE–DIMENSIONAL VISCOELASTIC PROBLEMS

HUANRONG LI

ABSTRACT. In this paper, generalized difference methods (GDM) for one–dimensional viscoelastic problems are proposed and analyzed. The new initial values are given in the generalized difference scheme, so we obtain optimal error estimates in $L^p$ and $W^{1,p}(2 \leq p \leq \infty)$ as well as some superconvergence estimates in $W^{1,p}(2 \leq p \leq \infty)$ between the GDM solution and the generalized Ritz–Volterra projection of the exact solution.

1. INTRODUCTION

Consider the following initial boundary value problem for the one–dimensional equation of viscoelasticity:

\begin{align}
(a) \quad & u_{tt} = \frac{\partial}{\partial x} \left( a(x,t) \frac{\partial u_t}{\partial x} + b(x,t) \frac{\partial u}{\partial x} \right) + f(x,t), \quad (x,t) \in (a,b) \times (0,T], \\
(b) \quad & u(a,t) = 0, \quad u(b,t) = 0, \quad t \in [0,T], \\
(c) \quad & u(x,0) = u_0(x), \quad u_t(x,0) = u_1(x), \quad x \in I = [a,b].
\end{align}

(1.1)

where $u_t = \frac{\partial u}{\partial t}$, $u_{tt} = \frac{\partial^2 u}{\partial t^2}$. $a(x,t)$, $b(x,t)$, $f(x,t)$, $u_0(x)$ and $u_1(x)$ are smooth enough to ensure the analysis validity and $a(x,t)$ is bounded from above and below:

\begin{align}
0 < a_0 \leq a(x,t) \leq M, \quad (x,t) \in (a,b) \times [0,T].
\end{align}

(1.2)

Since we shall show that the approximate solution is uniformly convergent to the exact solution of (1.1), the above assumptions only need to hold in a neighborhood of the exact solution.

The problem (1.1) describes many physical processes such as heat transfer with memory\cite{1,2}, gas diffusion\cite{3}, propagation of sound in viscous media\cite{4,5} and fluid dynamics.

The finite element methods to problem (1.1) have been studied by several authors. Y.P. Lin and Cannon\cite{6} demonstrated optimal order error estimates in the $L^2$ norms and $L^p$ norms error estimates in $\mathbb{R}^d (d \leq 4)$. Optimal maximum norm estimates are given by other author. However, the generalized difference methods haven’t been used.


Key words and phrases: Generalized Difference Methods, One–Dimensional Viscoelastic Problems, Optimal Error Estimates, Superconvergence.

© The Korean Society for Industrial and Applied Mathematics, 2005
to deal with the viscoelastic problem (1.1). In fact, the generalized difference methods have the same convergence orders as the corresponding finite element methods, but they require less computational expenses, and keep the mass conservation\([7,8]\).

The aim of this paper is to provide a theory for the generalized difference methods for the two-dimensional problem (1.1) of viscoelasticity. We derive the optimal error estimates in \(L^p\) and \(W^{1,p}\) for \(2 \leq p \leq \infty\). Moreover, some superconvergence is also obtained.

The paper is organized in the following way. In section 2, the new initial values are given and the semi-discrete generalized difference schemes are formulated in piecewise linear finite element spaces. Some important lemmas are introduced in section 3, which are essential in our analysis. Main results of this paper are given in section 4.

### 2. Semi-Discrete Generalized Difference Schemes

In this paper we will follow the notations and symbols in [7]. For examples, \(T_h = \{I_i; I_i = [x_{i-1}, x_i], 1 \leq i \leq n\}\), and \(T_h^* = \{I_i^*; I_i^* = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], 1 \leq i \leq n - 1, I_0^* = [x_0, x_1], I_n^* = [x_{n-\frac{1}{2}}, x_n]\}\) denote the primal partition and its dual partition, respectively. Let \(h_i = x_i - x_{i-1}, h = \max\{h_i; 1 \leq i \leq n\}\). The partitions are assumed to be regular, that is, there exists a constant \(\mu > 0\) such that \(h_i \geq \mu h, i = 1, 2, \ldots, n\).

The trial function space \(U_h \subset H^1_0(I) \equiv \{u \in H^1(I); u(a) = u(b) = 0\}\) is defined as a piecewise linear function space over \(T_h\) and \(U_h = \text{span}\{\varphi_i(x), 1 \leq i \leq n - 1\}\). The test function space \(V_h = \text{span}\{\psi_i(x), 1 \leq i \leq n - 1\} \subset L^2(I)\) is defined as a piecewise constant function space over \(T_h^*\).

For numerical analysis, we need to introduce the interpolation operators \(\Pi_h\) from \(H^1_0(I) \cap C(I)\) to \(U_h\) defined by

\[
\Pi_h w = \sum_{i=1}^{n-1} w(x_i) \varphi_i(x), w \in H^1_0(I),
\]

and \(\Pi_h^* : H^1_0(I) \cap C(I) \mapsto V_h\), defined by

\[
\Pi_h^* w = \sum_{i=1}^{n-1} w(x_i) \psi_i(x), w \in H^1_0(I).
\]

Using the interpolation theory, we have

\[
\begin{align*}
(a) & \quad |w - \Pi_h w|_{m,p} \leq C h^{k-m} |w|_{k,p}, & m = 0, 1, k = 1, 2, 1 \leq p \leq \infty, \\
(b) & \quad \|w - \Pi_h^* w\|_{0,p} \leq C h |w|_{1,p}, & 1 \leq p \leq \infty.
\end{align*}
\]

(2.3)

where \(\cdot|_{m,p}\) and \(\|\cdot\|_{m,p}\) stand for the semi-norm and norm of the Sobolev space \(W^{m,p}(I)\) respectively, \(\cdot|_{m}\) and \(\|\cdot\|_{m}\) stand for the semi-norm and norm of the Sobolev space \(H^m(I) = W^{m,2}(I)\) respectively, and \(C\) is a positive constant independent of \(h\).
Let’s define, for any \( u, v \in H^1_0(I) \), \( u_h \in U_h \) and \( v_h \in V_h \), some bilinear forms as follows:

\[
\begin{align*}
a(u, v) &= \int_a^b a(x, t)u'v'dx; \\
b(u, v) &= \int_a^b b(x, t)u'v'dx; \\
c(u, v) &= b(u, v) - a_t(u, v); \\
a^*(u_h, v_h) &= \sum_{j=1}^{n-1} v_j a^*(u_h, \psi_j); \\
b^*(u_h, v_h) &= \sum_{j=1}^{n-1} v_j b^*(u_h, \psi_j); \\
c^*(u_h, v_h) &= b^*(u_h, v_h) - a_t^*(u_h, v_h),
\end{align*}
\]

(2.4)

where \( a^*(u_h, \psi_j) \) and \( b^*(u_h, \psi_j) \) defined by

\[
\begin{align*}
a^*(u_h, \psi_j) &= a_{j-\frac{1}{2}}u'_h(x_{j-\frac{1}{2}}) - a_{j+\frac{1}{2}}u'_h(x_{j+\frac{1}{2}}), \\
b^*(u_h, \psi_j) &= b_{j-\frac{1}{2}}u'_h(x_{j-\frac{1}{2}}) - b_{j+\frac{1}{2}}u'_h(x_{j+\frac{1}{2}})
\end{align*}
\]

with \( u'_h(x_{j-\frac{1}{2}}) = \frac{u_j - u_{j-1}}{h_j}, u' = \frac{\partial u}{\partial x}, v' = \frac{\partial v}{\partial x}, u_j = u_h(x_j), v_j = v_h(x_j), u_0 = 0, u_n = 0, x_{j-\frac{1}{2}} = \frac{1}{2}(x_{j-1} + x_j), a_{j-\frac{1}{2}} = a(x_{j-\frac{1}{2}}, t), b_{j-\frac{1}{2}} = b(x_{j-\frac{1}{2}}, t) \) and the coefficients of \( a_t(\cdot, \cdot, \cdot), a^*_t(\cdot, \cdot, \cdot) \) and \( b_t(\cdot, \cdot, \cdot), b^*_t(\cdot, \cdot, \cdot) \) which appear in the following are obtained from differentiating the corresponding coefficients of \( a(\cdot, \cdot, \cdot), a^*(\cdot, \cdot, \cdot), b(\cdot, \cdot, \cdot) \) and \( b^*(\cdot, \cdot, \cdot) \) with respect to \( t \), respectively.

The generalized weak form of (1.1) is to find a map \( u(t) : [0, T] \rightarrow H^1_0(I) \), such that

\[
\begin{align*}
(a) \quad (u_t(t), v) + a(u_t, v) + b(u, v) &= (f, v), \quad \forall \, v \in H^1_0(I), \\
(b) \quad u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x), \quad x \in I.
\end{align*}
\]

(2.5)

For error estimates, we next introduce the Ritz projection operator \( R_h = R_h(t) : H^1_0(I) \rightarrow U_h, 0 \leq t \leq T \), defined by

\[
a(u - R_hu, v_h) = 0, \quad \forall \, v_h \in U_h,
\]

(2.6)

the generalized Ritz projection operator \( R^*_h = R^*_h(t) : H^1_0(I) \rightarrow U_h, 0 \leq t \leq T \), defined by

\[
a^*(u - R^*_h u, v_h) = 0, \quad \forall \, v_h \in V_h,
\]

(2.7)

and the generalized Ritz-Volterra projection operator \( V^*_h = V^*_h(t) : H^1_0(I) \rightarrow U_h, 0 \leq t \leq T \), defined by

\[
a^*(u - V^*_h u, v_h) + \int_0^t c^*(u - V^*_h u, v_h) d\tau = 0, \quad \forall \, v_h \in V_h,
\]

(2.8)

Differentiating (2.8) with respect to \( t \), we can obtain the equivalence of (2.8):

\[
\begin{align*}
\left\{ \begin{array}{ll}
  a^*((u - V^*_h u)_t, v_h) + b^*(u - V^*_h u, v_h) = 0, & \forall \, v_h \in V_h, \\
  a^*(u(0) - V^*_h u(0), v_h) = 0, & \forall \, v_h \in V_h.
\end{array} \right.
\]

(2.8')
Obviously, $V_h^*(0) = R_h^*(0)$. Then, the semi-discrete generalized difference schemes of (1.1) is to find a map $u_h(t) : [0, T] \rightarrow U_h$, such that

\begin{align}
(a)\quad & (u_{h,t,t}, v_h) + a^*(u_{h,t}, v_h) + b^*(u_h, v_h) = (f, v_h), \quad \forall \, v_h \in V_h, \\
(b)\quad & u_h(0) = u_{0h}, \quad u_{h,t}(0) = u_{1h}, \quad x \in I.
\end{align}

where $u_{0h} = V_h^* u_0 = R_h^* u_0$, $u_{1h} \in U_h$ satisfies

\[ a^*((u_{1h}, v_h)) = (f(0), v_h) - b^*((u_{0h}, v_h)) - ((V_h^* u)_{tt}(0), v_h), \quad \forall \, v_h \in V_h, \]

here, $(V_h^* u)_{tt}(0)$ satisfies: $\forall \, v_h \in V_h$,

\[ a_t^*((u - V_h^* u)_t(0), v_h) + a^*((u - V_h^* u)_{tt}(0), v_h) + b^*((u - V_h^* u)_t(0), v_h) \]

\[ + b_t^*((u - V_h^* u)(0), v_h) = 0, \]

and $u_{tt}(0) = \frac{\partial}{\partial x} \{a(x, 0) \frac{\partial u}{\partial x} + b(x, 0) \frac{\partial u}{\partial x} \} + f(x, 0)$, $(V_h^* u)_t(0)$ is uniquely determined by

\[ a^*((u - V_h^* u)_t(0), v_h) + b^*((u - V_h^* u)(0), v_h) = 0, \quad \forall \, v_h \in V_h. \]

3. SOME LEMMAS

Noting that for any $u_h \in U_h$, we have, by (2.2)

\[ |u_h|_{1,p} = \left( \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} |u_h'|^p \, dx \right)^{\frac{1}{p}} = \left\{ \sum_{i=1}^{n} h_i \left( \frac{u_i - u_{i-1}}{h_i} \right)^p \right\}^{\frac{1}{p}}. \]

Define some discrete norms in $U_h$:

\[ \|u_h\|_{0,h} = \left\{ \sum_{i=1}^{n} h_i (u_i^2 + u_{i-1}^2) \right\}^{\frac{1}{2}}, \]

\[ |u_h|_{1,h} = \left\{ \sum_{i=1}^{n} \left( \frac{(u_i - u_{i-1})^2}{h_i} \right) \right\}^{\frac{1}{2}}, \]

\[ \|u_h\|_{1,h} = (\|u_h\|_{0,h}^2 + |u_h|_{1,h}^2)^{\frac{1}{2}}. \]

Then we can easily prove the following lemmas.

**Lemma 3.1** (See [7,8]) There exist two positive constants $C_1$ and $C_2$, independent of $h$, such that for any $u_h \in U_h$,

\begin{align}
(a)\quad & |u_h|_{1,h} = |u_h|_1; \\
(b)\quad & C_1 \|u_h\|_{0,h} \leq \|u_h\|_0 \leq C_2 \|u_h\|_{0,h}; \\
(c)\quad & C_1 \|u_h\|_{1,h} \leq \|u_h\|_1 \leq C_2 \|u_h\|_{1,h}.
\end{align}

**Lemma 3.2** (See [7,8]) For $\forall \, u_h, w_h \in U_h$,

\begin{enumerate}
    \item $(u_h, \Pi_h^* w_h) = (w_h, \Pi_h^* u_h)$
    \item Let $||u_h||_0 = (u_h, \Pi_h^* u_h)$, then $|| \cdot ||$ is equivalent to $|| \cdot ||_0$ in $U_h$
\end{enumerate}
According to the technique given in [7,8], it is easy to derive the following conclusions:

**Lemma 3.3** There exist four positive constants \( \alpha, \bar{M}_1, \bar{M}_2 \) and \( M \), independent of \( h \), such that for \( \forall \, u, \, v \in U \),

\[
\begin{align*}
(a) & \quad a(u, u) \geq \tilde{\alpha}||u||_1^2; \\
(b) & \quad |a(u, v)| \leq \bar{M}_1||u||_1||v||_1; \\
(c) & \quad |b(u, v)| \leq \bar{M}_2||u||_1||v||_1; \\
(d) & \quad |c(u, v)| \leq M||u||_1||v||_1.
\end{align*}
\]

**Lemma 3.4** There exist four positive constants \( \alpha, M_1, M_2 \) and \( M \), independent of \( h \), such that

\[
\begin{align*}
(a) & \quad a^*(u_h, \Pi_h^* u) \geq \alpha||u_h||_1^2, \quad \forall \, u_h \in U_h; \\
(b) & \quad |a^*(u_h, \Pi_h^* v_h)| \leq M_1||u_h||_1||v_h||_1, \quad \forall \, u_h, \, v_h \in U_h; \\
(c) & \quad |b^*(u_h, \Pi_h^* v_h)| \leq M_2||u_h||_1||v_h||_1, \quad \forall \, u_h, \, v_h \in U_h; \\
(d) & \quad |c^*(u_h, \Pi_h^* v_h)| \leq M||u_h||_1||v_h||_1, \quad \forall \, u_h, \, v_h \in U_h.
\end{align*}
\]

For simplicity, we set

\[
\begin{align*}
d_1(u - u_h, w_h) &= a(u - u_h, w_h) - a^*(u - u_h, \Pi_h^* w_h), \\
d_2(u - u_h, w_h) &= b(u - u_h, w_h) - b^*(u - u_h, \Pi_h^* w_h).
\end{align*}
\]

We now present a very useful lemma:

**Lemma 3.5** If \( u \in W^{3,p}(I) \), for any \( u_h, \, v_h, \, w_h \in U_h \), we have

\[
\begin{align*}
(a) & \quad |d_1(u - u_h, w_h)| \leq Ch^2(h^{-1}|u - u_h|_{1,p} + |u|_{3,p})||w_h||_{1,p'}; \\
(b) & \quad |d_2(v_h, w_h)| \leq Ch||v_h||_{1,p}||w_h||_{1,p'}; \\
(c) & \quad |d_2(u - u_h, w_h)| \leq Ch^2(h^{-1}|u - u_h|_{1,p} + |u|_{3,p})||w_h||_{1,p'}; \\
(d) & \quad |d_2(v_h, w_h)| \leq Ch||v_h||_{1,p}||w_h||_{1,p'}.
\end{align*}
\]

where \( 1 \leq p \leq \infty, \quad \frac{1}{p} + \frac{1}{p'} = 1. \)

**Lemma 3.6** For any \( u_h, \, v_h, \, w_h \in U_h \), we can get

\[
\begin{align*}
(a) & \quad |a^*(u_h, \Pi_h^* w_h) - a^*(w_h, \Pi_h^* u_h)| \leq Ch||u_h||_1||w_h||_1; \\
(b) & \quad |b^*(u_h, \Pi_h^* w_h) - b^*(w_h, \Pi_h^* u_h)| \leq Ch||u_h||_1||w_h||_1.
\end{align*}
\]

**Remark:** If the coefficients of the bilinear forms \( a(\cdot, \cdot), b(\cdot, \cdot), a^*(\cdot, \cdot) \) and \( b^*(\cdot, \cdot) \) are replaced by other functions, the lemmas 3.3–3.6 are still valid.

Let \( X \) be a Banach space with norm \( || \cdot ||_X \) and \( \phi: [0,T] \rightarrow X \). Define

\[
||\phi||_{L^2(X)}^2 = \int_0^T ||\phi(t)||_X^2 \, dt \quad \text{and} \quad ||\phi||_{L^\infty(X)} = ess \sup_{0 \leq t \leq T} ||\phi(t)||_X.
\]
Let the space $H^k(W^{s,p})$ be defined by

$$H^k(0, T; W^{s,p}) = \{ u \in W^{s,p}; \frac{\partial^j u}{\partial t^j} \in L^2(0, T; W^{s,p}), \ j = 0, 1, \ldots, k \}$$

and for any $u \in H^k(W^{s,p})$, we set

$$\| u(t) \|_{k, s, p} = \sum_{j=1}^k \left\{ \| \frac{\partial^j u}{\partial t^j} \|_{s, p} + \int_0^t \| \frac{\partial^j u}{\partial t^j} \|_{s, p} dt \right\}, \ t \in [0, T].$$

Similar to the proof given in [7,9,10], we can deduce the properties of the generalized Ritz – Volterra projection.

**Lemma 3.7** $V_h^* u$ is defined by (2.8) or (2.8)', then

$$\begin{align*}
(a) \quad & \| D^l u (u - V_h^* u) \|_{1,p} \leq Ch \| u \|_{l, 2,p}, \ l = 0, 1, 2, 3, \ 2 \leq p \leq \infty; \\
(b) \quad & \| D^l (u - V_h^* u) \|_{0,p} \leq Ch^2 \| u \|_{l, 3,p}, \ l = 0, 1, 2, 3, \ 2 \leq p \leq \infty. \quad (3.6)
\end{align*}$$

For convenience, we write $u_h - u = (u_h - V_h^* u) + (V_h^* u - u) = \xi + \eta$ in this paper.

**Lemma 3.8** If $u_0, u_1$ and $u_{0h}, u_{1h}$ are the initial values of (1.1) and (2.9), respectively, then

$$\begin{align*}
(a) \quad & \xi(0) = \xi_{tt}(0) = 0; \\
(b) \quad & \| \xi_t(0) \|_1 \leq Ch^2(\| u_0 \|_3 + \| u_1 \|_3 + \| u_{tt}(0) \|_3). \quad (3.7)
\end{align*}$$

**Proof.** Obviously, $\xi(0) = u_{0h} - V_h^* u_0 = 0$. By noting that

$$((V_h^* u)_{ttt} - (u_h - a^*(u_{1h}, v_h) - b^*(u_{0h}, v_h)) = (u_{h,t} - a^*(u_{1h}, v_h) - b^*(u_{0h}, v_h), v_h), \ \forall v_h \in V_h,$

we can know $(V_h^* u)_{ttt}(0) = u_{h,ttt}(0)$, i.e. $\xi_{tt}(0) = 0$, the conclusion of (3.7a) is proved.

To show (3.7b), apply (2.5), (2.9) and (2.8') to get the error equation:

$$(\xi_{tt}, v_h) + a^*(\xi_t, v_h) + b^*(\xi, v_h) = -(\eta_{tt}, v_h), \ \forall v_h \in V_h, \quad (3.8)$$

Integrating (3.8) with respect to $t$ and noting $\xi(0) = 0$, we can obtain the equivalence of (3.8), by (2.4b)

$$((\xi_t, v_h) + a^*(\xi_t, v_h) + \int_0^t c^*(\xi_t, v_h) dt = -(\eta_t - \xi_t(0) - \eta_t(0), v_h), \ \forall v_h \in V_h, \quad (3.8')$$

Setting $t = 0$ and $v_h = \Pi_h^* \xi_t(0)$ in (3.8), we have, by (3.7a)

$$a^*(\xi_t(0), \Pi_h^* \xi_t(0)) = -(\eta_t(0), \Pi_h^* \xi_t(0)), \quad (3.8')$$

Also from (3.3a) and (3.5b),

$$a^*\| \xi_t(0) \|^2_1 \leq \| \eta_t(0) \|_1 \| \xi_t(0) \| \leq Ch^2(\| u_0 \|_3 + \| u_1 \|_3 + \| u_{tt}(0) \|_3)\| \xi_t(0) \|_1.$$ 

Hence, this completes the proof of (3.7b).

**Lemma 3.9** If $u$ and $u_h$ are the solution of (1.1) and (2.9), respectively, then

$$\| \xi_t \|_1 + \| \xi \|_1 \leq Ch^2\{\| u_0 \|_3 + \| u_1 \|_3 + \| u_{tt}(0) \|_3 + \int_0^t (\| u_* \|_3 + \| u_t \|_3 + \| u_{tt} \|_3) dt\}. \quad (3.9)$$
Proof. Taking \( v_h = \Pi_h^* \xi_t \) in (3.8), we have
\[
\langle \xi_{tt}, \Pi_h^* \xi_t \rangle + a^*(\xi_t, \Pi_h^* \xi_t) + b^*(\xi, \Pi_h^* \xi_t) = -(\eta_{tt}, \Pi_h^* \xi_t),
\]
The above is also written as
\[
\frac{1}{2} \frac{d}{dt} \|\xi_t\|^2 + a^*(\xi_t, \Pi_h^* \xi_t) + \frac{1}{2} \frac{d}{dt} b^*(\xi, \Pi_h^* \xi_t) = -(\eta_{tt}, \Pi_h^* \xi_t) + \frac{1}{2} b^*(\xi, \Pi_h^* \xi_t) - b^*(\xi, \Pi_h^* \xi_t) + \frac{1}{2} b^*_t(\xi, \Pi_h^* \xi_t),
\]
(3.10)
Noting \( \xi(0) = 0 \), lemmas 3.4 and 3.6 and integrating from 0 to \( t \), we get
\[
\|\xi_t\|^2 + \|\xi\|^2 \leq \|\xi_t(0)\|^2 + C \int_0^t \|\eta_{tt}\| \|\xi_t\| d\tau + C \int_0^t h \|\xi_t\|_1 \|\xi\|_1 d\tau + C \int_0^t \|\xi\|^2_d\tau
\]
\[
\leq \|\xi_t(0)\|^2 + C \int_0^t \|\eta_{tt}\|^2 d\tau + C \int_0^t \|\xi_t\|^2 d\tau + C \int_0^t \|\xi\|^2 d\tau,
\]
(3.11)
here we have applied the inverse properties of the finite element space and the inequality
\[
\|\xi_t\|_1 \leq Ch^{-1} \|\xi_t\|.
\]
Thus, the conclusion follows from Gronwall's Lemma, and lemmas 3.7 and 3.8. \( \square \)

Finally, in order to conclude maximum norm estimates, we introduce Green function \( \partial_z G_z^h \in U_h \) and pre-Green function \( \partial_z G_z^* \in H_0^1(I) \):
\[
\begin{align*}
(a) \quad a(\partial_z G_z^h, v_h) &= \partial_z v_h(z), \quad \forall v_h \in U_h; \\
(b) \quad a(\partial_z G_z^*, v) &= \partial_z P_h v(z), \quad \forall v \in H_0^1(I).
\end{align*}
\]
(3.12)
where \( P_h: L^2(I) \rightarrow U_h \) is \( L^2 \) projection operator, and we have the following (see[11])
\[
\|P_h u\|_{s,q} \leq C\|u\|_{s,q}, \quad s = 0, 1, \quad 2 \leq q \leq \infty.
\]
(3.13)

**Lemma 3.10**[11] For Green functions defined in (3.12), we know
\[
\begin{align*}
(a) \quad \|\partial_z G_z^* - \partial_z G_z^h\|_{1,1} + h \|\partial_z G_z^h\|_1 + \|\partial_z G_z^*\|_{0,1} &\leq C; \\
(b) \quad \|\partial_z G_z^h\| &\leq C.
\end{align*}
\]
(3.14)

4. MAIN RESULTS

We next demonstrate a superconvergence results of \( u_h - V_h^* u \).

**Theorem 4.1** Under the conditions of lemma 3.9, for \( h \) sufficiently small, we can deduce
\[
\|\xi\|_{1,p} \leq Ch^2 \{\|u(0)\|_{2,3,p} + \|u\|_{2,3,p}\}, \quad 2 \leq p \leq \infty.
\]
(4.1)

**Proof.** (i) Let us consider the case of \( 2 \leq p < \infty \).

We now introduce an auxiliary problem. Denote \( \phi_x \) to be the derivative of \( \phi \) and let \( \Phi \in H_0^1(I) \) be the solution of
\[
a(v, \Phi) = -(v, \phi_x), \quad v \in H_0^1(I),
\]
(4.2)
and there is a priori estimate
\[ \| \Phi \|_{1,p'} \leq C \| \phi \|_{0,p'}, \quad p' = \frac{p}{p-1}. \]  
(4.3)

By virtue of Green formula, (2.6), and (3.8'),
\[ (\xi_x, \phi) = a(\xi, \Phi) \]
\[ = a(\xi, R_h\Phi) \]
\[ = d_1(\xi, R_h\Phi) + a^*(\xi, \Pi_h^*R_h\Phi) \]
\[ = d_1(\xi, R_h\Phi) - (\xi_t + \eta_t - \xi_t(0) - \eta_t(0), \Pi_h^*R_h\Phi) \]
\[ - \int_0^t c^*(\xi, \Pi_h^*R_h\Phi) d\tau \]
\[ = I_1 + I_2 + I_3. \]

Now it suffices to estimate each term in the above.
Noting that lemmas 3.5 and 3.4, and \( \| R_h\Phi \|_{1,p'} \leq C \| \Phi \|_{1,p'} \), we easily get
\[ |I_1| \leq Ch\| \xi \|_{1,p} \| R_h\Phi \|_{1,p'} \leq Ch\| \xi \|_{1,p} \| \Phi \|_{1,p'} \]

and
\[ |I_3| \leq C \int_0^t \| \xi \|_{1,p} \| \Phi \|_{1,p'}. \]

For \( I_2 \), from Sobolev’s imbedding inequalities, we have
\[ |I_2| \leq (\| \xi_t \|_{1} + \| \xi_t(0) \|_{1}) \| R_h\Phi \| + (\| \eta_t \|_{0,p} + \| \eta_t(0) \|_{0,p}) \| R_h\Phi \|_{0,p'} \]
\[ \leq C(\| \xi_t \|_{1} + \| \xi_t(0) \|_{1} + \| \eta_t \|_{0,p} + \| \eta_t(0) \|_{0,p}) \| \Phi \|_{1,p'}. \]

Combining the estimates of \( I_1 - I_4 \), we obtain also by (4.3) that
\[ \| \xi \|_{1,p} \leq C \sup_{\Phi \in L^{p'}} \| (\xi_x, \phi) \|_{0,p'} \]
\[ \leq Ch\| \xi \|_{1,p} + C(\| \xi_t \|_{1} + \| \xi_t(0) \|_{1} + \| \eta_t \|_{0,p} + \| \eta_t(0) \|_{0,p}) + C \int_0^t \| \xi \|_{1,p} d\tau \]

By letting \( h \) sufficiently small such that \( Ch \leq \frac{1}{2} \), the results for \( 2 \leq p < \infty \) now follows by Gronwall’s Lemma and lemmas 3.7–3.9.

(ii) Let us next consider the case of \( p = \infty \).

Applying the definition (3.12a) of Green function, we have
\[ \partial_z \xi(z) = a(\xi, \partial_z G_z^h) \]
\[ = d_1(\xi, \partial_z G_z^h) - \int_0^t c^*(\xi, \Pi_h^*\partial_z G_z^h) d\tau - (\xi_t + \eta_t - \xi_t(0) - \eta_t(0), \Pi_h^*\partial_z G_z^h) \]
\[ = d_1(\xi, \partial_z G_z^h) - \int_0^t [d_2(\xi, \partial_z G_z^h) - d_t(\xi, \partial_z G_z^h)] d\tau - \int_0^t c(\xi, \partial_z G_z^h - \partial_z G_z^* d\tau \]
\[ - \int_0^t c(\xi, \partial_z G_z^*) d\tau - (\xi_t + \eta_t - \xi_t(0) - \eta_t(0), \Pi_h^*\partial_z G_z^h) \]
\[ = J_1 + \cdots + J_5, \]

Now we proceed to estimate these $J_i$ one by one. From lemmas 3.5 and 3.3, and (3.14a), we get

$$|J_1| \leq C \bar{h} \|\xi\|_1 \|\partial_z G_z^h\|_1 \leq C \|\xi\|_1,$$

$$|J_2| \leq C \int_0^t \|\xi\|_1 d\tau \leq C \int_0^t \|\xi\|_{1,\infty} d\tau,$$

and

$$|J_3| \leq C \int_0^t \|\xi\|_{1,\infty} d\tau \|\partial_z G_z^* - \partial_z G_z^h\|_{1,1} \leq C \int_0^t \|\xi\|_{1,\infty} d\tau.$$

As for $J_4$, it follows from (3.12b), by integration by parts and (3.14a), that

$$|J_4| = \left| \int_0^t \left[ \int_a^b b(x, t) \frac{\partial \xi}{\partial x} \frac{\partial}{\partial x} \left( \partial_z G_z^* \right) dx - \int_a^b a_t(x, t) \frac{\partial \xi}{\partial x} \frac{\partial}{\partial x} \left( \partial_z G_z^* \right) dx \right] d\tau \right|$$

$$= \left| \int_0^t \left\{ \int_a^b \left[ \frac{a_t(x, t)}{a(x, t)} \xi - a_t(x, t) \frac{\partial}{\partial x} \left( \frac{a(x, t)}{a_t(x, t)} \xi \right) \frac{\partial}{\partial x} \left( \partial_z G_z^* \right) dx \right] \right\} d\tau \right|$$

$$= \left| \int_0^t \partial_z P \left( \frac{b(x, t)}{a(x, t)} \xi \right) d\tau - \int_0^t \int_a^b \frac{\partial}{\partial x} \left[ a(x, t) \frac{\partial}{\partial x} \left( \frac{b(x, t)}{a(x, t)} \xi \right) \partial_z G_z^* dx d\tau \right.$$

$$\left. - \int_0^t \partial_z P \left( \frac{a_t(x, t)}{a(x, t)} \xi \right) d\tau + \int_0^t \int_a^b \frac{\partial}{\partial x} \left[ a(x, t) \frac{\partial}{\partial x} \left( \frac{a_t(x, t)}{a(x, t)} \xi \right) \partial_z G_z^* dx d\tau \right] \right|$$

$$\leq C \int_0^t \|\xi\|_{1,\infty} d\tau + C \int_0^t \|\xi\|_{1,\infty} d\tau \|\partial_z G_z^*\|_{0,1}$$

$$\leq C \int_0^t \|\xi\|_{1,\infty} d\tau.$$

Lastly, it is easy to see, by Sobolev's imbedding inequalities and (3.14b), that

$$|J_5| \leq C (\|\xi_t\| + \|\eta_t\| + \|\xi_t(0)\| + \|\eta_t(0)\|) \|\partial_z G_z^h\|$$

$$\leq C (\|\xi\| + \|\eta_t\| + \|\xi_t(0)\| + \|\eta_t(0)\|).$$

Combining the estimates of $J_1 - J_5$, we have

$$\|\xi\|_{1,\infty} \leq C (\|\xi\| + \|\eta_t\| + \|\xi_t(0)\| + \|\eta_t(0)\|) + C \int_0^t \|\xi\|_{1,\infty} d\tau.$$

which together with Gronwall's Lemma, and lemmas 3.7 - 3.9 completes the proof of $p = \infty$. 
Finally, we can deduce the $L^p$ and $W^{1,p}$ norm error estimates of $u - u_h$, by using 
$$\|\xi\|_{0,p} \leq C\|\xi\|_{1,p}$$ and lemma 3.7. \qed

**Theorem 4.2** Under the conditions of theorem 4.1, we can conclude that

$$(a) \quad \|u - u_h\|_{0,p} \leq Ch^2\{\|u(0)\|_{2,3,p} + \|u\|_{2,3,p}\}, \quad 2 \leq p < \infty,$$

$$(c) \quad \|u - u_h\|_{1,p} \leq Ch\{\|u(0)\|_{2,3,p} + \|u\|_{2,3,p}\}, \quad 2 \leq p \leq \infty. \quad (4.4)$$

**References**


Huanrong Li:
DEPARTMENT OF MATHEMATICS, CAPITAL NORMAL UNIVERSITY, BEIJING, 100037, P.R.CHINA
E-mail: lihuanrong1979@126.com
EXISTENCE OF SOLUTIONS OF FUZZY DELAY INTEGRODIFFERENTIAL EQUATIONS WITH NONLOCAL CONDITION

K.BALACHANDRAN AND K.KANAGARAJAN

ABSTRACT. In this paper we prove the existence of solutions of fuzzy delay integrodifferential equations with nonlocal condition. The results are obtained by using the fixed point principles.

1. Introduction

Several authors [3-7,11,12] have studied the fuzzy differential equations by using the H-differentiability for the fuzzy valued mappings of a real variable whose values are normal, convex, upper semi continuous and compactly supported fuzzy sets in $R^n$. Seikkala [10] defined the fuzzy derivative which is generalization of the Hukuhara derivative in [8]. For the Cauchy problem $x' = f(t, x), \ x(t_0) = x_0$, the local existence theorems are proved in [11], and the existence theorems under compactness-type conditions are investigated in [12] when the fuzzy valued mapping $f$ satisfies the generalized Lipschitz condition. Park et al [7] studied the fuzzy differential equation with nonlocal condition. Nieto [6] proved an existence theorem for fuzzy differential equations on the metric space $\ (E^n, D)$. Balachandran and Prakash [2] proved the existence of solutions of fuzzy delay differential equations with nonlocal condition of the form

$$x'(t) = f(t, x(\sigma_1(t)), x(\sigma_2(t)), \ldots, x(\sigma_n(t))), \quad t \in J = [0, a],$$

$$x(0) - g(t_1, t_2, \ldots, t_p, x(\cdot)) = x_0.$$

In this paper we study the existence of solutions of fuzzy delay integrodifferential equations with nonlocal condition of the form

$$(1) \quad x'(t) = f\left(t, x(\sigma_1(t)), \int_0^t h\left(t, s, x(\sigma_2(s)), \int_0^s k(s, \tau, x(\sigma_3(\tau)))d\tau\right)ds\right),$$

$$(2) \quad x(0) - g(t_1, t_2, \ldots, t_p, x(\cdot)) = x_0,$$

where $f : J \times E^n \times E^n \rightarrow E^n, h : J \times J \times E^n \times E^n \rightarrow E^n$ and $k : J \times J \times E^n \rightarrow E^n$ are levelwise continuous functions, $g : J^p \times E^n \rightarrow E^n$ satisfies the Lipschitz condition.

2000 Mathematics Subject Classification: 45G10.
Key words and phrases: Fuzzy delay integrodifferential equations, Nonlocal condition.
© The Korean Society for Industrial and Applied Mathematics, 2005

65
and $\sigma_i : J \to J, i = 1, 2, 3$ are continuous functions, $\sigma_i(t) \leq t$ for all $t \in J$. The existence of solutions for non-fuzzy case of the problem (1)-(2) has been discussed in [5]. The symbol $g(t_1, t_2, \cdots t_p, x(\cdot))$ is used in the sense that in the place of $f(\cdot)$, we can substitute only elements of the set $\{t_1, t_2, \cdots, t_p\}$. For example, $g(t_1, t_2, \cdots, t_p, x(\cdot))$ can be defined by the formula

$$g(t_1, t_2, \cdots, t_p, x(\cdot)) = c_1 x(t_1) + c_2 x(t_2) + \cdots + c_p x(t_p),$$

where $c_i (i = 1, 2, \cdots, p)$ are given constants.

2. Preliminaries

Let $P_K(R^n)$ denote the family of all nonempty, compact, convex subsets of $R^n$. Addition and scalar multiplication in $P_K(R^n)$ are defined as usual. Let $A$ and $B$ be two nonempty bounded subsets of $R^n$. The distance between $A$ and $B$ is defined by the Hausdorff metric

$$d(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} ||a - b||, \sup_{b \in B} \inf_{a \in A} ||a - b|| \right\},$$

where $\cdot$ denote the usual Euclidean norm in $R^n$. Then it is clear that $(P_K(R^n), d)$ becomes a metric space. Let $I = [t_0, t_0 + a] \subset R$ $(a > 0)$ be a compact interval and let $E^n$ be the set of all $u : R^n \to [0, 1]$ such that $u$ satisfies the following conditions:

(i) $u$ is normal, that is, there exists an $x_0 \in R^n$ such that $u(x_0) = 1$,

(ii) $u$ is fuzzy convex, that is, $u(\lambda x + (1 - \lambda) y) \geq \min\{u(x), u(y)\}$, for any $x, y \in R^n$ and $0 \leq \lambda \leq 1$,

(iii) $u$ is upper semicontinuous,

(iv) $[u]^\alpha = \text{cl}\{x \in R^n : u(x) > 0\}$ is compact.

If $u \in E^n$, then $u$ is called a fuzzy number, and $E^n$ is said to be a fuzzy number space. For $0 < \alpha \leq 1$, denote $[u]^\alpha = \{x \in R^n : u(x) \geq \alpha\}$. Then from (i)-(iv), it follows that the $\alpha$-level set $[u]^\alpha \in P_K(R^n)$ for all $0 \leq \alpha \leq 1$.

If $g : R^n \times R^n \to R^n$ is a function, then using Zadeh’s extension principle we can extend $g$ to $E^n \times E^n \to E^n$ by the equation

$$\hat{g}(u, v)(z) = \sup_{z = g(x, y)} \min \{u(x), v(y)\}.$$

It is well known that $[\hat{g}(u, v)]^\alpha = g([u]^\alpha, [v]^\alpha)$ for all $u, v \in E^n$, $0 \leq \alpha \leq 1$ and continuous function $g$. Further, we have $[u + v]^\alpha = [u]^\alpha + [v]^\alpha, [ku]^\alpha = k[u]^\alpha$, where $k \in R$. Define $D : E^n \times E^n \to [0, \infty)$ by the relation $D(u, v) = \sup_{0 \leq \alpha \leq 1} d([u]^\alpha, [v]^\alpha)$,

where $d$ is the Hausdorff metric defined in $P_K(R^n)$. Then $D$ is a metric in $E^n$.

Further we know that [9]

(i) $(E^n, D)$ is a complete metric space,

(ii) $D(u + w, v + w) = D(u, v)$ for all $u, v, w \in E^n$,

(iii) $D(\lambda u, \lambda v) = |\lambda|D(u, v)$ for all $u, v \in E^n$ and $\lambda \in R$. 


It can be proved that \( D(u + v, w + z) \leq D(u, w) + D(v, z) \) for \( u, v, w \) and \( z \in E^n \).

**Definition 2.1.** [3] A mapping \( F : I \to E^n \) is strongly measurable if for all \( \alpha \in [0, 1] \) the set-valued map \( F_\alpha : I \to P_K(R^n) \) defined by \( F_\alpha(t) = [F(t)]^\alpha \) is Lebesgue measurable when \( P_K(R^n) \) has the topology induced by the Hausdorff metric \( d \).

**Definition 2.2.** [3] A mapping \( F : I \to E^n \) is said to be integrably bounded if there is an integrable function \( h(t) \) such that \( \|x(t)\| \leq h(t) \) for every \( x(t) \in F_0(t) \).

**Definition 2.3.** The integral of a fuzzy mapping \( F : I \to E^n \) is defined levelwise by \( \int_I F_\alpha(t)dt = \int_I F(t)dt = \text{The set of all}\ \int_I f(t)dt \text{ such that } f : I \to R^n \text{ is a measurable}\ )\text{ selection for } F_\alpha \text{ for all } \alpha \in [0, 1] \).

**Definition 2.4.** [1] A strongly measurable and integrably bounded mapping \( F : I \to E^n \) is said to be integrable over \( I \) if \( \int_I F(t)dt \in E^n \).

Note that if \( F : I \to E^n \) is strongly measurable and integrably bounded, then \( F \) is integrable. Further if \( F : I \to E^n \) is continuous, then it is integrable.

**Proposition 2.1.** Let \( F, G : I \to E^n \) be integrable and \( c \in I, \lambda \in R \). Then

1. \( \int_{t_0}^{t_0 + \alpha} F(t)dt = \int_{t_0}^{c} F(t)dt + \int_{c}^{t_0 + \alpha} F(t)dt \)
2. \( \int_I (F(t) + G(t))dt = \int_I F(t)dt + \int_I G(t)dt \)
3. \( \int_I \lambda F(t)dt = \lambda \int_I F(t)dt \)
4. \( D(F, G) \) is integrable,
5. \( D\left( \int_I F(t)dt, \int_I G(t)dt \right) \leq \int_I D(F(t), G(t))dt \).

**Definition 2.5** A mapping \( F : I \to E^n \) is Hukuhara differentiable at \( t_0 \in I \) if for some \( h_0 > 0 \) the Hukuhara differences

\[
F(t_0 + \Delta t) - h F(t_0), \quad F(t_0) - h F(t_0 - \Delta t)
\]

exist in \( E^n \) for all \( 0 < \Delta t < h_0 \) and there exists an \( F'(t_0) \in E^n \) such that

\[
\lim_{\Delta t \to 0^+} D((F(t_0 + \Delta t) - h F(t_0))/\Delta t, F'(t_0)) = 0
\]

and

\[
\lim_{\Delta t \to 0^+} D((F(t_0) - h F(t_0 - \Delta t))/\Delta t, F'(t_0)) = 0.
\]

Here \( F'(t) \) is called the Hukuhara derivative of \( F \) at \( t_0 \).

**Definition 2.6.** A mapping \( F : I \to E^n \) is called differentiable at a \( t_0 \in I \) if, for any \( \alpha \in [0, 1] \), the set-valued mapping \( F_\alpha(t) = [F(t)]^\alpha \) is Hukuhara differentiable at point \( t_0 \) with \( DF_\alpha(t_0) \) and the family \( \{DF_\alpha(t_0) : \alpha \in [0, 1]\} \) define a fuzzy number \( F(t_0) \in E^n \).
If \( F : I \to E^n \) is differentiable at \( t_0 \in I \), then we say that \( F'(t_0) \) is the fuzzy derivative of \( F(t) \) at the point \( t_0 \).

**Theorem 2.1.** Let \( F : I \to E^n \) be differentiable. Denote \( F_\alpha(t) = [f_\alpha(t), g_\alpha(t)] \). Then \( f_\alpha \) and \( g_\alpha \) are differentiable and \( [F'(t)]^\alpha = [f'_\alpha(t), g'_\alpha(t)] \).

**Theorem 2.2.** Let \( F : I \to E^n \) be differentiable and assume that the derivative \( F' \) is integrable over \( I \). Then, for each \( s \in I \), we have
\[
F'(s) = F(a) + \int_a^s F'(t) dt.
\]

**Definition 2.7.** A mapping \( f : I \times E^n \to E^n \) is called levelwise continuous at a point \( (t_0, x_0) \in I \times E^n \) provided, for any fixed \( \alpha \in [0,1] \) and arbitrary \( \epsilon > 0 \), there exists a \( \delta(\epsilon, \alpha) > 0 \) such that
\[
d([f(t, x)]^\alpha, [f(t_0, x_0)]^\alpha) < \epsilon
\]
whenever \( |t - t_0| < \delta(\epsilon, \alpha) \) and \( d([x]^\alpha, [x_0]^\alpha) < \delta(\epsilon, \alpha) \) for all \( t \in I, x \in E^n \).

**Corollary 2.1** [2] Suppose that \( F : I \to E^n \) is continuous. Then the function
\[
G(t) = \int_a^t F(s) ds, \quad t \in I
\]
is differentiable and \( G'(t) = F(t) \).

Now, if \( F \) is continuously differentiable on \( I \), then we have the following mean value theorem
\[
D(F(b), F(a)) \leq (b - a) \cdot \sup\{D(F'(t), \hat{0}), t \in I\}.
\]
As a consequence, we have that
\[
D(G(b), G(a)) \leq (b - a) \cdot \sup\{D(F(t), \hat{0}), t \in I\}.
\]

**Theorem 2.3.** Let \( X \) be a compact metric space and \( Y \) any metric space. A subset \( \Omega \) of the space \( C(X, Y) \) of continuous mappings of \( X \) into \( Y \) is totally bounded in the metric of uniform convergence if and only if \( \Omega \) is equicontinuous on \( X \), and \( \Omega(x) = \{\phi(x) : \phi \in \Omega\} \) is a totally bounded subset of \( Y \) for each \( x \in X \).

### 3. Main Results

**Definition 3.1.** A mapping \( x : J \to E^n \) is a solution to the problem (1)-(2) if and only if it is levelwise continuous and satisfies the integral equation
\[
x(t) = x_0 + g(t_1, t_2, \ldots, t_p, x(\cdot)) \\
+ \int_0^t \left( \int_0^s f \left( s, x(\sigma_1(s)), \int_0^\tau h \left( s, \tau, x(\sigma_2(\tau)) , \int_0^\theta k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds \\
\text{for all } t \in J.
\]
Let $M + Na = b$, a positive number, where

$$M = \max D \left( f \left( t, x(\sigma_1(t)), \int_0^t h \left( t, s, x(\sigma_2(s)), \int_0^s k(s, \tau, x(\sigma_3(\tau))) d\tau \right) ds \right), 0 \right)$$

and

$$N = D(g(t_1, t_2, \cdots, t_p, x(\cdot)), 0), \ 0 \in E^n.$$ 

Let $Y = \{ \xi \in E^n : H(\xi, x_0) \leq b \}$ be the space of continuous functions with $H(\xi, \psi) = \sup_{0 \leq t \leq a} D(\xi(t), \psi(t))$.

**Theorem 3.1.** Assume that:

(i) The mapping $f : J \times Y \rightarrow E^n$ is levelwise continuous in $t$ on $J$ and there exists a constant $G_0$ such that

$$D(f(t, x_1, x_2), f(t, y_1, y_2)) \leq G_0[D(x_1, y_1) + D(x_2, y_2)]$$

(ii) The mapping $h : J \times J \times Y \rightarrow E^n$ is levelwise continuous and there exists a constant $G_1$ such that

$$D(h(t, s, x_1, x_2), h(t, s, y_1, y_2)) \leq G_1[D(x_1, y_1) + D(x_2, y_2)]$$

(iii) The mapping $k : J \times J \times Y \rightarrow E^n$ is levelwise continuous and there exists a constant $G_2$ such that

$$D(k(t, s, x), k(t, s, y)) \leq G_2 D(x, y)$$

(iv) There exists a constant $G_3$ such that for all $x, y \in Y$ and $\sigma_i : J \rightarrow J, \ i = 1, 2, 3$

$$D(x(\sigma_i(t)), y(\sigma_i(t))) \leq G_3 D(x(t), y(t))$$

(v) $g : J^p \times Y \rightarrow E^n$ is a function and there exists a constant $G_4 > 0$ such that

$$D(g(t_1, t_2, \cdots, t_p, x(\cdot)), g(t_1, t_2, \cdots, t_p, y(\cdot))) \leq G_4 D(x, y).$$

Then there exists a unique solution $x(t)$ of (1)-(2) defined on the interval $[0, a]$.

**Proof.** Define an operator $\Phi : Y \rightarrow Y$ by

$$\Phi x(t) = x_0 + g(t_1, t_2, \cdots, t_p, x(\cdot)) + \int_0^t f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds.$$  

(4)

First, we show that $\Phi : Y \rightarrow Y$ is continuous whenever $\xi \in Y$ and that $H(\Phi \xi, x_0) \leq b$.

$$D(\Phi \xi(t+h), \Phi \xi(t))$$

$$= D(\int_0^{t+h} f \left( s, \xi(\sigma_1(s)), \int_0^s h \left( s, \tau, \xi(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, \xi(\sigma_3(\theta))) d\theta \right) d\tau \right) ds,$$

$$x_0 + g(t_1, t_2, \cdots, t_p, \xi(\cdot))$$
\[ + \int_0^t f \left( s, \xi(\sigma_1(s)), \int_0^s h \left( s, \tau, \xi(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, \xi(\sigma_3(\theta))) d\theta \right) d\tau \right) ds \]
\[ \leq D \left( \int_0^{t+h} f \left( s, \xi(\sigma_1(s)), \int_0^s h \left( s, \tau, \xi(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, \xi(\sigma_3(\theta))) d\theta \right) d\tau \right) ds, \right.
\[ + \int_0^t f \left( s, \xi(\sigma_1(s)), \int_0^s h \left( s, \tau, \xi(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, \xi(\sigma_3(\theta))) d\theta \right) d\tau \right) ds \]
\[ \leq \int_t^{t+h} D \left( f \left( s, \xi(\sigma_1(s)), \int_0^s h \left( s, \tau, \xi(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, \xi(\sigma_3(\theta))) d\theta \right) d\tau \right) ds, \right) \]
\[ \leq hM \rightarrow 0 \text{ as } h \rightarrow 0. \]

That is, the map \( \Phi \) is continuous. Now
\[ D(\Phi \xi(t), x_0) \]
\[ = D \left( x_0 + g(t_1, t_2, \ldots, t_p, \xi(\cdot)) \right) \]
\[ + \int_0^t f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds, \]
\[ \leq D \left( g(t_1, t_2, \ldots, t_p, \xi(\cdot)), \right) \]
\[ + \int_0^t \left[ D \left( f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right), \right) \right] ds \]
\[ \leq N + M t \]

and so
\[ H(\Phi \xi, x_0) = \sup_{0 \leq t \leq a} D(\Phi \xi(t), x_0) \leq N + Ma \leq b. \]

Thus \( \Phi \) is a mapping from \( Y \) into \( Y \). Since \( C([0, a], E^n) \) is a complete metric space with the metric \( H \), we only show that \( Y \) is a closed subset of \( C([0, a], E^n) \). Let \( \{ \psi_n \} \) be a sequence in \( Y \) such that \( \psi_n \rightarrow \psi \in C([0, a], E^n) \) as \( n \rightarrow \infty \). Then
\[ D(\psi(t), x_0) \leq D(\psi(t), \psi_n(t)) + D(\psi_n(t), x_0), \]

that is,
\[ H(\psi, x_0) = \sup_{0 \leq t \leq a} D(\psi(t), x_0) \leq H(\psi, \psi_n) + H(\psi_n, x_0) \leq \epsilon + b \]

for sufficiently large \( n \) and arbitrary \( \epsilon > 0 \). So \( \psi \in Y \). This implies that \( Y \) is closed subset of \( C([0, a], E^n) \). Therefore \( Y \) is a complete metric space.

By using Proposition 2.1 and assumptions (i)-(v), we will show that \( \Phi \) is a contraction mapping. For \( \xi, \psi \in Y \),
\[ D(\Phi \xi(t), \Phi \psi(t)) \]
\[ = D \left( x_0 + g(t_1, t_2, \ldots, t_p, \xi(\cdot)) \right) \]
EXISTENCE OF SOLUTIONS OF FUZZY DELAY INTEGRODIFFERENTIAL EQUATIONS

\[ + \int_{0}^{t} f \left( s, \xi(\sigma_{1}(s)), \int_{0}^{s} h \left( s, \tau, \xi(\sigma_{2}(\tau)), \int_{0}^{\tau} k(\tau, \theta, \xi(\sigma_{3}(\theta))) d\theta \right) d\tau \right) ds, \]

\[ x_{0} + g(t_{1}, t_{2}, \cdots, t_{p}, \psi(\cdot)) \]

\[ + \int_{0}^{t} f \left( s, \psi(\sigma_{1}(s)), \int_{0}^{s} h \left( s, \tau, \psi(\sigma_{2}(\tau)), \int_{0}^{\tau} k(\tau, \theta, \psi(\sigma_{3}(\theta))) d\theta \right) d\tau \right) ds \]

\[ \leq D(g(t_{1}, t_{2}, \cdots, t_{p}, \xi(\cdot)), g(t_{1}, t_{2}, \cdots, t_{p}, \psi(\cdot))) \]

\[ + \int_{0}^{t} D \left( f \left( s, \xi(\sigma_{1}(s)), \int_{0}^{s} h \left( s, \tau, \xi(\sigma_{2}(\tau)), \int_{0}^{\tau} k(\tau, \theta, \xi(\sigma_{3}(\theta))) d\theta \right) d\tau \right) ds, \]

\[ f \left( s, \psi(\sigma_{1}(s)), \int_{0}^{s} h \left( s, \tau, \psi(\sigma_{2}(\tau)), \int_{0}^{\tau} k(\tau, \theta, \psi(\sigma_{3}(\theta))) d\theta \right) d\tau \right) ds \]

\[ \leq G_{4}D(\xi(\cdot), \psi(\cdot)) + G_{0} \int_{0}^{t} D(\xi(\sigma_{1}(s)), \psi(\sigma_{1}(s))) ds \]

\[ + G_{0} \int_{0}^{t} D \left( \int_{0}^{s} h \left( s, \tau, \xi(\sigma_{2}(\tau)), \int_{0}^{\tau} k(\tau, \theta, \xi(\sigma_{3}(\theta))) d\theta \right) d\tau, \right. \]

\[ \left. \int_{0}^{s} h \left( s, \tau, \psi(\sigma_{2}(\tau)), \int_{0}^{\tau} k(\tau, \theta, \psi(\sigma_{3}(\theta))) d\theta \right) d\tau \right) ds \]

\[ \leq G_{4}D(\xi(\cdot), \psi(\cdot)) + G_{0}G_{3} \int_{0}^{t} D(\xi(s), \psi(s)) ds + G_{0}G_{1}G_{3} \int_{0}^{t} \int_{0}^{s} D(\xi(\tau), \psi(\tau)) d\tau ds \]

\[ + G_{0}G_{1}G_{2}G_{3} \int_{0}^{t} \int_{0}^{s} \int_{0}^{\tau} D(\xi(\theta), \psi(\theta)) d\theta d\tau ds. \]

Then we obtain

\[ H(\Phi_{x}, \Phi_{\psi}) \leq \sup_{0 \leq t \leq a} \left\{ G_{4}D(\xi(\cdot), \psi(\cdot)) + G_{0}G_{3} \int_{0}^{t} D(\xi(s), \psi(s)) ds \right. \]

\[ + G_{0}G_{1}G_{3} \int_{0}^{t} \int_{0}^{s} D(\xi(\tau), \psi(\tau)) d\tau ds \]

\[ + G_{0}G_{1}G_{2}G_{3} \int_{0}^{t} \int_{0}^{s} \int_{0}^{\tau} D(\xi(\theta), \psi(\theta)) d\theta d\tau ds \}

\[ \leq G_{4}D(\xi(\cdot), \psi(\cdot)) + aG_{0}G_{3}D(\xi(t), \psi(t)) \]

\[ + a^{2}G_{0}G_{1}G_{3}D(\xi(t), \psi(t)) + a^{3}G_{0}G_{1}G_{2}G_{3}D(\xi(t), \psi(t)) \]

\[ \leq pH(\xi, \psi), \]

where the constant \( p = G_{4} + G_{0}G_{3}a + G_{0}G_{1}G_{3}a^{2} + G_{0}G_{1}G_{2}G_{3}a^{3} \). Taking sufficiently small \( a \) such that \( p < 1 \), we obtain \( \Phi \) to be a contraction mapping. Therefore \( \Phi \) has a unique fixed point \( x \in C([0, a], E^{n}) \) such that \( \Phi x = x \), that is,

\[ x(t) = x_{0} + g(t_{1}, t_{2}, \cdots, t_{p}, x(\cdot)) \]
\[ + \int_0^t f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta)))d\theta \right) d\tau \right) ds. \]

**Theorem 3.2.** Let \( f, h, k, \sigma \) and \( g \) be as in Theorem 3.1. Denote by \( x(t, x_0), y(t, y_0) \) the solutions of equation (1) corresponding to \( x_0, y_0 \), respectively. Then there exists constant \( q > 0 \) such that
\[ H(x(\cdot, x_0), y(\cdot, y_0)) \leq qD(x_0, y_0) \]
for any \( x_0, y_0 \in E^n \) and \( q = 1/(1 - p) \).

**Proof.** Let \( x(t, x_0), y(t, y_0) \) be solutions of equations (1) corresponding to \( x_0, y_0 \), respectively. Then
\[
D(x(t, x_0), y(t, y_0))
= \begin{align*}
D & \left( x_0 + g(t_1, t_2, \cdots, t_p, x(\cdot)) \right) \\
& + \int_0^t f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta)))d\theta \right) d\tau \right) ds, \\
& \quad y_0 + g(t_1, t_2, \cdots, t_p, y(\cdot)) \\
& + \int_0^t f \left( s, y(\sigma_1(s)), \int_0^s h \left( s, \tau, y(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, y(\sigma_3(\theta)))d\theta \right) d\tau \right) ds \\
& \leq D(x_0, y_0) + D(g(t_1, t_2, \cdots, t_p, x(\cdot)), g(t_1, t_2, \cdots, t_p, y(\cdot))) \\
& + \int_0^t D \left( f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta)))d\theta \right) d\tau \right), \\
& \quad f \left( s, y(\sigma_1(s)), \int_0^s h \left( s, \tau, y(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, y(\sigma_3(\theta)))d\theta \right) d\tau \right) \right) ds \\
& \leq D(x_0, y_0) + G_4D(x(\cdot, y(\cdot)) + G_0G_3 \int_0^t D(x(s), y(s))ds \\
& + G_0G_1G_3 \int_0^t \int_0^s D(x(\tau), y(\tau))d\tau ds + G_0G_1G_2G_3 \int_0^t \int_0^s \int_0^\tau D(x(\theta), y(\theta))d\theta d\tau ds.
\end{align*}
\]
Thus \( H(x(\cdot, x_0), y(\cdot, y_0)) \leq D(x_0, y_0) + pH(x(\cdot, x_0), y(\cdot, y_0)) \). That is,
\[ H(x(\cdot, x_0), y(\cdot, y_0)) \leq 1/(1 - p)D(x_0, y_0). \]
This completes the proof of the theorem. \( \square \)

Next we generalize the above theorem for the fuzzy delay integrodifferential equation (1)-(2) with nonlocal condition.

**Theorem 3.3.** Suppose that \( f : J \times E^n \to E^n \), \( h : J \times J \times E^n \times E^n \to E^n \) and \( k : J \times J \times E^n \to E^n \) are level wise continuous and bounded, \( \sigma_i : J \to J \ (i = 1, 2, 3) \) and \( g : J^p \times E^n \to E^n \) are continuous. Then the initial value problem (1)-(2) possesses at least one solution on the interval \( J \).
Proof. Since \( f, h, k \) are continuous and bounded and \( g \) is a continuous function there exists \( r \geq 0 \) such that
\[
D \left( f \left( t, x(\sigma_1(t)), \int_0^t h \left( t, s, x(\sigma_2(s)), \int_0^s k(s, \tau, x(\sigma_3(\tau))) d\tau \right) ds \right), \hat{0} \right) \leq r, \: t \in J, \: x \in E^n.
\]
Let \( B \) be a bounded set in \( C(J, E^n) \). The set \( \Phi B = \{ \Phi x : x \in B \} \) is totally bounded if and only if it is equicontinuous and for every \( t \in J \), the set \( \Phi B(t) = \{ \Phi x(t) : t \in J \} \) is a totally bounded subset of \( E^n \). For \( t_0, t_1 \in J \) with \( t_0 \leq t_1 \) and \( x \in B \) we have that
\[
D(\Phi x(t_0), \Phi x(t_1)) \\
= D \left( x_0 + g(t_1, t_2, \cdots, t_p, x(\cdot)) + \int_0^{t_0} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds, \right. \\
\left. x_0 + g(t_1, t_2, \cdots, t_p, x(\cdot)) + \int_0^{t_1} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds \right)
\leq D \left( \int_0^{t_0} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds, \right.
\left. \int_0^{t_1} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds \right)
\leq \int_{t_0}^{t_1} D \left( f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right), \hat{0} \right) ds
\leq |t_1 - t_0| \cdot \sup \left\{ D \left( f \left( t, x(\sigma_1(t)), \int_0^t h \left( t, s, x(\sigma_2(s)), \int_0^s k(s, \tau, x(\sigma_3(\tau))) d\tau \right) ds \right), \hat{0} \right) \right\}
\leq |t_1 - t_0| \cdot r.
\]
This shows that \( \Phi B \) is equicontinuous. Now, for \( t \in J \) fixed, we have
\[
D(\Phi x(t), \Phi x(t')) \leq |t - t'| \cdot r, \quad \text{for every} \quad t' \in J, \: x \in B.
\]
Consequently, the set \( \{ \Phi x(t) : x \in B \} \) is totally bounded in \( E^n \). By Ascoli’s theorem we conclude that \( \Phi B \) is a relatively compact subset of \( C(J, E^n) \). Then \( \Phi \) is compact, that is, \( \Phi \) transforms bounded sets into relatively compact sets.

We know that \( x \in C(J, E^n) \) is a solution of (1)-(2) if and only if \( x \) is a fixed point of the operator \( \Phi \) defined by (4).

Now, in the metric space \( (C(J, E^n), H) \), consider the ball
\[
B = \{ \xi \in C(J, E^n), H(\xi, \hat{0}) \leq m \}, \quad m = a \cdot r.
\]
Thus, \( \Phi B \subset B \). Indeed, for \( x \in C(J, E^n) \),
\[
D(\Phi x(t), \Phi x(0)) \\
= D \left( x_0 + g(t_1, t_2, \cdots, t_p, x(\cdot)) + \int_0^{t_0} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds, \right.
\left. x_0 + g(t_1, t_2, \cdots, t_p, x(\cdot)) + \int_0^{t_1} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds \right)
\leq D \left( \int_0^{t_0} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds, \right.
\left. \int_0^{t_1} f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds \right)
\leq \int_{t_0}^{t_1} D \left( f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right), \hat{0} \right) ds
\leq |t_1 - t_0| \cdot \sup \left\{ D \left( f \left( t, x(\sigma_1(t)), \int_0^t h \left( t, s, x(\sigma_2(s)), \int_0^s k(s, \tau, x(\sigma_3(\tau))) d\tau \right) ds \right), \hat{0} \right) \right\}
\leq |t_1 - t_0| \cdot r.
\]
\begin{align*}
+ \int_0^t f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right) ds,
& x_0 + g(t_1, t_2, \ldots, t_p, x(\cdot)) \\
\leq \int_0^t D \left( f \left( s, x(\sigma_1(s)), \int_0^s h \left( s, \tau, x(\sigma_2(\tau)), \int_0^\tau k(\tau, \theta, x(\sigma_3(\theta))) d\theta \right) d\tau \right), \hat{0} \right) ds \\
\leq |t| \cdot r \leq a \cdot r.
\end{align*}

Therefore, defining \( \hat{0} : J \to E^n \), \( \hat{0}(t) = \hat{0} \), \( t \in J \) we have

\[ H(\Phi x, \Phi \hat{0}) = \sup \{ D(\Phi x(t), \Phi \hat{0}(t)) : t \in J \}. \]

Therefore \( \Phi \) is compact and, in consequence, it has a fixed point \( x \in B \). This fixed point is a solution of the initial value problem (1)-(2).

\[ \square \]

References


K.Balachandran:

DEPARTMENT OF MATHEMATICS BHARATHIAR UNIVERSITY, COIMBATORE 641 046, INDIA
E-mail: balachandran.k@lycos.com

K.Kanagarajan:

KARPAGAM COLLEGE OF ENGINEERING, COIMBATORE 641 032, INDIA
INVERSE HEAT CONDUCTION PROBLEM IN A THIN CIRCULAR PLATE AND ITS THERMAL DEFLECTION

A. K. Tikhe, K. C. Deshmukh

Abstract An inverse problem of transient heat conduction in a thin finite circular plate with the given temperature distribution on the interior surface of a thin circular plate being a function of both time and position has been solved with the help of integral transform technique and also determine the thermal deflection on the outer curved surface of a thin circular plate defined as \( 0 \leq r \leq a, \ 0 \leq z \leq h \). The results, obtained in the series form in terms of Bessel's functions, are illustrated numerically.

1. Introduction

An inverse thermoelastic problem consists of determination of the temperature of the heating medium, the heat flux on the boundary surfaces of the solid when the conditions of the displacement and stresses are known at the some points of the solid under consideration. Grysa and Cialkowski [1], Grysa and Kozlowski [2] investigated one dimensional transient thermoelastic problems and derived the heating temperature and the heat flux on the surface of an isotropic infinite slab. Noda N. [3] studied inverse problem of coupled thermal stress fields in a thick plate. Ashida F. et.al. [4] studied the inverse problem of two-dimensional Piezothermoelasticity in an orthotropic plate exhibiting crystal class mm\(^2\). Noda N. et al. [5] attempted an inverse thermoelastic problem in an isotropic plate associated with a Piezoelectric ceramic plate. The direct Problems of normal deflection of an axisymmetrically heated circular plate in the case of fixed and simply supported edges have been considered by Boley and Weiner [6]. Further, Roy Choudhuri [7] has succeeded in determining the normal deflection of a thin clamped circular plate due to ramp-type heating of a concentric circular region of the upper face. In this paper, we reconstruct the problem studied by Roy Choudhuri [7] and deals with the inverse thermoelastic problem of a thin clamped circular plate. The temperature distribution, the unknown temperature gradient and the thermal deflection on the outer curved surface of a thin circular plate of thickness \( h \) defined as \( D: 0 \leq r \leq a, \ 0 \leq z \leq h \) are discuss. No one has attempted this problem so far. The results, obtained in a series form involving Bessel's functions, are illustrated numerically.

2. Statement of the Inverse Heat Conduction Problem

Consider a thin circular plate of thickness \( h \) occupying the space \( D: 0 \leq r \leq a, \ 0 \leq z \leq h \). The plate is initially at zero temperature with temperature distribution on the interior surface of a thin circular plate and the faces \(( z = 0, \ z = h )\) of a thin circular plate are kept at zero temperature.

Key Words: Quasi-static, Transient problem, Thermal deflection, Inverse problem.
AMS No.: 35-XX, 44-XX, 80-XX.
© THE KOREAN SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS, 2005
The differential equation governing the temperature function $T(r, z, t)$, is as

$$
\frac{\partial^2 T}{\partial t^2} + \frac{1}{r} \frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2} = \frac{1}{K} \frac{\partial T}{\partial t},
$$

subject to the initial condition, boundary conditions and interior condition, respectively as

$$
T(r, z, t) \bigg|_{t=0} = 0,
$$

$$
T(r, z, t) \bigg|_{z=0} = 0,
$$

$$
T(r, z, t) \bigg|_{z=h} = 0,
$$

$$
T(r, z, t) \bigg|_{r=a} = g(z, t) \text{ (unknown)}
$$

and

$$
T(r, z, t) \bigg|_{r=\xi} = f(z, t); \quad 0 < \xi < a, \quad 0 < z < h \quad \text{(known)}
$$

where $K$ is thermal diffusivity of the material of the circular plate. The equations (1) to (6) constitute the mathematical formulation of the inverse heat conduction problem.

### 3. Solution of the Inverse Heat Conduction Problem

#### Determination of the Temperature Function $T(r, z, t)$ and the Unknown Temperature Gradient $g(z, t)$

On applying the finite Fourier sine transform and Laplace transform to the equations (1) to (6) and then applying their inversions to the resultant equations, one obtains, the expressions for temperature distribution and unknown temperature gradient, respectively as

$$
T(r, z, t) = 4K \sum_{m=1 \text{ odd}}^{\infty} \sum_{n=1}^{\infty} \left[ \sin \left( \frac{m\pi z}{h} \right) \frac{\beta_n I_0(\beta_n r)}{I_1(\beta_n \xi)} \phi_m(n, t) \right]
$$

and

$$
g(z, t) = 4K \sum_{m=1 \text{ odd}}^{\infty} \sum_{n=1}^{\infty} \left[ \sin \left( \frac{m\pi z}{h} \right) \frac{\beta_n I_0(\beta_n a)}{I_1(\beta_n \xi)} \phi_m(n, t) \right]
$$

where $m$ and $n$ are positive integers and $\beta_n$ is the $n^\text{th}$ positive root of the transcendental equation

$$
I_0(\beta_n a) = 0
$$

and

$$
\phi_m(n, t) = \int_0^t f_i(m, t') \cdot \exp(-K(\beta_n^2 + p^2)(t - t')) dt'
$$

#### Determination Of Quasi-Static Thermal Deflection

Differential equation satisfied by deflection $\omega(r, t)$ as in [7] is

$$
D \nabla^2 \omega = -(\nabla^2 M_T)/(1-\nu)
$$

where $M_T$ is the thermal moment of the plate, $\nu$ is the Poisson’s ratio of the plate material and $D$ is the flexural rigidity of the plate,

$$
D = \frac{E h^3}{12(1-\nu^2)}
$$
and \[ \nabla^2 \frac{\partial^2 w}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} \]

Since, the edge of the circular plate is fixed and clamped
\[ \partial w \frac{\partial}{\partial r} = 0 \text{ at } r = a \] (13)

Assume the solution of (11) satisfying conditions (13) as
\[ w(r, t) = \sum_{n=1}^{\infty} C_n(t) \left[ 2aI_0(\beta_n r) - 2aI_0(\beta_n a) \right] + \beta_n(r^2 - a^2)J_1(\beta_n a) \] (14)
where \( C_n(t) \) are selected in such a way that \( w(r, t) \) satisfies equation (11).

The thermal moment \( M_r \) is defined as
\[ M_r(r, t) = aE \int_0^h z T(r, z, t) \, dz \] (15)
where \( a \) and \( E \) are the coefficient of linear thermal expansion and Young's modulus, respectively.

On substituting the value of the temperature (7) in (15), one obtains
\[ M_r(r, t) = -\frac{4K\alpha Eh}{\pi^2} \sum_{n=1}^{\infty} \frac{\sin(m \pi)}{1} \cos(m \pi) \frac{\beta_n}{\beta_n} \phi_m(m, t) f_0(\beta_n r) \]
(16)

Using (14), (16) and the well-known result
\[ \left( \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} \right) f_0(\beta_n r) = -\beta_n^2 f_0(\beta_n r) \]
(17)
in equation (11), one obtains the constant \( C_n(t) \) as
\[ C_n(t) = -\frac{2K\alpha Eh}{\pi^2 Da \pi^2(1-v)} \sum_{m=1}^{\infty} \frac{\sin(m \pi)}{1} \cos(m \pi) \frac{1}{\beta_n} \phi_m(m, t) \]
(18)

Finally, substituting equation (18) in equation (14), one obtains the expression for the quasi-static thermal deflection \( w(r, t) \) as
\[ w(r, t) = -\frac{2K\alpha Eh}{\pi^2 Da \pi^2(1-v)} \sum_{m=1}^{\infty} \frac{\sin(m \pi)}{1} \cos(m \pi) \frac{1}{\beta_n} \phi_m(m, t) \]
\[ \cdot \left[ 2a(f_0(\beta_n r) - f_0(\beta_n a)) + \beta_n(r^2 - a^2)J_1(\beta_n a) \right] \]
(19)

Special Case

Set \( f(z, t) = T_0(1-e^{-At}) \cdot z(z-h) \) (20)
where \( T_0 > 0, A > 0 \) are constants.

On applying finite Fourier sine transform, one obtains
\[ \tilde{f}(m, t') = -4T_0 \frac{h^3}{m^2} \frac{1}{m} \left( 1 - e^{-At'} \right) \text{, if } m \text{ is odd} \] (21)

Substitute (21) in (10), one obtains
\[ \phi_m(n, t) = -4T_0 \frac{h^3}{m^2} \left[ \frac{1}{m^2} \right] \left[ 1 - e^{-Kbeta^2 + p^2 \mu} + e^{-At'} - e^{-Kbeta^2 + p^2 \mu} \right] \]
(22)

On using (22) in (7), (8) and (19), one obtains the expressions for the temperature distribution, the unknown temperature gradient and the thermal deflection, respectively as
\[ T(r, z, t) = \frac{-16T_0 K h^2}{\pi^2 \xi} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \left[ \frac{1}{m^4} \sin \left( \frac{m \pi z}{h} \right) \cdot \frac{\beta_n J_0(\beta_n r)}{I_1(\beta_n \xi)} \right] \]

\[ \left[ \frac{1 - e^{-K(\beta_n^2 + p^2) t}}{K(\beta_n^2 + p^2)} + \frac{e^{-A t} - e^{-K(\beta_n^2 + p^2) t}}{A - K(\beta_n^2 + p^2)} \right] \]

\[ g(z, t) = \frac{-16T_0 K h^2}{\pi^2 \xi} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \left[ \frac{1}{m^4} \sin \left( \frac{m \pi z}{h} \right) \cdot \frac{\beta_n J_0(\beta_n a)}{I_1(\beta_n \xi)} \right] \]

\[ \left[ \frac{1 - e^{-K(\beta_n^2 + p^2) t}}{K(\beta_n^2 + p^2)} + \frac{e^{-A t} - e^{-K(\beta_n^2 + p^2) t}}{A - K(\beta_n^2 + p^2)} \right] \]

\[ \omega(r, t) = \frac{-8T_0 K_a E h^4}{D a \pi^4 \xi (1 - \nu)} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \left[ \frac{1}{m^4} \left[ 2a(\beta_n J_0(\beta_n r) - J_0(\beta_n a)) + \beta_n(r^2 - a^2) \right] \right] \]

\[ \left[ \frac{1 - e^{-K(\beta_n^2 + p^2) t}}{K(\beta_n^2 + p^2)} + \frac{e^{-A t} - e^{-K(\beta_n^2 + p^2) t}}{A - K(\beta_n^2 + p^2)} \right] \]

4. NUMERICAL RESULTS AND DISCUSSIONS

For computational work, we take a steel material (SN50C) as the example for which the material constants are as follows:

\[ \alpha = 11.6 \times 10^{-6} \text{ [K}^{-1}], \quad K = 15.9 \times 10^{-6} \text{ [m}^2\text{s}^{-1}], \quad E = 215 \text{ [GPa]}, \]

\[ \nu = 0.281, \quad a = 5 \text{ m}, \quad h = 1 \text{ m}, \quad A = 0.3, \quad \xi = 1 \text{ m}, \quad \pi = 3.14, \quad t \text{ in second} \]

The plate is thin due to one fifth thickness of the largest dimension in the work of Nowinski [8] and the first five roots of the transcendental equation \( J_0(\beta n) = 0 \) as in [9]

are \( \beta_1 = 0.4809, \beta_2 = 1.1040, \beta_3 = 1.7307, \beta_4 = 2.3583, \beta_5 = 2.9861 \). Using the numerical values of the above material constants and roots of the transcendental equation, temperature distribution, unknown heating temperature and thermal deflection are evaluated. The variations are shown in the figures (1) to (6).

Figure 1 shows that the temperature oscillates for different times.
Figure 2 shows that fluctuation in temperature for the different radii.
Figure 3 shows that temperature increases up to certain limit and then start decline for different radii.
Figure 4 shows that the unknown temperature distribution is going on increasing up to a certain limit and then starts decline and reaches to the boundary for the different times.
Figure 5 shows that the thermal deflection goes on increasing as the time increases for the radii of the plate.
Figure 6 shows that initially the thermal deflection goes on increasing for small radii and when the radii increases the deflection starts decreasing for different times.
FIGURE 1: Variation of T versus r (r = 0, 1, 2, ..., 5) for different time t = 0.02, 0.04, ..., 0.1

FIGURE 2: Variation of T versus t (t = 0, 0.01, 0.02, ..., 0.1) for different radii r = 0, 1, ..., 5

FIGURE 3: Variation of T versus z (z = 0, 0.1, 0.2, ..., 1) for different values of r = 0, 1, ..., 5
FIGURE 4: Variation of $g$ versus $z = (0, 0.1, 0.2, ..., 1)$ for different values of $t = 0, 0.02, 0.04, ..., 0.1$

FIGURE 5: Variation of $\omega$ versus $t$ ($t = 0, 0.1, 0.2, ..., 0.5$) for different values of $r = 0, 1, ..., 5$

FIGURE 6: Variation of $\omega$ versus $r$ ($r = 0, 1, ..., 5$) for different values of $t = 0, 0.1, ..., 0.5$
5. CONCLUDING REMARKS

Roy Choudhuri [7] who studied the direct problem and deals with quasi-static thermal deflection of a thin clamped circular plate due to ramp-type of heating of a concentric circular region of the upper face, while we modify the work of Roy Choudhuri [7] and deals with the inverse heat conduction problem on the outer curved surface of a thin clamped circular plate. As a special case, mathematical model is constructed and determined the expressions for temperature distribution, unknown heating temperature and thermal deflection on the outer curved surface of a thin circular plate and illustrated numerically. The results, obtained here mainly applicable in engineering problems, particularly for industrial machines subjected to the heating such as the main shaft of a lathe, turbines and the roll of rolling mill. Any particular case of the special interest can be derived by assigning suitable values to the parameters and functions in the expressions (7), (8) and (19).

ACKNOWLEDGEMENT

The authors express their sincere thanks to Dr. P. C. Wankhede, Ex. Prof. and Head, Department of Mathematics, Nagpur University, Nagpur, India and the UGC, New Delhi, India for financial support under the major research project scheme.

REFERENCES

A. K. Tikhe  
Department of Mathematics  
Late. K. Z. S. Science College  
Bramhani - 441501 (Kalmeshwar)  
Dist. Nagpur (India)  
E-mail:

K. C. Deshmukh  
Department of Mathematics  
Post Graduate Department of Mathematics  
Nagpur University, Nagpur – 440010 (India)  
E-mail: kcndeshmukh2000@rediffmail.com
A NEW ALGORITHM OF EVOLVING ARTIFICIAL NEURAL NETWORKS VIA GENE EXPRESSION PROGRAMMING

KANGSHUN LI, YUANXIANG LI, HAIFANG MO, ZHANGXIN CHEN

ABSTRACT. In this paper a new algorithm of learning and evolving artificial neural networks using gene expression programming (GEP) is presented. Compared with other traditional algorithms, this new algorithm has more advantages in self-learning and self-organizing, and can find optimal solutions of artificial neural networks more efficiently and elegantly. Simulation experiments show that the algorithm of evolving weights or thresholds can easily find the perfect architecture of artificial neural networks, and obviously improves previous traditional evolving methods of artificial neural networks because the GEP algorithm imitates the evolution of the natural neural system of biology according to genotype schemes of biology to crossover and mutate the genes or chromosomes to generate the next generation, and the optimal architecture of artificial neural networks with evolved weights or thresholds is finally achieved.

1. INTRODUCTION

The artificial neural network (ANN) [3, 4] imitates natural neural networks of biology of animal brains to solve complex social problems, such as a classifier problem of diseases. The technology of ANNs is an interdisciplinary area that involves neuroscience, mathematics, statistics, physics, computer science, and engineering. ANNs are divided into two classes: supervised and unsupervised. They have been applied to almost all the fields, such as modeling, time series analysis, pattern identification, and signal process and control. A biological neuron may have million different inputs, and may send its outputs to many other neurons; neurons act in a three-dimensional space pattern. Therefore, the neural networks of biological brains are much more complex than any artificial neural network. Despite this fact, the artificial neural network can handle many difficult, nonlinear, and complex processes.

To simulate a biological neural network more closely, we need to design an artificial neural network by using algorithms that are random and close to the nature. There

**Key words and phrases:** gene expression programming, evolutionary, artificial neural network, genotype, phenotype.

This work is partly supported by the National Natural Science Key Foundation of China with Grant No. 60133010, the Research Project of Science and Technology of Education Department of Jiangxi Province with Grant No. Gan-Jiao-Ji-Zi [2005] 150, and the Key Laboratory of High Performance Computing Technology of Jiangxi Province with Grant No. JXHC-2005-003.

© THE KOREAN SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS, 2005

83
are many traditional designing algorithms of artificial neural networks, such as mathematical, numerical, and evolutionary algorithms and other hybrid algorithms. In this paper a new algorithm is introduced to design weights or thresholds of artificial neural networks to generate an optimal architecture of ANNs. This algorithm is closer to biological neural networks that use gene schemes, chromosomes, multi-genes, and multi-chromosomes to crossover and mutate each other than any other traditional algorithm. Experiments indicate that this algorithm is very simple and can easily and accurately construct an optimal architecture of ANNs.

The paper is organized as follows. In section 2, a basic GEP theory will be introduced. The algorithm of evolving weights and architectures of ANNs via GEP is discussed in section 3. Then, in section 4 experiments are given to demonstrate the algorithm of solving optimal artificial neural networks via GEP. Finally, in section 5 conclusions are given.

2. Basic GEP Theory

GEP [1, 2] was introduced by Ferreira in 2001. It belongs to the research field of evolutionary computations [6, 9] like a genetic algorithm (GA) and genetic programming (GP) [8], and it is a natural development of GAs and GP. It also simulates the natural biological evolution to generate new individuals of population through crossover and mutation gene codes, i.e., chromosomes of the individuals at random. GEP has been applied to many practical areas, such as solving system regression problems, fitting economic prediction curves, building mathematical models to approximate complex functions, solving parameter identification problems, and solving data mining problems. The goal of this paper is to focus on how to evolve weights or thresholds to build an optimal architecture of artificial neural networks and to perform the corresponding algorithm efficiently.

2.1. GEP structure. In GEP, a genome or chromosome consists of a linear and symbolic string of fixed length composed of one or more genes. Despite their fixed length it will be shown that GEP chromosomes can code expression trees with different sizes and shapes. The structural organization of GEP genes is better understood in terms of open reading frames (ORFs). In biology, an ORF or coding sequence of a gene begins with the "start" codon, continues with the amino acid codons, and ends at a termination codon. However, a gene possesses more than a respective ORF, with sequences upstream from the start codon and sequences downstream from the stop codon.

In addition, genes are structurally organized in a head that consists of arguments including function symbols, variables, and constants, and in a tail that includes only variables or constants. It is this structural and functional organization of GEP genes that always guarantees production of valid programs or expression trees, no matter how much and how profoundly we modify the chromosomes.
2.2. GEP encoding and decoding. GEP uses the same kind of diagram representation as GP [5, 7], but the entities (expression trees or parse trees) decoded via GEP are the expression of a genome, which is encoded using GEP rules and is different from GP. These kinds of expression trees decoded via GEP are phenotype representations of genome’s genotype of GEP individuals. Through crossover and mutation of genes, chromosomes, multi-genes, and multi-chromosomes, optimal weights of artificial neural networks can be produced, and the optimal architecture of artificial neural networks can be obtained.

2.3. GEP representation. We consider an example of an algebraic expression:

\[
\sqrt{\sin a + \cos b} \ast (a - b).
\]

We call function (1) the original problem of GEP, and then convert it into an expression tree as in Fig. 1.

In Fig. 1, “Q”, “S”, and “C” represent the square root, sin, and cos functions, respectively. We define this kind of diagram as a representation of the phenotype of GEP individuals of function (1). Function (1) is the chromosome function of GEP, and the following gene sequence is the genotype of GEP individuals of function (1) according to the GEP rule:

\[
1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 0
\]

\[
Q \ \ast \ + \ - \ S \ C \ a \ b \ a \ b
\]

where the first line denotes the order of the arguments of a GEP gene; i.e., there are ten arguments in this gene. The corresponding organization method between the phenotype in Fig. 1 and the genotype of expression (2) will be given below.

We define a decoding method from the genotype like expression (2) into the phenotype as in Fig. 1 with the reading rules of a GEP gene sequence: The genotype of a GEP gene sequence is transferred from the straightforward reading of the expression tree from left to right and from top to bottom.
2.4. Genotype and phenotype of ANNs. In order to illustrate a GEP representation of ANNs, we introduce a logical expression tree as shown in Fig. 2 as a phenotype of a complex ANN with the transferring threshold function from a genotype of ANN, which is a single gene consisting of 12 places of head and 12 places of tail in the same way as follows:

\[
\begin{array}{cccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 & 1 & 2 & 3 & 4 \\
Q & T & T & D & a & b & a & D & b & b & a & Q & b & a & a & b & b & a & b & b & a & b & b & h & e & a & d & t & a & i & l
\end{array}
\]

where “Q” represents a node with four argument inputs, “T” a node with three argument inputs, “D” a node with two argument inputs, and the a’s and b’s are external inputs (terminal inputs) of an ANN neuron. In Fig. 2, each number from 0 to 17 represents the subscript of a weight, which has been obtained at random in the interval range and represents the time order of generating random numbers of weights.

3. A New Algorithm

We consider a complex feedforward multi-layer artificial neural network consisting of a set of processing elements [10], also known as neurons or nodes, which are interconnected. It can be described as a directed graph in which each node \( i \) performs a transfer function \( f_i \) of the form

\[
y_i^{m+1} = f_i^{m+1} \left( \sum_{j=1}^{n} w_{ij}^m x_j^m + b_j^m \right), \quad m = 0, 1, \ldots, M - 1, \quad i = 1, 2, \ldots, n,
\]

where \( M \) is the number of layers in the network, \( y_i^{m+1} \) is the output of node \( i \) in the \( m + 1 \)th layer, \( x_j^m \) is the \( j \)th input to the node in the \( m \)th layer, \( w_{ij}^m \) is the connection weight between nodes \( i \) and \( j \) in the \( m \)th layer, and \( b_j^m \) is the threshold (or bias) of the node in the \( m \)th layer. Usually, \( f_i^{m+1} \) is nonlinear, such as a heaviside, sigmoid, or Gaussian function.

Assume the training set \( \{(X_i, y_i) : i = 1, 2, \ldots, k\} \), where \( X_i \) are the input vectors and \( y_i \) are the target outputs (the gene size is 30 with head size 8 and tail size 22).

In order to see the advantages of evolving ANNs in a better way via GEP, we set the fitness function

\[
\text{fitness} = \sum_{i=1}^{k} (M - |\hat{y}_i - y_i|),
\]

where \( M = 100 \), \( y_i \) are the target outputs, and \( \hat{y}_i \) are the computing outputs. If \(|\hat{y}_i - y_i| \leq \varepsilon \) for some \( i = 1, 2, \ldots, k \), change \(|\hat{y}_i - y_i| \) to 0. The sizes of \( M \) and \( \varepsilon \) depend on an ANN problem. Thus we can easily see that the larger the fitness value, the better
the evolved ANN architecture. Therefore, the algorithm can be described according to the genotype and phenotype representation of ANNs in section 2.4 as follows:

1. Create an initial population of an ANN genotype using the gene expression (3) and the corresponding initial population of weights and thresholds according to the number of weights of each GEP individual at random.
2. Transfer all the genotypes into phenotypes of ANN genes, and calculate all the corresponding fitness values.
3. Select a single point gene (or chromosome) or multi-point gene (or chromosome, multi-gene or multi-chromosome) of two ANN gene parents to cross, and then select two weights of every individual to mutate randomly to generate new individuals.
4. Select a single point gene (or chromosome) or multi-point gene (or chromosome, multi-gene or multi-chromosome) of an ANN gene parent to mutate, and then select a weight of this individual to mutate to produce new individuals.
5. Recalculate the fitness value of the evolved individual if the value is larger than the smallest fitness value of individuals in the population, then replace the bad individual with the new individual, and generate the next population.
6. If the stop criterion is reached, i.e., if the fitness value of the best individual in the population or the iteration number is more than a prescribed number, save the best individual as the best architecture of an ANN. Otherwise, go to step 3.

4. Simulation Experiments

In order to demonstrate the validity of constructing the optimal ANN architecture using GEP, only the exclusive-or function (XOR) is used to perform our experiments in this paper.

The XOR is a simple Boolean function of two activities with values 0 or 1, and when the values of two activities are equal, the output of the final layer is 0; otherwise, the output of the final layer is 1. It is known that this kind of function can be solved easily using linear encoded neural networks. However, if we evolve this neural network using GEP, all the structural and non-structural ANN architectures satisfying the XOR, the inputs, and the final layer outputs can be obtained. This is the main difference between a GEP algorithm and other mathematical algorithms.

In the GEP evolving, we set $T = 100$, the function set is $\{Q, T, D\}$, the population size is 20, the terminal set is $\{a, b\}$, the weights range in the interval $[-2,2]$, the head length is 8, the tail length is 22, $\varepsilon = 0.01$, and we use a heaviside transfer function to calculate all the outputs. We run the GEP program 30 times consecutively, and then find the perfect solution as the best individual as follows:

$$1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0$$
$$Q T Q Q D D b a a a a b b b a a b a b a b a b a a a a b b a$$
where

Subscripts of Weights = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17\}
Random weights = \{1.51, 1.79, 1.16, 1.74, -1.97, -0.84, -1.16, -0.2, -0.44, -1.11, 1.61, -1.74, -1.84, -1.91, 0.55, -0.79, -0.16, -1.67, -1.15\}.

According to the rule of transferring the genotype to the phenotype, expression (6) can be written in the expression tree as in Fig. 3.

In the above experiments, if the function set is changed to \{D\} and other parameters remain the same, then a two-layer artificial neural network can be obtained, a shorter perfect ANN gene solution can be generated, and its expression tree is shown as in Fig. 4. Hence we can easily see from the experiments that the ANN architecture obtained using GEP is not unique, its architecture and shape depend on the function set, the number of layers, the head length the tail length, etc.

5. Conclusions

From the above theoretical and experimental analysis of GEP, we see that the algorithm of solving ANNs optimal architecture via GEP is easier than other algorithms. In addition to the above experiments, some evolving experiments of complex ANNs architectures have been also performed and in general we can fins an optimal solution of ANNs in less than 100 iterations. Compared with GP, the convergence speed and shape of evolving ANNs using GEP are much better; the main reason is that GP's
genotype and phenotype are non-linearly represented and the evolved genes sometimes have syntax mistakes, but GEP has overcome these shortcomings and linearly represents the genotype and phenotype. Furthermore, the evolved genes are legal at any time. Therefore, the GEP evolving algorithm has a wider application range in solving the architectures and weights and thresholds of ANNs. We have used GEP to train neural network ensembles to broaden its applications.

References


Kangshun Li:  
**School of Information Engineering**, **Jiangxi University of Science & Technology**, **Jiangxi** **341000**, **China**.  
**E-mail**: lks@public1.gzptt.jx.cn

Yuanxiang Li and Haifang Mo:  
**State Key Laboratory of Software Engineering**, **Wuhan University**, **Wuhan 430072**, **China**.

Zhangxin Chen:  
**Center for Scientific Computation and Department of Mathematics**, **Southern Methodist University**, **Dallas, TX 75275-0156 USA**.  
**E-mail**: zchen@mail.smu.edu
MOD M NORMALITY OF $\beta$-EXPANSIONS

YOUNG-HO AHN

ABSTRACT. If $\beta > 1$, then every non-negative number $x$ has a $\beta$-expansion, i.e.,

$$x = \varepsilon_0(x) + \frac{\varepsilon_1(x)}{\beta} + \frac{\varepsilon_2(x)}{\beta^2} + \cdots$$

where $\varepsilon_0(x) = [x]$, $\varepsilon_1(x) = [\beta(x)]$, $\varepsilon_2(x) = [\beta((\beta x))]$, and so on ($[x]$ denotes the integral part and $(x)$ the fractional part of the real number $x$). Let $T$ be a transformation on $[0, 1]$ defined by $x \to ((\beta x))$. It is well known that the relative frequency of $k \in \{0, 1, \cdots, [\beta]\}$ in $\beta$-expansion of $x$ is described by the $T$-invariant absolutely continuous measure $\mu_\beta$. In this paper, we show the mod $M$ normality of the sequence $\{\varepsilon_n(x)\}$.

1. INTRODUCTION

Let $(X, B, \mu)$ be a probability space and $T$ a measure preserving transformation on $X$. A transformation $T$ on $X$ is called ergodic if the constant function is the only $T$-invariant function and it is called weakly mixing if the constant function is the only eigenfunction with respect to $T$. A measure preserving transformation $T$ is called an exact transformation if $\cap_{n=0}^{\infty} T^{-n}B$ is the trivial $\sigma$-algebra consisting of empty set and whole set modulo measure zero sets. So exact transformation are as far from being invertible as possible. Recall that if a transformation is exact then that transformation is weakly mixing [10].

Let $X = \{x \mid 0 \leq x < 1\}$ be the compact group of real numbers modulo 1, and let $\theta \in X$ be irrational. The numbers $j\theta, j = 0, \pm 1, \cdots, n$, comprise a dense subgroup of $X$. For each interval $I \subset X$ and $n > 0$ define $S_n = S_n(\theta, I)$ to be the numbers of integers $j, 1 \leq j \leq n$, such that $j\theta \in I$. By Kronecker-Weyl theorem $\lim_{n \to \infty} \frac{S_n}{n} = \mu(I)$, where $\mu$ is Lebesgue measure on $X$. Veech [9] is interested in the behavior of the sequence $\{d_n\}$ of parities of $\{S_n\}$. That is, $d_n$ is 0 or 1 as $S_n$ is even or odd. i.e., He investigates the existence of the limit

$$\mu_\theta(I) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} d_n,$$

2000 Mathematics Subject Classification: 28D05, 47A35.
Key words and phrases: $\beta$-expansions, $\beta$-transformation, coboundary, Mod $M$ normality.
Partially supported by Korea Research Foundation Grant(KRF-2003-037-C00007).
© THE KOREAN SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS, 2005
and he shows that a necessary and sufficient condition for $\mu_\theta(I)$ to exist for every interval $I \subset X$ is that $\theta$ has bounded partial quotients and show that $d_n$ is evenly distributed if the length of the interval is not an integral multiple of $\theta$ modulo 1.

For a given $\beta > 1$, consider a $\beta$-transformation, $T$ on $[0, 1)$ defined by $x \mapsto (\beta x)$.

In this paper, we are interested in the uniform distribution of the sequence $d_n \in \{0, \cdots, M - 1\}$ defined by

$$d_n(x) \equiv \sum_{k=0}^{n-1} 1_E(T^k x) \pmod M,$$

for $\beta$-transformations on the interval where $1_E(x)$ is an indicator function of finite union of intervals, $E$. If $E$ is the form of $[\frac{k}{\beta}, \frac{k+1}{\beta})$ when $\frac{k+1}{\beta} < 1$ or $[\frac{k}{\beta}, 1)$ otherwise, then the distributions of $d_n(x)$ are just the mod $M$ normality of the sequence $\{\epsilon_n(x)\}$.

In [1], Ahn and Choe consider the case when transformations defined by $x \mapsto (Lx)$ with $L \in \mathbb{N}$ on $X = [0, 1)$ and $M = 2$, and show that the sequence $\{d_n\}$ is evenly distributed if $\exp(\pi i 1_E(x))$ has finite $L$-adic discontinuity points $\frac{1}{L} \leq t_1 < \cdots < t_n \leq 1$. Recently, Choe, Hamachi and Nakada[4] show that $\{d_n\}$ is evenly distributed for more general sets and that $\mathbb{Z}_2$-extension induced by $\phi(x) = \exp(\pi i 1_B(x))$ where $1_B$ is a characteristic function of $B$, is ergodic. In this paper, we show that for all $\beta$-transformations on the unit interval, the sequence $\{d_n\}$ is uniformly distributed and that the corresponding compact group extension of $\beta$-transformation is weakly mixing. Hence the compact group extension by $\phi(x)$ is exact.

2. Properties of $\beta$-expansions

In this section, we recall the basic properties of $\beta$-expansions. For more related results, see [3, 7, 8]. For every $\beta > 1$, there is a unique $T_\beta$-invariant absolutely continuous normalized measure $\mu_\beta$ and $T_\beta$ is an exact transformation on $(X, B, \mu_\beta)$. By the Radon-Nikodym theorem, there is a measurable function $h_\beta(x)$, essentially unique, such that

$$\mu_\beta(E) = \int_E h_\beta(x) \, dx.$$

We call $h_\beta(x)$ as the density function of $\mu_\beta$.

**Theorem 1.** Let $T$ be the transformation on $[0, 1)$ defined by $T(x) = (\beta x)$. Then the density function of $T$-invariant absolutely continuous measure $h_\beta(x)$ is described by

$$h_\beta(x) = \frac{1}{F(\beta)} \sum_{n,T^n(1) > x} \frac{1}{\beta^n} \quad \text{where} \quad F(\beta) = \int_0^1 \left( \sum_{n,T^n(1) > x} \frac{1}{\beta^n} \right) \, dx.$$

**Proof.** Since $\mu_\beta E = \mu_\beta T^{-1} E$ for all Lebesgue measurable set $E$,

$$\mu_\beta[a, b] = \int_a^b h_\beta(x) \, dx = \mu_\beta T^{-1}[a, b] = \sum_{m=0}^{[\beta-b]} h_\beta \left( \frac{x + m}{\beta} \right) \quad \text{a.e.}$$
if \([\beta - b]\) is the largest integer \(m\) for which both \(\frac{a+m}{\beta}\) and \(\frac{b+m}{\beta}\) are less than 1, and either \(a < (\beta)\) and \(b < (\beta)\) or \(a > (\beta)\) and \(b > (\beta)\). In this case,

\[
\beta \left( \int_{a}^{b} \frac{h_\beta(x)}{b-a} \, dx \right) = \sum_{m=0}^{[\beta-b]} \left( \left( \int_{a+m}^{b+m} \frac{h_\beta(x)}{b-a} \, dx \right) / \left( \frac{b+m}{\beta} - \frac{a+m}{\beta} \right) \right).
\]

Hence we show that the density function of \(\mu_\beta\), \(h_\beta(x)\) satisfies the following.

\[
\beta h_\beta(x) = \sum_{T_\gamma=x}^{[\beta-x]} h_\beta(y) = \sum_{m=0}^{[\beta-x]} h_\beta \left( \frac{x+m}{\beta} \right).
\]

Now we will show that \(h_\beta(x)\) defined by

\[
h_\beta(x) = \sum_{n,T^n(1) > x}^{\infty} \frac{1}{\beta^n},
\]

satisfies \(\beta h_\beta(x) = \sum_{m=0}^{[\beta-x]} h_\beta \left( \frac{x+m}{\beta} \right)\).

Let

\[
a_{n,m} = \begin{cases} 1, & \text{if } \frac{x+m}{\beta} < T^n(1), \\ 0, & \text{otherwise} \end{cases}
\]

and

\[
a_{n,m} = \begin{cases} 1, & \text{if } x < T^n(1), \\ 0, & \text{otherwise}. \end{cases}
\]

Note that \(\sum_{m=0}^{[\beta-x]} a_{n,m}\) is the number of \(m\) among \(0, 1, \ldots, [\beta-x]\) with \(\frac{x+m}{\beta} < T^n(1)\).

Hence

\[
\sum_{m=0}^{[\beta-x]} a_{n,m} = \begin{cases} [\beta T^n_\beta(1)] + 1, & \text{if } x < (\beta T^n(1)) = T^{n+1}(1) \\ [\beta T^n_\beta(1)], & \text{otherwise}. \end{cases}
\]

Thus

\[
\sum_{m=0}^{[\beta-x]} h_\beta \left( \frac{x+m}{\beta} \right) = \sum_{n=0}^{\infty} \sum_{m=0}^{[\beta-x]} \frac{a_{n,m}}{\beta^n} = \sum_{n=0}^{\infty} \sum_{m=0}^{[\beta-x]} a_{n,m} \frac{1}{\beta^n} = \sum_{n=0}^{\infty} \frac{[\beta T^n_\beta(1)]}{\beta^n} + a_{n+1} \frac{1}{\beta^n}
\]

\[
= \sum_{n=0}^{\infty} \frac{c_n(\beta)}{\beta^n} + \beta \sum_{n=0}^{\infty} \frac{a_{n+1}}{\beta^{n+1}} = \beta + \beta \sum_{n=0}^{\infty} \frac{a_n}{\beta^n} - \beta a_0
\]

\[
= \beta h_\beta(x).
\]

Finally, consider the following formula,
\[ F(\beta) = \int_0^1 \left( \sum_{n,T^n(1) > x} \frac{1}{\beta^n} \right) \, dx = \int_0^1 \left( \sum_{n=0}^{\infty} \frac{a_n(x)}{\beta^n} \right) \, dx = \int_0^1 \left( \sum_{n=0}^{\infty} \frac{T^n(1)}{\beta^n} \right) \, dx = \sum_{n=0}^{\infty} \frac{1}{\beta^n} \int_0^1 a_n(x) \, dx = \sum_{n=0}^{\infty} \frac{T^n(1)}{\beta^n} \]

So we have

\[ 1 \leq F(\beta) \leq 1 + \frac{1}{\beta} + \frac{1}{\beta^2} + \cdots = \frac{\beta}{\beta - 1}. \]

Hence \( h_\beta(x) \) is well defined. By the ergodicity of \( T \), the proof is completed. \( \square \)

**Remark 1.** We call those \( \beta \) which have recurrent tails, i.e., \( \varepsilon_{n+k}(\beta) = \varepsilon_n(\beta) \) for all \( n \geq N \) in their \( \beta \)-expansions, as \( \beta \)-numbers. Those with zero tails, we call simple \( \beta \)-number. It is easy to know that \( h_\beta(x) \) is a step function with finite discontinuity if and only if \( \beta \) has a recurrent tail in its \( \beta \)-expansion by the previous Theorem.

**Example 1.** Let \( \beta = \frac{\sqrt{5} + 1}{2} \). Then \( (\beta) = \beta - 1 = \frac{1}{\beta} \). So \( T^2(1) = 1 \). Hence by the formula of the previous Theorem,

\[
h_\beta(x) = \begin{cases} 
\frac{5 + 3\sqrt{5}}{2}, & \text{for } 0 \leq x < \frac{\sqrt{5} - 1}{2}, \\
\frac{5 + \sqrt{5}}{2}, & \text{for } \frac{\sqrt{5} - 1}{2} \leq x < 1.
\end{cases}
\]

3. **Mod \( M \) normality and coboundary equations of \( \beta \)-transformations**

Let \( G \) be a finite subgroup of the circle group \( \mathbb{T} \) generated by \( \exp(\frac{2\pi i}{M}) \). To investigate the sequence \( \{d_n(x)\} \), we consider the behavior of the sequence \( \exp(\frac{2\pi i}{M} d_n(x)) \) and check whether this sequence is uniformly distributed on compact group \( G \) generated by \( \exp(\frac{2\pi i}{M}) \). Weyl’s criterion on uniform distribution says that the sequence \( \exp(\frac{2\pi i}{M} d_n(x)) \) is uniformly distributed if and only if

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \exp(k \frac{2\pi i}{M} d_n(x)) = 0
\]

for all \( 1 \leq k \leq M - 1 \) [5].

We investigate the problem from the viewpoint of spectral theory. Let \( (X, \mu) \) be a probability space and \( T \) an ergodic measure preserving transformation on \( X \), which is not necessarily invertible. Let \( \phi(x) \) be a \( G \)-valued function defined by \( \phi(x) = \exp(\frac{2\pi i}{M} 1_E(x)) \). Consider the skew product transformation \( T_\phi \) on \( X \times G \) defined by \( T_\phi(x, g) = (Tx, \phi(x)g) \). Then
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{1}^{N} \exp^{k} \left( \frac{2\pi i}{M} d_{n}(x) \right) \cdot z^{k} = \lim_{N \to \infty} \frac{1}{N} \sum_{1}^{N} (U_{T_{\phi}})^{n} f(x, z)
\]

where \( U_{T_{\phi}} \) is an isometry on \( L^{2}(X \times G) \) induced by \( T_{\phi} \) and \( f(x, z) = z^{k} \). Hence if \( T_{\phi} \) is ergodic, then \( \lim_{N \to \infty} \frac{1}{N} \sum_{1}^{N} \exp^{k} \left( \frac{2\pi i}{M} d_{n}(x) \right) = 0 \) by an application of the Birkhoff's Ergodic theorem to \( f(x, z) = z^{k} \). Recall that the dual group of \( G \) consists of the trivial homomorphism 1 and \( \gamma_{k} \) defined by \( \gamma_{k}(z) = z^{k} \) for \( 1 \leq k \leq M - 1 \). Hence \( L^{2}(X \times G) = \bigoplus_{k=0}^{L-1} L^{2}(X) \cdot z^{k} \) and each \( L^{2}(X) \cdot z^{k} \) is an invariant subspace of \( U_{T_{\phi}} \). If \( f(x, z) \) is an eigen-function with eigenvalue \( \lambda \) then \( f(x, z) = \sum_{k=0}^{L-1} f_{k}(x) \cdot z^{k} \) and
\[
U_{T_{\phi}} f(x, z) = \sum_{k=0}^{L-1} \phi^{k}(x) f_{k}(Tx) \cdot z^{k}.
\]

Thus \( \phi^{k}(x) f_{k}(Tx) = \lambda f_{k}(x) \) for each \( k \). Hence to check the ergodicity we only need to know whether there exist \( 0 \leq k \leq M - 1 \) and \( f(x) \) such that \( \phi^{k}(x) f(Tx) = f(x) \).

Recall that a function \( f(x) \) is called a quasicoboundary if \( f(x) = \lambda \cdot q(x) q(Tx) \), \( |q(x)| = 1 \), \( |\lambda| = 1 \) a.e. on \( X \). Specially if \( \lambda = 1 \) then \( f(x) \) is called a coboundary. Similarly a real valued function \( g(x) \) is called an additive quasicoboundary if \( g(x) = k + q(Tx) - q(x) \) \( k \in \mathbb{R} \). Hence if \( g(x) \) is an additive quasicoboundary then \( f(x) \equiv \exp(2\pi i g(x)) \) is a quasicoboundary.

**Proposition 1.** Let \( T \) be an ergodic transformation on \( X \) and \( \phi(x) \) be a \( G \)-valued function. Let \( T_{\phi} \) be the skew product transformation defined by \( T_{\phi}(x, g) = (Tx, \phi(x) \cdot g) \) on \( X \times G \). If \( \phi(x) h(Tx) = h(x) \), then there exists a \( G \)-valued function \( q(x) \) such that the following diagram commutes
\[
\begin{array}{ccc}
X \times G & \xrightarrow{T_{\phi}} & X \times G \\
\downarrow Q & & \downarrow Q \\
X \times G & \xrightarrow{S} & X \times G
\end{array}
\]

where \( Q(x, g) = (x, q(x) g) \) and \( S(x, g) = (Tx, g) \). Hence \( T_{\phi} \) has \( M \) ergodic components.

**Proof.** Since \( (\phi(x))^{M} = 1 \), \( (\phi(x))^{M}(h(Tx))^{M} = (h(x))^{M} \) is equivalent to \( (h(Tx))^{M} = (h(x))^{M} \). So we may assume that \( (h(x))^{M} = 1 \) by the ergodicity of \( T \). Hence there exist a \( G \)-valued function \( q(x) \) such that \( \phi(x) q(Tx) = q(x) \). For this \( q(x) \), it turns out that the diagram commutes by easy consideration.

**Lemma 1.** Let \( \tau \) be a piecewise twice continuously differentiable function such that \( \inf_{x \in J_{1}} |\tau'(x)| > 1 \) where \( J_{1} = \{ x \in X, \tau'(x) \text{ exists} \} \). If the number of discontinuity points of \( \tau \) or \( \tau' \) is finite, then there is a finite collection of sets \( L_{1}, \cdots, L_{n} \) and a set of invariant function \( \{ f_{1}, \cdots, f_{n} \} \) such that

1. each \( L_{i} (1 \leq i \leq n) \) is a finite union of closed intervals;
2. \( L_{i} \cap L_{j} \) contains at most a finite number of points when \( i \neq j \);
(3) \( f_i(x) = 0 \) for \( x \notin L_i, 1 \leq i \leq n \), and \( f_i(x) > 0 \) for a.e. \( x \) in \( L_i \);
(4) \( \int_{L_i} f_i(x) dx = 1 \) for \( 1 \leq i \leq n \);
(5) every \( \tau \) invariant function can be written as \( f = \sum_{i=1}^{n} a_i f_i \) with suitable chosen \( \{a_i\} \).

Proof. For the proof, See [2, 6]. \( \square \)

**Proposition 2.** For the \( \beta \)-transformation, if a \( G \)-valued function \( \phi(x) \) is a step function with finite discontinuity points and \( \phi(x)h(Tx) = h(x) \), then there exist \( G \)-valued function \( q(x) \) which is a step function with finite discontinuity points and \( \phi(x)q(Tx) = q(x) \). Hence we may assume that \( h(x) \) is a \( G \)-valued step function with finite discontinuity points.

Proof. Assume that \( \phi(x)h(Tx) = h(x) \). Without loss of generality assume that \( X = [0,1) \). Since \( X \times G = \bigcup_{k=0}^{M-1} \{X \times \exp(\frac{2\pi i}{M})\} \), let's identify \( \{X \times \exp(\frac{2\pi i}{M})\} \) with the unit interval \( [k,k+1) \), \( 0 \leq k < M \). Since \( \phi(x) \) is a \( G \)-valued step function with finite discontinuity points, we can regard \( T_{\phi} \) as a piecewise continuous map on \([0,M)\) satisfying the condition of Lemma 1. So there exist a \( G \)-valued function \( q(x) \) which is a step function with finite discontinuity points by Lemma 1 and Proposition 1. \( \square \)

To investigate the mod \( M \) normality of \( \beta \)-transformation, we consider a function \( \phi(x) = \exp(\frac{2\pi i}{M}1_E(x)) \). In the following two Lemmas, we consider more general functions \( \phi(x) \) with finite discontinuity points.

**Proposition 3.** For \( \beta \)-transformation, if a \( G \)-valued nonconstant function \( \phi(x) \) is a step function with finite discontinuity points \( \frac{1}{\beta} \leq t_1 < \cdots < t_n < 1 \), then \( \phi(x) \) is not coboundary.

Proof. Assume that \( \phi(x)h(Tx) = h(x) \). Since \( \phi(x) \) is a step function with finite discontinuity points, \( h(x) \) is also a step function with finite discontinuity points. Hence there exist \( 0 < r \leq \frac{1}{\beta} \) such that \( h(x) \) is constant on \([0,r)\). There exist \( x_0 \in (0,r) \) such that \( \beta x_0 \) is also in \((0,r)\). Since \( \phi(x)h(Tx) = h(x) \), we have \( \phi(x_0)h(Tx_0) = h(x_0) \). Hence \( \phi(x_0) = 1 \) on \([0,\frac{1}{\beta})\). Hence \( h(Tx) = h(x) \) for all \( x \in [0,\frac{1}{\beta}) \). Since \( T(0,\frac{1}{\beta}) = [0,1) \), \( h(x) \) has to be constant, i.e., \( \phi(x) \equiv 1 \). This contradicts to the assumption of \( \phi(x) \). \( \square \)

**Example 2.** Let \( \beta = \frac{\sqrt{5} + 1}{2} \). For the transformation on \([0,1)\) defined by \( x \rightarrow (\beta x) \), let's consider the following. Let \( I = [\frac{15}{\beta},1) \), \( F = \bigcup_{k=0}^{\infty} \frac{1}{\beta^k}I \) and \( E = F \triangle T^{-1}F \). Then \( \phi(x) = \exp(\pi i 1_E(x)) \) is a coboundary even if the discontinuity points of \( \phi(x) \) are contained in \([\frac{1}{\beta},1)\) where the cobounding function is \( h(x) = \exp(\pi i 1_F(x)) \). Hence the assumption of finite discontinuity points on \( \phi(x) \) can't be omitted.

Let \( \phi(x) = \exp(\frac{2\pi i}{M}1_E(x)) \) where \( E \) is the form of \([\frac{k}{\beta}, \frac{k+1}{\beta}) \) when \( \frac{k+1}{\beta} < 1 \) or \([\frac{k}{\beta},1) \) otherwise. By the previous Proposition, \( \phi(x) \) is not coboundary. Hence we have the following Theorem.
Theorem 2. Let $\beta > 1$ be given and we have a $\beta$-expansion for nonnegative number $x$, i.e.,

$$x = \varepsilon_0(x) + \frac{\varepsilon_1(x)}{\beta} + \frac{\varepsilon_2(x)}{\beta^2} + \ldots$$

where $\varepsilon_0(x) = [x]$, $\varepsilon_1(x) = [\beta(x)]$, $\varepsilon_2(x) = [\beta((\beta x))]$, and so on ($[x]$ denotes the integral part and $(x)$ the fractional part of the real number $x$). Then the sequence $\{\varepsilon_n(x)\}$ satisfies the mod $M$ normality almost everywhere.

References


Young-Ho Ahn:
Department of Mathematics, Mokpo National University, 534-729, Korea
E-mail: yhahn@mokpo.ac.kr
FUNCTION APPROXIMATION OVER TRIANGULAR DOMAIN
USING CONSTRAINED Legendre POLYNOMIALS

YOUNG JOON AHN

ABSTRACT. We present a relation between the orthogonality of the constrained Legendre polynomials over the triangular domain and the BB (Bézier -Bernstein) coefficients of the polynomials using the equivalence of orthogonal complements. Using it we also show that the best constrained degree reduction of polynomials in BB form equals the best approximation of weighted Euclidean norm of coefficients of given polynomial in BB form from the coefficients of polynomials of lower degree in BB form.

1. INTRODUCTION

Degree reduction of Bézier curves is one of the important problems in CAGD (Computer Aided Geometric Design) or CAD/CAM. In general, degree reduction cannot be done exactly so that it invokes approximation problems. Thus many efforts and proposals for dealing with the problems have been made in the recent twenty years or so. They are classified by different norm in which the distance between polynomials is measured, e.g., in $L_{\infty}$-norm [7, 13], in $L_2$-norm [15, 16, 17], in $L_1$-norm [11] or in $L_p$-norm [5, 10], etc. Furthermore, the constrained degree reduction of Bézier curves with $C^{a-1}$-continuity at both end points is developed in many previous literature [1, 2, 4, 6, 8, 12, 14, 17, 19, 20].

Recently, Lutterkort et al. [18] showed that the orthogonal complement of a subspace in the polynomial space of degree $n$ over the triangular domain with respect to the $L_2$-inner product and the Euclidean inner product of BB coefficients are equal. Using this fact they also showed that the best degree reduction of polynomial $f$ of degree $n$ over the triangular domain in $L_2$-norm is equivalent to the best approximation of the vector of BB coefficients of $f$ from all vector of BB coefficients of degree elevated polynomials of degree less than $n$ in the Euclidean norm of the vector. We follow their results in the case of constrained degree reduction over the triangular domain. We first show that the orthogonal complement of a subspace in the constrained polynomial space of degree $n$ over the triangular domain with respect to $L_2$-inner product and the

2000 Mathematics Subject Classification: 65D05,65D07,65D17.
Key words and phrases: Triangular domain, Legendre polynomial, Bernstein polynomial, Bézier surface, weights, Constrained degree reduction.
© THE KOREAN SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS, 2005
weighted Euclidean inner product of BB coefficients are equal for some weights. Using
the fact we also show that the best constrained degree reduction of $f$ of degree $n$ over
the triangular domain in $L_2$-norm is equal to the best approximation of the vector
of coefficients from all vectors of coefficients of degree elevated polynomials with the
constraint in weighted Euclidean norm of vectors. Finally we present a relation between
the orthogonality of the constrained Legendre polynomials over the triangular domain
and the BB (Bézier -Bernstein) coefficients of the polynomials using the equivalence
of orthogonal complements. The relation plays an important role to construct the
constrained Legendre polynomials in BB form.

The outline of this paper is as follows. In Section 2, we explain the constrained
polynomial space over triangular domain and the constrained Legendre polynomials.
In Section 3, we show that the orthogonal complement of a subspace in the constrained
polynomial space of degree $n$ over triangular domain with respect to $L_2$-inner product
and the weighted Euclidean inner product of BB coefficients are equal. In Section 4,
we present the properties of the best constrained degree reduction in BB form and the
constrained Legendre polynomials over triangular domain in BB form.

2. CONSTRAINED LEGENDRE POLYNOMIALS OVER TRIANGULAR DOMAIN

In this section we consider the constrained Legendre polynomial of degree $n$ over
triangular domain. Let $T$ be a triangle in the plane, defined by vertices $p_k = (x_k, y_k)$
for $k = 0, 1, 2$. If these vertices are not collinear, any point $p \in T$ can be written
uniquely in terms of barycentric coordinates $u, v, w$ where $u + v + w = 1$, with respect
to $T$:
\[ p = up_0 + vp_1 + wp_2. \]
Let $\mathbb{P}_n$ be the linear space of polynomials of degree less than or equal to $n$. It is conve-
nient to introduce the compact notation $\alpha = (\alpha_1, \alpha_2)$ to denote doubles of nonnegative
integers, and we write $|\alpha| = \alpha_1 + \alpha_2$. The Bernstein basis of degree $n$ over $T$ is denoted by
\[ B^\alpha_n(u, v) = \frac{n!}{\alpha_1!\alpha_2!(n - \alpha_1 - \alpha_2)!} u^{\alpha_1} v^{\alpha_2} (1 - u - v)^{(n - \alpha_1 - \alpha_2)}, \quad |\alpha| \leq n. \]
Thus $\mathbb{P}_n$ has exactly $(n + 1)(n + 2)/2$ basis functions. We collect the basis functions in
triangular arrays of size $n$
\[ B^n := [B^\alpha_n]_{|\alpha| \leq n} \]
and with $b = [b_\alpha]_{|\alpha| \leq n}$ a simplicial array of reals we write polynomials in BB form as
\[ B^n b = \sum_{|\alpha| \leq n} B^\alpha_n b_\alpha. \]
For the nonnegative integer $a \leq n/3$, $\mathbb{P}_m^n$ is denoted by the linear space of the con-
strained polynomials of degree less than or equal to $m$ over the triangle $T$ as follows:
\[ \mathbb{P}_m^n = \{ B^n b \in \mathbb{P}_m^n : b_\alpha = 0 \text{ for } \alpha \in J^n_m \}. \]
where $I^a_n$ and $J^a_n$ are the sets of double index $\alpha$ such that

\[ I^a_n = \{ |\alpha| \leq n : \alpha_1 \geq a, \alpha_2 \geq a, |\alpha| \leq n - a \} \]
\[ J^a_n = \{ |\alpha| \leq n : \alpha \not\in I^a_n \} \]

Then $P^a_m$ has exactly $(k+1)(k+2)/2$ basis functions, where $k = m - 3a$. Farouki et al. [9] constructed the basis of linearly independent and mutually orthogonal polynomials $L_{m,i}$, say Legendre polynomials, with hierarchical ordering in BB form over $T$. We also consider the constrained Legendre polynomials $L^a_{m,i}$ with hierarchical ordering in BB form over $T$. For example $a = 1$,

1 degree 3 basis functions $L^1_{3,0}$
2 degree 4 basis functions $L^1_{4,0}, L^1_{4,1}$
\[ \vdots \]
$m - 2$ degree $n$ basis functions $L^1_{m,0}, \ldots, L^1_{m,m-3}$

We present a relation between the orthogonality of the constrained Legendre polynomials and the BB (Bézier - Bernstein) coefficients of the polynomials in section 4 using the equivalence of orthogonal complements in section 3.

3. EQUIVALENCE OF ORTHOGONAL COMPLEMENTS

For $a \leq m/3$, let

\[ Q^a_m = \{ f(u,v) \in P_m : f(i,j) = 0 \text{ for } i,j,n-i-j = 0, \ldots, a - 1 \}, \]

which was also introduced for one variable by Ahn et al. [3]. Note that $P_m = P^0_m = Q^0_m$. We consider the Lagrange polynomials characterized by

\[ Q^a_\alpha(\beta) = \delta_{\alpha,\beta}, \quad |\alpha|, |\beta| \leq n. \]

Peters and Reif [18] was already introduced the notation of the Lagrange polynomials $Q^a_\alpha$. We collect the basis functions in triangular arrays of size $n$

\[ Q^n := [Q^a_\alpha]_{|\alpha| \leq n} \]

with $b = [b_\alpha]_{|\alpha| \leq n}$ a simplicial array of reals we write polynomials in Lagrange form as

\[ Q^n b = \sum_{|\alpha| \leq n} Q^n_\alpha b_\alpha. \]

The Lagrange form is used to relate a discrete polynomial dependence of the coefficients on the array index to a continuous polynomial. For example, if the coefficients $b_\alpha = (\alpha_1 \alpha_2 + \alpha_2)\alpha_1 \alpha_2(n - \alpha_1 - \alpha_2)$ depend quintically on the index $\alpha$, then $Q^5(u,v)b = (uv + v)uv(n - u - v)$ is the corresponding quintic polynomial. The following lemma is an extension of Lemma 2.1 in Lutterkort et al. [17], Lemma 3.1 in Ahn et al. [3] and Lemma 2.1 in Peters and Reif [18].
Lemma 3.1. A polynomial $B^n b$ is of degree $\leq m$ with $b_\alpha = 0$ for $\alpha \in J_n^a$ if and only if the triangular array of coefficients is a polynomial of degree $\leq m$ with zeros at $(i,j)$, $i,j,n-i-j=0,\ldots,a-1$ in its index, i.e.,

$$B^n b \in \mathbb{P}_m^a \Leftrightarrow Q^n b \in \mathbb{Q}_m^a.$$  

Proof. It is well-known [18] that

$$B^n b \in \mathbb{P}_m^0 \Leftrightarrow Q^n b \in \mathbb{Q}_m^0.$$  

Note that if the coefficients $b_\alpha$ depend on the index $\alpha$, then $Q^n(u,v)b$ is also the corresponding polynomial. Thus $b_\alpha = 0$ for $\alpha \in J_n^a$ is equivalent to $Q^n(i,j)b = 0$ for $i,j,n-i-j=0,\ldots,a-1$. Hence we have $B^n b \in \mathbb{P}_m^a$ if and only if $Q^n b \in \mathbb{Q}_m^a$. \hfill $\square$

Theorem 3.2. The orthogonal complements of $\mathbb{P}_m^a$ in $\mathbb{P}_n^a$ with respect to the $L_2$-inner product

$$\langle f,g \rangle_L := \int_T \int_T f(x)g(x)dx$$

and the weighted Euclidean inner product of the BB coefficients

$$\langle B^n b, B^n c \rangle_W := \sum_{\alpha \in I_n^a} b_\alpha c_\alpha w_\alpha$$

where

$$w_\alpha := \begin{cases} \frac{(2a_1)(2a_2)(2(n-|\alpha|))}{(a_1)(a_2)(n-|\alpha|)} & (\alpha \in I_n^a) \\ \frac{(2a_1)(2a_2)(2(n-|\alpha|))}{(a_1-a)(a_2-a)(n-|\alpha|-a)} & (\alpha \in J_n^a) \\ 1 & (\alpha \in J_n^a) \end{cases}$$

are equal.

Proof. Denote the orthogonal complement of $\mathbb{P}_m^a$ in $\mathbb{P}_n^a$ with respect to the weighted Euclidean inner product by $\mathbb{P}_{m,n}^a$, and let $\{B^n q^\alpha : m-3a < |\alpha| < n-3a\}$ be some basis of this space. By equality of dimensions it suffices to show that $\mathbb{P}_{m,n}^a$ is contained in the orthogonal complement with respect to the $L_2$-inner product, i.e., the polynomials $B^n w^\alpha$ have to be $L_2$-orthogonal to all polynomials in $\mathbb{P}_m^a$,

$$\langle B^n q^\alpha, u^{a+\beta_1}v^{a+\beta_2}(1-u-v)^a \rangle_L = 0, \quad (0 \leq |\beta| \leq m-3a < |\alpha| \leq n-3a).$$

Defining the triangular array $p_\alpha^\beta$ by

$$p_\alpha^\beta := \frac{1}{w_\alpha} \int_T \int_T B^\alpha(u,v) u^{a+\beta_1}v^{a+\beta_2}(1-u-v)^a dA,$$

clearly we have

$$\langle B^n q^\alpha, u^{a+\beta_1}v^{a+\beta_2}(1-u-v)^a \rangle_L = \langle B^n q^\alpha, B^n p^\beta \rangle_W.$$

By definition, the latter expression vanishes if and only if $B^n p^\beta \in \mathbb{P}_m^a$, and by Lemma 3.1, this is equivalent to $Q^n p^\beta \in \mathbb{Q}_m^a$. In other words, we have to show that $p_\alpha^\beta$ is a
polynomial in $\alpha$ of degree $\leq m$ with zeros at $\alpha \in J_n^a$, for all $\beta$ with $|\beta| \leq m - 3a$. Using the formula

$$\int \int_T B_\alpha^n(u, v) dA = \frac{1}{(n + 1)(n + 2)}.$$

we have

$$p_\alpha^\beta = \frac{1}{w_\alpha} \int \int_T B_\alpha^n(u, v) u^{a + \beta_1 a + \beta_2 (1 - u - v)^a} dA$$

$$= \frac{1}{w_\alpha} \left( \frac{n}{n + |\beta| + 3a} \right)^{\beta_1} \left( \frac{n}{n + |\beta| + 3a} \right)^{\beta_2} \int \int_T B_{n + |\beta| + 3a}^{n + (a, a)}(u, v) dA$$

$$= \frac{1}{w_\alpha} \left( \frac{n}{n + |\beta| + 3a} \right)^{\beta_1 + \beta_2 + a} \left( \frac{n}{n + |\beta| + 3a} \right)^{n + 1} \frac{1}{(n + |\beta| + 3a + 1)(n + |\beta| + 3a + 2)}$$

Now,

$$\frac{1}{w_\alpha} = \frac{(2\alpha_1)(2\alpha_2)(2(n - |\alpha|))}{(\alpha_1 - a)(\alpha_2 - a)(2(n - |\alpha|))} \frac{(2(n - |\alpha|))}{(n - |\alpha|)}$$

$$= \frac{\alpha_1! \alpha_2! (n - |\alpha|)!^2}{(\alpha_1 - a)! (\alpha_2 - a)! (n - |\alpha| - a)! (n - |\alpha| + a)!}$$

and

$$\frac{n}{n + |\beta| + 3a} = \frac{n!}{(n + |\beta| + 3a)!} \frac{(\alpha_1 + \beta_1 + a)! (\alpha_2 + \beta_2 + a)! (n - |\alpha| + a)!}{\alpha_1 \alpha_2! (n - |\alpha|)!}$$

so that

$$p_\alpha^\beta = \frac{n!}{(n + |\beta| + 3a + 2)!} \times \frac{\alpha_1! \alpha_2! (n - |\alpha|)!}{(\alpha_1 - a)! (\alpha_2 - a)! (n - |\alpha| - a)!} \times \frac{(\alpha_1 + \beta_1 + a)! (\alpha_2 + \beta_2 + a)!}{(\alpha_1 + a)! (\alpha_2 + a)!}$$

$$\times \prod_{l_1=1}^{a} (\alpha_1 - a + l_1) \prod_{l_2=1}^{a} (\alpha_2 - a + l_2) \prod_{l_3=1}^{a} (n - |\alpha| - a + l_3)$$

$$\times \prod_{r_1=1}^{\beta_1} (\alpha_1 + a + r_1) \prod_{r_2=1}^{\beta_2} (\alpha_2 + a + r_2).$$

Thus $p_\alpha^\beta$ is a polynomial of degree $\leq m$ and $p_\alpha^\beta = 0$ for $\alpha \in J_n^a$ i.e., $\alpha_1 = 0, \ldots, a - 1, \alpha_2 = 0, \ldots, a - 1$ and $n - |\alpha| = 0, \ldots, a - 1$. \hfill $\square$

In particular, in one dimensional case, i.e., $d = 1$, the weights are

$$w_\alpha = \frac{(2\alpha_1)(2(n - |\alpha|))}{(\alpha_1 - a)(n - |\alpha| - a)} = \frac{(2\alpha_1)! (2(n - |\alpha|))!}{\alpha_1!^2 (n - |\alpha|)!^2} \frac{(2n)!}{(\alpha_1 - a)! (\alpha_1 + a)! (n - |\alpha| - a)!} = \frac{\alpha_1^2}{\alpha_1 - a} \frac{n^2}{\alpha_1 + a}.$$
which are the same weights given by Ahn. et al. [3].

4. Properties of constrained Legendre polynomials in BB form

**Theorem 4.1.** Given a polynomial $B^n b$ of degree $n$, the approximation problem

$$\min_{p \in \mathbb{P}_m} \{ \| B^n b - p \| : p = B^n c \in \mathbb{P}_m, \ b_\alpha = c_\alpha \ for \ \alpha \in J_n^a \}$$

has the same minimizer for the norm induced either by the $L_2$-inner product (3.1) or the weighted Euclidean inner product (3.2).

**Proof.** Let $f^a = B^n d \in \mathbb{P}_m$ be a polynomial of degree $m$ satisfying

$$b_\alpha = d_\alpha$$

for $\alpha \in J_n^a$. Then the polynomial $B^n b - f^a \in \mathbb{P}_n$ can be decomposed uniquely according to

$$B^n b - f^a = p^a + q^a, \quad p^a \in \mathbb{P}_m, \quad q^a \in \mathbb{P}_{m,n}$$

and, by the orthogonality, $p^a$ is the minimizer of

$$\min_{p^a \in \mathbb{P}_m} \| B^n b - f^a - p^a \|$$

for both norm. For all $p = B^n c \in \mathbb{P}_m$ satisfying

$$b_\alpha = c_\alpha$$

for $\alpha \in J_n^a$, we have

$$\| B^n b - p \| = \| B^n b - f^a - (p - f^a) \| \geq \| B^n b - f^a - p^a \|$$

since $p - f^a \in \mathbb{P}_m$. Thus $p = p^a + f^a \in \mathbb{P}_m$ is the wanted solution for both norms. \qed

**Corollary 4.2.** Denote by $\mathcal{P}_{m,n}^a$ the linear operator mapping polynomials $B^n b \in \mathbb{P}_n$ to their best constrained $L_2$-norm or weighted Euclidean approximant $p \in \mathbb{P}$. Then

$$\mathcal{P}_{m,n}^a = \mathcal{P}_m^a \mathcal{P}_{m,n}^a, \quad m \leq \ell \leq n.$$

For $m \geq 1$, consider the space $\mathcal{L}_m$ of degree-$m$ polynomials that are orthogonal to all polynomials of degree $< m$ over $T$:

$$\mathcal{L}_m^a = \{ p \in \mathbb{P}_m : p \perp \mathbb{P}_{m-1}^a \}.$$

The following theorem is an extension of Lemma 4 in Farouki et al. [9].
Theorem 4.3. Let \( p = B^n c \in \mathbb{P}_n \). Then we have

\[
p \in \mathcal{L}^a_n \iff \sum_{\alpha \in I_n^a} b_\alpha c_\alpha w_\alpha = 0 \text{ for all } q = B^n b \in \mathbb{P}_{n-1}^a
\]

Proof. By definition of \( \mathcal{L}^a_n \), \( p \in \mathcal{L}^a_n \) if and only if \( p \perp q \) for all \( q \in \mathbb{P}_{n-1}^a \). It is equivalent to

\[
\int_T \int_T pqdA = 0.
\]

By Theorem 3.2, These are equivalent to

\[
\sum_{\alpha \in I_n^a} c_\alpha b_\alpha w_\alpha = 0.
\]

\( \square \)

Let \( f(u, v) \) be a given \( C^{a-1} \) function in \( T \). If \( q(u, v) \) is chosen in \( \mathbb{P}_n \) so that \( q \) is the \( C^{a-1} \)-interpolation of \( f \) at the boundaries, and \( P \in \mathbb{P}_n^a \) satisfies

\[
\min_{P \in \mathbb{P}_n^a} \int_T [(f(u, v) - q(u, v)) - P(u, v)]^2 dA,
\]

then the polynomial \( q(u, v) - P(u, v) \) is a constrained \( n \)-the degree polynomial approximation of \( f(u, v) \). If the constrained Legendre polynomials are constructed, then \( P(u, v) \) can be simply obtained as

\[
P(u, v) = \sum_{r=0}^{n} \sum_{i=0}^{r} \ell_{r,i} L_{r,i}(u, v)
\]

with

\[
\ell_{r,i} = \frac{< L_{r,i}, f - q >}{< L_{r,i}, L_{r,i} >}.
\]

ACKNOWLEDGEMENTS

This work was supported by Korea Research Foundation Grant (KRF-2004-002-C00035).

References


Young Joon Ahn:
DEPARTMENT OF MATHEMATICS EDUCATION, CHOSUN UNIVERSITY, GWANGJU, 501–759, SOUTH KOREA
E-mail: ahn@chosun.ac.kr