The relationship between the 0-tree and other trees in a linear nongroup cellular automata

Sung-Jin Cho

Abstract
We investigate the relationship between the 0-tree and other trees in linear nongroup cellular automata. And we show that given a 0-basic path of 0-tree and a nonzero attractor $\alpha$ of a multiple attractor linear cellular automata with two predecessor we construct an $\alpha$-tree of that multiple attractor linear cellular automata.

1 Introduction

An analysis of the state-transition behavior of group cellular automata (abbreviately, CA) was studied by many researchers ([1], [4], [7], [9]). The characteristic matrix of group CA is nonsingular. But the characteristic matrix of nongroup CA is singular. Although the study of nonsingular linear machines has received considerable attention from researchers, the study of the class of machines with singular characteristic matrix has not received due attention. However some properties of nonsingular CA have been employed in several applications ([5], [8], [9]). In this paper, we investigate the relationship between the 0-tree and other trees in linear nongroup cellular automata. Especially given a 0-basic path of 0-tree and all attractors of a multiple attractor linear CA (abbreviately, MALCA) with two predecessor we construct all trees of that MALCA.

2 Linear nongroup CA

In the rest of this paper, unless specified otherwise, a null boundary CA [7] is simply referred to as a CA.
Definition 2.1. A state with a self-loop in the state-transition diagram of a nongroup CA are referred to as an attractor.

Remark 2.2. The cycles with length \( l \geq 2 \) in the state-transition diagram of nongroup CA are not attractors.

Definition 2.3. [3] The nongroup CA for which the state-transition diagram consists of a set of disjoint components forming (inverted) tree-like structures rooted at attractors are referred to as multiple-attractor linear CA (MALCA).

Remark 2.4. In case the number of attractors is one we call MALCA single-attractor linear CA (SALCA).

The tree rooted at a cyclic state \( \alpha \) is denoted as \( \alpha \)-tree.

Definition 2.5. The depth of a CA is defined to be the minimum number of clock cycles required to reach the cyclic state from any nonreachable state in the state-transition diagram of the CA.

Lemma 2.6. The state 0 in a linear nongroup CA \( \mathbb{C} \) is an attractor.

Proof. Let \( T \) be the characteristic matrix of \( \mathbb{C} \). Then \( T0 = 0 \). Hence 0 is an attractor in \( \mathbb{C} \). \( \square \)

Since the 0-tree and another tree rooted at a nonzero cyclic state have very interesting relationships, the study of the 0-tree is necessary and very important.

Lemma 2.7. If \( d \) is the dimension of the null space \( N(T) \) of the characteristic matrix \( T \) of a nongroup CA \( \mathbb{C} \), the total number of 1-predecessors of the state 0 is \( 2^d \).

Proof. Let \( X \) be a predecessor of the state 0 in \( \mathbb{C} \). Then \( TX = 0 \). Since \( \dim N(T) = d \), \( TX = 0 \) has \( d \) free variables on \( GF(2) \). Hence the total number of 1-predecessors of the state 0 is \( 2^d \). \( \square \)

Theorem 2.8. The number of predecessors of a reachable state and those of the state 0 in a linear nongroup CA \( \mathbb{C} \) are equal.
Proof. Let $T$ be the characteristic matrix of $C$ and $\alpha$ be a reachable state in $C$. Also let $X$ be a state in the $\alpha$-tree. Then the equation $TX = \alpha$ has at least one solution. Therefore $|\{Y|TY = \alpha\}| = |\{X|TX = 0\}|$. Hence the proof is completed. 

**Definition 2.9.** A state $X$ at level $l$ ($l \leq \text{depth}$) of the $\alpha$-tree is a state lying on that tree and it evolves to the state $\alpha$ exactly after $l$-cycles ($l$ is the smallest possible integer for which $T^lX = \alpha$).

**Definition 2.10.** A state $Y$ of an $n$-cell CA is an $i$-predecessor ($1 \leq i \leq 2^n - 1$) of a state $X$ if $TY = X$, where $T$ is the characteristic matrix of the CA. A state $X_i$ in a cycle of length $l$ is called by the cyclic $r$-predecessor ($r < l$) of the state $X_i$ (written by $X_{i-r}$) if $T^rX_{i-r} = X_i$.

From the following theorem we obtain the relation between the minimal polynomial of the characteristic matrix of a CA and the depth of the state-transition diagram of the CA.

**Theorem 2.11.** Let $k$ be the largest integer such that $x^k$ divides the minimal polynomial of the characteristic matrix of an $n$-cell CA $C$. Then the depth of the state-transition diagram of the CA $C$ is $k$.

Proof. Let the characteristic matrix of $C$ be $T$. And let $m(x)$ (the minimal polynomial of $T$) be written as

$$m(x) = x^k\phi(x)$$

where $k$ is the largest integer for which $x^k$ is one of its factors and $\phi(x)$ is a polynomial not divisible by $x$. Since $x^k$ and $\phi(x)$ are co-prime polynomials, the whole space $S$ corresponding to $T$ is decomposed into two invariant subspaces as

$$S = R_1 + R_2$$

where $R_1$ is the invariant subspace with $x^k$ as the minimal polynomial and $R_2$ is the invariant subspace with $\phi(x)$ as the minimal polynomial [6]. Thus there exists at least one nonzero vector $y$ in $R_1$ which has $x^k$ as its minimal polynomial. Therefore $T^ky = 0$ for the minimum value of $k$. Hence the depth of the state-transition diagram of $C$ is $k$. 

\[\square\]
Example 2.12. Let $C$ be a four-cell linear nongroup CA with the rule $<102,102,60,60>$. Then

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

Now $m(x) = x^2(x^2 + 1)$, $S = \{0,1,\cdots,15\}$, $R_1 = \{0,5,10,15\}$ and $R_2 = \{0,1,8,9\}$. Hence the depth of the state-transition diagram of $C$ is 2.

Lemma 2.13. The number of $r$-predecessors($r > 0$) of any cyclic state is the same as that of the state 0.

Proof. Let $S_r$ be the set of all $r$-predecessors of the state 0 ($l > 0$). Then

$$S_r = \{Y|T^rY = 0\}$$

Let $X_0$ be a nonzero cyclic state in the state-transition diagram and let $A_r$ be the set of all $r$-predecessors of the state $X_0$. Then

$$A_r = \{Z|T^rZ = X_0\}$$

Since $A_r \neq \emptyset$,

$$|A_r| = |S_r|$$

This completes the proof. 

3. The relationship between 0-tree and other $\alpha(\neq 0)$-tree in linear nongroup CA

In this section we show that there are interesting relationships between the states in an $\alpha$-tree corresponding to each level state in the 0-tree.

Lemma 3.1. Let $X_l$ and $X_m$ be level $i$ states in an $\alpha$-tree. If there exists $j(\leq i)$ such that $j = \min\{k|T^kX_l = T^kX_m\}$, then $X_l \oplus X_m$ is one of level $j$ states in the 0-tree.

Proof. Since $T^jX_l = T^jX_m$, $T^j(X_l \oplus X_m) = 0$. Thus $X_l \oplus X_m$ is one of $j$-predecessors of the state 0. Suppose that $X_l \oplus X_m$ is a level $p(< j)$ state in the 0-tree. Then $T^p(X_l \oplus X_m) = 0$. Thus $T^pX_l = T^pX_m$. This is a contradiction. Hence $X_l \oplus X_m$ is one of level $j$ states in the 0-tree. 

}\end{document}
From Lemma 3.1 we obtain the following corollary.

**Corollary 3.2.** The sum of two different predecessors of any reachable state is a nonzero predecessor of the state 0.

**Theorem 3.3.** The sum of two states lying at different levels \( p \) and \( q(p > q) \) of the \( \alpha \)-tree is a state at level \( p \) of the 0-tree.

*Proof.* Let \( X_p \) be a level \( p \) state of the \( \alpha \)-tree, \( X_q \) a level \( q \) state of the \( \alpha \)-tree and \( X_0 = \alpha \) the attractor of the \( \alpha \)-tree. Then \( T^p X_p = X_0 \). Since

\[
T^p X_q = T^{p-q}(T^q X_q) = T^{p-q} X_0 = X_0
\]

Thus \( T^p(X_p \oplus X_q) = 0 \) and therefore \( X_p \oplus X_q \) is a \( p \)-predecessor of the state 0. Since \( X_p \) is a level \( p \) state of the \( \alpha \)-tree, \( T^{p-1}X_p \) is a nonzero predecessor of the state \( \alpha \), i.e., \( T^{p-1}X_p \) is a level 1 state of the \( \alpha \)-tree and \( T^{p-1}X_q = X_0 \). Therefore \( T^{p-1}(X_p \oplus X_q) = X_1 \oplus X_0 \).

Since \( X_1 \) and \( X_0 \) are different predecessors of the \( \alpha \)-tree, by Corollary 3.2 \( X_1 \oplus X_0 \) is a nonzero predecessor of the state 0. Hence \( X_p \oplus X_q \) is a state at level \( p \) of the 0-tree. \( \square \)

**Theorem 3.4.** Let the number of 1-predecessors of the state 0 in a linear nongroup CA be \( n \). If \( U_i = \{P_1, P_2, \ldots, P_n\} \) is the set of the \( i \)-predecessors with respect to the state 0 (with \( P_1 = 0 \)) and \( X_1 \) is one of the \( i \)-predecessors with respect to a nonzero cyclic state \( X \), then the set of the \( i \)-predecessors with respect to \( X \) is \( \{X_1 \oplus P_k | k = 1, 2, \ldots, n^i \} \).

*Proof.* We know that \( U_i = \{Y | T^i Y = 0 \} \). And let \( V_i = \{Z | T^i Z = X \} \). Since \( X_1 \in V_i, V_i \neq \emptyset \). Thus \( |V_i| = |U_i| = n^i \). Since \( T^i(X_1 \oplus P_k) = T^i X_1 \oplus T^i P_k = T^i X_1 \oplus 0 = X, X_1 \oplus P_k \) is a \( i \)-predecessor of the state \( X \) where \( k = 1, 2, \ldots, n^i \). Thus the set of the \( i \)-predecessors with respect to \( X \) is \( \{X_1 \oplus P_k | k = 1, 2, \ldots, n^i \} \). \( \square \)

**Theorem 3.5.** Let \( r \) be the number of \( i \)-predecessors of the state 0 in a linear nongroup CA, \( P_{ij}(j = 1, 2, \ldots, (r - 1)n^i - 1) \) be the level \( i \) states of the 0-tree and \( R_i \) the cyclic \( i \)-predecessor of a cyclic state \( X \). And let \( X_{ij}(j = 1, 2, \ldots, (r - 1)n^i - 1) \) be the level \( i \) states of the \( X \)-tree. Then

\[(*) \quad X_{ij} = R_i \oplus P_{ij} \quad \text{where} \quad 1 \leq i \leq \text{depth}, \quad j = 1, 2, \ldots, (r - 1)n^i - 1\]
Proof. We prove (\ast) by mathematical induction. The case where $i = 1$ is trivial by Theorem 3.4. Assume that $X_{kj} = R_k \oplus P_{kj} (j = 1, 2, \cdots, (r - 1)r^{k-1})$. Then $TR_{k+1} = R_k$. Since $T(X_{k+1,j} \oplus R_{k+1}) = X_{kj} \oplus R_k$, $X_{k+1,j} \oplus R_{k+1} = P_{k+1}$ for some $P_{k+1}$ such that $P_{k+1,j}$ is a level $k + 1$ state of the 0-tree. Thus $X_{k+1,j} \oplus R_{k+1} = P_{k+1,j}$ for some $P_{k+1,j}$ such that $P_{k+1,j}$ is a level $k + 1$ state of the 0-tree. Therefore (\ast) holds for $i = k + 1$. Hence the proof is completed. 

From the above theorem we obtain the following corollary.

**Corollary 3.6.** Let $C$ be a nongroup linear CA. Let $r$ be the number of 1-predecessors of the state 0, $P_{ij}(j = 1, 2, \cdots, (r - 1)r^{i-1})$ be the level $i$ states of the 0-tree and $\alpha$ an attractor of $C$. And let $X_{ij}(j = 1, 2, \cdots, (r - 1)r^{i-1})$ be the level $i$ states of the $X$-tree. Then $X_{ij} = \alpha \oplus P_{ij}$ where $1 \leq i \leq \text{depth}$, $j = 1, 2, \cdots, (r - 1)r^{i-1}$

**Example 3.7.** Let $C$ be a four-cell linear nongroup CA with the rule $<102, 102, 102, 60>$. Then $m(x) = x^2(x^2 + 1)$, rank = 3 and depth = 2. The number of predecessors is equal to 2 and cyclic states are $\{0\}, \{4, 12\}$ and $\{8\}$. Thus attractors are 0 and 8. We get the states of each nonzero tree as the following:

(i) In 4-tree, $14 = 10 \oplus 4, 1 = 5 \oplus 4$ and $3 = 15 \oplus 12$.

(ii) In 12-tree, $6 = 10 \oplus 12, 9 = 5 \oplus 12$ and $11 = 15 \oplus 4$.

(iii) In 8-tree, since state 8 is an attractor, the cyclic $r$-predecessor is always state 8.

Thus $13 = 5 \oplus 8, 2 = 10 \oplus 8$ and $7 = 15 \oplus 8$.

The state-transition diagram is as the following:

**Lemma 3.8.** Given a linear nongroup CA $C$, let $T$ be the characteristic matrix of $C$. Let $d$ be the depth of the 0-tree of $C$ and let $\text{dim}N(T) = r$. Then the number of states in the 0-tree is $2^rd$.

**Proof.** Since $\text{dim}N(T) = r$, the number of 1-predecessors of any reachable state is $2^r$ by Lemma 2.7. Let $a_i$ be the number of level $i$ states in the 0-tree. Then $a_{i+1} = 2^r a_i$. Therefore the number of level $i$ states of the 0-tree is $(2^r)^i - 1$. Thus the number of states in the 0-tree is

$$1 + (2^r - 1) + 2^r(2^r - 1) + \cdots + (2^r)^{d-1}(2^r - 1) = 2^rd$$
Theorem 3.9. Let \( C, T, d \) and \( r \) be in Lemma 3.8. Let \( m(x) \) be the minimal polynomial of \( T \). If \(|T + xI| = x^d(x + 1)^{n-d}\) and \( m(x) = x^d(x + 1) \), then the following hold:

(i) The number of states in the 0-tree is \( 2^d \).

(ii) The number of attractors is \( 2^{n-rd} \).

Proof. (i) By Lemma 2.7 the number of predecessors of the state 0 is \( 2^r \). And since \( m(x) = x^d(x + 1) \), the depth of the 0-tree is \( d \). Hence by Lemma 3.8 the number of states in the 0-tree is \( 2^d \).

(ii) By Lemma 2.13 the number of states in each attractor tree is the same as the number of states in the 0-tree. Therefore by (i) the number of attractors is \( 2^n/2^{rd} = 2^{n-rd} \).

From Theorem 3.9 we obtain the following two corollaries.

Corollary 3.10. The number of states in each attractor tree is \( 2^d \).

Corollary 3.11. Given a linear non-group CA \( C \), let \( T \) be the characteristic matrix of \( C \). Let \( d \) be the depth of the 0-tree of \( C \) and let \( \text{dim}N(T) = r \). Then the number of states in the 0-tree is \( 2^d \).

Theorem 3.12. [2] Let \( C \) be SALCA having two-predecessor. If the states of the state-transition diagram of \( C \) are labeled such that \( S_{l,k} \) be the \((k+1)\)-th state in the \( l \)-th level, then the following relation holds good:

\[
S_{l,k} = S_{l,0} \oplus \sum_{i=1}^{l-1} b_i S_{i,0}
\]

where \( b_{l-1}b_{l-2}\cdots b_1 \) is the binary representation of \( k \) and the maximum value of \( k \) is \( 2^{l-1} - 1 \).

Definition 3.13. Let \( C \) be MALCA with two-predecessor and the depth of \( C \) be \( d \). Let \( \beta \) be a nonreachable state of the \( \alpha \)-tree of \( C \). Then we call the \( \beta \rightarrow T\beta \rightarrow \cdots \rightarrow \alpha \) a \( \alpha \)-basic path of the \( \alpha \)-tree of \( C \).

Remark 3.14. Let \( C \) be the SALCA in Theorem 3.12 with the depth \( d \). Then \( S_{d,0} \rightarrow S_{d-1,0} \rightarrow \cdots \rightarrow S_{1,0} \rightarrow 0 \) is a 0-basic path of the 0-tree of \( C \).
Example 3.15. Let $C$ be a five-cell linear nongroup CA with the rule $< 204, 240, 240, 240, 240 >$. Then

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

The minimal polynomial $m(x)$ of $T$ is $m(x) = x^4(x + 1)$ and attractors are 0 and 31. The state-transition diagram is as the following:

Figure 1: The state-transition diagram of $C$
the l-th level of the α-tree (resp. 0-tree) in $C$ and $S_{l,0}^\alpha = S_{l,0} \oplus \alpha$, then the following hold:

$$S_{l,k}^\alpha = S_{l,0}^\alpha \oplus \sum_{i=1}^{l-1} b_i S_{i,0}$$

where $b_{l-1}b_{l-2}\cdots b_1$ is the binary representation of $k$ and the maximum value of $k$ is $2^{l-1} - 1$.

**Proof.** Since $S_{l,k} = S_{l,k} \oplus \alpha$ by Lemma 3.16,

$$S_{l,k}^\alpha = S_{l,k} \oplus \alpha$$

$$= (S_{l,0} \oplus \sum_{i=1}^{l-1} b_i S_{i,0}) \oplus \alpha \quad \text{by Theorem 3.12}$$

$$= (S_{l,0} \oplus \alpha) \oplus \sum_{i=1}^{l-1} b_i S_{i,0}$$

$$= S_{l,0}^\alpha \oplus \sum_{i=1}^{l-1} b_i S_{i,0}$$

From Theorem 3.18 we obtain the following corollary.

**Corollary 3.19.** Let $C$ be a MALCA with two-predecessor. Given a 0-basic path of the 0-tree of $C$ and the set of attractors, we can construct the state-transition diagram of $C$.

**References**


Department of Applied Mathematics
Pukyong National University
Pusan 608-737
KOREA
e-mail: sjcho@dolphin.pknu.ac.kr
SMOOTHING ANALYSIS IN MULTIGRID METHOD FOR THE LINEAR ELASTICITY FOR MIXED FORMULATION

KAB SEOK KANG

Abstract. We introduce an assumption about smoothing operator for mixed formulations and show that convergence of Multigrid method for the mixed finite element formulation for the Linear Elasticity. And we show that Richardson and Kaczmarz smoothing satisfy this assumption.

1. Introduction and Preliminary

We consider Multigrid method for the pure displacement and pure traction problems in planar linear elasticity by using mixed formulation. The resulting algebraic linear operators by discretization of mixed formulation for the linear elasticity are not positive definite but nonsingular. So, we cannot use the Jacobi and Gauss-Seidel smoother but can use Richardson type smoother and Kaczmarz smoother for solving algebraic linear system.

Richardson type smoother is a very simple and convergence of Multigrid method with this was easily shown([1],[2],[10],[11], [12], [13],[14], [18]), but Multigrid method with this has a slow convergence. For the positive definite problem, authors show that the convergence of Multigrid method with various smoothing by using some assumptions concerning smoothing and show that Jacobi smoothing and Gauss-Seidel smoothing satisfy these assumptions([3], [4], [5], [6], [7], [8], [9],[19]). In [16], authors introduce weaker assumptions and show that convergence of Multigrid method. In this paper, we introduce an assumption concerning smoother and show that Multigrid algorithm converge under this assumption and Richardson and Kaczmarz smoother satisfy this assumption.

From here and after, a boldfaces is used to denote vector-valued functions, operators, and their associated spaces. Upper characters and Greece characters are used for matrix-valued functions and operators. We define

\[\text{grad} p = \left( \begin{array}{c} \frac{\partial p}{\partial x_1} \\ \frac{\partial p}{\partial x_2} \end{array} \right), \quad \text{curl} p = \left( \begin{array}{c} \frac{\partial p}{\partial x_2} \\ -\frac{\partial p}{\partial x_1} \end{array} \right)\]
\[
\text{div}\tau = \left( \frac{\partial \tau_{11}}{\partial x_1} + \frac{\partial \tau_{12}}{\partial x_2} \right),
\]
\[
\text{div}v = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}, \quad \text{rot}v = -\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1},
\]
\[
\text{Grad}v = \left( \frac{\partial v_1}{\partial x_1}, \frac{\partial v_1}{\partial x_2} \right), \quad \text{Curl}v = \left( \frac{\partial v_2}{\partial x_1}, -\frac{\partial v_1}{\partial x_1} \right).
\]

We also define
\[
\delta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \chi = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad \text{tr}(\tau) = \tau : \delta,
\]
where
\[
\tau : \delta = \sum_{i=1}^{2} \sum_{j=1}^{2} \delta_{ij} \tau_{ij}.
\]

Finally, let
\[
\varepsilon(v) = \frac{1}{2}[\text{Grad}v + (\text{Grad}v)^t] = \text{Grad}v - \frac{1}{2}(\text{rot}v)\chi.
\]

Let \(\Omega\) be a bounded convex polygonal domain in \(\mathbb{R}^2\). \(u\) denotes the displacement, \(f\) the body force, \(g\) the boundary traction, \(\mu > 0, \lambda > 0\) the Lamé constants, and \(n\) is the outer normal. We assume that the Lamé constants \((\mu, \lambda)\) belong to the range \([\mu_1, \mu_2] \times [\lambda_0, \infty)\), where \(\mu_1, \mu_2, \lambda_0\) are fixed positive constants.

Here we define two elasticity problems and give the well known properties concerning solution of the problems. The pure displacement boundary value problem for planar linear elasticity is given by
\[
-\text{div}\{2\mu \varepsilon(u) + \lambda \text{tr}(\varepsilon(u))\delta\} = f, \quad \text{in } \Omega,
\]
\[
u = 0, \quad \text{on } \partial \Omega.
\]

It is well known ([13]) that, for \(f \in L^2(\Omega)\), equation (1.1) has a unique solution \(u \in H^2(\Omega) \cap H_0^1(\Omega)\). Moreover, there exists a positive constant \(C\) independent of \(\mu\) and \(\lambda\) such that
\[
\|u\|_{H^2(\Omega)} + \lambda \|\text{div}u\|_{H_0^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}.
\]

The pure traction boundary value problem for planar linear elasticity is given by
\[
-\text{div}\{2\mu \varepsilon(u) + \lambda \text{tr}(\varepsilon(u))\delta\} = f, \quad \text{in } \Omega,
\]
\[
(2\mu \varepsilon(u) + \lambda \text{tr}(\varepsilon(u))\delta) \cdot n = g, \quad \text{on } \partial \Omega.
\]

Since the domain \(\Omega\) is a polygon which has corners, the boundary conditions (1.2) must be carefully handled. We shall denote by \(S_i, 1 \leq i \leq n\), the vertices of \(\partial \Omega\); by \(\Gamma_i, 1 \leq i \leq n\), be open line segment joining \(S_i\) to \(S_{i+1}\); by \(t_i\) the positively oriented unit tangent along \(\Gamma_i\); and by \(n_i\) the unit outer normal along \(\Gamma_i\). Let \(p \in H^{1/2}(\Gamma_i)\) and \(q \in H^{1/2}(\Gamma_{i+1})\). We say that \(p \equiv q\) at \(S_{i+1}\) if
\[
\int_0^t |q(s) - \tilde{q}(s)| \frac{ds}{s} < \infty,
\]
where $s$ is the oriented arc length measured from $S_{i+1}$ and $t$ is a positive number less than $\min\{|\Gamma_i| : 1 \leq i \leq n\}$.

We are able to write equation (1.2) more precisely as

$$-\text{div} \{2\mu \varepsilon(u) + \lambda \text{tr} (\varepsilon(u)) \, \delta\} = f, \quad \text{in } \Omega,$$

$$(2\mu \varepsilon(u) + \lambda \text{tr} (\varepsilon(u)) \, \delta) \, n|_{\Gamma_i} = g_i, \quad 1 \leq i \leq n,$$

where $f \in L^2(\Omega)$, and $g_i \in H^{1/2}(\Gamma_i)$ satisfy

$$g_i \cdot n_{i+1} \equiv g_{i+1} \cdot n_i \quad \text{at } S_{i+1} \text{ for } 1 \leq i \leq n.$$

In order to exist a solution of (1.3), $f$ and $g_i$ must satisfy the compatibility condition

$$\int_\Omega f \cdot v \, dx + \sum_{i=1}^n \int_{\Gamma_i} g_i \cdot n_i \, ds = 0, \quad \forall v \in RM,$$

where $RM$, the space of rigid motions, is defined by

$$RM := \{ v : v = (a + bx, c - by), \quad a, b, c, \in \mathbb{R} \}.$$

When this compatibility condition holds, the pure traction boundary value problem (1.3) has a unique solution ([14],[15]) $u \in H^2_\perp(\Omega)$ where

$$H^k_\perp(\Omega) := \{ u \in H^k(\Omega) : \int_\Omega u \cdot v \, dx = 0, \quad \forall v \in RM \}.$$

Moreover, there exists a positive constant $C$ independent of $\mu$ and $\lambda$ such that

$$\|u\|_{H^2(\Omega)} + \lambda \|\text{div} \, u\|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}.$$

Here, $H^k(\Omega), k \geq 0$, denotes the usual $L^2$-based Sobolev spaces of vector-valued functions. The space $L^2_\perp(\Omega)$ is interpreted as $H^1_\perp(\Omega)$. Note that $|u|_{H^k(\Omega)}$ becomes a norm on $H^k_\perp(\Omega)$.

In Section 2, we consider the mixed formulations of (1.1) and (1.3) and its finite discretizations. In Section 3, we consider Multigrid methods and its convergence analysis with smoothing assumption. In Section 4, we show that simple Richardson type smoother and Kaczmarz smoother satisfy the above assumption concerning smoother. We give numerical experiment of (1.1) in Section 5.

2. Mixed Formulations and its Finite Discretizations

First, we consider the pure displacement boundary value problem. The boundary value problem (1.1) can be written as

$$-\mu \Delta u - (\mu + \lambda) \text{grad}(\text{div} \, u) = f, \quad \text{in } \Omega,$$

$$u = 0, \quad \text{on } \partial \Omega.$$

Hence, we have the following weak formulation:

Find $u \in H^1_0(\Omega)$ such that

$$(2.1) \quad \mu \int_\Omega \text{Grad} : \text{Gradv} \, dx + (\mu + \lambda) \int_\Omega (\text{div} \, u)(\text{div} \, v) \, dx = \int_\Omega f \cdot v \, dx,$$
for all $v \in H^1_0(\Omega)$. Let $\gamma = \frac{\mu + \lambda}{\mu}$ and $p = \gamma \div \mathbf{u}$. It is clear that (2.1) is equivalent to the following mixed formulation:

Find $(u, p) \in H^1_0(\Omega) \times L^2(\Omega)$ such that

$$
\int_{\Omega} \nabla u : \nabla v \, dx + \int_{\Omega} p \div v \, dx = \frac{1}{\mu} \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, dx, \quad \forall v \in H^1_0(\Omega),
$$

$$
\int_{\Omega} (\div u) q - \frac{1}{\gamma} \int_{\Omega} pq \, dx = 0, \quad q \in L^2(\Omega).
$$

Equation (2.2) can be written concisely as

$$(2.3) \quad \mathcal{B}((u, p), (v, q)) = \frac{1}{\mu} \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, dx, \quad \forall (v, q) \in H^1_0(\Omega) \times L^2(\Omega),$$

where the symmetric bilinear form $\mathcal{B}(\cdot, \cdot) : H^1_0(\Omega) \times L^2(\Omega) \rightarrow \mathbb{R}$ is defined by

$$
\mathcal{B}((v_1, q_1), (v_2, q_2)) := \int_{\Omega} \left\{ \nabla v_1 : \nabla v_2 + q_1 (\div v_2) + (\div v_1) q_2 - \frac{1}{\gamma} q_1 q_2 \right\} \, dx.
$$

It is clear from the definition of $\mathcal{B}$ that

$$
|\mathcal{B}((v_1, q_1), (v_2, q_2))| \leq \sqrt{2} (|v_1|_{H^1(\Omega)} + \|q_1\|_{L^2(\Omega)}) (|v_2|_{H^1(\Omega)} + \|q_2\|_{L^2(\Omega)}).
$$

Let $\mathcal{T}_k$ be a sequence of triangulations of $\Omega$, where $\mathcal{T}_{k+1}$ is obtained by connecting the midpoints of the triangles in $\mathcal{T}_k$. We will denote $\max\{\text{diam} T : T \in \mathcal{T}_k\}$ by $h_k$. Let

$$
Q_k = \{ q : q \in L^2(\Omega) \text{ and } q|T \text{ is a constant for all } T \in \mathcal{T}_k \}.
$$

The nonconforming finite element spaces $V_k$ are defined as follows.

$$
V_k = \{ v : v \in L^2(\Omega), v|_T \text{ is linear for all } T \in \mathcal{T}_k, \quad v \text{ is continuous at the midpoints of interelement boundaries} \}
$$

and $v = 0$ at the midpoints of edges along $\partial \Omega$.

The discretized problem for (2.3) is: Find $(u_k, p_k) \in V_k \times Q_k$ such that

$$(2.4) \quad \mathcal{B}_k((u_k, p_k), (v, q)) = \frac{1}{\mu} \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, dx, \quad \forall (v, q) \in V_k \times Q_k.$$

Here the symmetric bilinear form $\mathcal{B}_k(\cdot, \cdot) : (H^1_0(\Omega) + V_k) \times Q_k \rightarrow \mathbb{R}$ is defined by

$$
\mathcal{B}_k((v_1, q_1), (v_2, q_2)) := \int_{\Omega} \left\{ \nabla_k v_1 : \nabla_k v_2 + q_1 (\div_k v_2) + (\div_k v_1) q_2 - \frac{1}{\gamma} q_1 q_2 \right\} \, dx,
$$

where

$$
(\nabla_k v)|_T = \nabla (v|_T), \quad (\div_k v)|_T = \div (v|_T), \quad \forall T \in \mathcal{T}_k.
$$
In [13], author show that (2.4) is uniquely solvable and derive the following discretization error estimate:
\[
\|u - u_k\|_{L^2(\Omega)} + h_k (\|u - u_k\|_k + \|p - p_k\|_{L^2(\Omega)}) \leq Ch_k^2 \|f\|_{L^2(\Omega)},
\]
where the nonconforming energy norm \(\|\cdot\|_k\) on \(H^1_0(\Omega) + V_k\) is defined by
\[
\|v\|_k := \|\text{Grad}_k v\|_{L^2(\Omega)}.
\]

Second, we consider the pure traction problem. Let \(\gamma = \frac{\lambda}{2\mu}\) and \(p = \gamma \text{div} u\), we consider the mixed weak formulation for (1.3) as follows:
Find \((u, p)\) such that
\[
\int_{\Omega} \varepsilon(u) : \varepsilon(v) \, dx + \int_{\Omega} p(\text{div} v) \, dx = \frac{1}{2\mu} \left[ \int_{\Omega} f \cdot v \, dx + \sum_{i=1}^n \int_{\Gamma_i} g_i \cdot v_{|\Gamma_i} \, ds \right],
\]
\[
\int_{\Omega} (\text{div} u) q \, dx - \frac{1}{\gamma} \int_{\Omega} pq \, dx = 0,
\]
for all \((v, q)\) in \(H^1_\perp(\Omega) \times L^2(\Omega)\).
Replacing \(p\) and \(q\) by \(\sqrt{\omega} p\) and \(\sqrt{\omega} q\) (\(\omega \geq 1\)), respectively, we obtain the following formulation which is equivalent to (2.5):
Find \((u, p)\) in \(H^1_\perp(\Omega) \times L^2(\Omega)\) such that
\[
B_\omega ((u, p), (v, q)) = \frac{1}{2\mu} \left[ \int_{\Omega} f \cdot v \, dx + \sum_{i=1}^n \int_{\Gamma_i} g_i \cdot v_{|\Gamma_i} \, ds \right],
\]
for all \((v, q)\) in \(H^1_\perp(\Omega) \times L^2(\Omega)\), where
\[
B_\omega ((u, p), (v, q)) := \int_{\Omega} \left\{ \varepsilon(u) : \varepsilon(v) + \sqrt{\omega} p(\text{div} v) + \sqrt{\omega}(\text{div} u) q - \frac{\omega}{\gamma} pq \right\} \, dx.
\]
The quantity \(\omega\) is called the weighting factor. Equation (2.6) has a unique solution on \(H^1_\perp(\Omega) \times L^2(\Omega)\).

Let \(\{T^k\}\) be a family of triangulations of \(\Omega\), where \(T^{k+1}\) is obtained by connecting the midpoints of the edges of the triangles in \(T^k\). Let \(h_k := \max_{T \in T^k} \text{diam} T\), then \(h_k = 2h_{k+1}\). Now let us define the conforming finite element spaces.
\[
W_k := \{ u : u_{|T} \text{ is linear for all } T \in T^k, u \text{ is continuous on } \Omega \},
\]
\[
W^\perp_k := \left\{ u \in W_k : \int_{\Omega} u \cdot v \, dx = 0, \quad \forall v \in \text{RM} \right\}.
\]
To describe the mixed finite element method, we define
\[
Q_k := \{ q : q \in L^2(\Omega) \text{ and } q_{|T} \text{ is a constant for all } T \in T^k \}.
\]
For each \(k\), define the bilinear form \(B_{\omega,k}\) on \(H^1(\Omega) \times L^2(\Omega)\) by
\[
B_{\omega,k} ((u, p), (v, q)) := \int_{\Omega} \left\{ \varepsilon(u) : \varepsilon(v) + \sqrt{\omega} p_k(\text{div} v) + \sqrt{\omega}(P_{k-1} \text{div} u) q - \frac{\omega}{\gamma} pq \right\} \, dx,
\]
where $P_{k-1}$ is the $L^2$-orthogonal projection onto $Q_{k-1}$. Note that the bilinear forms $B_{\omega,k}$ are symmetric but indefinite.

The following discretization of (2.6) is a modification of one introduced by Falk in [15].

Find $(u_k, p_k) \in W^\perp_k \times Q_{k-1}$ such that

$$B_{\omega,k}((u_k, p_k), (v, q)) = \frac{1}{2\mu} \left[ \int_{\Omega} f \cdot v dx + \sum_{i=1}^n \int_{\Gamma_i} g_i \cdot v|_{\Gamma_i} ds \right]$$

for all $(v, q) \in W^\perp_k \times Q_{k-1}$.

In [17], author showed the uniqueness of the solution of the discretization (2.7) with $\omega = 1$ and derived the following discretization error estimate:

$$\|u - u_k\|_{L^2(\Omega)} + h_k (\|u - u_k\|_{H^1(\Omega)} + \|p - p_k\|_{L^2(\Omega)}) \leq C h_k^2 \left\{ \|f\|_{L^2(\Omega)} + \sum_{i=1}^n \|g_i\|_{H^{1/2}(\Gamma_i)} \right\}.$$

3. MULTIGRID ALGORITHM AND CONVERGENCE ANALYSIS

We define the intergrid transfer operators and the mesh-dependent norms for each problems. Next, we present Multigrid algorithm MG with Kaczmarz or Richardson type smoother and prove the convergence of the algorithm at the same time. Some of them are rephrases of lemmas in [13] and [18] and we give these lemmas without proof.

For the pure displace problem, since the $V_k$'s are nonconforming, the intergrid transfer operators are defined by averaging.

Let $J_{k-1}^k : V_{k-1} \rightarrow V_k$ be defined by

$$J_{k-1}^k(v)(m_e) = \left\{ \begin{array}{ll} v(m_e), & \text{if } m_e \in \text{int} T \\ \frac{1}{2} [v|_{T_1}(m_e) + v|_{T_2}(m_e)], & \text{if } e = T_1 \cap T_2 \end{array} \right.$$  

for some $T \in T_{k-1}$, at the midpoints $m_e$ of internal edges $e$ in $T_k$.

The coarse-to-fine operator $I_{k-1}^k : V_{k-1} \times Q_{k-1} \rightarrow V_k \times Q_k$ is defined by

$$I_{k-1}^k(v, q) = (J_{k-1}^k v, q).$$

Define the mesh dependent inner product by

$$(u, p), (v, q) \rightleftharpoons (u, v)_{L^2(\Omega)} + h_k^2 (p, q)_{L^2(\Omega)}.$$

The intergrid transfer operators $T_k^k : V_k \times Q_k \rightarrow V_{k-1} \times Q_{k-1}$ is defined by

$$\left( T_k^{k-1}(u, p), (v, q) \right)_{k-1} = \left( (u, p), I_{k-1}^k(v, q) \right)_k,$$

for all $(u, p) \in V_k \times Q_k$, and $(v, q) \in V_{k-1} \times Q_{k-1}$.

Let $Q_k = \{ q \in Q_k : \int_{\Omega} q dx = 0 \}$. 

Define the mesh dependent inner product by

$$(u, p), (v, q) \rightleftharpoons (u, v)_{L^2(\Omega)} + h_k^2 (p, q)_{L^2(\Omega)}.$$
The following three lemmas which concerned the pure displacement problems came from [13].

**Lemma 3.1.** The following properties of $I_{k-1}^k$ and $I_k^{k-1}$ hold.

(i) Given any $(v, q) \in V_k \times Q_k$, $(v, q) \in V_k \times \hat{Q}_k$ only if $((v, q), (0, 1))_k = 0$.

(ii) $(I_{k-1}^k(v, q), (0, 1))_k = \frac{1}{4}((v, q), (0, 1))_{k-1}$, for all $(v, q) \in V_{k-1} \times Q_{k-1}$.

(iii) $I_{k-1}^k : V_{k-1} \times \hat{Q}_{k-1} \to V_k \times \hat{Q}_k$.

(iv) $I_k^{k-1} : V_k \times \hat{Q}_k \to V_{k-1} \times \hat{Q}_{k-1}$.

Let $B_k : V_k \times Q_k \to V_k \times Q_k$ be defined by

$$(B_k(u, p), (v, q))_k = B_k((u, p), (v, q)), \quad \forall (u, p), (v, q) \in V_k \times Q_k.$$

**Lemma 3.2.** $B_k : V_k \times \hat{Q}_k \to V_k \times \hat{Q}_k$.

Let $\hat{B}_k = B_k|_{V_k \times \hat{Q}_k}$.

**Lemma 3.3.** The spectral radius of $\hat{B}_k \leq C_h^{-2}$ for $k = 1, 2, \ldots$.

For the pure traction problem, because the $W_k$'s are conforming, the intergrid transfer operators $I_{k-1}^k$ are defined by the natural way.

Define the mesh dependent inner product by

$$(u, v)_k := (u, v)_{L^2(\Omega)} + h_k^2(p, q)_{L^2(\Omega)}.$$

The intergrid transfer operators $I_k^{k-1} : W_k \times Q_{k-1} \to W_{k-1} \times Q_{k-2}$ is defined by

$$(I_k^{k-1}(u, p), (v, q))_{k-1} = ((u, p), (v, q))_k,$$

for all $(u, p) \in W_k \times Q_{k-1}$, and $(v, q) \in W_{k-1} \times Q_{k-2}$.

The following three lemmas which concerns the pure traction problems came from [18].

**Lemma 3.4.** (i) $RM \subset W_k$, $\forall k = 1, 2, \ldots$.

(ii) Given $(u, p) \in W_k \times Q_{k-1}$,

$$(u, p) \in W_k^t \times Q_{k-1} \iff ((u, p), (v, 0))_k, \quad \forall v \in RM.$$

(iii) $I_k^{k-1} : W_k^t \times Q_{k-1} \to W_{k-1}^t \times Q_{k-2}$.

Define $B_{k, \omega} : W_k \times Q_{k-1} \to W_k \times Q_{k-1}$ by

$$(B_{k, \omega}(u, p), (v, q))_k = B_{k, \omega}((u, p), (v, q)), \quad \forall (u, p), (v, q) \in V_k \times Q_{k-1}.$$

**Lemma 3.5.** $B_k : W_k^t \times Q_{k-1} \to W_k^t \times Q_{k-1}$.

Let $B_{k, \omega} = B_{k, \omega}|_{W_k^t \times Q_{k-1}}$.

**Lemma 3.6.** The spectral radius of $B_{k, \omega} \leq C h^{-2}$ for $k = 1, 2, \ldots$.

Here and after in this chapter, we only consider the case of the pure displacement problems. So we only use $V_k$, $\hat{Q}_k$ and $\hat{B}_k$, but the following lemmas and theorem are satisfied the case of pure traction problems by replacing $V_k$ as $W_k^t$, $\hat{Q}_k$ as $Q_{k-1}$ and $\hat{B}_k$ as $B_k^t$. 
The mesh-dependent norms on $\mathbf{V}_k \times \hat{Q}_k$ are defined as follows:

$$
|||(\mathbf{u}, p)|||_{s,k} := \sqrt{ \left( \frac{(\hat{B}_k^2)^{s/2}(\mathbf{u}, p), (\mathbf{u}, p)}{k} \right)}, \quad \forall (\mathbf{u}, p) \in \mathbf{V}_k \times \hat{Q}_k.
$$

This norm is well-defined. Moreover, for all $(\mathbf{u}, p), (\mathbf{v}, q) \in \mathbf{V}_k \times \hat{Q}_k$,

$$
|||(\mathbf{u}, p)|||_{0,k} = \sqrt{\|\mathbf{u}\|_{L^2(\Omega)}^2 + h_k^{-2} p^2_{L^2(\Omega)}},
$$

$$
|\mathcal{B}_k ((\mathbf{u}, p), (\mathbf{v}, q))| \leq |||(\mathbf{u}, p)|||_{2,k} |||(\mathbf{v}, q)|||_{0,k},
$$

and

$$
|||(\mathbf{u}, p)|||_{2,k} = \sup_{(\mathbf{v}, q) \in \mathbf{V}_k \times \hat{Q}_k - (0,0)} \|\mathcal{B}_k ((\mathbf{u}, p), (\mathbf{v}, q))\|. 
$$

Define $P_k^{-1} : \mathbf{V}_k \times \hat{Q}_k \to \mathbf{V}_{k-1} \times \hat{Q}_{k-1}$ by

$$
\mathcal{B}_{k-1} \left( P_{k}^{-1} (\mathbf{u}, p), (\mathbf{v}, q) \right) = \mathcal{B}_k \left( (\mathbf{u}, p), I_{k-1}^k (\mathbf{v}, q) \right),
$$

for all $(\mathbf{u}, p) \in \mathbf{V}_k \times \hat{Q}_k$ and $(\mathbf{v}, q) \in \mathbf{V}_{k-1} \times \hat{Q}_{k-1}$.

We are now ready to state the basic Lemmas which are essential in the approximation property of the multigrid algorithm.

**Lemma 3.7.** Given $\mathbf{w} \in \mathbf{L}^2(\Omega)$, let $(\mathbf{u}_k, p_k) \in \mathbf{V}_k \times \hat{Q}_k$ be the solution of

$$
\mathcal{B}_k ((\mathbf{u}_k, p_k), (\mathbf{v}, q)) = \int_{\Omega} \mathbf{w} \cdot \mathbf{v} \, dx, \quad \forall (\mathbf{v}, q) \in \mathbf{V}_k \times \hat{Q}_k
$$

and $(\mathbf{u}_{k-1}, p_{k-1}) \in \mathbf{V}_{k-1} \times \hat{Q}_{k-1}$ be the solution of

$$
\mathcal{B}_{k-1} ((\mathbf{u}_{k-1}, p_{k-1}), (\mathbf{v}, q)) = \int_{\Omega} \mathbf{w} \cdot \mathbf{v} \, dx, \quad \forall (\mathbf{v}, q) \in \mathbf{V}_{k-1} \times \hat{Q}_{k-1}.
$$

Then

$$
||| P_{k}^{-1} (\mathbf{u}_k, p_k) - (\mathbf{u}_{k-1}, p_{k-1}) |||_{0,k-1} \leq Ch_k^2 \|\mathbf{w}\|_{\mathbf{L}^2(\Omega)}.
$$

**Lemma 3.8.** Given $w \in \mathbf{L}^2(\Omega)$, let $(\mathbf{u}_k, p_k) \in \mathbf{V}_k \times \hat{Q}_k$ be the solution of

$$
\mathcal{B}_k ((\mathbf{u}_k, p_k), (\mathbf{v}, q)) = \int_{\Omega} w q \, dx, \quad \forall (\mathbf{v}, q) \in \mathbf{V}_k \times \hat{Q}_k
$$

and $(\mathbf{u}_{k-1}, p_{k-1}) \in \mathbf{V}_{k-1} \times \hat{Q}_{k-1}$ be the solution of

$$
\mathcal{B}_{k-1} ((\mathbf{u}_{k-1}, p_{k-1}), (\mathbf{v}, q)) = \int_{\Omega} w q \, dx, \quad \forall (\mathbf{v}, q) \in \mathbf{V}_{k-1} \times \hat{Q}_{k-1}.
$$

Then

$$
||| (\mathbf{u}_k, p_k) - I_{k-1}^k (\mathbf{u}_{k-1}, p_{k-1}) |||_{0,k} \leq Ch_k \|w\|_{\mathbf{L}^2(\Omega)}.
$$

Finally, to define the k-th level Multigrid algorithm MG$_k$, we need linear smoothing operators $R_k : \mathbf{V}_k \times \hat{Q}_k \to \mathbf{V}_k \times \hat{Q}_k$ for all $k$. For the analysis of Multigrid algorithms, we assume the following condition concerning the smoothing operators. To describe this, we first define $K_k = I - R_k \hat{B}_k$. 
Smoothing assumption (SM). There exists a constant $C$, independent of $h_k$ and $m$, such that

\begin{equation}
|||K^m_k(u,p)|||_{2,k} \leq Ch_k^{-2} \frac{1}{\sqrt{m}}|||(u,p)|||_{0,k}, \quad \forall (u,p) \in V_k \times \hat{Q}_k.
\end{equation}

Here, we only consider the $W$-cycle nonsymmetric multigrid algorithm.

The $k$-th level iteration scheme of Multigrid algorithm $\text{MG}_k$ : For $k = 1$, $\text{MG}_k((y_0,s_0),(w,r))$ is the solution obtained from a direct method, i.e.,

$$\text{MG}_1((y_0,s_0),(w,r)) = \left(\hat{B}_1\right)^{-1}(w,r).$$

We assume that $\text{MG}_{k-1}$ is defined. The $k$-th level iteration with initial guess $(y_0,s_0) \in V_k \times \hat{Q}_k$ yields $\text{MG}_k((y_0,s_0),(w,r))$ as an approximate solution to the following problem.

Find $(y,s) \in V_k \times \hat{Q}_k$ such that

$$\hat{B}_k(y,s) = (w,r), \quad \text{where} \quad (w,r) \in V_k \times \hat{Q}_k.$$

Smoothing Step: The approximation $(y_m,s_m) \in V_k \times \hat{Q}_k$ is constructed recursively from the initial guess $(y_0,s_0)$ and the equations

$$(y_1,s_1) = (y_{l-1},s_{l-1}) + R_kB_k((w,r) - B_k(y_{l-1},s_{l-1})), \quad 1 \leq l \leq m.$$  

Correction Step: The coarse-grid correction in $V_{k-1} \times \hat{Q}_{k-1}$ is obtained by applying the $(k-1)$-th level iteration twice. More precisely,

$$\begin{align*}
(v_0,q_0) &= (0,0) \quad \text{and} \\
(v_i,q_i) &= \text{MG}_{k-1}((v_{i-1},q_{i-1},(\bar{w},\bar{r}))), \quad i = 1,2
\end{align*}$$

where $(\bar{w},\bar{r}) \in V_{k-1} \times \hat{Q}_{k-1}$ is defined by $(\bar{w},\bar{r}) := I^{k-1}_k((w,r) - B_k(y_m,s_m))$.

Then

$$\text{MG}_k((y_0,s_0),(w,r)) = (y_m,s_m) + (v_2,q_2).$$

Let the final output of the two-grid algorithm be

$$(y^\sharp,s^\sharp) := (y_m,s_m) + (v^\sharp,q^\sharp)$$

where

$$(v^\sharp,q^\sharp) = \left(B_{k-1}\right)^{-1}(\bar{w},\bar{r})$$

$$= \left(\hat{B}_{k-1}\right)^{-1}I^{k-1}_k((w,r) - B_k(y_m,s_m))$$

$$= \left(\hat{B}_{k-1}\right)^{-1}I^{k-1}_kB_k(y - y_m,s - s_m).$$

The following two lemmas are found in [13].

**Lemma 3.9.**

$$(v^\sharp,q^\sharp) = P^{k-1}_k(y - y_m,s - s_m).$$
From the definitions, we have 
\[(y - y_m, s - s_m) = K_m^k(y - y_0, s - s_0),\]
\[(y - y^\sharp, s - s^\sharp) = (I - P_k^{k-1})K_m^k(y - y_0, s - s_0).\]

**Lemma 3.10.** There exists a constant $C$, independent of $h_k$ and $m$, such that
\[||(I - P_k^{k-1})(u, p)||_{0,k} \leq Ch_k^2||(u, p)||_{2,k}, \quad \forall (u, p) \in V_k \times \hat{Q}_k.\]

**Theorem 3.11.** [Convergence of the Two-Grid Algorithm] There exists a constant $C$, independent of $k$ and $m$, such that
\[||(y - y^\sharp, s - s^\sharp)||_{0,k} \leq \frac{C}{\sqrt{m}}||(y - y_0, s - s_0)||_{0,k}.\]

**Proof.** From the definition, (3.2), and (3.1), we get
\[||(y - y^\sharp, s - s^\sharp)||_{0,k} = \frac{C}{\sqrt{m}}||(y - y_0, s - s_0)||_{0,k} \leq C \sqrt{m}|||(y - y_0, s - s_0)|||_{0,k}.\]

**Theorem 3.12.** [Convergence of the $k$-th level Iteration] There exists a constant $C$, independent of $k$ and $m$, such that
\[||(y, s) - MG_k((y_0, s_0), (w, r))||_{0,k} \leq \frac{C}{\sqrt{m}}||(y - y_0, s - s_0)||_{0,k}.\]

### 4. Verification of the Smoothing Assumption

In this section, we consider two smoothing, one is a Richardson type smoothing for nonsymmetric or indefinite operator and other is a Kaczmarz smoothing.

The Richardson type smoothing is defined by
\[R_k := \frac{1}{\Lambda_k^2} \hat{B}_k^2,\]
where $\Lambda_k^2$ be the largest eigenvalue of $\hat{B}_k^2$. This smoothing are considered in [13] and [18].

**Lemma 4.1.** The above $R_k$ satisfy Smoothing assumption (SM).

Now, we consider the Kaczmarz smoothing. Let $\hat{B}_k = (b_{ij})^k_{i,j=1}$. Then Kaczmarz smoother is defined by the following algorithm.

**Kaczmarz Algorithm.** Let $(w, r) \in V_k \times \hat{Q}_k$. We define $R_k(w, r) \in V_k \times \hat{Q}_k$ as follows:

(i) Set $(\phi_0, \zeta_0) = (0, 0)$.

(ii) Define $(\phi_i, \zeta_i)$ for $i = 1, \ldots, l$ by
\[\frac{b_i}{b_i^t b_i}(b_i^t(\phi_{i-1}, \zeta_{i-1}) - (w, r))\]
where $b_i^t$ is $i^{th}$ row of $B_k$, i.e., $b_i^t = (b_{i1}, b_{i2}, \ldots, b_{il})$. 

(iii) Set $R_k(w, r) = (\phi_l, \zeta_l)$.

From the above algorithm, we obtain $K_k = I - \hat{B}_k(D + L)^{-1}\hat{B}_k$ where $B_kB_k^T = D + L + L^T$, $L$ is a strictly lower triangular matrix, and $D$ is a diagonal matrix.

The following theorem are in [6].

**Theorem 4.2.** Let $A_k$ be a sparse symmetric positive definite operator from $M_k$ to $M_k$ and let $A_k = l + d + l^t$ where $l$ is a lower triangular part and $d$ is a diagonal part of $A$. Let $R_k = (d + l)^{-1}$. Then $R_k$ satisfy the following property: There is a constant $C_R$ which does not depend on $k$ such that

$$\frac{\|u\|^2}{\lambda_k} \leq C_R(\bar{R}_ku, u)_k, \quad \text{for all } u \in M_k.$$

Here, $\bar{R}_k$ is either $(I - K_k^*K_k)A_k^{-1}$ or $(I - K_kK_k^*)A_k^*$ and $K_k = I - R_kA_k$. $\lambda_k$ is the largest eigenvalue of $A_k$.

**Lemma 4.3.** The Kaczmarz smoother is satisfied the smoothing assumption (SM).

**Proof.** Let $A_k = (\hat{B}_k)^2$ and $M_k = V_k \times \hat{Q}_k$ in Theorem 4.2, then we have

$$\frac{\|[(u, p)]^2}{A_k^2} \leq C_R\left[(\hat{B}_k)^2(u, p), (u, p)\right] - ((\hat{B}_k)^2(I - (D + L)^{-2}(\hat{B}_k)^2)(u, p), (I - (D + L)^{-2}(\hat{B}_k)^2)(u, p)),$$

i.e.,

$$\frac{\|\hat{B}_k(u, p)\|^2}{A_k^2} \leq \left[(\hat{B}_k(u, p), \hat{B}_k(u, p))\right] - ((I - \hat{B}_k(D + L)^{-2}\hat{B}_k)^2\hat{B}_k(u, p), (I - \hat{B}_k(D + L)^{-2}\hat{B}_k)^2\hat{B}_k(u, p))\right]$$

$$= \left((\hat{B}_k(u, p), \hat{B}_k(u, p)) - (K_k\hat{B}_k(u, p), K_k\hat{B}_k(u, p))\right).$$

In above, let $\hat{B}_k(u, p) = (v, q)$, then we have

$$\frac{\|([v, q])^2}{A_k^2} \leq C_R\left(((v, q), (v, q)) - (K_k(v, q), K_k(v, q))\right)$$

$$= C_R((I - K_k^*K_k)(v, q), (v, q)).$$

(4.1)
In (4.1), we let \((v, q) = K_k^m(w, s)\), then we have
\[
\|\| K_k^m(w, s) \|\|_{2,k}^2 \leq C_R \Lambda_k^2 \left( (I - K_k^t K_k)(K_k^t K_k)^m(w, s), (w, s) \right)
\]
\[
\leq C_R \Lambda_k^2 \frac{1}{m} \sum_{i=0}^{m-1} \left( (I - K_k^t K_k)(K_k^t K_k)^i(w, s), (w, s) \right)
\]
\[
= C_R \Lambda_k^2 \frac{1}{m} \| \| (w, s) \| \|_{0,k}^2
\]
because spectral radius of \(K_k\) is less than 1 and \(\Lambda_k = Ch_k^{-2}\).

5. Numerical Experiments

Multigrid algorithm described in Section 4 was applied to the pure displacement boundary value problem (1.1) with \(\mu = 1\). The domain \(\Omega\) is the unit square. In Table I and II, \(\nu = \lambda/(2(1 + \lambda))\) is the Poisson ratio, \(h\) represents the lengths of the horizontal and vertical sides of the triangles in the triangulation, the numbers \(n\) represent Multigrid iterations required to achieved an \(L^2\) relative error of less than 1% in the displacements. In the first row, smoothing number represent the number of smoothing steps in Multigrid algorithm. Table I represent the number of Multigrid iterations with Kaczmarz smoother and Table II represent the number Multigrid iterations with Richardson smoother.

The results clearly illustrate that the number of Multigrid iterations is independent of the Poisson ratio and Multigrid algorithm with Kaczmarz smoother is slightly better than Multigrid algorithm with Richardson smoother.

References

### Table I. Multigrid iterations with Kacmarz smoother.

<table>
<thead>
<tr>
<th>smoothing number</th>
<th>$h$</th>
<th>n</th>
<th>n</th>
<th>n</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(0.5)$^4$</td>
<td>38</td>
<td>34</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>74</td>
<td>67</td>
<td>62</td>
<td>60</td>
</tr>
<tr>
<td>0.45</td>
<td>(0.5)$^4$</td>
<td>25</td>
<td>23</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>49</td>
<td>46</td>
<td>43</td>
<td>41</td>
</tr>
<tr>
<td>0.495</td>
<td>(0.5)$^4$</td>
<td>22</td>
<td>21</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>43</td>
<td>43</td>
<td>42</td>
<td>40</td>
</tr>
<tr>
<td>0.4995</td>
<td>(0.5)$^4$</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>40</td>
<td>41</td>
<td>40</td>
<td>38</td>
</tr>
</tbody>
</table>

### Table II. Multigrid iterations with Richardson smoother.

<table>
<thead>
<tr>
<th>smoothing number</th>
<th>$h$</th>
<th>n</th>
<th>n</th>
<th>n</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(0.5)$^4$</td>
<td>63</td>
<td>49</td>
<td>44</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>125</td>
<td>98</td>
<td>89</td>
<td>85</td>
</tr>
<tr>
<td>0.45</td>
<td>(0.5)$^4$</td>
<td>71</td>
<td>51</td>
<td>42</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>136</td>
<td>99</td>
<td>84</td>
<td>75</td>
</tr>
<tr>
<td>0.495</td>
<td>(0.5)$^4$</td>
<td>67</td>
<td>44</td>
<td>35</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>125</td>
<td>81</td>
<td>65</td>
<td>58</td>
</tr>
<tr>
<td>0.4995</td>
<td>(0.5)$^4$</td>
<td>95</td>
<td>66</td>
<td>54</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>(0.5)$^5$</td>
<td>122</td>
<td>78</td>
<td>63</td>
<td>57</td>
</tr>
</tbody>
</table>


Korea Advanced Institute of Science and Technology, Taejon, Korea, 305-701
email: kks002@math.kaist.ac.kr
ERGODICITY AND RANDOM WALKS ON A COMPACT GROUP

GEON HO CHOE

Abstract. Let $G$ be a finite group with a probability measure. We investigate the random walks on $G$ in terms of ergodicity of the associated skew product transformation.

1. Introduction

Random shuffling of $n$ cards may be regarded as a random walk on the symmetric group on $n$ symbols. The speed of convergence to the perfect random shuffle has been studied using the Fourier analysis on the symmetric group $S_n$ of permutations on $n$ symbols, i.e., the group representations of $S_n$. When a probability distribution $\mu$ on $S_n$ is given, the probability of $g \in S_n$ is the probability of shuffling $n$ cards according to the rule given by $g$. The random walk given by a finite sequence $g_1g_2 \cdots g_k$ is the shuffling of $n$ cards obtained by consecutively applying $k$ shuffling methods $g_i$. We may say that a finitely many applications of shuffles produce perfect randomness if the probability of finding the walker at any point in $S_n$ is $1/|S_n|$. The probability distribution for the location of the walker after $k$ random applications of elements in $S_n$ is given by the convolution $\mu^*k$, hence the problem is to check the speed of the convergence of $\mu^*k$ to the Haar measure. See [1]. For additional information, see [3].

In this paper we investigate the properties of random walks on a general finite group from the viewpoint of ergodic theory, especially in terms of skew product transformations. Instead of considering the convolution of measures as done in other literature, we will focus on density functions, which are $L^2$-functions, hence our analysis will be simpler.

Let $G$ be a compact group with the right-invariant Haar measure $m$. We consider a measurable function $h(x)$ defined on $G$ satisfying (i) $h(x) \geq 0$, (ii) $\int_G h \, d\mu = 1$, (iii) $h$ is bounded. Put

$$\text{supp } h = \{x \in G : h(x) > 0\}.$$

The function $h$ may be regarded as a probability distribution on $G$: Put $d\nu = h(x) \, dm$. Then $\nu$ is a probability measure satisfying $\nu(\text{supp } h) = 1$. If $h = 1$ then $\nu = m$, and it is sometimes called the uniform distribution on $G$.

1991 Mathematics Subject Classification. Primary 94A17, secondary 60J10.

Key words and phrases. Random walk, finite group, group representation, skew product transformation.
We say that \( B \) is included in \( A \) modulo measure zero sets if \( m(B \setminus A) = 0 \) and that \( A \) and \( B \) are equal modulo measure zero sets if \( m(A \setminus B) + m(B \setminus A) = 0 \). If there is no danger of confusion, all set inclusions are understood as inclusions modulo measure zero sets. For finite groups there is no need to pay extra attention to such subtleties.

Recall that for \( f_1, f_2 \in L^1(G, m) \) their convolution is defined by
\[
f_1 \ast f_2(x) = \int_G f_1(xy^{-1})f_2(y) \, dm(y).
\]

Let \( f^{*n} \) denote the \( n \)-times convolution of \( f \) with itself, i.e., \( f^{*n} = \underbrace{f \ast f \ast \cdots \ast f}_n \).

Observe that \( f^{*n}(x) \geq 0 \) and \( \int_G f^{*n} \, dm = 1 \). On a finite group \( G \) we have \( f^{*n}(x) > 0 \) if and only if \( x \) is a product of exactly \( n \) elements in the support of \( f \).

## 2. Fourier analysis on a finite group

In this section we use results on the representations of a compact group. For the references, consult \([2], [9]\).

**Lemma 2.1.** Let \((X, \nu)\) be a probability space. (i) Let \( \psi : X \to \mathbb{C} \) be a measurable function such that \( |\psi(x)| \leq 1 \). If \( \int_X \psi(x) \, d\nu = 1 \), then \( \psi(x) \) is a constant of modulus \( 1 \) a.e. with respect to \( \nu \).

(ii) Let \( u : X \to \mathbb{C}^d \) be a vector-valued measurable function such that \( \| u(x) \| \leq 1 \). If \( \int_X u(x) \, d\nu = 1 \), then \( u(x) \) is a constant vector of norm \( 1 \) a.e. with respect to \( \nu \).

**Proof.** (i) Suppose \( \int_X \psi(x) \, d\nu = 1 \). Recall the Cauchy-Schwarz inequality \( \int f g \, d\nu \leq \int |f|^2 \, d\nu \int |g|^2 \, d\nu \). By taking \( f = \psi \), \( g = 1 \), we have
\[
1 = \left| \int_X f g \, d\nu \right|^2 \leq \int_X |f|^2 \, d\nu \int |g|^2 \, d\nu = \int_X |f|^2 \, d\nu \leq 1.
\]
Since the equality holds, \( f = \psi \) is a constant multiple of \( g = 1 \).

(ii) Suppose \( \| \int_X u(x) \, d\nu \| = 1 \). Put \( v = \int_X u(x) \, d\nu \). Then \( \| v \| \leq 1 \) and
\[
1 = \left\| \int_X u(x) \, d\nu \right\|^2 = (\int_X u(x) \, d\nu, v) = \int_X (u(x), v) \, d\nu.
\]
Now put \( \psi(x) = (u(x), v) \) and apply the part (i). Then \( (u(x), v) = \lambda, |\lambda| = 1 \), which is possible only if \( u(x) = \lambda v \). Incidentally, \( \lambda = 1 \) since \( v = \int u(x) \, d\nu \).

**Definition 2.2.** Given a group \( G \) and its subset \( S \), let \( \text{gen}(S) \) denote the set of all products of elements in \( S \). (Taking inverses of elements in \( S \) are not allowed.) We say that \( \text{gen}(S) \) is generated by \( S \).

For a finite group \( G \), \( \text{gen}(S) \) is a subgroup. To see why, observe that for every \( s \in S \) the elements \( s, s^2, s^3, \ldots, s^n, \ldots \) are contained in a finite set \( G \). So \( s^j = s^k \) for some \( j > k \geq 1 \), and \( s^{j-k} = e \in \text{gen}(S) \), hence \( s^{-1} = s^{j-k} \in \text{gen}(S) \). Therefore, if \( s_{i_1}^{p_1} \cdots s_{i_m}^{p_m} \) is in \( \text{gen}(S) \) then its inverse \( s_{i_1}^{-1} \cdots s_{i_m}^{-1} \) is also in \( \text{gen}(S) \).

In general, \( \text{gen}(S) \) need not be a subgroup if \( G \) is not finite. For example, if \( G = \{ z \in \mathbb{C} : |z| = 1 \} \) and \( S = \{ e^{2\pi i \theta} \}, \theta \) irrational, then \( \text{gen}(S) = \{ e^{2\pi in\theta} : n \geq 1 \} \), which does not contain the identity 1.
Given a compact group $G$ and a representation $\rho$ of $G$, $x \in G$, we define the operator norm of $\rho(x)$ by
\[
\|\rho(x)\| = \sup_{\vec{v} \neq 0} \frac{||\rho(x)\vec{v}||_2}{||\vec{v}||_2},
\]
for the Euclidean norm $|| \cdot ||_2$ on $\mathbb{C}^{d_\rho}$. Define $\hat{f}(\rho) = \int_G f(x)\rho(x)\,d\mu(x)$. If $\rho$ is unitary, then $||\hat{f}(\rho)|| \leq 1$ by the Minkowski inequality. Note that $\hat{f}_1 * \hat{f}_2(\rho) = \hat{f}_1(\rho)\hat{f}_2(\rho)$ and $\hat{f}^{\ast n}(\rho) = \hat{f}(\rho)^n$. For $\rho = 1$ we have $\hat{f}(1) = 1$. Notation: (i) Let $d_\rho$ denote the dimension of $\rho$, i.e., $\rho(x)$ is a $d_\rho \times d_\rho$ matrix. (ii) The set of all irreducible unitary representations of $G$ is denoted by $\hat{G}$.

Lemma 2.3. Let $\rho$ be a unitary representation of $(G,m)$. (i) If $\hat{f}(\rho)$ is an identity matrix for $\rho \neq 1$, then $\text{supp } f$ is contained in a closed normal subgroup $H$ of $G$, $H \neq G$.

(ii) If $||\hat{f}(\rho)|| = 1$ for $\rho \neq 1$, then $\text{supp } f$ is contained in a coset of a closed subgroup $H$ of $G$, $H \neq G$.

Proof. (i) If $\hat{f}(\rho) = \int_G f(x)\rho(x)\,d\mu = I$, then $\rho(x) = I$ on $\text{supp } f$, hence $\text{supp } f$ is included in $H = \{x : \rho(x) = I\}$.

(ii) If $||\int_G f(x)\rho(x)\,d\mu|| = 1$, then there exists a constant unitary matrix $A$ such that $\rho(x) = A \,\ast\, e.$ on $\text{supp } f$, hence $\text{supp } f \subset \{x : \rho(x) = A\}$. Choose $g \in \text{supp } f$ such that $\rho(g) = A$. Then $\text{supp } f$ is included in the coset $gH$ where $H = \{x : \rho(x) = I\}$. □

Theorem 2.4. Let $(G,m)$ be a compact group.

(i) If $\text{supp } f$ is not contained in any closed normal subgroup $H$ of $G$, $H \neq G$, then $\frac{1}{n}\sum_{k=1}^{n} f^{*k}$ converges to 1 in $L^2$ as $n \to \infty$. In this case, $m(\bigcup_{k=1}^{\infty} \text{supp } f^{*k}) = 1$.

(ii) If $\text{supp } f$ is not contained in any coset of any closed subgroup $H$ of $G$, $H \neq G$, then $f^{*n}$ converges to 1 in $L^2$ as $n \to \infty$. In this case, $m(\text{supp } f^{*n}) = 1$ for sufficiently large $n$ and
\[
||f^{*n}(x) - 1||_{L^2}^2 = \sum_{\rho \neq 1} d_\rho \text{Tr} \left( \hat{f}(\rho)^n (\hat{f}(\rho)^n)^* \right).
\]

Proof. Recall that the functions $\sqrt{d_\rho}\rho_{ij}$, $\rho \in \hat{G}$, form an orthonormal basis for $L^2(G,\mu)$. Part (ii) is simpler, and will be proved first. (ii) Lemma 2.3(ii) implies $||\hat{f}(\rho)|| < 1$ for $\rho \neq 1$. From the Fourier inversion formula, we have
\[
f^{*n}(x) = \sum_{\rho \in \hat{G}} \sqrt{d_\rho} \text{Tr} \left( \hat{f}(\rho)^n \sqrt{d_\rho} \rho(x^{-1}) \right),
\]
hence the Parseval’s relation gives
\[
||f^{*n}(x) - 1||_{L^2}^2 = \sum_{\rho \neq 1} d_\rho \left| \text{Tr} \left( \hat{f}(\rho)^n \sqrt{d_\rho} \rho(x^{-1}) \right) \right|^2.
\]
Since $\text{Tr}(AB) = \sum_i (AB)_{ii} = \sum_{ij} A_{ij} B_{ji}$, we have
\[
\text{Tr} \left( \hat{f}(\rho)^n \sqrt{d_\rho} \rho(x^{-1}) \right) = \sum_{ij} (\hat{f}(\rho)^n)_{ij} \sqrt{d_\rho} \rho_{ij}(x),
\]
and by the orthogonality relation we obtain
\[
\| \text{Tr} \left( \hat{f}(\rho)^n \sqrt{d_\rho} \rho(x^{-1}) \right) \|_{L^2}^2 = \sum_{ij} |(\hat{f}(\rho)^n)_{ij}|^2 = \text{Tr}[\hat{f}(\rho)^n(\hat{f}(\rho)^n)^*],
\]
which converges to zero as \( n \to \infty \) since \( \|\hat{f}(\rho)^n(\hat{f}(\rho)^n)^*\| \leq \|\hat{f}(\rho)\|^2n \) and \( \|\hat{f}(\rho)\| < 1 \).

(i) Lemma 2.3(i) implies \( \hat{f}(\rho) \neq I \) for \( \rho \neq 1 \). Put
\[
M_n = \frac{1}{n} \sum_{k=1}^n \hat{f}(\rho)^k = \frac{1}{n} \left[ I - \hat{f}(\rho)^{n+1} \right] \left[ I - \hat{f}(\rho) \right]^{-1}.
\]
Then \( \|M_n\| \leq \frac{1}{n} \cdot C \) for some constant \( C > 0 \). From the Parseval’s relation
\[
\left\| \frac{1}{n} \sum_{k=1}^n f^{*k} - 1 \right\|_{L^2}^2 = \sum_{\rho \neq 1} d_\rho \left\| \text{Tr} \left( M_n \sqrt{d_\rho} \rho(x^{-1}) \right) \right\|_{L^2}^2
= \sum_{\rho \neq 1} d_\rho \sum_{ij} |(M_n)_{ij}|^2
= \sum_{\rho \neq 1} d_\rho \text{Tr} \left( M_n M_n^* \right),
\]
which converges to zero as \( n \to \infty \) since \( \|M_n M_n^*\| \leq \|M_n\|^2 \leq C^2/n^2 \).

\[ \square \]

**Lemma 2.5.** Let \( \rho \neq 1 \) be an irreducible unitary representation of \( G \). If \( \text{supp} \ f \) is not contained in any closed normal subgroup \( H \) of \( G \), \( H \neq G \), then \( I - \hat{f}(\rho) \) is invertible.

**Proof.** Suppose not. Then there exists a nonzero vector \( v \) such that \( (I - \hat{f}(\rho))v = 0 \). Hence \( \hat{f}(\rho)v = v \) and \( \int_G f(x)\rho(x)v \, dm = v \). Since \( \|\rho(x)v\| = \|v\| \) for all \( x \in G \), we have \( \rho(x)v = v \) for a.e. \( x \in \text{supp} \ f \). From Theorem 2.4(i) we observe that \( \bigcup_{k=1}^\infty \text{supp} f^{*k} \) is a dense subset in \( G \). Since \( \rho: G \to \mathbb{C}^{d_\rho 	imes d_\rho} \) is continuous, we see that \( \rho(x)v = v \) for all \( x \in G \), which contradicts the irreducibility of \( \rho \). \( \square \)

For a finite group \( G \), \( L^2 \)-convergence can be replaced by pointwise convergence since the \( L^2 \)-norm satisfies \( \|h\|^2 = \frac{1}{|G|} \sum_{x \in G} |h(x)|^2 \). Or we may observe that \( L^p(G) \), \( 1 \leq p \leq \infty \), is isomorphic to \( \mathbb{C}^n \) where \( n = |G| \), the number of elements in \( G \) and that on a finite-dimensional space all the norms are equivalent.

Consider a probability distribution on \( G \), i.e., a nonnegative function \( f \) on \( G \) satisfying \( \frac{1}{|G|} \sum_{x \in G} f(x) = 1 \). For two arbitrary elements \( x, y \in G \), suppose that the transition probability of moving from \( x \) to \( y \) in one step is given by \( \frac{1}{|G|} f(yx^{-1}) \). For example, \( \frac{1}{|G|} f(y) \) is the probability of moving to \( y \) from the group identity element \( e \). Thus we have a Markov chain given by random walks on \( G \). A simple example is given by the 1-dimensional random walk on \( G = \mathbb{Z}_n = \{0, 1, \ldots, n-1\} \) where \( f(1) = \frac{n}{2} = f(-1) \). The random walker moves up or down, with probability \( \frac{1}{2} \) in each case, depending on the outcome of a fair coin tossing.
In [1] the probability measure \( Q(\{x\}) = \frac{1}{|G|} f(x) \) is used. In this case the \( L^1 \)-norm of \( Q^n - U \) is considered where \( U(\{x\}) = \frac{1}{|G|} \) is the Haar measure on \( G \) even though the \( L^2 \)-norm is easier to compute. The reason for this is the following: If \( |G| \) is even and if \( Q \) is uniformly distributed on half the points (that is, equal to \( \frac{2}{|G|} \)) and zero on the half. Then \( ||Q - U||_2 = \frac{1}{\sqrt{|G|}} \) is close to zero for \( |G| \) large. This makes it difficult to compare the convergence rates of \( Q^n \) to the uniformity as \( n \to \infty \) for different values of \( |G| \). But if we consider \( f \) instead of \( Q \), this problem disappears: In the previous example, \( |f(x) - 1| = 1 \) for every \( x \), hence \( ||f - 1||_2 = 1 \). Furthermore, we can easily extend our argument to groups with infinitely many elements.

Let \( P \) be the transition matrix associated with the random walk, i.e., \( P_{x,y} = \frac{1}{|G|} f(yx^{-1}) \). Then \( (P^2)_{x,z} = \sum_y P_{x,y} P_{y,z} \) is the probability to arrive at \( z \) from \( x \) in two steps, and so on. Note that

\[
(f * f)(x) = \frac{1}{|G|} \sum_{y \in G} f(xy^{-1}) f(y) = |G| \sum_{y \in G} P_{y,x} P_{e,y} = |G|(P^2)_{e,x},
\]

and similarly \( \frac{1}{|G|} f^{*k}(x) = (P^k)_{e,x} \) is the probability that the random walker is found at \( x \) at time \( k \) if he starts from \( e \). Observe that \( f^{*k}(x) > 0 \) if and only if \( x \) is a product of \( k \) elements \( g_1, \ldots, g_k \) such that \( f(g_i) > 0 \).

**Theorem 2.6.** Let \( G \) be a finite group. (i) If \( \text{supp } f \) is not contained in any normal subgroup \( H \) of \( G \), \( H \neq G \), then every \( g \in G \) is a product of elements in \( \text{supp } f \).

(ii) If \( \text{supp } f \) is not contained in a coset of any subgroup \( H \) of \( G \), \( H \neq G \), then there exists \( N \) such that every \( g \in G \) is a product of exactly \( n \) elements in \( \text{supp } f \) for any \( n \geq N \).

**Proof.** (i) For any small \( \epsilon > 0 \), Theorem 2.4(i) implies that for sufficiently large \( n \) such that \( \frac{1}{n} \sum_{k=1}^{n} f^{*k} - 1 | < \epsilon \) for every \( x \), hence \( \frac{1}{n} \sum_{k=1}^{n} f^{*k} > 1 - \epsilon \), which implies \( \bigcup_{k=1}^{n} \text{supp } f^{*k} = G \). Thus for every \( g \in G \) there exists \( k \) such that \( f^{*k}(g) > 0 \).

(ii) For any small \( \epsilon > 0 \), Theorem 2.4(ii) implies that there exists \( N \) such that if \( n \geq N \) then \( |f^{*n} - 1| < \epsilon \) for every \( x \), hence \( f^{*n}(x) > 1 - \epsilon \). \( \square \)

**Remark 2.7.** (i) The following statements are equivalent: (a) Markov chain is ergodic, (b) \( P \) is irreducible, and (c) \( \text{supp } f \) is not contained in any normal subgroup \( H \) of \( G \), \( H \neq G \). (ii) The following statements are equivalent: (a) Markov chain is weak-mixing, (b) Markov chain is mixing, (c) \( P \) is aperiodic, and (d) \( \text{supp } f \) is not contained in a coset of any subgroup \( H \) of \( G \), \( H \neq G \). The proofs are found in [7],[10].

When the group is the circle group and the random walk is the rotation by \( e^{2\pi i \theta} \) or \( e^{-2\pi i \theta} \), \( \theta \) irrational, depending on the outcome of coin tossing. This was studied in [8].
3. Skew product transformation

Let $G$ be a compact group with its right Haar measure $\lambda$, and $(X, \mu)$ a probability space and $T : X \rightarrow X$ an ergodic measure preserving transformation. Given a function $\phi : X \rightarrow G$, define a skew product transformation $T_\phi : G \times X \rightarrow G \times X$ by $(g, x) \mapsto (g \cdot \phi(x), Tx)$. Then $T_\phi$ preserves the product measure $\lambda \times \mu$. See [4]. The ergodicity of $T_\phi$ can be checked by the decomposition of $L^2(G \times X)$. It is known that the matrix coefficients of the irreducible unitary representations form an orthogonal basis for $L^2(G, \lambda)$. Take any irreducible unitary representation $\rho$ and let $(\rho_{ij})$ be its matrix representation. Then

$$U_{T_\phi}(\rho_{ij}(g)f(x)) = \rho_{ij}(g\phi(x))f(Tx) = \sum_k \rho_{ik}(g)\rho_{kj}(\phi(x))f(Tx).$$

Hence we have the following $U_{T_\phi}$-invariant orthogonal decomposition:

$$L^2(G \times X) = \oplus L^2_\rho(G \times X)$$

where the subspace $L^2_\rho(G \times X)$ is spanned by functions of the form $\rho_{ij}(g)f(x), f \in L^2(X)$.

For $\rho = 1$ two Hilbert spaces $L^2_\rho(G, X)$ and $L^2(X)$ are identical. Let $U_1$ be the operator restricted on $L^2_\rho(G, X)$. Then two operators $U_1$ on $L^2_\rho(G, X)$ and $U_T$ on $L^2(X)$ are unitarily equivalent. Since $T$ is ergodic, there is no nonconstant eigenfunction of $U_1$ for the eigenvalue 1. The following is known but its proof is included here for the sake of completeness and notational convenience. Consult [6] for more details.

Fact 3.1. (i) The skew product transformation $T_\phi : G \times X \rightarrow G \times X$ is not ergodic if and only if there exists an irreducible representation $\rho \neq 1$ satisfying $\rho(\phi(x))h(Tx) = h(x)$ for some $h = (h_i)_{1 \leq i \leq d}$, $h_i \in L^2(X)$, $h_i \neq 0$, where $d$ is the dimension of $\rho$.

(ii) It is not weak-mixing if and only if there exists an irreducible representation $\rho \neq 1$ and some constant $\lambda \in \mathbb{C}, |\lambda| = 1$, satisfying $\rho(\phi(x))h(Tx) = \lambda h(x)$ for some $h = (h_i)_{1 \leq i \leq d}$, $h_i \in L^2(X)$, $h_i \neq 0$, $f \neq 0$, where $d$ is the dimension of $\rho$.

Proof. (i) Suppose $T_\phi$ is not ergodic. Then there exists a nonconstant function $h(g, x)$ in $L^2_\rho(G \times X), \rho \neq 1$, such that $h(g\phi(x), Tx) = h(g, x)$. Put $h(g, x) = \sum_{i,j} \rho_{ij}(g)f_{ij}(x)$
and let $\rho^T$ denote the transpose of $\rho$. Then
\[
h(g\phi(x), Tx) = \sum_{i,j} \rho_{ij}(g\phi(x)) f_{ij}(Tx)
= \sum_{i,j} \left[ \rho(g(\phi(x))) \right]_{ij} f_{ij}(Tx)
= \sum_{i,j} \left( \sum_k \rho_{ik}(g) \rho_{kj}(\phi(x)) \right) f_{ij}(Tx)
= \sum_{i,k} \rho_{ik}(g) \left( \sum_j \rho_{kj}(\phi(x)) f_{ij}(Tx) \right)
= \sum_{i,j} \rho_{ij}(g) \left( \sum_k \rho_{jk}(\phi(x)) f_{ik}(Tx) \right),
\]
and for every $i, j$ we have
\[
f_{ij}(x) = \sum_k \rho_{jk}(\phi(x)) f_{ik}(Tx) = \sum_k f_{ik}(Tx) \rho_{kj}^T(\phi(x)).
\]

Define a matrix-valued function $F$ on $X$ by $F(x) = [f_{ij}(x)]$. Then $F(x) = F(Tx) \rho^T(\phi(x))$, and $F(x)^T = \rho(\phi(x)) F(Tx)^T$. Choose a nonzero column of $F(x)^T$ and call it $f(x)$.

Conversely, suppose that there exists an irreducible representation $\rho \neq 1$ and a nonzero vector-valued function $f(x)$ such that $f(x) = \rho(\phi(x)) f(Tx)$. Let $v(g, x)$ be a vector-valued function on $G \times X$ defined by $v(g, x) = \rho(g) f(x)$. It is not a constant function on $G \times X$ since $f(x) \neq 0$. Then
\[
v(g\phi(x), Tx) = \rho(g\phi(x)) f(Tx)
= \rho(g) \rho(\phi(x)) f(Tx)
= \rho(g) f(x) = v(g, x).
\]
Note that every component function of $v(g, x)$ is $T_\phi$-invariant and that not all of them are constant. Therefore $T_\phi$ is not ergodic.

(ii) The proof is almost identical with the case (i). \qed

Here is a connection between random walks and skew product transformations. Let $G$ be a finite group with a probability distribution $\mu$ such that the probability of random walk from $g \in G$ to $h \in G$ is given by $\mu(hg^{-1})$. Let $\{g_1, g_2, \ldots, g_k\} \subset G$ be the support of $\mu$ and put $p_i = \mu(g_i)$. Take an alphabet $A = \{1, 2, \ldots, k\}$ of $k$ symbols and define a Bernoulli shift space $X = \prod_{i=1}^\infty A$ with the product measure defined by the probability distribution $(p_1, \ldots, p_k)$. Define the left shift transformation $T$ by $(Tx)_i = x_{i+1}$ for $x \in X$. Recall that $T$ is a measure preserving ergodic transformation. Define $\phi : X \to G$ by $\phi(x) = g_{x_1}$ for $x = (x_1, x_2, x_3, \ldots)$. Note that $\phi$ depends only on the first component in $x$ so we write $\phi(x_1)$ to emphasize the fact. Using such $\phi$ we finally define the skew product transformation $T_\phi : G \times X \to G \times X$ by $T_\phi(g, x) = (g \cdot \phi(x), Tx)$.
bottom row of the following commutative diagram we see that the random walk under
consideration is the projection of the skew product transformation where the projection
\( \pi : G \times X \to G \) is defined by \( \pi(g, x) = g \).

\[
\begin{array}{ccc}
(g, x) & \xrightarrow{T_\phi^n} & (g \cdot \phi(x) \cdot \phi(Tx) \cdots \phi(T^{n-1}x), T^nx) \\
\downarrow \pi & & \downarrow \pi \\
g & \xrightarrow{\text{random walks}} & g \cdot gx_1 \cdot gx_2 \cdots gx_n
\end{array}
\]

In the following we consider an absolutely continuous measure \( f \, d\mu \). If a discrete
measure \( \nu \) is considered as in irrational random walks on the unit circle, we have
to replace the conditions of \( \text{supp} f \) by the weak convergence of \( \frac{1}{n} \sum_{k=1}^{n} \nu^k \) and \( \nu^n \),
respectively, to obtain the ergodicity and the weak-mixing property.

**Theorem 3.2.** (i) If \( \text{supp} f \) is not contained in any closed normal subgroup \( H \) of \( G \),
\( H \neq G \), then \( T_\phi \) is ergodic.

(ii) If \( \text{supp} f \) is not contained in a coset of any closed normal subgroup \( H \) of \( G \),
\( H \neq G \), then \( T_\phi \) is weak-mixing.

**Proof.** We prove the case (i). The proof for the second case is similar. Suppose \( T_\phi \)
is not ergodic. Then the coboundary condition holds: There exists an irreducible unitary
measure \( \nu \) with \( \nu(\mathcal{H}) = 1 \) and a nonzero vector-valued function \( \phi \) such that every
\( \phi : X \to \mathbb{C} \) is an \( L^2 \)-function where \( d \) is the dimension of \( \phi \) and \( \rho(\phi(x))q(Tx) = q(x) \).
For the sake of notational convenience in the remainder of the proof we write \( \phi(x_1) \) in
place of \( \phi(x) \). Then

\[
\rho(\phi(x_1))q(x_2, x_3, x_4, \ldots) = q(x_1, x_2, x_3, \ldots). \tag{*}
\]

By applying \( \rho(\phi(x_1)) \) to the both sides, we have

\[
\rho(\phi(x_1))\rho(\phi(x_2))q(x_3, x_4, x_5, \ldots) = \rho(\phi(x_1))q(x_2, x_3, x_4, \ldots)
= q(x_1, x_2, x_3, \ldots).
\]

In general,

\[
\rho(\phi(x_1) \cdots \phi(x_n))q(x_{n+1}, x_{n+2}, x_{n+3}, \ldots) = q(x_1, x_2, x_3, \ldots). \tag{**}
\]

Let \( [x_1, \ldots, x_n] \) denote the cylinder set \( \{ y \in X : y_1 = x_1, \ldots, y_n = x_n \} \) and define
\( u_n : \prod_{i=1}^{n} \{ 1, 2, \ldots, k \} \to \mathbb{C}^d \) by

\[
u_n(x_1, \ldots, x_n) = \int_{[x_1, \ldots, x_n]} q(x)
\]

where the integration of the vector-valued function \( q \) is done with respect to the
remaining variables \( x_{n+1}, x_{n+2}, \ldots \). If we integrate (**) on \([x_1, \ldots, x_n]\) with respect to the
remaining variables \( x_{n+1}, x_{n+2}, \ldots \), then

\[
\rho(\phi(x_1) \cdots \phi(x_n)) v_n = u_n(x_1, \ldots, x_n)
\]
where \( v_n \) is obtained by integrating \( u_n(x_1, \ldots, x_n) \) with respect to the variables \( x_1, \ldots, x_n \). In other words, \( v_n \) is a convex linear combination of \( u_n(x_1, \ldots, x_n) \) over \( (x_1, \ldots, x_n) \). Since \( \rho \) is unitary, for every \( (x_1, \ldots, x_n) \) we have
\[
\|u_n(x_1, \ldots, x_n)\|_{\mathbb{C}^d} = \|v_n\|_{\mathbb{C}^d},
\]
which is possible only if \( u_n(x_1, \ldots, x_n) \) is a constant vector that is nothing but \( v_n \). Therefore for any fixed \( n \) the average of \( q(x) \) on every cylinder set \([x_1, \ldots, x_n]\) is the same, hence \( q \) is a constant vector. Now we have \( \rho(\phi(x_1))q = q \) from (*) Since \( \text{supp } f \) generates almost every element in \( G \), we observe that \( \rho(g)q = q \) for every \( g \in G \), which in turn implies \( q = 0 \) by the irreducibility of \( \rho \), and we have a contradiction. \( \square \)

**Remark 3.3.** Kloss[5] considered the convergence of convolutions of a sequence of probability distributions on a compact group.

**References**


Department of Mathematics, Korea Advanced Institute of Science and Technology, Taejon, Korea
choe@euclid.kaist.ac.kr
ON THE OSTROWSKI INEQUALITY FOR THE
RIEMANN-STIELTJES INTEGRAL \( \int_a^b f(t) \, du(t) \), WHERE \( f \) IS OF
HÖLDER TYPE AND \( u \) IS OF BOUNDED VARIATION AND
APPLICATIONS

S. S. DRAGOMIR

Abstract. In this paper we point out an Ostrowski type inequality for the Riemann-
Stieltjes integral \( \int_a^b f(t) \, du(t) \), where \( f \) is of \( p-H \)-Hölder type on \([a, b]\), and \( u \) is
of bounded variation on \([a, b]\). Applications for the approximation problem of the
Riemann-Stieltjes integral in terms of Riemann-Stieltjes sums are also given.

1. Introduction

In 1938, A. Ostrowski proved the following integral inequality [1, p. 468]:

**Theorem 1.** Let \( f : [a, b] \to \mathbb{R} \) be continuous on \([a, b]\), differentiable on \((a, b)\), with its
first derivative \( f' : (a, b) \to \mathbb{R} \) bounded on \((a, b)\), that is, \( \|f'\|_{\infty} := \sup_{t \in (a, b)} |f'(t)| < \infty \).

Then

\[
\left| f(x) - \frac{1}{b-a} \int_a^b f(t) \, dt \right| \leq \left[ \frac{1}{4} + \left( \frac{x-a+b}{2} \right)^2 \right] \|f'\|_{\infty} (b-a),
\]

for all \( x \in [a, b] \).

The constant \( \frac{1}{4} \) is sharp in the sense that it cannot be replaced by a smaller one.

For a different proof than the original one provided by Ostrowski in 1938 as well as
applications for special means (identric mean, logarithmic mean, \( p \)-logarithmic mean,
etc.) and in Numerical Analysis for quadrature formulae of Riemann type, see the
recent paper [2].

In [3], the following version of Ostrowski’s inequality for the 1-norm of the first
derivatives has been given.

**Theorem 2.** Let \( f : [a, b] \to \mathbb{R} \) be continuous on \([a, b]\), differentiable on \((a, b)\), with its
first derivative \( f' : (a, b) \to \mathbb{R} \) integrable on \((a, b)\), that is, \( \|f'|_1 := \int_a^b |f'(t)| \, dt < \infty \).

Then

\[
\left| f(x) - \frac{1}{b-a} \int_a^b f(t) \, dt \right| \leq \left[ \frac{1}{2} + \frac{|x-a+b|}{2(b-a)} \right] \|f'|_1,
\]

1991 Mathematics Subject Classification. Primary 26D15, 26D20; Secondary 41A55.
Key words and phrases. Ostrowski inequality, Riemann-Stieltjes integral.
for all \( x \in [a, b] \).

The constant \( \frac{1}{2} \) is sharp.

Note that the sharpness of the constant \( \frac{1}{2} \) in the class of differentiable mappings whose derivatives are integrable on \((a, b)\) has been proven in the paper [5].

In [3], the authors applied (1.2) for special means and for quadrature formulae of Riemann type.

The following natural extension of Theorem 2 has been pointed out by S.S. Dragomir in [6].

**Theorem 3.** Let \( f : [a, b] \to \mathbb{R} \) be a mapping of bounded variation on \([a, b]\) and \( \nabla b a (f) \) its total variation on \([a, b]\). Then

\[
(1.3) \quad \left| f(x) - \frac{1}{b-a} \int_a^b f(t) \, dt \right| \leq \left[ \frac{1}{2} + \frac{|x-a+b|}{b-a} \right]^{\frac{b}{a}} \nabla (f),
\]

for all \( x \in [a, b] \). The constant \( \frac{1}{2} \) is sharp.

In [6], the author applied (1.3) for quadrature formulae of Riemann type as well as for Euler’s Beta mapping.

In this paper we point out some generalizations of (1.3) for the Riemann-Stieltjes integral \( \int_a^b f(t) \, du(t) \) where \( f \) is of Hölder type and \( u \) is of bounded variation. Applications to the problem of approximating the Riemann-Stieltjes integral in terms of Riemann-Stieltjes sums are also given.

### 2. Some Integral Inequalities

The following theorem holds.

**Theorem 4.** Let \( f : [a, b] \to \mathbb{R} \) be a \( p - H \)-Hölder type mapping, that is, it satisfies the condition

\[
(2.1) \quad |f(x) - f(y)| \leq H |x-y|^p, \quad \text{for all } x, y \in [a, b];
\]

where \( H > 0 \) and \( p \in (0, 1] \) are given, and \( u : [a, b] \to \mathbb{R} \) is a mapping of bounded variation on \([a, b]\). Then we have the inequality

\[
(2.2) \quad \left| f(x)(u(b) - u(a)) - \int_a^b f(t) \, du(t) \right| \leq H \left[ \frac{1}{2} (b-a) + \left| x - \frac{a+b}{2} \right| \right]^{\frac{p}{a}} \nabla (u),
\]

for all \( x \in [a, b] \), where \( \nabla a (u) \) denotes the total variation of \( u \) on \([a, b]\). Furthermore, the constant \( \frac{1}{2} \) is the best possible, for all \( p \in (0, 1] \).
Proof. It is well known that if \( g : [a, b] \to \mathbb{R} \) is continuous and \( v : [a, b] \to \mathbb{R} \) is of bounded variation, then the Riemann-Stieltjes integral \( \int_a^b g(t) \, dv(t) \) exists and the following inequality holds:

\[
(2.3) \quad \left| \int_a^b g(t) \, dv(t) \right| \leq \sup_{t \in [a,b]} |g(t)| \int_a^b (v).
\]

Using this property, we have

\[
(2.4) \quad \left| f(x)(u(b) - u(a)) - \int_a^b f(t) \, du(t) \right| = \left| \int_a^b (f(x) - f(t)) \, du(t) \right| \leq \sup_{t \in [a,b]} |f(x) - f(t)| \int_a^b (u).
\]

As \( f \) is of \( p - H \)-Hölder type, we have

\[
\sup_{t \in [a,b]} |f(x) - g(t)| \leq \sup_{t \in [a,b]} [H |x - t|^p] = H \max \{ (x - a)^p, (b - x)^p \} = H \max \{ x - a, b - x \}^p = H \left[ \frac{1}{2} (b - a) + \left| x - \frac{a + b}{2} \right| \right]^p.
\]

Using (2.4), we deduce (2.2).

To prove the sharpness of the constant \( \frac{1}{2} \) for any \( p \in (0, 1) \), assume that (2.2) holds with a constant \( C > 0 \), that is,

\[
(2.5) \quad \left| f(x)(u(b) - u(a)) - \int_a^b f(t) \, du(t) \right| \leq H \left[ C (b - a) + \left| x - \frac{a + b}{2} \right|^p \right] \int_a^b (u),
\]

for all \( f, p - H \)-Hölder type mappings on \([a, b]\) and \( u \) of bounded variation on the same interval.

Choose \( f(x) = x^p \ (p \in (0, 1)), \ x \in [0, 1] \) and \( u : [0, 1] \to [0, \infty) \) given by

\[
u(x) = \begin{cases} 0 & \text{if } x \in [0, 1) \ 1 & \text{if } x = 1 \end{cases}.
\]

As

\[
|f(x) - f(y)| = |x^p - y^p| \leq |x - y|^p
\]

for all \( x, y \in [0, 1], \ p \in (0, 1) \), it follows that \( f \) is of \( p - H \)-Hölder type with the constant \( H = 1 \).
By using the integration by parts formula for Riemann-Stieltjes integrals, we have:

$$\int_0^1 f(t) \, du(t) = f(t) \, u(t)|_0^1 - \int_0^1 u(t) \, df(t)$$

and

$$\int_0^1 u(t) \, df(t) = 1 - 0 = 1$$

Consequently, by (2.5), we get

$$|x^p - 1| \leq \left[ C + \left| x - \frac{1}{2} \right| \right]^p, \text{ for all } x \in [0, 1].$$

For $$x = 0$$, we get $$1 \leq (C + \frac{1}{2})^p$$, which implies that $$C \geq \frac{1}{2}$$, and the theorem is completely proved.

The following corollaries are natural.

**Corollary 1.** Let $$u$$ be as in Theorem 4 and $$f : [a, b] \to \mathbb{R}$$ an $$L$$-Lipschitzian mapping on $$[a, b]$$, that is,

(L) \quad $$|f(t) - f(s)| \leq L \, |t - s|$$ \text{ for all } t, s \in [a, b]$$

where $$L > 0$$ is fixed.

Then, for all $$x \in [a, b]$$, we have the inequality

$$|\Theta(f, u, x, a, b)| \leq L \left[ \frac{1}{2} (b - a) + \left| x - \frac{a + b}{2} \right| \right] \frac{b}{b} (u)$$

where

$$\Theta(f, u, x, a, b) = f(x) (u(b) - u(a)) - \int_a^b f(t) \, du(t)$$

is the Ostrowski’s functional associated to $$f$$ and $$u$$ as above. The constant $$\frac{1}{2}$$ is the best possible.

**Remark 1.** If $$u$$ is monotonic on $$[a, b]$$ and $$f$$ is of $$p - H$$-H"{o}lder type, then, by (2.2) we get

$$|\Theta(f, u, a, b)| \leq H \left[ \frac{1}{2} (b - a) + \left| x - \frac{a + b}{2} \right| \right] |u(b) - u(a)|, \quad x \in [a, b],$$

and if we assume that $$f$$ is $$L$$-Lipschitzian, then (2.6) becomes

$$|\Theta(f, u, a, b)| \leq L \left[ \frac{1}{2} (b - a) + \left| x - \frac{a + b}{2} \right| \right] |u(b) - u(a)|, \quad x \in [a, b].$$
Remark 2. If \( u \) is \( K \)-Lipschitzian, then obviously \( u \) is of bounded variation on \([a, b]\) and \( \sqrt{b} \int_a^b u \leq L(b-a) \). Consequently, if \( f \) is of \( p \)-H"older type, then
\[
\Theta(f, u, a, b) \leq HK \left[ \frac{1}{2} (b-a) + \left| x - \frac{a+b}{2} \right| \right]^p (b-a), \quad x \in [a, b]
\]
and if \( f \) is \( L \)-Lipschitzian, then
\[
\Theta(f, u, a, b) \leq LK \left[ \frac{1}{2} (b-a) + \left| x - \frac{a+b}{2} \right| \right] (b-a), \quad x \in [a, b].
\]

The following corollary concerning a generalization of the mid-point inequality holds:

**Corollary 2.** Let \( f \) and \( u \) be as defined in Theorem 4. Then we have the generalized mid-point formula
\[
|\Upsilon(f, u, a, b)| \leq H \left( \frac{2p}{b-a} \right)^p \sqrt{b} \int_a^b u,
\]
where
\[
\Upsilon(f, u, a, b) = f \left( \frac{a+b}{2} \right) (u(b) - u(a)) - \int_a^b f(t) du(t)
\]
is the mid point functional associated to \( f \) and \( u \) as above. In particular, if \( f \) is \( L \)-Lipschitzian, then
\[
|\Upsilon(f, u, a, b)| \leq \frac{L}{2} (b-a) \sqrt{b} \int_a^b u.
\]

Remark 3. Now, if in (2.11) and (2.12) we assume that \( u \) is monotonic, then we get the midpoint inequalities
\[
|\Upsilon(f, u, a, b)| \leq \frac{H}{2p} (b-a)^p |u(b) - u(a)|
\]
and
\[
|\Upsilon(f, u, a, b)| \leq \frac{L}{2} (b-a) |u(b) - u(a)|
\]
respectively.
In addition, if in (2.11) and (2.12) we assume that \( u \) is \( K \)-Lipschitzian, then we obtain the inequalities
\[
|\Upsilon(f, u, a, b)| \leq HK \left( \frac{2p}{b-a} \right)^{p+1}
\]
and
\[
|\Upsilon(f, u, a, b)| \leq LK \left( \frac{2}{b-a} \right)^2 .
\]

The following inequalities of “rectangle type” also hold:
Corollary 3. Let \( f \) and \( u \) be as in Theorem 4. Then we have the generalized “left rectangle” inequality

\[
\left| f (a) (u(b) - u(a)) - \int_a^b f(t) \, du(t) \right| \leq H (b - a)^p \sqrt[2p]{a} (u)
\]

and the “right rectangle” inequality

\[
\left| f (b) (u(b) - u(a)) - \int_a^b f(t) \, du(t) \right| \leq H (b - a)^p \sqrt[2p]{a} (u).
\]

Remark 4. If we add (2.17) and (2.18), then, by the triangle inequality, we end up with the following generalized trapezoidal inequality

\[
\left| f (a) + f (b) \frac{(u(b) - u(a)) - \int_a^b f(t) \, du(t)}{2} \right| \leq H (b - a)^p \sqrt[2p]{a} (u).
\]

In what follows, we point out some results for the Riemann integral of a product.

Corollary 4. Let \( f : [a, b] \to \mathbb{R} \) be a \( p - H \)–Hölder type mapping and \( g : [a, b] \to \mathbb{R} \) be continuous on \([a, b]\). Then we have the inequality

\[
\left| f (x) \int_a^b g(s) \, ds - \int_a^b f(t) \, g(t) \, dt \right| \leq H \left[ \frac{1}{2} (b - a) + \left| x - \frac{a + b}{2} \right| \right]^p \int_a^b |g(s)| \, ds
\]

for all \( x \in [a, b] \).

Proof. Define the mapping \( u : [a, b] \to \mathbb{R}, u(t) = \int_a^t g(s) \, ds \). Then \( u \) is differentiable on \((a, b)\) and \( u'(t) = g(t) \). Using the properties of the Riemann-Stieltjes integral, we have

\[
\int_a^b f(t) \, du(t) = \int_a^b f(t) \, g(t) \, dt
\]

and

\[
\sqrt[2p]{a} (u) = \int_a^b |u'(t)| \, dt = \int_a^b |g(t)| \, dt.
\]

Therefore, by the inequality (2.2), we deduce (2.20).

Remark 5. The best inequality we can get from (2.20) is that one for which \( x = \frac{a + b}{2} \), obtaining the midpoint inequality

\[
\left| f \left( \frac{a + b}{2} \right) \int_a^b g(s) \, ds - \int_a^b f(t) \, g(t) \, dt \right| \leq \frac{1}{2^p} H (b - a)^p \int_a^b |g(s)| \, ds.
\]

We now give some examples of weighted Ostrowski inequalities for some of the most popular weights.
Example 1. (Legendre) If $g(t) = 1$, and $t \in [a, b]$, then we get the following Ostrowski inequality for Hölder type mappings $f : [a, b] \to \mathbb{R}$

\[(b - a) f(x) - \int_a^b f(t) \, dt \leq H \left[ \frac{1}{2} (b - a) + \left| x - \frac{a + b}{2} \right| \right]^p (b - a)\]

for all $x \in [a, b]$, and, in particular, the mid-point inequality

\[(b - a) f \left( \frac{a + b}{2} \right) - \int_a^b f(t) \, dt \leq \frac{1}{2^p} H (b - a)^{p+1}.\]

Example 2. (Logarithm) If $g(t) = \ln \left( \frac{1}{t} \right)$, $t \in (0, 1]$, $f$ is of $p-$Hölder type on $[0, 1]$ and the integral $\int_0^1 f(t) \ln \left( \frac{1}{t} \right) \, dt$ is finite, then we have

\[f(x) - \int_0^1 f(t) \ln \left( \frac{1}{t} \right) \, dt \leq H \left[ \frac{1}{2} + \left| x - \frac{1}{2} \right| \right]^p\]

for all $x \in [0, 1]$ and, in particular,

\[f \left( \frac{1}{2} \right) - \int_0^1 f(t) \ln \left( \frac{1}{t} \right) \, dt \leq \frac{1}{2^p} H.\]

Example 3. (Jacobi) If $g(t) = \frac{1}{\sqrt{1-t^2}}$, $t \in (-1, 1]$, $f$ is as above and the integral $\int_0^1 f(t) \frac{1}{\sqrt{1-t^2}} \, dt$ is finite, then we have

\[f(x) - \frac{1}{2} \int_0^1 f(t) \frac{1}{\sqrt{1-t^2}} \, dt \leq H \left[ \frac{1}{2} + \left| x - \frac{1}{2} \right| \right]^p,\]

for all $x \in [0, 1]$ and, in particular,

\[f \left( \frac{1}{2} \right) - \frac{1}{2} \int_0^1 f(t) \frac{1}{\sqrt{1-t^2}} \, dt \leq \frac{1}{2^p} H.\]

Finally, we have the following:

Example 4. (Chebychev) If $g(t) = \frac{1}{\sqrt{1-t^2}}$, $t \in (-1, 1]$, $f$ is of $p-$Hölder type on $(-1, -1)$ and the integral $\int_{-1}^1 \frac{f(t)}{\sqrt{1-t^2}} \, dt$ is finite, then

\[f(x) - \frac{1}{\pi} \int_{-1}^1 \frac{f(t)}{\sqrt{1-t^2}} \, dt \leq H |1 + |x||^p\]

for all $x \in [-1, 1]$, and in particular,

\[f(0) - \frac{1}{\pi} \int_{-1}^1 \frac{f(t)}{\sqrt{1-t^2}} \, dt \leq H.\]
3. AN APPROXIMATION FOR THE RIEMANN-STIELTJES INTEGRAL

Consider $I_n : a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ to be a division of the interval $[a, b]$, $h_i := x_{i+1} - x_i$ ($i = 0, \ldots, n - 1$) and $\nu (h) := \max \{ h_i | i = 0, \ldots, n - 1 \}$. Define the general Riemann-Stieltjes sum

$$S(f, u, I_n, \xi) := \sum_{i=0}^{n-1} f(\xi_i) (u(x_{i+1}) - u(x_i)).$$

In what follows, we point out some upper bounds for the error approximation of the Riemann-Stieltjes integral $\int_a^b f(t) \, du(t)$ by its Riemann-Stieltjes sum $S(f, u, I_n, \xi)$.

**Theorem 5.** Let $u : [a, b] \rightarrow \mathbb{R}$ be a mapping of bounded variation on $[a, b]$ and $f : [a, b] \rightarrow \mathbb{R}$ a $p-H$–Hölder type mapping. Then

$$\int_a^b f(t) \, du(t) = S(f, u, I_n, \xi) + R(f, u, I_n, \xi),$$

where $S(f, u, I_n, \xi)$ is as given in (3.1) and the remainder $R(f, u, I_n, \xi)$ satisfies the bound

$$|R(f, u, I_n, \xi)| \leq H \left[ \frac{1}{2} \nu(h) + \max_{i=0, n-1} \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right]^p \int_a^b \left( u \right)^p \, du,$$

for all $i \in \{0, \ldots, n-1\}$.

Proof. We apply Theorem 4 on the subintervals $[x_i, x_{i+1}]$ ($i = 0, \ldots, n - 1$) to obtain

$$\left| f(\xi_i) (u(x_{i+1}) - u(x_i)) - \int_{x_i}^{x_{i+1}} f(t) \, du(t) \right| \leq H \left[ \frac{1}{2} h_i + \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right]^p \int_{x_i}^{x_{i+1}} \left( u \right)^p \, du,$$

for all $i \in \{0, \ldots, n-1\}$.

Summing over $i$ from 0 to $n-1$ and using the generalized triangle inequality, we deduce

$$|R(f, u, I_n, \xi)| \leq \sum_{i=0}^{n-1} \left| f(\xi_i) (u(x_{i+1}) - u(x_i)) - \int_{x_i}^{x_{i+1}} f(t) \, du(t) \right|$$

$$\leq H \sum_{i=0}^{n-1} \left[ \frac{1}{2} h_i + \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right]^p \int_{x_i}^{x_{i+1}} \left( u \right)^p \, du$$

$$\leq H \sup_{i=0, n-1} \left[ \frac{1}{2} h_i + \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right]^p \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \left( u \right)^p \, du.$$
However,
\[
\sup_{i=0, n-1} \left[ \frac{1}{2} h_i + \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right]^p \leq \left[ \frac{1}{2} \nu(h) + \sup_{i=0, n-1} \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right]^p
\]
and
\[
\sum_{i=0}^{n-1} \bigvee_{x_i}^b (u) = \bigvee_{a}^b (u),
\]
which completely proves the first inequality in (3.3).
For the second inequality, we observe that
\[
\left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \leq \frac{1}{2} \cdot h_i,
\]
for all \(i \in \{0, ..., n - 1\}\).
The theorem is thus proved. \(\blacksquare\)

The following corollaries are natural.

**Corollary 5.** Let \(u\) be as in Theorem 5 and \(f\) an \(L\)-Lipschitzian mapping. Then we have the formula (3.2) and the remainder \(R(f, u, I, \xi)\) satisfies the bound

\[
(3.5) \quad |R(f, u, I, \xi)| \leq L \left[ \frac{1}{2} \nu(h) + \max_{i=0, n-1} \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right] \bigvee_{a}^b (u)
\]

\[
\leq H \nu(h) \bigvee_{a}^b (u).
\]

**Remark 6.** If \(u\) is monotonic on \([a, b]\), then the error estimate (3.3) becomes

\[
(3.6) \quad |R(f, u, I, \xi)|
\]

\[
\leq H \left[ \frac{1}{2} \nu(h) + \max_{i=0, n-1} \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right]^p |u(b) - u(a)|
\]

\[
\leq H \nu(h)^p |u(b) - u(a)|
\]

and (3.5) becomes

\[
(3.7) \quad |R(f, u, I, \xi)|
\]

\[
\leq L \left[ \frac{1}{2} \nu(h) + \max_{i=0, n-1} \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right] |u(b) - u(a)|
\]

\[
\leq L \nu(h) |u(b) - u(a)|.
\]

Using Remark 2, we can state the following corollary.

**Corollary 6.** If \(u : [a, b] \rightarrow \mathbb{R}\) is Lipschitzian with the constant \(K\) and \(f : [a, b] \rightarrow \mathbb{R}\) is of \(p - H\)-Hölder type, then the formula (3.2) holds and the remainder \(R(f, u, I, \xi)\)
satisfies the bound

\begin{equation}
|R (f, u, I_n, \xi)| \leq HK \sum_{i=0}^{n-1} \left( \frac{1}{2} h_i + \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| \right)^p h_i
\end{equation}

\begin{equation}
\leq HK \sum_{i=0}^{n-1} h_i^{p+1} \leq HK (b-a) [\nu (h)]^p.
\end{equation}

In particular, if we assume that \( f \) is \( L \)-Lipschitzian, then

\begin{equation}
|R (f, u, I_n, \xi)| \leq \frac{1}{2} LK \sum_{i=0}^{n-1} h_i^2 + LK \sum_{i=0}^{n-1} \left| \xi_i - \frac{x_i + x_{i+1}}{2} \right| h_i
\end{equation}

\begin{equation}
\leq LK \sum_{i=0}^{n-1} h_i^2 \leq LK (b-a) \nu (h).
\end{equation}

The best quadrature formula we can get from Theorem 5 is that one for which \( \xi_i = \frac{x_i + x_{i+1}}{2} \) for all \( i \in \{0, ..., n-1\} \). Consequently, we can state the following corollary.

**Corollary 7.** Let \( f \) and \( u \) be as in Theorem 5. Then

\begin{equation}
\int_a^b f (t) \, du (t) = S_M (f, u, I_n) + R_M (f, u, I_n)
\end{equation}

where \( S_M (f, u, I_n) \) is the generalized midpoint formula, that is;

\[ S_M (f, u, I_n) := \sum_{i=0}^{n-1} f \left( \frac{x_i + x_{i+1}}{2} \right) (u (x_{i+1}) - u (x_i)) \]

and the remainder satisfies the estimate

\begin{equation}
|R_M (f, u, I_n)| \leq \frac{H}{2^p} [\nu (h)]^p \left( \frac{b}{a} \right)^p (u).
\end{equation}

In particular, if \( f \) is \( L \)-Lipschitzian, then we have the bound:

\begin{equation}
|R_M (f, u, I_n)| \leq \frac{H}{2^p} \nu (h) \left( \frac{b}{a} \right)^p (u).
\end{equation}

**Remark 7.** If in (3.11) and (3.12) we assume that \( u \) is monotonic, then we get the inequalities

\begin{equation}
|R_M (f, u, I_n)| \leq \frac{H}{2^p} [\nu (h)]^p |f (b) - f (a)|
\end{equation}

and

\begin{equation}
|R_M (f, u, I_n)| \leq \frac{H}{2^p} \nu (h) |f (b) - f (a)|.
\end{equation}

The case where \( f \) is \( K \)-Lipschitzian is embodied in the following corollary.
Corollary 8. Let $u$ and $f$ be as in Corollary 6. Then we have the quadrature formula (3.10) and the remainder satisfies the estimate

\begin{equation}
|R_M(f, u, I_n)| \leq \frac{HK}{2^p} \sum_{i=0}^{n-1} h_i^{p+1} \leq \frac{HK}{2^p} \nu(h)^p.
\end{equation}

In particular, if $f$ is $L-$ Lipschitzian, then we have the estimate

\begin{equation}
|R_M(f, u, I_n)| \leq \frac{1}{2} LK \sum_{i=0}^{n-1} h_i^2 \leq \frac{1}{2} LK (b-a) \nu(h).
\end{equation}

References


School of Communications and Informatics
Victoria University of Technology
PO Box 14428
Melbourne City
MC 8001 Australia
email:sever@matilda.vu.edu.au
urladdrhttp://melba.vu.edu.au/dragomirweb.html
ON APPROXIMATIONS BY IRRATIONAL SPLINES

Mikhail P. Levin

Abstract. A problem of approximation by irrational splines is considered. These splines have a constant curvature between interpolation nodes and need only one additional boundary condition for derivatives, which should be set only at one of two boundary nodes, that is impossible for usual polynomial splines required boundary conditions at both boundary nodal points. Some estimations for numerical differentiation and rounding error analysis are presented.

1. Introduction

Although in recent years new various approaches in approximation of data by convex and positivity preserving splines have been proposed (see for instance [1-3]) a problem of data interpolation by smooth functions with a constant curvature between interpolation nodes is important to date. This problem is especially topical in Computational Fluid Dynamics in transonic and supersonic cases and in some other applications. It is well-known that this interpolation problem can not be solved by usual polynomial splines because the curvature of cubic and other higher order polynomial splines between the interpolation is not constant. As to the quadratic polynomial splines it is known that these splines have a constant second derivative or curvature only between the splines nodes, but for these splines their nodes do not coincide with the interpolation nodes and usually are located at the middle of the interpolation nodes. This is necessary to provide a stability of the algorithm for evaluation of quadratic spline coefficients [4-5].

Another topical problem consists in setting of auxiliary boundary conditions for spline derivatives only at one of two boundary nodes. For polynomial splines this is impossible, because the algorithm for evaluation of spline coefficients is also unstable [4-5] in this case.

In this paper one class of irrational splines is considered. In this class on each segment, restricted by adjacent interpolation nodes, splines are described by circle arcs passing through the interpolation nodes. In all internal interpolation nodes the

AMS subject Classification: 65D07, 65D25. This work was supported by the BK21 project.
condition of smoothing for the first order spline derivatives is provided. It is shown that for considering splines, it is enough to set the boundary conditions for derivatives only at one of two boundary points.

2. Definitions

Let us consider any grid function \( y_i \), \( (i = 0, 1, 2, 3, ..., N) \) defined in nodes \( x_0, x_1, x_2, x_3, ..., x_N \). Let \( y'_0 \) be a known derivative of the data grid function \( y_i \). Then we will show that these data is enough to construct an irrational spline function.

Let us consider the \( i \)-th segment \([x_i, x_{i+1}]\) and define a circle arc passing through points \((x_i, y_i)\) and \((x_{i+1}, y_{i+1})\) with tangent \( y'_i \) at the first point \((x_i, y_i)\). We present the equation of the circle arc in usual form

\[
(x - x_{ci})^2 + (y - y_{ci})^2 = R_i^2 .
\]  

(1)

Here \((x_{ci}, y_{ci})\) is a center and \( R_i \) is a radius of the circle constructed for the \( i \)-th interpolation segment.

Implicit differentiation of equation (1) with respect to \( x \) yields

\[
(x - x_{ci}) + (y - y_{ci})y'_i = 0 .
\]  

(2)

Values \( x_{ci}, y_{ci} \) and \( R_i \) are unknown and our goal is to evaluate these values by the known values \( y_i, y_{i+1} \) at points \( x_i \) and \( y_{i+1} \).

Let us take two interpolation conditions satisfying to equation (1) at points \( x_i \) and \( x_{i+1} \) and the interpolation condition for expression (2) taken at point \( x_i \). Then we obtain a system of three equations

\[
(x - x_{ci})^2 + (y - y_{ci})^2 = R_i^2 .
\]  

(3)

\[
(x_{i+1} - x_{ci})^2 + (y_{i+1} - y_{ci})^2 = R_i^2 .
\]

\[
(x_i - x_{ci}) + (y_i - y_{ci})y'_i = 0 .
\]

To construct a solution of (3), let us introduce new searching variables \( \xi_i = x_{ci} - x_i \) and \( \eta_i = y_{ci} - y_i \) and denote \( h_i = x_{i+1} - x_i \), \( H_i = y_{i+1} - y_i \). Then formulas (3) can be presented as follows

\[
\xi_i^2 + \eta_i^2 - R_i^2 = 0 ,
\]

(4)

\[
(h_i - \xi_i)^2 + (H_i - \eta_i)^2 - R_i^2 = 0 ,
\]

\[
\xi_i + \eta_i y'_i = 0 .
\]

Solving this system we obtain

\[
\eta_i = \frac{h_i^2 + H_i^2}{2(H_i - h_i y'_i)} ,
\]

(5)

\[
\xi_i = -\frac{h_i^2 + H_i^2}{2(H_i - h_i y'_i)} y'_i ,
\]

\[
R_i^2 = \frac{(h_i^2 + H_i^2)^2[1 + (y'_i)^2]}{4(H_i - h_i y'_i)^2} .
\]
According to (5) we can find
\[ x_{ci} = x_i + \xi_i, \]
\[ y_{ci} = y_i + \eta_i. \]
and can present the equation of the circle arc in one of the following forms
\[ (x - x_i)(x - x_i - 2\xi_i) + (y - y_i)(y - y_i - 2\eta_i) = 0 \] (7a)
or
\[ \frac{y - y_i}{x - x_i} = -\frac{x - x_i - 2\xi_i}{y - y_i - 2\eta_i}. \] (7b)
or
\[ \frac{y - y_i}{x - x_i} = -\frac{(x - x_i)(H_i - h_i y_i') + (h_i^2 + H_i^2)y_i'}{(y - y_i)(H_i - h_i y_i') - (h_i^2 + H_i^2)}. \] (7c)

Therefore the solution of interpolation problem on the segment \([x_i, x_{i+1}]\) consists in solution of non-linear equation (7) for any data value \(x_s \in [x_i, x_{i+1}]\). For this purpose it is possible to apply, for instance, well-known Newton method or one of its modifications.

3. Numerical Differentiation

Now we consider a problem of evaluation the first derivative of the data grid function at the second point \(x_{i+1}\) at the considering segment \([x_i, x_{i+1}]\). For this purpose we use the following geometry property of the tangential lines passing through the first and through the end points of the circle arc and the secant line passing also through these points

\[ \varphi_{si} = \frac{1}{2}(\varphi_i + \varphi_{i+1}), \]

where \(\varphi_i = arctg(y_i')\), \(\varphi_{i+1} = arctg(y_{i+1}')\) are angles between tangential lines to the considering arc at points \(x_i\) and \(x_{i+1}\) and \(x\)-axes, \(\varphi_{si} = arctg(H_i h_i')\) is an angle between the secant line passing through the points \((x_i, y_i)\) and \((x_{i+1}, y_{i+1})\).

In this case, since
\[ tg(\varphi_i + \varphi_{i+1}) = tg(2\varphi_{si}), \]
\[ tg(2\varphi_{si}) = \frac{2tg(\varphi_{si})}{1 - tg^2(\varphi_{si})} = \frac{2H_i h_i}{h_i^2 - H_i^2}, \]
we can express the first derivative at the point \( i + 1 \) by the following formula

\[
y'_{i+1} = \frac{2H_i h_i + (H_i^2 - h_i^2)y'_i}{h_i^2 - H_i^2 + 2H_i h_i y'_i}.
\]

Thus, following to the formula (8) and step by step procedure starting from the first interpolation segment, we can construct the irrational spline for all considering interpolation segments \([x_i, x_{i+1}], i = 0, 1, 2, 3, ..., (N - 1)\). Since we take \( y'_i \) evaluated by (8) at the \((i - 1)\)-th step as initial data for calculation \( y'_{i+1} \) at the \( i \)-th step, we provide smoothing conditions for the first derivative of the considering spline at all internal interpolation nodes.

According to above mentioned we can see that considering irrational splines don’t need solution of linear algebra equations systems for evaluation of their coefficients as usual polynomial splines need. All coefficients of these splines can be computed by the recurrent formulas (5,6,8)

Using a Taylor-series expansion we can estimate the accuracy of the formula (8) intending for the numerical differentiation of the data grid function. As a result this estimation can be presented as follows

\[
y' - \tilde{y}' = \frac{\tilde{y}'(\tilde{y}'')^2}{2[1 + (\tilde{y}')^2]} - \frac{\tilde{y}''}{6} h^2 + O(h^3).
\]

Here \( \tilde{y}' \) is an exact value of the first derivative, \( \tilde{y}'' \) is an exact value of the second derivative and \( \tilde{y}''' \) is an exact value of the third derivative of the considering function taken at the nodal point \( i \). Thus the formula (8) has the second order approximation error.

4. Degeneration Case

Now we consider a degeneration case \( y'_i = \frac{H_i}{h_i} \). In this case denominators of fractions in right-hand sides of formulas (5) are equal to zero and hence the appropriate spline parameters \( \eta_i, \xi_i \) and \( R_i \) are undefined.

However, in this case according to the formula (8) we have

\[
y'_{i+1} = y'_i = \frac{H_i}{h_i}.
\]

This means that the considering circle arc element is singular with \( R_i = \infty \) and \( \eta_i = \xi_i = \infty \). Thus in considering case according to the formula (10) the circle arc spline element degenerates into the straight line element.
5. Estimation of Rounding Errors Spreading

Now we provide a rounding error analysis of the numerical differentiation formula (8). Let suppose that at any point \( i \) we know a perturbed value of the first derivative \( y'_i = \tilde{y}'_i + \varepsilon_i \). Here \( \tilde{y}'_i \) is the exact value of the first derivative and \( \varepsilon_i \) is an rounding error. Then we consider the grid value of the first derivative taken at the next point \( i + 1 \) and evaluated by the exact data

\[
\tilde{D}_{i+1} = \frac{2H_i h_i + (H_i^2 - h_i^2)\tilde{y}'_i}{h_i^2 - H_i^2 + 2H_i h_i \tilde{y}'_i}.
\]

For comparison we also consider the appropriate value evaluated by the perturbed data

\[
D_{i+1} = \frac{2H_i h_i + (H_i^2 - h_i^2)(\tilde{y}'_i + \varepsilon_i)}{h_i^2 - H_i^2 + 2H_i h_i (\tilde{y}'_i + \varepsilon_i)}.
\]

Applying the Taylor-series expansion of the last expression with respect to \( \varepsilon_i \), we obtain the following formula

\[
D_{i+1} - \tilde{D}_{i+1} = -\varepsilon_i \left( \frac{h_i^2 + H_i^2}{h_i^2 - H_i^2 + 2H_i h_i \tilde{y}'_i} \right)^2 + O(\varepsilon_i^2). \tag{11}
\]

According to the formula (11) perturbations in initial data or rounding errors are strictly damping, if the following inequality satisfies

\[
\left| \frac{h_i^2 + H_i^2}{h_i^2 - H_i^2 + 2H_i h_i \tilde{y}'_i} \right| < 1, \tag{12a}
\]

or

\[
-h_i^2 + H_i^2 - 2H_i h_i \tilde{y}'_i < h_i^2 + H_i^2 < h_i^2 + H_i^2 < h_i^2 - H_i^2 + 2H_i h_i \tilde{y}'_i. \tag{12b}
\]

The inequality (12) can be presented as a system of two inequalities

\[
H_i^2 < H_i h_i \tilde{y}'_i, \tag{13}
\]

\[-H_i h_i \tilde{y}'_i < h_i^2. \]

Let us suppose \( h_i > 0 \) for the considering grid, otherwise we could change enumeration of grid points to provide this condition. Then (13) turns as follows

\[
\frac{H_i^2}{h_i^2} < \frac{H_i}{h_i} \tilde{y}'_i, \tag{14}
\]

\[-\frac{H_i}{h_i} \tilde{y}'_i < 1. \]
Let us consider case $H_i \geq 0$. Then according to this condition and the first inequality of (14) we have

$$0 \leq \frac{H_i}{h_i} < \tilde{y}_i',$$

and then the second inequality of (14) is satisfied identically, because $-\tilde{y}_i' < 0$ and $\frac{h_i}{H_i} > 0$.

In case $H_i < 0$, we have $\frac{H_i}{h_i} < 0$, and dividing both inequalities (14) by this negative term, we obtain instead of the first inequality

$$0 > \frac{H_i}{h_i} > \tilde{y}_i'$$

and then the second inequality in the form $-\frac{h_i}{H_i} > \tilde{y}_i'$ is satisfied identically.

In both cases considered above, we have

$$|\frac{H_i}{h_i}| < |\tilde{y}_i'|,$$  \hspace{1cm} (15a)

or

$$|y_{si}'| < |\tilde{y}_i'|.$$  \hspace{1cm} (15b)

Here we denote a tangent of the secant line slope $tg(\varphi_{si}) = \frac{H_i}{h_i}$ as $y_{si}'$.

We could present $y_{si}'$ also as follows

$$y_{si}' = tg\left[\frac{1}{2}(arctg y_i' + arctg y_{i+1}')\right]$$

$$= \sqrt{\frac{[1 + (y_i')^2][1 + (y_{i+1}')^2] + y_i'y_{i+1}' - 1}{y_i' + y_{i+1}'}}.$$  \hspace{1cm} (16)

Let us write $y_{i+1}' = y_i' + (y_{i+1}' - y_i')$ and take a Taylor series expansion of right-hand side of the formula (16) with respect to $\triangle y_i' = y_{i+1}' - y_i'$. In result we obtain

$$y_{si}' = y_i' + \frac{1}{2} \triangle y_i' + O[(\triangle y_i')^2]$$

or

$$y_{si}' = \frac{1}{2}(y_i' + y_{i+1}') + O[(\triangle y_i')^2].$$
Since
\[ \frac{1}{2}(y_i' + y_{i+1}') + O[(\Delta y_i')^2] \leq \frac{1}{2}(|y_i'| + |y_{i+1}'|) + |O[(\Delta y_i')^2]|, \]
then the condition (12) is automatically satisfied, if the function \( y \) satisfies to the following condition
\[ \frac{1}{2}|y_{i+1}'| + O[(\Delta y_i')^2] < \frac{1}{2}|y_i'|. \]
Then taking a limit of the both sides of last inequality, divided by \( h_i \), as \( h_i \) goes to zero, we obtain the following condition
\[ \frac{dy'}{dx} < 0. \] (17)
Thus, if any function \( y(x) \) has a negative derivative of the absolute value of its first derivative, then the condition (12) holds and the rounding error arising in evaluation of the first derivative by formula (8) is strictly damping.

References

Comparing Two Approaches of Analyzing Mixed Finite Volume Methods

So-Hsiang Chou and Shengrong Tang

Abstract

Given the anisotropic Poisson equation $-\nabla \cdot K \nabla p = f$, one can convert it into a system of two first order PDEs: the Darcy law for the flux $u = -K \nabla p$ and conservation of mass $\nabla \cdot u = f$. A very natural mixed finite volume method for this system is to seek the pressure in the nonconforming $P_1$ space and the Darcy velocity in the lowest order Raviart-Thomas space. The equations for these variables are obtained by integrating the two first order systems over the triangular volumes. In this paper we show that such a method is really a standard finite element method with local recovery of the flux in disguise. As a consequence, we compare two approaches in analyzing finite volume methods (FVM) and shed light on the proper way of analyzing non co–volume type of FVM. Numerical results for Dirichlet and Neumann problems are included.

1 Introduction

Consider the variable-coefficient Poisson equation in a polygonal domain $\Omega \subset \mathbb{R}^2$

\[
\begin{aligned}
&\left\{\begin{array}{ll}
-\nabla \cdot K \nabla p = f & \text{in } \Omega, \\
p = 0 & \text{on } \partial \Omega,
\end{array}\right.
\end{aligned}
\]

where $K = K(x)$ is a symmetric positive definite matrix function such that there exist two positive constants $\alpha_1$ and $\alpha_2$ with

\[
\alpha_1 \xi^T \xi \leq K(x) \xi \xi^T \leq \alpha_2 \xi^T \xi \quad \forall \xi \in \mathbb{R}^2, x \in \bar{\Omega}.
\]

Now let us introduce a flux variable $u := -K \nabla p$ and write the above equation as the system of first order partial differential equations

\[
\begin{aligned}
&\left\{\begin{array}{ll}
\nabla \cdot u - f = 0 & \text{in } \Omega \\
u + K \nabla p = 0 & \text{in } \Omega \\
p = 0 & \text{on } \partial \Omega
\end{array}\right.
\end{aligned}
\]
Problems (1) and (3) are equivalent when the right hand side $f$, the diffusion tensor $K$ and the domain $\Omega$ (e.g. convex) are such that the solution is smooth enough, e.g., $u \in H^1(\Omega)^2$, $p \in H^1_0(\Omega) \cap H^2(\Omega)$.

This system can be interpreted as modeling an incompressible single phase flow in a reservoir; ignoring gravitational effects. The matrix $K$ is the mobility $\kappa/\mu$, the ratio of permeability tensor to viscosity of the fluid, $u$ is the Darcy velocity and $p$ the pressure. The second equation is the Darcy law and the first represents conservation of mass with $f$ standing for a source or sink term. Since $\kappa$ is in general discontinuous due to different rock formations, separating the Darcy law from the second order equation and discretizing it directly together with the mass conservation may lead to a better numerical treatment on the velocity than just computing it from the pressure via the Darcy law. This approach is well known in the finite element circle [20], but the same approach can be applied in conjunction with the finite volume method (termed mixed finite volume methods) as well [6, 9, 10, 11, 12, 20, 21]. For other similarly related issues, see also [17, 18, 22].

Let $T_h = \{ K_j \}_{j=1}^N$ be the usual non–overlapping finite element triangulation of the domain $\Omega = \cup_{K \in T_h} K$. Furthermore $T_h$ is assumed to be regular, that is, $\min_{K \in T_h} d(K)/\rho(K) \geq C$ for a constant $C$ independent of $h$. Here $\rho(K)$ is the diameter of triangle $K$; $d(K)$ the diameter of the inscribed circle of $K$, and $h = \max_{K \in T_h} \rho(K)$.

We denote the area of $K$ by $|K|$, by $\mathcal{A} = \mathcal{A}_i \cup \mathcal{A}_b$ the set of all edges of $T_h$ consisting of the interior edge set $\mathcal{A}_i$ and boundary edge set $\mathcal{A}_b$. We use $N_{\mathcal{A}_i}$ and $N_{\mathcal{A}_b}$ to denote the number of interior edges and the number of boundary edges, respectively. The total number of edges is $N_{\mathcal{A}} = N_{\mathcal{A}_i} + N_{\mathcal{A}_b}$.

Define the lowest order Raviart-Thomas space [19]

$$V_h = \{ u_h \in H(\text{div}; \Omega) : u_h|_K \in RT_0(K) \}$$

where $RT_0(K) = \{ u = (u^1, u^2) : u^1 = a + bx, u^2 = c + by \text{ in } K \}$ and the standard $P1$ nonconforming finite element space

$$Y_h = \{ p_h|_K \in P_1(K) : p_h \text{ continuous at the middle point of each } e \in \partial K \}.$$

Consider approximating $u$ by $u_h \in V_h$ and $p$ by $p_h \in Y_h$ via a mixed finite volume approach on (3):

Find $(u_h, p_h) \in V_h \times Y_h$ such that

$$\begin{cases} (\nabla \cdot u_h - f, \chi_K) = 0 & \text{for all } K \in T_h, \\ (u_h + K \nabla p_h, \chi_K) = 0 & \text{for all } K \in T_h \end{cases}$$

(4)

where $p_h = 0$ at all midpoints of boundary edges. Here $\chi_K$ is the characteristic function of triangle $K$ and $\chi_K$ is any constant vector multiple of $\chi_K$.

This method was introduced and analyzed for the case of $K = I$, the identity matrix, by Courbet & Croisille [15]. In this paper we study the general full tensor case. The
full tensor case is important in porous media flow applications since it represents the more natural anisotropic case. Mathematically, the full tensor case is by no means a trivial extension of the simple Poisson equation, especially in the finite volume case: the presence of the full tensor often results in a nonsymmetric system and is harder to give physical interpretations [1]. Furthermore, unlike in the mathematical analysis of finite element methods, convergence of finite volume methods for the isotropic and anisotropic cases in general cannot be handled in a unified way. See [16] and the references therein for the difficulties involved. Also see [10, 11] where a uniform treatment is possible, but in these co-volume schemes Chou et al. used two grid systems in discretization. It is therefore surprising that the extension of the Courbet & Croisille method has none of these drawbacks, having only one grid system for both variables.

This paper is the first in a series of two papers on comparing the analysis of the finite volume method from a finite element person’s or a “pure” finite volume person’s viewpoint. It grows out of the informal report [13]. We will present things in an elementary and non-terse way and the point will be how different viewpoints may lead to the same conclusions but in a roundabout way. The reader is referred to follow-up [14] for a more extensive and in-depth mathematical presentation.

First we adopt a pure finite volume person’s point of view. In Sec. 2, we derive the discrete system, and in Sec. 3 we compute explicitly the element and global stiffness matrices whose proper interpretations lead to Thm 4.1, which says that the pressure approximation can be re-interpreted as the solution to the system of the standard $P_1$ nonconforming finite elements applied to the second order elliptic problem (1). However, the process leading to that conclusion is long and technical. The advantage in this approach is that conservation laws are obvious from the start and the drawback is that the mathematical analysis is long and very uninspiring.

On the other hand, as shown in [8, 12, 11] it is fruitful to try to relate the analysis of a mixed finite volume method to a close finite element method. Adopting this viewpoint, we can prove Thm. 4.1 in a few short lines. The advantage of this approach is that mathematical analysis is neat, but conservation law is somewhat hidden. To keep the paper short, how to get the best of these two approaches is elucidated in the follow-up paper [14].

The remaining of this paper is organized as follows. In Sec. 5, we show the error estimates in $p$ and $u$. Finally, in the last section we provide numerical results for both Dirichlet and Neumann problems.

### 2 Problem formulation

Let us now represent the system (4) using proper basis functions. Notice that by divergence theorem, Eq. (4) can be written as

$$ \int_{\partial K} u_h \cdot n dx - |K| f_K = 0 $$

(5)
where \( f_K = \frac{1}{|K|} \int_K f \, dx \) is the average of \( f(x) \) over triangle \( K \). Since \( \nabla p_h \) is constant over \( K \), Eq. (4) can be written as

\[
\int_K u_h \, dx + |K| A_K \nabla p_h = 0
\]  

(6)

where the matrix \( A_K = \frac{1}{|K|} \int_K K \, dx \) is the average of matrix \( K(x) \) over triangle \( K \).

Figure 1: Local elements based on an interior edge \( e_1 \).

With reference to Fig. 1, let \( \lambda_S \) be the usual nodal linear basis function associated with the vertex \( S \) of \( K = K_L \). Recall that \( \lambda_S \) is one at \( S \) and zero at other two vertices and is also called barycentric or area coordinate function. For any \( p_h(x) \in Y_h \), we have the local representation on \( K \)

\[
p_h(x)|_K = \sum_{e \in \partial K} p_e \varphi_e(x)
\]  

(7)

where \( \varphi_e(x) = 1 - 2\lambda_S(x) \) is the local basis function of space \( Y_h \) on edge \( e \) with \( \lambda_S(x) \) being barycentric coordinate of \( x \) with respect to vertex \( S \) opposite to \( e \) in triangle \( K \). (Note that \( e = e_1 \) in Fig. 1.) It is easy to see \( \nabla \varphi_e(x) = \frac{|e|}{|K|} n_e = \text{const} \), where \( |e| \) is the length of edge \( e \).

Given any triangular element \( K \in T_h \), we always orient \( K \) counterclockwise as shown in Fig. 1 (e.g. \( K = K_L \) there). Then the three local basis functions associated
with the three edges are as follows. For example, for the edge \( e = e_1 \) of \( K = K_L \) in Fig. 1, we define \( P_{K,e} \)

\[
P_{K,e}(x) = \frac{1}{2|K|} \left[ \begin{array}{c} x - x_S \\ y - y_S \end{array} \right] \quad \forall \ (x, y) \in K.
\]

(8)

Note that

\[
P_{K,e}(x) \cdot n = \begin{cases} \frac{1}{|e|} & \forall \ x \in e = S'S'' \\ 0 & \forall \ x \in SS' \\ 0 & \forall \ x \in SS'' \end{cases}
\]

(9)

The other two basis functions \( P_{K,e_i}, i = 2, 3 \) are defined similarly.

For any \( u_h(x) \in V_h \), we have the local representation on \( K \)

\[
|K| P_{K,e}(x) \chi_K(x) = \sum_{e \in \partial K} u_e P_{K,e}(x) \quad (10)
\]

where \( u_e = \int_e u \cdot nds \) is the flux across edge \( e \). For any edge \( a \in A \) we define the global canonical basis of \( V_h \) associated with edge \( a \) as follows. If the edge \( a \) corresponds to edge \( S'S'' \) in the local ordering (cf. Fig. 1), then

\[
P_a(x) = P_{K_L,e}(x) \chi_{K_L}(x) - P_{K_R,e}(x) \chi_{K_R}(x)
\]

(11)

where \( a \) is oriented from \( K_L \) towards \( K_R \). The global basis functions based on boundary edges are defined similarly.

Finally, by Taylor's expansion at the barycenter \( B \) of \( K \) we have on \( K \)

\[
u_h(x) = u_K + (\nabla \cdot u_h)_K P_K(x)
\]

(12)

where \( u_K = \frac{1}{|K|} \int_K u_h dx = -A_K \nabla p_h \), the average of \( u_h(x) \) on \( K \), and

\[
P_K(x) = \frac{1}{2} \left[ \begin{array}{c} x - x_B \\ y - y_B \end{array} \right] = \frac{|K|}{3} \sum_{e \in \partial K} P_{K,e}(x) \quad \forall \ (x, y) \in K
\]

(13)

with \((x_B, y_B)\) being the coordinates of \( B \). Alternatively, one can also write

\[
u_h(x) = -A_K \nabla p_h + |K| f_K P_K(x).
\]

(14)

Note that \(|K| f_K P_K(x)\) has zero mean on triangle \( K \).

Let \( u_h \in V_h \) and \( p_h \in Y_h \) have the local representations (10) and (7), respectively. Then for each \( K \in T_h \), Eq. (5) implies

\[
\sum_{e \in \partial K} u_e - |K| f_K = 0 \quad \text{(NT equations)}
\]

(15)
and Eq. (6) becomes
\[\sum_{e \in \partial K} u_e \int_K P_{K,e}(x) dx + p_e |K| A_K \nabla \varphi_e(x) = 0.\]

Define \(Q_e = \int_K P_{K,e}(x) dx, N_e = |e| A_K n_e\), and recall that \(\nabla \varphi_e(x) = \frac{|e|}{|K|} n_e = \text{const}\) to get
\[\sum_{e \in \partial K} u_e Q_e + p_e N_e = 0 \quad \text{(2NT equations).} \quad (16)\]

Referring to Fig. 1, we see that by the one-point quadrature using the barycenter \(B_Q = P_{K,e}(B) |K| = 1/6(\overrightarrow{SS'} + \overrightarrow{SS''}) := 1/6(e_3 - e_2)\)
where \(\overrightarrow{SS'} = e_3\) and \(\overrightarrow{SS''} = e_2\). Hence
\[\sum_{e \in \partial K} Q_e = 0. \quad (17)\]

Also note
\[\sum_{e \in \partial K} N_e = A_K (\sum_i |e_i| n_{e_i}) = 0. \quad (18)\]

On the boundary
\[p_a = 0 \quad \text{(N.A_b equations).} \quad (19)\]

Clearly
\[3NT + N A_b = 2N A. \quad (20)\]

and we have as many equations as unknowns: the number of unknowns \((u_a, p_a)_{a \in A}\) being \(2N A\) and the total number of equations in (15), (16) and (19) being \(3NT + N A_b = 2N A\).

Combining (15), (16) and (19), we see that system (4) becomes: Find \(u_h = \sum_{a \in A} u_a P_a(x), p_h(x) = \sum_{a \in A} p_a \varphi_a(x)\) such that
\[\begin{cases}
    \sum_{e \in \partial K} u_e = |K| I_K & \forall K \in T_h, \\
    \sum_{e \in \partial K} (u_e Q_e + p_e N_e) = 0 & \forall K \in T_h, \\
    p_a = 0 & \forall a \in A_b.
\end{cases} \quad (21)\]

A remark about notation is in order here. We emphasize that the notation \(u_a, a \in A\) is reserved for the component with respect to the global basis whereas the notation \(u_{K,e}, e \in \partial K\) is the component with respect to the local basis. When there is no danger of confusion we simply use \(u_e\) instead of \(u_{K,e}\). Later in sec. 5, we shall show the existence and uniqueness of the above system.
3 Element and global stiffness matrices

In this section we shall work out the details of implementation of the resulting discrete systems.

Denote by $U = (u_a)_{a \in A}$ the global vector of $u_h(x)$ onto basis $\{P_a(x)\}$ and $P = (p_a)_{a \in A}$ the global vector of $p_h(x)$ onto basis $\{\varphi_a(x)\}$. Also define $U_K$ and $P_K$ to be the local vectors of $u_h(x)$ and $p_h(x)$ onto local basis functions in the triangle $K$

$$U_K = [u_{e_1}, u_{e_2}, u_{e_3}]^T,$$
$$P_K = [p_{e_1}, p_{e_2}, p_{e_3}]^T$$

where $e_1$, $e_2$ and $e_3$ are three edges of $K$.

Then system (21) can be written as:

$$\bar{L}_K U_K + \bar{M}_K P_K = \bar{F}_K \quad \forall K \in T_h$$

where $\bar{L}_K, \bar{M}_K \in \mathbb{R}^{3 \times 3}, \bar{F}_K \in \mathbb{R}^3$ with

$$\bar{L}_K = \begin{bmatrix} 1 & 1 & 1 \\ Q_{e_1} & Q_{e_2} & Q_{e_3} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ \frac{1}{6}(e_3 - e_2) & \frac{1}{6}(e_1 - e_3) & \frac{1}{6}(e_2 - e_1) \end{bmatrix},$$
$$\bar{M}_K = \begin{bmatrix} 0 & 0 & 0 \\ N_{e_1} & N_{e_2} & N_{e_3} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ |e_1|A_K n_{e_1} & |e_2|A_K n_{e_2} & |e_3|A_K n_{e_3} \end{bmatrix},$$
$$\bar{F}_K = \begin{bmatrix} |K|f_K \\ 0 \\ 0 \end{bmatrix}.$$

We know that the matrix $\bar{L}_K$ is nonsingular by (17) and hence (24) can be rewritten as

$$U_K = F_K - M_K P_K \quad \forall K \in T_h$$

where $M_K = \bar{L}_K^{-1} \bar{M}_K$ and $F_K = \bar{L}_K^{-1} \bar{F}_K$. It is easy to check

$$F_K = \frac{|K|}{3} f_K \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

We can eliminate the unknowns $(u_a)_{a \in A}$ to obtain a system in the pressures alone. If $a$ is an interior edge with orientation from $K_L(a)$ to $K_R(a)$, $a = e_L$ in $K_L(a)$, $a = e_R$ in $K_R(a)$, then the continuity of the flux across $a$ gives the identity $U_{K_L,e_L} = -U_{K_R,e_R}$ holds. (Here $U_{K_L,e_L}$ means the component of $U_{K_L}$ in (22) corresponding to the edge $e_L$.) Thus we have the (scalar) identity

$$[M_{K_L} P_{K_L}]_{e_L} + [M_{K_R} P_{K_R}]_{e_R} = F_{K_L,e_L} + F_{K_R,e_R} \quad \forall a \in A.$$

(27)
This assembly along with \( p_a = 0 \) for all \( a \in A \) then gives rise to an \( NA \times NA \) linear system in the unknown \( P = (p_a)_{a \in A} \)

\[
\mathcal{M} P = \mathcal{F}.
\] (28)

We denote \( P_i \in \mathbb{R}^{NA_i} \), \( \mathcal{F}_i \in \mathbb{R}^{NA_i} \), and \( \mathcal{M}_i \in \mathbb{R}^{NA_i \times NA_i} \) the sub-vectors or submatrix of \( P, \mathcal{F} \) and \( \mathcal{M} \) corresponding to interior edge set \( A_i \). The resulting system is \( \mathcal{M}_i P_i = \mathcal{F}_i \). Next, we will give a simple form of local stencil \( M_K \) which is a symmetric positive matrix. Then we will introduce a way to assemble global stiffness matrix \( \mathcal{M}_i \) from local stencil \( M_K \). Finally we show that the matrix \( \mathcal{M}_i \) is symmetric positive definite.

Lemma 3.1 The local stencil \( M_K \) is symmetric positive semi-definite.

Proof 1 Let us compute \( \tilde{L}_K^{-T} \) first. Referring to the notation in Fig. 1, we have

\[
\tilde{L}_K^{-T} = \begin{bmatrix}
1 & 1/6(e_3 - e_2)^T \\
1 & 1/6(e_1 - e_3)^T \\
1 & 1/6(e_2 - e_1)^T
\end{bmatrix}.
\]

Using the fact \( e_1 + e_2 + e_3 = 0 \) and elementary geometry, we can easily see that

\[
\tilde{L}_K^{-T} = \begin{bmatrix}
\frac{1}{3} \Vert e_1 \Vert n_{e_1} & \frac{1}{3} \Vert e_2 \Vert n_{e_2} & \frac{1}{3} \Vert e_3 \Vert n_{e_3}
\end{bmatrix}
\]

where \( n_{e_i} \) are the unit outward normal to \( e_i \). Thus \( M_K = \tilde{L}_K^{-1} \tilde{M}_K \) has entries

\[
(M_K)_{ij} = \frac{|e_i| \Vert e_j \Vert}{|K|} A_K n_{e_j} = \frac{|e_i| \Vert e_j \Vert}{|K|} n_{e_i}^T A_K n_{e_i} = (M_K)_{ji}
\]

and hence \( M_K \) is symmetric.

Now let \( R(\theta) \) be the rotation matrix through an angle of \( \theta \), then

\[
n_{e_i} = R(-\frac{\pi}{2})e_i/\|e_i\| \]

and so

\[
(M_K)_{ij} = \frac{1}{|K|} e_i^T \tilde{A}_K e_j
\] (29)

where

\[
\tilde{A}_K = \begin{bmatrix}
c & -b \\
-b & a
\end{bmatrix} \quad i f \quad A_K = \begin{bmatrix}
a & b \\
b & c
\end{bmatrix}.
\] (30)

Next we show \( M_K \) is positive semi-definite. Denote \( E = [e_1, e_2, e_3] \), then

\[
M_K = \frac{1}{|K|} E^T \tilde{A}_K E.
\] (31)
ANALYZING MIXED FINITE VOLUME METHODS

and

\[
\xi^T M_K \xi = \frac{1}{|K|} \xi^T E^T \tilde{A}_K E \xi = \frac{1}{|K|} \xi^T E^T R \left( -\frac{\pi}{2} \right)^T A_K R \left( -\frac{\pi}{2} \right) E \xi \\
\geq \alpha \frac{1}{|K|} |R \left( -\frac{\pi}{2} \right) E \xi|^2 = \alpha \frac{1}{|K|} |E \xi|^2.
\]

Therefore \( M_K \) is positive semi-definite.

Remark: Note if we take \( A_K = I \), from (29) a simple computation shows that

\[
M_K = 2 \begin{bmatrix} d_2 + d_3 & -d_3 & -d_2 \\
-d_3 & d_1 + d_3 & -d_1 \\
-d_2 & -d_1 & d_1 + d_2 \end{bmatrix}
\]

where \( d_i = \cot \theta_i \), where \( \theta_i \) is the angle opposite to the edge \( e_i \). It is 4 times the local stencil in the standard FVM for Poisson problem. (see [3]) and same as that of the mixed covolume scheme. (see [14])

Now let \( \mathcal{I} := \{1, 2, \ldots, N_A\} \) be a global ordering of the interior edges in \( A_i \), and for \( l = 1, \ldots, NT \), let \( \{1, 2, 3\} \) be a local ordering of the edges of triangular element \( K^{(l)} \). We use \( g_j^{(l)} \) to denote the global edge number of the edge in element \( K^{(l)} \) that has local edge number \( j \). Also we use the notation \( M^{(l)} \) for the local stencil \( M_K \) when \( K = K^{(l)} \). Suppose that \( a \in A_i \) is the intersection of two elements \( K^{(l)} \) and \( K^{(m)} \) with orientation from \( K^{(l)} \) to \( K^{(m)} \). Let \( a = e_s^{(m)} \) on \( K^{(m)} \) and \( a = e_t^{(l)} \) where \( s \) and \( t \) are the local edge numbers. Now let \( K_L = K^{(l)} \) and \( K_R = K^{(m)} \) in (27), then its left hand side can be expressed as

\[
LHS = \sum_{j=1}^{3} m_{ij}^{(l)} p_{g_{ij}^{(l)}} + \sum_{j=1}^{3} m_{sj}^{(m)} p_{g_{sj}^{(m)}}
\]

in terms of pressures \( p, r \in \mathcal{I} \) (globally indexed) and the entries \( m_{ij}^{(l)} \) of the three by three matrix \( M^{(l)} \) and so on. The above suggests we define a global matrix for each element as follows. For each \( l, 1 \leq l \leq NT \) define the matrix \( \tilde{M}_i = \Hat{M}_K \in \mathbb{R}^{N_A \times N_A} \) associated with \( K = K^{(l)} \) so that \( \Hat{m}_{ij} = 0 \) if the (global) edges \( i \) and \( j \) are not in \( K^{(l)} \). Otherwise we set \( \Hat{m}_{ij} = m_{ij}^{(l)} \) where \( i = g_s^{(l)} \) and \( j = g_t^{(l)} \). Obviously, \( \Hat{M}_K \) is a symmetric matrix since \( M_K \) is.

We can now write (32) in terms of this new matrix and global indices. It is then not hard to conclude that

\[
\mathcal{M}_i = \sum_{K \in \mathcal{T}_h} \tilde{M}_K, \quad (33)
\]

\[
\mathcal{F}_i = \sum_{K \in \mathcal{T}_h} \tilde{F}_K. \quad (34)
\]
In fact, let $\alpha \in I$ and the edge $e_\alpha = a = K^{(l)} \cap K^{(m)}$, then 

$$
\left( \sum_{K \in T_h} \hat{M}_K P_i \right)_\alpha = \sum_{\beta=1}^{N A_i} \sum_{k=1}^{N T} \hat{m}_{\alpha \beta}^{(k)} (p_i) \beta \\
= \sum_{\beta=1}^{N A_i} \left\{ \hat{m}_{\alpha \beta}^{(l)} (p_i) \beta + \hat{m}_{\alpha \beta}^{(m)} (p_i) \beta \right\} \\
= \sum_{j=1}^{3} m_{ij}^{(l)} (p_i) g_j^{(l)} + m_{sj}^{(m)} (p_i) g_j^{(m)} \\
= \langle M_i P_i \rangle_\alpha.
$$

(35)

Hence $M_i$ is symmetric. Moreover 

$$P_i^T M_i P_i = \sum_{K \in T_h} P_i^T \hat{M}_K P_i = \sum_{K \in T_h} P_i^T M_K P_K.
$$

(36)

So from Lemma (3.1), we have $P_i^T M_i P_i \geq 0$. Later in Theorem 5.2 we will show the uniqueness of solution for the system (21). Therefore we obtain the following theorem:

**Theorem 3.1** The global stiffness matrix $M_i$ corresponding to the interior edge set $\mathcal{A}_i$ is symmetric positive definite.

**Algorithm:**

```
for each $K \in T_h$
    evaluate $\hat{A}_K, f_K$
    evaluate $M_K$ by (29), $F_K$ by (26)
    assemble $M_K$ to $M_i, F_K$ to $F_i$
enddo
solve $M_i P_i = F_i$
evaluate $U$ from (25), with boundary condition.
```

4 a Mixed box method is a FEM plus local flux recovery

In this section we first prove in Thm. 4.1 equivalence between the pressure approximation in our box method and a $P1$ nonconforming finite element method applied to the elliptic problem (1) with a *modified* right hand side. Then we show this leads to a better understanding of the box method. It turns out that the mixed box method is nothing but a standard FEM with accurate local recovery of the flux. This is the content of Thm. 4.2 with zero absorption.
Throughout the rest of this paper, we use the standard notation $W^{1,p}$ for the usual Sobolev spaces and $| \cdot |_{m,K}$, $\| \cdot \|_{m,K}$ for the semi and full $H^m$–norm, $m = 0, 1, 2$. We omit the subscript $K$ when $K = \Omega$ and sometimes use $\| \cdot \|$ when writing an $L^2$ norm. Also $| \cdot |_{H(\text{div};\Omega)}$ is the $H(\text{div};\Omega)$ semi-norm.

We now show the equivalence theorem.

**Theorem 4.1** The linear system (28) is the same as the discrete system resulting from the standard $P_1$ nonconforming FEM: Find $p_h \in Y_{h,0}$ such that

$$ a_h(p_h, q_h) = (f_h, q_h) \quad \forall q_h \in Y_{h,0} \quad (37) $$

where $f_h = P_h f$ is the $L_2$ projection to the piecewise constant space $L_h$.

**Proof 2** It suffices to show the two methods have the same element stiffness matrix and the same right hand side. The element stiffness matrix associated with element $K$ from (41) can be obtained as follows. Let $\varphi_i = 1 - 2\lambda_i$, $i = 1, 2, 3$ with $\lambda_i$ being the barycentric coordinates. Noting that $\nabla \varphi_i = \frac{|e_i|}{|K|} \mathbf{n}_i$, we have

$$ \int_K (K \nabla \varphi_i) \cdot \nabla \varphi_j d\mathbf{x} = \int_K (\frac{|e_i|}{|K|} \mathbf{n}_i) \cdot (\frac{|e_j|}{|K|} \mathbf{n}_j) d\mathbf{x} $$

$$ = \frac{|e_i|}{|K|} \mathbf{n}_i^T (\int_K K d\mathbf{x}) \frac{|e_j|}{|K|} \mathbf{n}_j $$

$$ = \frac{1}{|K|} \mathbf{e}_i^T A_K \mathbf{e}_j $$

which is exactly (29).

Since $f_h = P_h f$, the $L_2$ projection, $f_h|_K = \frac{1}{|K|} \int_K f$. Thus

$$ (f_h, \varphi_j) = \int_K f_h \varphi_j d\mathbf{x} $$

$$ = \frac{f_h|_K}{|K|} $$

$$ = \frac{|K| f_K}{3} $$

which is exactly (26). This completes the proof.

**Remark.** We must emphasize this theorem is proved because of the hard work in the previous sections. Our line of thinking follows that of [15]. In other words, we try to generalize their method to the anisotropic case, but our crucial new observation is the above theorem. It says the mixed finite volume method is related to a finite element
method. The next theorem says the hard work in the previous sections can be avoided altogether if we have a change of viewpoint.

To bring out an important difference between the mixed box method (finite volume) method and the finite element method, we will do more and actually show the equivalence theorem for the original problem (1) with an added absorption term i.e.,

\[
\begin{aligned}
-\nabla \cdot K \nabla p + \alpha_0 p &= f & \text{in } \Omega, \\
p &= 0 & \text{on } \partial \Omega,
\end{aligned}
\] (38)

where \(\alpha_0\) is a nonnegative piecewise constant function with respect to \(T_h\). The associated weak formulation is to find \(p \in H^1_0\) such that

\[
a(p, q) = (f, q) \quad \forall q \in H^1_0
\] (39)

where

\[
a(p, q) := \int_\Omega (K \nabla p) \cdot \nabla q + \alpha_0 pq \, dx.
\] (40)

Let

\[
Y_{h,0} := \{ q \in L^2(\Omega) : q|_K \in P_1(K), \forall K \in T_h; q \text{ is continuous at the midpoints of interior edges and vanishes at the midpoints of boundary edges} \}.
\]

The standard \(P1\) nonconforming FEM discretization is: Find \(\tilde{p}_h \in Y_{h,0}\) such that

\[
a_h(\tilde{p}_h, q_h) = (f, q_h) \quad \forall q_h \in Y_{h,0}
\] (41)

where

\[
a_h(\tilde{p}_h, q_h) := \sum_{j=1}^{NT} (K \nabla \tilde{p}_h, \nabla q_h)_K + (\alpha_0 \tilde{p}_h, q_h)_K,
\] (42)

where \((\cdot, \cdot)_K\) is the \(L^2\) inner product on \(K\). Define the semi-norm

\[
|q|_h := \left( \sum_{K \in T_h} |q|_{1,K}^2 \right)^{1/2} \quad \forall q \in H^1_0 \oplus Y_{h,0}.
\] (43)

It is clear \(| \cdot |_h\) is a full norm on space \(Y_{h,0}\) in Dirichlet Case. It is well known [7, 4, 5] that the solution \(p_h\) of system (41) converges to solution \(p\) of system (39): there exists a constant \(C\) independent of \(h\) such that

\[
\|p - \tilde{p}_h\|_0 + h\|p - \tilde{p}_h\|_h \leq Ch^2\|p\|_2
\] (44)

provided that the problem data is smooth enough so that the elliptic regularity estimate \(\|p\|_2 \leq C\|f\|_0\) holds. For example, if \(f \in L^2\) and \(K \in C^1(\Omega)\) on a convex domain \(\Omega\), then \(p \in H^2\) is guaranteed. (See p. 4 of [4] and the references therein.) We now show the equivalence theorem.
Theorem 4.2  Consider the problem of finding \((u_h, p_h) \in V_h \times Y_{h,0}\) such that

\[
(K \nabla p_h, \nabla q_h) + \sum_K \alpha_K (p_K, q_h)_K = (f_h, q_h) \quad \forall q_h \in Y_{h,0}
\]  \hspace{1cm} (45)

and over each \(K\) in \(T_h\)

\[
u_h = -A_K \nabla p_h + (f_K - \alpha_K p_K) P_K
\]  \hspace{1cm} (46)

where \(f_h = P_h f\) is the \(L_2\) projection to the piecewise constant space \(L_h\) and \(p_K = \frac{1}{|K|} \int_K p_h dx, f_K = \frac{1}{|K|} \int_K f dx = f_h|_K\) along with the mixed box method of finding \((u_h, p_h) \in V_h \times Y_{h,0}\) such that

\[
(\nabla \cdot \nu_h + \alpha_0 p_h - f, \chi_K) = 0,
\]  \hspace{1cm} (47)

\[
(\nu_h + K \nabla p_h, \chi_K) = 0.
\]  \hspace{1cm} (48)

Then the two above problems are equivalent.

**Proof 3**  We first show that (45)–(46) implies (47)–(48).

Take divergence on (46), recall (13), and integrate against the characteristic function \(\chi_K\) to see (46) implies (47). Now integrate (46) against \(\chi_K\) and use the fact \((P_K, \chi_K) = 0\) (one point quadrature rule using the barycenter \(B\)) to get (48).

Secondly, we prove that (47)–(48) implies (45)–(46).

From (47) we see that on \(K\)

\[
\nabla \cdot \nu_h = f_K - \alpha_K p_K.
\]

Since \(\nu_h \in V_h\), by Taylor’s expansion

\[
u_h = \nu_K + \frac{1}{2} \nabla \cdot \nu_h (x - x_B),
\]

where \(\nu_K = \frac{1}{|K|} \int_K \nu_h dx = -A_K \nabla p_h\) by (48). So

\[
u_h = -A_K \nabla p_h + (f_K - \alpha_K p_K) P_K
\]

which is (46). On the other hand by (48), integration by parts and (47)

\[
\sum_K (K \nabla p_h, \nabla q_h)_K = \sum_K (-\nu_h, \nabla q_h)_K
\]

\[
= \sum_K (\nabla \cdot \nu_h, q_h)_K - (\nu_h \cdot n, q_h)_{\partial K}
\]

\[
= \sum_K (f_K - \alpha_K p_K, q_h)
\]

which is (45). Notice that the \((\cdot, \cdot)_{\partial K}\) terms cancelled upon summation since \(\nu_h \cdot n\) and \(q_h\) are continuous at midpoints of edges.
Remark. One notes that the mixed finite volume method is equivalent to the standard nonconforming method (with a modified right hand side) when there is no absorption term ($\alpha_0 = 0$). In the presence of the absorption term, the mixed method is equivalent to a nonconforming method with the absorption term evaluated by a one-point quadrature at the element barycenter. The approximate $p_h$ so produced by (45) generates automatically a flux with continuous normal components across edges via (46). In other words, the usual nonconforming method does not produce continuous flux unless $\alpha_0 = 0$.

5 Error estimates

In this section we will prove the existence and uniqueness of solution for the system (21) and some error estimates in an energy norm. Through this section the letter $C$ denotes a generic positive constant, independent of $h$ and not necessarily the same in each occurrence. Let us define two energy norms which by (2) are equivalent:

$$|q|_h^2 = \sum_K |q|_{h,K}^2 := \sum_K \int_K |\nabla q|^2 \, dx \quad \forall q \in H^1(\Omega) \oplus Y_h.$$  

$$|q|_E^2 = \sum_K |q|_{E,K}^2 := \sum_K \int_K \nabla q^T K(x) \nabla q \, dx \quad \forall q \in H^1(\Omega) \oplus Y_h.$$  

The next lemma shows they are actually full norms on the space $Y_{h,0} = \{ q_h \in Y_h, q_h = 0 \text{ at all midpoints of boundary edges} \}$.

Lemma 5.1 The discrete energy semi-norm $|q_h|_h$ is a norm on the space $Y_{h,0}$.

Proof 4 Let $q_h \in Y_{h,0}$ such that $|q_h|_h = 0$. The gradient of $q_h$ is zero in each cell $K \in T_h$. Hence $q_h$ is constant in each cell $K$. Since $q_h$ is continuous at the middle point of each edge $e$ of $T_h$ and $q_h = 0$ on $\partial \Omega$, we have $q_h = 0$ in $\Omega$.

The next theorem shows the existence and uniqueness of solution for system (21).

Theorem 5.1 The discrete system (21) has a unique solution $(u_h, p_h) \in V_h \times Y_{h,0}$.

Proof 5 Since the number of unknowns equals the number of equations in (21) we only need to show that $u_h = 0$ and $p_h = 0$ when $f = 0$.

By Thm. 4.1 or 4.2 and the preceding lemma $p_h = 0$ when $f = 0$. On the other hand, $u_h = u_K = -A_K \nabla p_h = 0$.

Also we have the following stability condition:
Lemma 5.2 If \((u_h, p_h) \in V_h \times Y_{h,0}\) is the solution of system (21), then there exist positive constants \(C_1\) and \(C_2\), independent of \(h\), such that

\[
C_1 |p_h|_h \leq \|u_h\|_{0,\Omega} \leq C_2 (|p_h|_h + h \|f\|_{0,\Omega}) \tag{49}
\]

\[
|u_h|_h := \left( \sum_K \|\nabla u_h\|_{0,K}^2 \right)^{1/2} \leq \frac{1}{\sqrt{2}} \|f\|_{0,\Omega} \tag{50}
\]

Proof 6 We first show (50). From (14) \(\nabla u_h = |K| f_K \nabla P_K(x)\) and from (13) \(\|\nabla P_K\|_{0,K}^2 = \frac{1}{2|K|}\), so

\[
|u_h|_h^2 = \sum_K \|\nabla u_h\|_{0,K}^2 = \sum_K |K|^2 |f_K|^2 \|\nabla P_K\|_{0,K}^2 \leq 1/2 \sum_K |K| |f_K|^2 \leq 1/2 \|f\|_{0,\Omega}^2
\]

and we obtain (50).

Next we show (49). By the linearity of \(p_h\), bounds (2) on \(A_K\), and the Cauchy-Schwarz inequality, we get

\[
|p_h|_h^2 = \sum_K \int_K |\nabla p_h|^2 dx = \sum_K |K| \|\nabla p_h\|^2 \leq \sum_K |K| (A_K^{-1} u_K)^T (A_K^{-1} u_K) \leq C \sum_K |K| u_K^T u_K \leq C \sum_K \int_K |u_h(x)|^2 dx = C \|u_h\|_{0,\Omega}^2,
\]

whence

\[
C_1 |p_h|_h \leq \|u_h\|_{0,\Omega}.
\]

Again by (14), bound (2) on \(A_K\),

\[
\|u_h\|_{0,K} \leq \|A_K \nabla p_h\|_{0,K} + |K| |f_K| \|P_K\|_{0,K} \leq \left( \int_K \nabla p_h^T A_K^2 \nabla p_h dx \right)^{1/2} + |K| |f_K| \|P_K\|_{0,K} \leq C \|\nabla p_h\|_{0,K} + |K| |f_K| \|P_K\|_{0,K} \leq C |p_h|_{h,K} + |K| |f_K| \|P_K\|_{0,K}
\]

\[\text{From (13) we have}\]

\[
\|P_K\|_{0,K}^2 = \frac{1}{4|K|^2} \int_K (x - x_B)^2 + (y - y_B)^2 dxdy \leq \frac{h_K^2}{4|K|}.
\]
By the regularity assumption on the triangulations we know $h_K/|K|^{1/2} \leq C$ for all $K$. Also notice $|f_K| \leq \|f\|_{0,K}/|K|^{1/2}$, we obtain
\[
\|u_h\|_{0,K} \leq C(|p_h|_{h,K} + |K|^{1/2}\|f\|_{0,K}) \leq C(|p_h|_{h,K} + h\|f\|_{0,K}).
\]
Summing over $K$, we get
\[
\|u_h\|_{0,\Omega} \leq C(|p_h|_h + h\|f\|_{0,\Omega}).
\]
This completes the proof.

Our main result in this section is the following error estimate theorem.

**Theorem 5.2** Let the problem data of (1) be smooth enough so that the pressure solution $p \in H^2 \cap H^1_0$ and $u(x) = -\mathcal{K}(x)\nabla p(x) \in H^1(\Omega)^2$. Then there exists a constant $C$ independent of $h$ such that
\[
\|p - p_h\|_0 \leq Ch^2(|f|_h + \|f\|_0), \quad \text{if } f \in H^1(K) \quad \forall K \in \mathcal{T}_h, \tag{51}
\]
\[
\|p - p_h\|_h \leq Ch\|f\|_0 \tag{52}
\]
\[
\|u - u_h\|_0 \leq Ch(\|f\|_0 + \|u\|_0) \tag{53}
\]
\[
|u - u_h|_{H(\text{div};\Omega)} \leq Ch|f|_h \quad \text{if } f \in H^1(K) \quad \forall K \in \mathcal{T}_h. \tag{54}
\]
Also we assume that $\mathcal{K} \in W^{1,\infty}$.

**Remark.** Note that the condition of the function $f \in H^1(K)$ for all $K$ can be satisfied if we impose it on the coarsest mesh in the context of local refinement.

**Proof 7** i) proof of (51) and (52). First we show that
\[
\|\tilde{p} - p\|_0 \leq Ch^2|f|_h
\]
where $\tilde{p} \in H^1_0$ is the solution of $a(\tilde{p}, q) = (f_h, q) \forall q \in H^1_0$ and $p$ is the solution of $a(p, q) = (f, q) \forall q \in H^1_0$. Subtracting these two bilinear forms, we get
\[
a(p - \tilde{p}, q) = (f_f - f_h, q) \forall q \in H^1_0.
\]
Noticing that $\int_K f - f_h dx = 0$, we have $(f_f - f_h, q) = (f_f, q - q_h)_K$ where $q_h|_K = 1/|K| \int_K q dx$ is constant on each $K$. Then by the Cauchy–Schwarz inequality and an interpolation theorem, we have
\[
|(f_f - f_h, q - q_h)| = \left| \sum_K (f_f - f_h, q - q_h)_K \right| \leq C h^2 \sum_K |f|_{1,K}|q|_{1,K} \leq C h^2 |f_f|_1 |q|_1
\]
Thus
\[ |a(p - \tilde{p}, q)| \leq Ch^2|f||q|. \]
Taking \( q = p - \tilde{p} \) and using the coercivity of \( a(\cdot, \cdot) \), we have
\[ |p - \tilde{p}|^2 \leq Ch^2|f||p - \tilde{p}| \]
and hence
\[ |p - \tilde{p}| \leq Ch^2|f|. \tag{56} \]
By the Poincaré inequality, we get
\[ \|p - \tilde{p}\|_0 \leq Ch^2|f|. \tag{57} \]
Now note that \( p_h \in Y_{h,0} \) is the solution of the nonconforming method
\[ a_h(p_h, q_h) = (f_h, q_h) \quad \forall q_h \in Y_{h,0} \]
associated with the variational problem of finding \( \tilde{p} \in H^1_0 \) such that
\[ a(\tilde{p}, q) = (f_h, q) \quad \forall q \in H^1_0. \]
(The right hand side contains \( f_h \) instead of \( f \)!) Hence
\[ \|p_h - \tilde{p}\|_0 + h|\tilde{p} - p_h|_h \leq Ch^2||\tilde{p}||_2, \tag{58} \]
where we have applied (44) with the right side \( f_h \) and all the remarks concerning regularity of the solution made there then also apply here.

Now by the triangle inequality, (57), (58), the stability condition and the fact that \( f_h \) is the \( L_2 \) projection of \( f \), we have
\[ \|p_h - p\|_0 \leq \|p_h - \tilde{p}\|_0 + ||\tilde{p} - p||_0 \]
\[ \leq Ch^2||\tilde{p}||_2 + Ch^2|f|h \]
\[ \leq Ch^2||f_h||_0 + Ch^2|f|h \]
\[ \leq Ch^2||f||_0 + Ch^2|f|h. \]
This completes the proof of (51).

As for the proof of (52), first note that from (55) we can use the fact that \( ||f - f_h||_{0,K} \leq ||f||_{0,K} \) and an interpolation theorem to obtain
\[ |(f - f_h, q - q_h)| \leq Ch \sum_K \|f\|_{0,K} |q|_{1,K} \]
\[ \leq Ch\|f\|_{0}|q|_{1} \]
and consequently instead of (56) we have \(|\tilde{p} - p|_{1} \leq Ch\|f\|_{0}\); arguing as before. Now
\[
|p_{h} - p|_{h} \leq |p_{h} - \tilde{p}|_{h} + |\tilde{p} - p|_{h} \\
\leq Ch||\tilde{p}||_{2} + Ch\|f\|_{0}.
\]

ii) proof of (53).
From (14) we have
\[
\|u_{h}(x) - u(x)\|_{K} = -A_{K}\nabla p_{h} + K(x)\nabla p + |K|f_{K}P_{K}(x).
\]
So
\[
\|u_{h} - u\|_{0,K} \leq \|A_{K}\nabla p_{h} - K(x)\nabla p\|_{0,K} + |K|\|f_{K}\|\|P_{K}\|_{0,K},
\]
and since \(|K|\|f_{K}\|\|P_{K}\|_{0,K} \leq Ch\|f\|_{0,K}\) as shown in Lemma 5.2 we have
\[
\|u_{h} - u\|_{0,K} \leq \|A_{K}\nabla p_{h} - K(x)\nabla p\|_{0,K} + Ch\|f\|_{0,K}
\]
whereas by the triangle inequality and the interpolation theorem
\[
\|A_{K}\nabla p_{h} - K(x)\nabla p\|_{0,K} \leq (\nabla p_{h}^{T} \int_{K} (A_{K} - K)^{2} d\mathbf{x} \nabla p_{h})^{1/2} + \int_{K} (K(\nabla p_{h} - \nabla p))^{2} d\mathbf{x})^{1/2} \\
\leq Ch|p_{h}|_{h,K} |K|_{1,\infty,K} + C|p_{h} - p|_{E,K} \\
\leq Ch\|u_{h}\|_{0,K} + C|p_{h} - p|_{E,K} \\
\leq C\{h\|u - u_{h}\|_{0,K} + h\|u\|_{0,K} + C|p_{h} - p|_{E,K}\}
\]
where we have used the stability (49) restricted on \(K\) as shown in its proof. Now taking \(h\) small enough to move the first term on the right side we have
\[
\|u_{h} - u\|_{0,K} \leq C\{|p_{h} - p|_{E,K} + h\|u\|_{0,K} + h\|f\|_{0,K}\}.
\]
Summing over \(K\) and using (52) we have
\[
\|u_{h} - u\|_{0} \leq Ch(\|f\|_{0} + \|u\|_{0}).
\]

iii) proof of (54).
By (4) \(\nabla \cdot u_{h}(x) = f_{K}\) and hence
\[
\|\nabla \cdot u_{h} - \nabla \cdot u\|_{0,K} = \|f - f_{K}\|_{0,K} \leq Ch|f|_{1,K}
\]
where an interpolation theorem is used. Summing over \(K\), we get
\[
|u - u_{h}|_{H(div;\Omega)} \leq \|\nabla \cdot u_{h} - \nabla \cdot u\|_{0,\Omega} \leq Ch|f|_{h}.
\]
6 Numerical Examples

Notice that the error estimate theorem and the algorithm are valid for both Dirichlet and Neumann problems. We present numerical results for both cases. We partition the unit square \([0, 1] \times [0, 1]\) into squares evenly in both directions with the diagonals running from the upper-left corner of each triangle to its lower-right corner. We use the incomplete LU preconditioned conjugate gradient method to solve all the problems. The integral of \(f\) over element \(K\) is computed by the midpoint rule using the three edges of the triangle. Our experiments suggest second order approximation in all cases. The discrete norms in which the errors are estimated are as follows.

6.1 Choice of discrete norms

In Thm. 5.2 we predicted first order convergence in the \(H(\text{div}; \Omega)\) norm for flux \(u\) and second order convergence in the \(L_2\) norm for the pressure \(p\). We need to choose proper discrete norms to measure the error between true solution and computed solution.

Let \((x_i, y_j)\) be the center of square \((i, j)\) with \(x_i = (i - 1/2)h, y_j = (j - 1/2)h, h = 1/n, i, j = 1, 2, \ldots, n\). Let \(p_{ij}\) be the computed pressure at \((x_i, y_j)\). We define

\[
p_{\text{Err}}\text{abs} = \|p - p_h\| := \left[ \sum_{i,j=1}^{n} h^2 (p(x_i, y_j) - p_{ij})^2 \right]^{1/2},
\]

i.e. a discrete \(L^2\)-norm of the error \(p - p_h\).

Due to (59), the \(H(\text{div}; \Omega)\)-seminorm of the error in the flux is directly related to \(f - f_K\), and so for the flux we will use only an equivalent \(L_2\) discrete norm:

\[
u_{\text{Err}}\text{abs} = \|u - u_h\| := \left[ \sum_K \left( \int_{\partial K} \left( (u - u_h) \cdot \mathbf{n} \right) ds \right)^2 \right]^{1/2},
\]

where the edge integrals are evaluated by the midpoint rule [11]. (Note that the \(x\) components are picked out by the vertical edges, the \(y\) components by the horizontal edges, etc.)

We also compute the relative error

\[
p_{\text{Err}}\text{rel} := p_{\text{Err}}\text{abs}/\|p\|
\]

and

\[
u_{\text{Err}}\text{rel} := u_{\text{Err}}\text{abs}/\|u\|.
\]
6.2 Dirichlet problems

We consider the following Dirichlet problem

\[
\begin{aligned}
-\nabla \cdot K \nabla p &= f \quad \text{in } \Omega, \\
p &= 0 \quad \text{on } \partial \Omega.
\end{aligned}
\] (60)

In examples 1–4, the true pressure is \( p = (x^2 - x)(y^2 - y) \) on the unit square.

**Example 1.** The mobility tensor \( K = \text{diag}(1 + 10x^2 + y^2, 1 + x^2 + 10y^2) \).

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Example 1.} & h = 1/16 & h = 1/32 & h = 1/64 & h = 1/128 & \text{order} \\
\hline
\text{pErr\_abs} & 8.7748e-5 & 2.2318e-5 & 5.6041e-6 & 1.4026e-6 & \approx 2 \\
\text{pErr\_rel} & 0.0026 & 6.6952e-4 & 1.6812e-4 & 4.2077e-5 & \approx 2 \\
\text{uErr\_abs} & 0.0113 & 0.0029 & 7.1931e-4 & 1.8029e-4 & \approx 2 \\
\text{uErr\_rel} & 0.0086 & 0.0021 & 5.1929e-4 & 1.2905e-4 & \approx 2 \\
\hline
\text{Length of } P & 736 & 3,008 & 12,160 & 48,896 & \\
\text{Length of } U & 736 & 3,008 & 12,160 & 48,896 & \\
\hline
\end{array}
\]

**Example 2.** The mobility tensor \( K = \text{diag}(10^4, 1) \).

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Example 2.} & h = 1/16 & h = 1/32 & h = 1/64 & h = 1/128 & \text{order} \\
\hline
\text{pErr\_abs} & 9.1815e-5 & 2.3286e-5 & 5.8558e-6 & 1.4657e-6 & \approx 2 \\
\text{pErr\_rel} & 0.0028 & 6.9859e-4 & 1.7567e-4 & 4.3971e-5 & \approx 2 \\
\text{uErr\_abs} & 8.0759 & 2.0653 & 0.5218 & 0.1311 & \approx 2 \\
\text{uErr\_rel} & 0.0057 & 0.0014 & 3.5417e-4 & 8.8436e-5 & \approx 2 \\
\hline
\text{Length of } P & 736 & 3,008 & 12,160 & 48,896 & \\
\text{Length of } U & 736 & 3,008 & 12,160 & 48,896 & \\
\hline
\end{array}
\]

**Example 3.** In this example, we study the effect of discontinuous mobility matrix \( K \).

Let

\[
K = \begin{bmatrix}
10000 & 0 \\
0 & 1
\end{bmatrix}
\]
on the left half unit square, and

\[
K = \begin{bmatrix}
1 & 0 \\
0 & 2
\end{bmatrix}
\]
on the right half unit square. We find the approximation is as good as continuous cases.

**Example 4.** Let

\[
K = \begin{bmatrix}
1 + 10x^2 + y^2 & 1/2 + x^2 + y^2 \\
1/2 + x^2 + y^2 & 1 + x^2 + 10y^2
\end{bmatrix}
\]
### Table 3: Error behavior for Dirichlet problem

<table>
<thead>
<tr>
<th>Example 3.</th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{Err}}_{\text{abs}}$</td>
<td>1.6425e-4</td>
<td>4.1581e-5</td>
<td>1.0439e-5</td>
<td>2.6128e-6</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$p_{\text{Err}}_{\text{rel}}$</td>
<td>0.0049</td>
<td>0.0012</td>
<td>3.1317e-4</td>
<td>7.8382e-5</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$u_{\text{Err}}_{\text{abs}}$</td>
<td>6.2337</td>
<td>1.5882</td>
<td>0.4005</td>
<td>0.1005</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$u_{\text{Err}}_{\text{rel}}$</td>
<td>0.0062</td>
<td>0.0015</td>
<td>3.8441e-4</td>
<td>9.5877e-5</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>Length of $P$</td>
<td>736</td>
<td>3,008</td>
<td>12,160</td>
<td>48,896</td>
<td></td>
</tr>
<tr>
<td>Length of $U$</td>
<td>736</td>
<td>3,008</td>
<td>12,160</td>
<td>48,896</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Error behavior for Dirichlet problem

<table>
<thead>
<tr>
<th>Example 4.</th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{Err}}_{\text{abs}}$</td>
<td>1.4595e-4</td>
<td>3.7458e-5</td>
<td>9.4305e-6</td>
<td>2.3620e-6</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$p_{\text{Err}}_{\text{rel}}$</td>
<td>0.0044</td>
<td>0.0011</td>
<td>2.8292e-4</td>
<td>7.0861e-5</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$u_{\text{Err}}_{\text{abs}}$</td>
<td>0.0168</td>
<td>0.0043</td>
<td>0.0011</td>
<td>2.7533e-4</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$u_{\text{Err}}_{\text{rel}}$</td>
<td>0.0114</td>
<td>0.0028</td>
<td>7.0927e-4</td>
<td>1.7784e-4</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>Length of $P$</td>
<td>736</td>
<td>3,008</td>
<td>12,160</td>
<td>48,896</td>
<td></td>
</tr>
<tr>
<td>Length of $U$</td>
<td>736</td>
<td>3,008</td>
<td>12,160</td>
<td>48,896</td>
<td></td>
</tr>
</tbody>
</table>

### 6.3 Neumann problems

\[
\begin{align*}
\begin{cases}
-\nabla \cdot K \nabla p &= f \text{ in } \Omega, \\
K \nabla p \cdot n &= 0 \text{ on } \partial \Omega.
\end{cases}
\end{align*}
\]

**Example 5.** The true pressure is $p = \cos(2\pi x) \cos(2\pi y)$ and $K = \text{diag}(\cos(2\pi y) + 2, \cos(2\pi x) + 2)$.

### Table 5: Error behavior for Neumann problem

<table>
<thead>
<tr>
<th>Example 5.</th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{Err}}_{\text{abs}}$</td>
<td>0.0110</td>
<td>0.0028</td>
<td>6.9570e-4</td>
<td>1.7399e-4</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$p_{\text{Err}}_{\text{rel}}$</td>
<td>0.0221</td>
<td>0.0056</td>
<td>0.0014</td>
<td>3.4798e-4</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$u_{\text{Err}}_{\text{abs}}$</td>
<td>0.1383</td>
<td>0.0353</td>
<td>0.0089</td>
<td>0.0022</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>$u_{\text{Err}}_{\text{rel}}$</td>
<td>0.0101</td>
<td>0.0026</td>
<td>6.4814e-4</td>
<td>1.6226e-4</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>Length of $P$</td>
<td>800</td>
<td>3,136</td>
<td>12,416</td>
<td>49,408</td>
<td></td>
</tr>
<tr>
<td>Length of $U$</td>
<td>800</td>
<td>3,136</td>
<td>12,416</td>
<td>49,408</td>
<td></td>
</tr>
</tbody>
</table>

**Example 6.** The true (oscillatory) pressure is $p = \cos(2\pi x) \cos(10\pi y)$ and $K = I$. 
Table 6: Error behavior for Neumann problem

<table>
<thead>
<tr>
<th>Example 6.</th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>pErr_abs</td>
<td>0.0441</td>
<td>0.0130</td>
<td>0.0034</td>
<td>8.4912e-4</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>pErr_rel</td>
<td>0.0882</td>
<td>0.0260</td>
<td>0.0067</td>
<td>0.0017</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>uErr_abs</td>
<td>4.1789</td>
<td>1.1274</td>
<td>0.2882</td>
<td>0.0725</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>uErr_rel</td>
<td>0.1845</td>
<td>0.0498</td>
<td>0.0127</td>
<td>0.0032</td>
<td>$\approx 2$</td>
</tr>
<tr>
<td>Length of $P$</td>
<td>800</td>
<td>3,136</td>
<td>12,416</td>
<td>49,408</td>
<td></td>
</tr>
<tr>
<td>Length of $U$</td>
<td>800</td>
<td>3,136</td>
<td>12,416</td>
<td>49,408</td>
<td></td>
</tr>
</tbody>
</table>

References


Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, Ohio, U.S.A
email: chou@zeus.bgsu.edu
http://www-math.bgsu.edu/chou/.

Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, Ohio, U.S.A
email: tang@bgnet.bgsu.edu
2개 반경을 갖는 Diamond Wheel의 절삭 공식에 관한 연구

김희섭, 김희렬, 조창목

요 약
We find a cutting formula of the Diamond wheel with two radii.

I. 서론

곡면을 가공하고자 하면 곡면식과 NC(Numerical Control) 공작기계가 필요하다. 자유곡면은 Ferguson Surface, Bezier Surface, B-spline Surface, Polynomial Surface 등으로 많이 표현된다. 어떤 곡면이든 지간에 미분기하에 의해 벡선 Vector를 표현할 수 있다. 공작기계에서 사용되는 공구는 보통 금속가공에 Ball-endmill을 쓰고 가공 Filleted-endmill도 쓴다. 하지만 유리가공이나 금형가공 중에서도 특수한 정도를 요구하는 것에선 Diamond wheel을 쓴다. 이 논문에서는 두 개의 반경을 갖는 Diamond wheel 공구를 쓸 경우 기존에 나와 있는 식과 다른 방식의 공식을 유도한다.

II. 본론

보통의 NC 기계의 가공에서의 Ball-endmill은 벡선 방향이 Endmill의 중심을 향하고 있어 이 벡선 벡터를 사용하여 공구 위치를 계산하지만, 정밀 가공용 NC의 경우 공구 형상이 보통의 NC 기계와는 다르므로 (그림 1 참조) 가공면의 벡선이 Diamond Wheel의 중심과는 일치하지 않는다. 그러므로 공구 보정을 위하여는 가공면에서 공구의 중심을 향하는 vector를 구하는 Algorithm이 필요하다. 공작기계가 최종적으로 위치를 인식하는 CL(Cutting Location) data는 가공하고자 하는 곡면과 공구의 접촉점이 주어지면 이에 해당하는 공구 위치를 계산함으로써 얻어진다. CL data를 구하려면 곡면상의 접촉면에서의 벡선 벡터를 계산하여야 한다. 곡면이 매개방정식 \( r(u, v) \) 인 경우 매개방정식을 \( u \) 방향 및 \( v \) 방향에 대하여 미분하고 두 미분 값을 곱을 취하여 구한다. 단위벡선 벡터는

\[
\hat{n} = \frac{r_u \times r_v}{|r_u \times r_v|}
\]

이다.

1991 Mathematics Subject Classification. 65D17
key-words and Phrases : Numerical control, 벡선벡터.
그림 1 가공 NC기계의 Diamond Wheel

다방식 \( z = f(x, y) \)의 경우는

\[
\vec{n} = \frac{(-f_x, -f_y, 1)}{\sqrt{1 + f_x^2 + f_y^2}}
\]

이다.

그림 2에 보인 바와 같이 Ball-endmill과 Filleted-endmill의 경우 CL데이터는 다음 식으로 표현된다.

\[
\vec{r}_{CL} = \vec{r} + R(\vec{n} - \vec{u})
\]

\[
\vec{r}_{CL} = \vec{r} + a(\vec{n} - \vec{u}) \frac{(R - a)}{\sqrt{1 - (\vec{n} \cdot \vec{u})^2}} (\vec{n} - (\vec{n} \cdot \vec{u})\vec{u}).
\]

여기서 \( \vec{r} \)는 접촉점의 위치 벡터, \( \vec{r}_{CL} \)는 공구 끝점의 위치벡터를 나타낸다. \( \vec{u} = (0, 0, 1) \), \( \vec{v} \)은 공면의 법선 벡터이고 \( R \)은 공구반경, \( a \)는 Filleted-endmill의 경우 작은 반경이다.

이제부터 이 논문에서 찾고자 하는 식을 유도한다. 그림 3에서 Torus 의 Parameter 표현법은

\[
\vec{r}(u, v) = (a + b \cos v) \cos u \vec{i} + (a + b \cos v) \sin u \vec{k} + b \sin v \vec{j}
\]

\( z < 0 \)인 영역에서 토러스의 외측을 밀으면 \( \pi \leq u \leq 2\pi, -\frac{\pi}{2} \leq v \leq \frac{\pi}{2} \) 이다. 여기서 \( x, y, z \)좌표는 실제로 공작기계에 Diamond Wheel이 작동되었을 때의 \( x, y, z \)좌표와 일치하도록 잡았기 때문에 (2)식과 같이 주어진다. \( \vec{r}_u, \vec{r}_v \)은 계산에 의하여

\[
\vec{r}_u = -(a + b \cos v) \sin u \vec{i} + (a + b \cos v) \cos u \vec{k} + 0 \cdot \vec{j}
\]

\[
\vec{r}_v = -b \sin v \cos u \vec{i} - b \sin v \sin u \vec{k} + b \cos v \cdot \vec{j}
\]
(a) Ball-endmill (b) Filleted-endmill

그림 2 공구 접촉점과 CL data

그림 3 Torus에의한 공구의 중심에서 접촉점까지의 거리
로 주어진다.
\[
\vec{r}_u \times \vec{r}_v = -b(a + b \cos v) \cos u \cos v \vec{i} - b(a + b \cos v) \sin u \cos v \vec{k} - b(a + b \cos v) \sin v \vec{j}
\]  
(3)

이고 (3)이 Diamond Wheel의 법선 벡터이다.
연속하는 면 \( r_c = (x_c, y_c, z_c) \)의 벡선 벡터가 \( \vec{r}_c = n_x \vec{i} + n_y \vec{j} + n_z \vec{k} \) 로 한다. 작업 면 벡선 vector 와 Diamond Wheel 벡선 vector를 일치시키면
\[
n_x = \frac{b(a + b \cos v) \cos v \sin u}{b(a + b \cos v) \cos v \cos u} = \tan u.
\]
그러므로 \( u = \tan^{-1} (\frac{n_x}{n_z}) \). 단, \( \pi \leq u \leq 2\pi \)이기 때문에 계산식에서는
\[
\begin{align*}
    u &= u + 2\pi : u \leq 0 \\
    u &= u + \pi : u > 0
\end{align*}
\]
이다.
\[
n_y = \frac{b(a + b \cos v) \sin v}{b(a + b \cos v) \cos v \sin u} = \frac{\tan v}{\sin u}.
\]
그러므로 \( v = \tan^{-1} (\frac{n_y}{n_x}) \)이다. 위에서 구한 \( u, v \)값을 식(2)에 대입하면 \( \vec{r} \)가 얻어진다. 식 (2) 에 의한 위치 vector \( \vec{r} \)을 구하면 Diamond Wheel의 중심 위치 \( \vec{r}_{CL} \)는
\[
\vec{r}_{CL} = \vec{r}_c - \vec{r}
\]
(4)
이다. 이것이 우리가 구하는 공작기계의 위치를 인식하는 공구인 Diamond Wheel의 최종 CL data이다.

III. 토의

Filleted-endmill의 식(1)과 새로운 식(4)는 모두 CL data를 나타내지만 식(4)는 2개 반경을 갖는 Diamond Wheel을 Torus의 바깥면으로 보고 계산한 것이다. 이 논문에서 정밀 가공을 필요로하는 금형 또는 유리가공에서 사용되는 새로운 식 (4)를 유도하였다.

참고 문헌

1. 최명규, 곡면 모형 및 3차원 NC가공, 백한 출판사.
2. 전자수, 자유곡면 NC가공에서의 공구간섭 방지, Ph. D Thesis, KAIST
Department of Mathematics and Information, 
Kyungwon University 
Songnam, Kyunggido, 461-701, Korea 
hskimm@mail.kyungwon.ac.kr

Department of Refrigeration and Air/Conditioning Engineering, 
Pukyong National University 
NamGu, Pusan, 608-737, Korea

Corporate R&D Center, 
SAMSUNG SDI CO., LTD 
575, Shin-Dong, Paldal-Gu, 
Suwon-City, Kyungki-Do, 
442-391, Korea
A Robust Preconditioner on the CRAY-T3E for Large Nonsymmetric Sparse Linear Systems

Sangback Ma and Jaeyoung Cho

Abstract

In this paper we propose a block-type parallel preconditioner for solving large sparse nonsymmetric linear systems, which we expect to be scalable. It is Multi-Color Block SOR preconditioner, combined with direct sparse matrix solver. For the Laplacian matrix the SOR method is known to have a nondeteriorating rate of convergence when used with Multi-Color ordering. Since most of the time is spent on the diagonal inversion, which is done on each processor, we expect it to be a good scalable preconditioner. Finally, due to the blocking effect, it will be effective for ill-conditioned problems. We compared it with four other preconditioners, which are ILU(0)-wavefront ordering, ILU(0)-Multi-Color ordering, SPAI(SParse Approximate Inverse), and SSOR preconditioner. Experiments were conducted for the Finite Difference discretizations of two problems with various meshsizes varying up to 1024 x 1024, and for an ill-conditioned matrix from the shell problem from the Harwell-Boeing collection. CRAY-T3E with 128 nodes was used. MPI library was used for interprocess communications. The results show that Multi-Color Block SOR and ILU(0) with Multi-Color ordering give the best performances for the finite difference matrices and for the shell problem only the Multi-Color Block SOR converges.

1 Introduction

Iterative solutions of large sparse linear systems require the use of preconditioning techniques in order to converge in a reasonable number of iterations. Reordering the equations through multi-coloring provides a parallelism of order \( N \), where \( N \) is the dimension of the given matrix. For example, if the matrix has property A, as is the case for the standard 5-point matrices obtained from centered Finite Difference (FD) discretizations of elliptic Partial Differential Equations (PDE’s), there is a partition of the grid-points in two disjoint subsets such that the unknowns of any one subset are only related to unknowns of the other subset. This enables to produce a reordered matrix having a block-tridiagonal matrix, where the diagonal blocks are diagonal matrices. There are several different ways of exploiting this structure. For example, the unknowns associated with one of the subsets can be easily eliminated, and the resulting reduced system is often well-conditioned. This ‘two-coloring’ often referred to as a red-black
or checkerboard ordering, can be generalized to arbitrary sparse matrices by using multi-coloring. But the drawback of this approach is that it often suffers from the deterioration of the rate of convergence with certain iterative methods, such as in preconditioned CG (Conjugate Gradient) method.

Wavefront or level scheduling [11] technique is still another way to achieve a parallelism while preserving the initial ordering. It does not suffer from the deterioration of the rate of convergence as with multi-coloring, but the maximum parallelism is determined by the lengths of the wavefront, which is often nonuniform.

In the SPAI approach a sparse approximate inverse is computed explicitly, and then applied as a preconditioner to an iterative method. The computation of the preconditioner is inherently parallel, and its application only requires a matrix-vector product. The sparsity pattern of the approximate inverse can be fixed in advance, or expanded for more accuracy. The SPAI computes the entries, M, so that it minimizes \( \| A M - I \| \) under suitable norm, often the Frobenius norm. It decomposes into independent least squares problems, which can be solved in parallel. There are many variations in the SPAI approach depending on the way the residual norm is minimized and the choice of sparsity pattern. In this paper we adopted the approach by Grote and Huckle [4].

The Multi-Color Block SOR method combines the Multi-Coloring with the Block SOR method. It is known that for the 5-point Laplacian the SOR method has the same rate of convergence even with the Multi-Coloring, while the ILU(0) preconditioned CG method has a low rate of convergence with the Multi-Coloring. Hence, we expect the Multi-Color Block SOR method to be a good choice.

The CRAY-T3E computer in ETRI, Korea is a massively parallel message-passing machine with the 136 individual processing node (PE)s interconnected in a 3D-Torus structure. Each PE, a DEC Alpha EV5.6 chip, is capable of delivering up to 900 Megaflops, amounting to 115 GigaFlops in total. Each PE has 128 MBs of core memory.

2 Multi-coloring and Wavefront reordering

2.1 Multi-color reordering

Given a mesh, multi-coloring consists of assigning a color to each point so that the couplings between two points of the same color are eliminated in the discretization matrix. For example, for the 5-point Laplacian on a square with two colors in the checkerboard fashion we can remove the coupling between any two points of the same color, so that the values at all points of one color can be updated simultaneously. Similarly, four colors are needed to color the grid points of the 9-point Laplacian. However, it has been known that the convergence rate for the reordered systems often deteriorates. For the model problem SSOR and preconditioned-CG with the Red/Black ordering have a worse convergence rate than with the natural ordering, while SOR has the same rate if optimal \( \omega \) is used. The table 1 contains the rates of convergence for SSOR with optimal \( \omega \), and ILU(0) preconditioned CG methods with natural and
Table 1: Rate of convergence when reordering is used. $h$ is the meshsize.

<table>
<thead>
<tr>
<th></th>
<th>SOR</th>
<th>SSOR</th>
<th>ILU-CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Ordering</td>
<td>$O(h)$</td>
<td>$O(h)$</td>
<td>$O(\sqrt{h})$</td>
</tr>
<tr>
<td>Red/Black Ordering</td>
<td>$O(h)$</td>
<td>$O(h^2)$</td>
<td>$O(h)$</td>
</tr>
</tbody>
</table>

The Red/Black ordering can be extended to Multi-Color ordering schemes. For the nine-point Laplacian with properly selected $\omega$ the convergence rate of SOR remains the same as with the natural ordering. O’Leary[7] has considered several other ordering schemes for the 9-point Laplacian and has shown that the convergence rate of SOR iteration is no worse than that with the natural ordering.

However, even though the performance is not as sensitive to $\omega$ as it is to $\rho$ in ADI, it is preferable to have an optimal $\omega$ for the best performance as a preconditioner. For model problems with the 5-point, Laplacian the optimal $\omega$ both with the natural ordering and red/black ordering has been determined, but the determination of such $\omega$ is very difficult, in general. Also, when the points of one color is being updated, the other points are idle. This could cause CPU idle in parallel execution, depending on the actual mapping between the grid points and the processors.

Regarding the blocked version of SOR, blocking in general increases the convergence rate while the cost of one solution per iteration increases. For the model problem with the 5-point Laplacian the convergence rate of the line-SOR method is $2\sqrt{2\pi h}$, $\sqrt{2}$ times that of point SOR method. If the square was divided into $L$ subsquares, each with size of $q \times q$, $q = n/\sqrt{L}$, the spectral radius of Block Jacobi has been found to be[2]

$$Sp(M_J) \approx 1 - \frac{q\pi^2h^2}{2},$$  \hspace{1cm} (1)$$$M_J$ is the Block Jacobi iteration matrix. From the relation about optimal $\omega$ and the spectral radius of the Block SOR iteration matrix [15] the optimal $\omega^* \approx 2/(1 + \sqrt{q}\pi h)$, and hence the rate of convergence is $2\sqrt{q}\pi h$, $\sqrt{q}$ times that of point SOR. On the other hand since the diagonal blocks are no longer diagonal matrices, inverting them involves some extra costs.

As preconditioners SOR and Block SOR methods are expected to perform well as good preconditioners, although analytic explanation is not available. We could combine Block SOR with Multi-coloring for parallel preconditioners.

Let the domain be divided into $L$ blocks. Suppose that we apply a multi-coloring technique, such as a greedy algorithm described in [14], to these blocks so that a block of one color has no coupling with a block of the same color. Let $D_j$ be the coupling within the block $j$, and the $E_{j,k}, k = 1, q, k \neq \text{color}(j)$ the coupling between the $j$-th block and the other blocks of color $k$, where $\text{color}(j)$ is the color of the block $j$.

Then, we can describe the Multi-Color Block SOR as follows.
Algorithm 2.1 Multi-Color Block SOR

Let \( q \) be the total number of colors, and \( \text{color}(i), i=1, L, \) be the array of the color for each block.
1. Choose \( u_0 \), and \( \omega > 0 \).
2. For \( i > 0 \) Until Convergence Do
3. For kolor = 1, \( q \) Do
4. For \( j = 1, L \) Do
5. if\( \text{color}(j) == \text{kolor} \) then
6. \( (u_{i+1}^{GS})_j = D_j^{-1}(b - \sum_{k=1}^{q-1} E_{j,k}u_i) \).
7. \( u_{i+1} = u_i + \omega * (u_{i+1}^{GS} - u_i) \).
8. endif
9. Endfor
10. Endfor
11. Endfor

The \( u_{i}^{GS} \) is the \( i \)-th update of Block Gauss-Seidel iteration. If \( \omega = 1 \) it is equivalent to Block Gauss-Seidel method. Note that the innermost loop in line six and seven can be executed in parallel.

2.2 Wavefront-ordering (Level scheduling)

Rather than pursuing the parallelisms through reordering, the wavefront technique exploits the structure of the given matrix. If the matrix comes from the discretizations of PDEs such as by FDM or FEM, the value of a certain node is usually dependent on only the values of its neighbors. Hence, once the values of its neighbors are known that node can be updated.

For example, let us assume that we have an ILU(0) factorization for the 5-point Laplacian for the Poisson equation, and for a preconditioner we need to solve

\[
Lz = y,
\]

and

\[
Ux = z
\]

Wavefront technique (or Level scheduling) is a process of finding new ordering of the nodes so that the new matrix would look like as in Fig. (1), where the \( L_i \)’s are diagonal blocks. This technique would work equally well for three dimensional problems as well as two dimensional problems. For references, see [11]
Figure 1: Block partitioning for $L$

3 SPAI Preconditioner

We look for $M$ such that $\|AM - I\|_F$ is minimized where $\|\cdot\|_F$ stands for the Frobenius norm. Then since

$$\|AM - I\|_F^2 = \sum_{k=1}^{n} \|(AM - I)e_k\|_2^2$$  \hspace{1cm} (2)

finding such $M$ separates into $n$ independent least squares problems

$$\min_{m_k} \|Am_k - e_k\|_2, \quad k = 1, \ldots, n,$$  \hspace{1cm} (3)

where $e_k$ has 1 in the $k$-th position and 0 elsewhere. If $M$ is sparse, Eq. (3) becomes $n$ small least squares problems, which can be solved quickly.

The following is a description of the SPAI method as in [4].

Let $\tau$ be a given parameter controlling how much the approximate inverse is to be close to the actual inverse.

(a) Choose an initial sparsity pattern $J$, where $m_k(j) \neq 0, \quad j \in J$.

(b) Compute the row indices $I$ such that $A(i, J)$ is not identically zero, $i \in I$. Let $\hat{A} = A(I, J), \quad \hat{e}_k = e_k(I)$. Find the QR decomposition of $\hat{A}$,

$$\hat{A} = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$$  \hspace{1cm} (4)

Then, solve

$$\min_{\hat{m}_k} \|\hat{A}\hat{m}_k - \hat{e}_k\|_2$$  \hspace{1cm} (5)
and the residual \( r = A(\cdot, J)m_k - e_k \).

While \( \|r\|_2 > \tau \):

(c) Set \( K \) equal to \( \{l | r(l) \neq 0\} \).

(d) Set \( \tilde{J} \) equal to the set of all new column indices of \( A \) that appear in all \( K \) rows but not in \( J \).

(e) For each \( j \in \tilde{J} \) find the direction \( u_j \) such that \( \|r + u_je_j\|_2 \) is minimized.

(f) For each \( j \in \tilde{J} \) compute the 2-norm of the new residual \( r + u_je_j \) with the above \( u_j \). Delete from \( \tilde{J} \) all but the indices leading to significant reduction in 2-norm.

(g) Determine the new indices \( \tilde{I} \) and update the QR decomposition accordingly. Then, solve the new least squares problem, compute the new residual \( r = Am_k - e_k \) and set \( I = I \cup \tilde{I}, J = J \cup \tilde{J} \).

For further details see [4].

4 ILU(0) factorization

Meijerink and Van. der Vorst[6] introduced a so-called Incomplete LU(ILU) preconditioner for symmetric matrices. The following is a modification of the original ILU for nonsymmetric matrices, as described in [3]. Let \( A = LU + N \), where \( L_{i,j} = U_{i,j} = 0 \) if \( A_{i,j} = 0 \) and \( N_{i,j} = 0 \) if \( A_{i,j} \neq 0 \). Let \( NZ(A) \), the nonzero pattern of \( A \), denote the set of pairs of \( (i,j) \) for which \( A_{i,j} \), the \((i,j)\) entry of \( A \), is nonzero.

Algorithm 4.1 ILU Factorization

1. For \( i = 1, \ldots, \) Until \( N \) Do
2. For \( j = 1, \ldots, \) Until \( N \) Do
3. If( \( (i,j) \) belongs to \( NZ(A) \) ) then
4. \( s_{i,j} = A_{i,j} - \sum_{t=1}^{\min(i,j)-1} L_{i,t}U_{t,j} \).
5. if(\( i \geq j \) ) then \( L_{i,j} = s_{i,j} \).
6. if(\( i < j \) ) then \( U_{i,j} = s_{i,j}/L_{i,i} \).
7. Endif
8. Endfor
9. Endfor

4.1 Point-SSOR algorithm

Algorithm 4.2 Point-SSOR Let \( A = D - E - F \), where \( D \) is the diagonal part, \( -E \), is the lowertriangular part and \( -F \) the uppertriangular part.

1. Choose \( x_0 \).
2. For $i=0, \ldots$ Do

\[
(D - \omega E)x_{i+\frac{1}{2}} = ((1 - \omega)D + \omega F)x_i + \omega b
\]

\[
(D - \omega F)x_{i+1} = ((1 - \omega)D + \omega E)x_{i+\frac{1}{2}} + \omega ba
\]

Endfor

5 Experiments

5.1 Test problems

- **Problem 1** Elman’s problem [3]

\[
-(bu_x)_x - (cu_y)_y + (du)_x + du_x + (eu)_y + eu_y + fu = g
\]

\[
\Omega = (0, 1) \times (0, 1)
\]

\[
u = 0 \text{ on } \partial \Omega
\]

where $b = \exp(-xy)$, $c = \exp(xy)$, $d = \beta(x + y)$,

$e = \gamma(x + y)$, $f = \frac{1}{(1+xy)}$,

and $g$ is such that exact solution $u = x \exp(xy) \sin(\pi x) \sin(\pi y)$

- **Problem 2** Convection-diffusion equation

\[
-\epsilon \Delta u + \cos \alpha u_x + \sin \alpha u_y = f,
\]

\[
\Omega = (0, 1) \times (0, 1)
\]

\[
u = 0 \text{ on } \partial \Omega
\]

- **Problem 3** Cylindrical shell problem from Harwell-Boeing Collection [1] The 's3dkq4m2.dat' from the CYLSPHE set.

We have set $\epsilon = 0.0001$, and $alpha = 15$ degree. Upwind scheme was adopted for the convection term. 5-point approximation was used for 2nd order derivatives in Problem 1, and 9-point Laplacian for Problem 2. For problem 3 $N = 90449$ and the total number of nonzeros is equal to 4820891. This matrix is very ill-conditioned with an estimated condition number of $1.35^{11}$. 
5.2 Domain mapping onto the processors

In this paper we assume that the domain is a square, which is further divided into \( p \) rectangle-shaped blocks, where \( p \) is the number of the available processors. Further we assume that there is a one-to-one correspondence between the \( p \) blocks and \( p \) processors. See Fig. (2).

5.3 Results

BICGSTAB\cite{17} was used as the outer iterative method. As for the manipulation and storing of the matrices the CSR(Compress Sparse Row) format\cite{13} was adopted, and the MSR(Modified Sparse Row) format for the ILU(0) factorization and forward/backward solves. Since CSR format was used, our experiments realistically simulates the unstructured problems even though we used the unit square as our domain. We used the wavefront-ordering for the point-SSOR method and the parameter \( \omega \) was set to 1.2. The \( \tau \) parameter for the SPAI, which controls how much the approximate inverse is going to be close the actual inverse was set to 0.4.

We set \( \gamma = 50 \), \( \beta = 1 \), \( \epsilon = 0.1 \), \( \alpha = 15 \), so that the resulting matrices for Problem 1 and 2 are nonsymmetric. For the multi-coloring we used a simple heuristic based on a greedy algorithm as in \cite{14}. Using this heuristic the number of colors required for Problem 1 and 2 are 2 and 4, respectively.

For the Multi-Color Block SOR method we used the MA48 package to invert the diagonal block. For the partitioning of the 's3dkq4m2' matrix of the shell problem we have used the Metis 4.0 library by V. Kumar.

MPI(Message Passing Interface) library was used for the communications. This enables the codes to be run independent of the machines.
The CPU time and the number of iterations are shown in the Tables 2 - 10. ‘X’ stands for the case where the memory was insufficient for the problem size. As we compare the number of iterations in wavefront order and that in multi-coloring order, we see little difference, unlike in the symmetric case, such as in the ILU(0)-preconditioned CG method. Hence, we are led to believe that for problems tested the multi-coloring order outperforms the wavefront order, which is verified in the tables. The SPAI preconditioner is hardly competitive. In all cases ILU(0) with the multi-coloring order outperforms the other preconditioners, except the Multi-Color Block SOR. Since the cost of inverting and back-solving by MA48 is approximately proportional to \((N/p)^2\), for a given \(N\) increasing \(p\) reduces the cost quadratically. Hence, we observe that with for a given \(N\) there ia a limit on \(p\) such that above this limit the Multi-Color Block SOR preconditioner outperforms the ILU(0) with the Multi-Color ordering. The CPU time does not include the preprocessing costs of generating the preconditioner matrices. For the 's3dkq4m2' matrix from the shell problem only the Multi-Color Block SOR converges. That matrix is veri ill-conditioned, with the condition number of \(10^{-11}\).
Table 2: Problem 1, with FDM, N=128x128

<table>
<thead>
<tr>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU time/Iterations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>2.41/6</td>
<td>0.71/6</td>
<td>0.33/6</td>
<td>0.18/6</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>1.42/62</td>
<td>1.04/61</td>
<td>1.07/61</td>
<td>1.36/63</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>0.99/50</td>
<td>0.80/53</td>
<td>0.90/54</td>
<td>1.07/50</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>0.95/50</td>
<td>0.70/53</td>
<td>0.66/54</td>
<td>0.69/51</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>2.82/64</td>
<td>1.69/62</td>
<td>1.23/64</td>
<td>1.32/62</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>4.28/144</td>
<td>2.41/142</td>
<td>1.78/137</td>
<td>1.50/135</td>
</tr>
</tbody>
</table>

Table 3: Problem 1 with FDM, N=256x256

<table>
<thead>
<tr>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU time/Iterations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>18.3/7</td>
<td>6.14/7</td>
<td>2.48/9</td>
<td>0.97/9</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>11.22/126</td>
<td>6.46/122</td>
<td>4.92/132</td>
<td>4.04/119</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>7.04/107</td>
<td>4.80/107</td>
<td>3.53/107</td>
<td>3.21/104</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>7.90/107</td>
<td>4.47/108</td>
<td>2.86/107</td>
<td>2.11/101</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>21.96/137</td>
<td>12.48/171</td>
<td>6.90/136</td>
<td>4.80/137</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>29.16/276</td>
<td>16.71/296</td>
<td>9.10/281</td>
<td>5.30/264</td>
</tr>
</tbody>
</table>

Table 4: Problem 1 with FDM, N=512x512

<table>
<thead>
<tr>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU time/Iterations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>X</td>
<td>X</td>
<td>21.0/12</td>
<td>7.6/12</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>X</td>
<td>43.01/258</td>
<td>24.75/256</td>
<td>17.63/257</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>32.05/234</td>
<td>18.91/234</td>
<td>13.12/224</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>30.42/230</td>
<td>17.20/238</td>
<td>9.89/227</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>214.24/330</td>
<td>95.58/322</td>
<td>53.38/312</td>
<td>30.02/302</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>244.74/571</td>
<td>117.24/538</td>
<td>66.16/572</td>
<td>52.89/562</td>
</tr>
</tbody>
</table>

Table 5: Problem 1 with FDM, N=1024x1024

<table>
<thead>
<tr>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU time/Iterations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>68.4/17</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>106.48/534</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>80.16/482</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>67.09/467</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>298.56/640</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>316.43/1373</td>
</tr>
</tbody>
</table>
### Table 6: Problem 2 with FDM, N=128x128

<table>
<thead>
<tr>
<th></th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cpu time/Iterations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>4.46/3</td>
<td>1.31/3</td>
<td>0.55/3</td>
<td>0.4/5</td>
<td>0.34/5</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>1.93/65</td>
<td>1.54/68</td>
<td>1.68/66</td>
<td>1.96/66</td>
<td>3.09/66</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>0.93/37</td>
<td>0.81/41</td>
<td>0.97/41</td>
<td>1.14/41</td>
<td>1.94/43</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>0.95/38</td>
<td>0.78/42</td>
<td>0.78/41</td>
<td>0.93/44</td>
<td>1.41/46</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>3.57/68</td>
<td>2.04/68</td>
<td>1.41/68</td>
<td>1.55/68</td>
<td>2.22/68</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>5.40/194</td>
<td>3.51/194</td>
<td>2.00/201</td>
<td>2.29/211</td>
<td>3.62/204</td>
</tr>
</tbody>
</table>

### Table 7: Problem 2 with FDM, N=256x256

<table>
<thead>
<tr>
<th></th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cpu time/Iterations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>39.5/3</td>
<td>12.4/3</td>
<td>4.37/3</td>
<td>1.55/4</td>
<td>0.73/4</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>14.63/127</td>
<td>7.86/123</td>
<td>6.26/126</td>
<td>5.95/128</td>
<td>7.54/129</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>7.08/73</td>
<td>4.31/80</td>
<td>3.35/76</td>
<td>3.49/83</td>
<td>4.51/82</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>7.11/73</td>
<td>4.12/81</td>
<td>2.87/79</td>
<td>2.39/78</td>
<td>8.22/155</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>28.08/141</td>
<td>14.82/141</td>
<td>8.24/141</td>
<td>5.13/141</td>
<td>4.91/141</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>41.86/388</td>
<td>24.09/426</td>
<td>11.28/376</td>
<td>7.08/389</td>
<td>7.87/411</td>
</tr>
</tbody>
</table>

### Table 8: Problem 2 with FDM, N=512x512

<table>
<thead>
<tr>
<th></th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cpu time/Iterations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>X</td>
<td>X</td>
<td>39.0/3</td>
<td>13.2/4</td>
<td>4.8/4</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>X</td>
<td>52.87/237</td>
<td>31.42/232</td>
<td>22.98/239</td>
<td>22.02/244</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>27.00/148</td>
<td>16.93/148</td>
<td>12.41/151</td>
<td>12.19/150</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>25.95/147</td>
<td>14.25/141</td>
<td>9.68/150</td>
<td>8.21/155</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>226.10/282</td>
<td>115.00/285</td>
<td>59.02/281</td>
<td>32.58/285</td>
<td>20.23/279</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>345.17/815</td>
<td>172.96/816</td>
<td>86.13/784</td>
<td>44.00/749</td>
<td>30.28/779</td>
</tr>
</tbody>
</table>

### Table 9: Problem 2 with FDM, N=1024x1024

<table>
<thead>
<tr>
<th></th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cpu time/Iterations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-BSOR(2)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>SLOW</td>
<td>42.4/5</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>128.16/471</td>
<td>88.64/459</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>61.40/270</td>
<td>46.19/279</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>55.69/285</td>
<td>34.92/289</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.2$</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>341.54/535</td>
<td>159.33/535</td>
</tr>
<tr>
<td>SPAI, $\tau = 0.4$</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>442.01/2000</td>
<td>202.47/1724</td>
</tr>
</tbody>
</table>
Table 10: 's3dkq4m2' from the Harwell-Boeing Collection

<table>
<thead>
<tr>
<th>Method</th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC-BSOR(2)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1286.0/524</td>
<td>457.0/497</td>
</tr>
<tr>
<td>SSOR-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>SL</td>
<td>SL</td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>SL</td>
<td>SL</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>SL</td>
<td>SL</td>
</tr>
<tr>
<td>SPAI, (\tau = 0.2)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>SL</td>
<td>SL</td>
</tr>
<tr>
<td>SPAI, (\tau = 0.4)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>SL</td>
<td>SL</td>
</tr>
</tbody>
</table>

Table 11: Problem 1 with FDM, \(N=256\times256\)

<table>
<thead>
<tr>
<th>Method</th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>66</td>
<td>109</td>
<td>149</td>
<td>159</td>
<td>125</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>66</td>
<td>119</td>
<td>184</td>
<td>234</td>
<td>221</td>
</tr>
</tbody>
</table>

Table 12: Problem 1 with FDM, \(N=512\times512\)

<table>
<thead>
<tr>
<th>Method</th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>144</td>
<td>243</td>
<td>336</td>
<td>366</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>149</td>
<td>272</td>
<td>451</td>
<td>595</td>
</tr>
</tbody>
</table>

Table 13: Problem 1 with FDM, \(N=1024\times1024\)

<table>
<thead>
<tr>
<th>Method</th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>473</td>
<td>683</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>547</td>
<td>943</td>
</tr>
</tbody>
</table>
Table 14: Problem 2 with FDM, N=256x256

<table>
<thead>
<tr>
<th></th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megaflops</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>78</td>
<td>140</td>
<td>171</td>
<td>179</td>
<td>137</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>77</td>
<td>148</td>
<td>207</td>
<td>246</td>
<td>586</td>
</tr>
</tbody>
</table>

Table 15: Problem 2 with FDM, N=512x512

<table>
<thead>
<tr>
<th></th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megaflops</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>165</td>
<td>263</td>
<td>366</td>
<td>371</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>171</td>
<td>298</td>
<td>467</td>
<td>568</td>
</tr>
</tbody>
</table>

Table 16: Problem 2 with FDM, N=1024x1024

<table>
<thead>
<tr>
<th></th>
<th>p = 4</th>
<th>p = 8</th>
<th>p = 16</th>
<th>p = 32</th>
<th>p = 64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megaflops</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILU(0)-WAVEFRONT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>530</td>
<td>728</td>
</tr>
<tr>
<td>ILU(0)-MULTICOL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>617</td>
<td>997</td>
</tr>
</tbody>
</table>
Figure 3: Speedup of problem 1, N= 512x512
6 Summary

- Unlike in the symmetric case the convergence rate of the BICGSTAB method does not deteriorate at all, compared with the wavefront-order. Hence, due to the parallelism of order(N) coming from the multi-coloring, for problems tested ILU(0) with the multicolor ordering outperforms the other preconditioners considered in this paper. The Multi-Color Block SOR gives the worst performances.

- Seeing from the speedup curves, the communication overheads in the CRAY-T3E turns out to be very high for the irregularly structured sparse matrices in the CSR format. One of the reasons is that we are sorting the adjacent processor list array to avoid deadlocks. For example, in matrix-vector product this causes the higher numbered processors to wait until all of the lower numbered processors are done.

- It is an interesting fact on its own that the convergence rate of the BICGSTAB method in the multi-coloring order does not deteriorate, compared to that of the natural ordering. In future we might need more research to explain this phenomenon.

Acknowledgements. The work was supported by Korea STEPI research fund, 97-NF-03-01-A-03. The author would like to acknowledge the numerous advices of Prof. Youcef Saad, and also the Korea ETRI, which provided the computer facilities and an excellent research environment to conduct this research.

References


Department of Computer science
Hanyang University, Sa-1 Dong 1271
Ansan, Kyungki-Do, S. Korea, 425-791
email :sangback@cse.hanyang.ac.kr