DEGREE ELEVATION OF B–SPLINE CURVES AND ITS MATRIX REPRESENTATION

BYUNG-GOOK LEE AND YUNBEOM PARK

ABSTRACT. An algorithmic approach to degree elevation of B-spline curves is presented. The new algorithms are based on the blossoming process and its matrix representation. The elevation method is introduced that consists of the following steps: (a) decompose the B–spline curve into piecewise Bézier curves, (b) degree elevate each Bézier piece, and (c) compose the piecewise Bézier curves into B–spline curve.

1. Introduction

Degree elevation of B–spline curves are well understood and several algorithms are published([10], [11], [12], [13], [14], [15]). From a software engineering point of view, it is desirable to implement a simple and easy-to-understand algorithm. This approach was taken by Piegl and Tiller([11], [12], [13]), who implemented the simplest algorithm; they decomposed the B–spline curve into piecewise Bézier curves, elevated the degree of each Bézier piece, and then composed the piecewise Bézier curves into B–spline curves. We describe here the modified form of Piegl and Tiller’s algorithm. The presented procedure allow to elevate the degree from \( n \) to \( m \) in one step. The new algorithms are based on the blossoming analysis ([2], [8], [16], [17]) and matrix representation of the process.

2. Degree Elevation

Since a B–spline curve is a piecewise polynomial curve, it must be possible to elevate its degree from \( p \) to \( p + r \). That is, there must exist control points \( \tilde{P} \) and a knot vector \( \tilde{U} \) such that

\[
C_n^p(u) = C_{\tilde{P}}^{\tilde{n}}(u) = \sum_{i=0}^{\tilde{n}} \tilde{P}_i N_{i,p+r}(u).
\]

The curve \( C_n^p(u) \) and \( C_{\tilde{P}}^{\tilde{n}}(u) \) are the same geometrically and parametrically. The computing of \( \tilde{n} \), \( \tilde{P} \) and \( \tilde{U} \) is referred to as degree elevation. The knot vector \( \tilde{U} \) and number of points \( n \) can easily be computed as follows.

1991 Mathematics Subject Classification. 65D17, 68U05.

Key words and phrases. Degree Elevation, B–spline, Blossoming.

This work was supported by Dongseo University and Seowon University, grant 1999.
Assume that $U$ has the form
\[ U = \{a,a,\ldots,a,u_1,\ldots,u_1,\ldots,u_s,\ldots,u_s,b,\ldots,b\} \]
where the end knots $a$ and $b$ are repeated with multiplicity $p+1$, the interior knots $u_i$ are repeated with multiplicity $m_i$ and $s$ is the number of distinct interior knots. Since the curve $C_p^m(u)$ is $C^{p-m}$-continuous at the knot of multiplicity $m_i$, $C_p^m(u)$ must have the same continuity. Consequently, the new vector must take the form
\[ \tilde{U} = \{a,a,\ldots,a,u_1,\ldots,u_1,\ldots,u_s,\ldots,u_s,b,\ldots,b\} \]
which gives $\tilde{n} = n + (s + 2)r$.

The computation of $P$ can be done in a number of different ways, including solving a system of linear equations([3]), combining some techniques of box splines with knot insertion([14]), using blossoming analysis([10]), or approaching aspect of software engineering ([11]). We provide a procedural method that combine the blossoming analysis and the software engineering aspect. The procedure can be summarized as follows:

1. Decompose the B-spline curve into piecewise Bézier curves.
2. Degree elevate each Bézier piece.
3. Make the B-spline curve from the piecewise Bézier segment.

3. Curve Decomposition

Curve decomposition is normally done via knot insertion. This is very convenient in curve design when it is necessary to have local control. In the case of a nonuniform B-spline curve a new knot can be inserted to increase the number of control points and the number of curve segments. Consider the cubic B-spline curve in Fig. 1 with blossoming notation. Here, the Bézier segments are extracted by inserting knot 1 two times. From these we can get first Bézier piece $B_G(0,0,0)$, $B_G(0,0,1)$, $B_G(0,1,1)$, and $B_G(1,1,1)$, and second Bézier piece $B_G(1,1,1)$, $B_G(1,1,3)$, $B_G(1,3,3)$, and $B_G(3,3,3)$.

The mutiaffine and symmetry property can be used to compute new blossom values from old ones. For example, consider finding the value $B_G(0,1,1)$, $B_G(1,1,1)$, and $B_G(1,1,3)$.

\[
\begin{align*}
B_G(0,1,1) &= 2/3B_G(0,0,0) + 1/3B_G(0,1,0) = 2/3B_G(0,0,1) + 1/3B_G(0,1,1) \\
B_G(1,1,1) &= 2/3B_G(1,0,1) + 1/3B_G(1,1,0) = 2/3B_G(0,1,1) + 1/3B_G(1,1,1) \\
&= 4/9B_G(0,0,0) + 4/9B_G(0,1,0) + 4/9B_G(0,0,1) + 4/9B_G(0,1,1) \\
B_G(1,1,3) &= 2/3B_G(0,1,3) + 1/3B_G(3,1,3) = 2/3B_G(0,1,3) + 1/3B_G(1,3,3).
\end{align*}
\]

The Decomposition algorithm in Fig. 1 is expressed in a matrix form as
\[ Q = M_dP \]
Figure 1. Curve decomposition of a cubic B–spline curve. The control polygon (circle) after inserting knot \( u = 1 \) two times into the knot vector \( (0,0,0,0,1,3,3,3,3) \).

where

\[
Q = \begin{pmatrix}
B_G[0,0,0] \\
B_G[0,0,1] \\
B_G[0,1,1] \\
B_G[1,1,1] \\
B_G[1,1,3] \\
B_G[1,3,3] \\
B_G[3,3,3]
\end{pmatrix}, \quad M_d = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 2/3 & 1/3 & 0 & 0 \\
0 & 4/9 & 4/9 & 1/9 & 0 \\
0 & 0 & 2/3 & 1/3 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]

and

\[
P = \begin{pmatrix}
B_G[0,0,0] \\
B_G[0,0,1] \\
B_G[0,1,1] \\
B_G[1,1,1] \\
B_G[1,1,3] \\
B_G[1,3,3] \\
B_G[3,3,3]
\end{pmatrix}.
\]

Next we give detailed pseudocode to compute the decomposition matrix \( M_d \). It uses a local array \( V \) of size \( s \) to store the \( i \)th distinct interior knot values and another one \( M \) of size \( s \) to store the multiplicity.

\begin{verbatim}
Make_DecomposeBsplineMatrix(U, m, n, p)
    // Input: Knots vector U = \{a,a,\ldots,a,u_1,\ldots,u_1,\ldots,u_s,\ldots,u_s,b,b,\ldots,b\},
    // number of knots \( m + 1 \),
    // number of control points \( n + 1 \) and degree \( p \)
    // Output: \((ps + p + 1) \times (n + 1)\) matrix \( M_d \)

    (* In case of Bezier curve *)
    if (n = p) exit;
    (* initialize some variables *)
    s = 1; t = 1; l = p;
\end{verbatim}
(* initialize $M_d$ matrix *)
for ($i = 0$ to $n$ by 1)
    for ($j = 0$ to $n$ by 1)
        if ($i = j$) $M_d[i][j] = 1$; else $M_d[i][j] = 0$;
endfor
endfor

(* compute knot multiplicity *)
$V[s] = U[p + 1]$; $M[s] = 1$;
for ($i = p + 2$ to $n$ by 1)
    if ($V[s] = U[i]$) $M[s] = M[s] + 1$;
    else $s = s + 1$; $V[s] = U[i]$; $M[s] = 1$
endif
endfor

(* make $M_d$ matrix *)
for ($i = 1$ to $s$ by 1)
    $l = l + M[s]$;
    for ($j = M[i]$ to $p - 1$ by 1)
        KnotsInsertion ($U$, $l$, $V[i]$, $n$, $n + t$, $p$);
        for ($k = m + t$ to $l + 1$ by $-1$) $U[k] = U[k - 1]$;
        $U[l] = V[i]$; $l = l + 1$; $t = t + 1$
    endfor
endfor

KnotsInsertion ($U$, $l$, $v$, $n$, $k$, $p$)
// Input: Knots vector $U$, new knot $v \in [u_i, u_{i+1})$, and degree $p$
// Output: $(n+1) \times (k+1)$ matrix $M_d$

for ($i = l - p + 1$ to $l - 1$ by $1$)
    $\alpha = (v - U[i])/(U[i] + p - U[i])$;
    for ($j = 0$ to $n$ by 1)
        $T[i][j] = (1 - \alpha)M_d[i - 1][j] + \alpha M_d[i][j]$;
    endfor
endfor
for ($i = k$ to $l$ by $-1$)
    for ($j = 0$ to $n$ by 1) $M_d[i][j] = M_d[i - 1][j]$;
for ($i = l - p + 1$ to $l - 1$ by $1$)
    for ($j = 0$ to $n$ by 1) $M_d[i][j] = T[i][j]$;

4. Degree Elevation of Bézier Curves
   The degree elevation of Bézier curves is well understood and well documented([7]).
   As a first step, consider raising the degree of the Bézier curve by one. We can show that new control points are obtained from the old polygon by piecewise linear interpolation
at the parameter values \( j/(n+1) \). We may repeat this process and obtain a sequence of control points. After \( r \) degree elevation, we have a linear system \( R = T_{n,r}Q \) where the \((n + r + 1) \times (n + 1)\) matrix \( T_{n,r} = \{ t_{i,j} \} \) has elements \([9]\)

\[
(1) \quad t_{i+j,i} = \binom{n}{i} \binom{r}{j} \frac{(n+r+1)}{i+j}, \quad \begin{cases} i = 0, 1, \ldots, n \\ j = 0, 1, \ldots, r.
\end{cases}
\]

Bézier curves are known to be a special polynomial type of B–spline curve with the knot vector given by \( n \) knots at \( a \) and \( n \) knots at \( b \). By the dual functional property the control vertices \( Q_i \) for the Bézier representation of the curve are given in terms of blossom by \( Q_i = B_G(a, a, \ldots, a, b, b, \ldots, b) \) where \( a \) appears as an argument \((n − i)\) times and \( b \) appears \( i \) times. For the sake of simplicity, consider the cubic case. We begin with the control vertices from the blossom representations \( B_G(a, a, a, a), B_G(a, a, a, b), B_G(a, b, b), \) and \( B_G(b, b, b) \). From these we wish to raise the degree of the Bézier curve by two. We can represent the new Bézier curve as having a knot vector with \( n + 2 \) knots at \( a \) and another \((n + 2)\) at \( b \) as shown in Fig. 2.

\[
\begin{align*}
R_0 &= B_G(a, a, a, a, a) = B_G(a, a, a) \\
R_1 &= B_G(a, a, a, a, b) = 2/5B_G(a, a, a) + 3/5B_G(a, a, b) \\
R_2 &= B_G(a, a, a, b, b) = 1/5B_G(a, a, a) + 3/5B_G(a, a, b) + 3/5B_G(a, a, b) + 1/5B_G(a, b, b) \\
R_3 &= B_G(a, a, b, b, b) = 3/5B_G(a, a, b) + 3/5B_G(a, b, b) + 1/5B_G(b, b, b) \\
R_4 &= B_G(a, b, b, b, b) = 3/5B_G(a, b, b) + 2/5B_G(b, b, b) \\
R_5 &= B_G(b, b, b, b, b) = B_G(b, b, b)
\end{align*}
\]

**Figure 2.** Degree elevation each Bézier piece (circle).
The degree elevation algorithm in Fig. 2 is expressed in a matrix form as

\[ R = M_e Q \]

where

\[
R = \begin{pmatrix}
B_c(0,0,0,0) & B_c(0,0,0,1) & B_c(0,0,1,1) & B_c(0,1,1,1) & B_c(1,1,1,1) & B_c(1,1,1,3) & B_c(1,1,3,3) & B_c(1,3,3,3) & B_c(3,3,3,3)
\end{pmatrix},
M_e = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2/5 & 3/5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/10 & 3/5 & 3/10 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 3/10 & 3/5 & 1/10 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 3/5 & 2/5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/10 & 3/5 & 3/10 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 3/10 & 3/5 & 2/5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 3/5 & 2/5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]

and \( Q = \begin{pmatrix}
B_c(0,0,0) & B_c(0,0,1) & B_c(0,1,1) & B_c(0,1,3) & B_c(0,3,3) & B_c(1,1,1) & B_c(1,1,3) & B_c(1,3,3) & B_c(3,3,3)
\end{pmatrix} \).

The following algorithm compute degree elevation \( M_e \) matrix.

**Make_ElevateBezierMatrix**\((p, r, s)\)

// Input: Degree \( p \), elevation degree \( r \) and number of distinct interior knots \( s \)
// Output: \((p + r)(s + 1) + 1\) \( \times \) \((ps + p + 1)\) matrix \( M_e \)

\( \text{div} = 1; \)
\[ \text{for } (i = 1 \text{ to } r \text{ by } 1) \text{ div } = \text{div} \times (p + i); \]

\[ M[0] = 1; \]
\[ \text{for } (i = 0 \text{ to } p \text{ by } 1) \]
\[ \quad \text{for } (j = 1 \text{ to } r \text{ by } 1) \ M[j] = M[j] + M[j - 1]; \]
\[ \quad \text{for } (j = 0 \text{ to } r \text{ by } 1) \]
\[ \quad \quad \text{left} = 1; \]
\[ \quad \quad \text{for } (k = 1 \text{ to } j \text{ by } 1) \ \text{left} = \text{left} \times (r - k + 1); \]
\[ \quad \quad \text{right} = 1; \]
\[ \quad \quad \text{for } (k = 1 \text{ to } r - j \text{ by } 1) \ \text{right} = \text{right} \times (p - i + k); \]
\[ \quad \quad T[i + j][i] = \text{left} \times M[j] \times \text{right} / \text{div}; \]
\[ \text{endfor} \]
\[ \text{endfor} \]

\[ M_e[0][0] = T[0][0]; \]
\[ \text{for } (i = 0 \text{ to } s \text{ by } 1) \]
\[ \quad \text{for } (j = 1 \text{ to } p + r \text{ by } 1) \]
\[ \quad \quad \text{for } (k = 0 \text{ to } p \text{ by } 1) \ M_e[j + (p + r)i][(k + p)i] = T[j][k]; \]

5. Curve Composition

Now, we consider the knot removal algorithm. It is the reverse process of inserting a knot. While knot insertion is a precise procedure, i.e. the knot-inserted curve is precisely the same as the original one, knot removal, in general, produces an approximation of the original curve. Clearly, after a knot is inserted, it can be removed precisely.
Consider the curve shown in Fig. 3. The knot \( u = 1 \) is removable two times since the curve is \( C^2 \)-continuous at \( u = 1 \). Curve composition is the inverse of decomposition. In Fig. 3, it can be expressed in a linear equation as

\[
R = M_D \tilde{P}
\]

where

\[
R = \begin{pmatrix}
B_0(0,0,0,0) \\
B_0(0,0,0,1) \\
B_0(0,0,1,1) \\
B_0(0,1,1,1) \\
B_0(1,1,1,3) \\
B_0(1,1,3,3) \\
B_0(1,3,3,3) \\
B_0(3,3,3,3)
\end{pmatrix}, \quad M_D = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 4/9 & 4/9 & 1/9 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2/3 & 1/3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}
\]

and \( \tilde{P} = \begin{pmatrix}
B_0(0,0,0,0) \\
B_0(0,0,0,1) \\
B_0(0,0,1,1) \\
B_0(0,1,1,1) \\
B_0(1,1,1,3) \\
B_0(1,1,3,3) \\
B_0(1,3,3,3) \\
B_0(3,3,3,3)
\end{pmatrix} \). Because \( M_D \) has full column rank and \( M_D^t M_D \) is non-singular, \( (M_D^t M_D)^{-1} M_D^t \) is a left inverse of the matrix \( M_D([18]) \). Therefore the consistent equation \( R = M_D \tilde{P} \) have the unique solution

\[
\tilde{P} = M_c R
\]

where \( M_c = (M_D^t M_D)^{-1} M_D^t \). This shows that

\[
R = M_D (M_D^t M_D)^{-1} M_D^t R.
\]
We write down the algorithm for the degree elevation of a B-spline curve in a shorthand notation.

(1) Make-DecomposeBsplineMatrix(knots, m, n, p) : $M_d$

with knots $\{a, a, \ldots, a, u_1, \ldots, u_i, \ldots, u_j, b, b, \ldots, b\}$

(2) Make-ElevateBezierMatrix(p, r, s) : $M_e$

(3) Make-DecomposeBsplineMatrix(knots, m+(s+2)r, n+(s+2)r, p+r) : $M_D$

with knots $\{a, a, \ldots, a, u_1, \ldots, u_i, \ldots, u_j, b, b, \ldots, b\}$

Then,

$$\hat{P} = MP$$

where $M = M_cM_eM_d$.

References


Department of Applied Mathematics,
Dongseo University,
Pusan 617-716, Korea

Department of Mathematics Education,
Seowon University,
Chongju 361-742, Korea
RELATIONSHIPS BETWEEN AMERICAN PUTS AND CALLS ON FUTURES CONTRACTS

SUK JOON BYUN AND IN JOON KIM

ABSTRACT. This paper presents a formula that relates the optimal exercise boundaries of American call and put options on futures contract. It is shown that the geometric mean of the optimal exercise boundaries for call and put written on the same futures contract with the same exercise price is equal to the exercise price which is time invariant. The paper also investigates the properties of American calls and puts on futures contract.

1. INTRODUCTION

The seminal paper of Black (1976) provided a groundwork for the valuation of options written on futures contracts and showed how to extend the Black-Scholes (1973) option pricing formula to the case of European options on futures contracts. However, all of the publicly traded futures options on the organized exchanges are of the American type; i.e., they allow the holder to exercise them before the expiration date. Assuming that the risk-free rate of interest is positive, American futures options are always subject to early exercise and so American futures options must be worth strictly more than their European counterparts. Therefore Black’s formula does not provide a correct value and there is no closed-form solution for American futures options.

Working from McKean’s (1965) formulation of the free boundary problem, integral formulations of American option values were independently derived by Kim (1990), Jacka (1991), and Carr, Jarrow, and Myneni (1992). Kim (1990) derived the valuation formulas by considering an option exercisable at a finite number of points in time and evaluating its continuous limit as the time intervals shrink to zero. Jacka (1991) obtains the same valuation formulas using the probability theory applied to the optimal stopping problem. Carr, Jarrow, and Myneni (1992) also obtain the formulas by considering the trading strategy which converts an American option into a European one. Kim and Yu (1996) gives a more simpler and intuitive proof of the valuation formulas. Valuation formulas for American futures options can be obtained as a special case of these formulas. (One very comprehensive reference on options is that by Duffie 1992;
it contains a considerable amount of background material on option valuation. For a rigorous survey on the theory of American option pricing, see Myneni 1992).

The integral formulations of American option values yield a pair of nonlinear integral equations, one is for calls and the other is for puts, that must be satisfied by the optimal exercise boundaries. The optimal exercise boundary is defined by the critical futures price at or above (below) which it is optimal to exercise the American calls (puts) on futures contracts. This paper presents a formula that relates the optimal exercise boundaries of American call and put futures options, by exploiting the special structure of these integral equations. It is shown that the geometric mean of the optimal exercise boundaries for American call and put options written on the same futures contract with the same exercise price is equal to the exercise price which is a time invariant constant. This relationship between the optimal exercise boundaries enables us to investigate the relationships between American call and put futures options.

The paper is organized as follows. Section 2 examines European options on futures contracts. Following Black (1976), this section gives an identity between Black’s formulas for European call and put futures options. In Section 3, American options on futures contracts are examined and the relationships between American call and put futures options are presented. Concluding remarks are in Section 4.

2. European Options on Futures Contracts

Consider European call and put options written on a futures contract with exercise price $K$ and expiry date $T$. The underlying futures contract expires at or after time $T$. Throughout the paper, the usual conditions are assumed that the markets are perfect with continuous trading, there are no-arbitrage opportunities, the risk-free rate of interest, $r$, is a positive constant, and the futures price $F_t$ satisfies a stochastic differential equation:

$$
\frac{dF_t}{F_t} = (\mu - r) dt + \sigma dW_t, \quad t \in [0, T]
$$

where $W$ is a standard Brownian motion on a complete probability space $(\Omega, \mathcal{F}, Q)$ and the coefficients $\mu$ and $\sigma$ are positive constants. Equation (2.1) implies that the futures price follows a lognormal diffusion process with volatility parameter $\sigma$.

In this setting, Black (1976) developed a variation of his earlier Black-Scholes (1973) model to value European futures options. Let us denote the value functions of European call and put at time $t$ by $c(F, \tau)$ and $p(F, \tau)$ defined on domain $D \equiv \{(F, \tau); 0 < F < \infty, 0 < \tau \leq T\}$, where $F$ is the futures price and $\tau = T - t$ is time to expiration. The formulas for the prices may be expressed as:

$$
c(F, \tau) = e^{-\tau r} \left[ F \Phi(d_1(F, \tau; K)) - K \Phi(d_2(F, \tau; K)) \right]
$$

$$
p(F, \tau) = e^{-\tau r} \left[ K \Phi(-d_2(F, \tau; K)) - F \Phi(-d_1(F, \tau; K)) \right]
$$
where \( \Phi(\cdot) \) is the cumulative standard normal distribution function and

\[
d_1(F, \tau; K) = \frac{\ln(F/K) + \frac{1}{2} \sigma^2 \tau}{\sigma \sqrt{\tau}} \\
d_2(F, \tau; K) = \frac{\ln(F/K) - \frac{1}{2} \sigma^2 \tau}{\sigma \sqrt{\tau}}
\]

We will first present a useful identity between \( d_1(\cdot, \cdot; \cdot) \) and \( d_2(\cdot, \cdot; \cdot) \).

**Lemma 2.1** For all \( x > 0, y > 0, \) and \( \tau > s \geq 0 \),

\[
d_1(x, \tau - s; y) = -d_2\left( \frac{K^2}{x}, \frac{K^2}{y}; \frac{K^2}{x} \right) \tag{2.6}
\]

\[
d_2(x, \tau - s; y) = -d_1\left( \frac{K^2}{x}, \frac{K^2}{y}; \frac{K^2}{x} \right) \tag{2.7}
\]

**Proof of (2.6).** Using equation (2.5), we see that

\[
d_2\left( \frac{K^2}{x}, \frac{K^2}{y}; \frac{K^2}{x} \right) = \frac{\ln\left( \frac{K^2}{x} / \frac{K^2}{y} \right) - \frac{1}{2} \sigma^2 (\tau - s)}{\sigma \sqrt{\tau - s}}
\]

\[
= \frac{\ln(y/x) - \frac{1}{2} \sigma^2 (\tau - s)}{\sigma \sqrt{\tau - s}}
\]

\[
= -d_1(x, \tau - s; y)
\]

where the last equality follows from equation (2.4).

**Proof of (2.7).** The result is immediate from equation (2.6). \( \square \)

We then give an identity between put and call value functions for European options on futures contracts.

**Proposition 2.1** For all \( x > 0 \) and \( \tau > 0 \),

\[
c(x, \tau) = \frac{x}{K} p\left( \frac{K^2}{x}, \tau \right) \tag{2.8}
\]

where \( K \) is the exercise price.

**Proof.** Black’s European put pricing formula (2.3) gives

\[
p\left( \frac{K^2}{x}, \tau \right) = e^{-r \tau} \left[ K \Phi(-d_2(K^2/x, \tau; K)) - \frac{K^2}{F} \Phi(-d_1(K^2/x, \tau; K)) \right] \tag{2.9}
\]

\[
= e^{-r \tau} \left[ K \Phi(d_1(x, \tau; K)) - \frac{K^2}{F} \Phi(d_2(x, \tau; K)) \right] \tag{2.10}
\]
where we have used Lemma 2.1 with \( y = K \). Multiplying both sides of equation (2.10) by \( x/K \), we have

\[
\frac{x}{K} P \left( \frac{K^2}{x}, \tau \right) = e^{-r \tau} [x \Phi(d_1(F, \tau; K)) - K \Phi(d_2(x, \tau; K))]
\]

\[
= c(x, \tau)
\]

where the last equality follows from Black’s European call pricing formula (2.2).

3. American Options on Futures Contracts

Continuing with the setup and notation of the previous section, now consider the corresponding American call and put options written on the same futures contract. With American futures options, there is always the possibility of early exercise whether the underlying asset on which the futures contract is written pays dividends or not as long as the risk-free rate of interest is positive. This implies that, for each time to maturity \( \tau \in (0, T] \), there exists a critical futures price \( G(\tau)(B(\tau)) \) at or above (below) which the American call (put) should be exercised immediately. The optimal exercise boundary is defined as the time path of critical futures prices.

Let us denote the value functions of American call and put at time \( t \) by \( C(F, \tau) \) and \( P(F, \tau) \) defined on the domain \( D \). If the futures price is above (below) the optimal exercise boundary, the American call (put) is dead and its value is defined to be \( C(F, \tau) = F - K \) (\( P(F, \tau) = K - F \)). Although no one yet has found a closed-form solution for the values of live American futures options, the integral formulations express the value of a live American option as the sum of the corresponding European option value and the early-exercise premium. In summary, the American call and put futures option values are given by:

\[
C(F, \tau) = \begin{cases} 
F - K & \text{if } F \geq G(\tau) \\
c(F, \tau) + \int_0^\tau \phi(F, \tau - s; G(s)) ds & \text{if } F < G(\tau)
\end{cases}
\]

\[
P(F, \tau) = \begin{cases} 
K - F & \text{if } F \leq B(\tau) \\
p(F, \tau) + \int_0^\tau \psi(F, \tau - s; B(s)) ds & \text{if } F > B(\tau)
\end{cases}
\]

where \( c(F, \tau) \) and \( p(F, \tau) \) denote Black’s formulas for European call and put futures options, respectively, and

\[
\phi(F, \tau - s; G(s)) = re^{-r(\tau - s)} [F \Phi(d_1(F, \tau - s; G(s))) - K \Phi(d_2(F, \tau - s; G(s)))]
\]

\[
\psi(F, \tau - s; B(s)) = re^{-r(\tau - s)} [K \Phi(d_2(F, \tau - s; B(s))) - F \Phi(-d_1(F, \tau - s; B(s)))]
\]

We will give an identity between \( \phi(\cdot, \cdot; \cdot) \) and \( \psi(\cdot, \cdot; \cdot) \).

\footnote{This fact has been well established in Rama\textsc{swamy} and Sundaresan (1985), Brenner, Courtadon, and Subrah\textsc{manymam} (1985), and Ball and Torous (1986).}
Lemma 3.1 For all $x > 0, y > 0,$ and $\tau > s \geq 0$,

$$\phi(x, \tau - s; y) = \frac{x}{K^2} \psi\left(\frac{K^2}{x}, \tau - s; \frac{K^2}{y}\right)$$

\textit{Proof.} From equation (3.4), we have

$$\psi\left(\frac{K^2}{x}, \tau - s; \frac{K^2}{y}\right)$$

$$= re^{-r(\tau-s)} \left[ KN\left(-d_2\left(\frac{K^2}{x}, \tau - s; \frac{K^2}{y}\right)\right) - \frac{K^2}{x} e^{-\tau} \left(-d_1\left(\frac{K^2}{x}, \tau - s; \frac{K^2}{y}\right)\right) \right]$$

$$= re^{-r(\tau-s)} \left[ K\left(d_1(x, \tau - s; y)\right) - \frac{K^2}{x} \left(d_2(x, \tau - s; y)\right) \right]$$

$$= \frac{K}{x} \phi(x, \tau - s; y)$$

where the second equality follows from Lemma 2.1 and the last equality follows from equation (3.3).

By imposing an optimality condition on valuation formulas (3.1) and (3.2), the following nonlinear integral equations are obtained that implicitly define the optimal exercise boundaries.\footnote{There are many other integral equations that define the unique optimal exercise boundary. See, for example, McKean (1965), Kim (1990), and Carr, Jarrow, and Myneni (1992)}

$$G(\tau) - K = e(G(\tau), \tau) + \int_0^\tau \phi(G(\tau), \tau - s; G(s))ds \quad (3.5)$$

$$K - B(\tau) = p(B(\tau), \tau) + \int_0^\tau \psi(B(\tau), \tau - s; B(s))ds \quad (3.6)$$

Jacka (1991) and van Moerbeke (1976) address the questions of uniqueness and regularity of the solution to integral equations (3.5) and (3.6). Although there is no explicit solution available for the optimal exercise boundary, we explicitly know the limiting behaviour of the boundary. It can be easily checked that $G(0)B(0) = G(\infty)B(\infty) = K^2$, where $G(0)$ and $B(0)$ represent the limits of $G(\tau)$ and $B(\tau)$ as $\tau$ tends to zero, and similarly $G(\infty)$ and $B(\infty)$ represent the limits of $G(\tau)$ and $B(\tau)$ as $\tau$ tends to infinity.\footnote{See Appendix for a proof of $G(\infty)B(\infty) = K^2$.}

Note that $G(\infty)$ and $B(\infty)$ also stand for the critical futures prices for perpetual American call and put with otherwise similar terms. The next proposition uses integral equations (3.5) and (3.6) to generate an explicit expression for $B(\cdot)$ in terms of $G(\cdot)$.

\textbf{PROPOSITION 3.1} Assume that the integral equation (3.5) has a unique continuous solution $G(\tau)$ for $\tau \in [0, T]$. Then the integral equation (3.6) possesses a unique continuous solution and the solution is simply given by

$$B(\tau) = \frac{K^2}{G(\tau)} \quad \text{for} \quad \tau \in [0, T] \quad (3.7)$$
where $K$ is the exercise price.

Proof. Let $G(\tau)$ be the unique continuous solution of integral equation (3.5). Consider then $\hat{B}(\tau)$ defined by

$$\hat{B}(\tau) = \frac{K^2}{G(\tau)} \quad \tau \in (0,T]$$

(3.8)

We will show that this $\hat{B}(\tau)$ satisfies integral equation (3.6). Now substitute $\hat{B}(\tau)$ into equation (3.6). Then

$$p(\hat{B}(\tau), \tau) + \int_0^\tau \psi(\hat{B}(\tau), \tau - s; \hat{B}(s)) ds$$

$= p(K^2/G(\tau), \tau) + \int_0^\tau \psi(K^2/G(\tau), \tau - s; K^2/G(s)) ds$

$= \frac{K}{G(\tau)} c(G(\tau), \tau) + \int_0^\tau \frac{K}{G(\tau)} \phi(G(\tau), \tau - s; G(s)) ds$

$= \frac{K}{G(\tau)} (G(\tau) - K)$

$= K - \hat{B}(\tau)$

where the second equality follows from Proposition 2.1 and Lemma 3.1 with $x = G(\tau)$ and $y = G(s)$, the third equality follows from equation (3.5), and the last equality follows from equation (3.8). This proves that $\hat{B}(\tau)$ defined by (3.8) satisfies equation (3.6).

To show that $\hat{B}(\tau)$ is the only continuous solution, suppose there exists another continuous solution $\hat{\hat{B}}(\tau)$ of equation (3.6). Then equation (3.5) would have another continuous solution $\hat{\hat{G}}(\tau) \equiv K^2/\hat{\hat{B}}(\tau) \neq G(\tau)$, which contradicts the hypothesis that integral equation (3.5) has a unique continuous solution. Therefore we must have $\hat{B}(\tau) = \hat{\hat{B}}(\tau), \tau \in (0,T]$, and there is only one continuous solution.

Proposition 3.1 says that the geometric mean of the optimal exercise boundaries $G(\tau)$ and $B(\tau)$ is equal to the exercise price; i.e., $\sqrt{G(\tau)B(\tau)} = K$, for $\tau \in [0,T]$. This relation gives a shortcut for computing the optimal exercise boundary of an American call/put given the optimal exercise boundary of its counterpart. By rearranging and taking logarithms on the relation (3.7) in Proposition 3.1, we obtain:

$$\ln\left(\frac{G(\tau)}{K}\right) + \ln\left(\frac{B(\tau)}{K}\right) = 0$$

(3.9)

This says that the arithmetic mean of the logarithms of the optimal exercise boundaries divided by the exercise price is equal to zero as shown in Figure 1. Note the symmetry with respect to the time to maturity axis ($\tau$-axis).

Following the same lines of arguments, Proposition 3.1 can be generalized for American options on dividend-paying assets as follows: $G(\tau; r, \delta)B(\tau; r, \delta) = K^2$ where $r$ is the riskfree rate and $\delta$ is the dividend yield.
Using Propositions 2.1 and 3.1, along with valuation formulas (3.1) and (3.2), the following identity between put and call value functions for American futures options, analogous to Proposition 2.1, obtains.

**Proposition 3.2**  For all $x > 0$ and $\tau \geq 0$,

$$ \begin{align*}
C(x, \tau) &= \frac{x}{K}P \left( \frac{K^2}{x}, \tau \right) 
\end{align*} 
$$

(3.10)

where $K$ is the exercise price.\(^5\)

*Proof.* American put valuation formula (3.2) gives

$$ P \left( \frac{K^2}{x}, \tau \right) = p \left( \frac{K^2}{x}, \tau \right) + \int_0^{\tau} \psi(K^2/x, \tau - s; B(s)) ds 
$$

$$ = \frac{K}{x}c(x, \tau) + \int_0^{\tau} \psi(K^2/x, \tau - s; K^2/G(s)) ds 
$$

$$ = \frac{K}{x}c(x, \tau) + \int_0^{\tau} \frac{K}{x} \phi(x, \tau - s; G(s)) ds 
$$

$$ = \frac{K}{x}C(x, \tau) 
$$

where the second equality follows from Propositions 2.1 and 3.1, the third equality follows from Lemma 3.1 with $y = G(s)$, and the last equality follows from American call valuation formula (3.1).

Proposition 3.2 implies that if we know the value function of an American call/put futures option we can easily determine the value function of its counterpart.

Option traders are interested not only in price but also in the hedge parameters such as delta, gamma, and theta, which are used to evaluate and manage the risks of options. Differentiating both sides of equation (3.10) with respect to $x$ or with respect to $\tau$, the following identities for the hedge parameters can also be obtained.

**Proposition 3.3**  For all $x > 0$ and $\tau \geq 0$,

$$ \begin{align*}
C_F(x, \tau) &= \frac{1}{K}P \left( \frac{K^2}{x}, \tau \right) - \frac{K}{x}P_F \left( \frac{K^2}{x}, \tau \right) \\
C_{FF}(x, \tau) &= \left( \frac{K}{x} \right)^3 P_{FF} \left( \frac{K^2}{x}, \tau \right) \\
C_\tau(x, \tau) &= \frac{x}{K}P_\tau \left( \frac{K^2}{x}, \tau \right) 
\end{align*} 
$$

(3.11) (3.12) (3.13)

---

where $K$ is the exercise price.

It should be noted that Propositions 3.2 and 3.3 are valid for not only live but also dead options.

Proposition 3.2 shows a relationship between values of American call and put futures options at different futures levels. Now consider a relationship between values at the same futures level. Using equations (3.1), (3.2), Proposition 3.1 and put-call parity for European futures options, we obtain the following put-call parity for American futures options in terms of the optimal exercise boundary.

\[
C + Ke^{-r\tau} = P + Fe^{-r\tau} + \int_{0}^{\tau} re^{-r(\tau-s)} \left[ F \left( \frac{\mathbb{E}(d_1)}{\mathbb{E}(d_2)} \right) \right] ds
\]

where $C$ and $P$ denote values for American call and put futures options at the same futures level $F$, respectively, and

\[
d_1 = d_1(F,\tau-s;G(s))
\]
\[
d_2 = d_2(F,\tau-s;G(s))
\]
\[
e = \ln \left( \frac{G(s)}{K} \right) \geq 0
\]

4. Conclusion

This paper presents a formula that relates the optimal exercise boundaries of American call and put options written on the same futures contract with the same exercise price. It is shown that the geometric mean of the put and call optimal exercise boundaries is equal to the exercise price which is a time invariant constant. In addition, this paper provides certain kinds of symmetrical relationships between put and call value functions available for both European and American futures options. The results are important not only for gaining insight into the qualitative behavior of the value functions and the optimal exercise boundaries, but are also useful in the design of effective numerical methods.

5. Appendix

Proof of $G(\infty)B(\infty) = K^2$. $G(\infty)$ and $B(\infty)$ are defined by

\[
G(\infty) = \frac{\beta K}{\beta - 1}, \quad \beta = \frac{1}{\sigma^2} + \sqrt{\frac{1}{\sigma^2} + 2\sigma^2r}
\]
\[
B(\infty) = \frac{\theta K}{\theta - 1}, \quad \theta = \frac{1}{\sigma^2} - \sqrt{\frac{1}{\sigma^2} + 2\sigma^2r}
\]
Straightforward multiplication yields

\[ B(\infty)G(\infty) = \frac{\beta \theta K^2}{(\beta - 1)(\theta - 1)} \]
\[ = \frac{\beta \theta K^2}{\beta \theta - (\beta + \theta) + 1} \]
\[ = K^2 \]

where we have used the fact that \( \beta + \theta = 1 \).

\[ \square \]

REFERENCES


Graduate School of Management
Korea Advanced Institute of Science and Technology
Chungryangri-Dong, Tongdaemun-Gu, Seoul 270-43, Korea
Analysis of One-dimensional cellular automata over $GF(q)$

Sung-Jin Cho, Han-Doo Kim and Un-Sook Choi

Abstract
We study theoretical aspects of one-dimensional cellular automata over $GF(q)$, where $q$ is a power of a prime. Some results about the characteristic polynomials of such cellular automata are given. Intermediate boundary cellular automata are defined and related to the more common null boundary cellular automata.

1 Introduction

One-dimensional linear hybrid cellular automata (CA) have been proposed as an alternative to linear feedback shift registers (LFSRs), in applications such as test pattern generation, pseudorandom number generation, cryptography and signature analysis. Uniform CA have been studied extensively, both over $GF(2)$ and the more general setting of $GF(q)$ ([4], [6], [10], [11]). An LFSM $M$ is an implementation of a linear operator $L$, and many properties of $M$ can be expressed as properties of $L$. Some properties of linear operators, and hence of LFSMs, are easier to identify by the study of minimal polynomials than by direct examination of linear operators. Determining the minimal polynomial of a linear operator is sometimes complicated, whereas deriving the characteristic polynomial of a linear operator is straightforward. Fortunately, the classes of linear operators that we study often have the property that their minimal polynomials equal their characteristic polynomials. Cattel and Muzio ([2]) studied theoretical aspects of one-dimensional linear hybrid cellular automata over a finite field, and they presented general results concerning the characteristic polynomials of such CA. Also they defined cyclic boundary CA and gave relations to the more common null boundary CA. In this paper we study theoretical foundation of one-dimensional cellular automata over a finite field. Some results about the characteristic polynomials of such cellular automata are given. Intermediate boundary cellular automata are defined and related to the more common null boundary cellular automata.

1991 Mathematics Subject Classification: 94
Key Words and Phrases: Cellular automata, null boundary cellular automata, cyclic cellular automata, intermediate boundary cellular automata, characteristic polynomials, minimal polynomials.
This work was supported by the 2000 Inje University Research Grant
2 Preliminaries

The structures under consideration are a particular type of linear machine defined over the finite field $GF(q)$, where $q$ is a power of a prime. For background on CA, see [1], [2], [3], [4], [5], [6], [7], [8], [9]. The transition matrix of a null boundary CA has the form

$$T_{NBCA} = \begin{pmatrix}
    d_1 & b_1 & 0 & \cdots & 0 & 0 & 0 \\
    c_2 & d_2 & b_2 & \cdots & 0 & 0 & 0 \\
    0 & c_3 & d_3 & \cdots & 0 & 0 & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
    0 & 0 & 0 & \cdots & c_{n-1} & d_{n-1} & b_{n-1} \\
    0 & 0 & 0 & \cdots & 0 & c_n & d_n
\end{pmatrix}.$$

A CA has a maximal length cycle if the sequence of states includes all $q^n - 1$ nonzero states for any nonzero starting state.

Example 2.1. Consider the three cell CA over $GF(2^2)$ with transition matrix

$$T = \begin{pmatrix}
    1 & 1 & 0 \\
    \alpha^2 & 1 & \alpha \\
    0 & \alpha^2 & \alpha
\end{pmatrix}$$

where $\alpha^2 + \alpha + 1 = 0$. This machine has a maximal length cycle of 63 distinct nonzero states.

An irreducible polynomial $p(x)$ of degree $n$ is primitive if it has a root $\alpha$ such that the set $\{\alpha^i : i = 0, 1, \cdots, q^n - 2\}$ equals the set of nonzero elements of $GF(q^n)$. Let the transition matrix of an LFSM be denoted $T_{LFSM}$. The characteristic polynomial of the LFSM is defined to

$$|xI - T_{LFSM}|$$

where $x$ is an indeterminate and $I$ is the identity matrix with the same dimension as $T_{LFSM}$. The characteristic polynomial is primitive if and only if the LFSM has a maximal length cycle([8]). We use $\Delta_{i,j}$ to denote the characteristic polynomial of $M_{i,j}$ and abbreviate $\Delta_{1,j}$ as $\Delta_j$. $\Delta_{i,j}$ is called a subpolynomial. The machine $M_{1,n}$ is written as $M$ with corresponding characteristic polynomial $\Delta$.

3 Null Boundary CA(NBCA) characteristic polynomials

Cattel and Muzio([2]) presented several theorems that show interrelationships among the subpolynomials of a NBCA. The following theorems and corollary are in [2].
Theorem 3.1. $\Delta_k$ satisfies the following recurrence:

\[
\begin{align*}
\Delta_{-1} &= 0, \\
\Delta_0 &= 1, \\
\Delta_k &= (x - d_k)\Delta_{k-1} - b_{k-1}c_k\Delta_{k-2}, \quad k \geq 1.
\end{align*}
\]

Corollary 3.2. The characteristic polynomial of a CA can be computed with $2n$ polynomial additions and $2n$ scalar polynomial multiplications.

Example 3.3. Consider the three cell CA over $GF(2^2)$ from Example 2.1, with transition matrix

\[
T = \begin{pmatrix} 1 & 1 & 0 \\ \alpha^2 & 1 & \alpha \\ 0 & \alpha^2 & \alpha \end{pmatrix}.
\]

We have

\[
\begin{align*}
\Delta_{-1} &= 0, \\
\Delta_0 &= 1, \\
\Delta_1 &= (x - 1) \cdot 1 + 0 = x + 1, \\
\Delta_2 &= (x - 1)(x + 1) - 1 \cdot \alpha^2 \cdot 1 = x^2 + \alpha^2 + 1, \\
\Delta_3 &= x^3 + \alpha x^2 + \alpha^2 x + \alpha
\end{align*}
\]

Theorem 3.4. For any $k$ with $0 \leq k \leq n$,

\[
\Delta_{1,n} = \Delta_{1,k} \Delta_{k+1,n} - b_k c_{k+1} \Delta_{1,k-1} \Delta_{k+2,n}.
\]

Theorem 3.5.

\[
\Delta_{1,n-1} \Delta_{2n} - \Delta_{1,n} \Delta_{2,n-1} = \prod_{i=1}^{n-1} b_i \prod_{i=2}^{n} c_i.
\]

4 Cyclic CA

Cattel and Muzio ([2]) considered cyclic CA which is a generalization of NBCA. The transition matrix of a cyclic CA has the form

\[
T = \begin{pmatrix} d_1 & b_1 & 0 & \cdots & 0 & 0 & c_1 \\ c_2 & d_2 & b_2 & \cdots & 0 & 0 & 0 \\ 0 & c_3 & d_3 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & c_{n-1} & d_{n-1} & b_{n-1} \\ b_n & 0 & 0 & \cdots & 0 & c_n & d_n \end{pmatrix}
\]

They defined the notion of a submachine $C_{i,j}$. $\Phi_{i,j}$ denotes the characteristic polynomial of the submachine $C_{i,j}$ with $\Phi = \Phi_{1,n}$. The following theorem is in [2].
\textbf{Theorem 4.1.} Let $T$ be a cyclic CA, with $\Phi$ and $\Delta_{i,j}$ as defined above. Then

$$\Phi_{1,n} = \Delta_{1,n} - c_1 b_n \Delta_{2,n-1} + (-1)^{n+1} \left( \prod_{k=1}^{n} b_k + \prod_{k=1}^{n} c_k \right).$$

We obtain the following theorem from Theorem 4.1.

\textbf{Theorem 4.2.} Let $T$ be a cyclic CA, with $\Phi$ and $\Delta_{i,j}$ as defined above. Then

$$\Phi_{1,n} = a_n \Phi_{2,n} - c_n b_n \Phi_{2,n-1} + (-1)^{n+1} \left( \prod_{k=1}^{n-1} b_k + \prod_{k=1}^{n-1} c_k \right) (a_n + b_n + c_n),$$

where $a_k = x - d_k$, $k = 1, \ldots, n$, $a_1 = a_n$, $b_1 = b_n$, and $c_1 = c_n$.

\textit{Proof.} From Theorem 4.1 we know

$$\Phi_{2,n} = \Delta_{2,n} - c_2 b_n \Delta_{3,n-1} + (-1)^n \left( \prod_{k=2}^{n} b_k + \prod_{k=2}^{n} c_k \right).$$

$$\Phi_{2,n-1} = \Delta_{2,n-1} - c_2 b_{n-1} \Delta_{3,n-2} + (-1)^{n-1} \left( \prod_{k=2}^{n-1} b_k + \prod_{k=2}^{n-1} c_k \right).$$

Using the above two results

$$\Phi_{1,n} = a_1 \Delta_{2,n} - b_1 c_2 \Delta_{3,n} - c_1 b_n \Delta_{2,n-1} + (-1)^{n+1} \left( \prod_{k=1}^{n} b_k + \prod_{k=1}^{n} c_k \right)$$

$$= a_1 (\Phi_{2,n} + c_2 b_n \Delta_{3,n-1} + (-1)^{n+1} \left( \prod_{k=2}^{n} b_k + \prod_{k=2}^{n} c_k \right) - b_1 c_2 \Delta_{3,n})$$

$$- c_1 b_n (\Phi_{2,n-1} + c_2 b_{n-1} \Delta_{3,n-2} + (-1)^{n} \left( \prod_{k=2}^{n-1} b_k + \prod_{k=2}^{n-1} c_k \right) + (-1)^{n+1} \left( \prod_{k=2}^{n-1} b_k + \prod_{k=2}^{n-1} c_k \right))$$

$$= a_n \Phi_{2,n} - c_n b_n \Phi_{2,n-1} + c_2 b_n (a_n \Delta_{3,n-1} - b_{n-1} c_n \Delta_{3,n-2} - \Delta_{3,n})$$

$$+ (-1)^{n+1} a_n (\prod_{k=2}^{n} b_k + \prod_{k=2}^{n} c_k) + (-1)^{n+1} b_n c_n (\prod_{k=2}^{n-1} b_k + \prod_{k=2}^{n-1} c_k)$$

$$+ (-1)^{n+1} \left( \prod_{k=1}^{n} b_k + \prod_{k=1}^{n} c_k \right) \text{ because } a_1 = a_n, c_1 = c_n$$

$$= a_n \Phi_{2,n} - c_n b_n \Phi_{2,n-1} + (-1)^{n+1} (a_n + b_n + c_n) (\prod_{k=1}^{n} b_k + \prod_{k=1}^{n} c_k)$$

$\blacksquare$
As a corollary we obtain the following result which is in [3].

**Corollary 4.3.** Let \( b_k = c_k = 1, d_k = 0 \) or 1 and \( d_1 = d_n \) for \( k = 1, \ldots, n \) in Theorem 4.2. Then

\[
\Phi_{1,n} = (x + d_1) \Phi_{2,n} + \Phi_{2,n-1}.
\]

If \( q = 2^n \), then we obtain the following result.

**Theorem 4.4.**

\[
\Phi_{1,n} = a_1 \Delta_2,n - b_1 c_2 \Delta_3,n - b_n c_1 \Delta_{2,n-1} + (-1)^{n+1} \left( \sum_{k=1}^{n} b_k + \prod_{k=1}^{n} c_k \right)
\]

where \( n \) is even and \( b_k = b, c_k = c, d_1 = d_3 = \cdots = d_{n-1} \) and \( d_2 = d_4 = \cdots = d_n \).

**Proof.**

\[
\Phi_{1,n} = a_1 \Delta_2,n - b_1 c_2 \Delta_3,n - b_n c_1 \Delta_{2,n-1} + (-1)^{n+1} \left( \sum_{k=1}^{n} b_k + \prod_{k=1}^{n} c_k \right)
\]

\[
= a_1 \Delta_2,n - b c \Delta_3,n - b c \Delta_{2,n-1} + b^n + c^n
\]

\[
= a_1 \Delta_2,n + b^n + c^n \quad \text{because} \quad \Delta_3,n = \Delta_{2,n-1}
\]

\[
= a_1 (a_2 \Delta_3,n + b c \Delta_{4,n}) + b^n + c^n
\]

\[
= a_1 a_2 \Delta_3,n + a_1 b c \Delta_{4,n} + b^n + c^n
\]

because \( a_2 = a_1 \) and \( a_1 \Delta_{4,n-1} + \Delta_{4,n} = \Delta_{4,n-2} \)

\[
= a_1 a_2 \Phi_{3,n} + (bc) \{ \Phi_{5,n} + b^{n-4} + c^{n-4} \} + a_1 b c (b^{n-2} + a_1 a_2 c^{n-2} + b^n + c^n)
\]

\[
= a_1 a_2 \Phi_{3,n} + (bc) \{ \Phi_{5,n} + b^{n-4} + c^{n-4} \} + a_1 b c (b^{n-2} + a_1 a_2 c^{n-2} + b^n + c^n)
\]

\[
= a_1 a_2 \Phi_{3,n} + (bc) \{ \Phi_{5,n} + b^{n-4} + c^{n-4} \} + a_1 b c (b^{n-2} + a_1 a_2 c^{n-2} + b^n + c^n)
\]

\[
= a_1 a_2 \Phi_{3,n} + (bc) \{ \Phi_{5,n} + b^{n-4} + c^{n-4} \} + a_1 b c (b^{n-2} + a_1 a_2 c^{n-2} + b^n + c^n)
\]

\[
= a_1 a_2 \Phi_{3,n} + (bc) \{ \Phi_{5,n} + b^{n-4} + c^{n-4} \} + a_1 b c (b^{n-2} + a_1 a_2 c^{n-2} + b^n + c^n)
\]

\[\square\]

## 5 Intermediate Boundary CA

In this section we consider Intermediate Boundary CA (IBCA) and the relations with NBCA and study the related properties. The transition matrix of an IBCA has the form

\[
T_{IBCA} = \begin{pmatrix}
  d_1 & b_1 & c_1 & \cdots & 0 & 0 & 0 \\
  c_2 & d_2 & b_2 & \cdots & 0 & 0 & 0 \\
  0 & c_3 & d_3 & \cdots & 0 & 0 & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
  0 & 0 & 0 & \cdots & c_{n-1} & d_{n-1} & b_{n-1} \\
  0 & 0 & 0 & \cdots & b_n & c_n & d_n
\end{pmatrix}
\]
Theorem 5.1. For every NBCA such that $b_2 \neq 0$ and $c_{n-1} \neq 0$ there exist at least one IBCA having the same characteristic polynomial.

Proof. The characteristic matrix of a NBCA can be represented as

$$T_{NBCA} = \begin{pmatrix} d_1 & b_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ c_2 & d_2 & b_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & c_3 & d_3 & 0 & \cdots & 0 & 0 & 0 \\ & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & d_{n-2} & b_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & c_{n-1} & d_{n-1} & b_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & c_n & d_n \end{pmatrix}$$

The characteristic polynomial of this is $|xI - T_{NBCA}|$.

$$xI - T_{NBCA} = \begin{pmatrix} x - d_1 & b_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ c_2 & x - d_2 & b_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & c_3 & x - d_3 & 0 & \cdots & 0 & 0 & 0 \\ & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & x - d_{n-2} & b_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & c_{n-1} & x - d_{n-1} & b_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & c_n & x - d_n \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} x - d_1 + c_2 & x - d_2 + b_1 & b_2 & 0 & \cdots & 0 & 0 & 0 \\ c_2 & x - d_2 & b_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & c_3 & x - d_3 & 0 & \cdots & 0 & 0 & 0 \\ & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & x - d_{n-2} & b_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & c_{n-1} & x - d_{n-1} & b_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & c_{n-1} & x - d_{n-1} + c_n & x - d_n + b_{n-1} \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} x - u & u - d_2 + b_1 & b_2 & 0 & \cdots & 0 & 0 & 0 \\ c_2 & x - d_2 - c_2 & b_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & c_3 & x - d_3 & 0 & \cdots & 0 & 0 & 0 \\ & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & x - d_{n-2} & b_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & c_{n-1} & x - d_{n-1} - b_{n-1} & b_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & c_{n-1} & v - d_{n-1} + c_n & x - v \end{pmatrix}$$
where \( u = d_1 - c_2 \) and \( v = d_n - b_{n-1} \)

\[
\begin{pmatrix}
d_1 - c_2 & d_2 - d_1 + c_2 - b_1 & -b_2 & 0 & \cdots & 0 & 0 & 0 \\
-c_2 & d_2 + c_2 & -b_2 & 0 & \cdots & 0 & 0 & 0 \\
0 & -c_3 & d_3 & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & d_{n-2} & -b_{n-2} & 0 \\
0 & 0 & 0 & 0 & \cdots & -c_{n-1} & d_{n-1} + b_{n-1} & -b_{n-1} \\
0 & 0 & 0 & 0 & \cdots & -c_{n-1} & d_{n-1} - d_n + b_{n-1} - c_n & d_n - b_{n-1}
\end{pmatrix}
\]

\( \Rightarrow xI = \begin{pmatrix} \Psi_1 \end{pmatrix} \)

\( \Rightarrow xI in T_{IBCA} \) \( T_{IBCA} \) is the matrix representation of IBCA.

Let \( b_k = c_k = 1 \) and \( d_k = 0 \) or \( 1 \) in Theorem 5.1. Then we obtain the following result which is in [4].

**Corollary 5.2.** For every 90/150 NBCA there exists at least one IBCA having the same characteristic polynomial.

**Theorem 5.3.** Let \( n \geq 7 \). Then

\[
\Psi_{1,n} = \Delta_{1,n} + \prod_{k=1}^{3} c_k \Delta_{4,n} + \prod_{k=n-2}^{n} b_k \Delta_{1,n-2} + \prod_{k=1}^{3} c_k \prod_{k=n-2}^{n} b_k \Delta_{4,n-3}.
\]

**Proof.** Let \( a_i = x - d_i \). Then

\[
\Psi = \begin{pmatrix}
a_1 & b_1 & c_1 & \cdots & 0 & 0 & 0 \\
c_2 & a_2 & b_2 & \cdots & 0 & 0 & 0 \\
0 & c_3 & a_3 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & c_{n-1} & a_{n-1} & b_{n-1} \\
0 & 0 & 0 & \cdots & b_n & c_n & a_n
\end{pmatrix}
\]

By expanding the determinant along the first row, we have

\[
\Psi_{1,n} = a_1 |A| - b_1 |B| + c_1 |C| \\
= a_1(a_n \Delta_{2,n-1} - c_n |D| + b_n |E|) - b_1 c_2 |F| + c_1 c_2 c_3 |G|.
\]
where

\[
A = \begin{pmatrix}
    a_2 & b_2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    a_3 & b_3 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    0 & a_4 & b_4 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    0 & 0 & 0 & 0 & \cdots & c_{n-3} & a_{n-3} & b_{n-3} & 0 & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & c_{n-2} & a_{n-2} & b_{n-2} & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\
    0 & 0 & 0 & 0 & \cdots & 0 & 0 & b_n & c_n & a_n \\
\end{pmatrix}
\]

\[
B = \begin{pmatrix}
    c_2 & b_2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    0 & a_3 & b_3 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    0 & c_4 & a_4 & b_4 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    0 & 0 & 0 & 0 & \cdots & c_{n-3} & a_{n-3} & b_{n-3} & 0 & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & c_{n-2} & a_{n-2} & b_{n-2} & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\
    0 & 0 & 0 & 0 & \cdots & 0 & 0 & b_n & c_n & a_n \\
\end{pmatrix}
\]

\[
C = \begin{pmatrix}
    c_2 & a_2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    0 & c_3 & b_3 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & a_4 & b_4 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    0 & 0 & 0 & 0 & \cdots & c_{n-3} & a_{n-3} & b_{n-3} & 0 & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & c_{n-2} & a_{n-2} & b_{n-2} & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\
    0 & 0 & 0 & 0 & \cdots & 0 & 0 & b_n & c_n & a_n \\
\end{pmatrix}
\]

\[
D = \begin{pmatrix}
    a_2 & b_2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    c_3 & a_3 & b_3 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    0 & c_4 & a_4 & b_4 & \cdots & 0 & 0 & 0 & 0 & 0 \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    0 & 0 & 0 & 0 & \cdots & c_{n-3} & a_{n-3} & b_{n-3} & 0 & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & c_{n-2} & a_{n-2} & b_{n-2} & 0 \\
    0 & 0 & 0 & 0 & \cdots & 0 & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\
\end{pmatrix}
\]
\[
E = \begin{pmatrix}
  a_2 & b_2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
  c_3 & a_3 & b_3 & 0 & \cdots & 0 & 0 & 0 & 0 \\
  0 & c_4 & a_4 & b_4 & \cdots & 0 & 0 & 0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\
  0 & 0 & 0 & 0 & \cdots & c_{n-3} & a_{n-3} & 0 & 0 \\
  0 & 0 & 0 & 0 & \cdots & 0 & c_{n-2} & b_{n-2} & 0 \\
  0 & 0 & 0 & 0 & \cdots & 0 & 0 & a_{n-1} & b_{n-1}
\end{pmatrix}
\]

\[
F = \begin{pmatrix}
  a_3 & b_3 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
  c_4 & a_4 & b_4 & \cdots & 0 & 0 & 0 & 0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\
  0 & 0 & 0 & \cdots & c_{n-3} & a_{n-3} & b_{n-3} & 0 & 0 \\
  0 & 0 & 0 & \cdots & 0 & c_{n-2} & a_{n-2} & b_{n-2} & 0 \\
  0 & 0 & 0 & \cdots & 0 & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\
  0 & 0 & 0 & \cdots & 0 & 0 & b_n & c_n & a_n
\end{pmatrix}
\]

\[
G = \begin{pmatrix}
  a_4 & b_4 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\
  0 & 0 & \cdots & c_{n-3} & a_{n-3} & b_{n-3} & 0 & 0 & 0 \\
  0 & 0 & \cdots & 0 & c_{n-2} & a_{n-2} & b_{n-2} & 0 & 0 \\
  0 & 0 & \cdots & 0 & 0 & c_{n-1} & a_{n-1} & b_{n-1} & 0 \\
  0 & 0 & \cdots & 0 & 0 & b_n & c_n & a_n & 0
\end{pmatrix}
\]

Since \(|A| = a_n \Delta_{2,n-1} - c_n |D| + b_n |E|\), \(|D| = b_{n-1} \Delta_{2,n-2}\) and \(|E| = b_{n-1} b_{n-2} \Delta_{2,n-3}\), by Theorem 3.1 we have

\[
\Psi_{1,n} = a_1 \{ a_n \Delta_{2,n-1} - c_n b_{n-1} \Delta_{2,n-2} + b_n b_{n-1} b_{n-2} \Delta_{2,n-3} \} \\
- b_1 b_2 (a_n \Delta_{3,n-1} - c_n b_{n-1} \Delta_{3,n-2} + b_n b_{n-1} b_{n-2} \Delta_{3,n-3}) \\
+ c_1 c_2 c_3 (a_n \Delta_{4,n-1} - c_n b_{n-1} \Delta_{4,n-2} + b_n b_{n-1} b_{n-2} \Delta_{4,n-3}) \\
= (a_1 \Delta_{2,n} - b_1 c_2 \Delta_{3,n} + c_1 c_2 c_3 \Delta_{4,n}) \\
+ (a_1 \Delta_{2,n-3} - b_1 c_2 \Delta_{3,n-3} + c_1 c_2 c_3 \Delta_{4,n-3}) b_n b_{n-1} b_{n-2} \\
= \Delta_{1,n} + c_1 c_2 c_3 \Delta_{4,n} + b_n b_{n-1} b_{n-2} \Delta_{1,n-3} + c_1 c_2 c_3 b_n b_{n-1} b_{n-2} \Delta_{4,n-3} \\
= \Delta_{1,n} + \prod_{k=1}^{3} c_k \Delta_{4,n} + \prod_{k+n-2}^{n} b_k \Delta_{1,n-3} + \prod_{k=1}^{3} c_k \prod_{k+n-2}^{n} b_k \Delta_{4,n-3}
\]
Corollary 5.4. [12] Let \( q = 2, n \geq 7 \) and \( b_k = c_k = 1, d_k = 0 \) or 1, where \( k = 1, \ldots, n \). Then
\[
\Psi_{1,n} = \Delta_{1,n} + \Delta_{2,n} + \Delta_{1,n-3} + \Delta_{4,n-3}.
\]

Theorem 5.5. Let \( a_1 = a_4, b_1 = b_4, c_1 = c_4, c_2 = c_5 \) and \( n \geq 7 \). Then
\[
\Psi_{1,n} = a_1\Psi_{2,n} - b_1c_2\Psi_{3,n}.
\]

Proof.
\[
\Psi_{1,n} = a_1\Delta_{2,n} - b_1c_2\Delta_{3,n} + c_1c_2c_3\Delta_{4,n} + a_1b_n b_{n-1} b_{n-2} \Delta_{2,n-3}
\]
\[
- b_1c_2b_n b_{n-1} b_{n-2} \Delta_{3,n-3} + c_1c_2c_3b_n b_{n-1} b_{n-2} \Delta_{4,n-3}
\]
\[
= a_1(\Delta_{2,n} + \sum_{k=n-2}^{n} b_k \Delta_{2,n-3}) - b_1c_2(\Delta_{3,n} + \sum_{k=n-2}^{n} b_k \Delta_{3,n-3})
\]
\[
+ \sum_{k=1}^{3} c_k (\Delta_{4,n} + \sum_{k=n-2}^{n} b_k \Delta_{4,n-3})
\]
\[
= a_1(\Psi_{2,n} + \sum_{k=1}^{3} c_k \Delta_{5,n} - \sum_{k=n-2}^{n} b_k \sum_{k=1}^{3} c_k \Delta_{5,n-3}) - b_1c_2(\Psi_{3,n} - \sum_{k=1}^{3} c_k \Delta_{6,n})
\]
\[
- \sum_{k=n-2}^{n} b_k \sum_{k=1}^{3} c_k \Delta_{6,n-3} + \sum_{k=1}^{3} c_k (\Delta_{4,n} + \sum_{k=n-2}^{n} b_k \Delta_{4,n-3})
\]
\[
= a_1\Psi_{2,n} - b_1c_2\Psi_{3,n} - \sum_{k=1}^{3} c_k (a_1\Delta_{5,n} - b_1c_2\Delta_{6,n} - \Delta_{4,n})
\]
\[
- \sum_{k=n-2}^{n} b_k \sum_{k=1}^{3} c_k (a_1\Delta_{5,n-3} - b_1c_2\Delta_{6,n-3} - \Delta_{4,n-3}) \text{ by Theorem 3.1}
\]
\[
= a_1\Psi_{2,n} - b_1c_2\Psi_{3,n}
\]

\[\square\]

Corollary 5.6. [12] Let \( q = 2 \) and \( b_k = c_k = 1, d_k = 0 \) or 1, where \( k = 1, \ldots, n \) and \( n \geq 7 \). Then
\[
\Psi_{1,n} = \Psi_{2,n} + \Psi_{3,n}.
\]
References


Department of Applied Mathematics
Pukyong National University
Pusan 608-737
KOREA
e-mail: sjcho@dolphin.pku.ac.kr

Department of Computational Mathematics
Inje University
Kimhae 621-749
KOREA
e-mail: mathkhd@ijnc.inje.ac.kr
A MODIFIED SELF-AVOIDING WALK MODEL
ON THE SQUARE LATTICE
WITH REFLECTING AND ABSORBING BARRIERS

JUNHO SONG

ABSTRACT. Well known is the directed self-avoiding walk model on the square lattice with reflecting and absorbing barriers. We consider two models, namely, a pyramid self-avoiding polygon model and a top and bottom pyramid polygon model, as subcollections of the model. We derive explicit formulas for the number of $2N$-step polygons in these models.

1. INTRODUCTION

An $N$-step self-avoiding walk $w$ on the $n$-dimensional hypercubic lattice $\mathbb{Z}^d$ is a sequence

$$w = (0 = w(0), w(1), \ldots, w(N))$$

with $w(i) \in \mathbb{Z}^d$, $|w(i+1) - w(i)| = 1$ and $w(i) \neq w(j)$ for $i \neq j$. Equal probability is assigned to each $N$-step self-avoiding walk. A self-avoiding polygon is any self-avoiding walk whose final site is a nearest neighbor of the initial site, augmented by the bond joining the final site to the initial site. The self-avoiding walk model was first introduced by chemists as a model of polymer molecules and has been studied by physicists as an interesting model of critical phenomena. Also, it is of interest to probabilists as a natural example of a non Markov process.

The mathematical problems of calculating the exact analytical properties of self-avoiding walks are formidable. In efforts to overcome the difficulties, a number of modified self-avoiding walk models have been actively investigated. It is hoped that the studies on such modified models may shed light on the properties of the full self-avoiding walk model, and provides useful clues on how to approach the extremely difficult self-avoiding walk problem.

Most notable examples are the “spiralling self-avoiding walks” and the “directed self-avoiding walks”. In the spiralling self-avoiding walks, turns to specific directions are prohibited, in addition to the self-avoiding condition. In the directed self-avoiding walks, steps to specific directions are prohibited. There are the most obvious shortcomings of the models. In the spiralling self-avoiding walk model, due to the very nature of the model, the walker cannot double back and the resulting walks are very dense.
Whereas, in the directed self-avoiding walk model, the walker cannot reach certain quarters of the plane: one quarter in “two-choice directed self-avoiding walks” and two quarters in “three-choice directed self-avoiding walks” [6].

Although the self-avoiding polygons are a small collection of the self-avoiding walks, the quantity of the ratio which is obtained by dividing the number of 2N-step self-avoiding polygons by the number of 2N-step self-avoiding walks is of particular interest since it represents the probability of return to the origin after 2N steps. There are three classes of polygon problems whose exact solutions now exist: these are staircase polygons [5], convex polygons [4] and row-convex polygons [2].

Consider self-avoiding walks in which the walker starts out from the origin and travels over more than two quarters of the plane, winding clockwise around the origin. Such a walk can be divided into several walks that traversed exactly two quarters. Thus, each subwalk that traversed exactly two quarters might be regarded as a “base” of the walk. That is the first reason why we would give a walker “two-choice direction” and “reflecting and absorbing barriers” as microscopic and macroscopic constraint in addition to the self-avoiding condition. On the other hand, “two-choice direction” constraint gives us a model in which most problems can be solved easily and exactly. Moreover, it is possible to derive directly the generating function for the walks consisting of several subwalks that traversed exactly two quarters.

In this paper, we consider a model that the walker may travel over some region on the square lattice with reflecting and absorbing barriers, performing two kinds of two-choice directed self-avoiding walks. The walker starts out from the origin and may take $x_+$ or $y_+$ steps until he comes across a reflecting barrier. Thereafter if the walker would arrive at a point of absorbing barrier, he should take $x_-$ or $y_-$ steps from a reflecting barrier to the last point. Suppose that the walker tries to begin again at the last point and goes on taking his $x_-$ or $y_-$ steps to another reflecting barrier and terminates at a horizontal line below the $x$-axis with his $x_-$ or $y_+$ steps. It means that such a walk could be connected with two kinds of two-choice directed self-avoiding walks with reflecting and absorbing barriers. And it is possible for the walker to travel over more than two quarters of the plane. Let the last point be a lattice point that is an intersection of the walks and a horizontal line $y = -1$, in which the walker should pass by the reflecting barrier, and let the endpoint be the origin. Then it becomes a polygon problem. Therefore, we will investigate both a pyramid self-avoiding polygon model and a top and bottom pyramid self-avoiding polygon model.

2. Preliminaries

The generating function $G_1(x_+)$ for one-dimensional self-avoiding walks restricted to the $x_+$ direction is

$$G_1(x_+)=\frac{1}{1-x_+}$$

and the generating function $G_2(x_+,y_+)$ for two-choice directed self-avoiding walks in the square lattice in which the walker is restricted only to take either $x_+$ or $y_+$ direction
is

\[ G_2(x_+, y_+) = \frac{1}{1 - (x_+ + y_+)} \]

It follows from (1) and (2) that

\[ G_2(x_+, y_+) = G_1(x_+) G_1[y_+ G_1(x_+)] \]

The equation (3) implies that two-choice directed self-avoiding walks in the square lattice are nothing but one-dimensional self-avoiding walks in which the walker is restricted only to take upstairs \([y_+ G_1(x_+)](r)\) direction. Let

\[ s_k(x_+, y_+) = G_1(x_+) [y_+ G_1(x_+)]^k \] for \( k = 0, 1, 2, \ldots \),

\[ T_r(x_+, y_+) = \sum_{k=0}^{r} s_k(x_+, y_+) \] for \( r = 0, 1, 2, \ldots \).

Thus \( s_k(x_+, y_+) \) is the generating function for the walks that start from the horizontal line \( y = i \) and reach the horizontal line \( y = i + k \). We will use the equation (5) later.

3. Our Model

In this section, we consider a model for two-choice directed self-avoiding walks on the square lattice with reflecting and absorbing barriers. Let the horizontal line \( y = r \) be a reflecting barrier and let the horizontal line \( y = -m - 1 \) be an absorbing barrier, where \( r \) and \( m \) are nonnegative integers. In this model, the walker who starts out from the origin may take either \( x_+ \) or \( y_+ \) direction. The walker does or does not visit the reflecting barrier \( y = r \). If the walker left the reflecting barrier, he should have taken \( x_+ \) direction in the previous step. Once he leaves the reflecting barrier, he must be at a point in the horizontal line \( y = r - 1 \). Thereafter he may take any of \( x_+ \) and \( y_- \) directions. If the walker arrived at the absorbing barrier, he should have been in the horizontal line \( y = -m \) in the previous step.

The Figure 1 is the diagram of a typical example of a two-choice directed self-avoiding walk on the square lattice with reflecting and absorbing barriers.

![Figure 1. An N-step walk](image)
It follows from the statements above that the generating function for the walks from the origin to the horizontal line \( y = -m \) is given by

\[
G(x_+, y_+, y_- | r, -m) = T_r(x_+, y_+) + \left[ s_{r}(x_+, y_+) \right] x_+ y_- [T_{r+m-1}(x_+, y_-)]
\]

if \( r \neq 0 \), \( m \neq 0 \) and

\[
G(x_+, y_+, y_- | 0, 0) = T_0(x_+, y_+)
\]

and that the generating function for the walks from the origin to the absorbing barrier is given by

\[
f(x_+, y_+, y_- | r, -m - 1) = \left[ s_{r}(x_+, y_+) \right] x_+ y_- \left[ s_{r+m-1}(x_+, y_-) \right] y_-
\]

if \( r \neq 0 \), \( m \neq 0 \) and

\[
f(x_+, y_+, y_- | 0, -1) = s_0(x_+, y_+) x_+ y_+.
\]

Thus the generating function for all walks in the model is

\[
U(x_+, y_+, y_- | r, -m - 1) = G(x_+, y_+, y_- | r, -m) + f(x_+, y_+, y_- | r, -m - 1).
\]

Putting \( z = x_+ = y_+ = y_- \), we obtain that

\[
G(z | r, -m) = \begin{cases} 
\frac{1}{1-z} \left( 1 - \frac{z^{r+1}}{(1-z)^r} - \frac{z^{2r+m+2}}{(1-z)^{2r+m+1}} \right) & \text{if } r \neq 0, m \neq 0 \\
\frac{1}{1-z} & \text{if } r = 0, m = 0,
\end{cases}
\]

and that

\[
f(z | r, -m - 1) = \frac{z^{2r+m+2}}{(1-z)^{2r+m+1}}.
\]

To obtain the number \( c_N(r, -m - 1) \) of \( N \)-step walks, we evaluate the contour integral

\[
c_N(r, -m - 1) = \frac{1}{2\pi i} \oint \frac{dz}{z^{N+1}} U(z | r, -m - 1)
\]

along a small circle around the origin. The following identity is well-known (see, for example, [p. 47, 7]):

\[
\sum_{k=0}^{n} \binom{r+k-1}{k} 2^{n-k} = \sum_{k=0}^{n} \binom{r+n}{r+k}.
\]

Using

\[
\lim_{z \to 0} \frac{1}{n!} \left( \frac{d}{dz} \right)^n [(1-2z)^{-1}(1-z)^{-r}] = \sum_{k=0}^{n} \binom{r+k-1}{k} 2^{n-k} = \sum_{k=0}^{n} \binom{r+n}{r+k}
\]

and

\[
2 \sum_{k=0}^{n} \binom{r+n}{r+k} - \binom{r+n}{r} = \sum_{k=0}^{n} \binom{r+1+n}{r+1+k},
\]
we have a reasonably explicit formula for the number of \( N \)-step walks as follows:

\[
c_N(r, -m - 1) = \binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{r} + \binom{N-1}{r+1} + \binom{N-1}{r+2} + \cdots + \binom{N-1}{2r+m} + \binom{N-2}{2r+m}.
\]

Therefore, if we let \( a_N(r, -m - 1) \) denote the number of \( N \)-step walks that terminate at a lattice point of the absorbing barrier, then we obtain

\[
a_N(r, -m - 1) = \binom{N-2}{2r+m}.
\]

4. **Pyramid self-avoiding polygons**

In the \((2N - 1)\)-step walks, the walker starts out from the origin and must pass by some reflecting barrier, \( y = r \), to visit the point \((N - r - 1, -1)\). Next he should go back to the vertical line \( x = 0 \) (actually, the point \((0, -1)\)) using his remaining \( N - r - 1 \) steps only in the \( x_- \) direction. Finally, taking one more step in the \( y_+ \) direction, the walker makes a \( 2N \)-step polygon. See a typical example of a \( 2N \)-step polygon in Figure 2.

![Pyramid self-avoiding polygon](image)

**Figure 2.** A \((2N - 1)\)-step walk

First, observe that in any of this polygon, the number of \( x_- \)-steps is just as many as the number of \( x_+ \)-steps. Substituting \( x_+x_- \) for \( x_+ \) in \( f(x_+, y_+, y_- | r, -1)y_+ \), summing it up over \( r \) from 0 to \( \infty \), putting \( z = x_+ = x_- = y_+ = y_- \), we obtain the generating function for the polygons as follows:

\[
G^{pap}(z) = \frac{z^4(1-z^2)}{(z^2 + z - 1)(z^2 - z - 1)},
\]

with a critical point at \( z_c = (\sqrt{5} - 1)/2 \). Notice that \( (7) \) coincides accidentally with one of the results found by Lin et al [4] when they studied convex polygons on the square.
lattice. Evaluating the residues yields
\[
a_N^{psap} = \frac{1 - (-1)^N}{2\sqrt{5}} 
\left( \frac{\sqrt{5} + 1}{2} \right)^N + \left( \frac{\sqrt{5} - 1}{2} \right)^N
\]
for \( N \geq 4 \).

The following identity is well-known (see, for example, [p. 10, 3]):
\[
\sum_{r=0}^{N-2} \binom{N+r-2}{2r} = \frac{1}{\sqrt{5}} \left( \frac{\sqrt{5} + 1}{2} \right)^N + \left( \frac{\sqrt{5} - 1}{2} \right)^N,
\]
which is the \( 2(N-2) \)-th Fibonacci number. We have seen that
\[
a_{2N}^{psap} = \sum_{r=0}^{N-2} a_{N+r}(r, -1)
\]
\[
= \sum_{r=0}^{N-2} \binom{N+r-2}{2r}.
\]
Therefore, the identity (8) holds. This is another way to prove the identity (8).

5. Top and bottom pyramid self-avoiding polygons

Let us consider \( 2N \) step walks that emanate from the origin, pass by two reflecting barriers \( y = r_1 \) and \( y = r_2 \), return to the origin. Such polygons always cross the horizontal line \( y = -1 \). We want to subdivide such a polygon into two parts, one from the origin to the last point on the line \( y = -1 \) and another one from the last point on the line to the origin. Suppose that the first part (or top pyramid) has \( N_1 \) steps and thus the second part (or bottom pyramid) has \( 2N - N_1 \) steps. And then the last point is just \((N_1 - 2r_1 - 1, -1)\) or \((2N - N_1 - 2r_2 - 1, -1)\).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig3.png}
\caption{A 2N-step top and bottom polygon}
\end{figure}
Then, using equation (6), we have a reasonably explicit formula for the number of $2N$-step polygons as follows:

$$a_{2N} = \sum_{r_1=0}^{N-2} \sum_{r_2=0}^{N-2-r_1} a_{N_1}(r_1, -1)a_{N-N_1}(r_2, -1)$$

$$= \sum_{r_1=0}^{N-2} \sum_{r_2=0}^{N-2-r_1} \left( \frac{N + r_1 - r_2 - 2}{2r_1} \right) \left( \frac{N + r_2 - r_1 - 2}{2r_2} \right).$$

Here is a table for $a_{2N}$. The entries for $2N \leq 10$ were verified by drawing all of the diagrams for $2N$-step top and bottom pyramid self-avoiding polygons.

<table>
<thead>
<tr>
<th>$2N$</th>
<th>$a_{2N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>12</td>
<td>136</td>
</tr>
<tr>
<td>14</td>
<td>528</td>
</tr>
<tr>
<td>16</td>
<td>2080</td>
</tr>
<tr>
<td>18</td>
<td>8256</td>
</tr>
<tr>
<td>20</td>
<td>32896</td>
</tr>
<tr>
<td>50</td>
<td>35184376283136</td>
</tr>
<tr>
<td>100</td>
<td>39614081257132309534260330496</td>
</tr>
</tbody>
</table>

References


Department of Mathematics
University of Seoul, Seoul 130–743 Korea
email: jsong@uoscc.uos.ac.kr
**Abstract**

In this paper, we consider first order generalized difference scheme for the two-point boundary value problem and one-dimensional second order parabolic type problem. The optimal error estimates in $L_p$ and $W^{1,p}$ (for $2 \leq p \leq \infty$) as well as some superconvergence estimates in $W^{1,p}$ (for $2 \leq p \leq \infty$) are obtained. The main results in this paper perfect the theory of GDM.

1 Introduction

We first consider the two-point boundary problem

\[
\begin{cases}
(a) & Lu = -\frac{d}{dx} \left( p \frac{du}{dx} \right) = f, \quad a < x < b, \\
(b) & u(a) = 0, \quad u(b) = 0,
\end{cases}
\]

(1.1)

where $p = p(x) \geq p_{\text{min}} > 0$, $p \in C^1(I)$, $f \in L^2(I)$, $I = [a, b]$. Secondly, we consider one-dimensional second order parabolic type problem

\[
\begin{cases}
(a) & \frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left( p \frac{\partial u}{\partial x} \right) = f(x, t), \quad (x, t) \in (a, b) \times (0, T), \\
(b) & u(x, 0) = 0, \quad x \in [a, b], \\
(c) & u(a, t) = 0, \quad u(b, t) = 0, \quad t \in [0, T],
\end{cases}
\]

(1.2)

where $p = p(x) \geq p_{\text{min}} > 0$, $p \in C^1(I)$, $f \in L^2(I)$, $I = [a, b]$. In the past several decades, Li and other authors did extensive and deep research on the theory and application of generalized difference methods (GDM for short), including constructing first order or higher order difference schemes on elliptic, parabolic and hyperbolic equations, establishing the optimal Sobolev norm estimates of errors, and applying GDM to underground fluid, electromagnetic field and other fields. Theoretical researches and realistic computations show that GDM not only keep the computational simplicity of

AMS(MOS) Subject Classification: 65L10, 65L12, 65L17, 65M06, 65M15.

Keywords: $L_p$ error estimates, $W^{1,p}$ error estimates, superconvergence, first order generalized difference methods.
difference methods, but also enjoy the accuracy of finite element methods. See [4] and [7] for more details.

Since the time of studying GDM is only several decades and establishing the error estimates is very difficult, the theory of GDM is not perfect. For example, the optimal order \( H^1 \), \( L_2 \) and maximum norm error estimates to problem (1.1) and (1.2) have been obtained (See [4], [5], [6], [7]), but the \( W^{1,p} \) and \( L_p \) \((2 < p < \infty)\) norm error estimates as well as some superconvergence estimates have not been derived before. We shall do this work in this paper. Moreover, by using Green functions, we will reduce the demand of the smoothness of \( u \) in the error estimates of maximum norms, and get a perfect statement combining the case of \( 2 \leq p \leq \infty \).

This paper is organized in the following way. In section 2 we do some preparations, including introducing the partitions of the interval \( I \), the trial and test function spaces, Green functions and some lemmas which are essential in our analysis. We consider problem (1.1) and establish optimal error estimates of \( u - u_h \) in \( W^{1,p}(I) \) and \( L_p(I) \) \((2 \leq p \leq \infty)\) as well as the superconvergence estimate of \( \bar{u} - u_h \) in \( W^{1,p}(I) \) \((2 \leq p \leq \infty)\) in section 3. Section 4 deals with the problem (1.2) and we will obtain the optimal error estimates of \( u - u_h \) in \( W^{1,p}(I) \) and \( L_p(I) \) \((2 \leq p \leq \infty)\) in addition to the superconvergence estimates of \( \bar{u} - u_h \) and \( R^h u - u_h \) in \( W^{1,p}(I) \) \((2 \leq p \leq \infty)\).

2 Preparations

Let \( U = H^1_0(I) \equiv \{ v \in H^1(I), v(a) = v(b) = 0 \} \). Then the weak form of (1.1) is to find \( u \in U \) such that

\[
a(u,v) = (f,v), \quad \forall v \in U,
\]

where

\[
a(u,v) = \int_a^b pu'v' \, dx, \quad (f,v) = \int_a^b fv \, dx.
\]

We first define the partition \( T_h \) of the interval \( I = [a, b] \) with nodes \( x_i, i = 1, 2, \ldots, n, \)

\[
a = x_0 < x_1 < x_2 < \cdots < x_n = b.
\]

Let \( h_i = x_i - x_{i-1} \) denote the length of the element \( I_i = [x_{i-1}, x_i], \)

\[
h = \max_{1 \leq i \leq n} h_i \text{ and let the partition } T_h \text{ of } I \text{ be regular, that is, there exists a constant } \mu > 0 \text{ such that}
\]

\[
h_i \geq \mu h, \quad i = 1, 2, \ldots, n.
\]

Corresponding to the partition \( T_h \), we then introduce its dual partition \( T_h^* \) with nodes \( x_{i+\frac{1}{2}}, i = 0, \cdots, n - 1, \)

\[
a = x_0 < x_{\frac{1}{2}} < x_1 < \cdots < x_{n-\frac{1}{2}} < x_n = b.
\]
$I_0^* = [x_0, x_{1/2}]$, $I_j^* = [x_{j-1/2}, x_{j+1/2}]$ ($j = 1, 2, \cdots, n-1$), and $I_n^* = [x_{n-1/2}, x_n]$ are said to be dual elements, here

$$x_{j-1/2} = \frac{1}{2}(x_{j-1} + x_j), \quad 1 \leq j \leq n.$$  

Corresponding to the partition $T_h$, we choose the trial function space $U_h$ be the space of continuous piecewise linear functions, and $U_h = \text{span}\{\phi_i(x), 1 \leq i \leq n-1\}$.

The basis function $\phi_i$ corresponding to the node $x_i$ is

$$\phi_i(x) = \begin{cases} 
1 - h_i^{-1}|x - x_i|, & x_{i-1} \leq x \leq x_i, \\
1 - h_i^{-1}|x - x_i|, & x_i \leq x \leq x_{i+1}, \\
0, & \text{otherwise}, 
\end{cases} \quad i = 1, 2, \cdots, n-1,$$

so, any $u_h \in U_h$ can be expressed uniquely in the following way:

$$u_h(x) = \sum_{i=1}^{n-1} u_i \phi_i(x),$$

where $u_i = u_h(x_i)$, and on each element $I_i$, $i = 1, 2, \cdots, n$,

$$u'_h(x) = \frac{u_i - u_{i-1}}{h_i}, \quad x_{i-1} \leq x \leq x_i. \quad (2.3)$$

Corresponding to the partition $T_h^*$, let the test function space $V_h$ be the space of piecewise constant functions. Then the basis of $V_h$ may be taken to be characteristic functions of elements $I_j^*$,

$$\psi_j(x) = \begin{cases} 
1, & x \in I_j^*; \\
0, & \text{otherwise}, 
\end{cases} \quad j = 1, 2, \cdots, n-1,$$

and each $v_h \in V_h$ can be expressed uniquely in the following way:

$$v_h(x) = \sum_{j=1}^{n-1} v_j \psi_j(x).$$

Obviously

$$U_h(x) \subset U \cap W^{1, \infty}(I), \quad V_h \subset L_2(I).$$

Let’s define, for any $u_h \in U_h$ and $v_h \in V_h$, a bilinear form as follows

(a) \quad a^*(u_h, v_h) = \sum_{j=1}^{n-1} v_j a^*(u_h, \psi_j),

(b) \quad a^*(u_h, \psi_j) = p_{j-1/2} u_h'(x_{j-1/2}) - p_{j+1/2} u'_h(x_{j+1/2}) \equiv p_{j-1/2} \frac{u_{j-1} - u_{j}}{h_j} - p_{j+1/2} \frac{u_{j+1} - u_{j}}{h_{j+1}}, \quad (2.4)
where \( u_j = u_h(x_j) \), \( v_j = v_h(x_j) \), \( p_j = p(x_j) \), and \( j = 1, 2, \ldots, n - 1 \).

For numerical analysis, we need to introduce the interpolation operators \( \Pi_h : U \to U_h \), defined by

\[
\Pi_h w = \sum_{i=1}^{n-1} w(x_i) \phi_h(x), \quad \forall w \in U,
\]

and \( \Pi_h^* : U \to V_h \), defined by

\[
\Pi_h^* w = \sum_{j=1}^{n-1} w(x_j) \psi_j(x), \quad \forall w \in U.
\]

Using the theory of Sobolev’s interpolation, we have

\[
|w - \Pi_h w|_{m,p} \leq C h^{k-m} |w|_{k,p}, \quad m = 0, 1, \quad k = 1, 2, \quad 1 \leq p \leq \infty, \quad (2.5)
\]

\[
||| \Pi_h w |||_{m,p} \leq C |w|_{m,p}, \quad m = 0, 1, \quad 2 \leq p \leq \infty, \quad (2.6)
\]

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 \( h \).

Noting that for any \( u_h \in U_h \), we have, by (2.4),

\[
|u_h|_1 = \left[ \int_a^b |u_h'|^2 dx \right]^{1/2} = \left[ \sum_{i=1}^{n} \frac{(u_i - u_{i-1})^2}{h_i} \right]^{1/2}. \quad (2.7)
\]

Define the discrete \( L^2 \)-norm

\[
\|u_h\|_{0,h} = \left[ \sum_{i=1}^{n} h_i(u_i^2 + u_{i-1}^2) \right]^{1/2}, \quad u_h \in U_h. \quad (2.8)
\]

Then we can easily prove the following lemmas (See [4, 7]).

**Lemma 2.1** There exist positive constants \( C_1 \) and \( C_2 \) such that

\[
C_1 \|u_h\|_{0,h} \leq \|u_h\|_0 \leq C_2 \|u_h\|_{0,h}, \quad \forall u_h \in U_h. \quad (2.9)
\]

**Lemma 2.2** For any \( u_h \in U_h \), the norms \( ||\Pi_h^* u_h|| \) and \( ||u_h||_0 = (u_h, \Pi_h^* u_h)^{1/2} \) are equivalent to the \( L_2 \) norm \( \|u_h\|_0 \).

**Lemma 2.3** For any \( u_h, w_h \in U_h \), one has

\[
a^*(u_h, \Pi_h^* w_h) = a^*(w_h, \Pi_h^* u_h), \quad (2.10)
\]

\[
(u_h, \Pi_h^* w_h) = (w_h, \Pi_h^* u_h). \quad (2.11)
\]
Lemma 2.4 The bilinear form $a^*(\cdot, \Pi_h^*)$ is bounded over $U_h \times U_h$, that is, there exists a constant $M > 0$ such that
\[
|a^*(u_h, \Pi_h^* w_h)| \leq M \|u_h\|_1 \|w_h\|_1, \quad \forall u_h, w_h \in U_h. \tag{2.12}
\]

Lemma 2.5 The bilinear form $a^*(\cdot, \Pi_h^*) : U_h \times U_h \to R$ is positive definite, that is, there exists a positive constant $\alpha$ such that, for $h$ sufficiently small
\[
a^*(u_h, \Pi_h^* u_h) \geq \alpha \|u_h\|_1^2, \quad \forall u_h \in U_h. \tag{2.13}
\]

We now present a very useful lemma with respect to the bilinear form $a(\cdot, \cdot)$. For simplicity, we write
\[
d(u - u_h, w_h) = a(u - u_h, w_h) + a^*(u - u_h, \Pi_h^* w_h), \quad \forall u_h, w_h \in U_h. \tag{2.14}
\]

Lemma 2.6 [4] Let $u \in W^{3,q}(I)$, then, for $u_h, w_h \in U_h$,
\[
|d(u - u_h, w_h)| \leq C h [u - u_h]_{1,p} [w_h]_{1,p'} + h [u]_{2,q} [w_h]_{1,q'},
\]
with $1 \leq p, q \leq \infty$, $\frac{1}{p} + \frac{1}{p'} = 1$, $\frac{1}{q} + \frac{1}{q'} = 1$.

We prove another useful lemma similar to Lemma 2.6.

Lemma 2.7 Let $u \in W^{2,q}(I)$, then, for $u_h, w_h \in U_h$,
\[
|d(u - u_h, w_h)| \leq C h [u - u_h]_{1,p} [w_h]_{1,p'} + [u]_{2,q} [w_h]_{1,q'},
\]
with $1 \leq p, q \leq \infty$, $\frac{1}{p} + \frac{1}{p'} = 1$, $\frac{1}{q} + \frac{1}{q'} = 1$.

Proof. By making use of (2.3), we have, for $u_h$, $w_h \in U_h$
\[
a(u - u_h, w_h) = \int_a^b p(u - u_h)' w_h' dx
\]
\[
= \sum_{j=1}^n \int_{x_{j-1}}^{x_j} (p - p_{j-\frac{1}{2}})(u - u_h)' w_h' dx
\]
\[
+ \int_{x_{j-1}}^{x_j} p_{j-\frac{1}{2}}(u - u_h)' (w_j - w_{j-1}) \cdot \frac{w_j - w_{j-1}}{h_j} dx
\]
\[
= \sum_{j=1}^n \int_{x_{j-1}}^{x_j} (p - p_{j-\frac{1}{2}})(u - u_h)' w_h' dx
\]
\[
+ \sum_{j=1}^n p_{j-\frac{1}{2}} [u(x_j) - u(x_{j-1}) - u_j + u_{j-1}] dx \cdot \frac{w_j - w_{j-1}}{h_j},
\]
where $u_j = u_h(x_j)$, $w_j = w_h(x_j)$. 

\[L_p \text{ AND } W^{1,p} \text{ ERROR ESTIMATES} \]
On the other hand, by (2.4), (2.3) and noting \( w_0 = u_n = 0 \), we have,

\[
a^*(u - u_h, \Pi_h w_h) = \sum_{j=1}^{n-1} w_j a^*(u - u_h, \psi_j)
\]

\[
= \sum_{j=1}^{n-1} \left[ p_{j-\frac{1}{2}} (u - u_h)'_{j-\frac{1}{2}} - p_{j+\frac{1}{2}} (u - u_h)'_{j+\frac{1}{2}} \right] w_j
\]

\[
= \sum_{j=1}^{n-1} p_{j-\frac{1}{2}} (u - u_h)'_{j-\frac{1}{2}} w_j - \sum_{j=2}^{n} p_{j-\frac{1}{2}} (u - u_h)'_{j-\frac{1}{2}} w_{j-1}
\]

\[
= \sum_{j=1}^{n} p_{j-\frac{1}{2}} (u - u_h)'_{j-\frac{1}{2}} (w_j - w_{j-1})
\]

\[
= \sum_{j=1}^{n} p_{j-\frac{1}{2}} \left[ h_j u'(x_{j-\frac{1}{2}}) - u_j + u_{j-1} \right] \frac{w_j - w_{j-1}}{h_j}.
\]

Substitute the above relations into (2.14) and express (2.14) by the following

\[
d(u - u_h, w_h) = \sum_{j=1}^{2} E_j (u - u_h, w_h),
\]

where

\[
E_1 (u - u_h, w_h) = \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} (p - p_{j-\frac{1}{2}}) (u - u_h)' w_h \, dx,
\]

\[
E_2 (u - u_h, w_h) = \sum_{j=1}^{n} p_{j-\frac{1}{2}} \left[ u(x_j) - u(x_{j-1}) - h_j u'(x_{j-\frac{1}{2}}) \right] \frac{w_j - w_{j-1}}{h_j}.
\]

Thus

\[
|E_1 (u - u_h, w_h)| \leq C h |u - u_h|_{1,p} |w_h|_{1,q'}.
\]

Applying Taylor’s formula with integral type remainder

\[
g(x_j) - g(x_{j-1}) = g'(x_{j-\frac{1}{2}}) h_j + \int_{x_{j-\frac{1}{2}}}^{x_j} g''(x) (x_j - x) \, dx
\]

\[
- \int_{x_{j-\frac{1}{2}}}^{x_{j-1}} g''(x) (x_{j-1} - x) \, dx,
\]

we have

\[
|E_2 (u - u_h, w_h)| \leq C h \int_a^b |u'' w_h'| \, dx \leq C h |u|_{2,q} |w_h|_{1,q'}.
\]
Hence the conclusion (2.16) is a consequence of combination of estimates for \( E_1 \) and \( E_2 \). This completes the proof.

In order to derive maximum norm error estimates, we need to define the Green functions associated with the bilinear form \( a(\cdot, \cdot) \) (See [9]). Let \( G^h_z \in U_h \) and \( G^*_z \in H^1_0(I) \) be the discrete Green function and pre-Green function respectively, \( \partial_z G^*_z \) the directional derivative of \( G^*_z \) along some direction with respect to \( z \). Then \( G^h_z \) and \( \partial_z G^h_z \) are the finite element approximations to \( G^*_z \) and \( \partial_z G^*_z \), respectively. From [9], we know that

\[
\| G^*_z \|_{2,1} + \| \partial_z G^*_z \|_{1,1} \leq C. \tag{2.17}
\]

**Lemma 2.8** The discrete Green’s function \( G^h_z \) possess the following property.

\[
\| G^h_z \|_{1,\infty} \leq C. \tag{2.18}
\]

**Proof.** By using of inverse property, (2.5) and noting that \( W^{2,1}(I) \hookrightarrow w^{1,\infty}(I) \), we obtain

\[
\| G^h_z \|_{1,\infty} \leq \| G^*_z \|_{1,\infty} + \| G^*_z - \Pi_h G^*_z \|_{1,\infty} + \| \Pi_h G^*_z - G^h_z \|_{1,\infty} \\
\leq C \| G^*_z \|_{1,\infty} + C h^{-\frac{1}{6}} \| \Pi_h G^*_z - G^h_z \|_1 \\
\leq C \| G^*_z \|_{1,\infty} + C h^{-\frac{1}{6}} (\| \Pi_h G^*_z - G^*_z \|_1 + \| G^*_z - G^h_z \|_1) \\
\leq C \| G^*_z \|_{2,1} + C h^{-\frac{1}{6}} \| G^*_z \|_2. \tag{2.19}
\]

Let \( \delta^h_z \in U_h \) be the discrete Delta function defined by

\[
(v, \delta^h_z) = v(z), \quad \forall v \in U_h,
\]

Then the above definition and the inverse property imply that

\[
\| \delta^h_z \|_0^2 = (\delta^h_z, \delta^h_z) = |\delta^h_z(z)| \\
\leq \| \delta^h_z \|_{1,\infty} \leq C h^{-\frac{1}{6}} \| \delta^h_z \|_0,
\]

So also from [9], we have

\[
\| G^*_z \|_2 \leq C \| \delta^h_z \|_0 \leq C h^{-\frac{1}{7}},
\]

which together with (2.17) and (2.19) completes the proof.

### 3 Two-boundary Value problem

In this section, we consider the problem (1.1).

The first order generalized difference scheme for problem (1.1) is to find \( u_h \in U_h \) such that

\[
a^*(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{3.1}
\]
Using the definition of the operator $\Pi_h^*$ in section 2, we can obtain that the generalized difference scheme (3.1) is equivalent to finding $u_h \in U_h$, such that
\[
a^*(u_h, \Pi_h^* w_h) = (f, \Pi_h^* w_h), \quad \forall w_h \in U_h. \tag{3.2}
\]

Combining the results of Lax-Milgram’s Lemma, Lemmas 2.4 and 2.5, we have the solvability theorem.

**Theorem 3.1** The first order generalized difference scheme (3.1) has exactly one solution for $h$ sufficiently small.

In this section, we denote $u \in H^1_0(I)$ the weak solution of problem (1.1) and $u_h \in U_h$ the solution of (3.1).

In view of the generalized Galerkin variational principle (See [7]), we have
\[
a^*(u, v_h) = (f, v_h), \quad \forall v_h \in V_h.
\]
So that, by (3.1),
\[
a^*(u - u_h, v_h) = 0, \quad \forall v_h \in V_h. \tag{3.3}
\]

Now, we show the error estimate of $u - u_h$ in $W^{1,p}(I)$ ($2 \leq p \leq \infty$).

**Theorem 3.2** If $u \in W^{2,p}(I)$ ($2 \leq p < \infty$) and $h$ sufficiently small, then the following error estimate holds:
\[
\|u - u_h\|_{1,p} \leq Ch\|u\|_{2,p}, \quad 2 \leq p < \infty. \tag{3.4}
\]

**Proof.** We first introduce an auxiliary problem. Denote $\phi_x$ to be the derivative of $\phi$ and let $\Phi \in H^1_0(I)$ be the solution of
\[
a(v, \Phi) = -(v, \phi_x), \quad \forall v \in H^1_0(I), \tag{3.5}
\]
and there is a priori estimate
\[
\|\Phi\|_{1,p'} \leq C\|\phi\|_{0,p'}, \quad p' = \frac{p}{p-1}. \tag{3.6}
\]

Let $\Phi$ and $\bar{u}$ denote the standard finite element solutions of the problem (3.5) and (1.1) respectively. Then we have[2]
\[
a(v_h, \Phi - \bar{\Phi}) = 0, \quad \forall v_h \in U_h, \tag{3.7}
\]
\[
a(u - \bar{u}, v_h) = 0, \quad \forall v_h \in U_h, \tag{3.8}
\]
\[
\|u - \bar{u}\|_{s,q} \leq Ch^{2-s}\|u\|_{2,q}, \quad s = 0, 1, \quad 2 \leq q \leq \infty. \tag{3.9}
\]
By virtue of Green formula, (3.5), (3.7), (3.8), (3.3), Lemma 2.7, (3.9) and (3.6), we obtain that

\[
(u - u_h, \phi) = a(u - u_h, \phi) - (u - u_h, \Phi) + a(u - u_h, \Phi') \\
\leq C\|u - u\|_1 \|\Phi\|_1 + h\|u - u_h\|_1 \|\Phi\|_1 + \|u\|_2 \|\Phi\|_1 + \|u\|_2 \|\Phi\|_1 \\
\leq Ch\|u - u_h\|_1 \|u\|_2 + \|u\|_2 \|\phi\|_2.
\]

Thus

\[
\|(u - u_h)_x\|_0 = \sup_{\phi \in L^p(I)} \frac{\|(u - u_h)_x, \phi\|_1}{\|\phi\|_0} \leq Ch\|u - u_h\|_1 \|u\|_2,
\]

Therefore

\[
\|u - u_h\|_1 \leq Ch\|u - u_h\|_1 + \|u\|_2,
\]

where we have used the equivalence of the norms \(\|\cdot\|_1, \|\cdot\|_2\). Let \(h\) sufficiently small, such that \(Ch \leq \frac{1}{2}\), then the theorem follows at once from (3.10).

**Theorem 3.3** If \(u \in W^{2,\infty}(I)\) and \(h\) sufficiently small, then we have the following error estimate:

\[
\|u - u_h\|_2 \leq Ch\|u\|_2.
\]

**Proof.** The definition of \(\partial_z G^h_{1,2}\), (3.8), (3.3), Lemma 2.7, (3.9), and (2.17) imply that

\[
\partial_z(\bar{u} - u_h)(z) = a(\bar{u} - u_h, \partial_z G^h_{1,2}) \\
= a(u - u_h, \partial_z G^h_{1,2}) \\
= d(u - u_h, \partial_z G^h_{1,2}) \\
\leq Ch\|u - u_h\|_1 + \|u\|_2 \|\partial_z G^h_{1,2}\|_1 + Ch\|u - u_h\|_1 + \|u\|_2 \|\partial_z G^h_{1,2}\|_1
\]

Hence

\[
\|\bar{u} - u_h\|_2 \leq C(h\|\bar{u} - u_h\|_1 + h^2\|u\|_2).
\]

By letting \(h\) sufficiently small in the above inequality and using the triangle inequality

\[
\|u - u_h\|_0 \leq \|u - \bar{u}\|_0 + \|\bar{u} - u_h\|_1,
\]

we complete the proof also from (3.9).

Combining Theorems 3.2 and 3.3, we immediately derive the following.
**Theorem 3.4** If \( u \in W^{2,p}(I) \) \((2 \leq p \leq \infty)\) and \( h \) sufficiently small, then we have the following error estimate:

\[
\|u - u_h\|_{1,p} \leq Ch\|u\|_{2,p}, \quad 2 \leq p \leq \infty.
\] (3.14)

We then demonstrate the estimates of \( u - u_h \) in \( L_p(I) \) \((2 \leq p \leq \infty)\). From [4] and [5], we know that

\[
\|u - u_h\|_0 \leq Ch^2\|u\|_{3,1},
\]

\[
\|u - u_h\|_{0,\infty} \leq Ch^2\|u\|_3.
\]

The result is not perfect. We will modify the proof of the case of \( p = \infty \) to reduce the demand of the smoothness of the function \( u \). Then the case of \( 2 \leq p < \infty \) is the straight result of the case of \( p = \infty \).

**Theorem 3.5** If \( u \in W^{3,1}(I) \) and \( h \) sufficiently small, the following error estimate holds:

\[
\|u - u_h\|_{0,\infty} \leq Ch^2\|u\|_{3,1}.
\] (3.15)

**Proof.** Noting that the definition of \( G^h_z \), (3.12), Lemma 2.6, (3.14), (2.18) and Sobolev imbedding inequality, we deduce that

\[
(\tilde{u} - u_h)(z) = a(\tilde{u} - u_h, G^h_z) = d(u - u_h, G^h_z) \\
\leq Ch(|u - u_h| + h\|u\|_{3,1})|G^h_z|_{1,\infty} \\
\leq Ch^2(|u| + |u|_{3,1}) \\
\leq Ch^2\|u\|_{3,1},
\]

which together with (3.9), Sobolev imbedding inequality and triangle inequality completes the proof.

An application of the above theorem and the inequality

\[
\|u - u_h\|_{0,p} \leq C\|u - u_h\|_{0,\infty}, \quad 2 \leq p < \infty.
\]

immediately yields the following result.

**Corollary 3.1** Under the hypotheses of Theorem 3.5, we have the following error estimate:

\[
\|u - u_h\|_{0,p} \leq Ch^2\|u\|_{3,1}, \quad 2 \leq p \leq \infty.
\] (3.16)

We now turn to the superconvergence estimate of \( \tilde{u} - u_h \) in \( W^{1,p}(I) \) \((2 \leq p \leq \infty)\).

**Theorem 3.6** If \( u \in W^{3,p}(I) \) \((2 \leq p \leq \infty)\) and \( h \) sufficiently small, then we have

\[
\|\tilde{u} - u_h\|_{1,p} \leq Ch^2\|u\|_{3,p}, \quad 2 \leq p \leq \infty.
\] (3.17)
Proof. We first consider the case of $2 \leq p < \infty$.

By Lemma 2.6, Theorem 3.2, and a similar analysis to that in Theorem 3.2, we obtain that

$$((\bar{u} - u_h)_x, \phi) = -(\bar{u} - u_h, \phi_x)$$
$$= a(\bar{u} - u_h, \Phi)$$
$$= a(\bar{u} - u_h, \bar{\Phi})$$
$$= d(u - u_h, \bar{\Phi})$$
$$\leq Ch ||u - u_h||_{1,p} + h||u||_{3,p}||\Phi||_{1,p'}$$
$$\leq Ch^2(||u||_{2,p} + ||u||_{3,p}||\Phi||_{1,p'}).$$

Accordingly, (3.17) is derived from the above inequality and (3.6), for $2 \leq p < \infty$.

As far as the case of $p = \infty$ is concerned, it suffices to see the proof of Theorem 3.3 and the difference is that we use Lemma 2.6 instead of Lemma 2.7.

4 One-Dimensional Parabolic Problem

In this section, we consider the problem (1.2) on the base of the results derived in section 3.

Define the semi-discrete generalized difference scheme for problem (1.2): Find $u_h(t) : [0, T] \to U_h$ such that

$$\begin{cases}
(a) & (u_{h,t}, v_h) + a^*(u, v_h) = (f, v_h), \quad \forall v_h \in V_h, \quad 0 < t \leq T, \\
(a) & u_h(0) = u_{0,h},
\end{cases}$$

(4.1)

where $u_{h,t} = \frac{u_{h,t} - u_{h,0}}{h}$, $u_{0,h} \in U_h$ is taken the generalized elliptic projection $R_h^* u_0$ of $u_0$ defined in (4.2) below.

It can be proved that (4.1) has a unique solution for any $f \in L_2(I)$ (See [4, 7]).

For later use, we introduce the generalized elliptic projection operator $R_h^* : H^2(I) \cap H_0^1(I) \to U_h$ defined by

$$a^*(R_h^* w, v_h) = a^*(w, v_h), \quad \forall v_h \in V_h.$$  

(4.2)

It is easily seen that, from Lemmas 2.4 and 2.5, $R_h^* w$ is uniquely determined by (4.2) for any given $w \in H^1(I) \cap H_0^1(I)$.

Applying Theorems 3.4 and 3.6, we have the following error estimates.

Lemma 4.1 Let $R_h^*$ be defined by (4.2), then

$$\begin{cases}
(a) & \|w - R_h^* w\|_{1,p} \leq Ch \|w\|_{2,p}, \quad 2 \leq p \leq \infty; \\
(b) & \|w - R_h^* w\|_{0,p} \leq Ch^2 \|w\|_{3,1}, \quad 2 \leq p \leq \infty.
\end{cases}$$

(4.3)

Throughout this section, we denote $u \in H_0^1(I)$ the weak solution of problem (1.2), $u_h \in U_h$ the solution of problem (4.1) and write

$$u - u_h = (u - R_h^* u) + (R_h^* u - u_h) = \eta + \xi.$$
Let \( \tilde{u} \) be the finite element solution of problem (1.2), that is, \( \tilde{u} \in U_h \) satisfy:

\[
\begin{cases}
(\tilde{u}_t, v_h) + a(\tilde{u}, v_h) = (f, v_h), & \forall v_h \in U_h, \\
\tilde{u}(0) = \tilde{u}_0,
\end{cases}
\]

where \( \tilde{u}_t = \frac{\partial \tilde{u}}{\partial t} \), \( \tilde{u}_0 \in U_h \) is taken the elliptic projection \( R_h u_0 \) of \( u_0 \) defined as follows: Find \( R_h u \in U_h \) such that

\[
a(u - R_h u, v_h) = 0, \quad \forall v_h \in U_h. \tag{4.4}
\]

First, we give the superconvergence estimate of \( \xi \) in \( W^{1,p}(I) \) \( (2 \leq p \leq \infty) \). In order to get that we need to deduce the error estimate of \( \xi_t \) in \( L_2(I) \).

**Lemma 4.2** If \( u_t(0) \in W^{3,1}(I) \), \( u_{tt} \in L_2(0, t; u^{3,1}(I)) \), then we have the following error estimate:

\[
\|\xi_t\|_0 \leq C h^2 \{ \|u_t(0)\|_{3,1} + \left( \int_0^t \|u_{tt}\|_{3,1}^2 \, dt \right)^{1/2} \}. \tag{4.5}
\]

**Proof.** Multiplying (1.2) by \( v_h \) and integrating by parts, we have

\[
(u_t, v_h) + a^*(u, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{4.6}
\]

Subtracting (4.6) from (4.1) and applying (4.2), we obtain

\[
(\xi_t, v_h) + a^*(\xi, v_h) = -(\eta_t, v_h), \quad \forall v_h \in V_h. \tag{4.7}
\]

Taking \( t = 0 \) in (4.7) and noting that \( u_h(0) = R_h^* u_h \) implies \( \xi(0) = 0 \), then taking \( v_h(0) = \Pi_h^* \xi(0) \), we have, also by Lemma 2.2 and \( \varepsilon \)-inequality, that

\[
\|\|\xi(0)\|_0^2 \|_0 \leq -\langle \eta(0), \Pi_h^* \xi(0) \rangle \\
\leq C \|\eta(0)\|_0 \|\Pi_h^* \xi(0)\|_0 \\
\leq C \|\eta(0)\|_0^2 + \frac{\varepsilon}{2} \|\xi(0)\|_0^2,
\]

where \( \|\| \cdot \|\| \) is defined in Lemma 2.2. Then by (4.3b)

\[
\|\|\xi(0)\|_0^2 \|_0 \leq C \|\eta(0)\|_0 \leq C h^2 \|u_t(0)\|_{3,1}. \tag{4.8}
\]

Differentiate (4.7) with respect to \( t \) and take \( v_h = \Pi_h^* \xi_t \) to get

\[
(\xi_{tt}, \Pi_h^* \xi_t) + a^*(\xi_t, \Pi_h^* \xi_t) = -(\eta_t, \Pi_h^* \xi_t). \tag{4.9}
\]

Lemmas 2.3, 2.5 and 2.2 now imply that

\[
\frac{d}{dt}(\xi, \Pi_h^* \xi) + C \|\xi_t\|_0^2 \leq \langle \xi_t, \Pi_h^* \xi_t \rangle + a^*(\xi_t, \Pi_h^* \xi_t) \leq C(\|\eta_t\|_0^2 + \|\xi_t\|_0^2). \tag{4.10}
\]
Integrating (4.10) with respect to $t$, we have
\[
\|\xi(t)\|_0^2 + \int_0^t \|\xi(t)\|_1^2 \, dt \leq C(\|\xi(0)\|_0^2 + \int_0^t \|\eta(t)\|_0^2 \, dt + \int_0^t \|\zeta(t)\|_0^2 \, dt).
\]
By Gronwall Lemma, (4.8) and (4.3b), we have
\[
\|\xi(t)\|_0^2 \leq C(\|\xi(0)\|_0^2 + \int_0^t \|\eta(t)\|_0^2 \, dt) \leq C h^4 (\|u(0)\|^2_{3,1} + \int_0^t \|u(t)\|^2_{3,1} \, dt). \tag{4.11}
\]
(4.11) and Lemma 2.2 imply (4.5).

**Theorem 4.1** If $u(0), u_{tt} \in W^{3,1}(I), u \in W^{3,p}(I)$ and $h$ sufficiently small, then the following superconvergence estimate of $\xi$ in $W^{1,p}(I) \ (2 \leq p \leq \infty)$ holds:
\[
\|\xi\|_{1,p} \leq C h^2 \{\|u\|_{3,p} + \|u(0)\|_{3,1} + (\int_0^t \|u(t)\|^2_{3,1} \, dt)^{\frac{1}{2}}\}, \quad 2 \leq p \leq \infty. \tag{4.12}
\]

**Proof.** (i) Let us first consider the case of $2 \leq p < \infty$.
Using Green formula, (3.5), (3.7), (4.2) and (4.7), we write
\[
(\xi_t, \phi) = -(\xi, \phi_t) = a(\xi, \Phi) = a(\xi, R_h \Phi) = a(R_h^\ast u - u, R_h \Phi) + a(u - u_h, R_h \Phi) = a(R_h^\ast u - u, R_h \Phi) - a^\ast(R_h^\ast u - u, \Pi_h \Phi) + a(u - u_h, R_h \Phi) - a^\ast(u - u_h, \Pi_h \Phi) = d(R_h^\ast u - u, R_h \Phi) + d(u - u_h, R_h \Phi) - (\eta_t + \xi_t, \Pi_h \Phi) = E_1 + E_2 + E_3. \tag{4.13}
\]
Applying Lemma 2.6 and (4.3a), we get
\[
E_1 \leq C h (|R_h^\ast u - u|_{1,p} + h |u|_{3,p}) |R_h \Phi|_{1,p'} \leq C h^2 \|u\|_{3,p} |\Phi|_{1,p'},
\]
and
\[
E_2 \leq C h (|u - u_h|_{1,p} + h |u|_{3,p}) |R_h \Phi|_{1,p'} \leq C h (|u - R_h^\ast u|_{1,p} + |\xi|_{3,p} + h |u|_{3,p}) |\Phi|_{1,p'} \leq C (h |\xi|_{1,p} + h^2 |u|_{3,p}) |\Phi|_{1,p'}.
\]
Also we know, from (4.3b), that
\[
\|\eta_t\|_0 \leq \|\eta(0)\|_0 + \int_0^t \|\eta(t)\|_0 \, dt \leq C h^2 \{\|u(t)\|_{3,1} + \int_0^t \|u(t)\|_{3,1} \, dt\} \leq C h^2 \{\|u(t)\|_{3,1} + (\int_0^t \|u(t)\|^2_{3,1} \, dt)^{\frac{1}{2}}\},
\]
which together with (4.5) and Sobolev imbedding inequality implies that

\[ E_3 \leq (\|u_t\|_0 + \|\xi_t\|_0)\|\Pi_h^\ast R_h \Phi\|_0 \]
\[ \leq Ch^2 \{ (\|u_t(0)\|_{3,1} + \left( \int_0^t \|u_t\|_{3,1}^2 d\tau \right) \} \|\Phi\|_{1,p'}. \]

Combining the estimates of \( E_1 - E_3 \), we obtain also by (3.6) and (4.13) that

\[ \|\xi\|_{1,p} \leq C \sup_{\phi \in L^p(I)} \frac{(\xi, \phi)}{\|\phi\|_{0,p'}} \]
\[ \leq C \{ h\|\xi\|_{1,p} + h^2 [\|u\|_{3,p} + \|u_t(0)\|_{3,1} + \left( \int_0^t \|u_t\|_{3,1}^2 d\tau \right) ] \}. \]

By letting \( h \) sufficiently small such that \( Ch \leq \frac{1}{2} \), we can complete the proof of the case of \( 2 \leq p < \infty \).

(ii) Let us next consider the case of \( p = \infty \). By virtue of the definition of \( \partial_2 G^h_2 \), we have

\[ \partial_2 \xi(z) = a(\xi, \partial_2 G^h_2), \]

consequently, upon replacing \( R_h \Phi \) by \( \partial_2 G^h_2 \), \( p \) by \( \infty \), \( p' \) by 1 in part (i), likewise, we obtain the conclusion.

Arguing as in the proof of Lemma 4.2, we find that if we use (4.3a) instead of (4.3b) in (4.8) and (4.11), we can get the following lemma.

**Lemma 4.3** If \( u_t(0) \in W^{2,p}(I), \ u_{tt} \in L^2(0,t; W^{2,p}(I)) \), then the following error estimate holds:

\[ \|\xi_t\|_0 \leq Ch \{ \|u_t(0)\|_{2,p} + \left( \int_0^t \|u_t\|_{2,p}^2 d\tau \right) \}, \ 2 \leq p \leq \infty. \quad (4.14) \]

As to the error estimate of \( u - u_h \) in \( W^{1,p}(I) \), we obtain it by the similar way to getting Theorem 4.1. In the proof of Theorem 4.1, we use Lemma 2.7 instead of Lemma 2.6 and (4.14) instead of (4.5) for the error estimates of \( E_1 - E_3 \) to derive

\[ \|\xi\|_{1,p} \leq Ch \{ \|u_t(0)\|_{2,p} + \|u\|_{2,p} + \left( \int_0^t \|u_{tt}\|_{2,p}^2 d\tau \right) \}, \ 2 \leq p \leq \infty, \quad (4.15) \]

which together with (4.3a), by using triangle inequality leads to the following theorem.

**Theorem 4.2** If \( u_t(0), u \in W^{2,p}(I), \ u_{tt} \in L^2(0,t; W^{2,p}(I)) \) and \( h \) sufficiently small, then we have the following error estimate, for \( 2 \leq p \leq \infty \)

\[ \|u - u_h\|_{1,p} \leq Ch \{ \|u_t(0)\|_{2,p} + \|u\|_{2,p} + \left( \int_0^t \|u_{tt}\|_{2,p}^2 d\tau \right) \}. \quad (4.16) \]
Now we turn to the error estimates of \( u-u_h \) in \( L_p(I) \) \((2 \leq p \leq \infty)\). We only demonstrate the case of \( p=\infty \) and the case of \( 2 \leq p < \infty \) is immediate result.

**Theorem 4.3** If \( u(0), u \in W^{3,1}(I), u \in L_2(0,t; w^{3,1}(I)) \) and \( h \) sufficiently small, then the following error estimate holds:

\[
\|u-u_h\|_{0,\infty} \leq Ch^2 \{\|u(0)\|_{3,1} + \|u\|_{3,1} + \left( \int_0^t \|u_{tt}\|^2_{3,1} d\tau \right)^{\frac{1}{2}}\}.
\]  

(4.17)

**Proof.** In view of the definition of \( G^h \) and (4.13), we obtain that

\[
\xi(z) = a(\xi,G^h)
= d(R^h u - u,G^h) + d(u - u_h, G^h) - (\eta_t + \xi_t, \Pi^h G^h)
= Q_1 + Q_2 + Q_3.
\]

Lemma 2.6, (4.3a), (2.18) and Sobolev imbedding inequality imply that

\[
Q_1 \leq Ch(\|R^h u - u\|_1 + h\|u\|_{3,1})|G^h|_{1,\infty}
\leq Ch^2\|u\|_{\cdot,3,1},
\]

and that, also from (4.16)

\[
Q_2 \leq Ch(\|u-u_h\|_1 + h\|u\|_{3,1})|G^h|_{1,\infty}
\leq Ch^2\{\|u(0)\|_2 + \|u\|_2 + \left( \int_0^t \|u_{tt}\|^2_2 d\tau \right)^{\frac{1}{2}} + |u|_{3,1}\}
\leq Ch^2\{\|u(0)\|_{3,1} + \|u\|_{3,1} + \left( \int_0^t \|u_{tt}\|^2_{3,1} d\tau \right)^{\frac{1}{2}}\}.
\]

Taking into account (4.3b), (4.5) and Sobolev imbedding inequality, we find that

\[
Q_3 \leq (\|\eta_0\|_0 + \|\xi_0\|_0)|G^h|_{1,\infty}
\leq Ch^2\{\|u_t(0)\|_{3,1} + \left( \int_0^t \|u_{tt}\|^2_{3,1} d\tau \right)^{\frac{1}{2}}\}.
\]

The inequality (4.17) follows by combining the error estimates of \( Q_1-Q_3 \), (4.3b) and triangle inequality.

**Corollary 4.1** Under the hypotheses of Theorem 4.3, we have the following error estimate, for \( 2 \leq p \leq \infty \)

\[
\|u-u_h\|_{0,p} \leq Ch^2\{\|u(0)\|_{3,1} + \|u\|_{3,1} + \left( \int_0^t \|u_{tt}\|^2_{3,1} d\tau \right)^{\frac{1}{2}}\}.
\]

(4.18)
Finally, we deduce the superconvergence estimates of $\tilde{u} - u_h$ and $R_h^* u - \tilde{u}$ in $W^{1,p}(I)$ ($2 \leq p \leq \infty$).

From [3], we know that

$$
\parallel \tilde{u} - R_h u \parallel_{1,p} \leq C h^2 \left( \int_0^t \parallel u_t \parallel_{2,p}^2 d\tau \right)^{1/2}, \quad 2 \leq p \leq \infty. \quad (4.19)
$$

**Theorem 4.4** The following superconvergence estimate holds, for $2 \leq p \leq \infty$

$$
\parallel \tilde{u} - u_h \parallel_{1,p} \leq \begin{aligned}
&Ch^2 \{ \parallel u(0) \parallel_{3,p} + \parallel u_t(0) \parallel_{3,1} \\
&\quad + \left( \int_0^t \parallel u_t \parallel_{3,p}^2 d\tau \right)^{1/2} + \left( \int_0^t \parallel u \parallel_{3,1}^2 d\tau \right)^{1/2} \}.
\end{aligned} \quad (4.20)
$$

**Proof.** Using (3.5) and (4.4), we have

$$
((\tilde{u} - u_h)_x, \phi) = -((\tilde{u} - u_h, \phi_x)) = a(\tilde{u} - u_h, \tilde{\phi}) = a(\tilde{u} - u_h, R_b \tilde{\phi})
$$

$$
= a(\tilde{u} - R_b u, R_b \tilde{\phi}) + a(\tilde{u} - u_h, R_b \tilde{\phi})
$$

$$
\leq \begin{aligned}
&Ch^2 \{ \parallel u(0) \parallel_{1,p} \parallel R_b \tilde{\phi} \parallel_{1,p} \\
&\quad + \parallel u_t(0) \parallel_{3,1} \}.
\end{aligned}
$$

Therefore, similar to Theorem 4.1, the proof is easily complete also by (4.19).

Theorems 4.1 and 4.4 together with

$$
\parallel R_h^* u - \tilde{u} \parallel_{1,p} \leq \parallel \xi \parallel_{1,p} + \parallel \tilde{u} - u_h \parallel_{1,p},
$$

yield the following

**Corollary 4.2** The following superconvergence estimate holds, for $2 \leq p \leq \infty$

$$
\parallel R_h^* u - \tilde{u} \parallel_{1,p} \leq \begin{aligned}
&Ch^2 \{ \parallel u(0) \parallel_{3,p} + \parallel u_t(0) \parallel_{3,1} \\
&\quad + \left( \int_0^t \parallel u_t \parallel_{3,p}^2 d\tau \right)^{1/2} + \left( \int_0^t \parallel u \parallel_{3,1}^2 d\tau \right)^{1/2} \}.
\end{aligned} \quad (4.21)
$$

**References**


Department of Mathematics
Shandong Normal University
Jinan, Shandong, 250014, P. R. China
e-mail: li_qian@163.net
OPTIMALITY FOR MULTIOBJECTIVE
FRAC TIONAL VARIATIONAL PROGRAMMING

Cheonglai Jo and Dosang Kim

Abstract. We consider a multiobjective fractional variational programming problem (P) involving vector valued functions. By using the concept of proper efficiency, a relationship between the primal problem and parametric multiobjective variational problem is indicated.

1. Introduction and problems

Programs with several conflicting objectives have been extensively studied in literatures [1]-[11]. Introducing the concept of proper efficiency of solutions, Geoffrion[5] proved an equivalence between a multiobjective program with convex functions and a related parametric objective program. Using this equivalence, optimality and duality for multiobjective variational problems have been of much interest in recent years, and contributions have been made to its development. Bector and Husain[2] formulated a dual program for a multiobjective variational program having properly efficient solutions. Also, using parametric equivalence, Bector and Husain[1] formulated a dual program for a multiobjective fractional program having continuously differentiable convex functions. Further, Mishra and Mukherjee[6] considered the duality of multiobjective fractional variational problems by relating the primal problem to a parametric multiobjective variational problem.

Motivated by the above results, in this paper we propose studying optimality for multiobjective fractional variational problems having properly efficient solutions.

We consider a multiobjective fractional variational programming problem (P) involving vector valued functions.

\[
\text{(P) \ Minimize} \quad \frac{\int_a^b f(t, x(t), \dot{x}(t)) dt}{\int_a^b g(t, x(t), \dot{x}(t)) dt}
\]

\[
= \left[ \frac{\int_a^b f^1(t, x(t), \dot{x}(t)) dt}{\int_a^b g^1(t, x(t), \dot{x}(t)) dt}, \ldots, \frac{\int_a^b f^n(t, x(t), \dot{x}(t)) dt}{\int_a^b g^n(t, x(t), \dot{x}(t)) dt} \right]
\]

This research was supported by Pukyong National University supporting association.
subject to
\[ x(a) = \alpha, \quad x(b) = \beta \]  
\[ h^j(t, x(t), \dot{x}(t)) \leq 0, \quad t \in I, \quad j = 1, 2, \ldots, m, \]  
\[ x \in S, \quad \dot{x} \in C(I, \mathbb{R}^n). \]  

We assume that \( g^i(t, x(t), \dot{x}(t)) > 0 \) and \( f^i(t, x(t), \dot{x}(t)) \geq 0 \) whenever \( g^i(x) \) is not linear for all \( i = 1, 2, \ldots, p \).

To optimize \((P)\) is to find properly efficient solutions.

Geoffrion\[5\] introduced the definition of the properly efficient solution in order to eliminate the efficient solutions causing unbounded trade-offs between objective functions.

Corresponding to \((P)\), we consider the following parametric vector variational problem \((P_v)\).

\((P_v)\) \ Minimize \[ \int_a^b \{ f^1(t, x(t), \dot{x}(t)) - v_1 g^1(t, x(t), \dot{x}(t)) \} dt \]
\[ , \ldots \]
\[ \int_a^b \{ f^p(t, x(t), \dot{x}(t)) - v_p g^p(t, x(t), \dot{x}(t)) \} dt \]

subject to
\[ x(a) = \alpha, \quad x(b) = \beta, \]  
\[ h^j(t, x(t), \dot{x}(t)) \leq 0, \quad t \in I, \quad j = 1, 2, \ldots, m, \]  
\[ x \in S, \quad \dot{x} \in C(I, \mathbb{R}^n). \]  

In this paper, we prove that \((P)\) and \((P_v)\) have equivalent properly efficient solutions. We can obtain optimality and duality for \((P)\) by using of this equivalent relation.

We give some definitions and results from [2] and [7], which are used subsequently in our later results.

Let \( I = [a, b] \) be a real interval and \( f = (f^1, \ldots, f^p) : I \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^p \), \( g = (g^1, \ldots, g^p) : I \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^p \) and \( h = (h^1, \ldots, h^m) : I \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^m \) be continuously differentiable functions. In order to consider \( f^i(t, x(t), \dot{x}(t)) \), where \( x : I \rightarrow \mathbb{R}^n \) with derivative \( \dot{x} \), denote the partial derivative of \( f^i \) with respect to \( t, x \) and \( \dot{x} \) respectively, by \( \frac{\partial f^i}{\partial t}, \frac{\partial f^i}{\partial x}, \frac{\partial f^i}{\partial \dot{x}} \) such that
\[ f^i_{t} = \left( \frac{\partial f^i}{\partial t}, \frac{\partial f^i}{\partial x}, \ldots, \frac{\partial f^i}{\partial x_n} \right), \quad f^i_{x} = \left( \frac{\partial f^i}{\partial x_1}, \frac{\partial f^i}{\partial x_2}, \ldots, \frac{\partial f^i}{\partial x_n} \right), \]
\[ f^i_{\dot{x}} = \left( \frac{\partial f^i}{\partial \dot{x}_1}, \frac{\partial f^i}{\partial \dot{x}_2}, \ldots, \frac{\partial f^i}{\partial \dot{x}_n} \right). \]

Similarly, we write the partial derivatives of the vector functions \( g \) and \( h \) using matrices \( p \times n \) and \( m \times n \), respectively.
Let \( C(I, R^n) \) denote the space of piecewise smooth functions \( x \) with norm \( \| x \| = \| x \|_{\infty} + \| Dx \|_{\infty} \), where the differentiation operator \( D \) is given by

\[
u = Dx \iff x(t) = \alpha + \int_{\alpha}^t u(s) \, ds,
\]

where \( \alpha \) is a given boundary value.

Therefore, \( D = \frac{d}{dt} \) except at discontinuities.

Let \( S \subseteq R^n \) be open.

Let \( X \) the set of all feasible solutions of \((P) \) be given by

\[
X = \{ x \in C(I, R^n) \mid x(a) = \alpha, \ x(b) = \beta, \ h^j(t, x(t), \dot{x}(t)) \leq 0, \quad t \in I, \quad j = 1, 2, \cdots, m \}.
\]

In the sequel we shall always denote the set \( \{1, 2, \cdots, p\} \) and \( \{1, 2, \cdots, m\} \) by \( \bar{p} \) and \( \bar{m} \) respectively.

**Definition 1.** A feasible solution \( x^* \) of \((P) \) is an efficient solution of \((P) \) if there exist no other feasible \( x \) for \((P) \) such that

\[
\frac{\int_a^b f^i(t, x(t), \dot{x}(t)) \, dt}{\int_a^b g^i(t, x(t), \dot{x}(t)) \, dt} \leq \frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t)) \, dt} \quad \text{for all } i \in \bar{p} \tag{3}
\]

and

\[
\frac{\int_a^b f^i(t, x(t), \dot{x}(t)) \, dt}{\int_a^b g^i(t, x(t), \dot{x}(t)) \, dt} < \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t)) \, dt} \quad \text{for some } j \in \bar{p}. \tag{4}
\]

By eliminating efficient solutions causing unbounded trade-off between objective functions, we can define the proper efficient solutions as follows.

**Definition 2[5].** A feasible solution \( x^* \) of \((P) \) is a properly efficient solution of \((P) \) if it is efficient and if there exists a scalar \( M > 0 \) such that, for each \( i, \)

\[
\frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t)) \, dt} - \frac{\int_a^b f^i(t, x(t), \dot{x}(t)) \, dt}{\int_a^b g^i(t, x(t), \dot{x}(t)) \, dt} \leq M
\]

for some \( j \) such that

\[
\frac{\int_a^b f^j(t, x(t), \dot{x}(t)) \, dt}{\int_a^b g^j(t, x(t), \dot{x}(t)) \, dt} > \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t)) \, dt}
\]
whenever $x$ is feasible for (P) and

\[
\frac{\int_a^b f^i(t, x(t), \dot{x}(t)) dt}{\int_a^b g^i(t, x(t), \dot{x}(t)) dt} < \frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t)) dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t)) dt}
\]

An efficient point that is not properly efficient is said to be improperly efficient. Thus for $x^*$ to be improperly efficient means that for every scalar $M > 0$ (no matter how large) there is feasible point $x$ and an $i$ such that

\[
\frac{\int_a^b f^i(t, x(t), \dot{x}(t)) dt}{\int_a^b g^i(t, x(t), \dot{x}(t)) dt} < \frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t)) dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t)) dt}
\]

and

\[
\frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t)) dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t)) dt} - \frac{\int_a^b f^j(t, x(t), \dot{x}(t)) dt}{\int_a^b g^j(t, x(t), \dot{x}(t)) dt} > M
\]

for all $j$ such that

\[
\frac{\int_a^b f^j(t, x(t), \dot{x}(t)) dt}{\int_a^b g^j(t, x(t), \dot{x}(t)) dt} > \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t)) dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t)) dt}.
\]

2. Main Result

The following theorem connects (P) and (P_v) with $v = v^*$.

**Theorem 3** $x^*$ is a properly efficient solution of (P) if and only if there exists

\[
v_j^* = \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t)) dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t)) dt},
\]

for some $j \in \tilde{p}$ such that $x^*$ is a properly efficient solution of (P_v) with $v = v^*$.

**Proof.** Let $x^*$ be a properly efficient solution of (P) and let

\[
v_j^* = \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t)) dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t)) dt} \text{ for some } j \in \tilde{p}.
\]

If $x^*$ is not an efficient solution of (P_v) with $v = v^*$, then there exists a feasible solution $x$ of (P_v) with $v = v^*$, such that

\[
\int_a^b \{ f^i(t, x(t), \dot{x}(t)) - v_i^* g_i(t, x(t), \dot{x}(t)) \} dt \\
\leq \int_a^b \{ f^i(t, x^*(t), \dot{x}^*(t)) - v_i^* g_i(t, x^*(t), \dot{x}^*(t)) \} dt \text{ for all } i \in \tilde{p}
\]

for some $M > 0$. Therefore, $x^*$ is improperly efficient.
and
\[
\int_a^b \{ f^j(t, x(t), \dot{x}(t)) - v_j^* g^j(t, x(t), \dot{x}(t)) \} \, dt \\
< \int_a^b \{ f^j(t, x^*(t), \dot{x}^*(t)) - v_j^* g^j(t, x^*(t), \dot{x}^*(t)) \} \, dt \quad \text{for some } j \in \bar{P}.
\]

It follows that
\[
\frac{\int_a^b f^i(t, x(t), \dot{x}(t)) \, dt}{\int_a^b g^i(t, x(t), \dot{x}(t)) \, dt} \leq \frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t)) \, dt} \quad \text{for all } i \in \bar{P} \tag{6}
\]
and
\[
\frac{\int_a^b f^j(t, x(t), \dot{x}(t)) \, dt}{\int_a^b g^j(t, x(t), \dot{x}(t)) \, dt} < \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t)) \, dt} \quad \text{for some } j \in \bar{P} \tag{7}
\]
Contradicting the efficiency of \( x^* \) in (P). Hence \( x^* \) is an efficient solution of \((P_v)\) with \( v = v^* \).

Now we shall show \( x^* \) is a properly efficient solution of \((P_v)\) with \( v = v^* \). If \( x^* \) is not properly efficient for \((P_v)\) with \( v = v^* \), then, for every sufficiently large scalar \( M > 0 \), there is \( x \in X \) and an \( i \) such that
\[
\int_a^b \{ f^i(t, x(t), \dot{x}(t)) - v_i^* g^i(t, x(t), \dot{x}(t)) \} \, dt < 0 \tag{8}
\]
and
\[
\frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t)) - v_i^* g^i(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t)) \, dt} - \int_a^b \{ f^i(t, x(t), \dot{x}(t)) - v_i^* g^i(t, x(t), \dot{x}(t)) \} \, dt \\
\int_a^b \{ f^i(t, x(t), \dot{x}(t)) - v_i^* g^i(t, x(t), \dot{x}(t)) \} \, dt - \int_a^b \{ f^i(t, x^*(t), \dot{x}^*(t)) - v_i^* g^i(t, x^*(t), \dot{x}^*(t)) \} \, dt \\
\int_a^b \{ f^i(t, x(t), \dot{x}(t)) - v_i^* g^i(t, x(t), \dot{x}(t)) \} \, dt - \int_a^b \{ f^i(t, x^*(t), \dot{x}^*(t)) - v_i^* g^i(t, x^*(t), \dot{x}^*(t)) \} \, dt
\]
\( > M \tag{9} \)
for all \( j \) such that
\[
\int_a^b \{ f^j(t, x(t), \dot{x}(t)) - v_j^* g^j(t, x(t), \dot{x}(t)) \} \, dt > 0 \tag{10}
\]
i.e.,
\[
\frac{\int_a^b f^i(t, x(t), \dot{x}(t)) \, dt}{\int_a^b g^i(t, x(t), \dot{x}(t)) \, dt} < \frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t)) \, dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t)) \, dt} \tag{8'}
\]
for all $j$ such that

$$
\frac{\int_a^b f^j(t, x(t), \dot{x}(t))dt}{\int_a^b g^j(t, x(t), \dot{x}(t))dt} > \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t))dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t))dt}
$$

(10')

Now (9') can be rewritten as

$$
\frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t))dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t))dt} - \frac{\int_a^b f^i(t, x(t), \dot{x}(t))dt}{\int_a^b g^i(t, x(t), \dot{x}(t))dt} > M.
$$

So (8'), (9') and (10') imply that $x^*$ is not properly efficient for $(P)$. Hence $x^*$ is properly efficient in $(P_v)$ with $v = v^*$.

Conversely, let $x^*$ be a properly efficient solution of $(P_v)$ with $v = v^*$ where

$$
v_j^* = \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t))dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t))dt} \text{ for } j \in \bar{p}.
$$

(5)

then we shall show that $x^*$ is properly efficient for $(P)$.

If $x^*$ is not an efficient solution of $(P)$, then there exists a feasible solution $x$ for $(P)$ such that

$$
\frac{\int_a^b f^i(t, x(t), \dot{x}(t))dt}{\int_a^b g^i(t, x(t), \dot{x}(t))dt} \leq \frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t))dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t))dt} \text{ for all } i \in \bar{p}.
$$

(6)

and

$$
\frac{\int_a^b f^j(t, x(t), \dot{x}(t))dt}{\int_a^b g^j(t, x(t), \dot{x}(t))dt} < \frac{\int_a^b f^j(t, x^*(t), \dot{x}^*(t))dt}{\int_a^b g^j(t, x^*(t), \dot{x}^*(t))dt} \text{ for some } j \in \bar{p}.
$$

(7)

Now (5) together with (6) and (7) contradict the efficiency of $x^*$ in $(P_v)$ with $v = v^*$.

Thus $x^*$ is an efficient solution of $(P)$.

Now we shall show that $x^*$ is a properly efficient solution of $(P)$.

If $x^*$ is not properly efficient for $(P)$, then, there is an $x \in X$ and an $i \in \bar{p}$ such that

$$
\frac{\int_a^b f^i(t, x(t), \dot{x}(t))dt}{\int_a^b g^i(t, x(t), \dot{x}(t))dt} < \frac{\int_a^b f^i(t, x^*(t), \dot{x}^*(t))dt}{\int_a^b g^i(t, x^*(t), \dot{x}^*(t))dt}
$$

(11)
and
\[
\frac{\int_a^b f^i(t,x^*(t),\dot{x}^*(t))dt}{\int_a^b g^i(t,x(t),\dot{x}(t))dt} - \frac{\int_a^b f^i(t,x(t),\dot{x}(t))dt}{\int_a^b g^i(t,x(t),\dot{x}(t))dt} > M. \tag{12}
\]

for all \( M > 0 \) and for all \( j \) such that
\[
\frac{\int_a^b f^j(t,x(t),\dot{x}(t))dt}{\int_a^b g^j(t,x(t),\dot{x}(t))dt} > \frac{\int_a^b f^j(t,x^*(t),\dot{x}^*(t))dt}{\int_a^b g^j(t,x^*(t),\dot{x}^*(t))dt} \tag{13}
\]
i.e.,
\[
\int_a^b f^i(t,x(t),\dot{x}(t))dt - v_i^* \int_a^b g^i(t,x(t),\dot{x}(t))dt < 0. \tag{11'}
\]

and
\[
-\int_a^b f^j(t,x(t),\dot{x}(t))dt + v_j^* \int_a^b g^j(t,x(t),\dot{x}(t))dt > M \tag{12'}
\]
for all \( j \) such that
\[
\int_a^b f^j(t,x(t),\dot{x}(t))dt - v_j^* \int_a^b g^j(t,x(t),\dot{x}(t))dt > 0. \tag{13'}
\]

So (11'), (12') and (13') imply that \( x^* \) is not properly efficient for \( (P_v) \) with \( v = v^* \). Hence \( x^* \) is properly efficient in \( (P) \).

**REFERENCES**


Dept. of Applied Mathematics
Pukyong National University
Pusan 608-737, Republic of Korea
Boundary Controllability of Delay Integrodifferential Systems in Banach Spaces

K. Balachandran and E. R. Anandhi

Abstract
Sufficient conditions for boundary controllability of time varying delay integrodifferential systems in Banach spaces are established. The results are obtained by using the strongly continuous semigroup theory and the Banach contraction principle.

1. Introduction

Controllability of nonlinear systems represented by ordinary differential equations in Banach spaces has been extensively studied by several authors. Balachandran et al. [1] studied the controllability of nonlinear integrodifferential systems in Banach spaces whereas in [2] they have investigated the local null controllability of nonlinear functional differential systems. Controllability of nonlinear functional integrodifferential systems in Banach spaces has been studied by Park and Han [11]. Kwon et al. [9] discussed the approximate controllability for delay Volterra systems while Balachandran and Sankthivel [3] established a set of sufficient conditions for the controllability of delay integrodifferential systems in Banach spaces.

Several abstract settings have been developed to describe the distributed control systems on a domain $\Omega$ in which the control is acted through the boundary $\Gamma$. But in these approaches one can encounter the difficulty for the existence of sufficiently regular solution to state space system, the control must be taken in a space of sufficiently smooth functions. A semigroup approach to boundary input problems for linear differential equations was first presented by Fattorini [7]. This approach was extended by Balakrishnan [4] where he showed that the solution of a parabolic boundary control equation with $L^2$ controls can be expressed as a mild solution to an operator equation. Barbu and Precupanu [5] studied a class of convex control problems governed by linear evolution systems covering the principal boundary control systems of parabolic type. In [6] Barbu investigated a class of boundary-distributed linear control systems in Banach spaces. Lasiecka [10] established the regularity of optimal boundary controls for

2000 AMS Subject Classification: 93B05
Key Words: Boundary controllability, delay integrodifferential system, semigroup theory, fixed point theorem.
parabolic equations with quadratic cost criterion. Recently Han and Park [8] derived a set of sufficient conditions for the boundary controllability of a semilinear system with a nonlocal condition. The purpose of this paper is to study the boundary controllability of time varying delay integrodifferential systems in Banach spaces by using the Banach fixed point theorem.

2. Preliminaries

Let $E$ and $U$ be a pair of real Banach spaces with $\|\cdot\|$ and $|\cdot|$, respectively. Let $\sigma$ be a linear closed and densely defined operator with $D(\sigma) \subseteq E$ and $R(\sigma) \subseteq E$ and let $\theta$ be a linear operator with $D(\theta) \subseteq E$ and $R(\theta) \subseteq X$, a Banach space.

Consider the boundary control delay system of the form

\[
\begin{align*}
\dot{x}(t) &= \sigma x(t) + f(t, x(\gamma_1(t)), x(\gamma_2(t)), \ldots, x(\gamma_n(t))), \quad t \in J = [0, b], \\
\theta x(t) &= B_1 u(t), \\
x(0) &= x_0,
\end{align*}
\]

where $\gamma_i(t), i = 1, 2, \ldots, n$ are continuous functions, the state $x(.)$ takes values in the Banach space $E$, $B_1 : U \to X$ is a linear continuous operator, the control function $u \in L^1(J, U)$, a Banach space of admissible control functions and the nonlinear operator $f : J \times E^n \to E$ is continuous.

Let $A : E \to E$ be a linear operator defined by

\[
D(A) = \{x \in D(\sigma); \theta x = 0\}, \quad Ax = \sigma x, \quad \text{for} \quad x \in D(A).
\]

Let $B_r = \{y \in E : \|y\| \leq r\}$, for some $r > 0$. We shall make the following hypotheses:

(A$_1$) $D(\sigma) \subset D(\theta)$ and the restriction of $\theta$ to $D(\sigma)$ is continuous relative to graph norm of $D(\sigma)$.

(A$_2$) The operator $A$ is the infinitesimal generator of a $C_0$ semigroup $T(t)$ and there exists a constant $M > 0$ such that $\|T(t)\| \leq M$.

(A$_3$) There exists a linear continuous operator $B : U \to E$ such that $\sigma B \in L(U, E)$, $\theta(Bu) = B_1 u$, for all $u \in U$. Also $Bu(t)$ is continuously differentiable and $\|Bu\| \leq C\|B_1 u\|$ for all $u \in U$, where $C$ is a constant.

(A$_4$) For all $t \in (0, b]$ and $u \in U$, $T(t)Bu \in D(A)$. Moreover, there exists a positive function $\nu \in L^1(0, b)$ such that $\|AT(t)B\| \leq \nu(t)$, a.e. $t \in (0, b)$ and choose a constant $K > 0$ such that $\int_0^b \nu(t)dt \leq K.$
If $x(t)$ is the solution of (1), then we can define a function $z(t) = x(t) - Bu(t)$ and from our assumption we see that $z(t) \in D(A)$. Hence (1) can be written in terms of $A$ and $B$ as

$$
\begin{align*}
\dot{z}(t) &= Az(t) + \sigma Bu(t) + f(t, x(\gamma_1(t)), x(\gamma_2(t)), \ldots, x(\gamma_n(t))), \quad t \in J \\
x(t) &= z(t) + Bu(t), \\
x(0) &= x_0,
\end{align*}
$$

If $u$ is continuously differentiable on $[0, b]$ then $z$ can be defined as a mild solution to the Cauchy problem

$$
\begin{align*}
\dot{z}(t) &= Az(t) + \sigma Bu(t) - B\dot{u}(t) + f(t, x(\gamma_1(t)), x(\gamma_2(t)), \ldots, x(\gamma_n(t))), \\
z(0) &= x_0 - Bu(0),
\end{align*}
$$

and the solution of (1) is given by

$$
\begin{align*}
x(t) &= T(t)[x_0 - Bu(0)] + Bu(t) + \int_0^t T(t - s)[\sigma Bu(s) - B\dot{u}(s)]
+ f(s, x(\gamma_1(s)), x(\gamma_2(s)), \ldots, x(\gamma_n(s)))]ds. \\
& \quad + \int_0^t T(t - s)f(s, x(\gamma_1(s)), x(\gamma_2(s)), \ldots, x(\gamma_n(s)))]ds.
\end{align*}
$$

Since the differentiability of the control $u$ represents an unrealistic and severe requirement, it is necessary to extend the concept of the solution for the general inputs $u \in L^1(J, U)$. Integrating (3) by parts, we get

$$
\begin{align*}
x(t) &= T(t)x_0 + \int_0^t [T(t - s)\sigma - AT(t - s)]Bu(s)ds \\
& \quad + \int_0^t T(t - s)f(s, x(\gamma_1(s)), x(\gamma_2(s)), \ldots, x(\gamma_n(s)))]ds. \\
& \quad + \int_0^t T(t - s)f(s, x(\gamma_1(s)), x(\gamma_2(s)), \ldots, x(\gamma_n(s)))]ds.
\end{align*}
$$

Thus (4) is well defined and it is called a mild solution of the system (1).

**Definition:** The system (1) is said to be controllable on the interval $J$ if for every $x_0, x_1 \in E$, there exists a control $u \in L^2(J, U)$ such that the solution $x(\cdot)$ of (1) satisfies $x(b) = x_1$.

We further assume the following conditions:

(A_5) The linear operator $W$ from $L^2(J, U)$ into $E$ defined by

$$
Wu = \int_0^b [T(b - s)\sigma - AT(b - s)]Bu(s)ds
$$

induces an invertible operator $\tilde{W}$ defined on $L^2(J, U)/\ker W$ and there exists a positive constant $K_1 > 0$ such that $\|\tilde{W}^{-1}\| \leq K_1$. 

(i) $f : J \times E^n \to E$ is continuous and there exist constants $M_1$ and $M_2$ such that for all $v_i, w_i \in B_r, i = 1, 2, \ldots, n$ we have
\[ \| f(t, v_1, v_2, \ldots, v_n) - f(t, w_1, w_2, \ldots, w_n) \| \leq M_1 \sum_{i=1}^{n} \| v_i - w_i \| \]
and
\[ M_2 = \max_{t \in J} \| f(t, 0, \ldots, 0) \|. \]

(ii) There exists a constant $q$ such that for all $x_1, x_2 \in E$
\[ \| x_1(\gamma_i(t)) - x_2(\gamma_i(t)) \| \leq q \| x_1(t) - x_2(t) \|, \quad \text{for } i = 1, 2, \ldots, n. \]

(iii) $M \| x_0 \| + K_1 [bM \| \sigma B \| + K_2] \| x_1 \| + M \| x_0 \| + L \leq r$, where $L = bM(Mnr + M_2)$.

(iv) Let $p = nqbmM_1[1 + (bM \| \sigma B \| + K)K_1]$ be such that $0 < p < 1$.

3. Controllability of Delay System

**Theorem 3.1** If the hypotheses (A1)-(A5) and (i) - (iv) are satisfied, then the boundary control delay system (1) is controllable on $J$.

**Proof:** Let $Y = C(J, B_r)$. Using the hypothesis (A5), for an arbitrary function $x(.)$ define the control
\[ u(t) = \bar{W}^{-1}[x_1 - T(b)x_0 - \int_0^b T(b - s)f(s, x(\gamma_1(s)), x(\gamma_2(s)), \ldots, x(\gamma_n(s)))ds](t). \quad (5) \]
We shall show that, when using this control, the operator $\Psi$ defined on $Y$ by
\[ \Psi x(t) = T(t)x_0 + \int_0^t [T(t - s)\sigma - AT(t - s)]B\bar{W}^{-1}[x_1 - T(b)x_0 \]
\[ - \int_0^b T(b - \tau)f(\tau, x(\gamma_1(\tau)), x(\gamma_2(\tau)), \ldots, x(\gamma_n(\tau)))d\tau]ds \]
\[ + \int_0^t T(t - s)f(s, x(\gamma_1(s)), x(\gamma_2(s)), \ldots, x(\gamma_n(s)))ds \]
has a fixed point. This fixed point is then a solution of (1). Clearly $\Psi x(b) = x_1$, which means that the control $u$ steers the delay system (1) from the initial state $x_0$ to $x_1$ in time $b$ provided we can obtain a fixed point of the operator $\Psi$.

First we show that $\Psi$ maps $Y$ into itself. For $x \in Y$,
\[ \| \Psi x(t) \| \leq \| T(t)x_0 \| + \| \int_0^t [T(t - s)\sigma - AT(t - s)]B\bar{W}^{-1}[x_1 - T(b)x_0 \]

Thus $\Psi$ maps $Y$ into itself. Now, for $x_1, x_2 \in Y$ we have

\[
\left\| \Psi x_1(t) - \Psi x_2(t) \right\| 
\leq \int_0^t \left\| T(t-s) \right\| \| \sigma B \| \| \tilde{W}^{-1} \| \left\| f(\tau, x(\gamma_1(\tau)), x(\gamma_2(\tau)), \ldots, x(\gamma_n(\tau))) \right\| d\tau \left\| \int_0^b \left\| T(b-\tau) \right\| \| f(\tau, x(\gamma_1(\tau)), x(\gamma_2(\tau)), \ldots, x(\gamma_n(\tau))) \right\| - f(\tau, 0, \ldots, 0) \| d\tau \right\| ds 
\leq b[M\|\sigma B\| + K_1bMM_1\left\| x_1(\gamma_1(\tau)) - x_2(\gamma_1(\tau)) \right\| + \|x_1(\gamma_2(\tau)) - x_2(\gamma_2(\tau))\| + \ldots + \|x_1(\gamma_n(\tau)) - x_2(\gamma_n(\tau))\|] 
\leq bMM_1[1 + (bM\|\sigma B\| + K_1b) \sup_{t \in J} \left\| x_1(\gamma_1(t)) - x_2(\gamma_1(t)) \right\|] 
\]


\[ + \| x_1(\gamma_2(t)) - x_2(\gamma_2(t)) \| + \ldots + \| x_1(\gamma_n(t)) - x_2(\gamma_n(t)) \| \]
\[ \leq nqBMM_1[1 + (bM\|B\| + K)K_1]\|x_1(t) - x_2(t)\| \]
\[ \leq p\|x_1(t) - x_2(t)\|. \]

Therefore, \( \Psi \) is a contraction mapping and hence there exists a unique fixed point \( x \in Y \) such that \( \Psi x(t) = x(t) \). Any fixed point of \( \Psi \) is a mild solution of (1) on \( J \) which satisfies \( x(b) = x_1 \). Thus the system (1) is controllable on \( J \).

4. Controllability of Delay Integrodifferential System

Consider the boundary control delay integrodifferential system of the form

\[ \dot{x}(t) = \sigma x(t) + f(t, x(\gamma_1(t)), \int_0^t k(t, s)g(s, x(\gamma_2(s)))ds), \quad t \in J = [0, b], \]
\[ \tau x(t) = B_1 u(t), \]
\[ x(0) = x_0, \quad (6) \]

where the nonlinear operators \( f : J \times E \times E \to E \), \( g : J \times E \to E \) and \( k : J \times J \to R \) are given.

To establish the results we shall assume the following conditions:

(a) \( f : J \times E \times E \to E \) is continuous and there exist constants \( N_1 \) and \( N_2 \) such that for all \( v_1, v_2 \in B_r \) and \( w_1, w_2 \in E \) we have

\[ \| f(t, v_1, w_1) - f(t, v_2, w_2) \| \leq N_1[\|v_1 - v_2\| + \|w_1 - w_2\|] \]

and

\[ N_2 = \max_{t \in J} \| f(t, 0, 0) \|. \]

(b) \( g : J \times E \to E \) is continuous and there exist constants \( L_1 \) and \( L_2 \) such that for all \( v_1, v_2 \in B_r \) we have

\[ \| g(t, v_1) - g(t, v_2) \| \leq L_1\|v_1 - v_2\| \]

and

\[ L_2 = \max_{t \in J} \| g(t, 0) \|. \]

(c) There exists a constant \( L \) such that

\[ \| k(t, s) \| \leq L, \quad \text{for} \quad (t, s) \in J \times J. \]

(d) There exists a constant \( q \) such that for all \( x_1, x_2 \in E \)

\[ \| x_1(\gamma_i(t)) - x_2(\gamma_i(t)) \| \leq q\|x_1(t) - x_2(t)\|, \quad \text{for} \quad i = 1, 2. \]
(e) $M\|x_0\| + K_1[bM\|\sigma B\| + K]\|x_1\| + M\|x_0\| + N] + N \leq r$
where $N = bM[N_1(r + bL(L_1r + L_2)) + N_2]$. 

(f) Let $a = [(bM\|\sigma B\| + K)K_1 + 1]bqMN_1[1 + bLL_1]$ be such that $0 \leq a < 1$.

Using the similar argument as in the previous section we can obtain a mild solution of (6) and it can be written as

$$x(t) = T(t)x_0 + \int_0^t [T(t-s)\sigma - AT(t-s)]Bu(s)ds$$
$$+ \int_0^t T(t-s)f(s, x(\gamma_1(s)), \int_0^s k(s, \tau)g(\tau, x(\gamma_2(\tau)))d\tau)ds.$$

(7)

**Theorem 4.1** If the hypotheses $(A_1)-(A_5)$ and (a) - (f) are satisfied, then the boundary control delay integrodifferential system (6) is controllable on $J$.

**Proof:** Using the hypothesis $(A_5)$, for an arbitrary function $x(.)$ define the control

$$u(t) = W^{-1}[x_1 - T(b)x_0 - \int_0^b T(b-s)f(s, x(\gamma_1(s)), \int_0^s k(s, \tau)g(\tau, x(\gamma_2(\tau)))d\tau)ds](t).$$

We shall show that, when using this control, the operator $\Phi$ defined on $Y$ by

$$\Phi x(t) = T(t)x_0 + \int_0^t [T(t-s)\sigma - AT(t-s)]BW^{-1}[x_1 - T(b)x_0$$

$$+ \int_0^b T(b-\tau)f(\tau, x(\gamma_1(\tau)), \int_0^\tau k(\tau, \eta)g(\eta, x(\gamma_2(\eta)))d\eta)](s)ds$$

$$+ \int_0^t T(t-s)f(s, x(\gamma_1(s)), \int_0^s k(s, \tau)g(\tau, x(\gamma_2(\tau)))d\tau)ds$$

has a fixed point. This fixed point is then a solution of (6). Clearly $\Phi x(b) = x_1$, which means that the control $u$ steers the delay integrodifferential system (6) from the initial state $x_0$ to $x_1$ in time $b$ provided we can obtain a fixed point of the nonlinear operator $\Phi$.

First we show that $\Phi$ maps $Y$ into itself. For $x \in Y$,

$$\|\Phi x(t)\| \leq \|T(t)x_0\| + \int_0^t \|T(t-s)\sigma - AT(t-s)]BW^{-1}[x_1 - T(b)x_0$$

$$- \int_0^b T(b-\tau)f(\tau, x(\gamma_1(\tau)), \int_0^\tau k(\tau, \eta)g(\eta, x(\gamma_2(\eta)))d\eta)d\tau(s)ds$$

$$+ \int_0^t T(t-s)f(s, x(\gamma_1(s)), \int_0^s k(s, \tau)g(\tau, x(\gamma_2(\tau)))d\tau)ds$$

$$\leq \|T(t)x_0\| + \int_0^t \|T(t-s]\|\|\sigma B\|\|W^{-1}\|[\|x_1\| + \|T(b)x_0\|$
Thus \( \Phi \) maps \( Y \) into itself. Now, for \( x_1, x_2 \in Y \) we have

\[
\| \Phi x_1(t) - \Phi x_2(t) \| \\
\leq \int_0^t \| T(t - s) \| \| \sigma B \| + \| AT(t - s)B \| \| \bar{W}^{-1} \| \| \int_0^b \| T(b - \tau) \| \\
\| f(\tau, x_1(\gamma_1(\tau)), \int_0^\tau k(\tau, \eta)g(\eta, x_1(\gamma_2(\eta)))d\eta) \\
- f(\tau, x_2(\gamma_1(\tau)), \int_0^\tau k(\tau, \eta)g(\eta, x_2(\gamma_2(\eta)))d\eta) \| d\tau \| ds \\
+ \int_0^t \| T(t - s) \| \| f(s, x_1(\gamma_1(s)), \int_0^s k(s, \tau)g(\tau, x_1(\gamma_2(\tau)))d\tau) \\
- f(s, x_2(\gamma_1(s)), \int_0^s k(s, \tau)g(\tau, x_2(\gamma_2(\tau)))d\tau) \| ds \\
\leq b[M\| \sigma B \| + K]K_1bMN_1\| x_1(\gamma_1(\tau)) - x_2(\gamma_1(\tau)) \| \\
+ bLL_1\| x_1(\gamma_2(\eta)) - x_2(\gamma_2(\eta)) \| \\
+ bMN_1\| x_1(\gamma_1(s)) - x_2(\gamma_1(s)) \| \\
+ bLL_1\| x_1(\gamma_2(s)) - x_2(\gamma_2(s)) \| \\
\leq [(bM\| \sigma B \| + K)K_1 + 1]bMN_1\| x_1(\gamma_1(t)) - x_2(\gamma_1(t)) \| \\
+ bLL_1\sup_{t \in J}\| x_1(\gamma_2(t)) - x_2(\gamma_2(t)) \| \\
\leq [(bM\| \sigma B \| + K)K_1 + 1]bMN_1[1 + bLL_1]q\| x_1(t) - x_2(t) \| \\
\leq a\| x_1(t) - x_2(t) \| .
\]
Therefore, $\Phi$ is a contraction mapping and hence there exists a unique fixed point $x \in Y$ such that $\Phi x(t) = x(t)$. Any fixed point of $\Phi$ is a mild solution of (6) on $J$ which satisfies $x(b) = x_1$. Thus the system (6) is controllable on $J$.

References


HEDGING OPTION PORTFOLIOS WITH TRANSACTION COSTS AND BANDWIDTH

SEKI KIM

ABSTRACT. Black-Scholes equation arising from option pricing in the presence of cost in trading the underlying asset is derived. The transaction cost is chosen precisely and generalized to reflect the trade in the real world. Furthermore, the concept of the bandwidth is introduced to obtain the better rehedging. The model with bandwidth derived in this paper can be used to calculate the more accurate option price numerically even if it is nonlinear and more complicated than the models shown before.

1. INTRODUCTION

The option pricing model, derived by Black and Scholes [3], assumes perfect markets. This assumption is obviously unrealistic. In recent years, there are several works investigating the effects of incorporating more realistic assumptions about the underlying markets in which trading occurs into the option valuation process. The most important effects are arising from the inclusion of nonzero transaction costs of the underlying assets. This approach was started by Leland [9] and extended by Boyle & Vorst [4], Hoggard, Whalley & Wilmott [8], Avellaneda & Panas [2], Toft [10], Whalley & Wilmott [11], and Henrotte [6]. The first five of these assume hedging takes place at given discrete time intervals and the last two assume flexible but prescribed trading rules. These involve a band around the ideal value of $\Delta$, within which the number of assets actually held in the portfolio is allowed to vary. These can be also expressed by partial differential equations for the value of option, $V$ hereafter, which are similar to the Black-Scholes equation except an extra term representing the effect of the transaction costs.

\[ V_t + \frac{\sigma^2 V^2}{2} V_{SS} - rV + rSV_S = K(\nu, S), \]

where $\nu$ is generally a nonlinear functional of the Gamma, $\Gamma = V_{SS}$, the second derivative of the option with respect to the asset price, and consequently these equations are nonlinear.

Leland [9] included transaction costs proportional to the value of the shares traded into the Black-Scholes model using a fixed revision interval between trades by including the transaction cost term in a modified value for the variance. Boyle and Vorst [4] incorporate proportional transaction costs directly into a binomial framework, and obtain a limiting variance adjustment to Black-Scholes in the case of a long call option.

Key words and phrases. option pricing, transaction costs, bandwidth.

This paper was supported by Faculty Research Fund, SungKyunKwan University, 1999.
that is similar to but different from Leland’s due to the different assumptions made about the distribution of movements in the underlying asset.

As a different direction, the global-in-time models have been introduced by Hodges & Neuberger [7] and Davis, Panas & Zariphopoulou [5]. Because they are using the utility maximization function, the optimality can be achieved [1]. But such models are slow to compute since they usually result in three- or four-dimensional free boundary problems [12].

In this paper, we generalize the Hoggard, Whalley and Wilmott [8] and Henrotte [6] which are concentrated on the analysis of the transaction costs and bandwidth rehedging policy. In section 2, we derive the nonlinear partial differential equation which the option price with more general and precise transaction cost satisfies. In section 3, we introduce the bandwidth for the better rehedging and obtain the partial differential equation under the given bandwidth.

2. A Generalized Transaction Cost

In this section we derive the model with dividends taxed at rate $\tau$ under the specific transaction costs. This model is the generalization of Hoggard, Whalley, and Wilmott model [8]. Before deriving the mathematical model which gives the value of the option with transaction costs, we need to mention the assumptions.

- The portfolio is revised every $\delta t$ where $\delta t$ is a finite, fixed and small interval.

$$\delta S = \mu S \delta t + \sigma S \phi \delta t^{1/2}$$

where $\phi$ is drawn from a standardized normal distribution and $\mu$ is the stock price’s instantaneous expected return and $\sigma$ is the instantaneous variance of stock price’s return.

- Short selling is allowed and the assets are divisible.

- The risk-free interest rate $r$ and the asset volatility $\sigma$ are known, deterministic functions of time over the life of the option as constant.

- No arbitrage opportunities (The hedged portfolio has an expected return equal to that from a bank deposit).

- The constant dividend yield is $\eta$ and this dividend is taxed at rate $\tau$.

We proceed as in the Black-Scholes model and consider a portfolio $\Pi$ consisting of derivative products, whose total value is $V$, and a number $\Delta$, which will be determined later, of shares.

$$\Pi = V - \Delta S.$$  

Using Itô’s lemma, we obtain the change in value of this portfolio, $\delta \Pi$, by expanding $V$ about $(S, t)$ with a transaction costs $K(\nu, S)$ where $\nu$ is the number of shares traded, as

$$\delta \Pi = \left( \mu S \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \phi^2 \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t} - \mu \Delta S \right) \delta t$$

$$+ \sigma S \left( \frac{\partial V}{\partial S} - \Delta \right) \phi \delta t^{1/2} - K(\nu, S) - \eta(1 - \tau) \Delta S \delta t$$
Eliminating the risk associated with the stochastic movements or adopting a strategy of delta hedging, we choose the number of shares as
\[ \Delta = \frac{\partial V}{\partial S}. \]
Hence we hold the number of assets as
\[ -\frac{\partial V}{\partial S}(S, t). \]
After a time step \( \delta t \) and rehedging, the number of assets we hold becomes
\[ -\frac{\partial V}{\partial S}(S + \delta S, t + \delta t). \]
Hence we can find the number of assets we trade, \( \nu \), which is given by
\[ \nu = -\frac{\partial V}{\partial S}(S + \delta S, t + \delta t) + \frac{\partial V}{\partial S}(S, t). \]
Using Itô’s lemma, this number can be obtained as
\[ \nu = \sigma S \frac{\partial^2 V}{\partial S^2} \phi \delta t^{1/2}. \]
We do not know beforehand how many shares will be traded, but we can calculate the expected number.

The transaction costs \( K(\nu, S) \) is expressed as the sum of three terms, a fixed cost, a cost proportional to volume traded and a cost proportional to the value traded.

\[ K(\nu, S) = k_1 + k_2 |\nu| + \left( \sum_{i=1}^{n} (\zeta_i - \zeta_{i-1}) U(|\nu|S - x_i) \right) |\nu|S \]

where \( k_i \) and \( \zeta_i \) are constant, and \( x_i \) represents the level of the amount \( |\nu|S \) which has the proportional constant \( \zeta_i \) and \( U \) is the Heaviside function as
\[ U(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \]

**Theorem 1.** The value of the option with the transaction cost function as

\[ K(\nu, S) = k_1 + k_2 |\nu| + \left( \sum_{i=1}^{n} (\zeta_i - \zeta_{i-1}) U(|\nu|S - x_i) \right) |\nu|S \]

satisfies the following nonlinear partial differential equation:

\[ \frac{\partial V}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r - \eta)(1 - \tau)S \frac{\partial V}{\partial S} - r(1 - \tau)V \]
\[ = \frac{k_1}{\delta t} + \sqrt{\frac{2}{\pi \delta t}} k_2 \sigma S \left| \frac{\partial^2 V}{\partial S^2} \right| \]
\[ + \sqrt{\frac{2}{\pi \delta t}} \sigma S^2 \left| \frac{\partial^2 V}{\partial S^2} \right| \sum_{i=1}^{n} (\zeta_i - \zeta_{i-1}) \exp \left[ -\frac{x_i^2}{2 \sigma^2 S^4 \left( \frac{\partial^2 V}{\partial S^2} \right)^2 \delta t} \right]. \]
\textbf{Proof.} From the assumption on no arbitrage opportunities,
\begin{equation}
E[\delta \Pi] = r(1 - \tau)\Pi \delta t
= r(1 - \tau)(V - S \frac{\partial V}{\partial S}) \delta t.
\end{equation}

Equation (2) becomes
\begin{equation*}
E[\delta \Pi] = \left( \frac{\partial V}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \right) \delta t - E[K(\nu, S)] - \eta(1 - \tau)S \frac{\partial V}{\partial S} \delta t.
\end{equation*}

Here
\begin{align*}
E[K(\nu, S)] &= E \left[ k_1 + k_2 |\nu| + \left( \sum_{i=1}^{n} (\zeta_i - \zeta_{i-1}) U(|\nu|S - x_i) \right) |\nu|S \right] \\
&= E[k_1] + k_2 E[|\nu|] + S \sum_{i=1}^{n} (\zeta_i - \zeta_{i-1}) E[|\nu|U(|\nu|S - x_i)].
\end{align*}

We know the fact that
\begin{align}
E[|\nu|] &= E \left[ \sigma S \left| \frac{\partial V}{\partial S} \right| |\phi| \delta t^{1/2} \right] \\
&= \sigma S \left| \frac{\partial V}{\partial S} \right| \delta t^{1/2} E[|\phi|] \\
&= \sqrt{\frac{2}{\pi}} \sigma S \left| \frac{\partial V}{\partial S} \right| \delta t^{1/2}.
\end{align}

Similarly,
\begin{align*}
E[|\phi|U(|\nu|S - x_i)] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |y|U(R|y| - x_i) e^{-\frac{1}{2}y^2} dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x_i/R} |y|e^{-\frac{1}{2}y^2} dy + \frac{1}{\sqrt{2\pi}} \int_{x_i/R}^{\infty} |y|e^{-\frac{1}{2}y^2} dy \\
&= \sqrt{\frac{2}{\pi}} \int_{x_i/R}^{\infty} |y|e^{-\frac{1}{2}y^2} dy \\
&= \sqrt{\frac{2}{\pi}} \exp \left[ -\frac{1}{2} \frac{x_i^2}{R^2} \right].
\end{align*}

Hence
\begin{equation}
E[|\nu|SU(|\nu|S - x_i)] = \sqrt{\frac{2}{\pi}} \sigma S \left| \frac{\partial V}{\partial S} \right| \delta t^{1/2} \exp \left[ -\frac{x_i^2}{2\sigma^2 S^4 \left( \frac{\partial V}{\partial S} \right)^2} \right].
\end{equation}
From (7) and (8), the expected value of the transaction cost is given by
\[
E[K(\nu, S)] = k_1 + k_2 \sqrt{\frac{2}{\pi} \sigma S} \left| \frac{\partial^2 V}{\partial S^2} \right| \delta t^{1/2} + \sum_{i=1}^{n} (\zeta_i - \zeta_{i-1}) \sqrt{\frac{2}{\pi} \sigma S} \left| \frac{\partial^2 V}{\partial S^2} \right| \delta t^{1/2} \exp \left[ - \frac{x_i^2}{2 \sigma^2 S^4 \left( \frac{\partial^2 V}{\partial S^2} \right)^2} \right].
\]

Finally we have
\[
(9) \quad E[\delta \Pi] = \left( \frac{\partial V}{\partial t} + \frac{1}{2} \sigma^2 S \frac{\partial^2 V}{\partial S^2} - \eta(1 - \tau) S \frac{\partial V}{\partial S} \right) \delta t + k_1 \sqrt{\frac{2}{\pi} \sigma S} \left| \frac{\partial^2 V}{\partial S^2} \right| \delta t^{1/2} + \sqrt{\frac{2}{\pi} \sigma^2 S} \left| \frac{\partial^2 V}{\partial S^2} \right| \delta t^{1/2} \sum_{i=1}^{n} (\zeta_i - \zeta_{i-1}) \exp \left[ - \frac{x_i^2}{2 \sigma^2 S^4 \left( \frac{\partial^2 V}{\partial S^2} \right)^2} \right].
\]

The equations (6) and (9) complete the proof. □

It is difficult to obtain the approximate solution of the nonlinear partial differential equation in Theorem 1 because of its complicated nonlinearity although we generalize the transaction costs here. But the procedure of this section will be used to derive the partial differential equation with bandwidth in the next section, from which we have the more accurate numerical solution easier than from that of Theorem 1. If \( n = 1 \) and \( x_i = 0 \), then we have the simple partial differential equation
\[
\frac{\partial V}{\delta t} + \frac{1}{2} \sigma^2 S \frac{\partial^2 V}{\partial S^2} + r S \frac{\partial V}{\partial S} - r V = \frac{k_1}{\delta t} + \sqrt{\frac{2}{\pi} \sigma \delta t} \sigma S (k_2 + \zeta_1 S) \left| \frac{\partial^2 V}{\partial S^2} \right|
\]
which is similar to the model derived in [8] with no dividends and the simple transaction cost
\[
K(\nu, S) = k_1 + k_2 |\nu| + |\zeta_1| |\nu| S.
\]

3. HEDGING TO THE BANDWIDTH

In the previous section we derive the model of option price with realistic transaction costs when hedging takes place at fixed intervals of time. But the better strategy is to rehedge whenever the position becomes too far out of line with the perfect hedge position. Hedging has to take place discretely while prices of options are monitored continuously.

The perfect Black-Scholes hedge is
\[
\Delta = \frac{\partial V}{\partial S}.
\]
Because we can not perfectly hedge the portfolio, we will try to reduce the risk from mishedging. Suppose that we hold $-\Xi$ of the underlying asset but we do not want to accept the extra cost of trading to reposition the hedge. Then the portfolio is given by

$$\Pi = V - \Xi S.$$  

Using Itô’s lemma, we can write the change in value of this portfolio from $(S + \delta S, t + \delta t), \delta \Pi$, as

$$\delta \Pi = \left( \mu S \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \phi^2 \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t} - \mu \Xi S \right) \delta t$$

$$+ \sigma S \left( \frac{\partial V}{\partial S} - \Xi \right) \phi \delta t^{1/2} - \eta(1 - \tau) \Xi S \delta t.$$  

The risk, as measured by the variance over a timestep $\delta t$ of the imperfectly hedged position is, to leading order,

$$\sigma^2 S^2 \left( \frac{\partial V}{\partial S} - \Xi \right)^2 \delta t$$

since

$$E[\delta \Pi] = \left( \frac{1}{2} \sigma^2 S^2 \phi^2 \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t} + \left( \frac{\partial V}{\partial S} - \Xi \right) \mu S - \eta(1 - \tau) \Xi S \right) \delta t$$

and

$$E[\delta \Pi^2] = \sigma^2 S^2 \left( \frac{\partial V}{\partial S} - \Xi \right)^2 \delta t + O(\delta t^2).$$

When $\Xi = \partial V / \partial S$, the above variance (11) becomes zero. It is a natural hedging strategy that bounding the variance within a given tolerance and it can be expressed by

$$\sigma S \left| \frac{\partial V}{\partial S} - \Xi \right| \leq \Lambda$$

where $\Lambda$ is a measure of the maximum expected risk in the portfolio. Whenever (12) is violated (i.e. the perfect hedge $\partial V / \partial S$ and the current hedge $\Xi$ go out of line), the position must be rebalanced. Consequently this inequality (12) defines the bandwidth of the hedging position.

We now specify the maximum risk as below,

$$\Lambda = S^2 \left( \frac{\partial V}{\partial S} - \Xi \right)^2.$$  

This gives us the number we trade when the bandwidth is violated,

$$|\nu| = \left| \frac{\partial V}{\partial S} - \Xi \right| = \frac{\Lambda^{1/2}}{S}.$$
**Theorem 2.** The option’s price $V$ with the transaction cost (4) and the bandwidth (13) satisfies the nonlinear partial differential equation

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r - \eta)(1 - \tau)S \frac{\partial V}{\partial S} - r(1 - \tau)V$$

where $\Gamma$ is the option’s gamma $\frac{\partial^2 V}{\partial S^2}$.

**Proof.** From (3), we have

$$\nu^2 = \sigma^2 S^2 \phi^2 \Gamma^2 \delta t.$$

Using (14),

$$\nu^2 = \frac{\Lambda}{S^2}.$$

These two equations give us

$$\delta t = \frac{\Lambda}{\sigma^2 S^4 \phi^2 \Gamma^2};$$

hence we have the expected value of $1/\delta t$ as

$$E\left[\frac{1}{\delta t}\right] = \frac{\sigma^2 S^4 \Gamma^2}{\Lambda}.$$

Following the procedure of the proof in Theorem 1,

$$E\left[\frac{K(\nu, S)}{\delta t}\right] = E\left[\frac{k_1 + k_2 \nu + (\sum_{i=1}^{n}(\zeta_i - \zeta_{i-1})U(\nu | S - x_i)) |\nu| S}{\delta t}\right]$$

$$= \frac{\sigma^2 S^4 \Gamma^2}{\Lambda} \left( k_1 + \left( k_2 + S \sum_{i=1}^{n}(\zeta_i - \zeta_{i-1})U(\Lambda^{1/2} - x_i) \right) \frac{\Lambda^{1/2}}{S} \right).$$

$\square$

**References**


Department of Mathematics
SungKyunKwan University
email: skim@yurim.skku.ac.kr
ON THE CONVERGENCE OF QUADRATURE RULE FOR SINGULAR INTEGRAL EQUATIONS

SEKI KIM

Abstract. A quadrature rule for the solution of Cauchy singular integral equation is constructed and investigated. This method to calculate numerically singular integrals uses classical Jacobi quadratures adopting Hunter’s method. The proposed method is convergent under a reasonable assumption on the smoothness of the solution.

1. Introduction

The singular integral equation with Cauchy kernel most often considered has the form

$$a(x) \phi(x) + \frac{b(x)}{\pi} \int_{-1}^{1} \frac{\phi(t) \, dt}{t - x} + \int_{-1}^{1} k(x, t) \phi(t) \, dt = f(x) \quad -1 < x < 1.$$  

The first integral term is understood to be a Cauchy principal value integral. It is possible to reduce singular integral equations (hereafter SIE’s) to Fredholm integral equations (indirect method), but direct solution methods are preferred in practice. Also it is proved that when the Gauss numerical integration rule is used that both numerical methods are equivalent in the sense that they provide the same numerical results for the same number of abscissae used in numerical integrations [15].

Usually the unknown function is replaced by the product of a smooth function times a function taken as the weight of the quadrature. For variable coefficients SIE’s, this is non classical and the nodes and weights of the quadrature rule must be constructed from scratch. But for constant coefficients SIE’s this reduces to Jacobi quadrature. In this paper, we want to analyze the replacement of the possibly nonclassical weights and nodes, by the weights and zeros of Jacobi polynomials. This is a quite simpler approach than methods using nonclassical weights and nodes.

We mention some methods for the variable coefficient SIE’s. Theocaris and Tsamasphyros [14] attempt to apply a Gauss-Jacobi quadrature rule directly, but this results in the need to compute the zeros of a second kind of Jacobi function. Dow and Elliott [4] have developed an algorithm with error analysis, for solving an approximate solution to (1) by replacing $f$ and $k$ by polynomial approximations. In [12] the solvability of the discrete system is proved for arbitrary selection of quadrature and collocation nodes, but no error analysis is given there. Here we propose a simpler method and consider nodes which are well known zeros of Jacobi polynomial and want to develop the error analysis for the proposed method. This study concerns only global polynomial approximation.

In the error analysis, a restrictive assumption is used, which is a bound on the size of the coefficients. Since we are unable to find a closed form inverse from the matrix of
the discretized system, we perform the error analysis by treating the singular operator as a perturbation of the identity.

2. Preliminaries

The second kind singular integral equation with variable coefficients can be written as

\[ a(x) \phi(x) + \frac{1}{\pi} \int_{-1}^{1} \frac{K(x,t)}{t-x} \phi(t) dt = f(x) \quad -1 < x < 1 \]

This equation is reduced to (1) by setting \( K(x,t) = (K(x,t) - K(x,x)) + K(x,x) \) and

\[ k(x,t) = \begin{cases} \frac{(K(x,t) - K(x,x))/(t-x)}{t \neq x} \\ K'(x,t) \\ t = x \end{cases} \]

The singular integral in (1) is interpreted in the Cauchy principal value sense. And the equation

\[ a(x) \phi(x) + \frac{b(x)}{\pi} \int_{-1}^{1} \frac{\phi(t)}{t-x} dt = f(x) \quad -1 < x < 1 \]

is called the dominant equation of the equation (1) [10]. We have solutions for SIE’s under the following assumptions in general.

- The functions \( a, b, f \) and \( k \) are Hölder continuous in each independent variable on \([-1,1]\).
- The functions \( S(x) = a(x) + b(x) \) and \( D(x) = a(x) - b(x) \) do not vanish anywhere on \([-1,1]\).

Also it is not restrictive to assume the coefficients to satisfy \( r(x)^2 = a(x)^2 + b(x)^2 = 1 \) and \( b(x) \neq 0 \) on \((-1,1)\). For the latter case, \( b(x) \) may vanish at a finite number of isolated points in \((-1,1)\) as long as it remains of one sign. However, we will assume \( b(x) \) not vanishing in \((-1,1)\) for simplicity. Following [4], let us define the continuous function

\[
\theta(t) = \frac{1}{2\pi i} \ln \frac{a(t) - ib(t)}{a(t) + ib(t)} = \frac{1}{\pi} \arctan \frac{b(t)}{a(t)} + N(t)
\]

where \( N \) takes only integer values and may have discontinuities at the zeros of \( a/b \) and \(-\pi/2 < \arctan x < \pi/2\). In order to apply Elliott-Parget quadrature [2, 3], we assume \( a(x) \) doesn’t have zeros in \((-1,1)\), or we can choose Hunter’s quadrature[8, 1]. The fundamental function \( Z \) is defined as

\[
Z(t) = (1+t)^{n_1}(1-t)^{n_2} \exp \left( -\int_{-1}^{1} \frac{\theta(t)}{t-\tau} d\tau \right) \quad \text{for } t \in (-1,1)
\]

where \( n_1 \) and \( n_2 \) are integers. This function can be rewritten, after fixing \( n_1 \) and \( n_2 \), as

\[
Z(t) = (1+t)^{n_1-\theta(1)}(1-t)^{n_2+\theta(-1)} \Omega(t)
\]
say, where
\[ \Omega(t) = \exp\left\{ (\theta(1) - \theta(t)) \ln(1-t) + (\theta(t) - \theta(-1)) \ln(1+t) \right\} - \int_{-1}^{1} \frac{\theta(\tau) - \theta(t)}{\tau - t} d\tau. \]

Here define \( \alpha = n_2 - \theta(1) \) and \( \beta = n_1 + \theta(-1) \). The behavior of \( Z \) near the end points 
\( -1 \) and \( 1 \) can be considered as following:
\( Z \) is bounded and \( 1/Z \) is infinite, but integrable if \( 0 < \alpha, \beta < 1 \)
\( Z \) is infinite, but integrable and \( 1/Z \) is bounded if \( -1 < \alpha, \beta < 0 \)
In all cases we define the index \( \kappa \) of the singular operator of the dominant equation by
\[ \kappa = -(n_1 + n_2). \]

It turns out that \( \kappa \) can have up to three values depending upon whether \( Z \) is chosen
to be bounded or unbounded at non-special ends i.e. those for which \( \theta(1) \) (or \( \theta(-1) \)) is
not an integer. The largest of these values of \( \kappa \) is the index of singular operator \([5]\).
It is shown that this index can attain only three values \(-1, 0 \) and \( 1 \) if \( b(x) \neq 0 \) in \([-1, 1] \).
If \( \kappa = 1 \), we need an extra condition, to get a unique solution,
\[ \frac{1}{\pi} \int_{-1}^{1} \phi(t) dt = C_0. \]
When \( \kappa = -1 \), the consistency condition for the existence of solution of (2) is
\[ \int_{-1}^{1} \frac{f(x)}{Z(x)} dx = 0. \]

We use \( \rho(x) = (1 - x)^\alpha (1 + x)^\beta \) as a Jacobi weight function. Then \( Z(x) = \rho(x) \Omega(x) \)
where \( \Omega(x) \) is a positive continuous function on \([-1, 1] \). \( \phi(x) \) can be rewritten in terms
of \( \rho(x) \), which expresses explicitly the singular behavior at the end points, and a new
unknown function \( y^s(x) \) such that
\[ \phi(x) = Z(x) \varphi(x) = \rho(x) \Omega(x) \varphi(x) = \rho(x) y^s(x). \]

We consider here the case \( \kappa = 1 \) i.e. \( -1 < \alpha, \beta < 0 \). Let \( H_\mu[-1,1] \) denote the class
of Hölder continuous function of order \( \mu \) on \([-1, 1] \). Then clearly \( \rho(x) \in H_\mu[-1,1] \)
with \( \mu = \min(\alpha, -\beta) \) [10]. The integral (2) can be discretized by a classical Gaussian
quadrature. Let \( t_i \) be the zeros of \( P^{(\alpha, \beta)}_n(x) \) and \( s_j \) be the zeros of \( P^{(-\alpha, -\beta)}_{n-1}(x) \) where
\( P^{(\alpha, \beta)}_n(x) \) denotes the Jacobi polynomial of degree \( n \) relative to the weight function
\( \rho(x) \), and \( P^{(-\alpha, -\beta)}_{n-1}(x) \) the one relative to \( 1/\rho(x) \). Since \( b(x) \neq 0 \), we can rewrite the
equation (2) as
\[ \pi \frac{a(x)}{b(x)} \rho(x) y^s(x) + \int_{-1}^{1} \frac{\rho(t) y^s(t)}{t - x} dt = \frac{\pi f(x)}{b(x)}. \]
If we use \( a_s(x) \) for \( \pi a(x) / b(x) \) and \( f_s(x) \) for \( \pi f(x) / b(x) \), then (4) becomes
\[ a_s(x) \rho(x) y^s(x) + \int_{-1}^{1} \frac{\rho(t) y^s(t)}{t - x} dt = f_s(x). \]
To evaluate the singular integral in (5), we will use two kinds of Gauss-Jacobi quadrature: Hunter’s and Elliott-Paget’s, where both methods use Lagrange interpolations to derive quadrature rules. Let

\[ \psi_n^{(\alpha,\beta)}(z) = \int_{-1}^{1} \frac{\rho(t) P_n^{(\alpha,\beta)}(t)}{t - z} dt = 2(z - 1)^\alpha (z + 1)^\beta q_n^{(\alpha,\beta)}(z) \text{ for } z \notin [-1, 1] \]

where \( q_n^{(\alpha,\beta)} \) represents the so called Jacobi function of the second kind. We can define the values of the function \( \psi_n^{(\alpha,\beta)}(x) \) on the interval \([-1, 1]\) as follows

\[ \psi_n^{(\alpha,\beta)}(x) = \frac{1}{2} \left\{ \psi_n^{(\alpha,\beta)}(x + i0) + \psi_n^{(\alpha,\beta)}(x - i0) \right\} . \]

It can be expressed explicitly by means of the hypergeometric function [13]. Then Hunter’s method has the form

\[ Q_n(y^*, x) = \sum_{i=1}^{n} \frac{u_i y^*(t_i)}{t_i - x} + \frac{\psi_n^{(\alpha,\beta)}(x) y^*(x)}{P_n^{(\alpha,\beta)}(x)} \]

where

\[ \int_{-1}^{1} \frac{\rho(t) y^*(t)}{t - x} dt = Q_n(y^*, x) + \epsilon_{G_h} \text{ for } x \in [-1, 1]. \]

The Elliott-Paget method for singular integral is of the form

\[ Q_n(y^*, x) = \sum_{i=1}^{n} \left[ w_i - \frac{\psi_n^{(\alpha,\beta)}(x)}{P_n^{(\alpha,\beta)}(t_i)} \right] \frac{y^*(t_i)}{t_i - x} \]

where

\[ \int_{-1}^{1} \frac{\rho(t) y^*(t)}{t - x} dt = Q_n(y^*, x) + \epsilon_{G_e} \text{ for } x \in [-1, 1]. \]

Let us remark that through the paper, \( C_i, i = 0, 1, 2, \ldots, 9 \) represent different positive constants.

### 3. Numerical Scheme

In this section, by applying Hunter’s method to (5) and (3) and collocating at the node points \( s_j \), we have

\[ a_s(s_j) \rho(s_j) + \frac{\psi_n^{(\alpha,\beta)}(s_j)}{P_n^{(\alpha,\beta)}(s_j)} y^*(s_j) + \sum_{i=1}^{n} \frac{u_i y^*(t_i)}{t_i - s_j} + \epsilon_{G_1} = f_s(s_j) \]

\[ j = 1, \ldots, n - 1 \]

and

\[ \sum_{i=1}^{n} u_i y^*(t_i) = C_o \]
where (i) $\epsilon_{G_1}$ is the error of Gauss-Jacobi quadrature on $s_j$ and

\[(8) \quad w_j = \int_{-1}^{1} \frac{\rho(t)}{(t - t_j)} \frac{P_n^{(\alpha, \beta)}(t)}{P_n^{(\alpha, \beta)'(t_j)}} dt \]

are the weights of Gauss-Jacobi quadrature. (7) is from the normalization condition (3).

Recall that this equation is of index 1. We transform (5) into the following equation

\[a_s(x)\rho(x)y^s(x) + \int_{-1}^{1} \frac{1}{\rho(t)} \frac{y^s(t)p^2(t)}{t-x} dt = f_s(x)\]

and then proceed similarly, using as weight $1/\rho(t)$ and collocating at $t_i$,

\[(9) \quad \left[ a_s(t_i)\rho(t_i) + \frac{\psi_n^{(-\alpha, -\beta)}(t_i)}{P_n^{(-\alpha, -\beta)'(t_i)}} \right] y^s(t_i) + \sum_{j=1}^{n-1} \frac{w_j}{s_j - t_i} \frac{y^s(s_j)}{p^2(s_j)} + \epsilon_{G_2} = f_s(t_i) \quad i = 1, 2, \ldots, n\]

where (i) $\epsilon_{G_2}$ is the error of Gauss-Jacobi quadrature on $t_i$ and

\[(ii) \quad w_j^s = \int_{-1}^{1} \frac{1}{\rho(t)} \frac{P_n^{(-\alpha, -\beta)}(t)}{(t - s_j)P_n^{(-\alpha, -\beta)'(s_j)}} dt \quad j = 1, \ldots, n - 1\]

are the weights of Gauss-Jacobi quadrature.

In (9), we applied a different weight function $1/\rho(x)$ to the exactly same equation (5) and we are looking for the solution in $(-1, 1)$ because we already knew the end point behavior of the solution. Note that $\rho(x)$ is smooth on $(-1, 1)$ i.e. $y^s(x)p^2(x)$ can be assumed to be smooth on $(-1, 1)$ if $y^s(x)$ is smooth.

After dropping the error terms from (6) and (9), we add nonzero constants $l$ (see the matrix $D_1$ and the bottom of the proof of Theorem 5) and $-h_0$ (see the last column of $G$) to both sides of (7) and (9) respectively. Then we obtain a square system for the unknown vector

\[\underline{y} = (y(s_1), \ldots, y(s_{n-1}), 1, y(t_1), \ldots, y(t_n))\]

approximating the exact solution

\[\underline{y}^s = (y^s(s_1), \ldots, y^s(s_{n-1}), 1, y^s(t_1), \ldots, y^s(t_n)).\]

The system can be written as

\[(10) \quad M\underline{y} = \begin{bmatrix} D_1 & A \\ G & D_2 \end{bmatrix} \underline{y} = \underline{f}\]

where $\underline{f} = (f_s(s_1), \ldots, f_s(s_{n-1}), C_0 + l, f_s(t_1) - h_0, \ldots, f_s(t_n) - h_0)$, and $D_1$ is a diagonal matrix with

\[
\begin{cases}
(D_1)_{jj} = a_s(s_j)\rho(s_j) + \frac{\psi_n^{(\alpha, \beta)}(s_j)}{P_n^{(\alpha, \beta)'(s_j)}} \\
(D_1)_{nn} = l 
\end{cases}
\quad \text{for } j = 1, \ldots, n - 1.
We will use $1/w_n$ for $l$ and $1/|b(1)|$ for $h_o$ later (see [6]).

$$D_2 = \text{diag} \left( a_s(t_i)\rho(t_i) + \frac{\psi_{n-1}^{(-a,-\beta)}(t_i)}{p_{n-1}^{(-a,-\beta)}(t_i)}\rho^2(t_i) \right) \quad i = 1, \ldots, n$$

We assume $(D_1)_{ii}$ and $(D_2)_{ii}$ are not zero for all $i$.

$$A_{ij} = \begin{cases} 
\frac{w_i}{t_j - s_i} & i \neq n \\
\frac{w_i}{w_j} & i = n 
\end{cases} \quad j = 1, \ldots, n$$

$$G_{ij} = \begin{cases} 
\frac{w_i^*}{s_j - t_i}\rho^2(s_j) & j \neq n \\
-h_o & j = n 
\end{cases} \quad i = 1, 2, \ldots, n$$

Note that the system is of order $2n$. Also we remark that $G$ can be written as the product of two matrices $B$ and $D_3$.

$$G = BD_3$$

where

$$B_{ij} = \begin{cases} 
\frac{w_i^*}{s_j - t_i} & j \neq n \\
-h_o & j = n 
\end{cases} \quad i = 1, 2, \ldots, n$$

and $D_3$ is the diagonal matrix with

$$\begin{cases} 
(D_3)_{jj} = \rho^2(s_j) & j = 1, 2, \ldots, n - 1 \\
(D_3)_{nn} = 1 & j = n 
\end{cases}$$

Let us introduce the matrices:

$$N_1 = \begin{pmatrix} w_1 & \cdots & w_{n-1} \\
\frac{1}{t_1 - s_1} & \cdots & \frac{1}{t_{n-1} - s_1} \\
\frac{1}{t_1 - s_{n-1}} & \frac{1}{t_2 - s_{n-1}} & \cdots & \frac{1}{t_{n-1} - s_{n-1}} \\
1 & \frac{1}{t_1 - s_1} & \cdots & \frac{1}{t_{n-2} - s_1} \\
1 & 1 & \cdots & 1 
\end{pmatrix}, \quad N_2 = \begin{pmatrix} w_1^* & \cdots & w_{n-1}^* \\
1 & \frac{1}{|b(1)|} \end{pmatrix}$$

$$S = \begin{pmatrix} w_1 & \cdots & w_{n-1} \\
\frac{1}{t_1 - s_1} & \cdots & \frac{1}{t_{n-1} - s_1} \\
\frac{1}{t_1 - s_{n-1}} & \frac{1}{t_2 - s_{n-1}} & \cdots & \frac{1}{t_{n-2} - s_1} \\
1 & \frac{1}{t_1 - s_1} & \cdots & \frac{1}{t_{n-3} - s_1} \\
1 & 1 & \cdots & 1 
\end{pmatrix}$$

With these matrices, the matrices $A$ and $B$ can be rewritten by

$$A = SN_1, \quad B = -S^t N_2$$

In this notation, the marix $M$ which is the main matrix in discretized system is expressed in form of two matrices’ multiplication.

$$M = \begin{pmatrix} D_1 & S \cdot N_1 \\
-S^t \cdot N_2 \cdot D_3 & D_2 
\end{pmatrix} = \begin{pmatrix} D_1 D_3^{-1} N_2^{-1} & S \\
-S^t & D_2 N_1^{-1} \end{pmatrix} \begin{pmatrix} N_2 D_3 & 0 \\
0 & N_1 \end{pmatrix}. $$
Here we use notations $P_1$ and $P_2$ for $D_1D_3^{-1}N_2^{-1}$ and $D_2N_1^{-1}$ respectively which are diagonal matrices. Then

$$M = RT$$

where

$$T = \begin{pmatrix} N_2D_3 & 0 \\ 0 & N_1 \end{pmatrix}, \quad R = \begin{pmatrix} P_1 & S \\ -S^t & P_2 \end{pmatrix}$$

**Lemma 1.** (i)\[ \frac{\psi_n^{(\alpha,\beta)}(s_j)}{P_n^{(\alpha,\beta)}(s_j)} \sim \rho(s_j) \]

(ii) \[ \frac{\psi_{n-1}^{(-\alpha,-\beta)}(t_i)}{P_{n-1}^{(-\alpha,-\beta)}(t_i)} \rho(t_i) \sim O(1) \]

**Proof.** From [13], we have

$$\{(\frac{d}{dx})^k p_n^{(\alpha,\beta)}(x)\}_{x=\cos \theta} = \begin{cases} \theta^{-\alpha-k-1/2}O(n^{k+1/2}) & c/n \leq \theta \leq \pi/2 \\ O(n^{2k+\alpha}) & 0 \leq \theta \leq c/n. \end{cases}$$

For positive $s_j$, we have the following estimate:

$$\frac{\psi_n^{(\alpha,\beta)}(s_j)}{P_n^{(\alpha,\beta)}(s_j)} \sim \frac{\theta^{-\alpha-k-1/2}n^{-\alpha}}{\theta^{-\alpha-k-1/2}n^{\alpha}} \sim \left( \frac{1}{n} \right)^{2\alpha} \sim \rho(s_j).$$

If $s_j$ is negative, we use $\beta$ instead of $\alpha$. This still gives us the same estimate $\rho(s_j)$. Similarly,

$$\frac{\psi_{n-1}^{(-\alpha,-\beta)}(t_i)}{P_{n-1}^{(-\alpha,-\beta)}(t_i)} \rho(t_i) \sim \left( \frac{1}{n} \right)^{2\alpha} \sim O(1).$$

In matrices $D_1$ and $D_2$, if the value of $\alpha+\beta$ is large enough that it exceeds the value of $\psi_n/P_n$, then we may have positive values in $P_1$ and $P_2$ since $D_3$, $N_1$ and $N_2$ are positive diagonal matrices. Also, since $\alpha_+(x)$ doesn’t have any zero in $(-1, 1)$, we make $\alpha_+(x)$ have positive values without loss of generality.

Hence we may assume $P_1$ and $P_2$ are positive diagonal matrices.

**Theorem 2.** The system (10) is nonsingular, i.e., the matrix $M$ is nonsingular.

**Proof.** It suffices to show $R$ is invertible since $M = RT$ and $T$ is a diagonal matrix with positive entries. $R$ can be decomposed as below.

$$R = \begin{pmatrix} P_1 & S \\ -S^t & P_2 \end{pmatrix}$$

$$= \begin{pmatrix} P_1^\frac{1}{2} & 0 \\ 0 & P_2^\frac{1}{2} \end{pmatrix} \begin{pmatrix} I_n & P_1^{-\frac{1}{2}}SP_2^{-\frac{1}{2}} \\ -P_2^{-\frac{1}{2}}S^tP_1^{-\frac{1}{2}} & I_n \end{pmatrix} \begin{pmatrix} P_1^\frac{1}{2} & 0 \\ 0 & P_2^\frac{1}{2} \end{pmatrix}.$$
Here we observe that the middle matrix $Q$ in the above expression can be written $Q = I + K$ where $K$ is skew-symmetric.

\[
Q = \begin{pmatrix}
    I_n & P_1^{-\frac{1}{2}}SP_2^{-\frac{1}{2}} \\
    -P_2^{-\frac{1}{2}}SP_1^{-\frac{1}{2}} & I_n
\end{pmatrix}
\]

\[
= I_{2n} + K
\]

where

\[
K = \begin{pmatrix}
    0 & P_1^{-\frac{1}{2}}SP_2^{-\frac{1}{2}} \\
    -P_2^{-\frac{1}{2}}SP_1^{-\frac{1}{2}} & 0
\end{pmatrix}.
\]

Then $Q$ is nonsingular since $K$ is skew-symmetric. Note that $K$ has only pure imaginary eigenvalues. This shows eigenvalues of $Q$ cannot be zero. Consequently we have the Theorem hold. \(
\)

We have some properties of the matrices $A$ and $B$ if $\alpha + \beta = -1$. In this case $AB = BA = \frac{1}{b'(1)} I_n$ i.e. $A$ is the inverse of $B$ if $b^2(1) = -1$. Also $AN_{1}^{-1}A^t$ becomes a diagonal matrix $\frac{1}{b'(1)}N_{2}^{-1}$.

4. Error Analysis

In this section, we will show the convergence of the method proposed and the error bound of this method in uniform norm. Let $\epsilon_{G_1}$ be the vector of quadrature errors at the node points $t_i$ and $\epsilon_{G_2}$ the one at $s_j$. Then the system (10) can be rewritten with the errors $\epsilon_{G_1}$ and $\epsilon_{G_2}$ in (6) and (9) respectively as follows:

\[
M \ y^* = f^* = (f_1^T - \epsilon_{G_1}^T, f_2^T - \epsilon_{G_2}^T)
\]

where $y^*$ is the exact solution and

\[
\begin{align*}
  f_1^T &= [f_s(s_1), \ldots, f_s(s_{n-1}), C_o + h] \\
  f_2^T &= [f_s(t_1) - h_o, \ldots, f_s(t_n) - h_o].
\end{align*}
\]

Let us define the error vector

\[
\epsilon = y^* - y.
\]

We have then

\[
M\epsilon = (-\epsilon_{G_1}^T, -\epsilon_{G_2}^T)^T
\]

This system can be written as

\[
RT\epsilon = \xi
\]

where \( \xi = \begin{pmatrix} -\epsilon_{G_1}^T \\ -\epsilon_{G_2}^T \end{pmatrix} \).

Hence we have the exact error expression by Theorem 2 as follows:

\[
\epsilon = T^{-1}R^{-1}\xi.
\]

Taking the Euclidean norms, we have the estimate

\[
||\epsilon||_2 \leq ||T^{-1}||_2 \ ||R^{-1}||_2 \ ||\xi||_2.
\]

Lemma 3. $R$ is nonsymmetric positive definite.
Proof. For any nonzero \( x = (x_1^t, x_2^t)^t \in \mathbb{R}^{2n} \),
\[
x^t R x = (x_1^t, x_2^t) \begin{pmatrix}
P_1 & S \\
-S^t & P_2
\end{pmatrix} \begin{pmatrix}
x_1 \\
x_2
\end{pmatrix} = (x_1^t, x_2^t) \begin{pmatrix}
P_1 x_1 + S x_2 \\
-S^t x_1 + P_2 x_2
\end{pmatrix}
\]
\[
= x_1^t P_1 x_1 + x_1^t S x_2 - x_2^t S^t x_1 + x_2^t P_2 x_2
\]
\[
= P_1 \|x_1\|^2 + P_2 \|x_2\|^2 > 0
\]
since \( P_1 \) and \( P_2 \) are positive diagonal matrices and \( x_1^t S x_2 - x_2^t S^t x_1 = 0 \) as below.
\[
x_1^t S x_2 = (S x_2, x_1) = (x_1, S x_2) = (S x_2)^t x_1 = x_2^t S^t x_1
\]

Now let us define the symmetric part of matrix \( R \) by
\[
U = \frac{1}{2} (R + R^t) = \begin{pmatrix}
P_1 & 0 \\
0 & P_2
\end{pmatrix}.
\]

Note that \( U \) is symmetric positive definite by Lemma 3.

**Lemma 4.**
\[
\|R^{-1}\|_2 \leq \|U^{-1}\|_2
\]

**Proof.** It suffices to show
\[
\sigma_n(R) \geq \lambda_n(U)
\]
where \( \sigma_n(R) \) is the smallest singular value of \( R \) and \( \lambda_n(U) \) is the smallest eigenvalue of \( U \). Since \( U \) is symmetric positive definite, \( \lambda_n(U) \) is positive i.e.,
\[
\sigma_n(R) \geq \lambda_n(U) > 0.
\]

This shows
\[
\frac{1}{\sigma_n(R)} \leq \frac{1}{\lambda_n(U)}.
\]

Consequently, this lemma holds. For any \( x \in \mathbb{R}^{2n} \) with \( \|x\|_2 = 1 \),
\[
x^t U x = \frac{1}{2} (x^t R x + x^t R^t x) = x^t R x \leq \|x\|_2 \|R x\|_2.
\]

By Courant-Weyl theorem,
\[
\lambda_n(U) = \min_{w_1, ..., w_{n-1} \in \mathbb{R}^{2n}} \max_{x \in \mathbb{R}^{2n}} x^t U x
\]
\[
\leq \min_{w_1, ..., w_{n-1} \in \mathbb{R}^{2n}} \max_{x \in \mathbb{R}^{2n}} \|R x\|_2
\]
\[
= \sigma_n(R)
\]

By using Lemma 4, the error bound (12) becomes
\[
\|e\|_2 \leq \|T^{-1}\|_2 \|U^{-1}\|_2 \|\epsilon\|_2.
\]

Hence we have the error in uniform norm in the following Theorem.
Theorem 5.
\[ \|\epsilon\|_\infty \leq C_1 n^{1/2} \|\epsilon\|_\infty \]
where
\[ \|\epsilon\|_\infty = \max(\|\epsilon_{G_1}\|_\infty, \|\epsilon_{G_2}\|_\infty). \]

Proof. Since \( T \) and \( U \) are digonal matrices, from (13),
\[
\|\epsilon\|_2 \leq \|T^{-1}\|_2 \|U^{-1}\|_2 \|\epsilon\|_2 \\
= \|T^{-1}\|_\infty \|U^{-1}\|_\infty \|\epsilon\|_2 \\
\leq \|T^{-1}\|_\infty \|U^{-1}\|_\infty n^{1/2} \|\epsilon\|_\infty.
\]
Here we can get the uniform norm of \( T^{-1} \) which is expressed in (11). From [13], we have
(14) \[ w_i = O(n^{-1}) \quad \text{and} \quad w_j^* \cdot \rho^2(s_j) \sim w_j \]

since
\[
w_j^* \cdot \rho^2(s_j) \sim j^{-2\alpha+1} \cdot n^{2\alpha-2} \cdot \left( \frac{j}{n} \right)^{4\alpha} \\
= j^{2\alpha+1} \cdot n^{-2\alpha-2} \\
\sim w_j.
\]
This gives us
(15) \[ \|T^{-1}\|_\infty = O(n) \]

On the other hand,
\[
U^{-1} = \begin{pmatrix} P^{-1}_1 & 0 \\ 0 & P^{-1}_2 \end{pmatrix} = \begin{pmatrix} N_2 D_3 D^{-1}_1 & 0 \\ 0 & N_1 D^{-1}_2 \end{pmatrix}
\]
By Lemma 1, \((D_2)_{ii} \sim \rho(t_i)\), thus for the positive \( t_i \) (If we use the negative \( t_i \), then we replace \( \alpha \) by \( \beta \)),
\[
(N_1 D^{-1}_2)_{ii} \sim w_i \cdot \rho^{-1}(s_j) \\
\sim i^{2\alpha+1} \cdot n^{-2\alpha-2} \cdot \left( \frac{i}{n} \right)^{-2\alpha} \\
= i \cdot n^{-2}.
\]
This gives us the following,
(16) \[ \|N_1 D^{-1}_2\|_\infty = O(n^{-1}). \]

Also, for \( i \neq n \),
\[
(N_2 D_3 D^{-1}_1)_{ii} \sim w_i^* \cdot \rho^2(s_i) \cdot \rho^{-1}(s_j) = w_i^* \cdot \rho(s_i) \sim w_i \cdot \rho^{-1}(s_i)
\]
by (14). In case \( i = n \),
\[
(N_2 D_3 D^{-1}_1)_{nn} = |b(1)|^{-1} \cdot \frac{1}{l} = |b(1)|^{-1} \cdot w_n = O(n^{-1}).
\]
Similarly,
\begin{equation}
\|N_2 D_3 D_1^{-1}\|_{\infty} = O(n^{-1}).
\end{equation}

From (16) and (17),
\begin{equation}
\|U^{-1}\|_{\infty} = O(n^{-1}).
\end{equation}

Here we have the following bound,
\[\|e\|_{\infty} \leq C_3 \cdot n^{1/2} \cdot \|e\|_{\infty}\]

since \[\|T^{-1}\|_{\infty} \|U^{-1}\|_{\infty} \leq C_4\] from (15) and (18).

To get the error bound of the system, we need to know the error of Hunter’s method. It can be obtained from [2, 3, 16]. Now we have the error for Hunter’s method as follows.

**Lemma 6.** If \(y^{(m)} \in H_\mu\) with \(m + \mu \geq 1\), then for \(x \in (-1, 1)\)
\[|e|_G = O\left(\frac{\ln n}{n^{\mu-2\nu}}\right)\]

where \(\nu\) is any positive number such that \(2\nu < \mu\).

Lemma 6 and Theorem 5 let us have the convergence of the proposed method and its convergence rate is given by the following Theorem.

**Theorem 7.** \(y^{(m)} \in H_\mu\) with \(m + \mu \geq 1\), the following estimate holds
\[\|e\|_{\infty} \leq C_6 n^{-(m+\mu-1)/2-2\nu-\varepsilon}\]

with \(\varepsilon > 0\) arbitrarily small.

In this procedure, we need \(y^* \in H^1\) to obtain the convergence [11, 3]. If we choose Elliott-Paget method [3] instead of Hunter’s, and proceed in a similar way, we have the convergence and its convergence rate is given as below

**Corollary 8.** If \(y^{(m)} \in H_\mu\) with \(m + \mu > 1/2\),
\[\|e\|_{\infty} \leq C_6 n^{-(m+\mu-1/2-2\nu-\varepsilon)}\]

Finally, we construct the approximate solution represented by the Lagrange interpolatory polynomial \(P_{2n-1}(x)\) on the nodes \(\{t_1, s_1, \cdots, s_{n-1}, t_n\}\),
\[P_{2n-1}(x) = \frac{1}{2} \left[ \sum_{i=1}^{n} y(t_i) - \frac{P_n^{(a,\beta)}(x)}{(x-t_i) P_n^{(a,\beta)'}(t_i)} + \sum_{j=1}^{n-1} y(s_j) - \frac{P_{n-1}^{(1-a,\beta)}(x)}{(x-s_j) P_{n-1}^{(1-a,\beta)'}(s_j)} \right].\]
Let $P_{2n-1}^s(x)$ be the polynomial of degree $2n - 1$ interpolating on the exact values of $y^*(x)$. Then

$$
|P_{2n-1}^s(x) - P_{2n-1}(x)| \leq \frac{1}{2} \left[ \sum_{i=1}^{n} |y^*(t_i) - y(t_i)| \left| \frac{P_{n}^{[\alpha, \beta]}(x)}{(x - t_i) P_{n}^{[\alpha, \beta]}(t_i)} \right| + \sum_{j=1}^{n-1} |y^*(s_j) - y(s_j)| \left| \frac{P_{n-1}^{[-\alpha, -\beta]}(x)}{(x - s_j) P_{n-1}^{[-\alpha, -\beta]}(s_j)} \right| \right]
$$

where $\Lambda_p$ is the Lebesgue constant [13] and

$$
\Lambda_p = O(\log n).
$$

Consequently, we have an estimate between the exact solution and the approximate solution as follows.

$$
\| y^* - P_{2n-1} \|_\infty \leq \| y^* - P_{2n-1}^s \|_\infty + \| P_{2n-1}^s - P_{2n-1} \|_\infty
\leq C_7\omega(y^*, \frac{1}{n}) + C_8 \| e \|_\infty \log n
$$

where $\omega$ is the modulus of continuity. This ensures the rate convergence under the result of Theorem 7.

**Theorem 9.** If $y^{*(m)} \in H_\mu$ with $m + \mu \geq 1$, then

$$
\| y^* - P_{2n-1} \|_\infty \leq C_9 \cdot n^{-(m+\mu-1)/2-\nu-\varepsilon}
$$

The above result can be strengthened with the Elliott-Paget method i.e. we have a weaker condition such as $m + \mu > 1/2$. The rates of convergence of two methods are almost same even if they have different restrictions on the smoothness of the function $y^*$.

**References**


Department of Mathematics
SungKyunKwan University
email: skim@yurim.skku.ac.kr
A Parallel Iterative Algorithm for Solving The Eigenvalue Problem of Symmetric matrices

Ran Baik

Abstract

This paper is devoted to the parallelism of a numerical matrix eigenvalue problem. The eigenproblem arises in a variety of applications, including engineering, statistics, and economics. Especially we try to approach the industrial techniques from mathematical modeling. This paper has developed a parallel algorithm to find all eigenvalues. It is contributed to solve a specific practical problem, a vibration problem in the industry. Also we compare the runtime between the serial algorithm and the parallel algorithm for the given problems.

1 Introduction

The problem of finding eigenvalues and eigenvectors arises in a wide variety of practical applications. It arises in all branches of science and engineering. The mathematical models of many engineering problems are systems of differential and difference equations, and the solutions of these equations are often expressed in terms of the eigenvalues and eigenvectors of matrices of these systems. Furthermore, many important characteristics of physical and engineering systems, such as stability, can often determined only by knowing the nature and location of eigenvalues.

keywords and phrases: Eigenproblem, Gap*, Group Homotopy, Modified Newton Method, Parallel Computing
1991 Mathematics Subject classification by the American Mathematical Society
The author wishes to acknowledge the financial support of Honam university made in the program year of 2000
1.1 Stability problems for differential equations

A homogeneous linear systems of differential equations with constant coefficients of the form

\[
\begin{align*}
\frac{d}{dt} x_1(t) &= a_{11}x_1(t) + a_{12}x_2(t) + \cdots + a_{1n}x_n(t) \\
\frac{d}{dt} x_2(t) &= a_{21}x_1(t) + a_{22}x_2(t) + \cdots + a_{2n}x_n(t) \\
\vdots & \quad \vdots \quad \vdots \\
\frac{d}{dt} x_n(t) &= a_{n1}x_1(t) + a_{n2}x_2(t) + \cdots + a_{nn}x_n(t)
\end{align*}
\]

or in matrix form

\[
\dot{x}(t) = Ax(t)
\]

where \( A = (a_{ij})_{n \times n} \) and \( \dot{x}(t) = \frac{d}{dt} \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix} \) arises in a wide variety of physical and engineering systems. The solution of this system is intimately to the eigenvalue problem for the matrix A. To see this, assume that (1) has a solution \( x(t) = ve^\lambda t \), where \( \lambda \) is not dependent on \( t \). Then from (1) we must have \( \lambda ve^\lambda t = Ave^\lambda t \) that is \( Av = \lambda v \) showing that is an eigenvalue of A and is corresponding eigenvector. Thus the eigenpair \( (\lambda, v) \) of A can be used to compute a solution \( x(t) \) of (1). If A has \( n \) linearly independent eigenvectors (which will happen, as we saw when the eigenvalues of A are all distinct), then the general solution of the system can be written as

\[
x(t) = c_1v_1 e^{\lambda_1 t} + c_2v_2 e^{\lambda_2 t} + \cdots + c_nv_n e^{\lambda_n t}
\]

where \( \lambda_1, \lambda_2, \cdots, \lambda_n \) are the eigenvalues of A and \( v_1, v_2, \cdots, v_n \) are the corresponding eigenvectors.

1.2 Vibration Problem

A problem such as an analysis of a vibration of structures often gives rises to a system of second-order differential equations of the form

\[
By'' + Ay = 0
\]

where \( y'' = (y_1''(t) \ y_2''(t) \ \cdots \ y_n''(t))^T \). The solution of such a system leads to the solution of an eigenvalue problem of the type:

\[
Ax = \lambda Bx
\]
This can be seen as follows: Assume that solutions are of the form $y = xe^{\lambda t}$. Then from (3) we must have $\omega^2 Bx = Ax$. Writing $\lambda = \omega^2$, this becomes $Ax = \lambda Bx$. Such an eigenvalue problem is called a generalized eigenvalue problem. The number $\lambda$ is called an eigenvalue, and $x$ is the corresponding eigenvector of the generalized eigenvalue problem. Very often, in practice, the matrices $A$ and $B$ are symmetric and positive definite. The eigenvalues of the generalized eigenvalue problem in this case are all real and positive, and eigenvectors can be taken to be linearly independent. In such a case, an explicit solution of the system can be written down in terms of the eigenvalues and eigenvectors. Writing $\omega = \pm \sqrt{\lambda}$, we have the following result.

<Eigenvalue-Eigenvector Solution of $By'' + Ay = 0$ with $A$, $B$ Positive Definite >

Let $x_1, x_2, \ldots, x_n$ be the $n$ independent eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of (4) with $A$ and $B$ positive definite. Then the general solution of (3) is given by

$$y = \sum_{k=1}^{n} \left( c_k e^{i\sqrt{\lambda_k} t} + d_k e^{-i\sqrt{\lambda_k} t} \right) x_k$$

where $c_k$ and $d_k$ are arbitrary constants \([9],34-45\).

In vibration problems, the matrices $B$ and $A$ are called the mass and stiffness matrices, respectively, and are denoted by $M$ and $K$, giving rise to the symmetric generalized eigenvalue problem: $Kx = \lambda Mx$

**Definition** 1.2.1: The quantities $\omega_i = \sqrt{\lambda_i}, i = 1, \ldots, n$, are called natural frequencies, and $x_1, x_2, \ldots, x_n$ are called the modes of a vibration of the system.

The frequencies can be used to determine the periods $T_i$ for the vibrations. Thus $T_i = \frac{2\pi}{\omega_i}$ is the period of vibration for the $i$th mode. As we will see, the behavior of a vibrating system can be analyzed by knowing the natural frequencies and the modes. We will give a simple example to illustrate this.

**Example** Consider a two-story building(Figure 1.1 (a)) with a rigid floor. It is assumed the weight distribution of the building can be represented in the form of concentrated weight at each floor level, as shown in Figure 1.1(b), and the stiffness of supporting columns are represented by the spring constants $k_i$. The equation of motion for this system can be written as

$$m_1 y_1'' + (k_1 + k_2) y_1 - k_2 y_2 = 0$$
$$m_2 y_2'' - k_2 y_1 + k_2 y_2 = 0$$
or

\[
\begin{pmatrix}
  m_1 & 0 \\
  0 & m_2
\end{pmatrix}
\begin{pmatrix}
  y_1' \\
  y_2'
\end{pmatrix} + \begin{pmatrix}
  k_1 + k_2 & -k_2 \\
  -k_2 & k_2
\end{pmatrix}
= 0
\]  

Fig 1.1 (a) \hspace{1cm} Fig 1.1 (b)

Define the mass matrix

\[
M = \begin{pmatrix}
  m_1 & 0 \\
  0 & m_2
\end{pmatrix}
\]

and the stiffness matrix

\[
K = \begin{pmatrix}
  k_1 + k_2 & -k_2 \\
  -k_2 & k_2
\end{pmatrix}
\]

Equation (5) becomes \(My'' + Ky = 0\) where \(y = (y_1 \ y_2)^T\). In a vibration analysis of structures, a common engineering practice is just to compute the first few smallest eigenvalues(frequencies) and the corresponding eigenvectors(mode), because it has been in practice that the largest eigenvalues and eigenvectors contribute very little to the total response of the system.

2 Methodology for the Eigenvalue Problem

In this section we briefly describe the method, group Homotopy that we have developed for finding the dominant eigenvalues and the corresponding eigenvectors of a matrix on a vibration problem. The vibration problem provides a symmetric positive definite for solving the eigenvalue problem.
A general Homotopy method called the group Homotopy method has attractive features that it preserves the structure of the original matrix and finds a specific eigenvalue without computing any other eigenvalues, in contrast with the existing methods which destroy the structure during similarity transformation. Since the method preserves the special structure of matrices, it is quite suitable to obtain eigenpairs of a class of matrices which possess some special characteristics that we might take advantage. In this paper, we present the method how to find extreme eigenvalues of a Symmetric matrix. Recall that a Homotopy method for the eigenvalue problem is in the following. Let

$$A = U \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} U^T$$

where $\lambda_k \in \mathbb{R}$ are the eigenvalues and $U_k \in \mathbb{R}^n$ are the corresponding eigenvectors of $A$ for $k=1,\ldots,n$. We denote by

$$\begin{pmatrix} u_k \\ \lambda_k \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}$$

the eigenpair of $A$ for $k=1,\ldots,n$. Note that since $\{U_k\}_{k=1,\ldots,n}$ is linearly independent, $\left\{ \begin{pmatrix} U_k \\ \lambda_k \end{pmatrix} \right\}$ is linearly independent. We denote by $\sigma(A)$ the set of eigenvalues of $A$ (including multiplicities). Suppose the eigenvalue $\lambda_k \in \sigma(A)$ has the algebraic multiplicity $p$, i.e., $\lambda_k = \lambda_{k+1} = \cdots = \lambda_{k+p}$. Then we note that any $X \in \text{span}\{U_i\}_{i=k,\ldots,k+p}$ is an eigenvector for $\lambda_k$. Thus an eigenpair $\begin{pmatrix} U_k \\ \lambda_k \end{pmatrix}$ is not unique. Let $S \in \mathbb{M}_n$ be the initial Symmetric matrix with eigenpairs $\begin{pmatrix} U_k \\ \lambda_k \end{pmatrix}$ for $k=1,\ldots,n$ that are known. Let $A \in \mathbb{M}_n$ be the objective Symmetric matrix with eigenpairs $\begin{pmatrix} U_k \\ \lambda_k \end{pmatrix}$ for $k=1,\ldots,n$ are to be obtained. Define a mapping $H : \mathbb{R} \rightarrow \mathbb{M}_n$ such that

$$H(t) = A(t) \equiv (1-t)S + tA, \quad t \in [0,1].$$

Note that $A(t) = S + t(A - S)$ so that the matrix $A(t)$ can be made close to $S$ by choosing $t$ small enough. The objective is to obtain the set of all the eigenpairs

$$\left\{ \begin{pmatrix} U_k \\ \lambda_k \end{pmatrix} \right\}, \quad k = 1,\ldots,n \text{ of } A(1) = A \text{ by successively obtaining the set of eigenpairs}$$

$$\left\{ \begin{pmatrix} U_k^{(i)} \\ \lambda_k^{(i)} \end{pmatrix} \right\}, \quad k = 1,\ldots,n \text{ of } A(t_i),0 = t_0 < t_1 < \cdots < t_n = 1,$$
eigenpairs \( \left\{ \left( \frac{x_k}{\alpha_k} \right) \right\}, k = 1, \ldots, n \) of \( A(0) = S \). The procedure is called the Homotopy method [4]. From the existing Homotopy method, it didn’t give us any criterion of the step size \( \Delta t \) and any results for clustering eigenvalues cases. To overcome the absence of any definite criterion to choose a step size that guarantees the convergence of the method and to define a step size that the eigenvalues of \( A(t_i) \) is clustering and, a new Homotopy method called, the group Homotopy method, is developed [1,2].

We describe the method to compute all eigenvalues of a matrix based on the group homotopy. We consider a definite criterion of the stepsize under the group homotopy method for the eigenproblem of a symmetric matrix. Suppose \( \Delta t_i \) is determined via \( Gap^* \) i.e., \( \Delta t_i = t_{i+1} - t_i = \frac{Gap^*}{q\|A - S\|_2} \), to obtain all eigenpairs of \( A(t_{i+1}) \) from \( A(t_i) \). It is given that the clustered eigenvalues of \( A(t_i) \) are grouped together whenever the gap between adjacent eigenvalues is less than \( Gap^* \). We show that there are some strong connections between eigenvalues of \( A(t_i) \) and \( A(t_{i+1}) \) under the assumption \( \|A - S\|_2 \leq \frac{Gap^*}{q} \). Suppose \( \sigma(A(t_i)) = \{\alpha_1 \cdots \alpha_n\}, \alpha_1 \geq \cdots \geq \alpha_n \) is grouped together under the clustering criterion determined by \( Gap^* \), \( \sigma(A(t_i)) = \bigcup_{k=1}^{g} G_k, G_k = \{\alpha_{k_1} \geq \cdots \geq \alpha_{k_s}\} \) is a group of clustering eigenvalues of \( A(t_i), s_1 + s_2 + \cdots + s_g = n \). We devise \( \sigma(A(t_{i+1})) = \{\lambda_1 \cdots \lambda_n\}, \lambda_1 \geq \cdots \geq \lambda_n \) to form the groups correspond to the grouping in \( \sigma(A(t_i)) \) such that \( \sigma(A(t_i)) = \bigcup_{k=1}^{g'} G'_k, G'_k = \{\lambda_{k_1} \geq \cdots \geq \lambda_{k_s}\}, s_1 + s_2 + \cdots + s_g = n \).

We call \( G'_k \) the counterpart of \( G_k \). Suppose \( \|A - S\|_2 \leq \frac{Gap^*}{q} \). Then we claim that the counterpart \( G' \) is close to \( G \) for a large enough \( q \).

**Lemma 2.1:** Suppose \( A, S \in M_n \) are symmetric and let \( \lambda_1 \geq \cdots \geq \lambda_n \) and \( \alpha_1 \geq \cdots \geq \alpha_n \) be the eigenvalues of \( A \) and \( S \) respectively. Then \( \|A - S\|_2 \leq \max_k \{\lambda_k - \alpha_k, (A - S)\} \) where \( \rho(A - S) = \max_k \{\lambda_k (A - S)\}/\lambda_k (A - S) \) are the eigenvalues of \( (A - S) \). Consequently, if \( \|A - S\|_2 \leq \frac{Gap^*}{q} \), then \( |\lambda_k - \alpha_k| \leq \frac{Gap^*}{q} \) for all \( k = 1, \ldots, n \). Now we make the following observation. Note that \( \|A(t_{i+1}) - A(t_i)\|_2 = \|t_{i+1} - t_i\||A - S||_2 = \Delta t_i \|A - S||_2 \leq \frac{Gap^*}{q\|A - S\|_2} \cdot \|A - S\|_2 = \frac{Gap^*}{q} \).

Thus, \( \max_k \{\lambda_k - \alpha_k\} \leq \frac{Gap^*}{q} \) by Lemma 2.1, where \( \lambda_k \) and \( \alpha_k \) are the eigenvalues of \( A(t_{i+1}) \) and \( A(t_i) \), respectively. Therefore, the \( k \)-th eigenvalue of \( A(t_{i+1}) \) must be within \( \frac{Gap^*}{q} \) distance from the \( k \)-th eigenvalue of \( A(t_i) \) for all \( k = 1, \ldots, n \).

Now we verify the claim statement.
Lemma 2.2: Suppose \( \|A - S\|_2 = \frac{\text{Gap}^*}{q} \), \( q > 1 \) and let \( G_1 \) and \( G_2 \) be two distinct groups of clustering eigenvalues of \( A(t_i) \). If \( G'_1 \) and \( G'_2 \) are the groups of eigenvalues of \( A(t_{i+1}) \) counterpart to \( G_1 \) and \( G_2 \) respectively, then \( \text{dist}(G'_1, G'_2) \geq (1 - \frac{2}{q})\text{Gap}^* \), where \( \text{dist}(G'_1, G'_2) = \min\{|\lambda_i - \lambda_j|/\lambda_i \in G'_1 \text{ and } \lambda_j \in G'_2\} \).

Proof: Suppose \( G_1 \) and \( G_2 \) are two distinct groups in \( A(t_i) \). Then \( |\alpha_s - \alpha_t| > \text{Gap}^* \) for all \( \lambda_s \in G_1 \) and \( \lambda_t \in G_2 \). Now for \( \alpha_s \in G'_1 \) and \( \alpha_t \in G'_2 \), \( \text{Gap}^* < |\alpha_s - \alpha_t| = |\alpha_s - \lambda_s + \lambda_s - \lambda_t + \lambda_t - \alpha_t| \leq |\alpha_s - \lambda_s| + |\alpha_t - \lambda_t| + |\lambda_s - \lambda_t| \). Since \( |\alpha_s - \alpha_t| < \frac{\text{Gap}^*}{q} \) and \( |\alpha_t - \alpha_s| \), by Lemma 2.1, \( \text{Gap}^* < \frac{2\text{Gap}^*}{q} + |\lambda_s - \lambda_t| \), or \( |\lambda_s - \lambda_t| > (1 - \frac{2}{q})\text{Gap}^* \). Therefore Lemma 2.2 verifies that for a large enough \( q > 1 \), any two distinct groups \( G'_1 \) and \( G'_2 \) are separated by at least \((1 - \frac{2}{q})\text{Gap}^* \) distance. Now we verify that the above is enough to guarantee the convergence of \( \alpha \) to some \( \lambda \in G' \).

Theorem 2.3: Suppose \( \|A - S\|_2 = \frac{\text{Gap}^*}{q} \), \( q > 1 \) and let \( G = \{\alpha_{i_1}, \ldots, \alpha_{i_t}\} \) be a group of clustering eigenvalues, \( \alpha_{i_1} \leq \cdots \leq \alpha_{i_t} \). Let \( B = (\alpha_{i_s} - \frac{\text{Gap}^*}{q}, \alpha_{i_1} + \frac{\text{Gap}^*}{q}) \subseteq R \) be a closed interval. Then the iteration \( \{\alpha^{(i)}\} \) under the modified Newton method starting at \( \alpha^{(0)} = \alpha \in B \) remain in \( B \), i.e., for all \( i \geq 2 \).

Since the modified Newton method is guaranteed to converge, theorem 2.3 asserts that each \( \alpha \in G \) must converge to some eigenvalue \( \lambda \in G' \) of \( A(t_{i+1}) \). Now, it remains to show a way to obtain all eigenpairs of \( A(t_{i+1}) \). One simple way to achieve the goal is by employing the Gram-Schmidt orthogonalization process.

3 Parallel Algorithm for the Eigenvalue Problem

In this section, we develop the parallel algorithm and its experiments of the given problem. Since the modified Newton method is guaranteed to converge, Theorem 2.3 asserts that each \( \alpha \in G \) must converge to some eigenvalue \( \lambda \in G' \) of \( A(t_{i+1}) \). We have discussed about the basic idea of the group homotopy in the section 2. Now we want to apply the group homotopy method for computing all eigenvalues of a Symmetric matrix in parallel. A parallel computing is instead of a single CPU managing a serial program control flow, several(many) CPUs work together, each running a certain part of the overall computation. Goal is to decrease the runtime for the solution to a problem and to increase the size of the problem that can be solved. In our research, it is based on the parallel computer shared memory(GS320), SMP cluster(HPC320), hybrid
architectures (HPC160). Also in the parallel computing, there are the program models; distributed message passing parallelism (MPI, PVM), data parallelism (HPF, SIMD), shared memory parallel directives (OpenMP). In this paper, we are going to develop the parallel algorithm with a program model, OpenMP, on HPC320 system. OpenMP specification for programming shared memory multiprocessor computers specifies a portable and simple to use directive-based parallelization model with several advantages over other approaches: (i) Multi-platform, multi-OS support, (ii) Incremental parallelization of sequential codes, (iii) Single source code solution, (iv) Easier to use than threads or message passing, (v) Allows verifiable correctness of parallel programs. Those merits provide us another efficient results for solving for the eigen problem of the industrial techniques.

3.1 Algorithm

Now, we describe the algorithm in parallel on HPC 320. In our program, since we have the independent eigenpath of the modified newton method at each step $A(t_i)$, it is a big role to develop the parallelism of the algorithm in the modified newton method.

**Step 1:** Let $A = (a_{ij}) \in M_n$ be a given symmetric matrix with eigenpairs for $k = 1, \ldots, n$. Choose $S = \text{diag}(\alpha_1^{(0)}, \ldots, \alpha_n^{(0)})$ be to the initial matrix where the diagonal elements of $A$ are arranged in decreasing order.

Set the initial eigenpairs as $D = \begin{pmatrix} X_k^{(0)} \\ \alpha_k^{(0)} \end{pmatrix}$ for $k = 1, \ldots, n$ where $X_k^{(0)}$ is $e_k$.

**Step 2:** Compute for $k = 1, \ldots, n$ (in parallel with 4 processors)

(i) $\|(A - \alpha_k^{(0)} I)X_k^{(0)}\|_2$.

(ii) If $\|(A - \alpha_k^{(0)} I)X_k^{(0)}\|_2 < \epsilon, \epsilon = 10^{-8}$, then go to Step 5, otherwise go to Step 3.

**Step 3:** Apply the modified Newton method.

Modified Newton’s Iteration: (in parallel with 4 processors)

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

For $i = 1, \ldots$.

(i) Solve $(A - \alpha_k^{(i)} I)Y_k^{(i)} = X_k^{(i)}$.

(ii) Compute $\beta_k^{(i)} = (X_k^{(i)})^T Y_k^{(i)}$.

(iii) Compute $\alpha_k^{(i+1)} = \|Y_k^{(i)}\|_2$.

(iv) $X_k^{(i+1)} = \frac{Y_k^{(i)}}{\beta_k^{(i)}}$.

(v) $\alpha_k^{(i+1)} = \frac{\alpha_k^{(i)}}{\beta_k^{(i)}}$. 
(vi) \( \alpha_k^{(i)} > \alpha_k^{(i+1)} \), \( X_k^{(i)} > X_k^{(i+1)} \).

(vii) Check \( \| (A - \alpha_k^{(i)} I) X_k^{(i)} \|_2 < \epsilon \). Otherwise, go to Step 3 - (i).

**Step 4:** Let \( m \) be the number of eigenpairs obtained from Step 3. If \( m = n \), then all the \( n \) eigenpairs are obtained. If \( m < n \), then obtain the \( (n - m) \) eigenpairs as follows: Denote by \( \{ G_k \}_{j=1}^m \) the \( m \) groups of eigenpairs as defined below. \( G_k = \left\{ \left( X_k^{(0)} \atop \alpha_k^{(0)} \right) / \left( X_k^{(0)} \atop \alpha_k^{(0)} \right) \right\} \) converges to \( \left( U_k \atop \lambda_k \right) \) for \( k = 1, \cdots, m \) where \( m_1 + \cdots + m_m = n \), \( m_i \) is the number eigenpairs in each group.

For \( k = 1, \cdots, m \) (in Parallel with 4 processors)

(i) Compute

\[
\min_{1 \leq j \leq m_k} \| (A - \alpha_k^{(0)} I) X_k^{(0)} \|_2.
\]

(Note that for each group there is one vector which satisfies the above relation.)

(ii) Orthogonalize the other vectors.

For \( k = 1, \cdots, m \) (in Parallel with 4 processors):

For \( j = 1, \cdots, (m_k - 1) \),

Orthogonalize using \( \{ U_1, U_2, \cdots, U_l, X_k^{(0)} \} \rightarrow \{ U_1, U_2, \cdots, U_l, X_k^{(0)} \} \) using the modified Gram-Schmidt process.

(iii) Using Step 3 with eigenpair \( \left( X_k^{(0)} \atop \alpha_k^{(0)} \right) \), obtain a new set of eigenpairs \( \left( X_k^{(i)} \atop \alpha_k^{(i)} \right) \).

(iv) Set \( \left( X_k^{(0)} \atop \alpha_k^{(0)} \right) \rightarrow \left( U_k \atop \lambda_k \right) \).

**Step 5:** End.

### 3.2 Numerical Experiments

In this section, we provide the numerical experiments of two kinds types on HPC320, (i) the serial algorithm to compute all eigenvalues of a symmetric matrix, (ii) the comparison of the serial and the parallel computing algorithm with increasing the size. In the implementation, we consider to use \( \| A - S \|_F \) to compute \( \Delta t_i \) in the following examples since \( \| \cdot \|_2 \leq \| \cdot \|_F \) where \( \| \cdot \|_F \) is the Frobenius matrix norm and \( \| \cdot \|_F \) is easy to compute.

\( M \) = the number of eigenpairs obtained using the modified Newton iteration.

\( B \) = the number of eigenpairs recovered with modified Gram-Schmidt process using the modified Newton iteration.
C = the maximum number of iterations at each step.
D = the maximum number of iterations needed for the orthogonalization.

**Experiment 3.2.1:** Consider the matrix

\[ A = [a_{ij}] \in M_n \quad a_{ij} = i \quad \text{if } i > j, \]
\[ a_{ij} = (-1)^{i+j} \quad \text{if } i = j, \]
\[ a_{ij} = j \quad \text{if } j > i. \]

Choose the initial matrix \( S \) such that \( A(0) = S = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn}) \).

\[ \|A - S\|_F = 1779.2. \]

**Experiments 3.2.2:**

We compare with the serial algorithm and the parallel algorithm increasing the size of \( A \).

| Table 3.2.1 (Serial Algorithm for finding all eigenvalues of \( A \)) |
|---|---|---|---|---|---|---|---|
| \( t_i \) | 0 | 0.25 | 0.5 | 0.75 | 1.0 | Exact |
| e-values | A(0) | A(0.25) | A(0.5) | A(0.75) | A(1.0) | E-values |
| 1st | 150.00 | 457.8705 | 879.5562 | 1308.3704 | 1739.0537 | 1739.0537 |
| : | : | : | : | : | : | : |
| 20th | 36.00 | 32.0262 | 32.6110 | 27.9100 | 23.4161 | 23.4161 |
| : | : | : | : | : | : | : |
| 30th | -27.00 | -27.6036 | -30.4504 | -33.3424 | -36.2029 | -36.2029 |
| : | : | : | : | : | : | : |
| 40th | -87.00 | 94.4645 | -96.4081 | -104.6880 | -112.8265 | -112.8265 |
| : | : | : | : | : | : | : |
| 50th | -147.00 | -164.1900 | -215.4465 | -285.9817 | -363.2445 | -363.2445 |
| : | : | : | : | : | : | : |
| M | 46 | 44 | 45 | 45 | | |
| B | 4 | 6 | 5 | 5 | | |
| C(D) | 8(6) | 8(5) | 10(4) | 8(4) | | |

| Table 3.2.2 (CPU TIME(sec): Parallel Algorithm & Serial Algorithm) |
|---|---|---|---|---|---|---|
| \( n \) | 32 | 64 | 128 | 256 | 512 | 1024 |
| Serial | 0.32891 | 3.78102 | 38.46221 | 426.19090 | 3639.7295 | 53269.86037 |
| Parallel | 0.19715 | 2.53077 | 25.25010 | 340.94313 | 3119.51563 | 43292.95508 |
In our method, the choice of an initial matrix can be arbitrary even though a good choice of the initial matrix might facilitate convergence. The group homotopy is a general homotopy method in the following senses: (i) Unlike in the existing homotopy method[4], it gives a definite criterion of how to choose a step size that will guarantee the convergence in the modified Newton Iteration[1,2]. (ii) The method works regardless of the choice of the initial matrix, since the modified newton method is globally convergent at each step. But the good choice of the initial matrix makes the iteration decrease. (iii) The method guarantees a big enough step size so that the group homotopy method terminates in a finite number of steps and gives all the eigenpairs of the objective matrix. (iv) We can find the specific eigenvalue without calculating the other eigenvalues. (v) It has a natural parallelism. Since our Algorithm has many kinds of the merits, we apply the method on the super computer, i.e., HPC320. This parallel computer has a shared memory and it has 8-nodes SMP cluster. Each node has 4 cpu’s. From Table 3.2.2, we have an comparison experiment through the parallel algorithm by using OpenMP of the group homotopy. In this result, we can see that OpenMp programming is not enough to decrease the CPU time: In parallel algorithm, we are expecting to get the one - fourth of CPU time of the serial algorithm, but we can only decrease the one-fifth of the CPU time of the serial algorithm. This paper is the first parallel algorithm of the eigenproblem of the symmetric matrices as OpenMP with 4 processors. In our future of the parallel computing, we need to try another approach to develop the MPI that we have to concern the structure between the structure of the super computer. Also we try to apply it for the better quality super computer, GS320 which has 32 cpu’s (this is not available now).

Reference

[8] Yousef Saad, Iterative methods for sparse Linear system, pws publishing company,
1996.

Department of Mathematics
Natural Science College
Honam university
baik@honam.honam.ac.kr
Legendre Tau Method for the 2-D Stokes Problem

SeRan Jun, Sungkwon Kang† and YongHoon Kwon

Abstract

A Legendre spectral tau approximation scheme for solving the two-dimensional stationary incompressible Stokes equations is considered. Based on the vorticity-stream function formulation and variational forms, boundary value and normal derivative of vorticity are computed. A factorization technique for matrix stems based on the Schur decomposition is derived. Several numerical experiments are performed.

1 Introduction

The two-dimensional stationary Stokes equations describing the motion of an incompressible fluid in a bounded domain $\Omega \subset \mathbb{R}^2$ with the boundary $\Gamma$ can be written, in terms of primitive variables, as

\begin{align*}
-\nu \Delta u + \nabla p &= f \quad \text{in } \Omega, \\
\nabla \cdot u &= 0 \quad \text{in } \Omega, \\
u u &= g \quad \text{on } \Gamma,
\end{align*}

(1.1)

where $u$ is the velocity, $p$ is the pressure, $f$ is a field of given body forces, $\nu$ is the kinematic viscosity of the fluid, and $g$ is a given field defined on $\Gamma$ satisfying the global conservation property:

$$\int_{\Gamma} g \cdot \mathbf{n} \, ds = 0,$$

AMS subject classifications: 65M10, 76D07

Key words: Stokes equations, vorticity-stream function formulation, Legendre tau method, Schur decomposition.

†The research of this author was supported by the Chosun University Research Funds 1998.
where $\mathbf{n}$ is the unit outer normal vector on $\Gamma$. We assume, in this paper, that $\Omega = (-1,1) \times (-1,1)$ and $\Gamma_i, i = 1, 2, 3, 4$, are edges of the boundary $\Gamma$, starting from the south and turning counterclockwise (see Figure 1).

By applying the curl operation, the Stokes system (1.1), in terms of vorticity $\omega$ and stream function $\psi$, becomes

\begin{align}
-\nu \Delta \omega &= \text{curl} \mathbf{f} \quad \text{in } \Omega, \\
-\Delta \psi &= \omega \quad \text{in } \Omega. 
\end{align}

(1.2)

Since $\psi$ is unique up to a constant, the following boundary conditions corresponding to those of (1.1) are considered: for $i = 1, 2, 3, 4$,

\begin{align}
\psi(x) &= h_i := \int_{x_0x} \mathbf{g} \cdot \mathbf{n} \, ds \quad \text{on } \Gamma_i, \\
\frac{\partial \psi}{\partial \mathbf{n}} &= g_i := -\mathbf{g} \cdot \mathbf{s} \quad \text{on } \Gamma_i, 
\end{align}

(1.3)

where $\mathbf{s}$ is the tangential vector on $\Gamma$, $\psi(x_0) = 0$ for some $x_0 \in \Gamma$ and $x_0x$ is the path from $x_0$ to $x$ along $\Gamma$.

The advantage of the vorticity-stream function formulation (1.2)-(1.3) is that we do not need to deal with the divergence free condition $\nabla \cdot \mathbf{u} = 0$ and the pressure $p$. Note that the divergence condition is automatically satisfied and the pressure is dropped in (1.2). These lead to a low cost discretization in numerical implementation. Also, the velocity and the pressure can be easily recovered from the stream function. However, a critical drawback of the formulation (1.2)-(1.3) is the lack of boundary conditions on the vorticity $\omega$ while there are two boundary conditions on the stream function $\psi$. A well-known way to overcome this difficulty in finite difference or finite element methods is to define the boundary conditions of vorticity from the relation $\omega = -\Delta \psi$. In this paper, we derive an efficient method for finding the traces of the vorticity based on variational forms, Green’s formula and the Schur decomposition through a Legendre tau approximation. The ideas are similar to those proposed by Glowinski and Pironneau[8]. However, we only deal with a sparse, symmetric matrix system in which each column of the governing matrix is obtained by solving one Laplace equation through a Legendre tau approximation (see Section 3) instead of solving a full matrix system in which each column of the corresponding matrix is computed by solving two Laplace equations as in [8].
In recent years, a number of algorithms using spectral methods have been implemented for solving the Stokes and the Navier-Stokes equations. Meanwhile, various theoretical and numerical results dealing with spectral Galerkin and spectral collocation methods have been established (see, e.g., [1-5, 13-14] and references therein). However, to the authors' knowledge, the spectral tau methods seem to be less studied, although they are frequently used in practice because of their efficiencies in solving, for example, Helmholtz type equations (see, e.g., [10] for the fast Helmholtz solver and [9] for a Chebyshev tau solver). Legendre tau formulations for the Stokes problem can be founded in [12] and [15] in which the formulations are based on the velocity-pressure form with homogeneous boundary conditions.

An outline of this work is as follows. In Section 2, we introduce a decomposition of the system (1.2)-(1.3) and abstract variational forms. A Legendre tau approximation scheme and a factorization method are given in Section 3, and numerical results are presented in Section 4.

Throughout this paper, $H^s(\Omega)$ is the standard Sobolev space with the standard norm $\| \cdot \|_{H^s(\Omega)}$. We shall denote by $(\cdot, \cdot)$ the usual inner product of $L^2(\Omega)$. For any Banach (or Hilbert) spaces $X$ and $Y$, let $\mathcal{L}(U, V)$ be the space of all bounded linear operators from $U$ to $V$ and let $U'$ be the dual space of $U$.

## 2 Decomposition and Variational formulation

In this section, we decompose the system (1.2)-(1.3) into two systems and derive variational forms for solving the decomposed systems.

Let $X$ be the subspace of $\prod_{i=1}^4 H^2(\Gamma_i)$ defined by the matching condition:

$$h_i(e_{i+1}) = h_{i+1}(e_{i+1}) \quad \text{for } 1 \leq i \leq 4,$$

where $e_i$'s are the vertices of $\Gamma$ with the convention that $e_5 = e_1$. We shall denote by $\ll \cdot, \cdot \gg (\text{resp., } \langle \cdot, \cdot \rangle)$ the bilinear form of the duality between $X'$ and $X$ (resp., $(\prod_{i=1}^4 H^2(\Gamma_i))'$ and $\prod_{i=1}^4 H^2(\Gamma_i)$) which is defined by $\ll L, v \gg := L(v), \quad L \in X', \quad v \in X$ (resp., $\langle L, v \rangle := L(v), \quad L \in (\prod_{i=1}^4 H^2(\Gamma_i))'$, $v \in \prod_{i=1}^4 H^2(\Gamma_i)$). The bilinear form $\ll \cdot, \cdot \gg$ is an extension of $(\cdot, \cdot)_{L^2(\Gamma)}$: $\ll w, v \gg = \int_{\Gamma} wv \, ds$ for all $v \in X, w \in L^2(\Gamma)$. 
Consider the following spaces:
\[
H(\Omega) = \{u \in L^2(\Omega) \mid \Delta u \in L^2(\Omega)\},
\]
\[
\mathcal{H} = \{u \in L^2(\Omega) \mid \Delta u = 0\},
\]
\[
L^0_0(\Omega) = \{u \in L^2(\Omega) \mid (u,1) = 0\},
\]
\[
V = \{u \in H^2(\Omega) \mid \frac{\partial u}{\partial n} = 0\},
\]
\[
G = \{q \in X' \mid \ll q,1 \gg = 0\}.
\]

We now decompose the system (1.2)-(1.3) into the following two problems. Let \(\tilde{\psi}, \tilde{\omega}\) be the solutions of the problem:

\[
\begin{align*}
-\nu\Delta \tilde{\omega} &= f_1 & \text{in } \Omega, \\
-\Delta \tilde{\psi} &= \tilde{\omega} & \text{in } \Omega, \\
\tilde{\psi} &= 0 & \text{on } \Gamma_i, \\
\frac{\partial \tilde{\psi}}{\partial n} &= g_i & \text{on } \Gamma_i \text{ for } 1 \leq i \leq 4,
\end{align*}
\]  
(2.1)

and \(\tilde{\psi}, \tilde{\omega}\) be the solutions of the problem:

\[
\begin{align*}
-\nu\Delta \tilde{\omega} &= f_2 & \text{in } \Omega, \\
-\Delta \tilde{\psi} &= \tilde{\omega} & \text{in } \Omega, \\
\tilde{\psi} &= h_i & \text{on } \Gamma_i, \\
\frac{\partial \tilde{\psi}}{\partial n} &= 0 & \text{on } \Gamma_i \text{ for } 1 \leq i \leq 4,
\end{align*}
\]  
(2.2)

where \(\text{curl} f = f_1 + f_2, f_1 \in L^2(\Omega), f_2 \in L^0_0(\Omega)\). Then \(\psi = \tilde{\psi} + \tilde{\psi}\) and \(\omega = \tilde{\omega} + \tilde{\omega}\).

We now derive variational forms for the systems (2.1) and (2.2). Let \((-\Delta_d)^{-1} \in \mathcal{L}(L^2(\Omega), H^2(\Omega) \cap H^1_0(\Omega))\) denote Green’s operator related to the Dirichlet boundary value problem for \(-\Delta \text{ in } \Omega \subset \mathbb{R}^2\), i.e., for \(f \in L^2(\Omega)\), \(u = (-\Delta_d)^{-1} f\) is the solution of

\[
-\Delta u = f & \text{ in } \Omega, \quad u = 0 & \text{ on } \Gamma,
\]

and \((-\Delta_n)^{-1} \in \mathcal{L}(L^2_0(\Omega), V \cap L^2_0(\Omega))\) denote Green’s operator related to the Neumann boundary value problem for \(-\Delta \text{ in } \Omega \subset \mathbb{R}^2\); for \(f \in L^2_0(\Omega)\), \(u = (-\Delta_n)^{-1} f\) is the solution of

\[
-\Delta u = f & \text{ in } \Omega, \quad \frac{\partial u}{\partial n} = 0 & \text{ on } \Gamma.
\]
Let \( \gamma_0, \gamma_1 \) be the following trace operators:
\[
\gamma_0 v = (v|_{x_4}, \cdots, v|_{x_4}), \quad \gamma_1 v = \left( \frac{\partial v}{\partial n} |_{x_4}, \cdots, \frac{\partial v}{\partial n} |_{x_4} \right).
\]

Assume that curl \( f \in L^2(\Omega), (h_1, \cdots, h_4) \in X \) and \((g_1, \cdots, g_4) \in \prod_{i=1}^4 H^\frac{1}{2}(\Gamma_i)\). Then (1.1) is equivalent to (1.2)–(1.3).

We define an operator \( E \) on \( L^2(\Omega) \) by
\[
E \phi = \gamma_1((-\Delta_d)^{-1} \phi) \text{ for } \phi \in L^2(\Omega),
\]
and an operator \( T \) on \( L^2_0(\Omega) \) by
\[
T \phi = \gamma_0((-\Delta_n)^{-1} \phi) \text{ for } \phi \in L^2_0(\Omega).
\]

Then the operator \( E \) (resp., \( T \)) is a continuous linear operator from \( L^2(\Omega) \) to \( \prod_{i=1}^4 H^\frac{1}{2}(\Gamma_i) \) (resp., from \( L^2_0(\Omega) \) to \( X \)). Hence, the adjoint operator \( E^* \) of \( E \) is from \( (\prod_{i=1}^4 H^\frac{1}{2}(\Gamma_i))' \) to \( L^2(\Omega) \), and it is given by
\[
(E^* \mu, \phi) = \langle \mu, E \phi \rangle \text{ for } \mu \in (\prod_{i=1}^4 H^\frac{1}{2}(\Gamma_i))', \phi \in L^2(\Omega),
\]
and the adjoint operator \( T^* : X' \to L^2_0(\Omega)' \) of \( T \) is given by
\[
(T^* \mu)(\phi) = \langle \mu, T \phi \rangle \text{ for } \mu \in X', \phi \in L^2_0(\Omega).
\]

Then we have the following.

1. For any \( \mu \in (\prod_{i=1}^4 H^\frac{1}{2}(\Gamma_i))' \), let \( \tilde{\mu} \) be the unique solution of the problem:
\[
\begin{align*}
\Delta \tilde{\mu} &= 0 \text{ in } \Omega, \\
\gamma_0 \tilde{\mu} &= \mu \text{ on } \Gamma,
\end{align*}
\]
then \( E^* \mu = -\tilde{\mu} \).

2. For any \( \eta \in G \), let \( \tilde{\eta} \) be the unique solution of the problem:
\[
\begin{align*}
\Delta \tilde{\eta} &= 0 \text{ in } \Omega, \\
\gamma_1 \tilde{\eta} &= \eta \text{ on } \Gamma, \\
(\tilde{\eta}, 1) &= 0,
\end{align*}
\]
then \( T^* \eta = \tilde{\eta} \) in the \( (L^2_0(\Omega))' \)-sense.
By applying \((-\Delta_d)^{-1}\phi\) (resp., \((-\Delta_n)^{-1}\phi\)), and then Green’s second identity and the duality of \(E\) (resp., \(T\)), we have the variational form for (2.1) (resp., (2.2)):

\[
(\bar{\omega}, \phi) = -(E^* q, \phi) + \left(\frac{1}{\nu}(\Delta_d)^{-1} f_1, \phi\right) \text{ for any } \phi \in L^2(\Omega),
\]

\[
(\bar{\omega}, \phi) = (T^* p, \phi) + \left(\frac{1}{\nu}(\Delta_n)^{-1} f_2, \phi\right) \text{ for any } \phi \in L^2(\Omega),
\]

where \(q := \gamma_0 \bar{\omega} \in (\prod_{i=1}^{4} H^{\frac{1}{2}}(\Gamma_i))^\prime\) and \(p := \gamma_1 \bar{\omega} \in G\). Here, \(q\) and \(p\) satisfy the following linear variational equations:

\[
(E^* q, E^* \mu) = \left(\frac{1}{\nu}(\Delta_d)^{-1} f_1, E^* \mu\right) - \mu((g_1, g_2, g_3, g_4))
\]

\[
\text{for any } \mu \in (\prod_{i=1}^{4} H^{\frac{1}{2}}(\Gamma_i))^\prime,
\]

where \(\mu((g_1, g_2, g_3, g_4)) = \langle \mu, (g_1, g_2, g_3, g_4) \rangle\),

and

\[
(T^* p, T^* \eta) = -(\frac{1}{\nu}(\Delta_n)^{-1} f_2, T^* \eta) + \eta((h_1, h_2, h_3, h_4)) \text{ for any } \eta \in G,
\]

where \(\eta((h_1, h_2, h_3, h_4)) = \langle \eta, (h_1, h_2, h_3, h_4) \rangle\).

Therefore, the solution procedure for (2.1)–(2.2) is following:

1. Compute \(\frac{1}{\nu}(\Delta_d)^{-1} f_1\) and \(\frac{1}{\nu}(\Delta_n)^{-1} f_2\).

2. Compute \(q\), \(p\) from (2.7) and (2.8).

3. Finally, compute \(\bar{\omega}\), \(\bar{\phi}\) and \(\bar{\psi}\), \(\bar{\psi}\) from (2.5), (2.6) and (2.9)

\[
-\Delta \bar{\psi} = \bar{\omega}, \quad \bar{\psi} |_{\partial \Omega} = 0,
\]

\[
\bar{\psi} = (-\Delta_n)^{-1} \bar{\omega} + \frac{1}{4} \left(\sum_{i=1}^{4} \int_{\Gamma_i} h_i \, ds - \langle \gamma_0 ((-\Delta_n)^{-1} \bar{\omega}), 1 \rangle\right).
\]

Then the problem (2.1) (resp., (2.2)) is equivalent to (2.5), (2.7) and (2.9) (resp., (2.6), (2.8) and (2.10)). An application of these abstract forms to sine approximation, see [11].
Remark. Since \( q = \gamma_0 \tilde{\omega} \) and \( p = \gamma_1 \tilde{\omega} \) are computed from the boundary data \( g_h \)s and \( h_i \)s and the actions of \( E^* \) and \( T^* \) (see equations (2.7) and (2.8)), once \( p \) and \( q \) are obtained, \( \tilde{\omega} \), \( \tilde{\psi} \), \( \tilde{\omega} \) and \( \tilde{\psi} \) can be computed directly from equations (2.5), (2.6), (2.9) and (2.10). Thus, the main problem to be solved is to compute \( q \) and \( p \), in other words, the construction \( E^* \) and \( T^* \) through a Legendre tau approximation.

### 3 Legendre tau approximation scheme

In this section we describe a Legendre tau approximation and factorization scheme. Since the approximation scheme for (2.5), (2.7) and (2.9) can be described in a similar way, we present only the approximation scheme for (2.6), (2.8) and (2.10).

Let \( D \) be a subset in \( \mathbb{R} \) or \( \mathbb{R}^2 \). For any nonnegative integer \( M \) we denote by \( S_M(D) \) the space of all polynomials on \( D \) of degree \( \leq M \) in each variables. Denoting by \( S_{M,0}^1(D) \) the subspace of \( S_M(D) \) of all polynomials whose derivatives vanish on \( \partial \Omega \).

The Legendre polynomial \( L_k(x), \ k \geq 0, \) is orthogonal to any Legendre polynomial \( L_l, l \neq k, \) in \( L^2(-1,1) \), it has degree \( k \), \( L_k(1) = 1 \), and satisfies \( \int_{-1}^{1} L_k^2(x) dx = \frac{2}{2k+1} \) and \( \frac{\partial L_{k}(\pm 1)}{\partial x} = (\pm 1)^k \frac{k(k+1)}{2} \). Let \( P_M \) be the orthogonal projection operator in \( L^2(D) \) onto \( S_M(D) \). To simplify our expression, we assume that \( M \) is even.

We first consider a Legendre tau scheme for the following problem:

\[
\begin{align*}
-\Delta u &= f \quad \text{in } \Omega, \\
\frac{\partial u}{\partial n} &= 0 \quad \text{on } \Gamma.
\end{align*}
\]

(3.1)

Let the tau approximate solution for (3.1) be

\[
 u_M(x,y) = \sum_{k=0}^{M} \sum_{l=0}^{M} u_{kl} L_k(x)L_l(y).
\]

Note that the test functions do not satisfy the boundary conditions individually. Thus, it is necessary to have weighted residual conditions for both the PDE and the boundary
conditions. From the weighted residual conditions for the boundary conditions, we have

\[
\begin{align*}
    u_{k,M-1} \frac{(M-1)M}{2} &= - \sum_{l=1,l\text{ odd}}^{M-3} u_{kl} \frac{l(l+1)}{2}, \quad k = 0, \ldots, M, \\
    u_{k,M} \frac{M(M+1)}{2} &= - \sum_{l=2,l\text{ even}}^{M-2} u_{kl} \frac{l(l+1)}{2}, \quad k = 0, \ldots, M, \\
    u_{M-1,l} \frac{(M-1)M}{2} &= - \sum_{k=1,k\text{ odd}}^{M-3} u_{kl} \frac{k(k+1)}{2}, \quad l = 0, \ldots, M, \\
    u_{M,l} \frac{M(M+1)}{2} &= - \sum_{k=2,k\text{ even}}^{M-2} u_{kl} \frac{k(k+1)}{2}, \quad l = 0, \ldots, M.
\end{align*}
\]

(3.2)

Boundary equations (3.2) give only \(4M\) independent equations since the four corners of the square have been counted twice. From (3.2), the coefficients \(u_{M-1,l}, u_{k,M-1}, u_{Ml}, u_{kl}, k, l = 0, \ldots, M\) are determined by the coefficients \(u_{kl}, k, l = 0, \ldots, M - 2\), so that we need \((M - 1)^2 = (M + 1)^2 - 4M\) equations to determine the unknown coefficients \(u_{kl}, k, l = 0, \ldots, M,\) completely. Therefore, the spectral tau approximation of the problem (3.1) is equivalent to

\[
\begin{align*}
    \text{find} \quad & u_M \in S_M^{3,0}(\Omega) \cap L_2^0(\Omega) \text{ such that} \\
    & (-\Delta u_M, \phi) = (f, \phi) \text{ for any } \phi \in S_{M-2}(\Omega).
\end{align*}
\]

(3.3)

To introduce our diagonalization technique for the Legendre tau approximation scheme, we represent (3.3) as a linear system. We recall some properties (see [6]) which will be of constant use. The formal expansion of the first derivative of a function \(v(x) = \sum_{k=0}^{\infty} v_k L_k(x)\) can be written as

\[
\frac{dv(x)}{dx} = \sum_{m=0}^{\infty} \hat{v}_m^1 L_m(x),
\]

(3.4)

where

\[
\hat{v}_m^1 = (2m + 1) \sum_{k=m+1k+m\text{ odd}}^{\infty} \hat{v}_k.
\]

Actually, this formula can be justified for every \(v(x) \in H^1(-1,1)\) (see [7]). First, we shall construct the matrix \(A\) with size of \(M \times (M + 1)\) representing the differential operator \(\frac{d}{dx}\) on the space \(S_M(-1,1)\). Let \(v(x) := \sum_{k=0}^{M} v_k L_k(x)\) and \(\frac{dv(x)}{dx} := \sum_{k=0}^{M-1} a_k L_k(x)\).
Then, from relation (3.4), define the matrix $A$ by

$$
A := \begin{bmatrix}
 v_0 & 0 & 0 & 1 & 0 & \cdots & 1 & 0 \\
v_1 & 0 & 0 & 3 & 0 & \cdots & 0 & 3 \\
v_2 & 0 & 0 & 5 & 0 & \cdots & 5 & 0 \\
v_3 & 0 & 0 & 0 & 7 & \cdots & 0 & 7 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
v_{M-1} & 0 & 0 & 0 & 0 & \cdots & 2M - 3 & 0 \\
v_M & 0 & 0 & 0 & 0 & \cdots & 0 & 2M - 1
\end{bmatrix}
= \begin{bmatrix}
 a_0 & a_1 & a_2 & a_3 & \cdots & a_{M-2} & a_{M-1}
\end{bmatrix}^T.
$$

Second, we shall construct the matrix $A_0$ of size $M \times (M-1)$ representing the differential operator $\frac{d}{dx}$ on the space $S_M^{1,0}(-1,1)$. Let $v \in S_M^{1,0}(-1,1)$. Since $\frac{dv}{dx}(\pm 1) = 0$, we have the following relation.

$$
v_{M-1}^{M-1-M+1} = -(v_1 \frac{M+1}{2} + v_3 \frac{3-1}{2} + v_5 \frac{5-1}{2} + \cdots + v_M \frac{M-1-M+1}{2}),
$$

$$
v_{M-1}^{M-1-M} = -(v_1 \frac{M+1}{2} + v_3 \frac{3-1}{2} + v_5 \frac{5-1}{2} + \cdots + v_M \frac{M-3-M-2}{2}).
$$
By using (3.6), define the matrix $A_0$ by

$$A_0 := \begin{bmatrix}
  v_0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & \cdots & 0 \\
  v_1 & 0 & 0 & 3 & 0 & 3 & 0 & \cdots & 3 \\
  v_2 & 0 & 0 & 0 & 5 & 0 & 5 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  v_{M-2} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
  v_{M-2} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  v_{M-2} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  \end{bmatrix}
$$

(3.7) Then there exists a

$$B := \begin{bmatrix}
  1 & 0 & 0 & 0 & \cdots & 0 \\
  0 & 5 & 0 & \cdots & \cdots & \cdots \\
  0 & \vdots & \ddots & \vdots & \vdots & \vdots \\
  0 & \vdots & \vdots & \ddots & \vdots & \vdots \\
  0 & \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & \vdots & \vdots & \vdots & \vdots & \ddots \\
  2M - 3 & 0 & \cdots & \cdots & \cdots & \cdots \\
  2M - 3 & 0 & \cdots & \cdots & \cdots & \cdots \\
  2M - 5 & 0 & \cdots & \cdots & \cdots & \cdots \\
  2M - 1 & 0 & \cdots & \cdots & \cdots & \cdots \\
  \end{bmatrix}
$$

Third, we shall construct the matrix $B$ representing the differential operator $-\frac{d^2}{dx^2}$ on the space $S_{M}^{1,0}(-1,1)$. The matrix $B$ must have size of $(M - 1) \times (M - 1)$. Let $v(x) \in S_{M}^{1,0}(-1,1)$ and $\phi(x) := \sum_{k=0}^{M-2} \phi_k L_k(x) \in S_{M-2}(-1,1)$. Then there exists a
We define the matrices

\[
\psi(x) := \phi(x) + \alpha L_{M-1}(x) + \beta L_M(x) \in S^{1,0}_M(-1,1) \text{ so that}
\]

\[
(3.8) \quad \int_{-1}^{1} - \frac{d^2 v(x)}{dx^2} \phi(x) dx = \int_{-1}^{1} - \frac{d^2 v(x)}{dx^2} \psi(x) dx = \int_{-1}^{1} \frac{dv(x)}{dx} \frac{d\psi(x)}{dx} dx.
\]

Let \( \psi(x) := \sum_{k=0}^{M} b_k L_k(x) \). Then \( \psi_k = \phi_k, 0 \leq k \leq M - 2 \), and \( \psi_{M-1} = \alpha, \psi_M = \beta \). If we let

\[
\frac{dv(x)}{dx} := \sum_{k=0}^{M-1} b_k L_k(x),
\]

\[
[\bar{v}] := [v_0 \quad v_1 \quad \cdots \quad v_{M-2}]^T,
\]

\[
[\bar{\psi}] := [\psi_0 \quad \psi_1 \quad \cdots \quad \psi_{M-2}]^T,
\]

then from (3.7), we have

\[
(3.9) \quad A_0[\bar{v}] = [a_0 \quad a_1 \quad \cdots \quad a_{M-1}]^T, \quad A_0[\bar{\psi}] = [b_0 \quad b_1 \quad \cdots \quad b_{M-1}]^T.
\]

If we let

\[
- \frac{d^2 v(x)}{dx^2} := \sum_{k=0}^{M-2} \beta_k L_k(x),
\]

then by orthogonality, we have

\[
(3.10) \quad \int_{-1}^{1} - \frac{d^2 v(x)}{dx^2} \phi(x) dx = \sum_{k=0}^{M-2} \beta_k \phi_k \frac{2}{2k+1},
\]

\[
\int_{-1}^{1} \frac{dv(x)}{dx} \frac{d\psi(x)}{dx} dx = \sum_{k=0}^{M-1} a_k \beta_k \frac{2}{2k+1}.
\]

We define the matrices \( Q \) and \( \bar{Q} \) by

\[
Q := \begin{bmatrix}
0 & 0 & 0 & 0 & \cdots & 0 \\
\frac{2}{3} & 0 & 0 & \cdots & 0 \\
0 & \frac{2}{3} & 0 & \cdots & 0 \\
0 & 0 & \frac{2}{5} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & \cdots & \frac{2}{2M-1}
\end{bmatrix}_{M \times M},
\]

\[
\bar{Q} := \begin{bmatrix}
0 & 0 & 0 & 0 & \cdots & 0 \\
0 & \frac{2}{3} & 0 & 0 & \cdots & 0 \\
0 & 0 & \frac{2}{5} & 0 & \cdots & 0 \\
0 & 0 & 0 & \frac{2}{7} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & \cdots & \frac{2}{2M-1}
\end{bmatrix}_{(M-1) \times (M-1)}.
\]
From (3.8), (3.9) and (3.10), we have
\[
\begin{bmatrix}
v_0^T \\
v_1^T \\
\vdots^T \\
v_{M-2}^T
\end{bmatrix}
A_0^T QA_0
\begin{bmatrix}
\psi_0 \\
\psi_1 \\
\vdots \\
\psi_{M-2}
\end{bmatrix}
= \begin{bmatrix}
\beta_0^T \\
\beta_1^T \\
\vdots^T \\
\beta_{M-2}^T
\end{bmatrix}
\begin{bmatrix}
\phi_0 \\
\phi_1 \\
\vdots \\
\phi_{M-2}
\end{bmatrix}
.
\]

Since
\[B[\tilde{v}] = [\beta_0 \beta_1 \cdots \beta_{M-2}]^T,\]
we have
\[
\begin{bmatrix}
v_0^T \\
v_1^T \\
\vdots^T \\
v_{M-2}^T
\end{bmatrix}
A_0^T QA_0
\begin{bmatrix}
\psi_0 \\
\psi_1 \\
\vdots \\
\psi_{M-2}
\end{bmatrix}
= [B]
\begin{bmatrix}
\psi_0 \\
\psi_1 \\
\vdots \\
\psi_{M-2}
\end{bmatrix}
\begin{bmatrix}
\phi_0 \\
\phi_1 \\
\vdots \\
\phi_{M-2}
\end{bmatrix}
= [B]^T \tilde{Q}
\begin{bmatrix}
\phi_0 \\
\phi_1 \\
\vdots \\
\phi_{M-2}
\end{bmatrix}.
\]
Since \(v(x)\) and \(\phi(x)\) are arbitrary, and \(\psi_k = \phi_k, 0 \leq k \leq M - 2\), we obtain
\[B^T \tilde{Q} = A_0^T QA_0,\]
so that
\[B = \tilde{Q}^{-1} A_0^T QA_0.\]

Now we are ready to describe the problem (3.3) into a linear system. Let \(u_M(x, y) := \sum_{k=0}^{M} \sum_{l=0}^{M} u_{kl} L_k(x) L_l(y)\) with \(u_{00} = 0\), be the tau solution of the problem (3.3). Let \(\tilde{U}\) and \(U\) be defined by
\[
\tilde{U} = \begin{bmatrix}
u_0 & \cdots & v_{0M} \\
\vdots & \ddots & \vdots \\
u_{M0} & \cdots & v_{MM}
\end{bmatrix},
U = \begin{bmatrix}
u_0 & \cdots & v_{0, M-2} \\
\vdots & \ddots & \vdots \\
u_{M-2, 0} & \cdots & v_{M-2, M-2}
\end{bmatrix}.
\]
hence, from the orthogonal property of Legendre polynomials, we have

\[ P_{M-2}(-\Delta u_M(x, y)) = \sum_{k=0}^{M-2} \sum_{l=0}^{M-2} \alpha_{kl} L_k(x) L_l(y), \]

where

\[
\begin{bmatrix}
\alpha_{00} & \cdots & \alpha_{0,M-2} \\
\vdots & \ddots & \vdots \\
\alpha_{M-2,0} & \cdots & \alpha_{M-2,M-2}
\end{bmatrix} = B\tilde{U} + \tilde{U}B^T.
\]

Let \( P_{M-2} f(x, y) := \sum_{k=0}^{M-2} \sum_{l=0}^{M-2} f_{kl} L_k(x) L_l(y) \) and let \( f_{00} = 0 \). From (3.3) and (3.12), we have a spectral tau solver for the two-dimensional Poisson equation:

\[ B\tilde{U} + \tilde{U}B^T = F, \]

where

\[ F := [f_{kl}]_{(M-1) \times (M-1)}. \]

We now present our diagonalization technique based on Schur decomposition for the linear system (3.13). The successful implementation requires the previous procedures to keep up the merits and to avoid the faults of a matrix diagonalization and Schur decomposition. Let \( H = A_0^T Q A_0 \), so that \( H \) is a symmetric matrix. Then (3.13) can be expressed as

\[ [\tilde{Q}^{-1} H] \tilde{U} + \tilde{U} [\tilde{Q}^{-1} H]^T = F. \]

Let

\[ \tilde{H} := \tilde{Q}^{-1} H \tilde{Q}^{-1}, \quad \tilde{U} := \tilde{Q}^{1/2} U \tilde{Q}^{1/2}, \quad \tilde{F} := \tilde{Q}^{1/2} F \tilde{Q}^{1/2}, \]

and multiply both sides of (3.14) by \( \tilde{Q}^{1/2} \), then we have

\[ \tilde{H} \tilde{U} + \tilde{U} \tilde{H} = \tilde{F}. \]

By construction of \( \tilde{H} \), there exist an orthogonal matrix \( V \) and a diagonal matrix \( D = \text{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{M-1}) \), \( \lambda_1 = 0, \lambda_i > 0, i = 2, \ldots, M - 1 \), such that \( \tilde{H} V = V D \). If we let

\[ \tilde{U} := V^T \tilde{U}, \quad \tilde{F} := V^T \tilde{F}, \]
(3.15) becomes

\[ D\tilde{u} + \tilde{U}D = \tilde{F}. \]

Let \( \tilde{F} := [f_{kl}]_{(M-1) \times (M-1)} \) and \( \tilde{U} = [\tilde{u}_{kl}] \), from (3.16), we have the following relation:

\[
\begin{cases}
\tilde{u}_{kl} = 0 & \text{if } k = l = 0, \\
\tilde{u}_{kl} = \frac{j_{kl}}{\lambda_{k+\lambda_l}} & \text{otherwise}. 
\end{cases}
\]

Since \( \tilde{U} = V\tilde{U}V^T \) and \( \tilde{U} = \bar{Q}^{-\frac{1}{2}}\tilde{U}\bar{Q}^{-\frac{1}{2}} \), the coefficients \( u_{kl}, 0 \leq k, l \leq M - 2 \), are obtained through the following matrix multiplication only without solving the linear system (3.13) or (3.15).

\[ \tilde{U} = \bar{Q}^{-\frac{1}{2}}V\tilde{U}V^T \bar{Q}^{-\frac{1}{2}} = [\bar{Q}^{-\frac{1}{2}}V]\tilde{U}^{\frac{1}{2}}V^T. \]

Let \( C \) be the matrix representing the discrete operator \( T^*_M \) of \( T^n \). To obtain the matrix \( C \), we extend the tau solver based on our diagonalization technique for the homogeneous case to the case with inhomogeneous boundary condition. The matrix \( C \) consists of four column blocks. To obtain, for example, the \( n \)th column of the first block, we consider the following problem:

\[ \begin{cases}
\Delta u = 0 & \text{in } \Omega, \\
\frac{\partial u}{\partial n} = L_n(x)\chi_{\Gamma_1} & \text{on } \Gamma.
\end{cases} \]

Let \( u_M := \sum_{k=0}^{M} \sum_{l=0}^{M} u_{kl}L_k(x)L_l(y) \) be the approximate solution for the problem (3.18). Since the test functions do not satisfy the boundary conditions individually,
the weighted residual conditions corresponding to the boundary conditions are

\[
\begin{align*}
    u_{k, M-1} \frac{M(M-1)}{2} &= - \sum_{l=1, l \text{ odd}}^{M-3} u_{kl} \frac{l(l+1)}{2} \quad \text{for } k = 0, \ldots, M, \ k \neq n, \\
    u_{k, M} \frac{M(M+1)}{2} &= - \sum_{l=2, l \text{ even}}^{M-2} u_{kl} \frac{l(l+1)}{2} \quad \text{for } k = 0, \ldots, M, \ k \neq n, \\
    u_{M-1, l} \frac{M(M-1)}{2} &= - \sum_{k=1, k \text{ odd}}^{M-3} u_{kl} \frac{k(k+1)}{2} \quad \text{for } l = 0, \ldots, M, \\
    u_{M, M} \frac{M(M+1)}{2} &= - \sum_{k=2, k \text{ even}}^{M-2} u_{kl} \frac{k(k+1)}{2} \quad \text{for } l = 0, \ldots, M, \\
    u_{n, M-1} \frac{M(M-1)}{2} &= \frac{1}{2} - \sum_{l=1, l \text{ odd}}^{N-3} u_{nl} \frac{l(l+1)}{2}, \\
    u_{n, M} \frac{M(M+1)}{2} &= -\frac{1}{2} - \sum_{l=2, l \text{ even}}^{N-2} u_{nl} \frac{l(l+1)}{2}. 
\end{align*}
\]

(3.19)

Let

\[
u_n^2 = \frac{1}{M(M-1)} L_n(x) L_{M-1}(y) - \frac{1}{M(M-1)} L_n(x) L_M(y) \in L_0^2(\Omega),
\]

and \(u_M^1\) be the Legendre tau approximate solution of the following problem:

\[
\begin{align*}
    -\Delta u^1 = \Delta u_M^2 & \quad \text{in } \Omega, \\
    \frac{\partial u^1}{\partial n} = 0 & \quad \text{on } \Gamma.
\end{align*}
\]

(3.20)

Then \(u_M^1 \in S_{M-1}^{1,0}(\Omega) \cap L_0^2(\Omega)\) satisfies

\[
\begin{align*}
    -\Delta u_M^1, \phi &= (\Delta u_M^2, \phi) \quad \text{for } \phi \in S_{M-2}(\Omega), \\
\end{align*}
\]

(3.21)

and for \(n = M - 1, M,\)

\[
\begin{align*}
    (\Delta u_M^2, \phi) &= 0 \quad \text{for } \phi \in S_{M-2}(\Omega).
\end{align*}
\]

(3.22)

From (3.22), we have a trivial solution \(u_M^1 = 0\) for \(n = M - 1, M\). To avoid the trivial solution, for \(N \leq M - 2\), we define the boundary element space \(B_N \subset L_0^2(\Gamma)\) on \(\Gamma\) by

\[
B_N := \{L_k(x)\chi_{\Gamma_1}, L_l(y)\chi_{\Gamma_2} \mid k, l = 1, \ldots, N, i = 1, 3, j = 2, 4\}.
\]

Let \(P_N^B\) be the \(L^2\)-projection operator from \(L^2(\Gamma)\) to \(B_N\). From the definition of the boundary element space \(B_N,\)

\[
P_N^B \gamma_1(u_M^1 + u_M^2) = P_N^B \gamma_1(u_M^2) = L_n(x) \chi_{\Gamma_1},
\]

(3.23)
From (3.21) and (3.23), \( u_1^M + u_2^M \) becomes the spectral Legendre tau approximate solution of the problem (3.18); \( u_M = u_1^M + u_2^M \). Similarly, we can construct the other columns and those of other blocks of the matrix \( C \).

Now we are ready to describe the discretized formulations for (2.6), (2.8) and (2.10) by using the Laplace solver for the homogeneous and inhomogeneous boundary conditions. Define the discrete operator \( (-\Delta_n)^{-1}_M \) of \( (-\Delta_n)^{-1} \) by

\[
(-\Delta_n)^{-1}_M(f) := u_M \text{ for } f \in L^2_0(\Omega),
\]

where \( u_M \in S^1_0(\Omega) \cap L^2_0(\Omega) \) satisfies:

\[
(-\Delta u_M, \phi) = (f, \phi) \text{ for } \phi \in S_{M-2}(\Omega),
\]

and define the discrete operator \( T^*_M \) of \( T^* \) by

\[
T^*_M \lambda_N := (-\Delta_n)^{-1}_M(\Delta u_2) + u_2 \text{ for } \lambda_N \in B_N,
\]

where \( u_1 \in S^1_0(\Omega) \cap L^2_0(\Omega) \) and \( u_2 \in S_0(\Omega) \cap L^2_0(\Omega) \) satisfy

\[
\begin{cases}
(-\Delta u_1, \phi) = (\Delta u_2, \phi) \text{ for } \phi \in S_{M-2}(\Omega), \\
P^*_N \gamma_1(u_2) = \lambda_N.
\end{cases}
\]

Therefore, we consider the following discretized formulations for the discrete normal derivative \( p_N \) of the vorticity, the discrete vorticity \( \omega_M \) and stream function \( \psi_M \):

\[
\begin{aligned}
\text{find } p_N &\in B_N \text{ such that } \\
(T^*_M p_N, T^*_M \lambda_N) &= -\left(\frac{1}{\rho}(-\Delta_n)^{-1}_M(f_2), T^*_M \lambda_N\right) + \sum_{i=1}^4 \int_{\Gamma_i} h_i \lambda_N \ ds \\
&\quad \text{for any } \lambda_N \in B_N, \\
\text{find } \omega_M &\in S_{M-2} \text{ such that } \\
(\omega_M, \phi) &= (T^*_M p_N, \phi) + \left(\frac{1}{\rho}(-\Delta_n)^{-1}_M(f_2), \phi\right) \text{ for any } \phi \in S_{M-2}(\Omega), \\
\text{find } \psi_M &\in S_M(\Omega) \text{ such that } \\
\psi_M &= (-\Delta_n)^{-1}_M(\omega_M) + \left(\frac{1}{\rho} \sum_{i=1}^4 \int_{\Gamma_i} h_i \ ds - \int_{\Gamma_1} h_i \ ds\right) \left(-\Delta_n)^{-1}_M(\omega_M) \right) \ ds).
\end{aligned}
\]

4 Numerical results

The first example is the case with \( u = v = 0 \) on the sides \( \Gamma_1, \Gamma_2, \Gamma_4 \), and \( u = -(x + 1)^2(x-1)^2, v = 0 \) on \( \Gamma_3 \) (see Figure 1 for \( \Gamma_i \)). This problem has a regular solution. Table
1 shows the convergence for the discrete boundary value \( q_N \), the discrete vorticity \( \omega_M \) and the discrete stream function \( \psi_M \) for \( \text{curl} f = 0 \). In Figures 2.1 and 2.2, streamlines and vector fields for this case are shown for \( M = 16, N = 12 \).

Table 1. Convergence of \( \omega_M, \psi_M \) and \( q_N \): \( \text{curl} f = 0, u = v = 0 \) on \( \Gamma_1, \Gamma_2, \Gamma_4 \), 
\[ u = -(x + 1)^2(x - 1)^2, \quad v = 0 \) on \( \Gamma_3 \), and \( M_i = N_i + 4, i = 1, 2 \).

<table>
<thead>
<tr>
<th>( M_1 = 20 )</th>
<th>( M_2 = 24 )</th>
<th>( |\omega_{M_1} - \omega_{M_2}|_{L^2(\Omega)} )</th>
<th>( |\psi_{M_1} - \psi_{M_2}|_{L^2(\Omega)} )</th>
<th>( |q_{N_1} - q_{N_2}|_{L^2(\Gamma)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1061e-3</td>
<td>0.3307e-7</td>
<td>0.4807e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0436e-3</td>
<td>0.0636e-7</td>
<td>0.1488e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0208e-3</td>
<td>0.0265e-7</td>
<td>0.1114e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0110e-3</td>
<td>0.0113e-7</td>
<td>0.0739e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0049e-3</td>
<td>0.0047e-7</td>
<td>0.0494e-3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The second example is the case with \( \psi = 1, \frac{\partial \psi}{\partial n} = 0 \) on \( \Gamma \) and \( \text{curl} f = L_2(x)L_2(y) \), where \( L_2(x) \) and \( L_2(y) \) are the second order Legendre polynomials. Table 2 shows the convergence for the discrete normal derivative \( p_N \), the discrete vorticity \( \omega_M \) and the discrete stream function \( \psi_M \). In Figures 3.1 and 3.2, streamlines and vector fields for the second example are shown for \( M = 16, N = 12 \).

Table 2. Convergence of \( \omega_M, \psi_M \) and \( p_N \): \( \text{curl} f = L_2(x)L_2(y), \psi = 1, \frac{\partial \psi}{\partial n} = 0 \) on \( \Gamma \) and \( M_i = N_i + 4, i = 1, 2 \).

<table>
<thead>
<tr>
<th>( M_1 = 20 )</th>
<th>( M_2 = 24 )</th>
<th>( |\omega_{M_1} - \omega_{M_2}|_{L^2(\Omega)} )</th>
<th>( |\psi_{M_1} - \psi_{M_2}|_{L^2(\Omega)} )</th>
<th>( |p_{N_1} - p_{N_2}|_{L^2(\Gamma)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1885e-5</td>
<td>0.8187e-6</td>
<td>0.9586e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0625e-5</td>
<td>0.3614e-6</td>
<td>0.6143e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0292e-5</td>
<td>0.1668e-6</td>
<td>0.3665e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0190e-5</td>
<td>0.0842e-6</td>
<td>0.2231e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0131e-5</td>
<td>0.0458e-6</td>
<td>0.1460e-3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The third example is the case with \( u = -1 \) on \( \Gamma_3 \). Other boundary conditions are the same as in the first example. We do not expect that \( \psi \) belongs to \( H^2(\Omega) \),
because of singularities at the two top corners. However, for the discretized Stokes and Navier-Stokes problems, we expect a $H^2(\Omega)$ solution due to implementation of boundary conditions. This means that $u$ drops to zero continuously along the vertical sides $x = -1, x = 1$ from $y = 1$ to $y = 1 - \varepsilon$ for a small $\varepsilon > 0$. Here, we chose $\varepsilon = 0.15$ for this example. Physically, this smoothing indicates that a small amount of fluid enters the cavity through the point (1, 1) and the same amount of fluid leaves the cavity through the point (−1, 1). This guarantees that the compatibility condition $\int_{\Gamma} \mathbf{u} \cdot \mathbf{n} ds = 0$ is satisfied. In Figures 4.1 and 4.2, streamlines and vector fields for the third example are shown for $M = 16, N = 12$.

Figure 1. Geometry of the driven cavity problem.
Figure 2.1. Contour of the stream lines (case 1).

Figure 2.2. Velocity vector fields (case 1).
Figure 3.1. Contour of the stream lines (case 2).

Figure 3.2. Velocity vector fields (case 2).
Figure 4.1. Contour of the stream lines (case 3).

Figure 4.2. Velocity vector fields (case 3).
References


Department of Mathematics
Pohang University of Science and Technology
Pohang 790-784, Korea.
Current Address: Samsung Advanced Institute of Technology
Suwon 440-600, Korea
email: srjun@radon.sait.samsung.co.kr

† Department of Mathematics
Chosun University
Kwangju 501-759 Korea
sgkang@mail.chosun.ac.kr

Department of Mathematics
Pohang University of Science and Technology
Pohang 790–784, Korea
e-mail: kwony@euclid.postech.ac.kr
Comparison of multigrid performance for higher order scheme with 5-point scheme

Mun. S. Han, Do Y. Kwak and Jun S. Lee

Abstract

We consider a multigrid algorithm for higher order finite difference scheme for the Poisson problem on rectangular domain. Several smoothers including Jacobi, Red-black Gauss-Seidel are tested and compared. Since higher order scheme gives much more accurate result then 5-point scheme, one may use small number of levels with higher order scheme and thus the overall cost is reduced quite a lot. The numerical experiment compares the two cases.

1 Introduction

The purpose of this paper is to consider the convergence of a multigrid algorithm for a higher order finite difference scheme and compare overall computational cost with five point scheme.

Multigrid methods have been proven very fast and robust for most of elliptic problems with many types of discretizations. See [1],[2],[3], for example. For a various convergence proofs, we refer to [4],[5],[6],[9],[10].

Usually, multigrid algorithm consists of a sequence $M_1 \subset \cdots \subset M_J$ of nested finite dimensional spaces, smoothing process for each level and some transfer operator between the spaces. First, a smoothing(presmoothing) is performed on the finest level, then the residual is passed onto the next coarser level, where the slowly varying component is relaxed, the process repeated until down to the coarsest grid, where the problem is solved exactly. The resulting quantity then is passed back to the finer levels, sometimes with additional smoothing(postsmoothing). Such an algorithm is typical and called a V-cycle, while other variants, such as using smoothing as we go down only(backslash-cycle) or many smoothings for each level are possible. Another variation is to use two correction step at each level. This is called W-cycle. Usually W-cycle converges faster(with more work) and easier to prove its convergence. However, V-cycle is simpler and advocated by practitioners and the convergence theory is mathematically more challenging.

AMS(MOS) subject classifications: primary 65N30, secondary 65F10.
Key words: multigrid methods, higher-order finite difference method.
This research was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Science and Technology.
Commonly used discretization method for simple elliptic problems is 5-point formula, which has second order accuracy. However to achieve very accuracy, one needs to refine the meshes, which sometimes require large memory and time. In most laboratory job, one takes maximum 8 levels, which mounts to 256 × 256 grids. This is quite a memory and time consuming for a model problem. For realistic problems, one may need more levels and the problems are coupled with other systems, which requires still faster algorithms for elliptic part. Instead, higher order scheme requires lesser number of levels to achieve the same accuracy. To resolve this problem, one adapts higher order accuracy [11] which has $h^4$ order accuracy. However, when the right hand side is smooth, one can use newly developed higher order scheme in [16] to have $h^6$ order accuracy. This is an enormous gain, since one can use only one third number of levels than five point formula to get the same accuracy.

For finite difference case, there is a common belief that the sum of the order of prolongation and restriction should be greater than 2 [10] so that one uses for example bilinear prolongation with trivial restriction [15] or variants of such algorithms. However, we believe that the restriction operator as the adjoint of prolongation is most natural. It makes the whole algorithm symmetric and also fits the theory developed by Bramble et al. [9], [8]. The performance of such nonsymmetric operator pairs will be shown in later section.

2 Multigrid Method for 9 point compact scheme

In this section, we briefly describe 9 point compact scheme and introduce multigrid algorithm for this scheme. We first consider the following Poisson Equation:

$$\begin{cases}
-\Delta u(x, y) = f(x, y) \text{ in } \Omega \\
u(x, y) = g(x, y) \text{ on } \partial \Omega.
\end{cases} \quad (2.1)$$

Here, $\Omega$ can be any region in $\mathbb{R}^2$ covered by squares. For simplicity, we assume $\Omega$ is the unit square. For $k = 1, 2, \ldots, J$, let $h_k = 2^{-k}$ be a mesh size of level $k$. Define $\Omega_k$ be a space of points $(x_i, y_j) = (ih_k, jh_k)$ for $i, j = 0, 1, \ldots, 2^k$ and $V_k$ be a vector space of function evaluated at $\Omega_k$. A fourth-order compact scheme is written as follows:

$$\begin{align*}
\frac{1}{6h^2} & \left[ 20u_{i,j} - 4(u_{i-1,j} + u_{i,j-1} + u_{i+1,j} + u_{i,j+1}) \\
& - (u_{i-1,j-1} + u_{i+1,j-1} + u_{i+1,j+1} + u_{i-1,j+1}) \right] \\
& = \frac{1}{12} \left[ f_{i+1,j} + f_{i-1,j} + f_{i,j+1} + f_{i,j-1} + f_{i-1,j+1} + f_{i+1,j-1} + 8f_{i,j} \right].
\end{align*} \quad (2.2)$$

When the solution of the equation (2.1) is sufficiently smooth (having continuous sixth-order partial derivatives) one may use the newly developed scheme, whose convergence rate is of $O(h^8)$. The stencil of this high-order finite difference scheme is written as
follows[16]:
\[
\frac{1}{6h^2} \left[ 20u_{i,j} - 4(u_{i-1,j} + u_{i,j-1} + u_{i+1,j} + u_{i,j+1}) \\
- (u_{i-1,j-1} + u_{i+1,j-1} + u_{i+1,j+1} + u_{i-1,j+1}) \right]
- \left[ 1 + \frac{2h^2}{4!} \Delta + \frac{2h^4}{6!} \left( \Delta^2 + 2 \frac{\partial^4}{\partial x^2 \partial y^2} \right) \right] f_{i,j}
\]
\[ (2.3) \]
where \( h = h_k \). This nine-point discretization gives a truncation error of \( O(h^8) \) over a square mesh (i.e., a convergence rate of \( O(h^6) \)). We obtain a system of linear equation of the form
\[
A_k u = f,
\]
(2.4)
where \( A_k \) is a sparse, \( n \times n \), symmetric, positive definite matrix and \( u \) is the vector whose entries are \( u_{i,j} \), and \( f \) is the vector whose entries are \( f(x_i, y_j) \).

To describe the multigrid algorithm for this problem, we need certain intergrid transfer operators (called prolongation) between two grids. Assuming we are given a certain prolongation operator \( I^k_{k-1} : V_{k-1} \rightarrow V_k \), we define the restriction operator \( I^{k-1}_k : V_k \rightarrow V_{k-1} \) as its adjoint with respect to \( (\cdot, \cdot) \):
\[
(I^{k-1}_k u, v)_{k-1} = (u, I^k_{k-1} v)_k \quad \forall u \in V_k, \forall v \in V_{k-1}.
\]
Now the multigrid algorithm for solving (2.4) is defined as follows:

**Multigrid Algorithm. W(m,m)** Set \( B_1 = A_1^{-1} \). For \( 1 < k \leq J \), assume that \( B_{k-1} \) has been defined and define \( B_k f \) for \( f \in V_k \) as follows:

1. Set \( x^0 = 0 \) and \( q^0 = 0 \).

2. Define \( x^l \) for \( l = 1, \ldots, m \) by
\[
x^l = x^{l-1} + R^l_{k+m}(f - A_k x^{l-1}).
\]

3. Define \( y^m = x^m + I_k q^m \), where \( q^l \) is defined by
\[
q^l = q^{l-1} + B_{k-1} \left[ P_{k-1}^l (f - A_k x^m) - A_{k-1} q^{l-1} \right] \quad i = 1, \ldots, p
\]

4. Define \( y^l \) for \( l = m + 1, \ldots, 2m \) by
\[
y^l = y^{l-1} + R^l_{k+m}(f - A_k y^{l-1}).
\]

5. Set \( B_k f = y^{2m} \).

For comparison, we shall also consider multigrid algorithm with nonsymmetric operators. We will not repeat the algorithm here, instead, they will be considered in later section.
Fix $k$. Let $u_{i,j}, u_{i+1,j}, u_{i,j+1}, u_{i+1,j+1}$ be points of level $k-1$ and $E_{i,j}$ be a cell having them as its vertex. (We borrow the term "cell" from cell-centered method.) Let $u_{i,j}^1, u_{i,j}^2, u_{i,j}^3, u_{i,j}^4$ and $u_{i,j}^5$ be points of level $k$ defined as in Figure 1. Note that $u_{i,j}^1 = u_{i,j-1}^1$, $u_{i,j}^2 = u_{i+1,j}^1$, $u_{i,j}^3 = u_{i,j+1}^1$ and $u_{i,j}^4 = u_{i-1,j}^1$. Divide $E_{i,j}$ into four subcells, labeling them counterclockwise as $e_{i,j}^1, e_{i,j}^2, e_{i,j}^3, e_{i,j}^4$ at level $k$. (See Figure 1.)

Now, we define the prolongation operator $I_{k-1}^k$ be bilinear interpolation of four points $u_{i,j}, u_{i+1,j}, u_{i,j+1}$ and $u_{i+1,j+1}$. First, $u_{i,j}, u_{i+1,j}, u_{i,j+1}$ and $u_{i+1,j+1}$ of level $k$ are the same value of level $k-1$ respectively. The mid points $u_{i,j}^1, u_{i,j}^2, u_{i,j}^3$ and $u_{i,j}^4$ can be written as follows:

\[
\begin{align*}
    u_{i,j}^1 &= \frac{u_{i,j} + u_{i+1,j}}{2}, &
    u_{i,j}^2 &= \frac{u_{i,j} + u_{i+1,j+1}}{2}, \\
    u_{i,j}^3 &= \frac{u_{i,j+1} + u_{i+1,j+1}}{2}, &
    u_{i,j}^4 &= \frac{u_{i,j} + u_{i+1,j+1}}{2}.
\end{align*}
\]  

(2.5)

The value of center $u_{i,j}^5$ is the average of $u_{i,j}, u_{i+1,j}, u_{i,j+1}$ and $u_{i+1,j+1}$:

\[
    u_{i,j}^5 = \frac{u_{i,j} + u_{i+1,j} + u_{i,j+1} + u_{i+1,j+1}}{4}.
\]  

(2.6)

**Theorem 2.1** We have

\[
    A_k(I_{k-1}^k u, I_{k-1}^k u) \leq C A_{k-1}(u, u), \quad \forall u \in V_{k-1}.
\]  

(2.7)
Proof 1 Using the symmetry, we have

\[
A_{k-1}(u, u) = \frac{1}{6} \sum_{i,j}^{2^{k-1}} [20u_{i,j} - 4(u_{i-1,j} + u_{i,j-1} + u_{i+1,j} + u_{i,j+1})
- (u_{i-1,j-1} + u_{i+1,j-1} + u_{i+1,j+1} + u_{i-1,j+1})] u_{i,j}
\]

\[
= \frac{4}{6} \sum_{i,j}^{2^{k-1}} [(u_{i,j} - u_{i-1,j})^2 + (u_{i,j} - u_{i,j-1})^2
+ (u_{i,j} - u_{i+1,j})^2 + (u_{i,j} - u_{i,j+1})^2]
\]

\[
= \frac{1}{6} \sum_{i,j}^{2^{k-1}} [(u_{i,j} - u_{i-1,j-1})^2 + (u_{i,j} - u_{i-1,j+1})^2
+ (u_{i,j} - u_{i+1,j-1})^2 + (u_{i,j} - u_{i+1,j+1})^2].
\]

Set \( v = P_{k-1}^h u \) then, similarly, we have

\[
A_k(v, v) = \frac{4}{6} \sum_{m,n}^{2^k} [(v_{m,n} - v_{m-1,n})^2 + (v_{m,n} - v_{m,n-1})^2
+ (v_{m,n} - v_{m,n+1})^2 + (v_{m,n} - v_{m,n-1})^2]
\]

\[
= \frac{1}{6} \sum_{m,n}^{2^k} [(v_{m,n} - v_{m-1,n-1})^2 + (v_{m,n} - v_{m-1,n+1})^2
+ (v_{m,n} - v_{m+1,n-1})^2 + (v_{m,n} - v_{m+1,n+1})^2].
\]

Each terms in (2.9) is bounded by the terms in (2.8). For example, if \( m = 2i \) and \( n = 2j \) then \( v_{m,n} = u_{i,j} \) and \( v_{m,n-1} = (u_{i,j} - u_{i,j-1})/2 \) by the definition of our interpolation. So we have

\[
(v_{m,n} - v_{m,n-1})^2 \leq (u_{i,j} - u_{i,j-1})^2/4.
\]

Note that the difference \( (u_{i,j} - u_{i,j-1})^2 \) appears finite times (at most 10 times). Thus there exists a positive constant \( C \) satisfying the inequality (2.7).

To prove multigrid convergence theory, we need following property, so-called, “approximation and regularity”: There exist a number \( 0 < \alpha \leq 1 \) and a constant \( C_\alpha \) such that for all \( k = 1, \cdots, J \),

\[
A_k((I - P_{k-1}^h)u, u) \leq C_\alpha \left( \frac{\|A_ku\|_k^2}{\lambda_k} \right)^\alpha A_k(u, u)^{1-\alpha}, \quad \forall u \in V_k.
\]

Here, \( \lambda_k \) is the largest eigenvalue of \( A_k \) and \( P_{k-1}^h \) is the elliptic projection defined by

\[
A_{k-1}(P_{k-1}^h u, v) = A_k(u, P_{k-1}^h v), \quad \forall u \in V_k, v \in V_{k-1}.
\]
\begin{tabular}{cccccc}
\hline
$h_J$ & $\lambda_{\text{min}}$ & $\lambda_{\text{max}}$ & $K$ & $\delta$ \\
\hline
1/16 & 0.826 & 0.999 & 1.211 & 0.030 \\
1/32 & 0.823 & 0.999 & 1.215 & 0.030 \\
1/64 & 0.823 & 0.999 & 1.215 & 0.031 \\
1/128 & 0.823 & 0.999 & 1.216 & 0.031 \\
\hline
\end{tabular}

Table 1: high-order scheme with Gauss-Seidel smoothing and 9-point interpolation

The following result can be proved as in [9].

\textbf{Lemma 2.1} Let the operator $P_{k-1}$ be defined by (2.12). Then (2.11) holds for $\alpha = \frac{1}{2}$.

With these preliminaries, we can prove the W-cycle result by the framework of [9].

\textbf{Theorem 2.2} Let $E_k = I - B_k A_k$ in algorithm $W(m,m)$. Then we have

$$A_k(E_k u, u) \leq \delta_k A_k(u, u), \quad \forall u \in V_k,$$

(2.13)

where $\delta_k < 1$.

\section{Numerical Experiments}

We consider the following problem on the unit square:

$$-\nabla \cdot p \nabla \tilde{u} = f \quad \text{in } \Omega = (0,1)^2,$$

$$\tilde{u} = 0 \quad \text{on } \partial\Omega.$$  

(3.1)

First, we report the maximum, minimum eigenvalues, condition numbers and contractions of both algorithms with Gauss-Seidel smoothing and 9-point interpolation. Numerical experiment shows that multigrid algorithm of high-order scheme converges faster than that of 5-point scheme. Both algorithms contract independent of the mesh size $h$ and number of levels $J$.

Next, we report the number of iteration and discrete $L^2$ errors of test problem.

\section*{References}


<table>
<thead>
<tr>
<th>$h_J$</th>
<th>$\lambda_{\min}$</th>
<th>$\lambda_{\max}$</th>
<th>$K$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/16</td>
<td>0.776</td>
<td>0.999</td>
<td>1.290</td>
<td>0.050</td>
</tr>
<tr>
<td>1/32</td>
<td>0.765</td>
<td>0.999</td>
<td>1.307</td>
<td>0.055</td>
</tr>
<tr>
<td>1/64</td>
<td>0.761</td>
<td>0.999</td>
<td>1.314</td>
<td>0.057</td>
</tr>
<tr>
<td>1/128</td>
<td>0.756</td>
<td>0.999</td>
<td>1.321</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Table 2: 5-point scheme with Gauss-Seidel smoothing and 9-point interpolation

<table>
<thead>
<tr>
<th>$h_J$</th>
<th>high-order scheme</th>
<th>5-pt scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high-order scheme</td>
<td>5-pt scheme</td>
</tr>
<tr>
<td></td>
<td>$L^2$ error</td>
<td>$L^2$ error</td>
</tr>
<tr>
<td>1/16</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1.47(-06)</td>
<td>7.93(-04)</td>
</tr>
<tr>
<td>1/32</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>9.22(-08)</td>
<td>1.98(-04)</td>
</tr>
<tr>
<td>1/64</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>5.77(-09)</td>
<td>4.95(-05)</td>
</tr>
<tr>
<td>1/128</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>3.65(-10)</td>
<td>1.24(-05)</td>
</tr>
</tbody>
</table>

Table 3: high-order scheme with Gauss-Seidel smoothing and 9-point interpolation


ETRI, Taejon, Korea.
email: msh@etri.re.kr

Department of Mathematics
KAIST, Taejon, Korea, 305-701
email: dykwak@math.kaist.ac.kr

Department of Mathematics
KAIST, Taejon, Korea 305-701
email: jslee@math.kaist.ac.kr
FATIGUE ANALYSIS OF ELECTROMYOGRAPHIC SIGNAL BASED ON STATIONARY WAVELET TRANSFORM

Young Seock Lee, Jin Lee

Abstract

As muscular contraction is sustained, the Fourier spectrum of the myoelectric signal is shifted toward the lower frequency. This spectral density is associated with muscle fatigue. This paper describes a quantitative measurement method that performs the measurement of localized muscle fatigue by tracking changes of median frequency based on stationary wavelet transform. Applying to the human masseter muscle, the proposed method offers the much information for muscle fatigue, comparing with the conventional FFT-based method for muscle fatigue measurement.

1. Muscle fatigue analysis

When a muscle contraction continuously at a high force, it is seen that the frequency content of its electromyographic(EMG) signal gradually shifts to lower frequencies in the course of several minutes. This effect is commonly associated with the fatigue that manifests at same time[1]. It is generally accepted that the mean or median frequency of the power spectral density function effectively track the spectral scaling due to fatigue. The fatigue analysis of EMG signal offers much information about human muscle mechanics such as the firing rate of the motor unit, inter-pulse interval of impulse function which represents the time events of motor unit action potentials(MUAPs), and the frequency contents of time-dependent MUAPs and fiber conduction velocity(FCV). Especially significant parameter linked to FCV is the median frequency of the EMG spectrum \( f_m \), which is defined as the frequency that divides the spectrum into equal areas as

\[
\frac{\int_{0}^{f_m} G_e(f) df}{\int_{0}^{\infty} G_e(f) df} = \frac{1}{2}
\]

Keywords : Electromyographic signal, Stationary wavelet transform, Muscle fatigue, Median frequency

This research was supported by the Chungwoon university research fund in 1998.
and it can be computed directly by fast Fourier transform (FFT) [2]. In fact, traditional spectral fatigue estimation requires the signal to be wide sense stationary within observation window. When muscle contractions depart from this constraint, as during dynamic contraction, the stationarity of the EMG signals can no longer be assumed.

Although there are numerous situations in which isometric and constant force contraction paradigm is rather uncommon in most daily muscle activities [3]. In this paper, we propose a different possible approach to spectral analysis of nonstationary EMG signal using stationary wavelet transform which is not the decimation of coefficient sequences at each stage and its property offers the much finer time resolution comparing with standard wavelet transform.

II. Stationary wavelet transform procedure

Discrete wavelet transform

Suppose we are given a sequence \( \{c_i\} = c_0, c_1, \ldots, c_{N-1} \) where \( N = 2^J \) for some integer \( J \). The standard discrete wavelet transform (DWT) is based on filters \( H \) and \( G \) and on binary decimation operator \( D_0 \). The filter \( H \) is a low pass filter, defined by a sequence conventionally denoted \( \{h_n\} \) which is the discrete version of continuous mother wavelet \( \phi(t) \). Typically, only a small number of the \( \{h_n\} \) are non-zero. The action of the low pass filter on a doubly finite sequence \( \{\ldots, x_{-1}, x_0, x_1, x_2, \ldots\} \) is defined by

\[
(Hx)_k = \sum_n h_{n-k} x_n
\]  

(2)

The definitions for sequences of finite length depend on a choice treatment at the boundaries. In this paper periodic boundary conditions will be used.

The filter assumed to satisfy the internal orthogonality relation

\[
\sum_n h_n h_{n+2j} = 0
\]  

(3)

for all integers \( j \neq 0 \), and to have sum of squares \( \sum_n h_n^2 = 1 \). The high pass filter \( G \) is defined by the sequence

\[
G_n = (-1)^n h_{1-n}
\]  

(4)

for all integers \( n \). The filter \( G \) satisfies the same internal orthogonality relations as \( H \) and in addition that the filters obey the mutual orthogonality relation

\[
\sum_n h_n g_{n+2j} = 0
\]  

(5)
for all integers $j$. Filters constructed in this way are called quadrature mirror filters[4].

The binary decimation operator $D_0$ simply chooses every even number of a sequence, so that

$$(D_0 x)_j = x_{2j}$$

for all integers $j$. It follows the internal and mutual orthogonality properties of the quadrature mirror filters that the mapping of a sequence $x$ to the pair of sequences $(D_0 Gx, D_0 Hx)$ is an orthogonal transformation. If $x$ is a finite sequence of length $2^m$ with periodic boundary conditions applied, then each of $D_0 Gx$ and $D_0 Hx$ will be sequences of length $2^{m-1}$. The DWT is derived from a multiresolution analysis, performed as follows

$$c^j = D_0 H c^{j+1} \quad \text{and} \quad d^j = D_0 G c^{j+1}$$

for the smooth at level $j$, written $c^j$ to be the original data $c^j_n = c_n, \quad n=0,1,\ldots,2^{j-1}$ and for $j=J-1,J-2,\ldots,0$, recursively define the smooth $c^j$ at level $j$ and detail $d^j$ at level $j$. Because the mapping $(D_0 Gx, D_0 Hx)$ is an orthogonal transform it is inverted to find $c^{j+1}$ in terms $c^j$ and $d^j$ by describing the transform as matrix and taking its transpose

$$c^{j+1} = R_0(c^j, d^j) \quad \text{for all each } j$$

where $R_0$ denotes the inverse transform to recover $c^{j+1}$ from $c^j$ and $d^j$.

**DWT by even and odd number decimation**

To define stationary wavelet transform(SWT), The modified version of the standard DWT is required. In this paragraph, we describe the modified procedure of the standard DWT. Since the DWT is an orthogonal transform it corresponds to particular choice of basis for the space $\mathbb{R}^N$ in which the original sequence lies and the decomposition could equally be carried out by selecting every odd number of each sequence rather than every even number, i.e., for sequence $x$, the operator $D_1$ is defined by

$$(D_1 x)_j = x_{2j+1}$$

for all integers $j$. The mapping $(D_1 Gx, D_1 Hx)$ is still an orthogonal transform, and the multiresolution analysis can be carried out by successively applying this operation in (7)
instead of \((D_0 G x, D_0 H x)\). The results will not be same, but the overall transformation will still be an orthogonal transformation. The reconstruction can be obtained by successive application of the corresponding inverse operator, denoted \(R_1\).

Suppose that \(\varepsilon_{j-1}, \varepsilon_{j-2}, \ldots, \varepsilon_0\) is a sequence of 0's and 1's. We can then use the operator \(D_\varepsilon\) at level \(j\), and perform the reconstruction by using the corresponding sequence of operator \(R_\varepsilon\), called the \(\varepsilon\)-decimated DWT. This procedure can be obtained by considering shift operators. Let \(S\) be the shift operator defined by

\[
(Sx)_j = x_{j+1}
\]  

(10)

It is from the definitions that \(D_1 = D_0 S\) and that \(R_1 = S^{-1} R_0\). And also it is that \(SD_0 = D_0 S^2\) and that the operator \(S\) commutes with \(H\) and \(G\). Now define \(p\) to be the integer whose binary representation is \(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{j-1}\).

It can be shown that the coefficient sequences \(c^j\) and \(d^j\) yielded by the standard DWT applied to the shifted sequence \(S^p x\). To see this, consider any fixed \(j\) and let \(p_1\) and \(p_2\) be the integers with binary representations \(\varepsilon_0 \varepsilon_1 \cdots \varepsilon_{j-1}\) and \(\varepsilon_j \varepsilon_{j+1} \cdots \varepsilon_{j-1}\). In standard DWT, the sequence \(d^j = D_0 G (D_0 H)^{j-j-1} c^j\). In the \(\varepsilon\)-decimated case, the sequence is

\[
d^j = D_\varepsilon GD_\varepsilon \varepsilon H D_\varepsilon \varepsilon H \cdots D_\varepsilon \varepsilon H^j
\]

\[
= D_0 S^\varepsilon GD_0 S^\varepsilon H D_0 S^\varepsilon H \cdots D_0 S^\varepsilon H^j
\]

(11)

By repeatedly commuting the shift operators with \(G\) and \(H\) and applying the property \(SD_0 = D_0 S^2\), equation (11) is

\[
d^j = D_0 G (D_0 H)^{j-j-1} S^{p_1} c^j
\]

(12)

and applying the operator \(S^{p_1}\)

\[
S^{p_1} = D_0 G (D_0 H)^{j-j-1} S^{p} c^j
\]

(13)

since \(P = 2^{j-j-1} p_1 + p_2\). Thus \(d_j\) shifted by an amount \(p_1\) is the \(j\)th detail sequence of the standard DWT applied to the original data sequence by an amount \(P\).

It follows that the basis vectors of the \(\varepsilon\)-decimated DWT can be obtained from those of the standard DWT by applying the shift operator \(S^{P}\), and the choice of \(\varepsilon\) thus corresponds to a choice of origin with respect to which the basis function defined.
Properties of SWT

Unlike standard or $\varepsilon$-decimated DWT, the SWT do not decimate at each level and the two new sequences have the same length as the original sequence. Instead the filters at each level are modified by padding them with zeros.

Let $\mathcal{Z}$ be the operator that alternates a given sequence with zeros, so that, for all integers $j$, $(\mathcal{Z}x)_{2j} = x_j$ and $(\mathcal{Z}x)_{2j+1} = 0$. Define filters $H^{[r]}$ and $G^{[r]}$ to have weights $Z'^k$ and $Z'^g$ respectively. Thus the filter $H^{[r]}$ is obtained by inserting a zero between every adjacent pair of elements of the filter $H^{[r-1]}$, and similarly for $G^{[r]}$. It is that $H^{[r]}$ and $G^{[r]}$ commute with $\mathcal{S}$ and that

$$D_0H^{[r]} = HD_0' \quad \text{and} \quad D_0G^{[r]} = GD_0' \quad (14)$$

For the original data sequence $a^j$, the SWT is recursively defined

$$a^{j-1} = H^{[J-j]}a^j \quad \text{and} \quad b^{j-1} = G^{[J-j]}a^j \quad (15)$$

and for any given $\varepsilon$ and corresponding origin $S$ in the $\varepsilon$-decimated DWT, the detail at level $j$ are shifted version of $D_0^{j-1}S^p b_j$ and the data at level $j$ the same shifted version of $D_0^{j-1}S^p a_j$ by

$$S^{-k_i}D_0^{j-1}S^p b_j = D_0^{j-1}S^{k_i}G^{[J-j-1]}H^{[J-j-2]}...H^{[0]}c^j$$
$$= D_0GD_0HD_0^{j-2}...HS^{k_i}c^j$$
$$= D_0G(D_0H)^{j-1}S^{k_i}c^j = d^j \quad (16)$$

Equation (16) implies that the SWT contains the coefficients of the $\varepsilon$-decimated DWT for every choice of $\varepsilon$. And it means that the SWT fills in the gaps between the coefficients in any particular $\varepsilon$-decimated DWT.

For each fixed $j$, the sequence $b_k^j$ provides information about the original data at scale $2^{j-i}$ and position $k$ because that the mother wavelet $\psi(t)$ is band-limited in the frequency domain, so we can see that $b_k^j$ is a band-limited filtering of the original sequence. As $j$ varies, the frequency response function of the filter is dilated by a factor of $2^j$. And it implies that the frequencies allow through by the band-limited filter are in interval whose endpoints are proportional to $2^j$.

III. Experimental results

The conventional methods for the muscle fatigue analysis are mostly FFT-based spectral frequency estimations[5][6]. The typical procedure of the conventional method is following steps. The 1st step is the EMG signal acquisition under isometric contraction and constant
force conditions at applicable frequency band. The 2nd step divides the acquired EMG signal to the same data block and performs the FFT at each data block. The last step calculates the median frequency from the calculated FFT data of each data block and performs the linear regression of the calculated median frequencies at each data block. The problem of the FFT-based conventional method is that the frequency resolution of EMG data at each data block is decreasing, increasing the number of data block. It means that the decreased number of data block improves the frequency resolution at each data block, but diminishes the time resolution of the calculated median frequencies. Finally, the slope of linear regression curve of median frequencies is used for the muscle fatigue analysis.

To overcome these disadvantages of conventional methods, we proposed the stationary wavelet transform-based muscle fatigue analysis method. The procedure the proposed method is the same as the conventional method except that the stationary wavelet transform is applied between the 1st and the 2nd step. The stationary wavelet transform offers the same time resolution as the original data by its property, not decreasing of the corresponding frequency resolution.

So the $N$ level stationary wavelet transform of $M$-points EMG data produces the $M$-points data at each level and the FFT-based median frequencies can be calculated from the data blocks of each level.

![EMG data from the masseter muscle of human subject](image)

**Fig.1.** The acquired EMG data from the masseter muscle of human subject

This implies that the proposed method can perform the muscle fatigue analysis which is not the loss of the time and corresponding frequency resolutions.
Fig. 2. The results of 8-level stationary wavelet transform of acquired EMG data

To experiment, we acquire the EMG data from the masseter muscle of human subject which is used for masticating operation of tooth in the conditions of isometric contraction and constant force of 100% MVC (maximum voluntary contraction) for 30 seconds. Then
it is amplified 1000 times by the differential amplifier and digitalized by the sampling frequency of 1024Hz. Fig. 1 is shown the acquired EMG data. Applying the 8-level stationary wavelet transform, the acquired EMG data is divided at each level as fig. 2.

![Graphs showing data divided at different levels](image)

(a) level 8  
(b) level 7  
(c) level 6  
(d) level 5  
(e) level 4  
(f) level 3  
(g) level 2  
(h) level 1

Fig. 3. The curves of calculated median frequencies at each level

And also we calculate median frequency at each level, where the sample number of data block for calculating the median frequency is 1024 samples so that the number of block at each level is 30. The curves of calculated median frequencies are showing as Fig. 3. and Fig. 4 is the curve of median frequencies by the conventional method.
The table 1. is the results of linear regression of median frequency curve by the proposed and the conventional method.

<table>
<thead>
<tr>
<th>Level</th>
<th>Proposed method</th>
<th>Conventional method</th>
</tr>
</thead>
<tbody>
<tr>
<td>level 1</td>
<td>$y = -0.359x + 102.475$</td>
<td></td>
</tr>
<tr>
<td>level 2</td>
<td>$y = -0.433x + 114.813$</td>
<td></td>
</tr>
<tr>
<td>level 3</td>
<td>$y = -0.467x + 129.052$</td>
<td></td>
</tr>
<tr>
<td>level 4</td>
<td>$y = -0.506x + 148.018$</td>
<td>$y = -2.53x + 167.956$</td>
</tr>
<tr>
<td>level 5</td>
<td>$y = -0.806x + 178.381$</td>
<td></td>
</tr>
<tr>
<td>level 6</td>
<td>$y = -1.122x + 220.036$</td>
<td></td>
</tr>
<tr>
<td>level 7</td>
<td>$y = -1.545x + 293.514$</td>
<td></td>
</tr>
<tr>
<td>level 8</td>
<td>$y = -3.684x + 295.110$</td>
<td></td>
</tr>
</tbody>
</table>

The results of the proposed method show that the median frequencies of low frequency components of EMG signal which correspond to low level data of the stationary wavelet transform is slow changed. It means that the muscle fatigue is not affected by the low frequency components of EMG signal. But the median frequencies of high frequency components of EMG signal is fast changed and it is similar to the result by the conventional method. It means that muscle fatigue affected by high frequency components of the EMG signal.

The results of Fig.3 and Fig.4 may be interpreted in different manner as well. the conventional method can not represent the muscle fatigue in the total frequency band of EMG signal, which is only approximately represents muscle fatigue. But the proposed method offers the total muscle fatigue informations by median frequency curve at each level. So
the proposed method can be used for accurate clinical diagnosis of muscular disorders related with muscle fatigue.

IV. Conclusions

In this paper, we propose the new method for muscle fatigue analysis based on stationary wavelet transform, which offers the various spectrum information related with muscle fatigue comparing with the conventional method. And also we confirm that the proposed method can be used for muscle fatigue analysis instead of the conventional method. The advantage of the proposed method can analyze the muscle fatigue for EMG signal of all possible frequency bands but the conventional method offers approximately muscle fatigue information. As a result, the proposed method can be used for accurate clinical diagnosis of muscle fatigue instead of the conventional method.

References


Department of Electronics Engineering,
Chungwoon University,
Hongsung 350-800, Korea

*Department of Control & Instrumentation Engineering,
*Samchok National University
*Samchok 245-711, Korea
Heat Transfer from each surface for a 3-D Thermally Asymmetric Rectangular Fin

Hyung Suk Kang

Abstract

The non-dimensional convective heat losses from each surface are investigated as a function of the non-dimensional fin length, width and the ratio of upper surface Biot number to bottom surface Biot number (Bi2/Bi1) using the three-dimensional separation of variables method. Heat loss ratio in view of each surface with the variation of Bi2/Bi1 is presented. The variation of the non-dimensional temperature profile along the fin center line for a thermally asymmetric conditions is also presented.

1. Introduction

Fins are widely used to enhance the rate of heat transfer to a surrounding fluid in many industrial applications such as electronic equipments, many kind of heat exchangers, air craft and so on. As a result, a great deal of attention has been directed to fin problems and various shapes of fins have been studied. For example, Sen and Trinh (1986), Look (1988) have discussed rectangular, Burmeister (1979) and Abrate and Newnham (1995) were concerned with triangular while Kang and Look (1999) and Kraus et al. (1978) examined trapezoidal. Finally Kim and Kang (1998) presented parabolic fin and Ullmann and Kalman (1989) researched annular fins. Usually most of the studies on the fin assume that the heat transfer coefficients for all surfaces of the fin are the same. But no literature seems to be available which presents a rectangular fin with unequal heat transfer coefficients by using a three dimensional analysis.

This paper presents an analysis of heat transfer from each surface for a three dimensional thermally asymmetric rectangular fin. In this study the upper surface Biot number, Bi1, is equal to or larger than the bottom surface Biot number, Bi2 and the left surface Biot number, Bi3, is equal to or larger than the right surface Biot number, Bi4, and Bi5, at the fin tip, has various values. The non-dimensional heat losses from each surface are investigated as a function of the non-dimensional fin length, width and the Bi2/Bi1 ratio using the three-dimensional separation of variables method. Heat loss ratio in view of each surface with the variation of Bi2/Bi1 is presented. The variation of the non-dimensional

Key words: 3-D analytical method, Heat transfer coefficient, Heat loss, Biot number
temperature profile along the fin center line for a thermally asymmetric conditions is also presented. For simplicity, the root temperature and the thermal conductivity of the fin’s material are assumed constant as well as steady-state.

**Nomenclature**

- $\text{Bi}_1$ : fin upper surface Biot number, $h_1 l / k$
- $\text{Bi}_2$ : fin bottom surface Biot number, $h_2 l / k$
- $\text{Bi}_3$ : fin left surface Biot number, $h_3 l / k$
- $\text{Bi}_4$ : fin right surface Biot number, $h_4 l / k$
- $\text{Bi}_5$ : fin tip surface Biot number, $h_5 l / k$
- $h_1$ : fin upper surface heat transfer coefficient [W/m² °C]
- $h_2$ : fin bottom surface heat transfer coefficient [W/m² °C]
- $h_3$ : fin left surface heat transfer coefficient [W/m² °C]
- $h_4$ : fin right surface heat transfer coefficient [W/m² °C]
- $h_5$ : fin tip surface heat transfer coefficient [W/m² °C]
- $k$ : thermal conductivity [W/m °C]
- $l$ : one half fin height at the base [m]
- $L'$ : fin length (base to tip) [m]
- $L$ : non-dimensional fin length, $L'/l$
- $Q$ : heat loss from a rectangular fin [W]
- $Q_{\text{as}}$ : convective heat loss from upper surface [W]
- $Q_{\text{bs}}$ : convective heat loss from bottom surface [W]
- $Q_{\text{ls}}$ : convective heat loss from left surface [W]
- $Q_{\text{rs}}$ : convective heat loss from right surface [W]
- $Q_{\text{ts}}$ : convective heat loss from tip surface [W]
- $T$ : fin temperature [°C]
- $T_w$ : fin base temperature [°C]
- $T_\infty$ : ambient temperature [°C]
- $w'$ : one half fin width [m]
- $w$ : non-dimensional a half fin width, $w'/l$
- $x'$ : length directional variable [m]
- $x$ : non-dimensional length directional variable, $x'/l$
- $y'$ : height directional variable [m]
- $y$ : non-dimensional height directional variable, $y'/l$
- $z'$ : width directional variable [m]
- $z$ : non-dimensional width directional variable, $z'/l$
- $\theta_0$ : adjusted temperature, $(T_w - T_\infty)$
- $\theta$ : non-dimensional temperature, $(T - T_\infty)/(T_w - T_\infty )$
- $\lambda_m$ : eigenvalues ($m = 1, 2, 3, \ldots$)
\[ \mu_n : \text{eigenvalues (} n = 1, 2, 3, \cdots \text{)} \]
\[ \rho_{nm} : \text{eigenvalues (} = \sqrt{\lambda_m^2 + \mu_n^2} \text{)} \]

2. Three-Dimensional Analysis

Geometry of a rectangular fin with all different heat transfer coefficients is shown in Fig. 1. Three-dimensional governing differential equation under steady state for this figure is

\[ \frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} + \frac{\partial^2 \theta}{\partial z^2} = 0. \] (1)

Six boundary conditions are required to solve the equation (1). These conditions are shown as equations (2)~(7).

\[ \theta = 1 \quad \text{at} \quad x = 0 \] (2)
\[ \frac{\partial \theta}{\partial x} + Bi5 \cdot \theta = 0 \quad \text{at} \quad x = L \] (3)
\[ \frac{\partial \theta}{\partial y} + Bi1 \cdot \theta = 0 \quad \text{at} \quad y = 1 \] (4)
\[ \frac{\partial \theta}{\partial y} - Bi2 \cdot \theta = 0 \quad \text{at} \quad y = -1 \] (5)
\[ \frac{\partial \theta}{\partial z} + Bi3 \cdot \theta = 0 \quad \text{at} \quad z = w \] (6)

Fig. 1 Geometry of a thermally asymmetric rectangular fin
\[
\frac{\partial \theta}{\partial z} - Bi4 \cdot \theta = 0 \quad \text{at} \quad z = -w
\]  

(7)

The solution for the non-dimensional temperature distribution \( \theta(x, y, z) \) within the fin obtained with equations (2) \(-\) (5) is

\[
\theta(x, y, z) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} N_{nm} \cdot f(x) \cdot f(y) \cdot f(z)
\]

(8)

where

\[
N_{nm} = \frac{4 \sin(\lambda_n) \cdot \sin(\mu_m \cdot w)}{f_n \cdot g_m}
\]

(9)

\[
f(x) = \cosh(\rho_{nm} \cdot x) - C_{nm} \cdot \sinh(\rho_{nm} \cdot x)
\]

(10)

\[
C_{nm} = \frac{\rho_{nm} \cdot \tanh(\rho_{nm} \cdot L) + Bi5}{\rho_{nm} + Bi5 \cdot \tanh(\rho_{nm} \cdot L)}
\]

(11)

\[
\rho_{nm} = \sqrt{(\lambda_n^2 + \mu_m^2)}
\]

(12)

\[
f(y) = \cos(\lambda_n \cdot y) + A_n \cdot \sin(\lambda_n \cdot y)
\]

(13)

\[
A_n = \frac{\lambda_n \cdot \tan \lambda_n - Bi1}{\lambda_n + Bi1 \cdot \tan \lambda_n}
\]

(14)

\[
f(z) = \cos(\mu_m z) + B_m \cdot \sin(\mu_m z)
\]

(15)

\[
B_m = \frac{\mu_m \cdot \tan(\mu_m \cdot w) + Bi3}{\mu_m + Bi3 \cdot \sin(\mu_m \cdot w)}
\]

(16)

\[
f_n = \lambda_n + \frac{1}{2} \sin(2\lambda_n) + A_n^2 \cdot \left\{\lambda_n - \frac{1}{2} \sin(2\lambda_n)\right\}
\]

(17)

\[
g_m = \mu_m w + \frac{1}{2} \sin(\mu_m w) + B_m^2 \cdot \left\{\mu_m w - \frac{1}{2} \sin(\mu_m w)\right\}
\]

(18)

The eigenvalues \( \lambda_n \) can be obtained from equation (19) which comes from equation (4) and equation (5).

\[
\frac{\lambda_n \cdot \sin(\lambda_n) - Bi1 \cdot \cos(\lambda_n)}{\lambda_n \cdot \cos(\lambda_n) + Bi1 \cdot \sin(\lambda_n)} = \frac{Bi2 \cdot \cos(\lambda_n) - \lambda_n \cdot \sin(\lambda_n)}{\lambda_n \cdot \cos(\lambda_n) + Bi2 \cdot \sin(\lambda_n)}
\]

(19)

The eigenvalues \( \mu_m \) can be obtained from equation (20) which comes from equation (6) and equation (7).

\[
\frac{\mu_m \cdot \sin(\mu_m \cdot w) - Bi3 \cdot \cos(\mu_m \cdot w)}{\mu_m \cdot \cos(\mu_m \cdot w) + Bi3 \cdot \sin(\mu_m \cdot w)} = \frac{Bi4 \cdot \cos(\mu_m \cdot w) - \mu_m \cdot \sin(\mu_m \cdot w)}{\mu_m \cdot \cos(\mu_m \cdot w) + Bi4 \cdot \sin(\mu_m \cdot w)}
\]

(20)

By applying equation (8) to Fourier’s law, the heat loss rate conducted into the fin through the fin base is given by equation (21).
\[ Q = 4k \cdot l \cdot \theta_0 \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} N_{nm} \cdot \rho_{nm} \cdot C_{nm} \cdot \frac{\sin(\lambda_n)}{\lambda_n} \cdot \frac{\sin(\mu_m \cdot w)}{\mu_m} \] (21)

Convective heat losses from each surface for a thermally asymmetric rectangular fin can be expressed equation (21) through equation (25).

\[
\frac{Q_{qs}}{k\theta_0} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} Bi1 \cdot N_{nm} \cdot D_{nm} \cdot (\cos \lambda_n + A_n \cdot \sin \lambda_n) \cdot \frac{2 \sin(\mu_m \cdot w)}{\mu_m} \] (21)

\[
\frac{Q_{bs}}{k\theta_0} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} Bi2 \cdot N_{nm} \cdot D_{nm} \cdot (\cos \lambda_n - A_n \cdot \sin \lambda_n) \cdot \frac{2 \sin(\mu_m \cdot w)}{\mu_m} \] (22)

\[
\frac{Q_{ls}}{k\theta_0} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} Bi3 \cdot N_{nm} \cdot D_{nm} \cdot (\cos \mu_m + B_m \cdot \sin \mu_m) \cdot \frac{2 \sin(\lambda_n)}{\lambda_n} \] (23)

\[
\frac{Q_{rs}}{k\theta_0} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} Bi4 \cdot N_{nm} \cdot D_{nm} \cdot (\cos \mu_m - B_m \cdot \sin \mu_m) \cdot \frac{2 \sin(\lambda_n)}{\lambda_n} \] (24)

\[
\frac{Q_{is}}{k\theta_0} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} Bi5 \cdot N_{nm} \cdot E_{nm} \cdot \frac{\sin(\lambda_n)}{\lambda_n} \cdot \frac{\sin(\mu_m \cdot w)}{\mu_m} \] (25)

where

\[
D_{nm} = \frac{1}{\rho_{nm}} \cdot \frac{Bi5 \cdot (\cosh(\rho_{nm}L) - 1) + \rho_{nm} \cdot \sinh(\rho_{nm}L)}{\rho_{nm} \cdot \cosh(\rho_{nm}L) + Bi5 \cdot \sinh(\rho_{nm}L)} \] (26)

\[
E_{nm} = \frac{\rho_{nm}}{\rho_{nm} \cdot \cosh(\rho_{nm}L) + Bi5 \cdot \sinh(\rho_{nm}L)} \] (27)

3. Results and Discussions

Figure 2 presents the convective heat losses from each surface of a thermally asymmetric rectangular fin as a function of the non-dimensional fin length for Bi2/Bi1=0.8, Bi4/Bi3=0.9, Bi1=Bi3=Bi5=0.1 and w=0.5. It is shown that the heat loss from the tip decreases while heat losses from all the other surfaces increase as the non-dimensional fin length increases. It can be noted that heat losses from upper and bottom surfaces are less than those from left and right surfaces because the fin width is comparatively narrow relative to the fin height.

Figures 3(a) shows the non-dimensional heat losses a unit width from each surface versus the non-dimensional fin length for Bi1=Bi3=Bi5=0.01, Bi2/Bi1=Bi4/Bi3=0.9 and w=0.1. The heat loss magnitude order and variation trend is somewhat similar to Fig. 2. The same condition except w=2 is described in Fig. 3(b). Comparing this figure to Fig. 3(a), heat loss from each surface varies more rapidly at a rate as the non-dimensional fin length increases and the order of heat loss magnitude between upper, bottom surfaces and left, right surfaces is
reversed because of width change. Also it can be noted that heat loss from the fin tip is the largest until the non-dimensional fin length increases to about 2.

Figure 4 presents the non-dimensional heat losses from each surface versus the non-dimensional fin width for $\text{Bi}_1=\text{Bi}_3=\text{Bi}_5=0.01$, $\text{Bi}_2/\text{Bi}_1=\text{Bi}_4/\text{Bi}_3=0.9$ and $L=5$. Heat losses from left and right surfaces increase as $w$ increases from 0.1 to 1 and then decrease as $w$ increases from 1 to 10 while heat losses from upper, bottom and tip surfaces increase linearly as $w$ increases from 0.1 to 10. It can be explained physically that the effect of heat loss from left and right surfaces on the total heat loss decreases as the fin width increases when the fin length is fixed.

Figure 5 illustrates the variation of the non-dimensional temperature along the fin center line for three different asymmetric condition in case of $\text{Bi}_1=\text{Bi}_3=\text{Bi}_5=0.02$, $L=5$ and $w=0.5$. Average surrounding Biot numbers are equal for three different conditions. The temperature difference increases as $x$ coordinate increases. The temperature is the lowest in case of $\text{Bi}_2/\text{Bi}_1=0.2$, $\text{Bi}_4/\text{Bi}_3=1.0$ and it means physically that heat loss is the highest.

Table 1 lists the heat loss ratio in view of each surface with the
Fig. 3 Non-dimensional heat loss a unit width from each surface vs. $L$ for $Bi1=Bi3=Bi5=0.01$, $Bi2/Bi1=0.9$ and $Bi4/Bi3=0.9$
Fig. 4 Non-dimensional heat loss from each surface vs. \( w \) for \( L=5, \ Bi1=Bi3=Bi5=0.01, \) and \( Bi2/Bi1=Bi4/Bi3=0.9 \)

Fig. 5 Non-dimensional temperature along \( x \) coordinate for \( L=5, \ w=0.5 \) and \( Bi1=Bi3=Bi5=0.02 \)
Table 1 Heat loss ratio in view of each surface with the variation of Bi2/Bi1 for Bi1=Bi3=Bi5=0.05, Bi4/Bi3=0.8, L=5 and w=0.5.

<table>
<thead>
<tr>
<th>Bi2/Bi1</th>
<th>Qrs/Qls (%)</th>
<th>Qbs/QuS (%)</th>
<th>QuS/Qls (%)</th>
<th>Qts/Qls (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>80.35</td>
<td>61.18</td>
<td>49.38</td>
<td>11.27</td>
</tr>
<tr>
<td>0.7</td>
<td>80.35</td>
<td>71.03</td>
<td>49.46</td>
<td>11.17</td>
</tr>
<tr>
<td>0.8</td>
<td>80.35</td>
<td>80.78</td>
<td>49.55</td>
<td>11.08</td>
</tr>
<tr>
<td>0.9</td>
<td>80.35</td>
<td>90.44</td>
<td>49.63</td>
<td>10.99</td>
</tr>
<tr>
<td>1.0</td>
<td>80.35</td>
<td>100.00</td>
<td>49.72</td>
<td>10.91</td>
</tr>
</tbody>
</table>

Fig. 6 Non-dimensional heat loss from each surface vs. Bi2/Bi1 for Bi1=Bi3=Bi5=0.02, w=0.2, L=2 and Bi4/Bi3=1-Bi2/Bi1

variation of Bi2/Bi1 for Bi1=Bi3=Bi5=0.05, Bi4/Bi3=0.8, L=5 and w=0.5. First it can be noted that the ratio of heat loss from right surface to that from left surface remains 80.35 %, which is slightly over 80 %, with the variation of Bi2/Bi1 since Bi4/Bi3 remains 0.8. The ratio of heat loss from bottom surface to that from upper surface varies in proportion to the ratio of Bi2/Bi1 but heat loss ratio is slightly higher
than the Biot number ratio and the difference between these two ratio
decrease as Biot number ratio increases and it finally becomes 0 at
Bi2/Bi1=1. It also describes heat loss from upper surface slightly less
than half of that from left surface even though both surfaces have the
same Biot number.

The non-dimensional heat losses from each surface versus Bi2/Bi1 for
Bi1=Bi3=Bi5=0.02, Bi4/Bi3=1-Bi2/Bi1, w=0.2 and L=2 are illustrated in
Fig. 6. As expected, heat loss from bottom surface increases while that
from right surface decreases as Bi2/Bi1 increases since Bi4 decreases as
Bi2/Bi1 increases. But it is interesting to note that heat losses from
upper, left and tip surfaces increase as Bi2/Bi1 increases even though
Bi1, Bi3 and Bi5 are constants.

4. Conclusions

The following conclusions can be made from the results.
(1) Convective heat loss from fin tip increases as fin width increases
while it decreases as fin length increases.
(2) The ratio of heat loss from bottom surface to that from upper
surface slightly higher than the ratio of Biot number of bottom
surface to that of upper surface when this ratio \((\text{Bi2}/\text{Bi1})\) is not 1.
(3) Heat losses from each surface vary linearly with the variations of
Bi2/Bi1 and Bi4/Bi3.

Reference

1. Abrate, S. and Newnham, P., "Finite Element Analysis of Triangular
Fin Attached to a Thick Wall", Computer & Structures, Vol. 57, No.
6, 1995, pp. 945~957.

2. Burmeister, L. C., "Triangular Fin Performance by the Heat Balance

Methods for the Analysis of a Trapezoidal Fin", KSME Int. Journal,

Two Parabolic Fin Models using Two Different Numerical Methods",

5. Kraus, A. D., Snider, A. D. and Doty, L. F., "An Efficient Algorithm
for Evaluating Arrays of Extended Surface", ASME J. of Heat

6. Look, D. C., "2-D Fin Performance: Bi(top) > Bi(bottom)", ASME J.


Division of Mechanical and Mechatronics Engineering
College of Engineering
Kangwon National University
Chunchon, Korea
h kang@cc.kangwon.ac.kr