CHEBYSHEV PSEUDOSPECTRAL-FINITE ELEMENT METHOD FOR TWO-DIMENSIONAL UNSTEADY NAVIER-STOKES EQUATION

Guo Ben-yu *and Hou Jing-yu

Abstract A Chebyshev pseudospectral-finite element method is proposed for two-dimensional unsteady Navier-Stokes equation. The generalized stability and the convergence are proved strictly. The numerical results show the advantages of this method.

Keywords. Navier-Stokes equation, Chebyshev pseudospectral-finite element method.


1 Introduction.

Spectral and pseudospectral methods have the high accuracy. In particular, pseudospectral method is easier to be performed. But in most practical problems, the domains are not rectangular. This fact limits their applications. However, the sections of domains might be rectangular in certain directions, such as a cylindrical container. For solving such problems, it is natural to use Chebyshev pseudospectral-finite element approximation, see [1]. In this paper, we develop a mixed Chebyshev pseudospectral-finite element method for two-dimensional unsteady Navier-Stokes equation. It is easy to generalize this method to three-dimensional problems with complex geometry. In particular, it is easy to deal with the nonlinear terms. We also follow the idea in [2, 3] to calculate the pressure based on a Poisson equation. Therefore we avoid the difficult job of choosing the trial function space in which the divergence of every element vanishes. We construct the scheme and present the numerical results in Section 2 and 3. The numerical results show the advantages of this method. We list some lemmas in Section 4, and then prove the generalized stability and the convergence in the last two sections.

2 The Scheme.

Let $I_x = \{x / -1 < x < 1\}$, $I_y = \{y / 0 < y < 1\}$ and $\Omega = I_x \times I_y$. The speed and the pressure are denoted by $U = (U_1, U_2)$ and $P$, respectively. $\nu > 0$ is the kinetic viscosity.

*Dept. of Mathematics, POSTECH, Korea

†Lab. of Computers, Yunnan Univ. of Industry, China.
$U_0$ and $f$ are given functions. Let $T > 0$ and \( \partial_z = \frac{\partial}{\partial z} \), $z = t, x, y$. We consider the problem

\[
\begin{cases}
\partial_t U + d(U, U) + \nabla P - \nu \nabla^2 U = f, & \text{in } \Omega \times (0, T], \\
\nabla^2 P + \Phi(U) = \nabla \cdot f, & \text{in } \Omega \times (0, T], \\
U|_{t=0} = U_0, & \text{in } \Omega \cup \partial \Omega,
\end{cases}
\]

where

\[
d(U, V) = \partial_x(V_1 U) + \partial_y(V_2 U), \quad \Phi(U) = 2(\partial_y U_1 \partial_x U_2 - \partial_x U_1 \partial_y U_2).
\]

(2.1) is one of representations of Navier-Stokes equation. Suppose that the boundary is a non-slip wall and so $U = 0$ on $\partial \Omega$. There is no boundary condition for the pressure. But if we use the second formula of (2.1) to evaluate the pressure, we need a non-standard boundary condition. Sometimes, it is assumed approximately that $\frac{\partial P}{\partial n} = g(x)$ on $\partial \Omega$, see [2, 3]. For simplicity, let $g(x) = 0$ in the following discussions. In addition, for fixing the value of pressure, we require that for all $t \leq T$,

\[
\int \int_{\Omega} P(x, y, t)dxdy = 0.
\]

Since the derivation of (2.1) implies the incompressibility, we avoid the difficult job of constructing the trial function space in which the divergence of every element vanishes.

Let $D$ be an interval (or a domain) in $\mathbb{R}$ (or $\mathbb{R}^2$). $L^2(D)$, $H^r(D)$ and $H^0_0(D)$ $(r > 0)$ denote the usual Hilbert spaces with the usual inner products and norms. We also define

\[
L^2_0(D) = \{u \in L^2(D) \mid \int_D udD = 0\}.
\]

Let $\omega(x) = (1 - x^2)^{-\frac{3}{2}}$ and

\[
(u, v)_{\omega, I_x} = \int_{I_x} uv\omega dx, \quad \|u\|_{\omega, I_x}^2 = (u, u)_{\omega, I_x},
\]

\[
L^2_\omega(I_x) = \{u(x) \mid u \text{ is measurable on } I_x \text{ and } \|u\|_{\omega, I_x} < \infty\}.
\]

Furthermore

\[
(u, v)_{\omega} = \int \int_{\Omega} uv\omega dxdy, \quad \|u\|_{\omega}^2 = (u, u)_{\omega},
\]

\[
L^2_\omega(\Omega) = \{u(x, y) \mid u \text{ is measurable on } \Omega \text{ and } \|u\|_{\omega} < \infty\}.
\]

Let $a_\omega(u, v) = (\nabla u, \nabla (v\omega))_{L^2(\Omega)}$. The weak formulation of (2.1) is to find $U \in X^2(\Omega)$ and $P \in Y(\Omega)$ for all $t \leq T$ such that

\[
\begin{cases}
(\partial_t U + d(U, U) + \nabla P, v)_\omega + a_\omega(U, v) = (f, v)_\omega, & \forall v \in X^2(\Omega), \\
a_\omega(P, w) = (\Phi(U) - \nabla \cdot f, w)_\omega, & \forall w \in X(\Omega)
\end{cases}
\]

(2.2)
where
\[ \tilde{X}(\Omega) = \{ u / u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \in L^2_0(\Omega) \}, \]
\[ X(\Omega) = \tilde{X}(\Omega) \cap \{ u / u = 0 \text{ on } \partial \Omega \}, \]
\[ Y(\Omega) = \tilde{X}(\Omega) \cap L^2_0(\Omega) \cap \{ u / \frac{\partial u}{\partial x} = 0 \text{ for } |x| = 1 \}. \]

We now construct the trial function spaces. For the Chebyshev approximation, let \( N \) be any positive integer and \( \mathcal{P}_N(I_x) \) denote the set of all algebraic polynomials of degree \( \leq N \), defined on \( I_x \). Moreover
\[ V_N(I_x) = \{ u(x) \in \mathcal{P}_N(I_x) / u(-1) = u(1) = 0 \}, \]
\[ W_N(I_x) = \{ u(x) \in \mathcal{P}_N(I_x) / \frac{du}{dx}(-1) = \frac{du}{dx}(1) = 0 \}. \]

For the finite element approximation, let \( \tau_h \) be a class of regular decomposition of \( I_y \) with subintervals \( I_l = (y_{l-1}, y_l), 1 \leq l \leq M \) where \( 0 = y_0 < y_1 < \cdots < y_M = 1 \). Suppose that \( \tau_h \) satisfies the inverse assumption. Let \( h = \max_{1 \leq l \leq M} |y_l - y_{l-1}|, \bar{h} = \min_{1 \leq l \leq M} |y_l - y_{l-1}|. \) Then \( h/\bar{h} \leq d, d \) being a positive constant independent of \( h \). Moreover let \( k \geq 1 \) and
\[ \tilde{S}^k_h(I_y) = \{ v(y) / v|_{I_l} \in \mathcal{P}_k(I_l), 1 \leq l \leq M \}, \quad S^k_h(I_y) = \tilde{S}^k_h(I_y) \cap H^1(I_y). \]

Then we take the spaces \( X^k_{N,h}(\Omega) \) and \( Y^k_{N,h}(\Omega) \) as the trial function spaces for the speed and the pressure respectively, defined as
\[ X^k_{N,h}(\Omega) = V_N(I_x) \otimes \tilde{S}^k_h(I_y), \quad Y^k_{N,h}(\Omega) = [W_N(I_x) \otimes (\tilde{S}^k_h(I_y) \cap H^1(I_y))] \cap L^2_0(\Omega). \]

In addition,
\[ \tilde{X}^k_{N,h}(\Omega) = V_N(I_x) \otimes (\tilde{S}^k_h(I_y) \cap H^1(I_y)). \]

Next, let \( x^{(j)} \) and \( \omega^{(j)} \) be the nodes and weights of Gauss-Lobatto integration. The corresponding discrete inner products and norms are defined as
\[ (u,v)_{N,\omega} = \sum_{j=0}^{N} u(x^{(j)})v(x^{(j)})\omega^{(j)}, \quad \|u\|^2_{N,\omega} = (u,u)_{N,\omega}, \]
\[ (u,v)_{N,h,\omega} = \int_{I_y} (u,v)_{N,\omega} dy, \quad \|u\|^2_{N,h,\omega} = (u,u)_{N,h,\omega}. \]

Clearly, if \( uv \in \mathcal{P}_{2N-1}(I_x) \otimes L^2(I_y) \), then \( (u,v)_{N,h,\omega} = (u,v)_{\omega} \). Let \( P : C(I_x) \rightarrow \mathcal{P}_N(I_x) \) be the interpolation, i.e., \( P_j u(x^{(j)}) = u(x^{(j)}) \), for \( 0 \leq j \leq N \). Let \( \Pi^k_h : C(I_y) \rightarrow \tilde{S}^k_h(I_y) \cap H^1(I_y) \) be the piecewise Lagrange interpolation of degree \( k \) over each \( I_l \). Furthermore, let \( P_{N,h} : L^2_0(\Omega) \rightarrow X^k_{N,h}(\Omega) \) be the \( L^2_0(\Omega) \)-orthogonal projection.

Let \( \tau \) be the step size of time \( t \), \( \tilde{R}_t = \{ t = l\tau / 1 \leq l \leq \lceil T \rceil \} \) and \( R_t = \tilde{R}_t \cup \{0\} \). Set
\[ u(t) = \frac{1}{\tau}(u(t+\tau) - u(t)). \]
For approximating the terms in (2.2), we define

\[
d_c(u,v) = \partial_x P_1(v_1u) + \partial_y P_1(v_2u), \quad a_{N,h,\omega}(u,v) = -(\partial^2_x u, v)_{N,h,\omega} + (\partial_y u, \partial_y v)_{N,h,\omega},
\]
\[
\Phi_c(u) = 2[P_1(\partial_y u_1 \partial_y u_2) - P_1(\partial_x u_1 \partial_y u_2)].
\]

Let \( \lambda \geq 0, 0 \leq \sigma \leq 1, u \) and \( p \) denote the approximations to \( U \) and \( P \) respectively. The Chebyshev pseudospectral-finite element scheme for (2.2) is to find \( u \in (X_{N,h}^{k+\lambda}(\Omega))^2 \) and \( p \in Y_{N,h}^k(\Omega) \) such that for all \( t \in R_\tau, \)

\[
\begin{cases}
(u_t + d_c(u,u) + \nabla p, v)_{N,h,\omega} + \nu a_{N,h,\omega}(u + \sigma u_t, v) = (f,v)_{N,h,\omega}, \quad \forall v \in (X_{N,h}^{k+\lambda}(\Omega))^2, \\
a_{N,h,\omega}(p,w) = (\Phi_c(u) - \nabla \cdot f, w)_{N,h,\omega}, \quad \forall w \in X_{N,h}^k(\Omega), \\
u(0) = P_{N,h}U_0.
\end{cases}
\]  

(2.3)

3 Numerical Results.

We take the test functions

\[
U_1(x,y,t) = Ae^{Bt}((x^2 - 1)^2y(y - 1)(2y - 1),
\]
\[
U_2(x,y,t) = -2Ae^{Bt}(x^3 - x)y^2(y - 1)^2,
\]
\[
P(x,y,t) = 4Ae^{Bt}(x^3 - 3x)(2y^3 - 3y^2 + 0.5).
\]

We use scheme (2.3), in which the interval \( I_y \) is divided with the mesh size \( h_y = 1/M \).

For comparison, we also consider the finite element scheme (FEM). In this case, the domain is divided uniformly into rectangular subdomains with the length \( h_x = 2/N^t \) in \( x \)-direction and \( h_y = 1/M \) in \( y \)-direction, \( U \) is approximated by quadratic finite element and \( P \) by linear finite element. For describing the errors, let

\[
I_x = \{x_j / x_j = \cos \frac{j\pi}{N}, \ 1 \leq j \leq N - 1\}, \quad \text{for (2.3)},
\]
\[
I_x = \{x_j / x_j = -1 + jh_x, \ 1 \leq j \leq N^t - 1\}, \quad \text{for FEM},
\]
\[
I_y = \{y_j / y_j = jh_y, \ 1 \leq j \leq M - 1\}, \quad \text{for (2.3) and FEM}
\]

and

\[
E(U(t)) = \left( \frac{\sum_{q=1}^2 \sum_{x \in I_x} \sum_{y \in I_y} |U_q(x,y,t) - u_q(x,y,t)|^2}{\sum_{q=1}^2 \sum_{x \in I_x} \sum_{y \in I_y} |U_q(x,y,t)|^2} \right)^{1/2}.
\]
The error $E(P(t))$ is defined similarly. In calculations, $\nu = 0.0001$, $A = 0.2$, $B = 0.1$ and $k = 1$. We first take $N = 4$, $M = N^* = 10$, $\tau = 0.005$ and $\sigma = 0$. The numerical results are shown in the Table I. Clearly, scheme (2.3) gives better results than scheme FEM. We also take $N = 10$, $M = 10$, $\tau = 0.001$, $\sigma = 0$ and $\lambda = 1$ in scheme (2.3). The corresponding results are shown in Table II. We find that when $N$ increases and $\tau$ decreases, the better results follow. It shows the convergence of scheme (2.3).

<table>
<thead>
<tr>
<th>Scheme (2.3), $\lambda = 1$</th>
<th>Scheme FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>$E(U(t))$</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9167E-04</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1806E-03</td>
</tr>
<tr>
<td>1.5</td>
<td>0.2699E-03</td>
</tr>
<tr>
<td>2.0</td>
<td>0.3598E-03</td>
</tr>
<tr>
<td>2.5</td>
<td>0.4507E-03</td>
</tr>
</tbody>
</table>

| Table II. The errors of scheme (2.3). |
|-----------------------------|-------------|
| $t$ | $E(U(t))$ | $E(P(t))$ |
| 0.5 | 0.5144E-04 | 0.2061E-03 |
| 1.0 | 0.7531E-04 | 0.2120E-03 |
| 1.5 | 0.9928E-04 | 0.2173E-03 |
| 2.0 | 0.1234E-03 | 0.2230E-03 |
| 2.5 | 0.1478E-03 | 0.2288E-03 |

4 Some Lemmas.

We first introduce some Sobolev spaces with the weight $\omega(x)$. For any integer $r \geq 0$, set

$$
|u|_{r, \omega, L_x} = \| \partial_{\omega}^r u \|_{\omega, L_x}, \quad \|u\|_{r, \omega, L_x} = \left( \sum_{m=0}^r |u|_{m, \omega, L_x}^2 \right)^{1/2},
$$

$$
H_{\omega}^r(I_x) = \{ u(x) / \|u\|_{r, \omega, L_x} < \infty \}.
$$

For real $r > 0$, the space $H_{\omega}^r(I_x)$ is defined by the space interpolation. Furthermore,

$$
C^\infty_0(I_x) = \{ u(x) / u \text{ is infinitely differentiable and has a compact support in } I_x \}.
$$
Denote by $H^0_0(I_x)$ the closure of $C^\infty_0(I_x)$ in $H^r_0(I_x)$. Besides, $L^\infty(I_x)$ is the space of essentially bounded functions with the norm $\| \cdot \|_{\infty, L}$.

Next, let $B$ be a Banach space with the norm $\| \cdot \|_B$. Define

$$ \begin{align*}
L^2(D, B) &= \{u(z) : D \to B / u \text{ is strongly measurable and } \|u\|_{L^2(D, B)} < \infty\}, \\
C(D, B) &= \{u(z) : D \to B / u \text{ is strongly continuous and } \|u\|_{C(D, B)} < \infty\}
\end{align*} $$

where

$$ \|u\|_{L^2(D, B)} = \left( \int_D \|u(z)\|^2_B \, dz \right)^{1/2}, \quad \|u\|_{C(D, B)} = \max_{z \in D} \|u(z)\|_B. $$

For any integer $s \geq 0$, define

$$ H^s(D, B) = \{u(z) / \|u\|_{H^s(D, B)} < \infty\} $$

equipped with the semi-norm and norm

$$ \|u\|_{H^s(D, B)} = \|D^s u\|_{L^2(D, B)}, \quad \|u\|_{H^s(D, B)} = \left( \sum_{m=0}^{s} \|D^m u\|^2_{H^m(D, B)} \right)^{1/2}. $$

For real $s \geq 0$, the space $H^s(D, B)$ is defined by the space interpolation.

We now introduce the non-isotropic space

$$ H_0^{r, s}(\Omega) = L^2(I_y, H^r_0(I_x)) \bigcap H^s(I_y, L^2_0(I_x)), \quad r, s \geq 0 $$

with the norm

$$ \|u\|^2_{H_0^{r, s}(\Omega)} = \|u\|^2_{L^2(I_y, H^r_0(I_x))} + \|u\|^2_{H^s(I_y, L^2_0(I_x))}. $$

Also let

$$ \begin{align*}
M_0^{r, s}(\Omega) &= H_0^{r, s}(\Omega) \bigcap H^{s-1}(I_y, H^1_0(I_x)), \quad r, s \geq 1, \\
A_0^{r, s}(\Omega) &= H^r(I_y, H^s_0(I_x)) \bigcap H^s(I_y, H^{s+1}_0(I_x)) \bigcap H^{s+1}(I_y, H^r_0(I_x)), \quad r, s \geq 0, \\
Y_0^{r, s, \delta}(\Omega) &= H^r(I_y, H^s_0(I_x)) \bigcap H^{s+\delta}(I_y, H^{s+\delta}_0(I_x)) \bigcap H^{s+\delta}(I_y, H^{s+\delta}_0(I_x)), \quad r, s \geq 0, \quad \delta > 0, \\
Y_1^{r, s, \delta}(\Omega) &= H^{r+1}(I_y, H^s_0(I_x)) \bigcap H^1(I_y, H^r_0(I_x)) \bigcap H^{s+\delta}(I_y, H^{r+1}_0(I_x)) \bigcap H^{s+\delta}(I_y, H^{s+\delta}_0(I_x)) \bigcap H^{s+\delta}(I_y, H^{s+\delta}_0(I_x)) \bigcap H^{s+\delta}(I_y, H^{s+\delta}_0(I_x)), \quad r, s \geq 0, \quad \delta > 0.
\end{align*} $$

Their norms are defined similarly. Furthermore let $H_0^{r, s}(\Omega)$ be the closure of $C^\infty_0(\Omega)$ in $H_0^{r, s}(\Omega)$. If $r = s$, then $H_0^{r, s}(\Omega) = H^r_0(\Omega)$, and denote their semi-norm and norm by $\| \cdot \|_{r, \omega}$ and $\| \cdot \|_{r, \omega}$ respectively, etc.. Denote by $L^\infty(I_x)$, $L^\infty(\Omega)$ and $W^{1, \infty}(\Omega)$ the usual Sobolev spaces with the norms $\| \cdot \|_{\infty, I_x}$, $\| \cdot \|_{\infty}$ and $\| \cdot \|_{1, \infty}$, etc..
For simplicity of statements, let $\bar{s} = \min(s, k+1)$ and denote by $c$ a generic positive constant independent of $h$, $N$, $\tau$ and any function. In some lemmas, we require that there exist a suitably big positive constant $c_1$ and a positive constant $c_2$ such that

$$c_1 h^{-\frac{s}{2}} \leq N \leq c_2 h^{-\frac{s}{2}}.$$  \hspace{1cm} (4.1)

**Lemma 1.** If $u(x, y, t) \in C(\bar{I}_x) \times L^2(I_y)$ for all $t \in R_r$, then

$$2(u(t), u_t(t))_{N, h, \omega} = (\|u(t)\|_{N, h, \omega})^2 - \tau \|u_t(t)\|_{N, h, \omega}^2.$$

Let $P_N$ denote the Chebyshev truncated operator. We have that (see [4])

$$\|u - P_N u\|_{L^2(I_x)} \leq c\sigma_{N,q} N^{-m} \|u\|_{W^{2,s}(L_r)}, \quad m \geq 0, \quad 1 \leq q \leq \infty$$  \hspace{1cm} (4.2)

where $\sigma_{N,q} = 1 + \ln N$ for $q = 1$ or $q = \infty$, and $\sigma_{N,q} = 1$, otherwise.

**Lemma 2** (Lemma 1 of [1]). If $u \in C(\bar{I}_x) \times L^2(I_y)$ and $v \in P_N(\bar{I}_x) \times L^2(I_y)$, then

$$\|v\|_{\omega} \leq \|v\|_{N, h, \omega} \leq \sqrt{2} \|v\|_{\omega},$$

$$|(u, v)_{N, h, \omega} - (u, v)_{\omega}| \leq c(\|u - P_{N-1} u\|_{\omega} + \|u - P_N u\|_{\omega}) \|v\|_{\omega}.$$

**Lemma 3** (Lemma 2 of [1]). For any $u \in X^k_{N, h}(\Omega)$ with $k \geq 1$,

$$a_{N, h, \omega}(u, u) \geq \frac{1}{4} \|u\|_{1, \omega}^2.$$

**Lemma 4** (Lemma 5 of [5]). For any $u \in X^k_{N, h}(\Omega)$ with $k \geq 1$,

$$\|u\|_{\infty} \leq c(N/h)^{1/2} \|u\|_{\omega}.$$  

Moreover, for any $u \in H^1_\omega(\Omega)$,

$$\|P_N u\|_{\infty} \leq c(\ln N)^{1/2} \|u\|_{1, \omega}.$$  

**Lemma 5** (Lemma 2 of [5]). Let $u \in H^r_{0, \omega}(\Omega) \cap H^{r+s}_\omega(\Omega)$ with $0 \leq r \leq 1$, $s \geq 0$, or $u \in H^1_{0, \omega}(\Omega) \cap H^{r+s}_\omega(\Omega)$ with $r > 1$, $s \geq 0$. Then

$$\|u - P_N u\|_{\omega} \leq c(N-r^2 + h^s) \|u\|_{H^{r+s}_\omega(\Omega)}.$$  

**Lemma 6** (Lemma 5 of [1]). If $u \in H^{\beta}(I_y, H^\alpha(I_x))$ with $\beta \geq 0$, $r > 1/2$ and $0 \leq \alpha \leq r$, then

$$\|u - P_N u\|_{H^\beta(I_y, H^\alpha(I_x))} \leq cN^{2\alpha-r} \|u\|_{H^\beta(I_y, H^\alpha(I_x))}.$$  

We now introduce the projection $P^*_{N, h} : H^1_{0, \omega}(\Omega) \to X^k_{N, h}(\Omega)$ such that for any $u \in H^1_{0, \omega}(\Omega)$,

$$a_{\omega}(P^*_{N, h} u, v) = a_{\omega}(u, v), \quad \forall v \in X^k_{N, h}(\Omega),$$
and the projection $P_{N,h}^{k} : H^{1}_{0,\omega}(\Omega) \to X^{k}_{N,h}(\Omega)$ such that for any $u \in H^{1}_{0,\omega}(\Omega)$,

$$a_{N,h,\omega}(P_{N,h}^{k}u, v) = a_{\omega}(u, v), \quad \forall v \in X^{k}_{N,h}(\Omega).$$

**Lemma 7.** Let (4.1) hold and $k \geq 1$. If $u \in H^{1}_{0,\omega}(\Omega) \cap M_{r}^{s}(\Omega)$ with $r, s \geq 1$, then

$$\|u - P_{N,h}^{k}u\|_{1,\omega} \leq c(N^{1-r} + h^{r-1})\|u\|_{M_{r}^{s}(\Omega)},$$

$$\|u - P_{N,h}^{k}u\|_{1,\omega} \leq c(N^{1-r} + h^{r-1})\|u\|_{M_{r}^{s}(\Omega)}.$$  

If in addition $u \in M_{r}^{s+rac{1}{4},s}(\Omega)$, then

$$\|u - P_{N,h}^{k}u\|_{\omega} \leq c(N^{-r} + h^{\frac{s}{2}})\|u\|_{M_{r}^{s+rac{1}{4},s}(\Omega)},$$

$$\|u - P_{N,h}^{k}u\|_{\omega} \leq c(N^{-r} + h^{\frac{s}{2}})\|u\|_{M_{r}^{s+rac{1}{4},s}(\Omega)}.$$  

**Proof.** The first and the third conclusions come from Lemma 4 of [6]. Next, as in the proof of Lemma 7 of [1], we have

$$\|u - P_{N,h}^{k}u\|_{1,\omega} \leq c(N^{1-r} + h^{r-1})\|u\|_{M_{r}^{s}(\Omega)}.$$  

Finally, following the same line as in the proof of Lemma 7 of [1], we get

$$\|u - P_{N,h}^{k}u\|_{\omega} \leq c(N^{-1} + h)(N^{\frac{s}{4}} + h^{\frac{s}{2}-1})\|u\|_{M_{r}^{s+rac{1}{4},s}(\Omega)}.$$  

**Lemma 8.** Let (4.1) hold. If $u \in H^{1}_{0,\omega}(\Omega) \cap H^{\beta}(H^{\alpha}_{0,\omega}(\Omega) \cap H^{\alpha}(I_{y}, H^{\beta}_{0}(I_{x})))$ with $\alpha, \beta > 1/2$, then

$$\|P_{N,h}^{k}u\|_{\infty} \leq c\|u\|_{M_{r}^{s+rac{1}{4},s}(\Omega) \cap H^{\beta}(H^{\alpha}_{0,\omega}(\Omega) \cap H^{\alpha}(I_{y}, H^{\beta}_{0}(I_{x})))}.$$  

If, in addition, $u \in A_{\alpha,\beta}^{0,\beta}(\Omega)$ with $\alpha, \beta > 1/2$, then

$$\|P_{N,h}^{k}u\|_{1,\infty} \leq c\|u\|_{A_{\alpha,\beta}^{0,\beta}(\Omega)}.$$  

**Proof.** We have

$$\|P_{N,h}^{k}u\|_{\infty} \leq \|P_{N,h}^{k}u - \Pi_{h}^{k}P_{N}u\|_{\infty} + \|\Pi_{h}^{k}P_{N}u\|_{\infty}.$$  

By Lemma 4, Lemma 7, (4.2) and Theorem 3.2.1 of [7],

$$\|P_{N,h}^{k}u - \Pi_{h}^{k}P_{N}u\|_{\infty} \leq c\|u\|_{M_{r}^{s+rac{1}{4},s}(\Omega)}.$$  

By Lemma 5 of [6] and Theorem 3.1.5 of [7],

$$\|\Pi_{h}^{k}P_{N}u\|_{\infty} \leq c\|P_{N}u\|_{H^{\beta}(H^{\alpha}_{0,\omega}(\Omega) \cap H^{\alpha}(I_{y}, H^{\beta}_{0}(I_{x})))} \leq c\|u\|_{H^{\beta}(I_{y}, H^{\alpha}(I_{x}))}.$$
Next, we prove the second conclusion. Let $P_h^1 : H_0^1(I_y) \rightarrow S_h^1(I_y)$ be such that for any $u \in H_0^1(I_y)$,
\[
(\partial_y(P_h^1 u - u), \partial_y v)_{L^2(I_y)} = 0, \quad \forall v \in S_h^1(I_y).
\]
Then we have (see [8])
\[
\|u - P_h^1 u\|_{\mu, I_y} \leq c h^{1-\mu} |u|_{r, I_y}, \quad s \geq 1, \quad 0 \leq \mu \leq 1.
\]
Also let $P_N^1 : H_0^1(I_x) \rightarrow V_N(I_x)$ be such that for any $u \in H_0^1(I_x)$,
\[
(\partial_x(P_N^1 u - u), \partial_x v)_{L^2(I_x)} = 0, \quad \forall v \in V_N(I_x).
\]
We have (see (9.5.17) of [4]),
\[
\|u - P_N^1 u\|_{\mu, I_x} \leq c N^{1-\mu} \|u\|_{H_0^1(I_x)}, \quad 0 \leq \mu \leq 1, \quad r \geq \mu.
\]
Now, we begin to estimate $\|P_{N,h}^1 u\|_{1, \infty}$. In fact,
\[
|P_{N,h}^1 u|_{1, \infty} \leq |P_{N,h}^1 u - P_N^1 u|_{1, \infty} + |P_N^1 u|_{1, \infty}.
\]
By Lemma 4,
\[
|P_{N,h}^1 u - P_N^1 u|_{1, \infty} \leq c \sqrt{\frac{N}{h}} |P_{N,h}^1 u - P_N^1 u|_{1, \infty}.
\]
Let $\vartheta$ be the identity operator. By Lemma 3 and the definitions of $P_{N,h}^1$ and $P_N^1$,
\[
\|P_{N,h}^1 u - P_N^1 u\|_{1, \infty} \leq 4a_{N,h,\omega}(P_{N,h}^1 u - P_N^1 u, \vartheta(P_{N,h}^1 u - P_N^1 u))
\]
\[
\leq 4 \|((\vartheta - P_{N-1}) \partial_y P_{N,h}^1 u, \vartheta(P_{N,h}^1 u - P_N^1 u))_{\infty}.
\]
Therefore
\[
\|P_{N,h}^1 u - P_N^1 u\|_{1, \infty} \leq c \|((\vartheta - P_{N-1}) \partial_y P_{N,h}^1 u, \vartheta(P_{N,h}^1 u - P_N^1 u))_{\omega} + c \|((\vartheta - P_{N-1}) \partial_y P_h^1 P_N^1 u)_{\omega}.
\]
Furthermore
\[
\|((\vartheta - P_{N-1}) \partial_y P_{N,h}^1 u - \partial_y P_h^1 P_N^1 u)_{\infty} \leq c \|P_h^1 P_N^1 u - P_N^1 u\|_{1, \infty}.
\]
Let $W = P_h^1 P_N^1 u - P_N^1 u$. We have from Lemma 3 and the definition of $P_{N,h}^1$ that
\[
\|P_{N,h}^1 u - P_h^1 P_N^1 u\|_{1, \infty}^2 \leq 4a_{\omega}(P_{N,h}^1 u - P_h^1 P_N^1 u, W)_{\omega} + 4(\partial_y(u - P_h^1 P_N^1 u), \partial_y W)_{L^2(\Omega)}
\]
\[
= 4(\partial_y(u - P_h^1 P_N^1 u), \partial_y W)_{\omega} + 4(\partial_x(u - P_h^1 P_N^1 u), \partial_x(W)_{L^2(\Omega)}.
\]
Thus
\[ \| P^*_{N,h}u - P^1_h P_{N}^1 u \|_{1,\omega} \leq c(N^{-r} + h^{\frac{1}{2}}) \| u \|_{H^r(I_p, H^s(L_r))} \bigcap H^s(I_p, H^{s-\frac{1}{2}}(L_r)) \].
(4.7)

Similarly
\[ \| (\theta - P_{N-1}) \partial_y P^1_h P_{N}^1 u \|_{\omega} \leq cN^{-r} \| u \|_{H^r(I_p, H^s(L_r))}. \]
(4.8)

Thus we have from (4.5)–(4.8) that
\[ \| P^*_{N,h}u - P^1_h P_{N}^1 u \|_{1,\omega} \leq c(N^{-r} + h^{\frac{1}{2}}) \| u \|_{H^r(I_p, H^s(L_r))} \bigcap H^s(I_p, H^{s-\frac{1}{2}}(L_r)) \]
and so (4.4) implies that
\[ |P^*_{N,h}u - P^1_h P_{N}^1 u|_{1,\infty} \leq c \| u \|_{H^r(I_p, H^s(L_r))} \bigcap H^s(I_p, H^{s-\frac{1}{2}}(L_r)). \]
(4.9)

Now, we turn to estimate \( |P^*_{N,h}u|_{1,\infty} \). Clearly
\[ |P^*_{N,h}u|_{1,\infty} \leq c \sqrt{\frac{N}{h}} |P^*_{N,h}u - P^1_h P_{N}^1 u|_{1,\omega} + |P^1_h P_{N}^1 u|_{1,\infty}. \]
(4.10)

From (4.7),
\[ \sqrt{\frac{N}{h}} |P^*_{N,h}u - P^1_h P_{N}^1 u|_{1,\omega} \leq c \| u \|_{H^r(I_p, H^s(L_r))} \bigcap H^s(I_p, H^{s-\frac{1}{2}}(L_r)). \]
(4.11)

On the other hand,
\[ |P^1_h P_{N}^1 u|_{1,\infty} \leq \| \partial_y P^1_h P_{N}^1 u\|_{\infty} + \| \partial_x P^1_h P_{N}^1 u\|_{\infty}. \]

Furthermore
\[ \| \partial_y P^1_h P_{N}^1 u\|_{\infty} \leq \| \partial_y P^1_h P_{N}^1 u - \partial_y P^1_h P_{N}^1 u\|_{\infty} + \| \partial_y P^1_h P_{N}^1 u - \partial_y \Pi^1_h P_{N}^1 u\|_{\infty} + \| \partial_y \Pi^1_h P_{N}^1 u\|_{\infty} \]
\[ \leq c \| u \|_{H^r(I_p, H^s(L_r))} \bigcap H^{s+\frac{1}{2}}(I_p, H^{s-\frac{1}{2}}(L_r))}. \]
(4.12)

Similarly
\[ \| \partial_x P^1_h P_{N}^1 u\|_{\infty} \leq c \| u \|_{H^r(I_p, H^s(L_r))} \bigcap H^s(I_p, H^{s+\frac{1}{2}}(L_r))}. \]
(4.13)

and so
\[ |P^1_h P_{N}^1 u|_{1,\infty} \leq c \| u \|_{H^r(I_p, H^s(L_r))} \bigcap H^s(I_p, H^{s+\frac{1}{2}}(L_r)) \bigcap H^{s+\frac{1}{2}}(I_p, H^{s-\frac{1}{2}}(L_r)) \bigcap H^s(I_p, H^{s+\frac{1}{2}}(L_r))}. \]
(4.14)

Therefore we have from (4.10), (4.11) and (4.14) that \( |P^*_{N,h}u|_{1,\infty} \leq c \| u \|_{A^{r,s}_{\omega}}(\Omega) \).

**Lemma 9** (Lemma 9 of [1]). If \( u, v \in Y^r, s, \omega \) with \( r, s \geq 0 \) and \( \delta > 0 \), then
\[ \| uv \|_{H^r_{\omega} \cap H^s_{\omega}} \leq c \| u \|_{Y^r, s, \omega}(\Omega) \| v \|_{Y^r, s, \omega}(\Omega). \]
**Lemma 10** (Lemma 10 of [1]). If \( u \in L^2_\omega(\Omega) \) and \( v \in H^1_0(\Omega) \), then
\[
|\langle u, \partial_x(v \omega) \rangle|_{L^2_\omega(\Omega)} \leq 2\|u\|_{\omega} \|v\|_{\omega}.
\]

**Lemma 11** (Lemma 8 of [6]). There exists a positive constant \( c_d \) depending only on the value of \( d \), such that for all \( u \in \mathcal{P}_N(I_x) \otimes (H^1(I_y) \cap S_{\delta_k}(I_y)) \),
\[
|u|^2_{\omega} \leq (2N^4 + c_d h^{-2}) \|u\|_{\omega}^2.
\]

**Lemma 12.** Let (4.1) hold. \( u \in W_N(I_x) \times (S_{\delta_k}(I_y) \cap H^1(I_y)) \) and \( g \in \mathcal{P}_N(I_x) \times (S_{\delta_k}(I_y) \cap H^1(I_y)) \) satisfy the following equation
\[
a_{N,h,\omega}(u,v) = (g,v)_{N,h,\omega}, \quad \forall v \in \tilde{X}_{N,h}^k(\Omega).
\]
Then we have
\[
\|u\|_{1,\omega} \leq c\|g\|_{\omega}.
\]

**Proof.** Let \( \eta = (1 - x^2)^{1/2} \). Define the spaces \( L^2_\eta(I_x) \) with the norm \( \|u\|_{\eta,I_x} \) and \( H^1_\eta(I_x) \) with the norm \( \|u\|_{\eta,I_x} \) in the same way as \( L^2(I_x) \) and \( H^1(I_x) \), etc. From Lemma 3.1 of [9], for any \( \xi \in L^2_\eta(I_x) \) and \( \lambda > 0 \), there exists a unique function \( W \in H^1_\eta(I_x) \) such that
\[
\begin{cases}
LW = -\frac{\partial^2 W}{\partial x^2} + \lambda W = \xi, & \text{in } I_x, \\
\frac{\partial W}{\partial x}(-1) = \frac{\partial W}{\partial x}(1) = 0.
\end{cases}
\]
(4.16)

Let \( H^{-1}(I_x) \) be the dual space of \( H^1(I_x) \). We have from (4.16) that
\[
\|W\|_{H^{-1}(I_x)}^2 + \lambda \|W\|_{L^2(I_x)}^2 \leq c\|\xi\|_{H^{-1}(I_x)}^2.
\]
Furthermore, if \( \xi \in L^2(I_x) \), then by multiplying (4.16) by \( \frac{\partial W}{\partial x} \) and integrating by parts,
\[
\|W\|_{H^1(I_x)}^2 + \lambda \|W\|_{H^1(I_x)}^2 \leq c\|\xi\|_{L^2(I_x)}^2.
\]
By the space interpolation, if \( \xi \in H^{-4}(I_x) \) with \( 0 \leq s \leq 1 \), then \( W \in H^{2-s}(I_x) \) and
\[
\|W\|_{H^{2-s}(I_x)}^2 + \lambda \|W\|_{H^{2-s}(I_x)}^2 \leq c\|\xi\|_{H^{-4}(I_x)}^2.
\]
(4.17)

Since for any real \( s \geq 0 \), \( H^s(I_x) \subset H^{1+s/4}(I_x) \) (see Theorem 4.2 of [9]), we get
\[
\|W\|_{L^2(I_x)}^2 + \lambda \|W\|_{L^2(I_x)}^2 \leq c\|\xi\|_{H^{-1/4}(I_x)}^2 \leq c\|\xi\|_{\eta,I_x}^2.
\]
(4.18)

Moreover for any real \( s \geq 1/4 \), \( H^s(I_x) \subset H^{1/4}(I_x) \) (see Theorem 4.1 of [9]), and thus
\[
\|W\|_{L^2(I_x)}^2 + \lambda \|W\|_{L^2(I_x)}^2 \leq c\|\xi\|_{\eta,I_x}^2 = c\|LW\|_{\eta,I_x}^2.
\]
(4.19)
Next, consider an auxiliary problem. It is to find \( \lambda \in \mathcal{R} \) and \( \phi \in \tilde{S}_h^k(I_y) \cap H^1(I_y) \), such that
\[
\int_0^1 \frac{d\phi}{dx} dy = \lambda \int_0^1 \phi dy, \quad \forall \phi \in \tilde{S}_h^k(I_y) \cap H^1(I_y).
\] (4.20)

Then there exists a normalized \( L^2(I_y) \)-orthogonal eigenfunction system \( \{ \phi_l(y) \} \), \( l = 0, 1, 2, \cdots, M_h \). The corresponding eigenvalues are ranged as
\[
0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{M_h}.
\]

Let \( \phi_0(y) \equiv 1 \). By (4.20), we know that
\[
\int_0^1 \frac{d\phi_k}{dy} \frac{d\phi_l}{dy} dy = \lambda_k \delta_{kl}, \quad k, l = 0, 1, 2, \cdots, M_h.
\] (4.21)

Let \( \{ \lambda^{(l)} \} \) be the eigenvalues of the corresponding continuous problem, that is
\[
\begin{cases}
-\frac{d^2u}{dy^2} = \lambda u, \\
\frac{du}{dy}(0) = \frac{du}{dy}(1) = 0.
\end{cases}
\]

Then \( \lambda^{(l)} = \pi^2 l^2, \ l = 0, 1, \cdots \). We have that for sufficiently small \( h \) and certain positive constant \( \delta \) (see [2]),
\[
\lambda^{(l)} \leq \lambda_l \leq \lambda^{(l)} + 2\delta k^{2k} [\lambda^{(l)}]^{k+1}.
\] (4.22)

Now, we turn to prove the lemma. Since \( u \in W_N(I_x) \times (\tilde{S}_h^k(I_y) \cap H^1(I_y)) \) and \( g \in \mathcal{P}_N(I_x) \times (\tilde{S}_h^k(I_y) \cap H^1(I_y)) \), we put
\[
u = \sum_{l=0}^{M_h} u_l(x) \phi_l(y), \quad \varphi = \sum_{l=0}^{M_h} g_l(x) \phi_l(y)
\]
where \( u, g \in W_N(I_x) \), \( g \in \mathcal{P}_N(I_x) \). From (4.15),
\[
-(\sum_{l=0}^{M_h} \frac{d^2u_l}{dx^2} \phi_l(y), v)_{\omega} + (\sum_{l=0}^{M_h} u_l(x) \partial_x \phi_l(y), \partial_x v)_{N,h,\omega} = (\sum_{l=0}^{M_h} g_l(x) \phi_l(y), v)_{N,h,\omega}, \quad \forall \nu \in \tilde{X}_h^k(\Omega).
\] (4.23)

We first let \( l \neq 0 \), \( v = (1-x^2)z_l(x) \phi_l(y) \) and \( z_l(x) \in \mathcal{P}_{N-2}(I_x) \). Then
\[
-(\frac{d^2u_l}{dx^2}, z_l)_{\eta,L_x} + \lambda_l (u_l, (1-x^2)z_l)_{N,\omega} = (g_l, (1-x^2)z_l)_{N,\omega}
\]
and so
\[
(Lu, z_l)_{\eta,L_x} \equiv (\frac{d^2u_l}{dx^2} + \lambda_l u_l, z_l)_{\eta,L_x} = (g_l, z_l)_{\eta,L_x} + E_1(z_l) + E_2(z_l)
\] (4.24)
with

\[ E_1(z_l) = \lambda_l(u_l, (1 - x^2)z_l)_x - \lambda_l(u_l, (1 - x^2)z_l)_N, \]
\[ E_2(z_l) = (g_l, (1 - x^2)z_l)_N - (g_l, (1 - x^2)z_l)_x. \]

Let \( \tilde{P}_N \) be the \( L^2(I_x) \)-orthogonal projection, and \( z_l = \tilde{P}_{N-2}u_l \). Then we have

\[ z_l = Lu_l - \lambda_l(u_l - \tilde{P}_{N-2}u_l) \in \mathcal{P}_{N-2}(I_x). \]

So (4.24) reads

\[ \|Lu_l\|_{\mathcal{P}_{N-2}(I_x)}^2 - \lambda_l(Lu_l u_l - \tilde{P}_{N-2}u_l)_{\mathcal{P}_{N-2}(I_x)} = (g_l, Lu_l)_{\mathcal{P}_{N-2}(I_x)} - \lambda_l(g_l, u_l - \tilde{P}_{N-2}u_l)_{\mathcal{P}_{N-2}(I_x)} + E_1(P_{N-2}Lu_l) + E_2(P_{N-2}Lu_l). \]  

(4.25)

Furthermore, since \( (1 - x^2)z_l \in \mathcal{P}_N(I_x) \), we get from (9.3.5) of [4] that

\[ |E_1(P_{N-2}Lu_l)| \leq c\lambda_l u_l - \tilde{P}_{N-1}u_l \|_{\mathcal{P}_{N-2}(I_x)}(\|Lu_l\|_{\mathcal{P}_{N-2}(I_x)} + \lambda_l u_l - \tilde{P}_{N-2}u_l), \]
\[ |E_2(P_{N-2}Lu_l)| \leq c\|g_l\|_{\mathcal{P}_{N-2}(I_x)}(\|Lu_l\|_{\mathcal{P}_{N-2}(I_x)} + \lambda_l u_l - \tilde{P}_{N-2}u_l). \]

By substituting the above estimations into (4.25), we obtain

\[ \frac{1}{2}\|Lu_l\|_{\mathcal{P}_{N-2}(I_x)}^2 \leq c\|g_l\|_{\mathcal{P}_{N-2}(I_x)}^2 + c\lambda_l\|u_l - \tilde{P}_{N-1}u_l\|_{\mathcal{P}_{N-2}(I_x)} + \|u_l - \tilde{P}_{N-2}u_l\|_{\mathcal{P}_{N-2}(I_x)}^2. \]  

(4.26)

We know from (4.19) that

\[ \|u_l\|_{\mathcal{P}_N(I_x)}^2 + \lambda_l\|u_l\|_{\mathcal{P}_N(I_x)}^2 \leq c\|Lu_l\|_{\mathcal{P}_{N-2}(I_x)}^2. \]

Since

\[ \|u_l - \tilde{P}_{N-1}u_l\|_{\mathcal{P}_{N-2}(I_x)} \leq cN^{-r}\|u_l\|_{\mathcal{P}_N(I_x)}, \quad \text{for } r \geq 0, \]

we obtain from (4.26) that

\[ \|u_l\|_{\mathcal{P}_N(I_x)}^2 + \lambda_l\|u_l\|_{\mathcal{P}_N(I_x)}^2 \leq c\|g_l\|_{\mathcal{P}_N(I_x)}^2 + c\lambda_l^2\|u_l\|_{\mathcal{P}_N(I_x)}^2. \]  

(4.27)

We know from (4.22) that

\[ \lambda_l \leq \lambda_{M_0} \leq \lambda_{(M_b)}^{2h^2}[\lambda_{(M_b)}]^{k+1}, \quad \lambda_{(M_b)} = (M_b)^2\pi^2. \]

Since \( M_b = O(\frac{1}{h}) \), we have \( \lambda_l \leq ch^{-2} \). Thus

\[ 1 - c\lambda_l^2N^{-3} \geq 1 - c\lambda_{M_b}^2N^{-3} \geq 1 - ch^{-4}N^{-3}. \]

Thanks to condition (4.1), we have \( 1 - ch^{-4}N^{-3} \geq \alpha > 0 \). Hence (4.27) implies that

\[ \|u_l\|_{\mathcal{P}_N(I_x)}^2 + \lambda_l\|u_l\|_{\mathcal{P}_N(I_x)}^2 \leq c\|g_l\|_{\mathcal{P}_N(I_x)}^2 \]

and so for \( l \neq 0 \),

\[ \|u_l\|_{\mathcal{P}_N(I_x)}^2 + \lambda_l\|u_l\|_{\mathcal{P}_N(I_x)}^2 \leq c\|g_l\|_{\mathcal{P}_N(I_x)}^2. \]  

(4.28)
Next, we consider the case with \( l = 0 \). By taking 
\[ v = -(1 - x^2) \partial_y^2 u_0 \phi_0(y) = -(1 - x^2) \partial_y^2 u_0 \] in (4.23), we get 
\[ \| \partial_y^2 u_0 \|_{\Omega, L^2} = - (g_0, (1 - x^2) \partial_y^2 u_0)_{\Omega, L^2} \leq c \| g_0 \|_{\Omega, L^2} \| \partial_y^2 u_0 \|_{\Omega, L^2}. \]
Hence \( \| \partial_y^2 u_0 \|_{\Omega, L^2} \leq c \| g_0 \|_{\Omega, L^2} \). On the other hand, since \( u_0 \in W_N(I_x) \), we have 
\[ |u_0|^2 \leq \int_{I_x} \omega(x) (\int_{I_x} \eta(x) (\partial_y^2 u_0)^2 dx) (\int_{I_x} \omega(x) dx) dx \leq c \| \partial_y^2 u_0 \|_{\Omega, L^2}^2 \]
and so 
\[ |u_0|^2 \leq c \| g_0 \|_{\Omega, L^2}^2. \] (4.29)
Finally, we have from (4.28) and (4.29) that 
\[ \| u \|^2_{\Omega, L^2} = \sum_{l=0}^{M_N} (|u_l|^2_{\Omega, L^2} + \gamma_l \| u_l \|^2_{\Omega, L^2}) \leq c \| g \|_{\Omega, L^2}^2. \]

5 The Generalized Stability.

Assume that \( u(0) \) and \( f(t) \) have the errors \( \tilde{u}(0) \) and \( \tilde{f}(t) \) which induce the errors of \( u(t) \) and \( p(t) \), denoted by \( \tilde{u}(t) \) and \( \tilde{p}(t) \) respectively. Then 
\[ \begin{cases} 
(\tilde{u}_t, v)_{N, h} + (d_v(\tilde{u}, u + \tilde{u}) + d_v(u, \tilde{u}) + \nabla \tilde{p}, v)_{N, h} \\
\quad + \nu a_N(h, \tilde{u}, v) = (\tilde{f}, v)_{N, h}, \quad \forall v \in (X^{k+1})(\Omega)^2, \\
\quad a_N(h, \tilde{p}, w) = (\tilde{\Phi}(\tilde{u}) + \tilde{\Phi}_t(u, \tilde{u}) - \nabla \cdot \tilde{f}, w)_{N, h}, \quad \forall w \in \tilde{X}_N(h, \Omega) 
\end{cases} \]
where 
\[ \tilde{\Phi}_t(u, \tilde{u}) = 2[A_0(\partial_y u_1 \partial_x \tilde{u}_2 + P_1(\partial_y \tilde{u}_1 \partial_x u_2) - P_1(\partial_x u_1 \partial_y \tilde{u}_2) - P_1(\partial_x \tilde{u}_1 \partial_y u_2)]. \]

Let \( \varepsilon > 0 \) and \( m \) be an undetermined positive constant. By taking \( v = 2\tilde{u}(t) + m\tilde{u}_t(t) \) in the first formula of (5.1), we have from Lemma 1, Lemma 2 and Lemma 3 that 
\[ (|\tilde{u}|^2_{N, h})_t + \tau (m - 1 - \varepsilon)|\tilde{u}|^2_{N, h} + \frac{\nu m}{2} |\tilde{u}|^2_{N, h} + \frac{\nu m^2}{4} |\tilde{u}|^2_{N, h} + \nu (\tau + m) \]
\[ ((|\partial_x \tilde{u}|^2_{N, h})_t - \tau |\partial_x \tilde{u}|^2_{N, h})_t - \tau |\partial_x \tilde{u}|^2_{N, h} - \tau |\partial_y \tilde{u}|^2_{N, h} + A + B + \sum_{j=1}^{6} F_j \]
\[ \leq 2|\tilde{u}|^2_{N, h} + (2 + \frac{\tau m^2}{2}) \| P_1 \tilde{f} \|_{\Omega, L^2}^2 \]
where 
\[ A = 2\nu \tau (\partial_x \tilde{u}_1, x^2 \partial_y^2 \tilde{u}_2)_{\Omega, h}, \quad B = \nu \tau (\partial_x \tilde{u}_1, x \partial_y^2 \tilde{u}_1)_{\Omega, h}, \]
\[ F_1 = (d_v(\tilde{u}, u), 2\tilde{u} + m\tilde{u}_t)_{N, h}, \quad F_2 = (d_v(u, \tilde{u}), 2\tilde{u} + m\tilde{u}_t)_{N, h}, \]
\[ F_3 = 2(d_v(\tilde{u}, u), \tilde{u})_{N, h}, \quad F_4 = m\tau (d_v(\tilde{u}, u), \tilde{u})_{N, h}, \]
\[ F_5 = 2(\nabla \tilde{p}, \tilde{u})_{N, h}, \quad F_6 = m\tau (\nabla \tilde{p}, \tilde{u})_{N, h}. \]
We have from Lemma 1 of [10] that \( \|v\omega^2\|_{I_x} \leq \|v\|_{I_x} \) for any \( v \in H_0^1(I_x) \), and so
\[
|A| + |B| \leq \frac{\nu}{4}(\sigma + \frac{m}{2})\|\partial_\tau \tilde{u}\|_{\omega}^2 + 4\nu\tau^2(\sigma + \frac{m}{2})\|\partial_\tau \tilde{u}\|_{\omega}^2.
\]
Let \( |u|_{N,M}^2 = \|\partial_x \tilde{u}\|_{N,M}^2 + \|\partial_y \tilde{u}\|_{N,M}^2 \). Then (5.2) leads to
\[
(\|\tilde{u}\|_{N,M}^2) + \tau(m - 1 + \epsilon)\|\tilde{u}\|_{N,M}^2 + \frac{m}{2}(4 - m - 2\sigma)\|\tilde{u}\|_{N,M}^2 + \nu\tau(\sigma + \frac{m}{2})(\|\tilde{u}\|_{N,M}^2) + \frac{\nu\tau m^2}{4}\|\tilde{u}\|_{N,M}^2 - 5\nu\tau^2(\sigma + \frac{m}{2})\|\tilde{u}\|_{N,M}^2 + \sum_{\omega^2} F_j \leq 2\|\tilde{u}\|_{\omega}^2 + (2 + \frac{\tau m^2}{4\epsilon})\|P_c \tilde{f}\|_{\omega}^2.
\]
(5.3)

Now we estimate \( |F_j| \). Similarly, by Lemma 2 and Lemma 10,
\[
|F_1| \leq \frac{\epsilon\nu}{8}\|\tilde{u}\|_{\omega}^2 + \frac{\epsilon\nu\tau m^2}{6}\|\tilde{u}\|_{\omega}^2 + \frac{c(m+1)}{\epsilon\nu}\|u\|_{\omega}^2\|\tilde{u}\|_{\omega}^2.
\]
Similarly
\[
|F_2| \leq \frac{\epsilon\nu}{8}\|\tilde{u}\|_{\omega}^2 + \frac{\epsilon\nu\tau m^2}{6}\|\tilde{u}\|_{\omega}^2 + \frac{c(m+1)}{\epsilon\nu}\|u\|_{\omega}^2\|\tilde{u}\|_{\omega}^2.
\]
By Lemma 4 and Lemma 10,
\[
|F_j| \leq \frac{\epsilon\nu}{8}\|\tilde{u}\|_{\omega}^2 + \frac{cN}{\epsilon\nu}\|u\|_{\omega}^2\|\tilde{u}\|_{\omega}^2.
\]
Similarly
\[
|F_k| \leq \frac{\epsilon\nu\tau m^2}{6}\|\tilde{u}\|_{\omega}^2 + \frac{cN}{\epsilon\nu}\|u\|_{\omega}^2\|\tilde{u}\|_{\omega}^2.
\]
Now, we apply Lemma 12 to the second formula of (5.1) and obtain that
\[
|\tilde{p}|_{\omega} \leq c(\|{}^\alpha \Phi_c(u)\|_{\omega} + \|{}^\alpha \Phi_c^*(u, \tilde{u})\|_{\omega} + \|P_c(\nabla \cdot \tilde{f})\|_{\omega}).
\]
Moreover,
\[
\|{}^\alpha \Phi_c(u)\|_{\omega} \leq c\|u\|_{1,\omega} \|\tilde{u}\|_{1,\omega} \leq c\sqrt{\frac{N}{h}}\|\tilde{u}\|_{1,\omega}^2,
\]
\[
\|{}^\alpha \Phi_c^*(u, \tilde{u})\|_{\omega} \leq c\|u\|_{1,\omega} \|\tilde{u}\|_{1,\omega}.
\]
Thus
\[
|\tilde{p}|_{\omega} \leq c(\|u\|_{1,\omega} \|\tilde{u}\|_{1,\omega} + \sqrt{\frac{N}{h}}\|\tilde{u}\|_{1,\omega}^2 + \|P_c(\nabla \cdot \tilde{f})\|_{\omega})
\]
and
\[
|P_5| \leq \frac{\epsilon\nu}{8}\|\tilde{u}\|_{\omega}^2 + c(1 + \frac{1}{\epsilon\nu}\|u\|_{1,\omega})\|\tilde{u}\|_{\omega}^2 + \frac{cN}{\epsilon\nu}\|u\|_{\omega}^2\|\tilde{u}\|_{\omega}^2 + c\|P_c(\nabla \cdot \tilde{f})\|_{\omega}^2.
\]
By Lemma 11,
\[
|F_6| \leq \tau\|\tilde{u}\|_{\omega}^2 + \frac{c\nu\tau}{4}\|u\|_{1,\omega}^2\|\tilde{u}\|_{\omega}^2 + \frac{c\nu\tau}{4}\|\tilde{u}\|_{\omega}^2 + \frac{c\nu\tau N}{\epsilon\nu}(\|\tilde{u}\|_{\omega}^2\|\tilde{u}\|_{\omega}^2).
\]
Let \( \|u\|_{1,\infty} = \max_{t \in R_\varepsilon} \|u(t)\|_{1,\infty} \), etc. By substituting the above estimations into (5.3), we have from Lemma 2 and Lemma 12 that

\[
(\|\bar{u}\|_{N,h,\omega}^2)_{t} + \tau[m - 2\varepsilon - \nu(5(2\sigma + m) + \frac{\varepsilon m}{2})]\|\bar{u}\|_{\omega}^2
\]

\[
+ \frac{\nu}{8}(\frac{15}{4} - m - 2\sigma - 4\varepsilon)\|\bar{u}\|_{\omega}^2 + \nu(\sigma + \frac{\varepsilon}{2})(\|\bar{u}\|_{N,h,\omega}^2)_{t}
\]

\[
\leq M_1 \|\bar{u}\|_{\omega}^2 + B(\|\bar{u}\|_{\omega})\|\bar{u}\|_{\omega}^2 + G_1.
\]

where

\[
M_1 = c + \frac{c}{\varepsilon}[(1 + m)\|u\|_{\omega}^2 + \|u\|_{1,\infty}^2],
\]

\[
B(\|\bar{u}\|_{\omega}) = -\frac{\nu}{32} + \frac{cm^2}{\varepsilon}\|u\|_{1,\infty}^2
\]

\[
+ \frac{\nu}{4}(1 + m) \ln N + \frac{\varepsilon N}{\varepsilon N}(\frac{1}{\rho} + m^2\tau(2N^4 + c\varepsilon h^{-2}))\|\bar{u}\|_{\omega}^2,
\]

\[
G_1 = (2 + \frac{\varepsilon^2}{2})\|P\bar{f}\|_{\omega}^2 + c(1 + \frac{m^2}{\varepsilon})\|P\nabla \cdot \bar{f}\|_{\omega}^2.
\]

Let \( \varepsilon \) be suitably small and \( \mu \) suitably large. Suppose that

\[
\nu(\sigma + \frac{\varepsilon}{2}) < \frac{1}{\mu^{5 + \frac{\varepsilon}{2} - \frac{\varepsilon}{4}}}. \tag{5.5}
\]

We take

\[
m = \left(\frac{33}{32} + 2\varepsilon + 10\sigma\nu(2N^4 + c\varepsilon h^{-2})(1 - \frac{1}{\mu})^{-1}. \right.
\]

Then the coefficient of the term \( \|\bar{u}\|_{\omega}^2 \) in (5.4) is not less than \( \frac{\varepsilon}{32} \). Moreover, if

\[
\mu > \left(\frac{10\sigma}{5 + \frac{\varepsilon}{2} - \frac{\varepsilon}{4}} + \frac{7}{2} - 2\sigma - 4\varepsilon)(\frac{79}{32} - 2\sigma - 6\varepsilon)\right)^{-1}, \tag{5.6}
\]

then the coefficient of the term \( \|\bar{u}\|_{\omega}^2 \) in (5.4) is not less than \( \frac{\varepsilon}{32} \). Thus (5.4) reads

\[
(\|\bar{u}\|_{N,h,\omega}^2)_{t} + \frac{\varepsilon}{32}\|\bar{u}\|_{\omega}^2 + \frac{\varepsilon}{32}\|\bar{u}\|_{\omega}^2 + \nu(\sigma + \frac{\varepsilon}{2})(\|\bar{u}\|_{N,h,\omega}^2)_{t}
\]

\[
\leq M_1 \|\bar{u}\|_{\omega}^2 + B(\|\bar{u}\|_{\omega})\|\bar{u}\|_{\omega}^2 + G_1.
\]

Let

\[
E(t) = \|\bar{u}(t)\|_{\omega}^2 + \frac{\varepsilon}{32}\sum_{t' \leq t}(\tau\|\bar{u}(t')\|_{\omega}^2 + \nu\|\bar{u}(t')\|_{\omega}^2),
\]

\[
\rho(t) = 2\|\bar{u}(0)\|_{\omega}^2 + \nu(2\sigma + m)|\bar{u}(0)|_{\omega}^2 + \tau\sum_{t' \leq t}G_1(t').
\]

By summing (5.7) for \( t \in R_\varepsilon \), we have

\[
E(t) \leq \rho(t) + \tau \sum_{t' \leq t} (M_1 E(t') + B(E(t'))\|\bar{u}\|_{\omega}^2).
\]
Finally, we use a discrete inequality Lemma 4.16 of [2] to get the following conclusion.

**Theorem 1.** Assume that
(i) (5.5) and (5.6) hold;
(ii) for certain suitably small positive constant $c_3$, $\tau \|u\|_{\infty} < c_3 \nu$;
(iii) there exist positive constants $d_1$ and $d_2$ depending only on $\|u\|_{1, \infty}$ and $\nu$ such that $\rho(t_1)e^{d_1t_1} \leq \frac{d_2}{h}$ for some $t_1 \in R$.

Then for all $t \in R$, $t \leq t_1$, we have
$$E(t) \leq \rho(t)e^{d_1t}.$$

6. The Convergence

Let the $\tilde{P}^1_h$ be the operator such that for any $u \in H^1(I_y)$,
$$
\int_0^1 \partial_y u \partial_y v dy = \int_0^1 \partial_y (\tilde{P}^1_h u) \partial_y v dy, \quad \forall v \in \tilde{S}^1_h(I_y) \cap H^1(I_y),
$$
$$
\int_0^1 (u - \tilde{P}^1_h u) dy = 0.
$$

We have that for any $v \in H^1(I_y)$ with $s \geq 1$,
$$
\|u - \tilde{P}^1_h u\|_{H^p(I_y)} \leq c h^{3-p} \|u\|_{H^p(I_y)}, \quad 0 \leq \mu \leq 1, \mu \leq \bar{s}. \tag{6.1}
$$

Let $U^* = P_{N,h} U$ and $P^* \in P_N(I_x) \times (\tilde{S}^k_h(I_y) \cap H^1(I_y))$ defined by
$$
P^* = \tilde{P}^1_h P(-1, y) + \int_{-1}^1 \tilde{P}^1_h P_{N-1} \frac{\partial P}{\partial s}(s, y) ds.
$$

Set $\tilde{U} = u - U^*$ and $\tilde{P} = p - P^*$. Then by (2.2) and (2.3),
$$
\begin{align*}
&\begin{cases}
(\tilde{U}_t + d_c(\tilde{U}, U^* + \tilde{U}) + d_v(U^*, U) + \nabla \tilde{P}, v)_{N,h,\omega} + \\
\nu a_{N,h,\omega}(\tilde{U} + \sigma \tau \tilde{U}_t, v) = \sum_{j=1}^5 A_j(v), & \forall v \in (X^{k+\lambda}_{N,h}(\Omega))^2, \\
\end{cases} \\
&\begin{cases}
a_{N,h,\omega}(\tilde{P}, w) = (\Phi_c(\tilde{U}) + \Phi_v(U^*, \tilde{U}), w)_{N,h,\omega} + \sum_{j=6}^8 A_j(w), & \forall w \in \tilde{X}^k_{N,h}(\Omega), \\
\tilde{U}(0) = P_{N,h} U_0 - P_{N,h} U_0,
\end{cases}
\end{align*}
$$

where
$$
\begin{align*}
A_1(v) &= (\partial_t U, v) - (U^*, v)_{N,h,\omega}, & A_2(v) &= (d(U, U), v) - (d_c(U^*, U^*), v)_{N,h,\omega}, \\
A_3(v) &= -\nu \sigma \tau a_{N,h,\omega}(U^*, v), & A_4(v) &= (\nabla P, v)_{N,h,\omega} - (\nabla P^*, v)_{N,h,\omega}, \\
A_5(v) &= (f, v)_{N,h,\omega} - (f, v)_{\omega}, & A_6(w) &= a_{\omega}(P, w) - a_{N,h,\omega}(P^*, w), \\
A_7(w) &= -\Phi(U, w)_{\omega} + (\Phi_c(U^*), w)_{N,h,\omega}, & A_8(w) &= (\nabla \cdot f, w)_{N,h,\omega} - (\nabla \cdot f, w)_{N,h,\omega}.
\end{align*}
$$
Now, we turn to estimate $|A_j|$. Let $\bar{s}$ be the same as before and $\bar{s} = \min(k + \lambda + 1, s)$. Take $v = 2\bar{U}$ in $A_j$, $1 \leq j \leq 5$. We obtain from Lemma 2, Lemma 6 and lemma 7 that for $r, s \geq 1$,

$$|A_1(\bar{U})| \leq c(\tau^{1/2} \|U\|_{H^1(t, t + \tau; L^2(\Omega))} + (N^{-r} + h^{\bar{s}}) \|U\|_{M^{t, \delta}_{\alpha}(\omega)} + \|((\theta - P_{N-1})U_t \|_{\omega}) \|\bar{U}\|_{\omega}

\leq \|\bar{U}\|_{\omega}^2 + c(N^{-2r} + h^{2\bar{s}}) \|U\|_{\omega}^2 C_{\alpha}(0, T; M^{t, \delta}_{\alpha}(\omega)) + \frac{c\tau}{\tau} \|U\|_{H^1(t, t + \tau; L^2(\Omega))}^2, \n
$$

$$|A_2(\bar{U})| \leq c(N^{-r} + h^{\bar{s}})(\|U\|_{C_{\alpha}(0, T; M^{t, \delta}_{\alpha}(\omega))} + (\|U\|_{\omega} + \|U_t\|_{\omega} + \|U_{tt}\|_{\omega}) \|\bar{U}\|_{\omega}, \n
$$

$$|A_3(\bar{U})| \leq cN^r \|U\|_{C_{\alpha}(0, T; M^{t, \delta}_{\alpha}(\omega))} \|\bar{U}\|_{\omega}, \n
$$

$$|A_4(\bar{U})| \leq c(N^{-r} + h^{\bar{s}}) \|P\|_{H^1(I_p, H^1(I_p))} \|\bar{U}\|_{\omega}, \n
$$

$$|A_5(\bar{U})| \leq c\|\bar{U}\|_{\omega} + \|((\theta - P_{N-1})f \|_{\omega} \leq cN^{-r} \|f\|_{L^2(I_p, H^1(I_p))} \|\bar{U}\|_{\omega} \n
$$

|A_6(w)| \leq c(N^{-r} + h^{\bar{s}}) \|P\|_{H^1(I_p, H^1(I_p))} \|\bar{U}\|_{\omega} \leq cN^{-r} \|f\|_{L^2(I_p, H^1(I_p))} \|\bar{U}\|_{\omega}, \n
$$|A_7(w)| \leq c(N^{-r} + h^{\bar{s}}) \|U^t\|_1 \|U\|_{C_{\alpha}(0, T; M^{t, \delta}_{\alpha}(\omega))} \|\bar{U}\|_{\omega}, \n
$$|A_8(w)| \leq cN^{-r} \|f\|_{L^2(I_p, H^1(I_p))} \|\bar{U}\|_{\omega}.

Similarly, we take $v = m\bar{T}U_i$ in $A_j(v)$, $1 \leq j \leq 5$, and estimate them, such as

$$m\tau|A_1(\bar{U})| \leq c\tau \|\bar{U}\|_{\omega}^2 + \frac{c\tau^2}{\tau} (N^{-2r} + h^{2\bar{s}}) \|U\|_{\omega}^2 C_{\alpha}(0, T; M^{t, \delta}_{\alpha}(\omega)) + \frac{c\tau^2}{\tau} \|U\|_{H^1(t, t + \tau; L^2(\Omega))}^2.

$$

Moreover, by Lemma 5, Lemma 7 and Lemma 11,

$$\|\bar{U}(0)\|_{\omega}^2 + \tau|\bar{U}(0)|_{\omega}^2 \leq c(1 + 2\tau N^4 + c_d\tau h^{-2})(N^{-2r} + h^{2\bar{s}}) \|U_0\|_{\omega}^2 M^{t, \delta}_{\alpha}(\omega), \quad \text{for } \alpha, \beta > \frac{1}{2}.

$$

Besides, if (5.5) holds and $r > \frac{3}{8}, \alpha \geq \frac{3}{8}, \bar{s} > \frac{13}{6}, \beta > \frac{13}{6}$, then $\tau^2 + N^{-2r} + h^{2\bar{s}} + h^{2(\bar{s}-1)} = o(\frac{1}{\tau^2})$.

Finally, by an argument similar to the proof of Theorem 1, we have the following result.

**Theorem 2.** Assume that

(i) $\lambda \geq 1, (4.1)$ and condition (i) of Theorem 1 hold;

(ii) for $r \geq 1$, $s > \frac{13}{6}$ and $\alpha, \beta > 1/2, U \in C(0, T; M^{t, \delta}_{\alpha, \beta}(\Omega)) \cap H^1(\omega) \cap A^{*}_{\alpha, \beta}(\Omega) \cap Y^{t, \delta}_{1, \omega}(\Omega)

\cap C^1(0, T; M^{t, \delta}_{\alpha, \beta}(\Omega)) \cap H^2(0, T; L^2(\Omega));

(iii) for $r \geq 1$ and $s > \frac{3}{8}, P \in C(0, T; H^{*}(I_y, H^{2}(I_x))) \cap H^2(I_y, H^{2+1}(I_x)) \cap H^1(I_y, H^{2+2}(I_x))$ and $f \in C(0, T; L^2(I_y, H^{2+1}(I_x)) \cap H^1(I_y, H^{2+2}(I_x)))$. 

Then there exists a positive constant $d_3$ depending only on $\nu$ and the norms of $U$ and $P$ in the spaces mentioned above such that for all $t \leq T$,

$$\|U(t) - u(t)\|_\omega \leq d_3(\tau + N^{-\tau} + h^{\tau} + h^{\tau-1}).$$

References


A Modified Adams Predictor-Corrector Method for Differential Equations with Highly Oscillating Solutions

Byung Soo Moon
Korea Atomic Energy Research Institute
P.O. Box 105, Taeduk Science Town, Taejon, Korea 305-600
E-mail: bsmoon@nanum.kaeri.re.kr

Abstract

An algorithm for a solution of ordinary differential equations using a modified corrector in the Adams predictor-corrector method of order four is described. The Lagrange interpolation used in the corrector of the Adams method is replaced partially by the cubic spline interpolation satisfying the first derivative constraints at the two endpoints. By exhibiting three examples, we show that the proposed method is more efficient when the solution of a differential equation is highly oscillating.

1. Introduction

There are many studies on variations of the Adams predictor-corrector method. Some consider the accuracy measures [1] while most others consider the stability problems [2,3,4]. In this paper, we consider a variation of the method which is more efficient for a specific type of problems, i.e., the cases where the solutions are highly oscillating. Given a differential equation \( y' = f(x,y) \) with \( y_0 = y(x_0) \), we may write the solution as

\[
y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(t,y(t)) dt
\]

In evaluating the integral, the Adams-Bashforth formula [5] of order \( k \) at \( x_n \) uses a polynomial \( p_{k,n} \) interpolating the computed derivatives at the \( k \) preceding points, i.e.,

\[
p_{k,n}(x) = \sum_{j=1}^{k} l_j(x) f_{n+1-j} \quad \text{for } j = 1, 2, \cdots, k,
\]

where \( l_j(x) \)'s are the Lagrange polynomials defined on \( x_{n+1-j}, j = 1, 2, \cdots, k. \)

In the following, we consider the case of \( k = 4 \) and assume that the points \( x_n \)'s are equally spaced with \( h = x_n - x_{n-1}. \) The Adams predictor-corrector method of order four used in this paper is of the form [5,p88]

\[
(g_1, g_2, g_3, g_4, g_5) = (1, \frac{1}{2}, \frac{5}{12}, \frac{3}{8}, \frac{251}{720})
\]

AMS classification: 65L05; Adams method; Predictor-Corrector Algorithm; Cubic Splines
(\phi_1^i, \phi_2^i, \phi_3^i, \phi_4^i) = (f_n, f_n - f_{n-1}, f_n - 2f_{n-1} + f_{n-2}, f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3})

p_{n+1} = y_n + h \sum_{i=1}^{4} g_i \phi_i^0

\phi_1^0 = 0, \quad \phi_i^0 = \phi_i^{e1} + \phi_i^i, \quad i = 4, 3, 2, 1

f_{n+1}^0 = f(x_{n+1}, p_{n+1})

y_{n+1} = p_{n+1} + h g_1 (f_{n+1}^0 - \phi_1^0)

f_{n+1} = f(x_{n+1}, y_{n+1})

Our variation is to add one more step at the end of (2) to redefine \( y_{n+1} \) by \( y_n + \int_{x_n}^{x_{n+1}} S_n(t) dt \), where \( S_n(x) \) is the cubic spline interpolation which satisfies

\[ S_n(x_{n-2}) = f'(x_{n-2}, y_{n-2}) = f_x(x_{n-2}, y_{n-2}) + y'(x_{n-2}) f_y(x_{n-2}, y_{n-2}) \]

\[ S_n(x_{n+2-j}) = f_{n+2-j}, \quad j = 1, 2, 3, 4 \]

\[ S_n(x_{n+1}) = f'(x_{n+1}, y_{n+1}) = f_x(x_{n+1}, y_{n+1}) + y'(x_{n+1}) f_y(x_{n+1}, y_{n+1}) \]

Note that the cubic spline function \( S_n(x) \) interpolates not only the function \( f(t, y(t)) \) at the four points \( t_{n-2}, t_{n-1}, t_n, t_{n+1} \) but also its derivatives at \( t_{n-2}, t_{n+1} \).

For convenience, we let \( a = x_n \). Then the spline interpolation function \( S_n(x) \) can be written as

\[ S_n(x) = \sum_{j=1}^{6} c_j B_{a-(4-j)h}(x) \]  

where \( B_{a-(4-j)h}(x) \)'s are the B-spline functions[6] with support \([a-(6-j)h, a-(2-j)h]\). Using properties of B-splines, the five equations in (2) can be written in a matrix form as \( Ac = b \), where \( c = (c_1, c_2, \cdots, c_6)^T \), \( b = (f_{n-2}, f_{n-2}, f_{n-1}, f_n, f_{n+1}, f_{n+1})^T \), and \( A \) is the coefficient matrix

\[
A = \frac{1}{6} \begin{bmatrix}
-3h & 0 & 2h & 0 & 0 & 0 \\
1 & 4 & 1 & 0 & 0 & 0 \\
0 & 1 & 4 & 1 & 0 & 0 \\
0 & 0 & 1 & 4 & 1 & 0 \\
0 & 0 & 0 & -3h & 0 & 3h \\
0 & 0 & 0 & 0 & -3h & 3h
\end{bmatrix}
\]

When the integral of \( B_{a-(4-j)h}(x) \) over the interval \([x_n, x_{n+1}] = [a, a+h]\) is evaluated, we find that for \( j = 3, 4, 5 \),

\[
\int_{a}^{a+h} B_{a-h}(t) dt = \int_{a}^{a+h} B_{a+2h}(t) dt = \frac{h}{24}, \quad \int_{a}^{a+h} B_{a}(t) dt = \int_{a}^{a+h} B_{a+h}(t) dt = \frac{11h}{24}
\]
while the integrals of $B_{n-3}h(x)$ and $B_{n-2}h(x)$ are both zero. Therefore, the integral of the cubic spline interpolation function $S_n(x)$ can be written as

$$\int_a^{a+h} S_n(t) \, dt = \int_a^{a+h} \sum_{j=1}^{6} c_j B_{a-(j-1)}h(t) \, dt = \frac{h}{24}(c_3 + 11c_4 + 11c_5 + c_6) \quad (6)$$

2. A Modified Algorithm

The coefficients $c_j$'s of the spline interpolation function $S_n(x)$ in (6) can be computed by using the inverse of the matrix $A$ in (5) along with the relation $c = A^{-1}b$. When the inverse of $A$ is computed, we have

$$A^{-1} = \frac{1}{45} \begin{bmatrix} -97h & -21 & 84 & -24 & 6 & -2h \\ 26h & 78 & -42 & 12 & -3 & h \\ -7h & -21 & 84 & -24 & 6 & -2h \\ 2h & 6 & -24 & 84 & -21 & 7h \\ -h & -3 & 12 & -42 & 78 & -26h \\ 2h & 6 & -24 & 84 & -21 & 97h \end{bmatrix} \quad (7)$$

and hence the coefficients $c_j$'s of the B-splines become

$$c_3 = \frac{1}{45}(-7hf^t_{n-2} - 21fn_{-2} + 84fn_{-1} - 24fn + 6fn_{n+1} - 2hf^t_{n+1})$$
$$c_4 = \frac{1}{45}(2hf^t_{n-2} + 6fn_{-2} - 24fn_{-1} + 84fn = 21fn_{n+1} + 7hf^t_{n+1})$$
$$c_5 = \frac{1}{45}(-hf^t_{n-2} - 3fn_{-2} + 12fn_{-1} - 42fn + 78fn_{n+1} - 26hf^t_{n+1})$$
$$c_6 = \frac{1}{45}(2hf^t_{n-2} + 6fn_{-2} - 24fn_{-1} + 84fn - 21fn_{n+1} + 97hf^t_{n+1})$$

Substituting these into (6), we obtain

$$\int_a^{a+h} S_n(x) \, dx = \frac{h}{1080}(6hf^t_{n-2} + 18fn_{-2} - 72fn_{-1} + 522fn + 612fn_{n+1} - 114hf^t_{n+1}) \quad (8)$$

3. Examples

The following are some of the computed results using the relation (6) as a corrector for the Adams-Moulton’s method. The results show that when the solutions are highly oscillating, the modified corrector reduces the error substantially even though it does not for other cases.

Example 1. $y' = y + 10e^x \cos(10x), \quad y(0) = 0$.

Note that the analytical solution of the above is $y = e^x \sin(10x)$. We solved the above equation from $x = 0$ to $x = 10$ using both the standard Adams-Moulton’s
method and the modified method. Table 1 shows the maximum and average of the absolute errors for various values of $h$. For $h \geq 0.05$, both the maximum and average errors are reduced to less than one tenth of those from the standard method. In the case of $h < 0.05$, the differences become significant when the integration range gets larger than [0,10].

**Table 1. Comparison of Maximum and Average Errors - Example 1**

<table>
<thead>
<tr>
<th>$h$</th>
<th>0.2</th>
<th>0.1</th>
<th>0.05</th>
<th>0.025</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adams</td>
<td>Mod.</td>
<td>Adams</td>
<td>Mod.</td>
</tr>
<tr>
<td>Average</td>
<td>929.4</td>
<td>65.1</td>
<td>27.52</td>
<td>2.54</td>
</tr>
<tr>
<td>Max.</td>
<td>10224.4</td>
<td>1025.5</td>
<td>311.89</td>
<td>39.63</td>
</tr>
</tbody>
</table>

**Example 2.** $y' = \frac{y}{x} + 2x^2 \cos(x^2)$, $y(0) = 0$, $y'(0) = 0$

It is easy to check that the analytical solution for the above equation is $y = x \sin(x^2)$. When the above equation is solved from $x = 0$ to $x = 10$, 20, 30, respectively by using both the standard Adams-Moulton’s method and the modified method, we obtain the results shown in Table 2.

**Table 2. Comparison of Maximum and Average Errors - Example 2**

<table>
<thead>
<tr>
<th>Range</th>
<th>$x = 0$ to 10</th>
<th>$x = 0$ to 20</th>
<th>$x = 0$ to 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 0.1$</td>
<td>Adams</td>
<td>0.4106</td>
<td>3.731</td>
</tr>
<tr>
<td></td>
<td>Mod.</td>
<td>0.0426</td>
<td>0.348</td>
</tr>
<tr>
<td>$h = 0.025$</td>
<td>Adams</td>
<td>0.00066</td>
<td>0.00614</td>
</tr>
<tr>
<td></td>
<td>Mod.</td>
<td>0.00031</td>
<td>0.00135</td>
</tr>
</tbody>
</table>

**Example 3.** $y' = 2x \cos(x^2)$, $y(0) = 0$.

It is clear that the analytical solution of the above is $y = \sin(x^2)$. The equation is solved from $x = 0$ to $x = 10$, 20, 30, respectively by using both the standard Adams-Moulton’s method and the modified method to obtain the results shown in Table 3.

**Table 3. Comparison of Maximum and Average Errors - Example 3**

<table>
<thead>
<tr>
<th>Range</th>
<th>$x = 0$ to 10</th>
<th>$x = 0$ to 20</th>
<th>$x = 0$ to 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 0.1$</td>
<td>Adams</td>
<td>0.04777</td>
<td>0.3711</td>
</tr>
<tr>
<td></td>
<td>Mod.</td>
<td>0.00509</td>
<td>0.0373</td>
</tr>
<tr>
<td>$h = 0.025$</td>
<td>Adams</td>
<td>0.00006</td>
<td>0.00056</td>
</tr>
<tr>
<td></td>
<td>Mod.</td>
<td>0.00001</td>
<td>0.00009</td>
</tr>
</tbody>
</table>
A MODIFIED ADAMS PREDICTOR-CORRECTOR METHOD

Reference


Linear Operators which Preserve Pairs on which the Rank is Additive

LeRoy B. Beasley

December 14, 1998

Abstract

Let $A$ and $B$ be $m \times n$ matrices. A linear operator $T$ preserves the set of matrices on which the rank is additive if $\text{rank}(A + B) = \text{rank}(A) + \text{rank}(B)$ implies that $\text{rank}(T(A) + T(B)) = \text{rank}(T(A)) + \text{rank}(T(B))$. We characterize the set of all linear operators which preserve the set of pairs of $n \times n$ matrices on which the rank is additive.

1 Introduction.

Let $\mathcal{M}_{m,n}(\mathbb{F})$ denote the set of all $m \times n$ matrices over the field $\mathbb{F}$, and let $T : \mathcal{M}_{m,n}(\mathbb{F}) \rightarrow \mathcal{M}_{m,n}(\mathbb{F})$ be a linear operator. A common problem considered in linear algebra is called a preserver problem, that is, characterize those linear operators which “preserve” a function or a set. We say that $T$ preserves a subset $\mathcal{K}$ of $\mathcal{M}_{m,n}(\mathbb{F})$ whenever $A \in \mathcal{K}$ implies that $T(A) \in \mathcal{K}$; we say that $T$ preserves a function $f : \mathcal{M}_{m,n}(\mathbb{F}) \rightarrow \mathbb{F}$ whenever $f(T(A)) = f(A)$ for all $A \in \mathcal{M}_{m,n}(\mathbb{F})$; and we say that $T$ preserves a subset $\mathcal{P}$ of $\mathcal{M}_{m,n}(\mathbb{F}) \times \mathcal{M}_{m,n}(\mathbb{F})$ whenever $(A,B) \in \mathcal{P}$ implies that $(T(A),T(B)) \in \mathcal{P}$. For example, if $\mathcal{P}$ is the set $\{(A,B) : AB = BA\}$ and $T$ preserves $\mathcal{P}$ then we say that $T$ preserves the set of commuting pairs. The set of all linear operators which preserve the set of commuting pairs was characterized in [1, 8]. In this paper we shall characterize the set of all linear operators which preserve the set of pairs on which the rank is additive.

The classification of preserves began about 100 years ago. In 1897, Frobenius [4] characterized the linear operators on $\mathcal{M}_{n,n}(\mathbb{F})$ which preserve certain matrix functions: those linear operators on $\mathcal{M}_{n,n}(\mathbb{F})$ that preserve the determinant and those that preserve the characteristic polynomial.

After half a century of relative inactivity there was renewed interest in preserver problems. That interest was sparked by the investigation of rank preservers in 1959 by Marcus and Moyls [5]. They proved:

*Department of Mathematics, Utah State University, Logan, Utah 84322-3900, U.S.A. (lbessley@math.usu.edu)
If \( \mathbb{F} \) is a algebraically closed field and of characteristic 0 and \( T \) is a rank preserver, then there exist \( m \times m \) and \( n \times n \) matrices \( U \) and \( V \), respectively, such that either
\[
T(A) = UAV \text{ for all } A \in \mathcal{M}_{m,n}(\mathbb{F})
\]
(1)
or
\[
m = n \text{ and } T(A) = UA^tV \text{ for all } A \in \mathcal{M}_{n,n}(\mathbb{F})
\]
(2)
where \( X^t \) denotes the transpose operator.

Also in 1959, Marcus and Moyls [6] found that \( T \) is a rank preserver if and only if \( T \) is a rank-1 preserver, that is, \( T \) preserves the set of matrices whose rank is 1. In 1967, Westwick [9] generalized these results to matrices over arbitrary algebraically closed fields.

Characterizations of preservers have been appearing regularly over the past thirty years and an excellent summary of nearly all characterizations of linear preservers can be found in a special issue of *Linear and Multilinear Algebra*, edited by S. Pierce with input by leaders in the field [7]. This article includes a list of over 200 articles written on the subject.

2 Additive Rank Preservers.

Let \( \rho : \mathcal{M}_{m,n}(\mathbb{F}) \to \{0, 1, 2, \cdots, \min(m,n)\} \) be the rank function and let \( \mathcal{P} = \{(A, B) : A, B \in \mathcal{M}_{m,n}(\mathbb{F}) \text{ and } \rho(A + B) = \rho(A) + \rho(B)\} \). Let \( L_{UV} \) denote the linear operator of the form (1), and when \( m = n \), let \( L_{UV}^t \) denote the linear operator of the form (2) above. It is obvious that \( L_{UV} \) and \( L_{UV}^t \) (when \( m = n \)) preserve the set \( \mathcal{P} \) for any nonsingular \( U \) and \( V \).

Suppose \( \mathbb{F} \) is algebraically closed, \( T : \mathcal{M}_{m,n}(\mathbb{F}) \to \mathcal{M}_{m,n}(\mathbb{F}) \) preserves \( \mathcal{P} \) and that \( T \) is singular. Without loss of generality we may assume that \( T \left( \begin{array}{cc} I_s & O \\ O & O \end{array} \right) = O \), for some \( s \) such that \( s \geq 1 \). Then, given any \( r \) with \( 1 \leq r \leq s \), since \( \left( \begin{array}{cc} I_s & O \\ O & O \end{array} \right) = \left( \begin{array}{ccc} I_r & O & O \\ O & O & O \\ O & O & O \end{array} \right) + \left( \begin{array}{ccc} O & O & O \\ O & I_{s-r} & O \\ O & O & O \end{array} \right) \) and \( T \) preserves \( \mathcal{P} \), we have
\[
0 = \rho \left( T \left( \begin{array}{cc} I_s & O \\ O & O \end{array} \right) \right) = \rho \left( T \left( \begin{array}{ccc} I_r & O & O \\ O & O & O \\ O & O & O \end{array} \right) \right) + \rho \left( T \left( \begin{array}{ccc} O & O & O \\ O & I_{s-r} & O \\ O & O & O \end{array} \right) \right)
\]
Thus, \( T \left( \begin{array}{ccc} I_r & O & O \\ O & O & O \\ O & O & O \end{array} \right) = T \left( \begin{array}{ccc} O & O & O \\ O & I_{s-r} & O \\ O & O & O \end{array} \right) = 0 \) for all \( r \) with \( 1 \leq r \leq s \).
Now, if $X$ is in the image of $T$ and $\rho(X) = r \leq s$, then there are nonsingular $U$ and $V$ such that $UXV = \begin{pmatrix} I_r & O \\ O & O \end{pmatrix}$. Thus $T \circ L_{UV} \circ T$ preserves $P$ and the image of $T \circ L_{UV} \circ T$ has strictly lower dimension than does $T$. Thus, suppose that $T$ preserves $P$, and that $T$ has the lowest dimension image of any operator in the semigroup generated by $T$ and the set $\{L_{UV} : U, V \text{ are nonsingular}\}$ (together with $\{L^t_{UV} : U, V \text{ are nonsingular}\}$ if $m = n$). From the above observations we have that if $s$ is the largest integer such that there is a matrix of rank $s$ in the kernel of $T$, then the image of $T$ has no element of rank $1$ through rank $s$. I.e., if $X \in \text{Im} T$, then $\rho(X) > s$ or $X = O$. Thus, the dimension of the image is at most $(n - s)(m - s)$ (since $F$ is algebraically closed, see Westwick, Theorem 2.1 [10]). Further, $\ker T$ is a space of matrices of rank at most $s$. By [3], $\dim \ker T \leq ns$. It follows that $nm = (n - s)(m - s) + ns$. This is possible only if $s = 0$ or $s = m$. Thus, unless $T$ maps all of $M_{m,n}(F)$ to $\{O\}$, $T$ must be nonsingular.

We have established:

**Lemma 1.** If $F$ is algebraically closed, and $T$ is a linear mapping preserving $P$, then either $T$ is the zero map or $T$ is nonsingular.

(Note that in the above argument the only requirement on the field is that the maximum dimension of a subspace of $M_{m,n}(F)$ all of whose nonzero elements have rank at least $s + 1$ is strictly less than $n(m - s)$.)

**Lemma 2.** If $T$ is a nonsingular linear operator on $M_{m,n}(F)$ preserving $P$ then $T$ preserves the set of matrices of rank at most $m - 1$.

**Proof.** Suppose that $A$ is a matrix of rank $k \leq m - 1$ such that $\rho(T(A)) = m$. Without loss of generality, we may assume that $A = \begin{pmatrix} I_k & O \\ O & O \end{pmatrix}$. Then, since $T$ preserves $P$ we must have that

$$\rho \left( T \left( \begin{pmatrix} I_m & O \\ O & O \end{pmatrix} \right) \right) = \rho(T(A)) + \rho \left( T \left( \begin{pmatrix} O & O \\ O & I_{m-k} \end{pmatrix} \right) \right) = m + \rho \left( T \left( \begin{pmatrix} O & O \\ O & I_{m-k} \end{pmatrix} \right) \right) > m$$

since $T$ is nonsingular. But this is impossible, consequently, if $\rho(A) < m$, we must have $\rho(T(A)) > m$. \qed

The following result of Chan and Lim is needed:

**Lemma 3.** [2] If $T$ preserves the set of all matrices of rank at most $k$ then either the image of $T$ has no matrix of rank greater than $k$ or there exist nonsingular matrices $U$ and $V$ such that $T = L_{UV}$ or, when $m = n$, $T = L^t_{UV}$. 

**Proof.** Suppose that $A$ is a matrix of rank $k \leq m - 1$ such that $\rho(T(A)) = m$. Without loss of generality, we may assume that $A = \begin{pmatrix} I_k & O \\ O & O \end{pmatrix}$. Then, since $T$ preserves $P$ we must have that

$$\rho \left( T \left( \begin{pmatrix} I_m & O \\ O & O \end{pmatrix} \right) \right) = \rho(T(A)) + \rho \left( T \left( \begin{pmatrix} O & O \\ O & I_{m-k} \end{pmatrix} \right) \right) = m + \rho \left( T \left( \begin{pmatrix} O & O \\ O & I_{m-k} \end{pmatrix} \right) \right) > m$$

since $T$ is nonsingular. But this is impossible, consequently, if $\rho(A) < m$, we must have $\rho(T(A)) > m$. \qed
**Theorem.** If $T : \mathcal{M}_{m,n}(\mathbb{F}) \rightarrow \mathcal{M}_{m,n}(\mathbb{F})$ is a linear operator and whenever $\rho(A + B) = \rho(A) + \rho(B)$ we have that $\rho(T(A+B)) = \rho(T(A)) + \rho(T(B))$, then either $T(X) = 0$ for all $X \in \mathcal{M}_{m,n}(\mathbb{F})$, or there exist nonsingular matrices $U \in \mathcal{M}_m(\mathbb{F})$ and $V \in \mathcal{M}_n(\mathbb{F})$ such that $T(X) = UXU$ for all $X \in \mathcal{M}_{m,n}(\mathbb{F})$ or, $m = n$ and $T(X) = UXU$ for all $X \in \mathcal{M}_n(\mathbb{F})$.

**Proof.** Suppose $T \neq 0$. By Lemma 1, $T$ is nonsingular. By Lemma 2, $T$ preserves the set of matrices of rank at most $m - 1$. By Lemma 3 the theorem follows.

**References**


SPECTRAL ANALYSIS FOR HYPERBOLIC INTEGRO-DIFFERENTIAL EQUATIONS WITH A WEAKLY SINGULAR KERNEL

S. K. Chung and M. G. Park

Abstract. Spectral analysis by energy method is given for a fully discrete methods for hyperbolic integro-differential equations with a weakly singular kernel. Stability and error estimates in $H^1$-norm are derived.

1. Introduction.

We will consider an approximate solution using the spectral methods for the hyperbolic integro-differential equations with a singular kernel:

\begin{equation}
\begin{aligned}
&u_{tt} + Au = \int_0^t K(t-s)Bu(s) \, ds + f, \quad (x,t) \in \Omega \times (0,T] \\
\end{aligned}
\end{equation}

with a Dirichlet boundary condition

\begin{equation}
\begin{aligned}
&u(x,t) = 0, \quad (x,t) \in \partial\Omega \times (0,T],
\end{aligned}
\end{equation}

and initial conditions

\begin{equation}
\begin{aligned}
&u(x,0) = u_0(x) \quad \text{and} \quad u_t(x,0) = u_1(x), \quad x \in \Omega.
\end{aligned}
\end{equation}

Here $\Omega = (0,\pi)^2$, $A(x,t)$ is a linear, positive, symmetric, uniformly elliptic operator and $B(x)$ is a general partial differential operator of second order with smooth coefficients. Given functions $u_0(x)$, $u_1(x)$ and $f(x,t)$ are real-valued and sufficiently smooth. Further, $K(t)$ is a positive decreasing weakly singular kernel with the property:

\begin{equation}
\begin{aligned}
&K(t) \leq Ct^\alpha, \quad -1 < \alpha < 0, \quad t > 0.
\end{aligned}
\end{equation}

Integro-differential equation (1.1) arises in visco-elastic problems. For more references on problem of the type (1.1), we refer to Renardy, Hrusa and Nohel[13] and references therein. The problem (1.1) with smooth kernels has been studied in Dix and Torrejón[7], Torrejón and Yong[14], Yanik and Fairweather[15], where they showed

1991 Mathematics Subject Classification. 65M70, 65M15.

Key words and phrases. Spectral methods, integro-differential equations, fully discrete scheme.

This work is partially supported by SNU-DAEWOO Research Fund under grant 96-03-1057.

Typeset by \TeX
existence and uniqueness of solutions for smooth initial data using energy estimates. In [10], Hrusa and Renardy showed existence of discontinuous solution for nonsmooth initial data. Local and global weak solutions of (1.1) with singular kernels have been studied in Engler [8] and Hrusa and Renardy [9] using the limit of solutions with smoothing kernels. Galerkin approximate solutions of (1.1) with weakly singular kernels have been discussed in Choi and MacCamy [6], in which error estimates are given for the semidiscrete scheme. For the fully discrete scheme, Galerkin solutions have been discussed in Pani, Thomée and Wahlbin [12] in case of the kernel \( K(t) \equiv 1 \), where they also considered storage reduction. In spite of many works on (1.1) with singular kernels, to our best knowledge, the finite element solutions of (1.1) with time stepping (fully discrete scheme) appear to be untouched, even though it is crucial for real computation. In this paper, we discuss error estimates of backward Euler's fully discrete spectral approximate solutions for (1.1) with a weakly singular kernel. In section 2, error estimates for several projections, like \( L^2 \)-projection, Ritz projection and Ritz-Volterra projection, will be discussed. In section 3, error estimates for finite element solutions by spectral methods with time stepping will be discussed, where \( H^1 \)-error estimates of \( O(k + N^{-2}) \) are shown.

2. Error Estimates for Projections.

Let \( V_N = \text{span}\{ \psi_{ij} = \sin i_1 \sin j_2 : i, j = 1, 2, \ldots, N \} \) be the subspace of the usual Sobolev space \( V = H^1_0(\Omega) \). The weak solution for (1.1) is defined as a function \( u_N : (0, T] \rightarrow V_N \) such that for all \( \chi \in V_N \)

\[
(u_Nu, \chi) + a(u_N, \chi) = \int_0^t K(t - s)b(u_N(s), \chi) \, ds + (f, \chi),
\]

\[
(u_N(x, 0), \chi) = (u_0, \chi),
\]

\[
(u_N(t, 0), \chi) = (u_1, \chi).
\]

Here \( a(\cdot, \cdot) \) and \( b(\cdot, \cdot) \) are bilinear forms on \( H^1_0(\Omega) \times H^1_0(\Omega) \) associated with differential operators \( A \) and \( B \), respectively. The inner product \( (\cdot, \cdot) : H^1_0(\Omega) \times H^1_0(\Omega) \rightarrow R \) is defined as

\[
(\phi, \psi) = \int_\Omega \phi(x)\psi(x) \, dx, \quad \phi, \psi \in H^1_0(\Omega).
\]

Define a \( L^2 \)-projection \( P_N : L^2(\Omega) \rightarrow V_N \) by \( P_Nv = \sum_{i,j} a_{ij}(t)\psi_{ij} \) for \( v(x, t) = \sum_{i,j} a_{ij}(t)\psi_{ij} \). Then we have

\[
(v - P_Nv, \chi) = 0, \quad \forall \chi \in V_N.
\]

That is, \( P_N \) is an orthogonal projection operator. It is well known that the following error estimate holds for the the \( L^2 \)-projection (see [4]). Hereafter, a constant \( C \) will be used as a generic constant independent of \( N \) and mesh \( k \). For notational convenience, we omit dependent variables \( x \) and \( t \) if there is no confusion.
Lemma 2.1. There exists a constant $C$ such that
$$
\|v - P_N v\| \leq CN^{-2} \|v\|_2.
$$

The following version of Gronwall’s lemma will be frequently used for error estimates, whose proof can be found in Chen, Thomée and Wahlbin [5].

Lemma 2.2. Assume that $y$ is a nonnegative function in $L_1(0, T)$ and satisfies
$$
y(t) \leq z(t) + \beta \int_0^t K(t - s)y(s) \, ds, \quad 0 < t \leq T, \quad -1 < \alpha < 0,
$$
where $z(t) \geq 0, \beta \geq 0$. Then there is a constant $C_T$ such that
$$
y(t) \leq z(t) + C_T \int_0^t K(t - s)z(s) \, ds, \quad 0 < t \leq T.
$$

We now introduce the standard Ritz projection operator $R_N : H^1_0(\Omega) \rightarrow V_N(\Omega)$ with

$$
a(v - R_N v, \chi) = 0, \quad \forall \chi \in V_N(\Omega).
$$

The error estimate of Ritz projection can be found in Bressan and Quarteroni[3], Bernardi and Maday[2] for all $v \in V$ with $0 \leq \mu \leq r$ and $r \geq 1$ as

$$
\|v - R_N v\|_\mu \leq CN^{-(r-e(\mu))}\|v\|_r,
$$

where
$$
e(\mu) = \begin{cases} 
\mu, & \mu \leq 1 \\
2\mu - 1, & \mu > 1.
\end{cases}
$$

Further, we introduce the Ritz-Volterra projection operator $\Pi_N : V \rightarrow V_N$ by

$$
a((\Pi_N u - u)(t), \chi) = \int_0^t K(t - s)b((\Pi_N u - u)(s), \chi) \, ds, \quad \forall \chi \in V_N^0(\Omega),
$$
as in Lin, Thomée and Wahlbin[11]. Then we have the following error estimate for the Ritz-Volterra projection.

Lemma 2.3. There exists a constant $C$ such that for $u \in V$.
$$
\|\Pi_N u - u\| + N^{-1} \|\Pi_N u - u\|_1 \leq CN^{-2} \sup_{s \leq t} \|u(s)\|_2.
$$

Proof. Let $\rho_N = \Pi_N u - u$. We begin with an $H^1$ - estimate for $\rho_N$. From (2.3), we obtain
$$
\|R_N u - u\| + N^{-1} \|R_N u - u\|_1 \leq CN^{-2} \|u\|_2.
$$
It follows from the coercivity of $a(\cdot, \cdot)$ and the orthogonal projection that for $c_0 > 0,$
\[
c_0 \| \Pi_N u - R_N u \|_1^2 \\
\leq a(\Pi_N u - R_N u, \Pi_N u - R_N u) \\
= a(\rho_N, \Pi_N u - R_N u) + a(u - R_N u, \Pi_N u - R_N u) \\
= a(\rho_N, \Pi_N u - R_N u) \\
= \int_0^t K(t-s) b(\rho_N(s), (\Pi_N u - R_N u)(t)) \, ds \\
\leq C \| \Pi_N u - R_N u \|_1 \int_0^t K(t-s) \| \rho_N(s) \|_1 \, ds.
\]
Thus, we obtain
\[
\| \Pi_N u - R_N u \|_1 \leq C \int_0^t K(t-s) \| \rho_N(s) \|_1 \, ds
\]
and
\[
\| \rho_N \|_1 \leq \| \Pi_N u - R_N u \|_1 + \| R_N u - u \|_1 \leq C \int_0^t K(t-s) \| \rho_N(s) \|_1 \, ds + \| R_N u - u \|_1.
\]
It follows from Lemma 2.2 that
\[
\| \rho_N \|_1 \leq \sup_{s \leq t} \| (R_N u - u)(s) \|_1 \leq C N^{-1} \sup_{s \leq t} \| u(s) \|_2.
\]
We now consider the $L_2$-estimate for $\rho_N$ using the duality argument. For any $\phi \in L^2,$ let $\psi$ be the solution of
\[
(2.5) \quad A\psi = \phi \quad \text{in } \Omega, \quad \psi = 0 \quad \text{on } \partial \Omega.
\]
Then $\psi$ is a unique solution of (2.5) such that
\[
\| \psi \|_2 \leq C \| \phi \| = C.
\]
Note that, for $\chi \in V_N$,
\[
(2.6) \quad (\rho_N, \phi) = (\rho_N, A\psi) = a(\rho_N, \psi - \chi) + a(\rho_N, \chi).
\]
The last term on the right hand side of (2.6) becomes
\[
a(\rho_N, \chi) = \int_0^t K(t-s) b(\rho_N(s), \chi) \, ds \\
\leq \int_0^t K(t-s) b(\rho_N(s), \chi - \psi) \, ds + \int_0^t K(t-s) (\rho_N(s), B^*\psi) \, ds,
\]
where $B^*$ is the adjoint of $B$. Replacing $\chi = R_N \psi$ and using (2.6), we obtain
\[
(\rho_N, \phi) \leq C \{ \|R_N \psi - \psi\|_1 \sup_{s \leq t} \|\rho_N(s)\|_1 + \|\psi\|_2 \int_0^t K(t - s)\|\rho_N(s)\| ds \}
\[
\leq C \{ N^{-1} \|\psi\|_2 \cdot N^{-1} \sup_{s \leq t} \|u(s)\|_2 + \|\phi\| \int_0^t K(t - s)\|\rho_N(s)\| ds \}.
\]
Replacing $\phi = \rho_N / \|\rho_N\|$ in the above inequality, we obtain
\[
\|\rho_N\| \leq C \{ N^{-2} \sup_{s \leq t} \|u(s)\|_2 + \int_0^t K(t - s)\|\rho_N(s)\| ds \}.
\]
An application of Lemma 2.2 completes the proof. □

Following the proof of Lemma 2.3, we may also obtain
\[
(2.7) \quad \|\rho_{N,t}\| + N^{-1} \|\rho_{N,t}\|_1 \leq CN^{-2} \sup_{s \leq t} \|u_t(s)\|_2
\]
and
\[
(2.8) \quad \|\rho_{N,tt}\| + N^{-1} \|\rho_{N,tt}\|_1 \leq CN^{-2} \sup_{s \leq t} \|u_{tt}(s)\|_2.
\]

3. Discretization in Time Direction.

Let $M$ be a positive integer and $k = T/M$. We define difference operators
\[
\bar{\partial}_k \phi^n = \frac{\phi^n - \phi^{n-1}}{k}, \quad \bar{\partial}_k^2 \phi^n = \bar{\partial}_k(\bar{\partial}_k \phi^n)
\]
and
\[
\bar{\phi}(s) = \frac{1}{k} [(t_j - s)\phi(t_{j-1}) + (s - t_{j-1})\phi(t_j)], \quad t_{j-1} \leq s \leq t_j.
\]
Let $U_N^n \in V_N$ be a solution of
\[
(3.1a) \quad (\bar{\partial}_k^2 U_N^n, \chi) + a_n(U_N^n, \chi) = \int_0^t K(t_n - s) b(\bar{U}_N(s), \chi) ds + (f^n, \chi), \quad n \geq 2,
\]
\[
(3.1b) \quad (U_N^0, \chi) = (u_0, \chi),
\]
\[
(3.1c) \quad (\bar{\partial}_k U_N^1, \chi) = (u_1, \chi),
\]
for all $\chi \in V_N$. We use the trapezoidal rule as in Weiss[16] and Atkinson[1] for integrations in (3.1). Then we obtain the quadrature error
\[
(3.2) \quad \varepsilon_n(\phi) = \sum_{j=1}^n \{ \nu_{n,j} \phi(t_{j-1}) + \mu_{n,j} \phi(t_j) \} - \int_0^t K(t_n - s)\phi(s) ds,
\]
where
\[
(3.3a) \quad \nu_{n,j} = \int_{t_{j-1}}^{t_j} (t_j - s) K(t_n - s) ds,
\]
\[
(3.3b) \quad \mu_{n,j} = \int_{t_{j-1}}^{t_j} (s - t_{j-1}) K(t_n - s) ds.
\]
Lemma 3.1. There is a constant $C$ such that if $\phi \in L_\infty(0, T; L_2)$, then
\[ k \sum_{n=1}^{M} \|\varepsilon_n(\phi)\| \leq Ck^2 \max_{0 \leq s \leq t_n} \|\phi_k(s)\|. \]

Proof. Since we may rewrite $\phi(s)$ as
\[ \phi(s) = \begin{cases} 
\phi(t_{j-1}) + (s - t_{j-1})\phi_t(\xi_{j-1}), & t_{j-1} < \xi_{j-1} < s \\
\phi(t_j) + (s - t_j)\phi_t(\zeta_j), & s < \zeta_j < t_j,
\end{cases} \]
we obtain, from the Taylor’s theorem,
\[ |(\tilde{\phi} - \phi)(s)| \leq \frac{1}{k}(t_j - s)(s - t_{j-1})\{|\phi_t(\xi_{j-1})| + |\phi_t(\zeta_j)|\} \leq k\{ |\phi_t(\xi_{j-1})| + |\phi_t(\zeta_j)|\}. \]
Since $\varepsilon_n(\phi) = \int_0^{t_n} K(t_n - s)(\tilde{\phi} - \phi)(s) \, ds$, we obtain
\[ |\varepsilon_n(\phi)| \leq 2k \sum_{j=1}^{n} \max_{t_{j-1} \leq s \leq t_j} |\phi_t(s)| \int_{t_{j-1}}^{t_j} K(t_n - s) \, ds \leq Ck \max_{0 \leq s \leq t_n} |\phi_t(s)|. \]
This implies
\[ \|\varepsilon_n(\phi)\| \leq Ck \max_{0 \leq s \leq t_n} \|\phi_t(s)\|. \]
Hence the required result holds from summation of the above inequality. \( \square \)

The following lemma is a discrete version of Lemma 2.2, which will be used judiciously.

Lemma 3.2. Let $\nu_{n,j}$ and $\mu_{n,j}$ be defined as in (3.3). Assume that $y_n \geq 0$, $z_n \geq 0$ and $\beta \geq 0$. If, either $x_{n,j} = \nu_{n,j}$ or $x_{n,j} = \mu_{n,j},$
\[ y_n \leq z_n + \beta \sum_{j=1}^{n} x_{n,j}y_j, \quad n \geq 0, \]
holds, then there is a constant $C$ such that
\[ y_n \leq z_n + C \sum_{j=1}^{n} x_{n,j}z_j, \quad n \geq 0. \]

We now state the stability of approximate solutions in terms of a discrete energy norm $\|\cdot\|$ defined by
\[ \|\phi^n\|^2 = \|\tilde{\phi}_h^n\|^2 + \|\phi^n\|^2, \quad n \geq 1. \]
Let $U_N \in V_N$ be a solution of (3.1). Then there exists a constant $C$ such that
\[ \|U_N\|_1 \leq C \{ \|U_N^1\|_1 + k \sum_{n=2}^M \|f^n\| \} . \]

**Proof.** Replacing $\chi = \bar{\partial}_k U_N^n$ in (3.1a), we obtain
\begin{equation}
\begin{split}
(\bar{\partial}_k^2 U_N^n, \bar{\partial}_k U_N^n) + a_n(U_N^n, \bar{\partial}_k U_N^n)
\end{split}
\end{equation}
\[ = \sum_{j=1}^n \int_{t_{j-1}}^{t_j} K(t_n - s) b(\bar{U}_N(s), \bar{\partial}_k U_N^n) \, ds + (f^n, \bar{\partial}_k U_N^n)
\]
\[ \equiv I_1^n + I_2^n . \]

Note that
\[ (\bar{\partial}_k^2 U_N^n, \bar{\partial}_k U_N^n) = \frac{1}{2} \| \bar{\partial}_k U_N^n \|^2 + \frac{k}{2} \| \bar{\partial}_k U_N^n \|^2 \]
and
\[ a_n(U_N^n, \bar{\partial}_k U_N^n) = \frac{1}{2} \bar{\partial}_k(a_n(U_N^n, U_N^n)) - \frac{1}{2}(\bar{\partial}_k a_n)(U_N^{n-1}, U_N^{n-1}) + \frac{k}{2} a_n(\bar{\partial}_k U_N^n, \bar{\partial}_k U_N^n). \]

Multiplying both sides of (3.4) by $2k$ and summing from $n = 2$ to $M$, we obtain, for positive constants $c_0$ and $C$,
\[ \| \bar{\partial}_k U_N^M \|^2 + c_0 \| U_N^M \|^2_1 \leq \| \bar{\partial}_k U_N^1 \|^2 + C \| U_N^1 \|^2_1 + k \sum_{n=2}^M (\bar{\partial}_k a_n)(U_N^{n-1}, U_N^{n-1}) \]
\[ + 2k \sum_{n=2}^M (I_1^n + I_2^n) . \]

The above inequality can be rewritten using the discrete norm $\|v_{dot}\|$ as
\begin{equation}
\begin{split}
\|U_N^M\|_1^2 \leq C \{ \|U_N^1\|_1^2 + k \sum_{n=2}^M \|U_N^{n-1}\|_1^2 \} + k \sum_{n=2}^M \|f^n\| \cdot \|U_N\|_1^2 .
\end{split}
\end{equation}

Let $\|U_N\|_1^2 = \max_{0 \leq n \leq M} \|U_N^n\|_1^2$. Then
\[ k \sum_{n=2}^m |I_2^n| = k \sum_{n=2}^M \|f^n\| \cdot \|U_N\|_1^2 .
\]

On the other hand, we can rewrite $I_1^n$ using notations in (3.2)-(3.3) as
\[ I_1^n = \nu_{n,j} b(U_N^{j-1}, \bar{\partial}_k U_N^n) + \mu_{n,j} b(U_N^n, \bar{\partial}_k U_N^n)
\]
\[ + \sum_{j=1}^{n-1} \bar{\partial}_k \{ \nu_{n,j} b(U_N^{j-1}, U_N^n) + \mu_{n,j} b(U_N^n, U_N^n) \}
\]
\[ - \sum_{j=1}^{n-1} ((\bar{\partial}_k \nu_{n,j}) b(U_N^{j-1}, U_N^{n-1}) + (\bar{\partial}_k \mu_{n,j}) b(U_N^{j-1}, U_N^{n-1})) .
\]
Summation both sides from \( n = 2 \) to \( M \), we obtain
\[
\sum_{n=2}^{M} \sum_{k=2}^{n} I_1^k = \sum_{n=2}^{M} \left\{ \mu_{n,n} b(U_N^n, U_N^n) + (\nu_{n,n} - \mu_{n,n}) b(U_N^n, U_N^{n-1}) - \nu_{n,n} b(U_N^{n-1}, U_N^{n-1}) \right\} \\
+ \sum_{j=1}^{M-1} \left\{ \nu_{M,j} b(U_N^{j-1}, U_N^M) + \mu_{M,j} b(U_N^j, U_N^M) \right\} \\
- \sum_{j=1}^{M-1} \left\{ \nu_{j,j} b(U_N^{j-1}, U_N^j) + \mu_{j,j} b(U_N^j, U_N^1) \right\} \\
- k \sum_{j=1}^{M-1} \sum_{n=j+1}^{M} \left\{ (\partial_k \nu_{n,j}) b(U_N^{j-1}, U_N^{n-1}) + (\partial_k \mu_{n,j}) b(U_N^j, U_N^{n-1}) \right\}.
\]

Thus
\[
k \sum_{n=2}^{M} |I_1^n| \leq C \sum_{j=1}^{M} \left\{ |\nu_{j,j}| ||U_N^{j-1}||_1 + |\nu_{j,j} - \mu_{j,j}| ||U_N^j||_1 + |\nu_{j,j} - \mu_{j,j}|| |U_N^{j-1}||_1 \right\} ||U_N||_1: M \\
+ C \sum_{j=1}^{M} \left\{ |\nu_{M,j}| + |\nu_{j,j}| ||U_N^{j-1}||_1 + (|\mu_{M,j}| + |\mu_{j,j}|) ||U_N^j||_1 \right\} ||U_N||_1: M \\
+ Ck \sum_{j=1}^{M-1} \left\{ ||U_N^{j-1}||_1 \sum_{n=j+1}^{M} |\partial_k \nu_{n,j}| + ||U_N^j||_1 \sum_{n=j+1}^{M} |\partial_k \mu_{n,j}| \right\} ||U_N||_1: M.
\]

Since
\[
|\partial_k \nu_{n,j}| = \frac{1}{k} \int_{t_{j-1}}^{t_j} (t_j - s)[K(t_n - s) - K(t_{n-1} - s)] \, ds \\
\leq \int_{t_{j-1}}^{t_j} [K(t_{n-1} - s) - K(t_n - s)] \, ds,
\]
we obtain, by interchanging summation with integration,
\[
\sum_{n=j+1}^{M} |\partial_k \nu_{n,j}| \leq \sum_{n=j+1}^{M} \int_{t_{j-1}}^{t_j} [K(t_{n-1} - s) - K(t_n - s)] \, ds \\
= \int_{t_{j-1}}^{t_j} [K(t_j - s) - K(t_M - s)] \, ds \\
\leq C(||K||_{L_1(0,T)}).
\]

Similarly, we obtain
\[
\sum_{n=j+1}^{M} |\partial_k \mu_{n,j}| \leq C(||K||_{L_1(0,T)}).
\]
Noting that $|\nu_{n,j}| \leq k^{2+\alpha}$ and $|\mu_{n,j}| \leq k^{2+\alpha}$ for $j \leq n \leq M$ and dividing (3.5) by $\|U_N\|_{1:1}$, we obtain

$$\|U_N^M\|_1 \leq C \{ \|U_N^1\|_1 + k \sum_{n=1}^{M} \|U_N^n\|_1 + k \sum_{n=2}^{M} \|f^n\| \}.$$ 

Hence, the discrete Gronwall’s inequality completes the proof. □

Let $\theta^n = U_N^n - \Pi_N u(t_n)$. Then the error $e^n = U_N^n - u(t_n) = \theta^n + \rho^n_N$. Since we know the estimate of $\rho^n_N$, we have only to find the estimate for $\theta^n$.

**Theorem 3.2.** Let $u(t)$ be the solution of (1.1) and $U_N^k$ be a solution of (3.1). Then there exists a constant $C$ such that

$$\|\theta^n\|_1 \leq C \{ N^{-2} + \|\theta^0\|_1 \}.$$ 

**Proof.** From (2.4) and (3.1), we obtain

$$\|
\begin{array}{l}
(\partial_t^2 \theta^n, \chi) + a_n(\theta^n, \chi) = \int_0^{t_n} K(t_n - s) b(\theta^k, \chi) \, ds + \left( \sum_{i=1}^{3} J_i^n, \chi \right),
\end{array}$$

where

$$\begin{align*}
J_1^n &= u_{tt}(t_n) - \partial_t^2 \theta^n(t_n), \\
J_2^n &= -\partial_t^2 \rho_N^n, \\
J_3^n &= \int_0^{t_n} K(t_n - s) B\{ \Pi_N u(s) - \Pi_N u(t_n) \} \, ds.
\end{align*}$$

It follows from Theorem 3.1 that

$$\|\theta^M\|_1 \leq C \{ \|\theta^1\|_1 + k \sum_{n=2}^{M} \|J_i^n\| \}.$$ 

From the relation $J_3^n = \varepsilon_n(B\rho_N) + \varepsilon_n(Bu)$ and Lemma 3.1, we obtain

$$\|J_3^n\| \leq C k \{ \sup_{0 \leq s \leq t_n} \|B\rho_N(s)\| + \sup_{0 \leq s \leq t_n} \|Bu(s)\| \} \leq C k \|u_t\|_2.$$ 

Hence, we obtain

$$k \sum_{n=2}^{M} \|J_3^n\| \leq C k^2 \sum_{n=2}^{M} \left( \sup_{0 \leq s \leq t_n} \|u_t\|_2 \right) \leq C \int_0^{t_M} \|u_t(s)\|_2 \, ds.$$ 

Similarly, we obtain

$$k \sum_{n=2}^{M} \|J_2^n\| \leq C k^2 \sum_{n=2}^{M} \int_{t_{n-1}}^{t_n} \|u_{tttt}\| \, ds = C k^2 \int_{t_1}^{t_M} \|u_{tttt}\| \, ds.$$ 

For the estimate of $J_1^n$, it follows from (2.8) that

$$k \sum_{n=2}^{M} \|J_1^n\| \leq C \sum_{n=2}^{M} \int_{t_{n-1}}^{t_n} \|\rho_{Ntt}\| \, ds \leq C N^{-2}.$$ 

These complete the proof. □
Remarks. Spectral analysis is discussed for a hyperbolic integro-differential equation with a weakly singular kernel and error estimates of the spectral method is given. Because of the memory term and global bases functions, a storage problem in computation arises in the method. In order to overcome this problem, we will give an improved method elsewhere.

References


Department of Mathematics Education
Seoul National University
Seoul 151–742
SUBSTRUCTURING ALGORITHM FOR STRUCTURAL OPTIMIZATION USING THE FORCE METHOD

HO-JONG JANG
Department of Mathematics
Hanyang University, Seoul 133-791, Korea
hjang@fermat.hanyang.ac.kr

Abstract
We consider some numerical solution methods for equality-constrained quadratic problems in the context of structural analysis. Sparse orthogonal schemes for linear least squares problem are adapted to handle the solution step of the force method. We also examine these schemes with substructuring concepts.

1. Introduction

The problem of computing a sparse self-stress matrix from the equilibrium matrix of a structure arises in the force method of structural analysis. In this context, the system force vector \( f \) is a minimizer of the quadratic form corresponding to the potential energy \( \frac{1}{2} f^T A f \), subject to the equilibrium equation \( E f = p \). In these equations, \( E \) is the underdetermined equilibrium matrix, \( p \) is the vector of applied loads, and \( A \) is the symmetric, block diagonal, element flexibility matrix.

In the force method, the vector \( f \) is computed in two phases:

1. Compute a self-stress matrix \( N \) from the equilibrium matrix \( E \), and a particular solution \( f_p \) of the equilibrium equation.

2. Solve the system
   \[
   N^T A N f_0 = -N^T A f_p,
   \]
   and compute \( f = f_p + N f_0 \).

Since the self-stress matrix \( N \) is a basis for the nullspace of the equilibrium matrix, we call it a null basis. Methods of finding a sparse or structural basis of the nullspace of the equilibrium matrix have been the subject of extensive study over the past few years (see, for example [9]).

In our approach to compute the nullspace we use a parallel scheme by utilizing the graph theoretic ideas in what we call the substructuring method, which is introduced...

Key Words: Structural optimization, force method, substructuring method, null basis.
AMS subject classification: 65F05, 65F50.
This work was partially supported by Hanyang University Research Grants in 1998.
by Plemmons and White\cite{8}. This method successfully handles parallel nullspace computation. Once the substructures have been identified, then null basis is even readily computable by observation.

In fact nullspace computation of the force method has an important role in the entire force method computation. Let $k$ be the number of columns in $N$, then the system flexibility matrix $N^TAN$ is a square matrix of order $k$. In general, the matrix $N^TAN$ is unlikely to be sparse even when $N$ is sparse, and hence it is explicitly computed and stored only for small $k$. However, by using the special structure inherent in self-stress matrices of the physical structures based on the substructuring method, we can compute the system (1) more efficiently.

The focus of this paper is the use of the substructuring method for phase 2 of the force method while the system (1) is carried out using orthogonal factorization techniques developed for linear least squares problems.

2. Substructuring method

In general, for an equilibrium matrix with $m$ rows and $n$ columns, there exists a product of elementary matrices, $P$, such that

$$PE = [E_1, E_2] = E_1[I_m, E_1^{-1}E_2],$$

where $E_1$ is nonsingular. Consequently, the nullspace of $PE$, and hence $E$, is generated by the columns of the block matrix

$$N = \begin{bmatrix} E_1^{-1}E_2 \\ -I_{n-m} \end{bmatrix}.$$ 

With special interest in parallel computation, motivation for developing the substructuring method was that the nullspace computation (forming $N$) can often be done by appropriate ordering of the nodes and elements, extending certain results in\cite{2}. This ordering yields a matrix $E$ with a great deal of structure which can be exploited by multiprocessoring computers in forming $N$. We briefly consider two classes of structural problems in order to see how the substructuring method works.

**Example 1.** Substructuring with proper partition of a pin-jointed-truss leads to matrix $E$, which is $20 \times 62$, resembling that depicted in Figure 1.

**Fig. 1.** Pin-jointed-truss with 6 disjoint substructures and its equilibrium matrix
The blocks in $E$ are usually sparse (see Plemmons and White [8] for details). Note that each upper triangular diagonal block in $E_1$ has an inverse, and corresponds to stable substructures given by $S_1 = \{1, \{e_1\}\}, \ldots$, $S_3 = \{3, 4, 5, \{e_3, e_4, e_5\}\}, \ldots$, $S_6 = \{\{10\}, \{e_{10}\}\}$. The remaining elements, $e_{11}, \ldots, e_{31}$, connect these stable substructures. Note that the diagonal structure of $E_2$ which is a result of ordering the connecting elements from the left to the right. Because of this structure of $E_2$, we can also reduce a lot of work in parallel computation of $E^{-1}_1 E_2$.

**Example 2.** Consider a rigid frame which models a wheel with 8 spokes in Figure 2. Each spoke is a stable substructure and together they form a proper partition. The matrix $E$ is $96 \times 120$.

**Fig. 2.** Rigid frame which models a wheel and its equilibrium matrix

More discussion of the nullspace computation with various examples is in [8].

3. **The force method: phase 2**

Given a particular solution $f_p$ of the equilibrium equation, the main task of phase 2 of the force method is the computation of the redundant force vector $f_0$ which satisfies system (1). System (1) is simply the normal equation for the weighted least squares problem:

$$\minimize_{f_0} \|G^{-1}(N f_0 + f_p)\|_2, \quad (2)$$

where $G$ is Cholesky factor of element stiffness matrix $A^{-1}$. Several methods for solving problems of the form (2) are described in [1]. The traditional method of normal equations consists of the direct application of Cholesky's method to the symmetric positive definite matrix $N^T A N$. Unfortunately, explicitly forming the matrix $N^T A N$ can lead to loss of special block structure of $N$ and worsening of the conditioning of the problem. A better approach in this regard is to apply orthogonal transformations to the matrix $G^{-1} N$, leading to an algorithm of the following form:

$$P_1 G^{-1} N P_2^T = Q \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix}, \quad -Q^T P_1 G^{-1} f_p = \begin{bmatrix} c \\ d \end{bmatrix}, \quad f_0 = P_2^T R^{-1} c, \quad (3)$$

where $R$ is an upper triangular matrix of order $n - m$, $P_1$ and $P_2$ are permutation matrices of order $n$ and $n - m$, respectively, $Q$ is an orthogonal matrix of order $n$, and $c$ and $d$ are vectors of length $n - m$ and $m$, respectively.
We need to consider the shape of $G^{-1}N$ rather than $N^TA_N$ based on substructuring method. The calculation of the QR factorization of $G^{-1}N$ can be done in parallel. The shape of $G^{-1}N$ in Figure 3 is the same block structure as $N$ because $G$ has block diagonal structure.

Fig.3. $G^{-1}N$ for Figure 2.

The Givens transformations can be used in parallel by concurrently working on the eight column blocks. The block in row block 33 and column block 1 can be used to annihilate the components in row blocks 1-4 and column block 1. At the same time the terms in row blocks 5-8 and column 2 can be annihilated by the block in row block 34 and column block 2. The remainder of the top 32 row blocks can be annihilated concurrently in a similar manner.

Although orthogonal factorization is numerically superior to the normal equations, poor results may be obtained with either method when the element flexibility matrix $A$ is ill-conditioned. In a series of papers [6, 7], Paige has developed schemes which can considerably reduce this difficulty. Paige’s new schemes for linearly constrained sum-of-squares (LCSS) problem to (2) with trunback-QR factorization for nullspace computation are discussed in [3]. Furthermore, these new schemes successfully applied to rapid reanalysis of structures, and reported in [4]. In our approach, each of the two formulations of Paige’s method is reviewed by using the substructuring method with special structure of the nullspace $N$.

**Formulation I.** Following Paige [6], if we define the weighted residual vector

$$v = G^{-1}(N_{f_0} + f_p),$$

then problem (2) can be written in the equivalent form

$$\text{Min}_{v_{f_0}} v^T v \quad \text{subject to} \quad Gv = N_{f_0} + f_p.$$  \hspace{1cm} (4)

First, decompose $N$ in (4) as

$$Q^T N = \begin{bmatrix} Q_1^T N \\ Q_2^T N \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix},$$  \hspace{1cm} (5)
where $R$ is a nonsingular upper triangular matrix of order $n - m$, $Q = (Q_1, Q_2)$ is an orthogonal matrix of order $n$, and $Q_1$ and $Q_2$ are $n \times (n - m)$ and $n \times m$ matrices, respectively. The constraints in (4) then split into

$$Q_1^T Gv = Rf_0 + Q_1^T f_p$$

(6)

$$Q_2^T Gv = Q_2^T f_p.$$  

(7)

Since $R$ has full row rank, (6) can always be solved for $f_0$ once $v$ is given, and so (7) gives the constraints on $v$, and (4) becomes

$$\text{Min}_v v^T v \quad \text{subject to} \quad Q_2^T Gv = Q_2^T f_p.$$  

Next, apply the QR factorization to $(Q_2^T G)^T$ starting from the lower right components to decompose $Q_2^T G$ so that

$$Q_2^T GP = (0, L_2),$$

where $P = (P_1, P_2)$ is an orthogonal matrix of order $n$, and $P_1$ and $P_2$ are $n \times (n - m)$ and $n \times m$ matrices, respectively. Here $L_2$ has full column rank.

That is, decompose $Q^T G$ as

$$Q^T G = \begin{bmatrix} Q_1^T GP_1 & Q_2^T GP_2 \\ 0 & L_2 \end{bmatrix} = \begin{bmatrix} L_1 & L_{12} \\ 0 & L_2 \end{bmatrix},$$

(8)

Assuming $L_2$ is nonsingular we now obtain

$$v = P_2 L_2^{-1} Q_2^T f_p,$$

since $Q_2^T GP_2 = L_2$. Finally, $f_0$ is recovered from the triangular system (6).

The advantage of using substructuring in the formulation I occurs in (5). Since $N$ is more sparse than $G^{-1} N$, we need to have relatively less computation compare to (3). However, the first formulation does not take advantage of any special structure the matrix $G$ may have, $G$ will be triangular if it is computed by Cholesky factorization, and in our case $G$ has block diagonal structure. Indeed, that structure is in general destroyed by the computation (6)-(7). This problem can be resolved in Paige’s alternate formulation of LCSS.

**Formulation II.** Paige[7] has given a formulation II in which the two orthogonal transformations $U$ and $V$ (to the corresponding matrices $Q$ and $P$ in (8), respectively) simultaneously in a manner which retains the triangular structure of $G$ throughout the computations. For implementation details, see [7].

The result is a factorization of the form

$$U^T [f_p, N, GV] = \begin{bmatrix} f_{p1} & 0 & M_1 & 0 \\ f_{p2} & S^T & M_{21} & M_2 \end{bmatrix},$$

(9)

where $S^T$, $M_1$, and $M_2$ are lower triangular matrices of order $n - m$, $m$, and $n - m$, respectively, and $U = (U_1, U_2)$ and $V = (V_1, V_2)$ are orthogonal matrices of order $n$
and $m$, respectively. The matrices $U_1$ and $V_1$ are $n \times m$, and the matrices $U_2$ and $V_2$ are $n \times (n-m)$. We note that the matrices in (9) are not necessarily identical to the corresponding matrices in (8). Applying transformation (9) and using the change of variable

$$V^Tv = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

problem (4) becomes

$$\min_{v_1, v_2} v_1^Tv_1 + v_2^Tv_2 \quad \text{subject to} \quad \begin{bmatrix} M_1 & 0 \\ M_{21} & M_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ S^T \end{bmatrix} f_0 + \begin{bmatrix} f_{p1} \\ f_{p2} \end{bmatrix}.$$ 

Thus $v_1$ is completely determined by the equation

$$M_1v_1 = f_{p1},$$

and the functional is minimized by taking $v_2 = 0$. Finally, the solution $f_0$ may now be determined from the system

$$S^T f_0 = M_{21}v_1 - f_{p2}.$$ 

The reduction in (9) can be carried out in two main stages. The first stage consists of $n - m$ major steps in each which one computes the superdiagonal of the $(f_p, N)$ matrix is annihilated while the lower triangular form of the $G$ matrix is preserved, in our case $G$ is already triangular and block diagonal. For example with $m=1, n=4$, the second step is as in Figure 4.

**Fig. 4. Second step of the zero chasing for (9).**

The rotations are ordered $1, 1', 2, 2'$, and the nonzero element $i^*$, introduced by rotation $i$ from the left, is immediately made to be zero by rotation $i'$ from the right. When the first stage has been completed, each step of the second stage has the same form, eliminating one more complete diagonal of the $(f_p, N)$ matrix, and maintaining the lower triangular form of the $N$ matrix. Thus in the second stage there will be $m-1$ major steps. Since the shape of $N$ is block banded and there are not many nonzero components of the superdiagonal of the $N$, we do not need much computation in the first stage. For the second stage computation, the Givens transformations can be used in parallel by concurrently working on each column block. Hence, the lower triangular $S^T$ from $N$ based on the substructuring method can be computed in parallel without much of work.
4. Concluding remarks

In this paper we have attempted to describe the effects of using the substructuring and the force method for solving the constrained minimization problem arising from structural optimization. The special structure of the self-stress matrix \( N \) enables us to save a great deal of work in the computations, and also allows us to use parallel computational techniques.

The algorithm we have described are currently being implemented on current multiprocessor computer. The results of these numerical tests and comparisons will be reported in detail elsewhere.

In fact we also investigated the substructuring method for the solution step of the force method in the context of the incompressible fluid flow. A number of interesting characterizations of this problem and some parallel iterative schemes are reported in [5].

References


A Note on Bounds for the Estimation Error Variance of a Continuous Stream with Stationary Variogram

N.S. Barnett∗ and S.S. Dragomir†

Abstract

In this paper, by the use of an Ostrowski type integral inequality for double integrals, we establish an upper bound for the estimation error variance of a continuous stream with stationary variogram.

1 Introduction

In [1], the authors considered $X(t)$ as defining the quality of a product at time $t$ where $X(t)$ is a continuous time stochastic process which may be non-stationary. Typically, $X(t)$ represents a continuous stream industrial process such as is common in many areas of the chemical industry. The paper was concerned with issues related to sampling the stream with a view to estimating the mean quality characteristic of the flow, $\bar{X}$, over the interval $[0,d]$. Specifically, focus was on obtaining the sampling location, said to be optimal, which minimizes the estimation error variance, $E\left( (\bar{X} - X(t))^2 \right)$, $0 < t < d$.

Given that $t$ is as specified, the problem is to find the value of $t$ (the sampling location) that minimizes $E\left( (\bar{X} - X(t))^2 \right)$. It is shown that for constant stream flows the optimal sampling point is the mid point of $[0,d]$ for situations where the process variogram,

$$ V(u) = \frac{1}{2} E \left[ (X(t) - X(t+u))^2 \right], $$

$$ V(0) = 0, V(-u) = V(u), $$

is stationary (note that variogram stationarity is not equivalent to process stationarity).

The paper continues to consider optimal sampling locations for situations where the stream flow rate varies. The optimal sampling location is seen to depend on both the flow rate function and the form of the process variogram - some examples are given.

In this note, rather than focusing on the optimal sampling point, we focus on the actual value of the estimation error variance itself. In particular, we focus on obtaining

∗School of Communications and Informatics, Victoria University of Technology, PO Box 14428, MCMC, Melbourne, VICTORIA, 8001, AUSTRALIA. E-mail: neil@matilda.vut.edu.au.
†School of Communications and Informatics, Victoria University of Technology, PO Box 14428, MCMC, Melbourne, VICTORIA, 8001, AUSTRALIA. E-mail: sever@matilda.vut.edu.au.
an upper bound for its value. To do this, we use a result obtained in [2], an inequality of the Ostrowski type. For other results of this type see the Chapter XV of [7] and [3-6].

2 An Ostrowski Inequality for Double Integrals

Let \( f : [a, b] \times [c, d] \to \mathbb{R} \) be so that \( f(\cdot, \cdot) \) is integrable on \([a, b] \times [c, d]\), \( f(x, \cdot) \) is integrable on \([c, d]\) for any \( x \in [a, b] \) and \( f(\cdot, y) \) is integrable on \([a, b]\) for any \( y \in [c, d]\), \( f''_{x,y} = \frac{\partial^2 f}{\partial x \partial y} \) exists on \((a, b) \times (c, d)\) and is bounded, i.e.,

\[
\| f''_{s,t} \|_\infty := \sup_{(x,y) \in (a,b) \times (c,d)} \left| \frac{\partial^2 f(x,y)}{\partial x \partial y} \right| < \infty
\]

then we have the inequality:

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

(1)

for all \((x, y) \in [a, b] \times [c, d]\).

For the sake of completeness we give here a short proof of this inequality.

Integrating by parts successively, we have the equality:

\[
\int_a^x \int_c^y (s-a)(t-c) f''_{s,t}(s,t) \, dt \, ds 
\]

\[
= (y-c)(x-a)f(x,y) - (y-c) \int_a^x f(s,y) \, ds 
\]

\[
- (x-a) \int_c^y f(x,t) \, dt + \int_a^x \int_c^y f(s,t) \, ds \, dt.
\]

(2)
Also, by similar computations we have

\[
\int_a^x \int_y^d (s-a)(t-d) f''_{s,t}(s,t) \, ds \, dt = (x-a)(d-y) f(x,y) - (d-y) \int_a^x f(s,y) \, ds
\]

\[
- (x-a) \int_y^d f(x,t) \, dt + \int_a^x \int_c^y f(s,t) \, ds \, dt.
\]  \hspace{1cm} (3)

Now,

\[
\int_x^b \int_y^d (s-b)(t-d) f''_{s,t}(s,t) \, ds \, dt = (d-y)(b-x) f(x,y) - (d-y) \int_x^b f(s,y) \, ds
\]

\[
- (b-x) \int_y^d f(x,t) \, dt + \int_x^b \int_y^c f(s,t) \, ds \, dt.
\]  \hspace{1cm} (4)

and finally

\[
\int_x^b \int_c^y (s-b)(t-c) f''_{s,t}(s,t) \, ds \, dt = (y-c)(b-x) f(x,y) - (y-c) \int_x^b f(s,y) \, ds
\]

\[
- (b-x) \int_c^y f(x,t) \, dt + \int_x^b \int_c^y f(s,t) \, ds \, dt.
\]  \hspace{1cm} (5)
If we add the equalities (2) − (5) we get, in the right membership:

\[
[(y - c)(x - a) + (x - a)(d - y) + (d - y)(b - x) + (y - c)(b - x)]f(x, y)
\]

\[-(d - c) \int_a^x f(s, y) \, ds - (d - c) \int_x^b f(s, y) \, ds - (b - a) \int_c^y f(x, t) \, dt
\]

\[-(b - a) \int_y^d f(x, t) \, dt + \int_x^y \int_a^c f(s, t) \, ds \, dt + \int_x^y \int_a^d f(s, t) \, ds \, dt
\]

\[+ \int_x^y \int_a^d f(s, t) \, ds \, dt + \int_x^y \int_b^y f(s, t) \, ds \, dt\]

\[= (d - c)(b - a)f(x, y) - (d - c) \int_a^b f(s, y) \, ds
\]

\[-(b - a) \int_c^b f(x, t) \, dt + \int_a^b \int_c^d f(s, t) \, ds \, dt.
\]

For the first membership, let us define the kernels: \(p : [a, b]^2 \to \mathbb{R}, q : [c, d]^2 \to \mathbb{R}\) given by:

\[
p(x, s) := \begin{cases} 
  s - a & \text{if } s \in [a, x] \\
  s - b & \text{if } s \in (x, b]
\end{cases}
\]

and

\[
q(y, t) := \begin{cases} 
  t - c & \text{if } t \in [c, y] \\
  t - d & \text{if } t \in (y, d]
\end{cases}.
\]
Now, using these notations, we deduce that the left membership can be represented as:

$$\int_a^b \int_c^d p(x,s) q(y,t) f''_{s,t}(s,t) \, ds \, dt.$$ 

Consequently, we get the identity

$$\int_a^b \int_c^d p(x,s) q(y,t) f''_{s,t}(s,t) \, ds \, dt$$

$$= (d - c) (b - a) f(x,y) - (d - c) \int_a^b f(s,y) \, ds$$

$$- (b - a) \int_c^d f(x,t) \, dt + \int_a^b \int_c^d f(s,t) \, ds \, dt$$

(6)

for all \((x,y) \in [a,b] \times [c,d]\).

Now, using the identity (2.5) we get,

$$\left| \int_a^b \int_c^d f(s,t) \, ds \, dt - [(b - a) \int_c^d f(x,t) \, dt + (d - c) \int_a^b f(s,y) \, ds$$

$$- (d - c) (b - a) f(x,y)] \right|$$

$$\leq \int_a^b \int_c^d |p(x,s)||q(y,t)||f''_{s,t}(s,t)| \, ds \, dt$$

$$\leq \|f''_{s,t}\|_\infty \int_a^b \int_c^d |p(x,s)||q(y,t)| \, ds \, dt.$$

Observe that

$$\int_a^b |p(x,s)| \, ds = \frac{1}{4} (b - a)^2 + \left( x - \frac{a + b}{2} \right)^2$$

and,

$$\int_c^d |q(y,t)| \, dt = \frac{1}{4} (d - c)^2 + \left( y - \frac{c + d}{2} \right)^2.$$

Finally, using (6), we get the desired inequality (1).
3 Bound on Estimation Error Variance. Constant Flow Rate

From [1], using an identity given in [8], it can be shown that

\[
E \left[ (\bar{X} - X(t))^2 \right] = -\frac{1}{d^2} \int_0^d \int_0^d V(v-u)dv + \frac{2}{d^2} \left\{ \int_0^t V(u)du + \int_0^{d-t} V(u)du \right\}. \tag{7}
\]

Assume that \(V\) is twice differentiable on \((-d, d)\) and having the second derivative \(V''\) bounded on that interval.

Applying inequality (1) for the mapping \(f(u, v) = V(v-u)\) we can state the inequality

\[
\left| \int_0^d \int_0^d V(v-u)dv - [d \int_0^d V(v-x)dv + d \int_0^d V(y-u)du - d^2V(y-x)] \right| \leq \frac{1}{4} d^2 + (x - \frac{d}{2})^2 \|V''\|_\infty
\]

for all \(x, y \in [0, d]\).

Let \(x = y = t\). Then we get

\[
\left| \int_0^d \int_0^d V(v-u)dv - [d \int_0^d V(v-t)dv + d \int_0^d V(t-u)du] \right| \leq \frac{1}{4} d^2 + (t - \frac{d}{2})^2 \|V''\|_\infty \tag{8}
\]

as \(V(0) = 0\).

Now, observe that

\[
\int_0^d V(v-t)dv = \int_0^t V(u)du + \int_0^{d-t} V(u)du
\]

and

\[
\int_0^d V(t-u)du = \int_0^t V(u)du + \int_0^{d-t} V(u)du.
\]

By the inequality (8) we get that

\[
\left| \int_0^d \int_0^d V(v-u)dv - 2d \int_0^t V(v)dv + \int_0^{d-t} V(v)dv \right| \leq \frac{1}{4} d^2 + (t - \frac{d}{2})^2 \|V''\|_\infty
\]
and dividing by \( d^2 \)

\[
\left| \frac{1}{d^2} \int_0^d \int_0^d V(v - u)du dv - \frac{2}{d} \int_0^d V(v)dv + \int_0^{d-t} V(v)dv \right| \\
\leq \left[ \frac{1}{4} + \frac{(t - \frac{d}{2})^2}{d^2} \right]^2 d^2 \| V'' \|_\infty
\]

Using the equation (7) we conclude that the following inequality for the variance 
\[ E \left[ (\bar{X} - X(t))^2 \right] \] holds

\[ E \left[ (\bar{X} - X(t))^2 \right] \leq \left[ \frac{1}{4} + \frac{(t - \frac{d}{2})^2}{d^2} \right]^2 d^2 \| V'' \|_\infty. \] (9)

Note that the best inequality we can get from (9) is that one for which \( t = t_o = \frac{d}{2} \) giving the bound

\[ E \left[ (\bar{X} - X(t_o))^2 \right] \leq \frac{d^2}{16} \| V'' \|_\infty. \]

**Remark 1** It should be noted that this result requires double differentiability of \( V \) in \((-d, d)\) and that this condition does not hold for the case of a linear variogram, i.e.,

\[ V(u) = a \mid | u |, u \in R. \]

**REFERENCES**


ON SIMPSON’S QUADRATURE FORMULA FOR 
DIFFERENTIABLE MAPPINGS WHOSE DERIVATIVES 
BELONG TO $L^p$ – SPACES AND APPLICATIONS

Sever Silvestru Dragomir

Abstract

An estimation of remainder for Simpson’s quadrature formula for differentiable mappings whose derivatives belong to $L^p$–spaces and applications in theory of special means (logarithmic mean, identric mean etc...) are given.

1 INTRODUCTION

The following inequality is well known in the literature as the Simpson’s inequality :

$$\left| \int_a^b f(x)dx - \frac{b-a}{3} \left[ \frac{f(a) + f(b)}{2} + 2f \left( \frac{a+b}{2} \right) \right] \right| \leq \frac{1}{2880} \| f^{(4)} \|_\infty (b-a)^5 \quad (1.1)$$

where the mapping $f : [a, b] \rightarrow \mathbb{R}$ is supposed to be four time differentiable on the interval $(a, b)$ and having the fourth derivative bounded on $(a, b)$, that is

$$\| f^{(4)} \|_\infty := \sup_{x \in (a, b)} | f^{(4)}(x) | < \infty.$$

Now, if we assume that $I_h : a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ is a partition of the interval $[a, b]$ and $f$ is as above, then we have the Simpson’s quadrature formula:

$$\int_a^b f(x)dx = A_S(f, I_h) + R_S(f, I_h) \quad (1.2)$$

where $A_S(f, I_h)$ is the Simpson’s rule

$$A_S(f, I_h) =: \frac{1}{6} \sum_{i=0}^{n-1} [f(x_i) + f(x_{i+1})]h_i + \frac{2}{3} \sum_{i=0}^{n-1} f \left( \frac{x_i + x_{i+1}}{2} \right)h_i \quad (1.3)$$

\text{AMS Subject Classification: 26D16, 26D99, 41A55.}

Key Words and Phrases: Simpson’s formula, Special Means.
and the remainder term $R_S(f, I_h)$ satisfies the estimation

$$\left| R_S(f, I_h) \right| \leq \frac{1}{2880} \left\| f^{(4)} \right\|_{\infty} \sum_{i=0}^{n-1} h_i^5$$

(1.4)

where $h_i := x_{i+1} - x_i$ for $i = 0, ..., n - 1$.

When we have an equidistant partitioning of $[a, b]$ given by

$$I_n : x_i := a + \frac{b - a}{n} i, i = 0, ..., n$$

(1.5)

then we have the formula

$$\int_a^b f(x)dx = A_{S,n}(f) + R_{S,n}(f)$$

(1.6)

where

$$A_{S,n}(f) := \frac{b - a}{6n} \sum_{i=0}^{n-1} [f(a + \frac{b - a}{n} i) + f(a + \frac{b - a}{n} (i + 1))]$$

$$+ \frac{2(b - a)}{3n} \sum_{i=0}^{n-1} f(a + \frac{b - a}{n} \cdot \frac{2i + 1}{2})$$

(1.7)

and the remainder satisfies the estimation

$$\left| R_{S,n}(f) \right| \leq \frac{1}{2880} \cdot \frac{(b - a)^5}{n^4} \left\| f^{(4)} \right\|_{\infty} .$$

(1.8)

In the recent paper [1] the author proved the following result for lipschitzian mappings

THEOREM 1.1. Let $f : [a, b] \rightarrow R$ be an $L$-lipschitzian mapping on $[a, b]$. Then we have the inequality

$$\left| \int_a^b f(x)dx - \frac{b-a}{3} \left[ \frac{f(a) + f(b)}{2} + 2f\left(\frac{a+b}{2}\right) \right] \right| \leq \frac{5}{36} L(b-a)^2 .$$

(2.1)

The following corollary is useful in practice:

COROLLARY 1.2. Suppose that $f : [a, b] \rightarrow R$ is a differentiable mapping whose derivative is bounded on $(a, b)$, i.e.,

$$\left\| f' \right\|_{\infty} := \sup_{x \in (a, b)} | f'(x) | < \infty.$$
Then we have the inequality

\[ | \int_a^b f(x)dx - \frac{b-a}{3}\left[ \frac{f(a) + f(b)}{2} + 2f\left(\frac{a+b}{2}\right) \right] | \leq \frac{5}{36} \|f'\|_{\infty}(b-a)^2. \quad (2.5) \]

In the paper [2], S.S. Dragomir proved a version of Simpson’s inequality for mappings with bounded variation as follows:

**THEOREM 1.3.** Let \( f : [a, b] \to \mathbb{R} \) be a mapping with bounded variation on \([a, b]\). Then we have the inequality

\[ | \int_a^b f(x)dx - \frac{b-a}{3}\left[ \frac{f(a) + f(b)}{2} + 2f\left(\frac{a+b}{2}\right) \right] | \leq \frac{1}{3}(b-a)V_{ab}^b(f) \quad (2.1) \]

where \( V_{ab}^b(f) \) denotes the total variation of \( f \) on the interval \([a, b]\).

The constant \( \frac{1}{3} \) is the best possible one.

The following corollary is useful in practice

**COROLLARY 1.4.** Suppose that \( f : [a, b] \to \mathbb{R} \) is a differentiable mapping whose derivative is integrable on \((a, b)\), i.e.,

\[ \|f'\|_1 := \int_a^b |f'(x)| \, dx < \infty. \]

Then we have the inequality

\[ | \int_a^b f(x)dx - \frac{b-a}{3}\left[ \frac{f(a) + f(b)}{2} + 2f\left(\frac{a+b}{2}\right) \right] | \leq \frac{1}{3}\|f'\|_1(b-a)^2. \quad (2.5) \]

For some other integral inequalities see the recent book [3].

The main aim of this paper is to point out some bounds of the remainder in terms of \( p \)-norm of the derivative \( f' \) and apply them for composite quadrature formulae and for special means.
2 SIMPSON’S INEQUALITY IN TERMS OF p-NORMS

The following result holds:

**THEOREM 2.1.** Let \( f : [a, b] \to \mathbb{R} \) be a differentiable mapping on \((a, b)\) whose derivative belongs to \( L^p(a, b) \). Then we have the inequality

\[
\left| \int_a^b f(x)dx - \frac{b - a}{3} \left[ \frac{f(a) + f(b)}{2} + 2f\left(\frac{a + b}{2}\right) \right] \right|
\leq \frac{1}{6} \left[ \frac{2^{q+1} + 1}{3(q + 1)} \right]^{\frac{1}{q}} (b - a)^{1 + \frac{1}{q}} \left\| f' \right\|_p \tag{2.1}
\]

where \( \frac{1}{p} + \frac{1}{q} = 1, p > 1 \).

**Proof.** Using the integration by parts formula we have:

\[
\int_a^b s(x)f'(x)dx = \frac{b - a}{3} \left[ \frac{f(a) + f(b)}{2} + 2f\left(\frac{a + b}{2}\right) \right] - \int_a^b f(x)dx \tag{2.2}
\]

where

\[
s(x) := \begin{cases} 
  x - \frac{5a + b}{6}, & x \in [a, \frac{a + b}{2}) \\
  x - \frac{a + 5b}{6}, & x \in [\frac{a + b}{2}, b] 
\end{cases}
\]

Indeed,

\[
\int_a^b s(x)f'(x)dx = \int_a^{\frac{a + b}{2}} (x - \frac{5a + b}{6}) f'(x)dx + \int_{\frac{a + b}{2}}^b (x - \frac{a + 5b}{6}) f'(x)dx
\]

\[
= [(x - \frac{5a + b}{6})f(x)]_{\frac{a + b}{2}}^{\frac{a + b}{2}} + [(x - \frac{a + 5b}{6})f(x)]_{\frac{a + b}{2}}^b - \int_a^b f(x)dx
\]

\[
= \frac{b - a}{3} \left[ \frac{f(a) + f(b)}{2} + 2f\left(\frac{a + b}{2}\right) \right] - \int_a^b f(x)dx
\]

and the identity is proved.

Applying Hölder’s integral inequality we get

\[
\left| \int_a^b s(x)f'(x)dx \right| \leq \left( \int_a^b |s(x)|^q dx \right)^{\frac{1}{q}} \left\| f' \right\|_p . \tag{2.3}
\]
Let us compute
\[
\int_a^b |s(x)|^q \, dx = \int_a^{a+b/2} |x - \frac{5a+b}{6}|^q \, dx + \int_{a+b/2}^b |x - \frac{a+5b}{6}|^q \, dx
\]
\[
= \int_a^{a+b/6} (\frac{5a+b}{6} - x)^q \, dx + \int_{a+b/6}^{a+b/2} (x - \frac{5a+b}{6})^q \, dx
\]
\[
+ \int_{a+b/2}^{a+5b/6} (\frac{a+5b}{6} - x)^q \, dx + \int_{a+5b/6}^b (x - \frac{a+5b}{6})^q \, dx
\]
\[
= \frac{1}{q+1} \left[ (\frac{5a+b}{6} - x)^q \frac{a+b}{a-b} + (x - \frac{5a+b}{6})^q \frac{a+b}{a-b} \right]_a^{\frac{a+b}{6}}
\]
\[
= \frac{1}{q+1} \left[ (\frac{5a+b}{6} - a)^q + (\frac{a+b}{2} - \frac{5a+b}{6})^q \right]
\]
\[
+ (\frac{a+5b}{6} - \frac{a+b}{2} )^q + (b - \frac{a+5b}{6} )^q
\]
\[
= \frac{(2^{q+1} + 1)(b-a)^{q+1}}{3(q+1)6^q}
\]

Now, using the inequality (2.3) and the identity (2.2) we deduce the desired result (2.1).

The following corollary for Simpson’s composite formula holds:

**COROLLARY 2.2.** Let \( f \) and \( I_h \) be as above. Then we have Simpson’s rule (1.2) and the remainder \( R_S(f, I_h) \) satisfies the estimation
\[
| R_S(f, I_h) | \leq \frac{1}{6} \frac{2^{q+1} + 1}{3(q+1)} \| f' \|_p \left( \sum_{i=0}^{n-1} h_i^{1+q} \right)^{\frac{1}{q}}. (2.4)
\]

**Proof.** Apply Theorem 2.1 on the interval \([x_i, x_{i+1}]\) \((i = 0, \ldots, n - 1)\) to get
\[
| \int_{x_i}^{x_{i+1}} f(x) \, dx - \frac{h_i}{3} \left[ \frac{f(x_i) + f(x_{i+1})}{2} + 2f\left(\frac{x_i + x_{i+1}}{2}\right) \right] |
\]

\[
\leq \frac{1}{6} \left[ \frac{2^{q+1} + 1}{3(q + 1)} \right] \frac{1}{\eta} \left\| f' \right\|_p h_i^{1+\frac{1}{q}} \left( \int_{x_i}^{x_{i+1}} \left| f'(t) \right|^p \, dt \right)^{\frac{1}{p}}.
\]

Summing the above inequalities over \( i \) from 0 to \( n - 1 \), using the generalized triangle inequality and Hölder’s discrete inequality, we get

\[
| R_S(f, I_h) | \leq \sum_{i=0}^{n-1} | \int_{x_i}^{x_{i+1}} f(x) \, dx - \frac{h_i}{3} \left[ \frac{f(x_i) + f(x_{i+1})}{2} + 2f\left(\frac{x_i + x_{i+1}}{2}\right) \right] |
\]

\[
\leq \frac{1}{6} \left[ \frac{2^{q+1} + 1}{3(q + 1)} \right] \frac{1}{\eta} \sum_{i=0}^{n-1} h_i^{1+\frac{1}{q}} \left( \int_{x_i}^{x_{i+1}} \left| f'(t) \right|^p \, dt \right)^{\frac{1}{p}}
\]

\[
\leq \frac{1}{6} \left[ \frac{2^{q+1} + 1}{3(q + 1)} \right] \frac{1}{\eta} \left( \sum_{i=0}^{n-1} (h_i^{1+\frac{1}{q}})^q \right)^{\frac{1}{q}} \times \left( \sum_{i=0}^{n-1} \left( \frac{\int_{x_i}^{x_{i+1}} \left| f'(t) \right|^p \, dt}{h_i^{1+\frac{1}{q}}} \right)^{\frac{1}{p}} \right)^{\frac{1}{q}}
\]

\[
= \frac{1}{6} \left[ \frac{2^{q+1} + 1}{3(q + 1)} \right] \frac{1}{\eta} \left\| f' \right\|_p \left( \sum_{i=0}^{n-1} h_i^{1+q} \right)^{\frac{1}{q}}
\]

and the corollary is proved. ■

The case of equidistant partitioning is embodied in the following corollary:

COROLLARY 2.4. Let \( f \) be as above and if \( I_n \) is an equidistant partitioning of \([a, b]\), then we have the estimation

\[
| R_{S,n}(f) | \leq \frac{1}{6n} \left[ \frac{2^{q+1} + 1}{3(q + 1)} \right] \frac{1}{\eta} (b - a)^{1+\frac{1}{q}} \left\| f' \right\|_p.
\]

Remark 2.5. If we want to approximate the integral \( \int_a^b f(x) \, dx \) by Simpson’s formula \( A_{S,n}(f) \) with an accuracy less than \( \varepsilon > 0 \), we need at least \( n_\varepsilon \in \mathbb{N} \) points for the division \( I_n \), where

\[
n_\varepsilon := \left[ \frac{1}{6\varepsilon} \left( \frac{2^{q+1} + 1}{3(q + 1)} \right)^{\frac{1}{q}} (b - a)^{1+\frac{1}{q}} \right] \left\| f' \right\|_p + 1
\]
and \([r]\) denotes the integer part of \(r \in R\).

Comments 2.6. If the mapping \(f : [a, b] \rightarrow R\) is neither four time differentiable nor the fourth derivative is bounded on \((a, b)\), then we can not apply the classical estimation in Simpson’s formula using the fourth derivative. But if we assume that \(f' \in L_p(a, b)\), then we can use instead the formula (2.4).

We give here a class of mappings whose first derivatives belong to \(L_p(a, b)\) but having the fourth derivatives unbounded on the given interval.

Let \(f_s : [a, b] \rightarrow R, f_s(x) := (x-a)^s \) where \(s \in (3, 4)\). Then obviously

\[
f_s'(x) := s(x-a)^{s-1}, x \in (a, b)
\]

and

\[
f_s^{(4)}(x) = \frac{s(s-1)(s-2)(s-3)}{(x-a)^{1-s}}, x \in (a, b).
\]

It is clear that \(\lim_{x \to a+} f_s^{(4)}(x) = +\infty\) but \(\|f_s'\|_p = s \frac{(b-a)^{s-1+s-1}}{(s-1)p+1} < \infty\).

3 APPLICATIONS FOR SPECIAL MEANS

Let us recall the following means:

1. Arithmetic mean

\[
A = A(a, b) := \frac{a + b}{2}, a, b \geq 0;
\]

2. Geometric mean

\[
G = G(a, b) := \sqrt{ab}, a, b \geq 0;
\]

3. Harmonic mean

\[
H = H(a, b) := \frac{2}{\frac{1}{a} + \frac{1}{b}}, a, b > 0;
\]

4. Logarithmic mean
\[ L = L(a, b) := \frac{b - a}{\ln b - \ln a}, a, b > 0, a \neq b; \]

5. **Identric mean**

\[ I = I(a, b) := \frac{1}{e} \left( \frac{b}{a} \right)^{\frac{1}{\ln a}}, a, b > 0, a \neq b; \]

6. **p-Logarithmic mean**

\[ S_p = S_p(a, b) := \left[ \frac{b^{p+1} - a^{p+1}}{(p+1)(b-a)} \right]^{\frac{1}{p}}, p \in \mathbb{R}\{−1, 0\}, a, b > 0, a \neq b. \]

It is well known that \( S_p \) is monotonous nondecreasing over \( p \in \mathbb{R} \) with \( S_{−1} : = L \) and \( S_0 := I \). In particular, we have the following inequalities

\[ H \leq G \leq L \leq I \leq A. \quad (3.1) \]

In what follows, by the use of Theorem 2.1, we point out some new inequalities for the above means.

1. Let \( f : [a, b] \to R \ (0 < a < b), f(x) = x^s, s \in \mathbb{R}\{−1, 0\}. \) Then

\[ \frac{1}{b-a} \int_a^b f(x)dx = S_s^s(a, b), f\left( \frac{a+b}{2} \right) = A_s^s(a, b), \frac{f(a) + f(b)}{2} = A(a^s, b^s) \]

and \( \|f\|_p = |s| S_{s(−1)p}^s(b-a)^\frac{1}{p}. \)

Using the inequality (2.1) we get

\[ |S_s^s(a, b) - \frac{1}{3} A(a^s, b^s) - \frac{2}{3} A_s^s(a, b)| \leq \frac{1}{6} \left( \frac{2q+1}{3(q+1)} \right)^\frac{1}{q} |s| S_{s(−1)p}^s(a, b)(b-a), (3.2) \]

where \( \frac{1}{p} + \frac{1}{q} = 1, p > 1. \)

2. Let \( f : [a, b] \to R \ (0 < a < b), f(x) = \frac{1}{x}. \) Then

\[ \frac{1}{b-a} \int_a^b f(x)dx = L^{-1}(a, b), f\left( \frac{a+b}{2} \right) = A^{-1}(a, b), \]

\[ \frac{f(a) + f(b)}{2} = H^{-1}(a, b) \text{ and } \|f\|_p = S_{−2p}^{-2}(a, b)(b-a)^\frac{1}{p}. \]

Using the inequality (2.1) we get
| 3HA − LA − 2LH | ≤ \frac{1}{2}AHL\left(\frac{2q+1}{3(q+1)}\right)^{\frac{1}{q}}S_{-p}^{-1}(b-a) \quad (3.3).

3. Let \( f : [a, b] \rightarrow \mathbb{R} \) \((0 < a < b)\), \( f(x) = \ln x \). Then

\[
\frac{1}{b-a} \int_a^b f(x)dx = \ln I(a, b), f\left(\frac{a+b}{2}\right) = \ln A(a, b),
\]

\[
\frac{f(a) + f(b)}{2} = \ln A(a, b) \quad \text{and} \quad \|f'\|_p = S_{-p}^{-1}(a, b)(b-a)^{\frac{1}{p}}.
\]

Using the inequality (2.1) we get

\[
|\ln\frac{I}{G^{1/3}A^{2/3}}| \leq \frac{1}{6}\left(\frac{2q+1}{3(q+1)}\right)^{\frac{1}{q}}S_{-p}^{-1}(a, b)(b-a). \quad (3.4)
\]

REFERENCES


SINGULAR INTEGRAL EQUATIONS AND UNDERDETERMINED SYSTEMS

SEKI KIM

DEPARTMENT OF MATHEMATICS,
SUNGKYUNKWAN UNIVERSITY, SUWON 440-746, KOREA

skkim@math.skku.ac.kr

Abstract. In this paper the linear algebraic system obtained from a singular integral equation with variable coefficients by a quadrature-collocation method is considered. We study this underdetermined system by means of the Moore-Penrose generalized inverse. Convergence in compact subsets of \([-1, 1]\) can be shown under some assumptions on the coefficients of the equation.

Key words. Singular integral equation, underdetermined system, generalized inverse

AMS subject classification. 65R20

1. Introduction. @ <1111111111 <<

The purpose of this study is the investigation of some properties of the underdetermined linear algebraic system obtained from singular integral equations (SIE's) by direct quadrature-collocation based on nonclassical nodes. We are able to characterize some properties of the system.

Contrary to what happens for Fredholm equations, in SIE's because of the presence of the Cauchy principal value singularity the \(n\) quadrature nodes used to discretize the singular integral cannot be employed to collocate the functional equation resulting from the previous operation. Other nodes are thus necessary. If this other set of nodes is chosen appropriately, it turns out that they need to be the \(n - \kappa\) zeros of a certain orthogonal polynomial of a family related to the one that provides the quadrature nodes.

Recent investigations, see [15, 14, 8], have dealt with the problem of studying the effect of a "suboptimal" choice of the second set of nodes. In case of constant coefficients, we replace appropriate Gauss-Jacobi nodes with the more easily generated Chebyshev nodes [15, 14].

For variable coefficients, the quadrature and collocation nodes arise from nonclassical families of orthogonal polynomials, [4], and are difficult to construct. A recently proposed solution scheme replaces these nonclassical families by standard Gauss-Jacobi orthogonal polynomials, [8]. In both cases, the value of the unknown function at the collocation set is nonzero and hence cannot be ignored. Thus extra unknowns arise using this procedure. In order to obtain a square system, another formula needs to be used. It can be a quadrature of a different type, applied however using the same
nodes, or an interpolatory formula relating the value of the unknown function at the two sets of nodes. These procedures have the drawback that they double the size of the discretized linear algebraic system.

If we want to use a "standard approach", there is the need of using nonclassical orthogonal polynomials, whose efficient calculation is not yet available. An alternative is to try to avoid the use of these polynomials by looking only at the asymptotics of the equation, i.e. at the singular endpoint behavior, expressed by the constants $\alpha$ and $\beta$ defined in the next paragraph. The disadvantage however is the fact that for variable coefficient equations, the quantity $- (\alpha + \beta)$ is not an integer, the index of the equation, as it happens in the case of constant coefficients. We try here a different approach, by avoiding to look directly at the asymptotics, retaining however the use of classical Jacobi polynomials. Under suitable assumptions on the coefficients, we are able to show convergence in compact subsets of $[-1, 1]$, but the price we pay consists in the convergence rate being affected by the use of this "imprecise" endpoint singularity information and by the growth of the error constant.

The note is organized as follows. In the next section we give the mathematical description of the problem. Section 3 is devoted to the presentation of the numerical method, and the last section contains the results of the analysis.

2. Preliminaries. We consider here the dominant singular integral equation with variable coefficients

\begin{equation}
(1) \quad a(x) \phi(x) + \frac{b(x)}{\pi} \int_{-1}^{1} \frac{\phi(t)}{t-x} dt = f(x) \quad -1 < x < 1.
\end{equation}

Let $H_{\mu}[-1, 1]$ denote the class of Hölder continuous functions of exponent $\mu$ on $[-1, 1]$, i.e. the functions $y(x)$ that satisfy the Hölder condition ($H$-condition)

$$|y(x) - y(t)| \leq C^{0} |x - t|^{\mu},$$

for a suitable constant $C^{0} > 0$, and $0 < \mu \leq 1$. We assume that the coefficients are real valued functions on $H_{\mu}[-1, 1]$, satisfying $r^{2}(x) = a^{2}(x) + b^{2}(x) > 0$, for every $x \in [-1, 1]$. By choosing a continuous path for the function

$$\log \frac{a(x) - ib(x)}{a(x) + ib(x)},$$

it is then possible to find integers $M$ and $N$ such that the quantities $\alpha_{0}$ and $\beta_{0}$ given by

$$\alpha_{0} = \frac{1}{2\pi i} \log \frac{a(1) - ib(1)}{a(1) + ib(1)} + M, \quad \beta_{0} = -\frac{1}{2\pi i} \log \frac{a(-1) - ib(-1)}{a(-1) + ib(-1)} + N,$$

satisfy $|\alpha_{0}|, |\beta_{0}| < 1$, for more details see [4]. The index of the equation is then defined as $\chi = - (M + N)$. The fundamental function of the problem can be represented as

$$Z(x) = (1-x)^{\alpha_{0}} (1+x)^{\beta_{0}} \Omega(x),$$
where $\Omega(x)$ is a positive smooth function on $[-1, 1]$. Classically, a new unknown function $\varphi(x)$ is defined by
$$
\phi(x) = Z(x) \varphi(x).
$$
The singular operator in (1) transforms any function $\phi(x)$ locally satisfying the $H$-condition, into a new function $\zeta(x)$ which also locally satisfies the $H$-condition; hence the right hand side $f(x)$ must also locally satisfy the $H$-condition in order that the solution satisfies the $H$-condition, [11]. On the other hand, the solution $\phi(x)$ locally satisfies the $H$-condition if $f(x)$ is assumed to locally satisfy the $H$-condition [11]. Since both $a(x)$, $b(x) \in H_{\mu}[-1, 1]$, then also $\varphi(x) \in H_{\mu}[-1, 1]$. One more final assumption is made on the coefficient $b(x)$, namely that it is a polynomial. More general conditions leading to this situation are discussed in [5], [6] and [1].

In the standard approach, use is made of families of nonclassical orthogonal polynomials with respect to the weight function $Z(x)$. The related quadrature nodes and weights are difficult to compute. A major difference with respect to equations with constant coefficients consists in the fact that $- (a_0 + \beta_0) \neq \chi$.

Our basic approach here is still to try to exploit properties of classical Jacobi polynomials, at the expense of having to deal with less smooth unknown functions. We indeed rewrite the function $\phi(x)$ in terms of a suitable Jacobi weight $\rho(x)$, to be specified below, in order to express explicitly the singular behavior at the endpoints in a “classical” manner, and a new unknown function $y^*(x)$ such that
$$
\phi(x) = \rho(x) y^*(x).
$$
The introduction of the function $\rho(x)$ tries to capture the singular endpoint behavior of the solution, but not completely. Specifically, the exponents $\alpha$ and $\beta$ are chosen according to the following rule. Since $(a_0, \beta_0) \in (-1, 1) \times (-1, 1)$, we orthogonally project $(a_0, \beta_0)$ onto the line $y = -x + \kappa$, $\kappa \in (-1, 0, 1)$, to find the point $(\alpha, \beta)$. The integer $\kappa$ is chosen so that $\kappa = -1$, if $a_0$, $\beta_0 > 0$; $\kappa = 1$ if $a_0$, $\beta_0 < 0$, and $\kappa = 0$ otherwise. To be specific, we have
$$
\alpha = \frac{1}{2} (-\kappa + a_0 - \beta_0), \quad \beta = \frac{1}{2} (-\kappa + \beta_0 - a_0), \quad \rho(x) = (1 - x)^\alpha (1 + x)^\beta.
$$
Notice that with this choice,
$$
|a_0 - \alpha| < \frac{1}{2}, \quad |\beta_0 - \beta| < \frac{1}{2}, \quad \max(\alpha, \beta) \geq -\frac{1}{2}.
$$
It follows that
$$
y^*(x) = (1 - x)^{a_0 - \alpha} (1 + x)^{\beta_0 - \beta} u(x),
$$
and if $u$ is smooth,
$$
y^* \in H_{\lambda}[-1, 1], \quad \lambda = \min(|a_0 - \alpha|, |\beta_0 - \beta|).
$$
Recall indeed the inequality $|x^\nu - y^\nu| \leq |x - y|^\nu$, $-1 < \nu < 1$, see [10], p. 57.

The integral in (1) can be discretized by a classical Gaussian quadrature with weight $\rho(x)$. Let $P_n^{(\alpha,\beta)}(x)$ denote the Jacobi polynomial of degree $n$ relative to the weight function $\rho(x)$, and $P_{n-k}^{(-\alpha,-\beta)}(x)$ the one of degree $n-k$ relative to the reciprocal weight $\rho^{-1}(x)$. Let also $t_i$ represent the zeros of $P_n^{(\alpha,\beta)}(x)$ and $s_j$ those of $P_{n-k}^{(-\alpha,-\beta)}(x)$. We will need the following

**Lemma 1.** The Lebesgue constant for interpolation to $y^\nu$ on the zeros of the Jacobi polynomial grows like

$$n^{\max(\alpha,\beta)+\frac{1}{2}}, \text{ for } \max(\alpha,\beta) > -\frac{1}{2}, \log n, \text{ for } \max(\alpha,\beta) = -\frac{1}{2}.$$

**Proof.** See Theorem 14.4 of [13]. By (5) and (3) its assumptions hold. \qed

If $\chi \leq 0$, the theory of singular integral equations states that the solution is unique, but for existence, in case $\chi < 0$, the right hand side has to satisfy $\chi$ orthogonality conditions. If $\chi > 0$ instead, the solution is not unique. To determine it uniquely, we need extra conditions, which we take as

$$\frac{1}{\pi} \int_{-1}^1 P_k^{(\alpha,\beta)}(t) \phi(t) \, dt = K_k, \quad k = 0, \ldots, \chi - 1. \quad (6)$$

Equation (1) can be rewritten as follows

$$a(x) \rho(x) y^\nu(x) + \frac{b(x)}{\pi} \int_{-1}^1 \frac{\rho(t) y^\nu(t)}{t-x} \, dt = f(x). \quad (7)$$

To evaluate the singular integral in (7), we will use Hunter’s method [7]. Let

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

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

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

It can be expressed explicitly by means of the hypergeometric function, [13], but in this case using (2.1) of [9] it reduces to

$$\frac{1}{\pi} \int_{-1}^1 \frac{P_n^{(\alpha,\beta)}(t) \rho(t)}{t-x} \, dt = \cot(\pi\alpha) P_n^{(\alpha,\beta)}(x) \rho(x) - \frac{2^{-\kappa}}{\sin(\pi\alpha)} P_{n-k}^{(-\alpha,-\beta)}(x). \quad (8)$$

Then Hunter’s method takes the form

$$Q_n^*(y^\nu, x) = \sum_{i=1}^n \frac{w_i y^\nu(t_i)}{t_i-x} + \frac{\psi_n^{(\alpha,\beta)}(x) y^\nu(x)}{P_n^{(\alpha,\beta)}(x)}. \quad (9)$$
The singular integral is thus replaced by the above quadrature

\[ \int_{-1}^{1} \frac{\rho(t)y'(t)}{t-x} dt = Q_n(y^*, x) + \epsilon_G(x), \quad \text{for } x \in (-1, 1) \]

and \( \epsilon_G \) represents the quadrature error, and the quadrature weights have the explicit expressions

\[ w_i = \int_{-1}^{1} \rho(t) \frac{P_n^{(\alpha, \beta)}(t)}{(t-t_i)P_n^{(\alpha, \beta)'}(t_i)} dt = 2^{\alpha+\beta} \frac{\Gamma(n+\alpha)\Gamma(n+\beta)}{\Gamma(n)\Gamma(n+\alpha+\beta+1)} \frac{2n+\alpha+\beta}{P_n^{(\alpha, \beta)'}(t_i)P_n^{(\alpha, \beta)}(t_i)}, \quad i = 1, 2, \ldots, n \]

3. Discretization. Recall that \( s_j \neq t_i, j = 1, \ldots, n-\kappa, \) are the zeros of \( P_{n-\kappa}^{(\alpha, -\beta)}(x); \)
using (8) let us define for \( j = 1, \ldots, n-\kappa, \)

\[ d_j \equiv a(s_j) + b(s_j) \frac{\psi_n^{(\alpha, \beta)}(s_j)}{\pi P_n^{(\alpha, \beta)}(s_j)} \equiv \rho(s_j) \left[ a(s_j) + b(s_j) \cot(\pi\alpha) \right]. \]

From this, (10) and (9), collocating at the node points, we have

\[ d_j y^*(s_j) + \frac{b(s_j)}{\pi} \sum_{i=1}^{n} w_i y'(t_i) + \epsilon_{G,k} = f(s_j), \]

In case of positive index, the normalization conditions can also be discretized using standard Gauss-Jacobi quadrature over the same nodes and with the same weights. Thus

\[ \sum_{i=1}^{n} w_i P_k^{(\alpha, \beta)}(t_i) y'(t_i) + \epsilon_{G,k}^0 = K_k, \quad k = 0, \ldots, \chi - 1, \]

where \( \epsilon_{G,k}^0 \) are the components of the error \( \underline{\epsilon}_G^0 \) of standard Gauss-Jacobi quadrature.

Let us introduce the unknown vector

\[ y = [y(s_1), \ldots, y(s_{n-\kappa}), y(t_1), \ldots, y(t_n)]^T \]

approximating the exact solution vector

\[ y^* = [y^*(s_1), \ldots, y^*(s_{n-\kappa}), y^*(t_1), \ldots, y^*(t_n)]^T. \]

In other words, we denote by \( y^*(x) \) the exact solution of the original equation (7), together with (6), by \( y^* \) the vector of dimension \( 2n - \kappa \) of the function values of the exact solution \( y^* \) at the collocation points and the quadrature nodes. The vector \( y \) of dimension \( 2n - \kappa \) is the solution of the system (16) below.

After dropping the error term the discretized linear algebraic system of size \((n - \kappa + \chi) \times (2n - \kappa)\) can be written in block form as

\[ M y = \begin{bmatrix} D & A \end{bmatrix} y = f \]
where \( f = [f(s_1), \ldots, f(s_{n-K})]^t \). Here \( D \) is an \((n - \kappa + \chi) \times (n - \kappa)\) matrix with a special structure, where the only nonzero elements are
\[
(D)_{j,j} = d_j, \quad \text{for } j = 1, \ldots, n - \kappa.
\]
The matrix \( A \) has instead the following entries
\[
A_{ij} = \begin{cases} \frac{b(s_i)w_j}{s(t_j - s_i)} & i = 1, \ldots, n - \kappa \\ w_jP_{k}^{(\alpha, \beta)}(t_j) & i = n + k + 1 - \kappa, \quad k = 0, \ldots, \chi - 1 \end{cases} \quad j = 1, 2, \ldots, n.
\]
From the original equation it follows
\[
\tilde{M}y^* + \zeta_G = \underline{f},
\]
where \( \zeta_G \) denotes the consistency error vector, \( \zeta_G = [\epsilon_{G,1}, \ldots, \epsilon_{G,n-\chi}, \epsilon_{G,0}, \ldots, \epsilon_{G,\chi-1}]^t \).
Let \( I_n \) denote the \( n \times n \) identity matrix, and let us define the square matrices of sizes \( n \), and \( 2n - \kappa \) respectively
\[
N = \text{diag}(w_j^{-1}), \quad j = 1, \ldots, n; \quad B = \text{diag}\left(I_{n-\kappa}, N^\uparrow\right).
\]
Let us introduce
\[
M = \tilde{M}B = \begin{bmatrix} D & AN^\uparrow \end{bmatrix}.
\]
Rewrite the system (17) as follows
\[
MB^{-1}y^* + \zeta_G = \underline{f}
\]
and the system (16) as
\[
MB^{-1}y = \underline{f}.
\]
This is a rectangular, underdetermined system. On taking its Moore-Penrose generalized inverse, we find a solution \( y \) satisfying (16) in the sense of minimizing
\[
(B^{-1}y)^t(B^{-1}y)
\]
\[
y = BM^+\underline{f}.
\]
We can also define \( y^* \) as a solution of (19) in the sense of minimizing \((B^{-1}y^*)^t(B^{-1}y^*)\)
\[
y^* = BM^+(\underline{f} - \zeta_G).
\]
Define the error as \( e = y^* - y \). Subtracting (21) from (22)
\[
e = -BM^+\zeta_G.
\]
To determine convergence, we need estimates of the terms in the right hand side.
\[
\|e\| \leq \|B\|\|M^+\|\|\zeta_G\|.
\]
\section*{4. Estimates and Main Result.}

For the consistency error, and for some discussion of the norm of the Moore-Penrose generalized inverse, we need the results of \cite{2}. Their quadrature however differs from (9). It is

\begin{equation}
Q_n(y^*, x) = \sum_{i=1}^{n} W_i y^*(t_i) = \sum_{i=1}^{n} \frac{\psi_n^{(\alpha, \beta)}(t_i) - \psi_n^{(\alpha, \beta)}(x)}{P_n^{(\alpha, \beta)}(t_i)(t_i - x)} y^*(t_i).
\end{equation}

It can be easily recast in a form closely related to Hunter’s. Upon collocation at \( s_j \)

\begin{equation}
Q_n(y^*, s_j) = \sum_{i=1}^{n} \frac{w_i y^*(t_i)}{t_i - s_j} - \cot(\pi \alpha) \rho(s_j) \sum_{i=1}^{n} \frac{P_n^{(\alpha, \beta)}(s_j)}{P_n^{(\alpha, \beta)}(t_i)(t_i - s_j)} y^*(t_i),
\end{equation}

where in the first sum the weights are given by (12). The relationship between weights is then

\[ W_i = \frac{w_i}{t_i - s_j} - \cot(\pi \alpha) \rho(s_j) \frac{P_n^{(\alpha, \beta)}(s_j)}{P_n^{(\alpha, \beta)}(t_i)(t_i - s_j)}, \]

so that taking absolute values, summing and using lemma 1, for \( \max(\alpha, \beta) > -\frac{1}{2} \),

\begin{equation}
\sum_{i=1}^{n} \left| \frac{w_i}{t_i - s_j} \right| \leq \sum_{i=1}^{n} |W_i| + |\cot(\pi \alpha) \rho(s_j)| n^{\max(\alpha, \beta) + \frac{1}{2}}
\end{equation}

and similarly for \( \max(\alpha, \beta) = -\frac{1}{2} \). Lemma 3 of \cite{2} shows that the weights \( W_i \) in the quadrature behave like \( H + K \log n \). However the constants \( H \) and \( K \) depend on the location of the singularity, i.e. in our situation, on the collocation points. We have the following three different cases. For the first weight, from (3.6) of \cite{2}

\begin{equation}
|W_1| \leq \frac{B^*}{(1 - s_j)^{\left(\frac{4}{3} + \frac{3}{2}\right)}} \left\{ 1 + O\left( n^{-1} \right) \right\},
\end{equation}

\begin{equation}
B^* = 2^{\log_{10} \|B\| + 1} \left( 1 - s_j^2 \right)^{-\frac{3}{2}}.
\end{equation}

For the remaining ones, up to \( j - 1 \), (3.7) gives

\begin{equation}
\sum_{k=1}^{j-2} |W_k| \leq \frac{B^*}{(1 - s_j^2)^{\frac{3}{2}}} \left\{ D^* + (s_j + 2) \log n \right\} \left\{ 1 + O\left( n^{-1} \right) \right\},
\end{equation}

where the constant \( D^* \) is of the form

\[ D^* = C_1 + C_2 \log \left( 1 - s_j^2 \right). \]

In view of Theorem 8.9.1 of \cite{13}, we have

\[ \log \left( 1 - s_j^2 \right) \sim \alpha \log \left( j^\alpha n \right). \]
Substitution into (28) gives

\[ \sum_{k=1}^{j-2} |W_k| \leq \frac{B^*}{(1-x_j^2)^2} \left\{ C_1 + C_2 \alpha \log (j^\alpha n) \right\} \left\{ 1 + O\left( n^{-1} \right) \right\} \]

Finally for \( s_j-1 \), from (3.8) of [2]

\[ |W_{j-1}| \leq \frac{B^*}{(1-x_j^2)^2} \left\{ 1 + O\left( n^{-1} \right) \right\}. \]

Similar results hold for the remaining nodes, by substituting \( \alpha \) with \( \beta \). On using again theorem 8.9.1 of [13] on the previous estimates (27), (29) and (30), together with (26), we have

**Lemma 2.** For the weights in Hunter’s quadrature \( Q_n^\alpha (y^*, s_j) \) we have the bounds

\[ \left\| \frac{u_1}{t_1-s_j} \right\| \leq C_4 \left( \frac{j}{n} \right)^{-2-\frac{\beta}{2}} \left\{ 1 + O\left( n^{-1} \right) \right\} \]

\[ \sum_{i=1}^{j-2} \left\| \frac{u_i}{t_i-s_j} \right\| \leq C_5 \left( \frac{j}{n} \right)^{-2-\frac{\beta}{2}} \left\{ 1 + C_6 \log (n) \right\} \left\{ 1 + O\left( n^{-1} \right) \right\} \]

\[ \left\| \frac{u_{j-1}}{t_{j-1}-s_j} \right\| \leq C_7 \left( \frac{j}{n} \right)^{-2-\frac{\beta}{2}} \left\{ 1 + O\left( n^{-1} \right) \right\} \]

and similarly for the remaining weights.

Recall also formula (15.3.10) of [13] on the growth of the Christoffel numbers, i.e. the weights \( w_i \) relative to the nodes \( x_i \), in ordinary Gauss-Jacobi quadrature

\[ w_i \sim \frac{2^{\alpha+\beta+1}}{n^2} \left( \sin \frac{\theta_i}{2} \right)^{2n+1} \left( \cos \frac{\theta_i}{2} \right)^{2+1} \sim Cn^{-2(\alpha+\beta+1)}, \; x_i = \cos \theta_i, \]

in view also of theorem 8.9.1 of [13]. It then follows

**Lemma 3.** For the matrix \( B \), the following estimates hold

\[ \|B\|_\infty = \|B\|_1 = \|B\|_2 \sim n^{1-\kappa}. \]

*Proof.* Use (31) and the definitions of \( \kappa \) and of the diagonal matrix \( B \). \( \square \)

**Lemma 4.** We have also,

\[ \|M^t\|_\infty \leq n \|M^t\|_1 \leq n^{\frac{\kappa}{2}+\min(\alpha,\beta,\mu)-\kappa} \log n. \]

*Proof.* In fact, \( \|D_m\|_1 \leq C_{87}n^{\min(\alpha,\beta,\mu)} \), in view of the Hölder continuity of the function \( \rho \) and of \( a, b \). For \( N_m^{t/2} \) use (31) and for \( A_m^t \) use lemma 2, to get

\[ \|A_m^t\|_1 \leq n^{\frac{\kappa}{2}} \log n. \]
Hence, the claim.\[\square\]

**Lemma 5.** The matrix $\Lambda = ANA^T$ with entries given by (32), (33) and (34) below, has the following block structure, where $D_1$ and $D_2$ are diagonal matrices

$$\Lambda = \begin{bmatrix} D_1 & G^d \\ G & D_2 \end{bmatrix}.$$ 

**Proof.** Let us calculate $\Lambda_{i,j}$. Let $\lambda_i = \pi [b(s_i)]^{-1}$. Since $b$ is a polynomial with only finitely many zeros, for $n$ large enough we can assume $b(s_i) \neq 0$. There are three different cases to consider.

For $1 \leq i \leq n - \kappa$, $n - \kappa + 1 \leq j \leq n - \kappa + \chi$, letting $m = j - (n - \kappa + 1) \in \{0, \ldots, \chi - 1\}$, we have, by using the definition of the quadrature weights (12)

$$\lambda_i \Lambda_{i,j} = \sum_{k=1}^{n} \frac{1}{t_k - s_i} P_m^{(a,\beta)}(t_k) P_n^{(a,\beta)}(t_k) \frac{1}{\pi} \int_{-1}^{1} \frac{P_m^{(\alpha,\beta)}(t) \rho(t)}{t - t_k} dt = \frac{1}{\pi} \sum_{k=1}^{n} \frac{P_m^{(a,\beta)}(t_k)}{P_n^{(a,\beta)}(t_k)} \int_{-1}^{1} \frac{1}{t - t_k} + \frac{1}{t_k - s_i} \frac{P_n^{(\alpha,\beta)}(t) \rho(t)}{t - s_i} dt$$

Observe that the interpolation formula is exact on polynomials of degree up to $n$, so that

$$\sum_{k=1}^{n} P_m^{(a,\beta)}(t_k) P_n^{(a,\beta)}(t) = P_m^{(a,\beta)}(t), \quad \sum_{k=1}^{n} P_m^{(a,\beta)}(t_k) P_n^{(a,\beta)}(s_k) = P_m^{(a,\beta)}(s_k).$$

Substituting into the former expressions we have

$$\lambda_i \Lambda_{i,j} = \frac{1}{\pi} \int_{-1}^{1} \frac{P_m^{(a,\beta)}(t) \rho(t)}{(t - s_i)} dt = \frac{1}{\pi} \frac{P_n^{(\alpha,\beta)}(s_i)}{P_n^{(\alpha,\beta)}(s_i)} \int_{-1}^{1} \frac{P_m^{(\alpha,\beta)}(t) \rho(t)}{t - s_i} dt;$$

by using the second kind Jacobi function (8) and in view of the choice of the collocation nodes, $P_{n-k}^{(-\alpha,\beta)}(s_i) = 0$, so that for $i \neq j$

$$(32) \quad \Lambda_{i,j} \equiv \begin{bmatrix} G^d \end{bmatrix}_{i,m+1} = \frac{2^{-\kappa} b(s_i)}{\sin(\pi \alpha)} P_{m-k}^{(-\alpha,\beta)}(s_i).$$

where we have introduced the $\chi \times (n - \kappa)$ matrix $G$.

For $i, j \leq n - k$, and $i \neq j$

$$\sum_{k=1}^{n} \frac{w_k}{t_k - s_i} \frac{1}{w_k} \frac{w_k}{t_k - s_j} = \frac{1}{\pi} \sum_{k=1}^{n} \left[ \frac{w_k}{t_k - s_i} - \frac{w_k}{t_k - s_j} \right] = 0$$

in view of the previous computations, for the case $m \equiv 0$. For $i = j$ instead, we have, [8]

$$(33) \quad \Lambda_{i,i} = \frac{\beta^2(s_i)}{\pi} \sum_{k=1}^{n} \frac{w_k}{(t_k - s_i)^2} = \frac{\beta^2(s_i)}{w_i^2} \equiv a_i^2, \quad i = 1, \ldots, n - \kappa.$$
We now consider the case $n - \kappa + 1 \leq i, j \leq n - \kappa + \chi$; once again let $m = j - (n - \kappa)$, $q = i - (n - \kappa + 1)$. Using the expression of the weights, the Lagrange interpolation formula being exact on polynomials of degree $m + q < n$, and the exactness of the quadrature over polynomials of degree "low enough",

\[
\Lambda_{i,j} = \sum_{k=1}^{n} w_k P_q^{(\alpha,\beta)}(t_k) P_m^{(\alpha,\beta)}(t_k)
\]

\[
= \frac{1}{\pi} \sum_{k=1}^{n} \frac{1}{P_n^{(\alpha,\beta)}(t_k)} \int_{-1}^{1} P_n^{(\alpha,\beta)}(t) \rho(t) dt \frac{P_q^{(\alpha,\beta)}(t_k) P_m^{(\alpha,\beta)}(t_k)}{P_n^{(\alpha,\beta)}(t_k) (t - t_k)}
\]

\[
= \frac{1}{\pi} \int_{-1}^{1} \rho(t) \sum_{k=1}^{n} P_q^{(\alpha,\beta)}(t_k) P_m^{(\alpha,\beta)}(t_k) \frac{P_n^{(\alpha,\beta)}(t)}{P_n^{(\alpha,\beta)}(t_k) (t - t_k)} dt
\]

\[
= \frac{1}{\pi} \int_{-1}^{1} \rho(t) P_q^{(\alpha,\beta)}(t) P_m^{(\alpha,\beta)}(t) dt.
\]

From the orthogonality of the Jacobi polynomials, then

\[
\Lambda_{i,j} = \delta_{ij} \| P_q^{(\alpha,\beta)} \|_p^2 \equiv b_q^2, \quad q = 0, ..., \chi - 1,
\]

where the last symbol denotes the weighted two norm. \qed

Let us put $\tilde{\Lambda} = DD^t + ANA^t$. We will need an estimate for $\tilde{\Lambda}^{-1}$. Recall (7.32.2) of [13], for which

\[
\| P_n^{(\alpha,\beta)} \|_p^2 = \frac{2^{\alpha+\beta+1}}{2n + \alpha + \beta + 1} \frac{\Gamma(n + \alpha + 1) \Gamma(n + \beta + 1)}{\Gamma(n + 1) \Gamma(n + \alpha + \beta + 1)}
\]

so that

\[
\| P_n^{(\alpha,\beta)} \|_p^{-2} \sim n.
\]

Let us put $D_0 = \text{diag} (d_1, \ldots, d_{n-\kappa})$, $E \equiv DD^t = \text{diag} (D_0^2, 0, \ldots, 0)$, a square matrix of size $n - \kappa + \chi$, as is $\tilde{\Lambda}$. Then $\tilde{\Lambda} = \Lambda + E$. Now $\Lambda$ is a symmetric matrix, with the block form given in the lemma, where $G$ is the $\chi \times (n - \kappa)$ matrix with elements $G_{m,i}$, given by (32) and the diagonal matrix are square, of sizes respectively $n - \kappa$ and $\chi$, and whose elements are $D_2^2 = \text{diag} (a_1^2, \ldots, a_{n-\kappa}^2)$, see (33) and $D_2 = \text{diag} (v_0^2, \ldots, v_{\chi-1}^2)$, see (34). We will use a second decomposition of $\tilde{\Lambda}$, $\tilde{\Lambda} = \Lambda^* + \tilde{E}$

\[
\Lambda^* = \begin{bmatrix}
D_2^2 & G^t \\
G & D_2
\end{bmatrix}, \quad \tilde{E} = \begin{bmatrix}
D_0^2 & 0 \\
0 & O
\end{bmatrix}.
\]

We need a preliminary result.

**Lemma 6.** For $n$ large enough

\[
\tilde{\Lambda} \equiv GD_1^{-2}G^t = \gamma \text{diag} \left( \| P_{-\kappa}^{(-\alpha,-\beta)} \|_p^{-2}, \ldots, \| P_{\chi-1-\kappa}^{(-\alpha,-\beta)} \|_p^{-2} \right), \quad \text{with} \quad \gamma = \frac{2^{-2\kappa}}{\sin^2(\pi \alpha)}.
\]
Proof. Observe that for \( l, r = 0, \ldots, \chi - 1 \)

\[
\Delta_{l+1,r+1} = \sum_{k=1}^{n-k} \frac{2^{-k}}{\sin (\pi \alpha)} P_{l-k}^{[-\alpha, -\beta]} (s_k) w_k^* \frac{2^{-k}}{\sin (\pi \alpha)} P_{r-k-(n-k)}^{[-\alpha, -\beta]} (s_k)
\]

Using the definition of weights in the last expression we obtain

\[
\Delta_{l+1,r+1} = \gamma \sum_{k=1}^{n-k} \int_{-1}^{1} \rho^{-1} (t) \frac{P_{l-k}^{[-\alpha, -\beta]} (t) dt}{P_{n-k}^{[-\alpha, -\beta]} (s_k)^{t-l-k}} P_{r-k-(n-k)}^{[-\alpha, -\beta]} (s_k) P_{l-k}^{[-\alpha, -\beta]} (s_k) \frac{P_{l-k}^{[-\alpha, -\beta]} (t) dt}{P_{n-k}^{[-\alpha, -\beta]} (s_k) (t - s_k)}
\]

the last step being exact if \( n > l + r - k \), since we interpolate a polynomial. It follows

\[
\Delta_{l+1,r+1} = \gamma \delta_{l,r} \left\| P_{l-k}^{[-\alpha, -\beta]} \right\|_{\rho^{-1}}^2.
\]

By introducing the elementary matrix

\[
R = \begin{bmatrix} I & 0 \\ -GD_1^{-2} & I \end{bmatrix},
\]

we can diagonalize \( \Lambda^* \), since \( \Lambda^* = WR^t \), with \( W = \text{diag} [D_2^2, D_2 - \Delta] \). Thus the eigenvalues of \( \Lambda^* \) are just the entries on the diagonal of \( W \). Observe that for \( k = 0, \ldots, \chi - 1 \), by (35)

\[
(37) \quad \gamma \| P_{l-k}^{[-\alpha, -\beta]} \|_{\rho^{-1}}^2 = \frac{1}{\sin^2 (\pi \alpha)} \| P_{k}^{[\alpha, \beta]} \|_{\rho}^2
\]

so that the eigenvalues of \( \Lambda^* \) can be written down in an ordered fashion as follows

\[
(38) \gamma^* \| P_{l-k}^{[\alpha, \beta]} \|_\rho < < \gamma^* \| P_{\chi-k}^{[\alpha, \beta]} \|_\rho < 0 < \text{min} a_i^2 < \ldots < \text{max} a_i^2, \quad \gamma^* = 1 - \frac{1}{\sin^2 (\pi \alpha)}
\]

Since \( \Lambda \) is symmetric, so is its inverse. Their 2-norms are then given by their respective spectral radii. We are therefore interested in the eigenvalue of \( \Lambda^* \) which is closest to zero. In absence of more specific informations about the coefficients \( a(x) \) and \( b(x) \) of the original equation appearing in the terms \( a_i \) and \( d_i \), the above task seems to be a difficult one. We will then make some simplifying assumptions in the following discussion.
Fix now $0 < \epsilon < 1$, and let $\Delta_\epsilon \equiv [-1 + \epsilon, 1 - \epsilon]$. First of all, we assume that no $s_i$ tends to a zero of $b(x)$ as $n \to \infty$. In view of (31) it then follows that $a_i^2 \sim b(s_i) n^{2(1-\kappa)} \sim n^{2-2\kappa}$. We can then say that as $n$ increases all the $a_i$ increases as well; hence it is reasonable to take the eigenvalue of $A^*$ closest to zero to be $\omega_\lambda = \gamma^* \| P_{\lambda-1} \|_2$; alternatively, this certainly holds if

$$|\omega_\lambda| < \min_i a_i^2. \tag{39}$$

By the corollary to Weyl’s theorem, see [12] p. 193, and the fact that $\hat{E}$ is positive semidefinite, it follows that the eigenvalue of $A$ closest to zero is bounded below by $\omega_\lambda$ and by a standard result, [12] p. 191, it is bounded above by $\omega_\lambda + \| \hat{E} \|_2$. We would like this quantity still to be negative, as this ensures that $A$ is invertible and that this is still the eigenvalue of the matrix closest to zero. Hence together with (39) let us assume that

$$\omega_\lambda + \| \hat{E} \|_2 < 0. \tag{40}$$

To understand how strong this assumption could be, let us remark that $\| \hat{E} \|_2 = \max_i d_i^2(s_i)$. Use then (13); since the coefficients are continuous, the quantity in (13) within the brackets is bounded above; however if $\alpha$ or $\beta$ are negative the function $\rho$ blows up like $n^{-1-\min(\alpha, \beta)}$, when $s_i \to \pm 1$. To ensure then that (40) holds, in this case we need then to restrict attention to $s_i \in \Delta_\epsilon$. However condition (40) may be satisfied even in $[-1, 1]$ if both $\alpha, \beta \geq 0$. In this case (40) does not appear to be a very strict requirement. We may even relax it a bit as follows

$$|\omega_\lambda + \| \hat{E} \|_2| \leq C_\eta n^{-\zeta}, \quad > 0,$$

since more generally it involves the coefficients of the original equation. With this condition, in view of the above discussion, the following estimate holds

$$\| \bar{A}^{-1} \|_\infty \leq \sqrt{n} \| \bar{A}^{-1} \|_2 \leq C_{10} n^{\zeta + \frac{1}{2}}. \tag{41}$$

We have then

**Lemma 7.** For the consistency error the following estimates hold

$$\| \mathcal{L}_G \|_\infty \leq C_{11} n^{\max(\alpha, \beta) + \frac{1}{2} + \gamma}, \quad \text{for} \quad \max (\alpha, \beta) > -\frac{1}{2},$$

$$\| \mathcal{L}_G \|_{L_2} \leq C_{11} n^\gamma \log n, \quad \text{for}\quad \max (\alpha, \beta) = -\frac{1}{2}.$$ 

In $\Delta_\epsilon$, the above estimates can be written as

$$\| \mathcal{L}_G \|_{L_\infty} \leq C_{11} (\epsilon) n^{\max(\alpha, \beta) + \frac{1}{2} - \eta}, \quad \text{for} \quad \max (\alpha, \beta) > -\frac{1}{2},$$

$$\| \mathcal{L}_G \|_{L_\infty} \leq C_{11} (\epsilon) n^{-\eta} \log n, \quad \text{for} \quad \max (\alpha, \beta) = -\frac{1}{2}.$$
with $\eta$ “arbitrarily” high, where the notation emphasizes the fact that the bound itself depends on $\epsilon$, and the norm is taken in $\Delta$.

Proof. Standard techniques estimate the Gaussian quadrature error in terms of the Lebesgue constant times the best approximation error for the integrand, $E_n(y^*)$. Use lemma 1 and Jackson’s theorem, which gives the best approximation error in terms of the modulus of continuity of the integrand. Recalling section 2, the integrand function $y^*(x)$ however is only in $H_\lambda[-1, 1]$, see (5). This is not enough to ensure convergence for the method, as it will be clear in the proof of the theorem below. To obtain convergence, we are then forced once again to restrict the domain of interest to $\Delta$, since there $y^*(x)$ is analytic. We can then take $y^* \in C^\eta(\Delta)$, with $\eta$ “arbitrarily” high. We thus have

$$E_{n, \eta}(y^*) \leq C_{11}(\epsilon)n^{-\eta}.$$

Theorem 8. The proposed method is convergent in $\Delta$, with rate given by

$$\|\mathcal{L}\|_\infty \leq C_{12}(\epsilon)n^{\frac{\delta + \varepsilon + \max(\alpha, \beta) + \min(\alpha, \beta, \mu)}{2} - 2\epsilon - \eta \log n}.$$

Proof. We can recast the error equation (23) in the following form and use lemmas 3, 4 and 7 and (41) together with the equivalence of norms to get the claim

$$\|\mathcal{L}\|_\infty \leq \|B\|_\infty \|M\|_\infty \|\Lambda^{-1}\|_\infty \|\mathcal{L}\|_{e, \infty}.$$

Remarks.

1. In practice, at least for the nodes closest to the endpoints, the numerical convergence may very well be destroyed, since the error bound $C_{12}(\epsilon)$ does actually grow as $\epsilon$ tends to 0.

2. For nonpositive index equations the analysis simplifies considerably, since the matrix of the system is just given by (18). Then $\|\Lambda^{-1}\|_\infty = \|D_3^{-1}\|_\infty = n^{\min\{4\alpha, 4\beta, \xi\}}$, where $\xi > 0$ is the rate for which some $s_i$ tend to a zero of $b(x)$, if they do it at all.

3. The result of the theorem shows that convergence can be attained only for the “central” nodes of $[-1, 1]$; problems at the endpoints can in a certain sense be expected in view of our choice of the quadrature weight $\rho$, see the discussion after (2).

REFERENCES


THE DISCRETE SLOAN ITERATE FOR CAUCHY SINGULAR INTEGRAL EQUATIONS

SEKI KIM

DEPARTMENT OF MATHEMATICS,
SUNGKYUNKWAN UNIVERSITY, SUWON 440-746, KOREA
skkim@math.skku.ac.kr

Abstract. The superconvergence of the Sloan iterate obtained from a Galerkin method for the approximate solution of the singular integral equation based on the use of two sets of orthogonal polynomials is investigated. The discrete Sloan iterate using Gaussian quadrature to evaluate the integrals in the equation becomes the Nyström approximation obtained by the same rules. Consequently, it is impossible to expect the faster convergence of the Sloan iterate than the discrete Galerkin approximation in practice.

Key words. discrete Sloan iterate, Cauchy singular integral equation, Nyström approximation

AMS subject classification. 65R20, 45E05

1. Introduction.

Singular integral equations (SIEs) differ from Fredholm integral equations (FIEs) because they contain a Cauchy principal value integral. Because of this difference, the number of applicable numerical methods for SIEs is less than that for FIEs. Among numerical methods used for FIEs, the Sloan iterate is well-known to solve FIEs of the second kind with its superconvergence. It is taken as a better approximation than the Galerkin method as its general case [?, ?]. Recently the Sloan iterate of projection method for solving SIEs has been considered without practical examples in a number of papers [?, ?, ?]. They suggest the Sloan iterate as a very useful method for accelerating the convergence of some projection methods for SIEs. However, its superconvergence was not shown in actual calculations. In the special case of SIEs, the airfoil equation which is the SIEs of the first kind, the Sloan iterate does not converge at the rate no faster than the original approximation [?]. Here we will extend this result to the SIEs with variable coefficients. This shows we may not obtain the superconvergence of the Sloan iterate for SIEs.

The paper is organized as follows: after recalling some preliminary results, in section 3 we consider the quadrature methods used in SIEs including the Hunter’s method.
to evaluate numerically Cauchy principle value integrals, and derive some facts which are important in proving the main result. In section 4 we review the two sets of orthonormal polynomial-based Galerkin method for solving the SIEs with the order of convergence. Then we investigate the functional equation satisfied by the discrete Galerkin approximation, that is essential to obtain the discrete Sloan iterate. In the last section, we establish that the convergence rate of the Sloan iterate is exactly same to the rate of the Nyström approximation by examining the effects of numerical integration errors on the Galerkin solution.

2. Preliminaries. Consider the second kind singular integral equation on \([-1, 1)\)

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

with \(\tilde{a}(x)\) and \(\eta(x, t)\), real valued functions. We assume here that the kernel \(\eta\) is a Hölder continuous function on \([-1, 1] \times [-1, 1]\), and moreover a continuously differentiable function of the variable \(x\). We now consider the function of a single variable \(\eta(x, x)\), which is also Hölder continuous. Since it may have zeros at the points \(\lambda_i, i = 1, ..., N_\lambda\), each respectively of multiplicity \(\zeta_i\), we can define the polynomial

\[
\tilde{b}(x) = \prod_{i=1}^{N_\lambda} (x - \lambda_i)^{\zeta_i} \quad \text{with} \quad N = \sum_{i=1}^{N_\lambda} \zeta_i
\]

and consider the function

\[
v(x) = \frac{\eta(x, x)}{\tilde{b}(x)}
\]

which is of one sign, here taken to be positive, bounded near the endpoints \([-1, 1]\). Since \(\eta\) is Hölder continuous, \(v(x)\) must be integrable in \([-1, 1]\). After subtracting the singularity, the SIE then becomes

\[
\tilde{a}(x) \tilde{\mu}(x) + \tilde{b}(x) \int_{-1}^{1} \frac{1}{\pi} \frac{v(t) \tilde{\mu}(t)}{t-x} dt + \int_{-1}^{1} \tilde{k}(x, t) \tilde{\mu}(t) dt = \tilde{f}(x), \quad -1 < x < 1.
\]

where the coefficients \(\tilde{a}\) and \(\tilde{b}\) do not have common zeros \([?]\). Since \(v(x)\) is one-signed, we can then normalize them introducing the function

\[
r^2(x) = \frac{\tilde{a}^2(x)}{v^2(x)} + \tilde{b}^2(x) > 0, \quad \forall x \in [-1, 1],
\]

after appropriate simplifying and finally by changing the dependent variable

\[
\mu(x) = v(x) \tilde{\mu}(x),
\]
the SIE (??) then takes the standard form

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

(5)

where we recall that $b(x)$ is a polynomial of degree $N$ in (??).

We denote the space of Hölder continuous functions on $[-1, 1]$ by $H$, but we seek the solution in the space denoted by $H^*$, see [7], of Hölder continuous functions on $(-1, 1)$, such that near the endpoints their behavior is given by

$$\mu(x) = \frac{\sigma_{-1}(x)}{(1 + x)^{\tilde{\xi}_1}}, \quad \mu(x) = \frac{\sigma_1(x)}{(1 - x)^{\tilde{\xi}_1}}, \quad 0 \leq \tilde{\xi}_1, \tilde{\xi}_1 < 1,$$

(6)

with $\sigma_{-1}(x)$ and $\sigma_1(x)$ also in $H$. Also the real valued function $f$ is assumed to lie in $H^*$.

Now it is convenient to introduce $L_Z$ and $L_{Z^{-1}}$ Hilbert spaces of real-valued square-integrable functions with weights $Z$ and $1/Z$ respectively for Galerkin methods used in this study. The inner products on $L_Z$ and $L_{Z^{-1}}$ are denoted by $\langle, \rangle_Z$ and $\langle, \rangle_{Z^{-1}}$ where

$$\langle f, g \rangle_Z = \int_{-1}^{1} Z(t)f(t)g(t) \, dt \quad \text{and} \quad \langle f, g \rangle_{Z^{-1}} = \int_{-1}^{1} Z^{-1}(t)f(t)g(t) \, dt,$$

and the induced norms by $\| \|_Z$ and $\| \|_{Z^{-1}}$. The symbol $|| ||$ without subscript will be used for operator norms. In operator notation, equation (??) can be written as

$$\hat{S}\mu + \hat{K}\mu = f,$$

(7)

where the so-called dominant operator is

$$\hat{S}\mu(x) \equiv a(x)\mu(x) + \frac{b(x)}{\pi} \int_{-1}^{1} \frac{\mu(t)}{t-x} \, dt,$$

(8)

and

$$\hat{K}\mu(x) \equiv \int_{-1}^{1} k(x,t) \mu(t) \, dt$$

(9)

defines a compact operator from $L_Z$ to $L_{Z^{-1}}$. We assume that $\hat{S}\mu + \hat{K}\mu = 0$ has only the zero solution so that $(\hat{S} + \hat{K})^{-1} : L_{Z^{-1}} \to L_Z$ is bounded by the Fredholm alternative. Following [7] we introduce the “regularizing” operator $S^*$ on $H^*$ by

$$S^*\rho(x) \equiv a(x)\rho(x) - \frac{b(x)}{\pi} \int_{-1}^{1} \frac{\rho(t)}{Z(t)(t-x)} \, dt,$$

(10)

where $Z(x)$ represents the fundamental function of the SIE, see [7]. There is a relationship between the fundamental function and the canonical function $X(z)$ of the SIE, see e.g. [7]. Indeed letting
\[ \theta (t) = - \frac{1}{2\pi i} \ln \frac{a(t) - ib(t)}{a(t) + ib(t)} = \frac{1}{\pi} \arctan \frac{b(t)}{a(t)} + N(t), \]

where the function \( N(t) \) takes integer values, we have

\[ X(z) = (1 - z)^{-\kappa} \exp \left\{ - \int_{-1}^{1} \frac{\theta(t) \, dt}{t - z} \right\}, \quad z \notin [-1, 1], \]
\[ Z(x) = (1 - x)^{-\kappa} \exp \left\{ - \int_{-1}^{1} \frac{\theta(t) \, dt}{t - x} \right\}, \quad x \in (-1, 1). \]

Here \( \kappa \) denotes the index of the equation, to be defined below. Observe that \( Z(x) \) is positive and bounded in each closed subinterval of \((-1, 1)\), since in our case it can be written as follows,

\[ Z(x) = (1 - x)^{\gamma_2} \exp \left\{ - \int_{-1}^{1} \theta(t) \, dt \right\}, \quad x \in (-1, 1). \]

The classical theory \cite{?} shows that if \( \kappa \geq 0 \), the dominant equation \((??)\) where \( \hat{K} \equiv 0 \), is solvable for any right hand side \( f \). Uniqueness however is not guaranteed. To ensure it, we need the supplementary conditions

\[ \int_{-1}^{1} \frac{\mu(t) \, dt}{t - x} = C_l, \quad l = 0, 1, \ldots, \kappa - 1, \]

where \( C_l \) denote constants. For \( \kappa < 0 \), the dominant equation has a unique solution if and only if \( f \) satisfies the \(-\kappa\) orthogonality conditions

\[ \int_{-1}^{1} \frac{t^l \, f(t) \, dt}{Z(t)} = 0, \quad l = 0, 1, \ldots, -\kappa - 1. \]

Let \( \{\phi_n\}_{n=0}^{\infty} \) be the set of monic polynomials orthogonal with respect to the weight function \( Z \), and let \( \{\psi_n\}_{n=0}^{\infty} \) be the corresponding set of monic polynomials orthogonal with respect to \( 1/Z \). There is a special relationship between the two sets of orthogonal polynomials given by the dominant operator \( \hat{S} \) of \((??)\). Let us recall the following two results, Theorems 3.1 and 3.2 of \cite{?}:

**Lemma 2.1.** Let \( Q, R \) be functions such that

(i) \( QX - R \) is analytic in the deleted complex plane and zero at infinity.

(ii) on \((-1, 1)\), \( Q^+ (x) = Q^- (x) \), \( R^+ (x) = R^- (x) \) and the functions \( aQZ - R, bQZ \) are in the Hölder space \( H^* \). Then for \(-1 < x < 1\),

\[ \frac{1}{\pi} \int_{-1}^{1} \frac{b(t) \, Q(t) \, Z(t)}{t - x} \, dt = -a(x) Q(x) Z(x) + R(x). \]
By applying the above result to the two families of monic polynomials
then cause the operator \( \hat{S} \) always assume that by taking
\( s_n \) where
(16)
\[
\psi_n(x) = n
\]
From now on assume that
In what follows we will consider the polynomial
\( \phi_n \) by
Theorem 2.3. 
For our analysis the important tool is given by Theorem 9.14 of [?]. See also
theorems 3.2 and 3.4 of [?]. In our notation, recalling that \( N = \deg b(x) \), (??), it reads
\[
\|p_n\|Z = \|q_n\|Z^{-1}.
\]
From now on assume that \( n - \kappa > N - 1 \). Recall that \( \kappa \) is the index of \( \hat{S} \).
By applying the above result to the two families of monic polynomials \( \{\phi_n\}_{n=0}^{\infty} \) and
\( \{\psi_n\}_{n=0}^{\infty} \), orthogonal with respect to the weights \( Z \) and \( Z^{-1} \) respectively, we have
(15)
\[
a(x)\phi_n(x)Z(x) + \frac{b(x)}{\pi} \int_{-1}^{1} Z(t)\phi_n(t)dt = (-1)^n \psi_n(x),
\]
(16)
\[
a(x)\psi_n(x)Z(x) - \frac{b(x)}{\pi} \int_{-1}^{1} Z(t)\psi_n(t)dt = (-1)^n \phi_n(x),
\]
where \( n = \kappa + N, \kappa + N + 1, \ldots \). Also \( \phi_{-1} = \psi_{-1} = 0 \).
Note that two sets of polynomials mentioned above can be taken normalized because the operator \( \hat{S} \) is unitary if the index of \( \hat{S} \) is zero [?, ?].
In what follows we will consider the polynomial \( b(x) \) evaluated at the zeros \( t_i \) and
\( s_j \) of the orthogonal polynomials \( \phi_n \) and \( \psi_{n-\kappa} \) respectively. Since \( b \) has \( N \) zeros, we can always assume that by taking \( n \) large enough, the following conditions are satisfied
(17)
\[
b(t_i) \neq 0, \quad i = 1, \ldots, n; \quad b(s_j) \neq 0, \quad j = 1, \ldots, n - \kappa.
\]
3. Quadrature Methods in SIE. To set up a numerical method we define a
new unknown \( u \) using the fundamental function \( Z \)
(18)
\[
\mu(t) = Z(t)u(t).
\]
Indeed, the function $Z(t)$ contains the “bad features” of the unknown, and $u(t)$ is smooth. With this notation, we introduce simple operators by

$$
Su = \hat{S}Zu = \hat{S}\mu,
$$

$$
Ku = \hat{K}Zu = \hat{K}\mu.
$$

(19)

It is also necessary to introduce the new function $\Lambda_n$ associated with $\phi_n$

$$
\Lambda_n(x) = \frac{1}{\pi} \int_{-1}^{1} \frac{Z(t) \phi_n(t)}{t-x} dt.
$$

(20)

As quadrature formula we use Hunter’s method

$$
\frac{1}{\pi} \int_{-1}^{1} Z(t) \frac{u(t)}{t-x} dt = \frac{1}{\pi} \sum_{i=1}^{n} w_i u(t_i) + \frac{\Lambda_n(x)}{\phi_n(x)} u(x) + \epsilon_H
$$

(21)

where the $w_i$’s are the Gaussian quadrature weights associated with the weight $Z$, the $t_i$’s are the zeros of $\phi_n$ and $\epsilon_H$ represents the error term.

$$
\epsilon_H = \int_{-1}^{1} Z(t) R_n(x,t) \phi_n(t) dt, \quad R_n(x,t) = \frac{1}{2\pi i} \int_C \frac{u(z) dz}{(z-x)(z-t) \phi_n(z)}
$$

(22)

where $C$ denotes a contour in the complex plane, enclosing the interval $[-1, 1]$. In this notation (21) becomes

$$
b(x) \Lambda_n(x) = (-1)^{\kappa} \psi_{n-\kappa}(x) - a(x) Z(x) \phi_n(x)
$$

(23)

where $\kappa$ is the index of the singular integral equation. Note that (22) is exact when $u$ is a polynomial of degree less than $2n + 1$, see [?]. In a similar way by introducing the function

$$
\tilde{\Lambda}_n(x) = \frac{1}{\pi} \int_{-1}^{1} \frac{\psi_n(t)}{Z(t)(t-x)} dt,
$$

(24)

we can also write the “dual form” of Hunter’s quadrature (21), which is obtained by using the weight function $1/Z$ in place of $Z$, the weights $w_j^*$ in place of $w_j$ and where the polynomials $\psi_{n-\kappa}(x)$ and their zeros replace $\phi_n(x)$ and their zeros, and similar changes take place in the error term $\tilde{\epsilon}_H$,

$$
\frac{1}{\pi} \int_{-1}^{1} \frac{u(t)}{Z(t)(t-x)} dt = \frac{1}{\pi} \sum_{j=1}^{n-\kappa} \frac{w_j^* u(s_j)}{s_j-x} + \frac{\tilde{\Lambda}_{n-\kappa}(x)}{\psi_{n-\kappa}(x)} u(x) + \tilde{\epsilon}_H.
$$

(25)

Shifting the indices in (25) and rewriting it in this notation

$$
-b(x) \tilde{\Lambda}_{n-\kappa}(x) = (-1)^{\kappa} \phi_n(x) - a(x) Z(x) \phi_{n-\kappa}(x).
$$

(26)
For $i = 1, 2, \ldots, n$, the Christoffel number $w_i$ associated with the zeros $t_i$ of $\phi_n$, from (3.4.3) in [7] is

\begin{equation}
(27) \quad w_i = \int_{-1}^{1} \frac{Z(t) \phi_n(t)}{\phi_n'(t_i)} (t - t_i) dt = \pi \frac{\Lambda_n(t_i)}{\phi_n'(t_i)}.
\end{equation}

It follows then from (27)

\begin{equation}
(28) \quad b(t_i) w_i = \frac{\pi}{\phi_n'(t_i)} \left[ (-1)^{\kappa} \psi_{n-\kappa}(t_i) - a(t_i) Z(t_i) \phi_n(t_i) \right] = \pi (-1)^{\kappa} \frac{\psi_{n-\kappa}(t_i)}{\phi_n'(t_i)}.
\end{equation}

Similarly from (27) the value for the Christoffel numbers $w_j^*$ associated with the zeros $s_j$ of the orthogonal polynomials $\psi_{n-\kappa}$ is

\begin{equation}
(29) \quad -b(s_j) w_j^* \equiv -b(s_j) \frac{\tilde{\Lambda}_{n-\kappa}(s_j)}{\psi^*_{n-\kappa}(s_j)} = \pi (-1)^{\kappa} \frac{\phi_n(s_j)}{\psi^*_{n-\kappa}(s_j)} \quad \text{for} \quad j = 1, 2, \ldots, n - \kappa.
\end{equation}

Recall that the Christoffel numbers are all positive [7]. Using this fact with (28), (29) and (27), it follows then that $\{s_i\} \cap \{s_j\} = \emptyset$.

Recalling (28) and (29), since $t_i \in (-1, 1)$, we define $K_n u$ by discretizing $K u$ by means of a Gaussian quadrature with weight $Z(t)$ as follows

\begin{equation}
(30) \quad Ku(x) = \int_{-1}^{1} Z(t) k(x, t) u(t) dt \simeq \sum_{i=1}^{n} w_i k(x, t_i) u(t_i) \equiv K_n u(x).
\end{equation}

We now consider the regularized equation. Recalling (28), the FIE equivalent to the SIE (27) is given by (107.15) of [7],

\begin{equation}
(31) \quad \mu(x) + S^* K \mu(x) = S^* f(x) + Z(x) b(x) \tilde{p}_{\kappa-1}(x),
\end{equation}

where $\tilde{p}_{\kappa-1}$ represents an arbitrary polynomial of degree not greater than $\kappa - 1$, which is identically zero for nonpositive index.

We need some results which are used for last part of the paper. From [7], the canonical function $X(z)$ and its reciprocal have the expansions

\begin{equation}
(32) \quad X(z) = (-1)^{\kappa} \sum_{j=-\infty}^{\kappa} \alpha_j z^j \quad \text{and} \quad X^{-1}(z) = (-1)^{\kappa} \sum_{j=-\infty}^{\kappa} \beta_j z^j,
\end{equation}

where the coefficients $\alpha_j$ are given by

\begin{equation}
(33) \quad \alpha_j = \frac{1}{\pi} \int_{-1}^{1} \tau^{-j-1} b(\tau) Z(\tau) d\tau \quad \text{for} \quad j \leq \min(-\kappa, -1).
\end{equation}

The other coefficients can be obtained from $X X^{-1} \equiv 1$, which gives

\begin{equation}
(34) \quad \sum_{l=0}^{j} \alpha_{-\kappa+j+l} \beta_{\kappa-l} = \left\{ \begin{array}{ll}
0 & j < 0 \\
1 & j = 0.
\end{array} \right.
\end{equation}
On using this result twice in (38), we have
\[ X(x) \sim (-1)^\kappa x^{-\kappa}, \]
and from this we have \( \alpha_{-\kappa} = 1 \), so that \( \beta_\kappa = 1 \) as well. We define the polynomial \( \hat{p}_n(x) \) as follows
\[ \hat{p}_n(x) = pp\left((-1)^\kappa X^{-1}\right) = \sum_{i=0}^\kappa \beta_i x^i. \]
Moreover in (2.11) of [?] it is observed that the function \( R \) of lemma ?? represents the principal part of the canonical function, so that
\[ R(x) = \hat{p}_n(x). \]

**Theorem 3.4.**
\[ \frac{1}{\pi} \sum_{j=1}^{n-\kappa} \frac{w_j b(s_j)}{(t_i - s_j)(t_i - s_j)} + \hat{p}_n[t_i, t_i] = \begin{cases} 0 & i \neq l \\ \frac{\pi}{b(t_i)} w_i & i = l \end{cases} \]
where the last term represents the first divided difference of a certain polynomial \( \hat{p}_n(x) \), of degree \( \kappa \), which for \( \kappa = 0 \) becomes identically 0.

**Proof.** Using partial fractions, if \( l \neq i \)
\[ \frac{1}{\pi} \sum_{j=1}^{n-\kappa} \frac{w_j b(s_j)}{(t_i - s_j)(t_i - s_j)} = \frac{1}{\pi (t_i - t_i)} \sum_{j=1}^{n-\kappa} w_j b(s_j) \left[ \frac{1}{t_i - s_j} - \frac{1}{t_i - s_j} \right]. \]
Now we need to evaluate the right hand side of (38). Since \( b(x) \) is a polynomial of fixed degree \( N \), for \( n \) large enough the above quadrature is exact. The second term in the right hand side of (38) can now be rewritten using (35), while for the term on the left hand side we can use lemma ?? with \( Q \equiv 1 \) to get
\[ \frac{1}{\pi} \sum_{j=1}^{n-\kappa} \frac{w_j b(s_j)}{s_j - x} = \frac{a(x)}{Z(x)} - R(x) + \frac{1}{\psi_{n-\kappa}(x)} \left[ (-1)^\kappa \phi_n(x) - \frac{a(x)}{Z(x)} \psi_{n-\kappa}(x) \right] \]
\[ = -R(x) + \frac{(-1)^\kappa \phi_n(x)}{\psi_{n-\kappa}(x)}. \]
Use of (38) and collocation of (35) at the nodes \( t_i \) yields
\[ \frac{1}{\pi} \sum_{j=1}^{n-\kappa} \frac{w_j b(s_j)}{s_j - t_i} = -\hat{p}_n(t_i). \]

On using this result twice in (38) we have the first claim. For the case, \( i = l \), we start by differentiating (35) with respect to \( x \), to get
\[ \frac{1}{\pi} \sum_{j=1}^{n-\kappa} \frac{w_j b(s_j)}{(s_j - x)^2} = -R'(x) + (-1)^\kappa \left[ \frac{\phi_n'(x)}{\psi_{n-\kappa}(x)} - \frac{\psi_{n-\kappa}'(x) \phi_n(x)}{\psi_{n-\kappa}(x)^2} \right]. \]
Upon collocation at $t_l$, we have then
\[
\frac{1}{\pi} \sum_{j=1}^{n-\kappa} \frac{w_j b(s_j)}{(s_j - t_l)^2} = -\tilde{p}_{\kappa}'(t_l) + (-1)^\kappa \frac{\phi_n'(t_l)}{\psi_{n-\kappa}(t_l)}.
\]

The claim follows then using equation (40).

We can now characterize the coefficients of the arbitrary polynomial $\tilde{p}_{\kappa-1}(x)$ of (40) by using the supplementary normalization conditions (3).

**Proposition 3.5.** We have
\[
\tilde{p}_{\kappa-1}(x) = \sum_{i=0}^{\kappa-1} \xi_i x^i
\]
where
\[
\xi_i = C_i + \sum_{l=1}^{\kappa-1-i} \beta_{\kappa-l} C_{i+l},
\]
and the $C_i$’s are given by (3).

**Proof.** From the proof of Theorem 4.4 of [?] and by taking into account limits of sequences of functions, Theorem 5.3 of [?], $\tilde{p}_{\kappa-1}(x)$ has the following coefficients. For $0 \leq i \leq \kappa - 1$, on using (40)

\[
\xi_i = \frac{1}{\pi} \int_{-1}^{1} \tau^{\kappa-1-i} \phi(\tau) \, d\tau - \sum_{l=1}^{\kappa-1-i} \xi_{i+l} \frac{1}{\pi} \int_{-1}^{1} \tau^{\kappa-1+i+l} b(\tau) Z(\tau) \, d\tau
\]

\[
= C_i - \sum_{l=1}^{\kappa-1-i} \xi_{i+l} \alpha_{-\kappa-l}.
\]

Introduce the vectors $\xi = (\xi_0, \xi_1, \ldots, \xi_{\kappa-1})^T$, $C = (C_0, C_1, \ldots, C_{\kappa-1})^T$ and the $\kappa \times \kappa$ matrix $A = (A_{ij})$ with entries
\[
A_{ij} = \begin{cases} 
0 & i > j \\
1 & i = j \\
\alpha_{-\kappa-j+i} & i < j.
\end{cases}
\]

To get $\xi_i$, rearrange (40) and solve the triangular system $A\xi = C$. The triangular matrix $A$ is nonsingular since it has nonzero diagonal elements. Hence the inverse matrix $A^{-1} = B = (B_{ij})$ exists. By considering (41), we easily find its elements

\[
B_{ij} = \begin{cases} 
0 & i > j \\
1 & i = j \\
\beta_{-\kappa-j+i} & i < j.
\end{cases}
\]
We have then $\xi = B C$, and use of (??) leads to the claim.

Remark. These proofs in a certain sense use a “dual” argument of the one used in the proof of theorem 2.1 of [?].

We consider the case only the index $\kappa = 0$ for the remainder of the paper. In this case, the regularizing operator $S^*$ becomes the inverse $S^{-1}$ of the dominant operator $S$ because $\tilde{\rho}_{\kappa-1}$ becomes zero in (??).

4. The Galerkin Scheme. Approximating $u$ by a polynomial $u_n$ of degree $\leq n$ and setting the residual $r_n = Su_n + Ku_n - f$ orthogonal to $SU_n$ where $U_n$ is the set of polynomials of degree $\leq n$, which has a basis $\{\phi_l\}_{l=0}^n$,

$$u_n = \sum_{l=0}^n \eta_l \phi_l,$$

we obtain the linear equation to determine $\{\eta_l\}_{l=0}^n$,

$$\sum_{l=0}^n < S\phi_l, S\phi_j > _{Z^{-1}} \eta_l + \sum_{l=0}^n < K\phi_l, S\phi_j > _{Z^{-1}} \eta_l = < f, S\phi_j > _{Z^{-1}}$$

where $\{S\phi_l\}_{l=0}^n$ are orthonormal with respect to $1/Z$, (??) because $\kappa = 0$.

Let $P_n$ be the operator of orthogonal projection onto $V_n = SU_n$. Then $u_n$ satisfies

$$P_n Su_n = -P_n Ku_n + P_n f.$$ (43)

Since $Su_n \in V_n$, $P_n Su_n = Su_n$,

$$Su_n = -P_n Ku_n + P_n f.$$ (44)

From [?, ?, ?], the unique existence of $u_n$ for $n$ large enough and the convergence of $u_n$ to $u$ in $L_Z$ are obtained with

$$||u - u_n||_Z \leq C_g ||Su - P_n Su||_Z^{-1}$$

where $C_g$ is constant. For $f$ and $k \in C^{r,\alpha}$ which is the Hölder space of order $0 \leq \alpha < 1$ for the $r$th derivative and $r > 0$,

$$||u - u_n||_Z = O(n^{-(r+\alpha)}).$$ (45)

Discretizing (??) by $n$ point quadrature rules with the weight $Z^{-1}$ and (??), we obtain $v_n$, called a discrete Galerkin approximation to $u$, as the solution of the resulting functional equations and $v_n$ is given by

$$v_n = \sum_{l=0}^n \tilde{\eta}_l \phi_l,$$ (46)
where the \( \{ \tilde{x}_j \} \) satisfies
\[
\sum_{l=0}^{n} \left[ \sum_{d=1}^{n} w_d^l \psi_l(s_d) \psi_j(s_d) \right] \tilde{x}_l = 0.
\]

(47)
\[
+ \sum_{l=0}^{n} \left[ \sum_{d=1}^{n} \sum_{p=1}^{n} w_d^l w_p k(s_d, t_p) \psi_j(s_d) \phi_l(t_p) \right] \tilde{x}_l = 0
\]

\[
= \sum_{d=1}^{n} w_d^l f(s_d) \psi_j(s_d) \quad \text{for } 0 \leq j \leq n
\]

where the first term becomes \( \tilde{x}_j \| \psi_j \|_{\mathcal{S}^*_j}^2 = \tilde{x}_j \) because of the precision of the quadrature rule.

For the convergence of the discrete Galerkin approximation \( v_n \) to \( u \), we refer the following theorem[?, ?].

**Theorem 4.6.** For \( n \) large enough, the discrete Galerkin approximation \( v_n \) exists uniquely and converge to \( u \) in \( L_2 \) if \( f \) and \( k \in C^{\alpha, \infty} \), \( r + \alpha > 5/2 \). Furthermore \( ||u - v_n||_2 = O(n^{-1-(r+\alpha)}) \), and \( ||u - v_n||_\infty = O(n^{-1-(r+\alpha)+5/2}) \).

Let us define \( \Pi_n : C[-1, 1] \to V_n \) by

\[
\Pi_n u(x) = \sum_{k=0}^{n} \left[ \sum_{j=1}^{n} w_d^j u(s_j) \psi_k(s_j) \right] \psi_k(x).
\]

(48)
Then using (??) and the fact \( \kappa = 0 \), we have
\[
S v_n(x) = S \left[ \sum_{k=0}^{n} \tilde{x}_k \phi_k(x) \right] = \sum_{k=0}^{n} \tilde{x}_k \psi_k(x)
\]

(49)
and
\[
\Pi_n f(x) = \sum_{k=0}^{n} \left[ \sum_{d=1}^{n} w_d^k f(s_d) \psi_k(s_d) \right] \psi_k(x).
\]

(50)
Similarly
\[
\Pi_n K_n v_n(x) = \Pi_n \left[ \sum_{p=1}^{n} w_p k(x, t_p) v_n(t_p) \right]
\]
\[
= \sum_{p=1}^{n} w_p v_n(t_p) \Pi_n k(x, t_p)
\]
\[
= \sum_{p=1}^{n} w_p v_n(t_p) \left[ \sum_{l=0}^{n} \left[ \sum_{d=1}^{n} w_d^l k(s_d, t_p) \psi_l(s_d) \right] \psi_k(x) \right]
\]
\[
= \sum_{p=1}^{n} w_p \left[ \sum_{l=0}^{n} \tilde{x}_l \psi_l(t_p) \right] \left[ \sum_{k=0}^{n} \left[ \sum_{d=1}^{n} w_d^k k(s_d, t_p) \psi_k(s_d) \right] \psi_k(x) \right]
\]
\[
= \sum_{k=0}^{n} \left[ \sum_{l=0}^{n} \tilde{x}_l \left[ \sum_{p=1}^{n} \sum_{d=1}^{n} w_d^l w_p k(s_d, t_p) \psi_k(s_d) \phi_l(t_p) \right] \psi_k(x) \right].
\]
Here
\[
S v_n(x) + \Pi_n K_n v_n(x) - \Pi_n f(x) \\
= \sum_{k=0}^{\infty} \left[ \tilde{z}_k + \sum_{l=0}^{n} \sum_{d=1}^{\infty} w_d^l w_d k(s_d, t_p) \psi_l(s_d) \phi_l(t_p) \right] \tilde{z}_q - \sum_{d=1}^{n} w_d^l f(s_d) \phi_l(s_d) \psi_l(x) \\
= 0
\]
since the inside term of the middle is exactly same to (53). Consequently, the discrete Galerkin approximation \( v_n \) satisfies
\[
(52) \quad S v_n + \Pi_n K_n v_n - \Pi_n f = 0.
\]

5. The Sloan Iterate. In this section, we introduce the Sloan iterate [?, ?, ?] \( \bar{u}_n \) of \( u_n \) given by
\[
(53) \quad S \bar{u}_n = -K u_n + f
\]
where
\[
\bar{u}_n = -S^{-1} K u_n + S^{-1} f.
\]

Here applying the orthogonal projection operator \( P_n \) to (52), we have
\[
P_n S \bar{u}_n = -P_n K u_n + P_n f = P_n S u_n = S u_n.
\]

Therefore \( u_n = S^{-1} P_n S \bar{u}_n = Q_n \bar{u}_n \) where \( Q_n = S^{-1} P_n S \) is the orthogonal projection operator onto \( U_n \). Thus \( \bar{u}_n \) satisfies
\[
(54) \quad S \bar{u}_n = -K Q_n \bar{u}_n + f.
\]

**Lemma 5.7.** The equation (52) has a unique solution for all \( n \) large enough.

**Proof.** Because \( Q_n \) is orthogonal, \( ||K - K Q_n|| \to 0 \) as \( n \to \infty \). Letting \( T_n = S + K Q_n \) and \( T = S + K \), we have \( M = T - T_n = K - K Q_n \) and
\[
T_n = T_n - T + T = T - (T - T_n) \\
= T - M \\
= T(I - T^{-1} M).
\]

Here \( ||T^{-1} M|| < ||T^{-1}|| ||K - K Q_n|| < 1/2 \) for \( n \) large enough. Hence \( T_n^{-1} = (S + K Q_n)^{-1} \) exists since \( (I - T^{-1} M)^{-1} \) exists.

For \( n \) sufficiently large,
\[
(55) \quad u - \bar{u}_n = (S + K)^{-1} f - (S + K Q_n)^{-1} f \\
= (S + K Q_n)^{-1} (K Q_n - K)(S + K) f \\
= (S + K Q_n)^{-1} (K Q_n - K) u \\
= (S + K Q_n)^{-1} K(Q_n - I) u \\
= (S + K Q_n)^{-1} K(Q_n - I)(Q_n - I)(u - u_n) \\
= (S + K Q_n)^{-1} (K Q_n - K)(u - u_n)
\]
since \((Q_n - I)^2 = (Q_n - I)\) and \((I - Q_n)u_n = 0\). From (??), \(\|u - u_n\|_Z = O(n^{-(r+\alpha)})\) if \(k\) and \(f\) are \(C^{r,\alpha}\). Also Jackson’s theorem and the orthogonality of \(Q_n\) give \(\|K - KQ_n\| = O(n^{-(r+\alpha)})\) for \(k \in C^{r,\alpha}\). With these facts and (??), we have the following theorem.

**Theorem 5.8.** Let \(u_n\) be the solution of Galerkin method for \(n\) sufficiently large and the Sloan iterate \(\tilde{u}_n\) be defined by (??). Then \(\tilde{u}_n\) converges in \(L_Z\) to \(u\) and \(\|u - \tilde{u}_n\|_Z = O(n^{-2(r+\alpha)})\) if \(k\) and \(f\) are \(C^{r,\alpha}\) functions.

This theorem shows that the Sloan iterate converges at twice the rate of the Galerkin approximation if all integrals are evaluated exactly. This order of convergence can be obtained only theoretically.

In the remainder of the paper, the discrete Sloan iterate will be defined and compared with the discrete Galerkin approximation.

Ignoring the error term \(\epsilon_H\) in (??) with (??), we define

\[
S_n u(x) \equiv a(x)Z(x)u(x) + b(x) \left[ \sum_{i=1}^{n} \frac{w_i u(t_i)}{t_i - x} + \Lambda_n(x)u(x) \right]
\]

\[
= a(x)Z(x)u(x) + \frac{b(x)}{\pi} \sum_{i=1}^{n} \frac{w_i u(t_i)}{t_i - x}
+ \frac{u(x)}{\phi_n(x)} \left[ (-1)^{\nu} \psi_{n-k}(x) - a(x)Z(x)\phi_n(x) \right]
= \frac{b(x)}{\pi} \sum_{i=1}^{n} \frac{w_i u(t_i)}{t_i - x} + (-1)^{\nu} \frac{\psi_{n-k}(x)}{\phi_n(x)} u(x).
\]

And for the discrete Sloan iterate, we now define \(\hat{v}_n\) by

\[
S_n \hat{v}_n = -K_n v_n + f.
\]

Applying the operator \(\Pi_n\) with (??), then

\[
\Pi_n S_n \hat{v}_n = -\Pi_n K_n v_n + \Pi_n f
= S \hat{v}_n.
\]

Hence \(v_n = S^{-1} \Pi_n S_n \hat{v}_n\) so that

\[
S_n \hat{v}_n + K_n S^{-1} \Pi_n S_n \hat{v}_n = f.
\]

Now we give the main result showing that the discrete Sloan iterate becomes the Nyström quadrature approximation [?, ?].

**Theorem 5.9.** The discrete Sloan iterate \(\hat{v}_n\) of (??) also satisfies

\[
S_n \hat{v}_n + K_n \hat{v}_n = f
\]

which is the equation representing the Nyström approximation of \(u\).
Proof. Letting \( h_n = S^{-1} \Pi_n S_n \hat{v}_n \), we have
\[
K_n S^{-1} \Pi_n S_n \hat{v}_n(x) = K_n h_n
\]
\[
= \sum_{j=1}^{n} w_j k(x, t_j) h_n(t_j).
\]
Thus it suffices to show
\[
h_n(t_l) = \hat{v}_n(t_l) \quad \text{for } 1 \leq l \leq n.
\]

Let \( \tau = S_n \hat{v}_n \) and use (??) and (??) to have
\[
h_n(x) = S^{-1} \Pi_n \tau(x)
\]
\[
= \frac{a(x)}{Z(x)} \Pi_n \tau(x) - \frac{b(x)}{\pi} \int_{-1}^{1} \frac{\Pi_n \tau(t)}{Z(t)(t-x)} dt
\]
\[
= \frac{a(x)}{Z(x)} \Pi_n \tau(x) - b(x) \left[ \frac{1}{\pi} \sum_{j=1}^{n} \frac{w_j^* \Pi_n \tau(s_j)}{s_j - x} + \frac{\Lambda_n(x)}{\psi_n(x)} \Pi_n \tau(x) \right]
\]
since \( \Pi_n \) is a polynomial. From (??),
\[
h_n(x) = \frac{a(x)}{Z(x)} \Pi_n \tau(x) - \frac{b(x)}{\pi} \sum_{j=1}^{n} \frac{w_j^* \Pi_n \tau(s_j)}{s_j - x} + \frac{\Pi_n \tau(x)}{\psi_n(x)} \left[ \phi_n(x) - \frac{a(x)}{Z(x)} \psi_n(x) \right]
\]
\[
eq \frac{-b(x)}{\pi} \sum_{j=1}^{n} \frac{w_j^* \Pi_n \tau(s_j)}{s_j - x} + \frac{\phi_n(x)}{\psi_n(x)} \Pi_n \tau(x).
\]
Evaluating at the zero \( t_l \) of \( \phi_n \), we obtain
\[
h_n(t_l) = \frac{-b(t_l)}{\pi} \sum_{j=1}^{n} \frac{w_j^* \Pi_n \tau(s_j)}{s_j - t_l}
\]
since \( \Pi_n \tau(s_j) = \tau(s_j)(\Pi_n \) is the operator of polynomial interpolation on \( \{s_j\}_{j=1}^{n} \),
\[
= \frac{-b(t_l)}{\pi} \sum_{j=1}^{n} \frac{w_j^* S_n \hat{v}_n(s_j)}{s_j - t_l}.
\]

Hence we have, from (??)
\[
h_n(t_l) = \frac{-b(t_l)}{\pi} \sum_{j=1}^{n} \frac{w_j^*}{s_j - t_l} \left[ \frac{b(s_j)}{\pi} \sum_{i=1}^{n} \frac{w_i \hat{v}_n(t_i)}{t_i - s_j} \right]
\]
\[
= \frac{b(t_l)}{\pi} \sum_{i=1}^{n} w_i \hat{v}_n(t_i) \left[ \frac{1}{\pi} \sum_{j=1}^{n} \frac{w_j^* b(s_j)}{(t_i - s_j)(t_l - s_j)} \right]
\]
\[
= \hat{v}_n(t_l).
\]
The last equality comes from the theorem ??.

The discrete Sloan iterate obtained from a polynomial Galerkin approximation for solving SIEs becomes the Nyström approximation when all integrals are calculated by Gaussian quadratures using zeros of basis polynomials as their nodes. This reads that the discrete Galerkin approximation $v_n$ and the discrete Sloan iterate $\hat{v}_n$ agree at the quadrature nodes. Therefore we do not achieve the computational superconvergence in practice.

REFERENCES

A Heat Loss Comparison Between the Two Parabolic Fin Models Using Two Different Numerical Methods

K.T. Kim*, H.S. Kang**
* : Graduate School, Kangwon University
** : Department of Mechanical Engineering, Kangwon University

Abstract

A comparison of the two dimensional heat loss, computed using the analytical method and the finite difference method in two models (i.e. one is a parabolic fin whose parabolic curves meet at the fin center line and the other is a transformed parabolic fin whose tip cuts vertically), is made assuming the analytical method is correct. For these methods, the root temperature and surrounding convection coefficients of these fins are assumed as constants. The results show that the relative errors of the heat loss between the two methods for the parabolic fin whose tip cuts vertically are smaller than those for the one whose tip does not cut. In case of Bi=0.01, the values of the heat loss obtained using a finite difference method are close to those values obtained using the analytical method for both models. The values of the heat loss from both models calculated by using the analytical method are almost the same for given range of non-dimensional fin length in case of Bi = 0.01 and 0.1.

1. Introduction

There are many papers dealing with the fin so far. Many papers analyze the fin itself while some papers present the performance improvement of the machine by attaching the fin [1, 2] to the machine. It has been shown that the one-dimensional approach [3, 4, 5], is convenient, but may be in error for certain physical conditions. Various shapes of the fin (i.e. rectangular [4~7, 8], triangular [8, 11], trapezoidal [3, 11] and annular [12, 13] etc.) have been studied. No literature seems to be available which presents the analysis of a two dimensional parabolic fin by analytical method. The problem seems to be due to the difficulty in applying the boundary conditions to the fin.

This study performed two-dimensional analyses on the parabolic fins by using two different numerical methods. One is the analytical method and the other is a finite difference method. For parabolic fins, two models are chosen. The shape of model 1 is a parabolic fin whose parabolic curves meet at the fin center line and that of model

Key Words : Parabolic Fin, Heat Loss, Relative Error.
2 is a transformed parabolic fin whose fin tip cuts vertically. For the finite difference methods, 55 nodes for model 1 and 57 nodes for model 2 in the upper half of the fin was convenient. That is, checks run for several different configurations (number of nodes) indicated that these nodes were sufficient for the solutions to be consistent. Further, for this setup, the non-dimensional fin length is restricted to be less than 3 in order to prevent any error which might result due to $\Delta x$ being too large as the non-dimensional fin length increases. Also, the non-dimensional fin length was restricted to be less than 6 in order to decrease errors due to the number of iteration in case of the analytical method. Each analysis was based on the following assumptions: the root temperature, surrounding convection coefficient and fin thermal conductivity are constants and the condition is steady-state.

2. Two-Dimensional Numerical Analysis

2.1 Analytical Method

Based upon the fins illustrated in Fig. 1-a and Fig. 1-b, the two dimensional analysis is governed by the form of the 1st Law of thermodynamics shown as Eq. (1).

$$\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} = 0 \tag{1}$$

Three boundary conditions are represented as Eqs. (2)~(4) for both models and one energy balance equation is shown as Eq. (5) for model 1 and Eq. (6) for model 2.

$$\theta = \theta_0 \text{ at } x = 0 \tag{2}$$

$$\frac{\partial \theta}{\partial x} + Bi \cdot \theta = 0 \begin{cases} \text{at } x = L & \text{for model 1} \\ \text{at } x = LL & \text{for model 2} \end{cases} \tag{3}$$

$$\frac{\partial \theta}{\partial y} = 0 \text{ at } y = 0 \tag{4}$$

$$- \int_{0}^{1} \frac{\partial \theta}{\partial x} \bigg|_{y=0} dy = Bi \cdot \int_{0}^{1} \theta \cdot \sqrt{\frac{L^2}{4y} + 1} dy \tag{5}$$

$$- \int_{0}^{0.05} \frac{\partial \theta}{\partial x} \bigg|_{y=0} dy = Bi \cdot \int_{0.05}^{1} \theta \cdot \sqrt{\frac{L^2}{4y} + 1} dy - \int_{0}^{0.05} \frac{\partial \theta}{\partial x} \bigg|_{y=LL} dy \tag{6}$$

where,

$$x = \frac{x'}{l}, \ y = \frac{y'}{l}, \ L = \frac{L'}{l}, \ LL = 0.776393L, \ Bi = \frac{h \cdot l}{k}, \ \theta = T - T_{cc}, \ \text{and} \ \theta_0 = T_{w} - T_{cc}$$

The value of LL is chosen arbitrary. From Eq. (1) with three boundary conditions and one energy balance equation, the temperature profile can be obtained by the usual separation of variables procedure. The result is

$$\theta = \sum_{n=1}^{\infty} \hat{\theta}_n \cdot N_n \cdot f_n(x) \cdot f_n(y) \tag{7}$$

where,
A HEAT LOSS COMPARISON

\[ N_n = \frac{4 \theta_0 \cdot \sin \lambda_n}{2 \lambda_n + \sin (2 \lambda_n)} \]  

\[ f_n(x) = \frac{\lambda_n \cdot \cosh A_n + B_i \cdot \sinh A_n}{\lambda_n \cdot \cosh C_n + B_i \cdot \sinh C_n} \]  

\[ f_n(y) = \cos (\lambda_n \cdot y) \]

In Eq. (9) \( A_n = \lambda_n (L - x) \), \( C_n = \lambda_n \cdot L \) for model 1 and \( A_n = \lambda_n (LL - x) \), \( C_n = \lambda_n \cdot LL \) for model 2. Eigenvalues for model 1 are obtained by using Eq. (11) derived from Eq. (3) and Eq. (5).

\[ [\lambda_n \cdot \sinh C_n + B_i \cdot \cosh C_n] \cdot \sin (\lambda_n) = 2 \cdot A \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} A_{i,j} \cdot \lambda_n^{2i+2j-3} \]

\[ - A \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} A_{i,j,k} \cdot C_{i,j,k} \cdot D_{i,j,k} \cdot \lambda_n^{2i+2j-3} + C \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} A_{i,j} \cdot E_{i,j} \cdot \lambda_n^{2i+2j-3} \]

\[ + B^2 \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} F_{i,j,k} \cdot (G_{i,j,k} - H_{i,j,k}) \cdot \lambda_n^{2i+2j-3} \]  

(11)

where,

\[ A = B_i \cdot \sqrt{\left\lfloor \frac{L^2}{4} + 1 \right\rfloor} \]  

\[ C = B_i \cdot \ln \left[ \sqrt{\frac{L^2}{4} + 1} - 1 \right] \]  

\[ A_{i,j} = \frac{\Gamma(2i+1) \cdot \Gamma(2j+1) \cdot (2i+4j+5)}{\Gamma(2i+1) \cdot \Gamma(2j+1) \cdot (2i+4j+3)} \]  

\[ C_{i,j} = \frac{L \cdot \Gamma(2i+4j+4) \cdot \Gamma(2i+4j+3)}{4 \cdot \Gamma(2i+2j+2) \cdot \Gamma(i+2j+1)} \]  

\[ D_{i,j,k} = \frac{1}{\Gamma(k+1) \cdot \Gamma(k)} \]  

\[ E_{i,j} = \frac{L \cdot \Gamma(2i+4j+4) \cdot \Gamma(2i+4j+3)}{(-4)^{i+j+2} \cdot k \cdot \Gamma(2i+2j+2)} \]  

\[ F_{i,j} = \frac{(-1)^{i+j+1} \cdot L \cdot \Gamma(i+2j+2)}{\Gamma(2i+1) \cdot \Gamma(2j+1)} \]  

\[ G_{i,j,k} = \frac{2 \cdot \left( \frac{L^2}{4} \right)^{i+j+2} \cdot k \cdot \Gamma(i+2j+2)}{\Gamma(k) \cdot \Gamma(i+2j+2)} \]  

\[ H_{i,j,k} = \frac{(-1)^{i+j+1} \cdot \frac{L^2}{4} \cdot k \cdot \Gamma(i+2j+2)}{\Gamma(k) \cdot \Gamma(i+2j+2)} \]  

(19)  

(20)

Also, eigenvalues for model 2 can be obtained by using Eq. (21) derived from Eq. (3) and Eq. (6).
\[ [\lambda_n \cdot \sin C_n + B i \cdot \cosh C_n] \cdot \sin (\lambda_n) - B i \cdot \sin (0.05 \lambda_n) \]
\[ = AA_n \left[ 2 \cdot \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} A_{i,j} \cdot L_{i,j} \cdot \lambda_n^{2i+2j-4} \right. \]
\[ - \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \left. \left( i + 2j - 2 \right) \sum_{k=1}^{\infty} A_{i,j} \cdot C_{i,j} \cdot D_{i,j,k} \cdot \lambda_n^{2i+2j-4} \cdot M_k \right] \]
\[ + \left( \frac{C}{B} - D \right) \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} A_{i,j} \cdot E_{i,j} \cdot \lambda_n^{2i+2j-5} \right] \]
\[ + 2 \cdot CC_n \times \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} F_{i,j} \cdot (G_{i,j,k} - N_{i,j,k}) \cdot \lambda_n^{2i+2j-5} \right] \quad (21) \]

where,

\[ AA_n = B i \cdot \lambda_n \cdot \cosh D_n - B i^2 \cdot \sinh D_n \quad (22) \]
\[ CC_n = B i^2 \cdot \cosh D_n - B i \cdot \lambda_n \cdot \sinh D_n \quad (23) \]
\[ L_{i,j} = \sqrt{\frac{L^2}{4} + 1 - \frac{\sqrt{5L^2 + 1}}{20} \cdot 2^{-i-2j-2}} \quad (24) \]
\[ M_k = \sqrt{\frac{L^2}{4} + 1 - \frac{\sqrt{5L^2 + 1}}{20} \cdot 2^{-k}} \quad (25) \]
\[ D = \ln \left( \frac{\sqrt{5L^2 + 1}}{20} \right) \quad (26) \]
\[ N_{i,j,k} = \frac{L^2}{4} \cdot \frac{i + 2j - 2}{k + \frac{1}{2}} \cdot \frac{L^2}{4} + 0.05 
\[ D_n = \lambda_n \cdot (L - LL) \quad (28) \]

In Eqs. (14) to (20) and Eq. (27), \( \Gamma \) is a gamma function. Finally the heat loss can be calculated by Eq. (29) for both models.

\[ Q = \int_1^- k \cdot \frac{\partial \theta}{\partial x} \bigg|_{x=0} \cdot dy = 2k \cdot \theta_0 \sum_{n=1}^{\infty} f_n \cdot N_n \cdot \sin (\lambda_n) \quad (29) \]

where,

\[ f_n = \frac{B i \cdot \cosh C_n + \lambda_n \cdot \sinh C_n \cdot \lambda_n \cdot \cosh C_n + B i \cdot \sinh C_n}{\lambda_n \cdot \cosh C_n + B i \cdot \sinh C_n} \quad (30) \]

### 2.2 Finite Difference Method

For this method, 55 nodes were used for the upper half part of the model 1 (see Fig. 2-a) and the equations for each node are shown as Eq. (31) through Eq. (37).

For node 1 (11, 20, 28, 35, 41, 46, 50)
\[ \theta_{h} - [1 + f_1 \cdot (f_1 + f_2) + g_{1,2}] \cdot \theta_1 + [f_1 \cdot (f_1 + f_2)] \cdot \theta_2 + g_{1,2} \cdot \theta_1 = 0 \quad (31) \]

For node 2 (and a similar form for the other interior points - 3, 12 and so on)
\[ \theta_{h} + \frac{1}{2} \cdot [f_1 \cdot (f_1 + f_2)] \cdot \theta_1 - [1 + f_1 \cdot (f_1 + f_2) + g_{1,2}] \cdot \theta_2 \]
A HEAT LOSS COMPARISON

\[ + \frac{1}{2} \cdot [f_1 \cdot (f_1 + f_2)] \cdot \theta_1 + g_{1,2} \cdot \theta_{12} = 0 \] (32)

For node 9 (18, 26, 33, 39, 44, 48)
\[ \theta_0 + [ (f_1 + f_2) \cdot RA_{1,2} ] \cdot \theta_9 - [ 1 + 2 \cdot (f_1 + f_2) \cdot RA_{1,2} + g_{1,2} ] \cdot \theta_9 + \theta_{10} + g_{1,2} \cdot \theta_{19} = 0 \] (33)

For node 10 (19, 27, 34, 40, 45, 49, 52)
\[ \theta_0 + (f_1 \cdot RB_{1,2}) \cdot \theta_9 - [ 1 + (f_1 \cdot RB_{1,2}) + Bi \cdot \Delta k_1 \cdot \sqrt{(RB_{1,2})^2 + 1} ] \cdot \theta_{10} = 0 \] (34)

For node 53
\[ \theta_{50} - [ 1 + f_9 \cdot SA + g_{9,10} ] \cdot \theta_{53} + (f_9 \cdot SA) \cdot \theta_{54} + g_{9,10} \cdot \theta_{55} = 0 \] (35)

For node 54
\[ \theta_{51} + [ f_9 \cdot SB ] \cdot \theta_{53} - [ 1 + f_9 \cdot SB + Bi \cdot \Delta k_9 \cdot \sqrt{SB^2 + 1} ] \cdot \theta_{54} = 0 \] (36)

For node 55
\[ \theta_{53} - [ 1 + Bi \cdot \Delta k_{10} \cdot \sqrt{4 \cdot \left( \frac{\Delta k_{10}}{\Delta y} \right)^2 + 1} ] \cdot \theta_{55} = 0 \] (37)

where,
\[ f_i = \frac{\Delta k_i}{\Delta y} \quad g_{i,j} = \frac{\Delta k_i}{\Delta y} \quad RA_{i,j} = \frac{\Delta k_i}{\Delta y + \Delta y_j} \quad RB_{i,j} = \frac{\Delta k_i + \Delta k_j}{2 \Delta y + \Delta y_i + \Delta y_j} \]
\[ Z \cong 0.293 \quad SA = \frac{\Delta k_9 + 2Z}{\Delta y} \quad SB = \frac{\Delta k_9 + 2Z}{\Delta y} \]

In case of the model 2, 57 nodes were used for the upper half part (see Fig. 2-b) and the equations for each node are shown as Eq. (38) through Eq. (47). To obtain the value of the temperature of each node for both models, Gaussian elimination method is used.

For node 1 (11, 20, 28, 35, 41, 46)
\[ \theta_0 - [ 1 + (f_1 + f_2) \cdot t_{1.10,11} + g_{1,2} ] \cdot \theta_1 + [ (f_1 + f_2) \cdot t_{1.10,11} ] \cdot \theta_2 + g_{1,2} \cdot \theta_{11} = 0 \] (38)

For node 2 (12, 21, 29, 36, 42, 47, 51)
\[ \theta_0 + [ (f_1 + f_2) \cdot u_{1.9,10,11} ] \cdot \theta_1 - [ 1 + 2 \cdot (f_1 + f_2) \cdot u_{1.9,10,11} + g_{1,2} ] \cdot \theta_2 + [ (f_1 + f_2) \cdot u_{1.9,10,11} ] \cdot \theta_3 + g_{1,2} \cdot \theta_{12} = 0 \] (39)

For node 3 (and a similar form for the other interior points - 3, 12 and so on)
\[ \theta_0 + [ (f_1 + f_2) \cdot v_{1.8,9} ] \cdot \theta_1 - [ 1 + 2 \cdot (f_1 + f_2) \cdot v_{1.8,9} + g_{1,2} ] \cdot \theta_3 + [ (f_1 + f_2) \cdot v_{1.8,9} ] \cdot \theta_4 + g_{1,2} \cdot \theta_{13} = 0 \] (40)

For node 10 (19, 27, 34, 40, 45, 49, 52)
\[ \theta_0 + [ (f_1 + f_2) \cdot v_{1.1,2} ] \cdot \theta_0 \]
\[
[1 + (f_1 + f_2) \cdot v_{1,1,2} + B_i \cdot \Delta x_i \cdot \sqrt{(RB_{1,2})^2 + 1}] \cdot \theta_{10} = 0
\]  
(41)

For node 50
\[
\theta_{46} - \left[ 1 + (f_8 + f_9) \cdot V A_{8,10,11} + \frac{1}{2} \cdot \left( V B_{10,11} + V C_{10,11} \right) \cdot g_{8,9} \right] \cdot \theta_{50} + \left[ (f_8 + f_9) \cdot V A_{8,10,11} \right] \cdot \theta_{51} + \frac{1}{2} \cdot g_{8,9} \cdot \left( V B_{10,11} + V C_{10,11} \right) \cdot \theta_{54} = 0
\]  
(42)

For node 53
\[
\theta_{50} - \left[ 1 + V D_{9,10,11} \cdot \frac{4 \Delta x_i}{\Delta y_{11}} + g_{9,10} \right] \cdot \theta_{53} + \left[ V D_{9,10,11} \cdot \frac{4 \Delta x_i}{\Delta y_{11}} \right] \cdot \theta_{54} + g_{9,10} \cdot \theta_{56} = 0
\]  
(43)

For node 54
\[
\theta_{50} + \left[ V D_{9,10,11} \cdot V E_{9,10,11} \right] \cdot \theta_{53} - \left[ 1 + V D_{9,10,11} \cdot V E_{9,10,11} + V F_{9,10,11} \cdot g_{9,10} \right] \cdot \theta_{54} + \left[ V F_{9,10,11} \cdot V E_{9,10,11} \right] \cdot \theta_{55} + g_{9,10} \cdot \theta_{57} = 0
\]  
(44)

For node 55
\[
\theta_{51} + \left[ V F_{9,10,11} \cdot u_{9,9,10,11} \right] \cdot \theta_{56} - \left[ 1 + V F_{9,10,11} \cdot u_{9,9,10,11} + B_i \cdot \Delta x_i \cdot \sqrt{(u_{9,9,10,11})^2 + 1} \right] \cdot \theta_{55} = 0
\]  
(45)

For node 56
\[
\theta_{53} - \left[ 1 + 4 \cdot \left( \frac{\Delta x_i}{\Delta y_{11}} \right)^2 \right] + B_i \cdot \Delta x_i \cdot \theta_{56} + \left[ 4 \cdot \left( \frac{\Delta x_i}{\Delta y_{11}} \right)^2 \right] \cdot \theta_{57} = 0
\]  
(46)

For node 57
\[
\theta_{54} + \left[ \frac{\Delta x_i}{\Delta y_{11}} \cdot V E_{10,11} \right] \cdot \theta_{56} - \left[ 1 + \frac{\Delta x_i}{\Delta y_{11}} \cdot V E_{10,11} \right] + B_i \cdot \left( \frac{\Delta x_i}{\Delta y_{11}} \cdot \frac{4 \Delta x_i}{\Delta y_{11}} \right) \cdot \theta_{57} = 0
\]  
(47)

where,
\[
l_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k}, \quad u_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k}, \quad v_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k},
\]
\[
VA_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k}, \quad VB_{i,j} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k}, \quad VC_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k},
\]
\[
VD_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k}, \quad VE_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k}, \quad VF_{i,j,k} = \frac{\Delta x_i}{\Delta y_j} + \frac{\Delta y_j}{\Delta y_k},
\]

The definition of the values \( \Delta x_i, \Delta y_i \) (\( i = 1, 2 \)) which are used in this finite difference method is shown in Fig. 2-c and the rest \( \Delta x_i, \Delta y_i \) (\( i = 3, 4, \cdots \)) can be written in the same procedure.
3. Results and Discussions

Figure 3 presents the relative error of the heat loss in the finite difference method as compared to the analytical method as $L$ varies from 0.5 to 3.0 for several values of Biot number in case of the model 1. The relative error increases slowly and is less than 5% for given range of $L$ in case of $Bi = 0.01$. It is shown that the results become not so good as the value of $Bi$ and $L$ increase.

Results for the same conditions as in Fig. 3 except that the model 1 is changed to the model 2 are depicted in Fig. 4. The trend of the results are similar but the overall relative errors decrease considerably in comparison with Fig. 3. When $Bi = 0.01$, the relative error varies from 0.17% to 1.26% for given range of $L$. Also the relative error is less than 5% for given range of $L$ in case of $Bi = 0.05$ and for until $L = 2$ in case of $Bi = 0.1$. These improvement seem to be obtained by proper choosing the nodes for model 2.

Figure 5 shows the relative error of the heat loss in the finite difference method as compared to the analytical method as Biot number varies from 0.01 to 1.0 for $L = 0.5, 1.0, 2.0$ and 3.0 in case of the model 1. In the case of $L \leq 1.0$, the relative error increases regularly as the value of Biot number increases while it increases rapidly until $Bi = 0.1$ and varies somewhat irregularly in the range of $0.1 < Bi < 1.0$ for $L = 2.0$ and 3.0.

Figure 6 represents the same type of information as was presented as Fig. 5 but for model 2. For model 2, the relative errors increase somewhat regularly as Biot number increases for all values of $L$. The relative error for model 2 is less than that for model 1 overall and the relative error for $L = 0.5$ is less than 4% for given range of Biot number. This figure shows that both two methods can be used for model 2 because Biot number is considered to be less than 0.1 in our usual circumstance.

Figure 7 shows the variation of the heat loss from the two fin models for $0.5 \leq L \leq 6.0$ and $Bi = 0.01$ when the analytical method was used. To compare the heat loss from the two fin models, $L$ for model 2 is changed to $1.288L$ so the value of $LL$ for model 2 becomes to the same as the value of $L$ for model 1. The heat loss from the model 1 is almost the same that from the model 2 for given range of $L$. Precisely, the heat loss from the model 2 is slightly less than that from the model 1 until $L = 2.5$ and vice versa over $L = 2.5$. Figure 8 presents the same type of information as Fig. 7 except $Bi = 0.1$. In case of $Bi = 0.1$, the values of the heat loss increase almost linearly as $L$ increases while those values increase somewhat parabolically as $L$ increases for $Bi = 0.01$. This figure shows the heat loss from the model 1 is greater than that from the model 2 for given range of $L$ and the difference between two values is diminish as $L$ increases. Finally, last two figures show that the heat loss does not change remarkably as the model 1 changes to the model 2 for given range of Biot number and the non-dimensional fin length.

4. Conclusion

The relative error of the heat loss between analytical method and finite difference method for the fin model 2 diminishes considerably as comparing that for the fin model 1 for given range of the non-dimensional fin length and Biot number. These results seem to be due to the proper setting up the nodes for the fin model 2 when the finite difference method is used. So it seems to be that more exact results can be obtained if nodes are set up properly and the number of nodes increases in the case of using the finite difference method. Finally, even though the shape of the fin changed from the parabolic fin whose parabolic curves meet at the fin center line to the fin whose fin tip cuts vertically at the ratio $LL = 0.776L$, the value of the heat loss calculated by using the two-dimensional analytical method does not change remarkably for given range of Biot number and the non-dimensional fin length.
References


Fig. 1-a Geometry of a parabolic fin (model 1)

Fig. 1-b Geometry of a parabolic fin whose tip is cut (model 2)
Fig. 2-a Upper half fin geometries presenting 55 nodes for model 1

Fig. 2-b Upper half fin geometries presenting 57 nodes for model 2

Fig. 2-c The definition of notation for a finite difference method
FIG. 3 Relative error of the heat loss versus the non-dimensional fin length for model 1

FIG. 4 Relative error of the heat loss versus the non-dimensional fin length for model 2
Fig. 5 Relative error of the heat loss versus the Biot number for model 1

Fig. 6 Relative error of the heat loss versus the Biot number for model 2
Fig. 7 The heat loss obtained by using the analytical method versus the non-dimensional fin length for $Bi=0.01$

Fig. 8 The heat loss obtained by using the analytical method versus the non-dimensional fin length for $Bi=0.1$
An Algorithm for Performance Index of Telecommunications Network

YongYeon Shin, SoYeon Lee, JaiSang Koh
Electronics and Telecommunications Research Institute
161 Kajong-dong, Yusong-gu, Taejon, 305-350, Korea

Abstract
Performance index is a measure of telecommunications network integrating reliability and capacity simultaneously. This paper suggests a computerized algorithm evaluating a performance index for telecommunications network and compares this computerized algorithm with the algorithm[1] by experimenting on several benchmarks. A computerized algorithm proposed by this paper is superior to the algorithm[1] with respect to the computation time for most of the benchmarks.

AMS Mathematics Subject Classification: 90B12, 90B15
Key word and phrases: link capacity, performance index, telecommunications network

1. Introduction
Any telecommunications network can be modeled as a graph $G = (V, E)$, where $V$ is the set of nodes (or vertices) and $E$ is the set of links (or edges). There are two measures, network reliability and network capacity to evaluate the network performance. Traditionally, these two measures are used independently while neither is a true measure of the performance of the telecommunications network. Recent studies [1,2,3,5,7] have suggested a composite performance index of a network, integrating the important measures of reliability and capacity.

In this paper a manual technique which is proposed by Rushdi[5] has been computerized and a comparison of this computerized algorithm (henceforth computerized algorithm) with the Aggarwal’s algorithm[1] has been carried out by experimenting on published benchmarks. The computerized algorithm proposed in this paper outperforms the algorithm [1] with respect to the computation time.
for most of the benchmarks.

Section 2 introduces the assumptions and notations used in this paper. Section 3 presents a computerized algorithm to evaluate performance index of telecommunications network and section 4 provides an example to illustrate the computerized algorithm proposed in this paper. Section 5 shows the comparison of two algorithms. We conclude the paper in Section 6.

2. Assumptions

a. The telecommunications network is modeled by a graph \( G = (V, E) \) whose nodes are perfectly reliable and of unlimited capacities.

b. Each network link can have only two states, good or failed. The link failures are statistically independent.

c. Each network link is assigned specific values for its reliability and capacity. The link capacity is the upper bound on the link flow in either direction.

Notations

- \( s, t \): source, terminal node
- \( n \): number of links of the network
- \( m \): number of minimal paths of the network
- \( k \): number of minimal cutsets of the network
- \( MC_j \): Minimal cutset \( j \) of the network
- \( MP_i \): Minimal path \( i \) of the network
- \( p_l \): reliability of link \( l \)
- \( c_l \): capacity of link \( l \)

\[ C_{\text{max}} = \min_j \left\{ \sum_{l \in MC_j} c_l \right\} \]

PI Performance Index; the mean value of the source to terminal capacity normalized by its maximum:

\[ \text{PI} = \frac{E[C_{st}]}{C_{\text{max}}} \]

where \( C_{st} \) is a source to terminal network capacity
3. Algorithm

A manual technique[5] using a generalized cutset procedure has been computerized as follows:

Main Algorithm

1. Find all minimal cutsets \( \{MC_j\} j = 1, 2, \cdots, k \) of the network[6].

2. Find all minimal paths \( \{MP_i\} i = 1, 2, \cdots, m \) of the network. It is better (but not necessary) if the paths are enumerated in order of their cardinality [4].

3. Define an \( n \)-dimensional vector \( E_i = (e_{i1}, e_{i2}, \cdots, e_{im}), i = 1, \cdots, m \) corresponding to \( MP_i \) such that
   \[
   e_{il} = 1 \quad \text{if} \quad l \in MP_i \\
   e_{il} = 0 \quad \text{otherwise}
   \]

4. Define an \( n \)-dimensional vector \( V_j = (v_{j1}, v_{j2}, \cdots, v_{jn}), j = 1, \cdots, k \) corresponding to \( MC_j \) such that
   \[
   v_{jl} = c_l \quad \text{if} \quad l \in MC_j \\
   v_{jl} = 0 \quad \text{otherwise}
   \]

5. \( \text{persum} = 0 \)

6. For \( i = 1 \) To \( m \)

7. Separate each minimal path to disjoint paths (use sum of disjoint products method: see Appendix) : let \( d \) be the number of disjoint paths \( P_1, P_2, \cdots, P_d \)

8. For \( u = 1 \) To \( d \)

9. For \( j = 1 \) To \( k \)

   modified \( n+1 \) dimensional vector \( V_j = (b_{j1}, v_{j1}, v_{j2}, \cdots, v_{jn}) \) is derived from \( V_j^* \) as follows:
\[
\begin{align*}
\begin{cases}
  b_{jl} = \sum_{l \in D_u} v_{jl}^* \\
v_{jl} = 0 & \text{if } l \in D_u \cup \overline{D_u} \\
v_{jl} = v_{jl}^* & \text{if } l \in D_u \cup \overline{D_u}
\end{cases}
\]

where \( D_u \) is a set of links themselves belonging to \( P_u \)
\( \overline{D_u} \) is a set of links whose complement is belong to \( P_u \)

Next \( j \)

10. \( \text{persum} = \text{persum} + \text{SUB}(V) \times \text{Probability of disjoint path } P_u \)

where \( V = (V_1, V_2, \cdots, V_k)^t \)

Next \( u \)

Next \( i \)

11. Calculate \( C_{\text{max}} \)

12. \( PI = \text{persum} / C_{\text{max}} \)

\( \text{SUB}(V) \)

1. \( \text{per} = \min_j \{ b_{jl} \} \) and \( b_{jl} = b_{jl}^* - \min_j \{ b_{jl}^* \} \), \( j = 1, 2, \cdots, k \)

2. \( g = 0 \)

3. If there are \( V_j \)'s such that \( v_{jl} = 0 \) for all \( l = 1, 2, \cdots, n \) Then take one \( V_{j_0} \) with \( b_{j_0l}^* = \min_j \{ b_{jl}^* \} \) among them and delete all \( V_j \) such that \( b_{j_0l} \leq b_{jl} \) (In this case, \( V_j \) may have nonzero entries \( v_{jl} \)) and \( g = \) the number of deleted \( V_j \) (i.e. \( g = |V_j| \))

End If

4. \( k = k - g \)

5. If \( k > 1 \) Then

1) If there is \( V_j \) such that \( b_{jl} = 0 \) and all entries \( v_{jl} \) are zero except one \( v_{jh} \) (\( 1 \leq h \leq n \)) Then
take the reliability $p_h$ of link $h$ and

$$\text{per} = \text{per} + p_h \times \text{SUB}(W)$$

where $W_j = (b_{j1}, v_{j1}, v_{j2}, \ldots, v_{jn}), j = 1, \ldots, k$

is modified from $V_j$ as follows:

$$\begin{align*}
b_{j1} &= b_{j1} + v_{jh} \\
v_{jh} &= 0 \\
v_{jl} &= v_{jl}, \ l \neq h, l = 1, 2, \ldots, n
\end{align*}$$

2) Else

take the reliability $p_{l_0}$ of any link $l_0$ which $v_{jl_0} \neq 0$ and

$$\text{per} = \text{per} + p_{l_0} \times \text{SUB}(W) + (1 - p_{l_0}) \times \text{SUB}(\overline{W})$$

where $W_j = (b_{j1}, v_{j1}, v_{j2}, \ldots, v_{jn}), j = 1, \ldots, k$

is modified from $V_j$ as follows:

$$\begin{align*}
b_{j1} &= b_{j1} + v_{j0} \\
v_{j0} &= 0 \\
v_{jl} &= v_{jl}, \ l \neq l_0, l = 1, 2, \ldots, n
\end{align*}$$

and $\overline{W}_j = (b_{j1}, v_{j1}, v_{j2}, \ldots, v_{jn}), j = 1, \ldots, k$

is modified from $V_j$ as follows:

$$\begin{align*}
b_{j1} &= b_{j1} \\
v_{j0} &= 0 \\
v_{jl} &= v_{jl}, \ l \neq l_0, l = 1, 2, \ldots, n
\end{align*}$$

End If

End If

6. If $k = 1$ Then

$$\text{per} := \text{per} + b_{h1}$$

End If

7. $\text{SUB} = \text{per}$
4. Example

We now illustrate the preceding algorithm using bridge network of Fig.1, where the links are numbered as shown. Their probabilities and capacities are:

\[ p_l = 0.9 \text{ for all } l, \quad c_1 = 10, \quad c_2 = 4, \quad c_3 = 5, \quad c_4 = 3, \quad c_5 = 4 \]

![Bridge Network](image)

Fig.1. Bridge Network

The network of Fig. 1 has 4 minimal cutsets

\[ MC_1 = \{1, 2\}, \quad MC_2 = \{4, 5\}, \quad MC_3 = \{1, 3, 5\}, \quad MC_4 = \{2, 3, 4\} \]

and 4 minimal paths

\[ MP_1 = \{1, 4\}, \quad MP_2 = \{2, 5\}, \quad MP_3 = \{1, 3, 5\}, \quad MP_4 = \{2, 3, 4\}. \]

Define 5-dimensional vectors

\[ E_1 = (1, 0, 0, 1, 0), \quad E_2 = (0, 1, 0, 0, 1), \quad E_3 = (1, 0, 1, 0, 1), \quad E_4 = (0, 1, 1, 0, 1) \]

and

\[ V_1^* = (10, 4, 0, 0, 0), \quad V_2^* = (0, 0, 0, 3, 4), \quad V_3^* = (10, 0, 5, 0, 4), \quad V_4^* = (0, 4, 5, 3, 0). \]

Let persum = 0

For \( i = 1 \)

Separate minimal path \( MP_1 = \{1, 4\} \) to disjoint paths:

(In case \( i = 1 \), there is no predecessor minimal path to compare and skip step 7).

Write \( MP_1 \) for \( P_1 \); i.e. \( P_1 = \{1, 4\} \)

Then modified 5+1 dimensional vectors \( V = (V_1, V_2, V_3, V_4) \) are

\[ V_1 = (10, 0, 4, 0, 0, 0), \]
\[ V_2 = (3, 0, 0, 0, 0, 4), \]
\[ V_3 = (10, 0, 0, 5, 0, 4), \]
\[ V_4 = (3, 0, 4, 5, 0, 0) \]

persum \( = 0 + \text{SUB}(V)^n \cdot p_1 p_4 \)
ALGORITHM FOR PERFORMANCE INDEX OF TELECOMMUNICATIONS

\[
\text{SUB}(V) \quad \text{per} = 3 \quad \text{and} \quad b_{11} = 7, \quad b_{21} = 0, \quad b_{31} = 7, \quad b_{41} = 0
\]

\[
V_1 = (7, 0, 4, 0, 0, 0), \\
V_2 = (0, 0, 0, 0, 0, 4), \\
V_3 = (7, 0, 0, 5, 0, 4), \\
V_4 = (0, 0, 4, 5, 0, 0)
\]

Since there is no \( V_j \) deleted from \( V \), \( g = 0 \) and \( k = 4 \).

In \( V_2 \), all entries are zero except \( v_{25} = 4 \).
\[
\text{per} = \text{per} + p_5 \ast \text{SUB}(V) \quad \text{where} \quad V_1 = (7, 0, 4, 0, 0, 0), \\
V_2 = (4, 0, 0, 0, 0, 0), \\
V_3 = (11, 0, 0, 5, 0, 0), \\
V_4 = (0, 0, 4, 5, 0, 0)
\]

\[
\text{SUB}(V) \quad \text{per} = 0 \quad \text{and} \quad b_{11} = 7, \quad b_{21} = 4, \quad b_{31} = 11, \quad b_{41} = 0
\]

\[
V_1 = (7, 0, 4, 0, 0, 0), \\
V_2 = (4, 0, 0, 0, 0, 0), \\
V_3 = (11, 0, 0, 5, 0, 0), \\
V_4 = (0, 0, 4, 5, 0, 0)
\]

In \( V_2 \), \( v_{2l} = 0 \) (\( l = 1, 2, 3, 4, 5 \)) except \( b_{21} = 4 \).

Thus delete \( V_1, V_3 \) and \( g = 2 \) and \( k = 2 \).

Take the reliability \( p_2 \) of link 2 which \( v_{42} \neq 0 \)
\[
\text{per} = 0 + p_2 \ast \text{SUB}(V) + (1 - p_2) \ast \text{SUB}(\bar{V})
\]
where
\[
\begin{align*}
V_2 &= (4, 0, 0, 0, 0, 0) \\
V_4 &= (4, 0, 0, 5, 0, 0)
\end{align*}
\]
\[
\begin{align*}
\bar{V}_2 &= (4, 0, 0, 0, 0, 0) \\
\bar{V}_4 &= (0, 0, 0, 5, 0, 0)
\end{align*}
\]

\[
\text{SUB}(V) \quad \text{per} = 4
\]

\[
V_2 = (0, 0, 0, 0, 0, 0) \\
V_4 = (0, 0, 0, 5, 0, 0)
\]
Now take $V_2$ and delete $V_4$ and $g = 1$ and $k = 1$.

$$\text{per} = 4 + 0 = 4$$

$\text{SUB}(\vec{V})$

$$\text{per} = 0 \text{ and } g = 0 \text{ and } k = 2.$$  

$$\text{per} = 0 + p_3 \ast \text{SUB}(\vec{V})$$

where $\vec{V}_2 = (4, 0, 0, 0, 0, 0)$  

$\vec{V}_4 = (5, 0, 0, 0, 0, 0)$

$\text{SUB}(\vec{V})$

$$\text{per} = 4$$  

$\vec{V}_2 = (0, 0, 0, 0, 0)$  

$\vec{V}_3 = (1, 0, 0, 0, 0)$

Now take $V_2$ and delete $V_4$ and $g = 1$ and $k = 1$.

$$\text{per} = 4 + 0 = 4$$

Thus $\text{per} = 3 + p_3 \ast \text{SUB}(V)$

$$= 3 + p_3 \ast (p_2 \ast \text{SUB}(V) + (1 - p_2) \ast \text{SUB}(\vec{V}))$$

$$= 3 + p_3 \ast (p_2 \ast 4 + (1 - p_2) \ast (p_3 \ast \text{SUB}(\vec{V})))$$

$$= 3 + p_3 \ast (p_2 \ast 4 + (1 - p_2) \ast (p_3 \ast 4))$$

$$= 3 + 4p_3(p_2 + (1 - p_2)p_3)$$

$$\text{persum} = 0 + (3 + 4p_3(p_2 + (1 - p_2)p_3))p_1p_4$$

For $i = 2$

Separate minimal path $MP_2 = \{2, 5\}$ to disjoint paths $P_1 = \{\tilde{1}, 2, 5\}$ and $P_2 = \{1, \tilde{4}, 2, 5\}$.

For $d = 1$

modified 5+1 dimensional vectors $V = (V_1, V_2, V_3, V_4)^t$ are

$V_1 = (4, 0, 0, 0, 0, 0)$,

$V_2 = (4, 0, 0, 0, 3, 0)$,

$V_3 = (4, 0, 0, 5, 0, 0)$,

$V_4 = (4, 0, 0, 5, 3, 0)$

$$\text{persum} = 3 + 4p_3(p_2 + (1 - p_2)p_3))p_1p_4 + \text{SUB}(V)^* (1 - p_1)p_2p_3$$
SUB(V)

per = 4

\[ V_1 = (0, 0, 0, 0, 0, 0), \]
\[ V_2 = (0, 0, 0, 0, 3, 0), \]
\[ V_3 = (0, 0, 0, 5, 0, 0), \]
\[ V_4 = (0, 0, 0, 5, 3, 0) \]

Now take \( V_1 \) and delete \( V_2, V_3, V_4 \) and \( g = 3 \) and \( k = 1 \).

\[ \text{per} = 4 + 0 = 4 \]
\[ \text{persum} = (3 + 4p_5(p_2 + (1-p_2)p_3)) p_1 p_4 + 4(1-p_1)p_2 p_5 \]

For \( d = 2 \)

modified 5+1 dimensional vectors \( V = (V_1, V_2, V_3, V_4) \) are

\[ V_1 = (14, 0, 0, 0, 0, 0), \]
\[ V_2 = (4, 0, 0, 0, 0, 0), \]
\[ V_3 = (14, 0, 0, 5, 0, 0), \]
\[ V_4 = (4, 0, 0, 5, 0, 0) \]

\[ \text{persum} = (3 + 4p_5(p_2 + (1-p_2)p_3)) p_1 p_4 + 4(1-p_1)p_2 p_5 + \text{SUB}(V)^* p_1 (1-p_4)p_2 p_5 \]

SUB(V)

per = 4

\[ V_1 = (10, 0, 0, 0, 0, 0), \]
\[ V_2 = (0, 0, 0, 0, 0, 0), \]
\[ V_3 = (10, 0, 0, 5, 0, 0), \]
\[ V_4 = (0, 0, 0, 5, 0, 0) \]

Now take \( V_2 \) and delete \( V_1, V_3, V_4 \) and \( g = 3 \) and \( k = 1 \).

\[ \text{per} = 4 + 0 = 4 \]
\[ \text{persum} = (3 + 4p_5(p_2 + (1-p_2)p_3)) p_1 p_4 + 4(1-p_1)p_2 p_5 + 4 p_1 (1-p_4)p_2 p_5 \]

After the iterations \( i = 3, 4 \), we get

\[ \text{persum} = (3 + 4p_5(p_2 + (1-p_2)p_3)) p_1 p_4 + 4(1-p_1)p_2 p_5 + 4 p_1 (1-p_4)p_2 p_5 \]
\[ + 4 p_1 (1-p_2)p_3 (1-p_4)p_2 p_5 + 3 (1-p_1)p_2 p_3 p_4 (1-p_5) \]

\[ C_{max} = \min \{ 10+4, 3+4, 10+5+4, 4+5+3 \} = 7 \]
5. Comparison of Algorithms

We implemented the two algorithms, computerized algorithm and Aggarwal’s algorithm[1], in the visual basic language and executed them in IBM 586 PC using benchmarks as shown in Fig. 2. The performance of two algorithms are compared in terms of computation time. Table 1 shows that computerized algorithm proposed in this paper outperforms Aggarwal’s algorithm[1] in most of the benchmarks, as shown in Fig.2. We do not present the input data (link reliabilities, link capacities) of the benchmarks here; Details can be obtained by the authors.

<table>
<thead>
<tr>
<th>Network</th>
<th>Aggarwal’s Algorithm[1]</th>
<th>Computerized Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.11</td>
<td>0.17</td>
</tr>
<tr>
<td>B</td>
<td>2.08</td>
<td>0.50</td>
</tr>
<tr>
<td>C</td>
<td>2.75</td>
<td>0.71</td>
</tr>
<tr>
<td>D</td>
<td>1.04</td>
<td>0.38</td>
</tr>
<tr>
<td>E</td>
<td>13.02</td>
<td>2.58</td>
</tr>
<tr>
<td>F</td>
<td>20.10</td>
<td>2.80</td>
</tr>
<tr>
<td>G</td>
<td>81.45</td>
<td>9.28</td>
</tr>
<tr>
<td>H</td>
<td>43.17</td>
<td>4.06</td>
</tr>
<tr>
<td>I</td>
<td>1,525.61</td>
<td>339.93</td>
</tr>
<tr>
<td>J</td>
<td>27.14</td>
<td>2.30</td>
</tr>
<tr>
<td>K</td>
<td>44,695.70</td>
<td>11,169.38</td>
</tr>
</tbody>
</table>

Computation times can be changed a bit depend on the windows environment. We have tried ten times for each benchmark and the computation times on the Table 1 are the average of them.
Performance index is an important measure of telecommunications network integrating reliability and capacity simultaneously. This paper presents a computerized algorithm which evaluates a performance index for telecommunications network and compares this computerized algorithm with the Aggarwal’s algorithm[1] in terms of computation time. The results of running eleven problems in Table 1 show that even though the worst-case analysis of the number of subproblems (sum of disjoint product terms) in computerized algorithm has exponential time it is more efficient than [1] in that less computation times are needed. This computerized algorithm is quite useful for evaluating a performance index of telecommunications network.
References


Appendix

The procedure of main algorithm #7 (sum of disjoint products method) has been added to make the paper self-contained.

7. If there are any nonzero entries in $E_f$ ($f = 1, \ldots, i-1$), corresponding to zero entries in $E_i$

   1) Form a list of complemented products (as indicated by continuous overbars) for each $f = 1, \ldots, i-1$.

   2) Multiply the complemented products successively from left to right, using Boolean algebra theorems for simplification after each multiplication; let the disjoint terms obtained by the above be $R_1, R_2, \ldots, R_d$.

   3) Disjoint terms $R_1, R_2, \ldots, R_d$ are multiplied by $MP_i$ and obtain the disjoint product terms $MP_i R_1, MP_i R_2, \ldots, MP_i R_d$. 
write $MP_u R_u$ for $P_u$ in convenience for each $u = 1, \ldots, d$.

Example

We will explain the above procedure with Fig. 1.

5-dimensional vectors $E_i$ corresponding to 4 minimal paths

$MP_1 = \{1, 4\}, \ MP_2 = \{2, 5\}, \ MP_3 = \{1, 3, 5\}, \ MP_4 = \{2, 3, 4\}$ are:

$E_1 = (1, 0, 0, 1, 0), \ E_2 = (0, 1, 0, 0, 1), \ E_3 = (1, 0, 1, 0, 1), \ E_4 = (0, 1, 1, 0, 0)$

In case $i = 1$, there is no predecessor 5-dimensional vector to compare with $E_1$.

So we do not separate $MP_1 = \{1, 4\}$.

In case $i = 2$, nonzero entries in $E_1$ corresponding to zero entries in $E_2$ are \{1, 4\}.

Form a complemented product $14$ and $14 \equiv 1 + 14$ by boolean algebra.

Thus disjoint terms \{1\} , \{1, 4\} are multiplied by $MP_2$ and

obtain the disjoint product terms $P_1 = \{1, 2, 5\}$ and $P_2 = \{1, 4, 2, 5\}$.

In case $i = 3$, nonzero entries in $E_1$ corresponding to zero entries in $E_3$ are \{4\} and nonzero entries in $E_2$ corresponding to zero entries in $E_3$ are \{2\}.

Multiply the complemented products \{4\} \{2\} and \{4\} \{2\} is multiplied by $MP_3$ and

obtain the disjoint product terms $P_1 = \{4, 2, 1, 3, 5\}$.

In case $i = 4$, nonzero entries in $E_1$ corresponding to zero entries in $E_4$ are \{1\} and nonzero entries in $E_2$ corresponding to zero entries in $E_4$ are \{5\} and nonzero entries in $E_3$ corresponding to zero entries in $E_4$ are \{1, 5\}

Multiply the complemented products \{1\} \{5\} \{15\} and \{15\} is deleted by boolean algebra.

\{1\} \{5\} is multiplied by $MP_4$ and obtain the disjoint product terms $P_1 = \{1, 5, 2, 3, 4\}$.
ON MONOTONICITY OF ENTROPY

YOUNGSOO LEE

ABSTRACT. In this paper we define the entropy rate and stationary Markov chain and we show the monotonicity of entropy per element and prove that the random tree $T_n$ grows linearly with $n$.

1. INTRODUCTION

The asymptotic equipartition property (A.E.P) states that $\frac{1}{n} \log \frac{1}{p(X_1, X_2, \cdots, X_n)}$ is close the entropy $H$, where $X_1, X_2, \cdots, X_n$ are independent identically distribution (i.i.d) random variables and $p(X_1, X_2, \cdots, X_n)$ is the probability of observing the sequence $X_1, X_2, \cdots, X_n$, $p(X_1, X_2, \cdots, X_n)$ is close to $\sum 2^{-nH}$ with high probability.

Let $B_\delta^{(n)} < \omega^n$ be any set with $\Pr\{B_{\delta}^{(n)}\} \geq 1 - \delta$ and let $X_1, X_2, \cdots, X_n$ be i.i.d. Then the theorem 3.4 see that $\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta$ for $n$ sufficiently large.

When the limit exists, we define two definitions of entropy rate for a stochastic process as follows

$H(\omega) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n),$

$H'(\omega) = \lim_{n \to \infty} H(X_n \mid X_{n-1}, X_{n-2}, \cdots, X_1).$

In particular, for a stationary stochastic process,

$H(\omega) = H'(\omega).$

In this paper we will show that the theorem 4.4 is established. In detail, the contents of this paper is as follows. In 2, we explain the terminology of typical set and entropy. In 3, we define typical set and we prove the theorem 3.4. In 4, we define entropy rate and we prove the theorem 4.4, theorem 4.5.

Key words and phrases: Entropy, information theory

This research was supported by Woosuk University Research Grant.

Typeset by AMS-\TeX
2. Preliminary

Let $X$ be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function by $p(x)$. Then the entropy $H(X)$ of a discrete random variable $X$ is defined by

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x).$$

We often denote the $H(X)$ as $H(p)$ and entropy is expressed in bits. $H(X,Y)$ of a pair of discrete random variable $(X,Y)$ with a joint distribution $p(x,y)$ is called the joint entropy and it is defined as

$$H(X,Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y).$$

Also $H(Y|X)$ is called the conditional entropy as

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

$$= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$

$$= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x).$$

The relative entropy between two probability mass function $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$ 

The relative entropy between the joint distributions and the product distributions $p(x), p(y)$ are called the mutual information and it is represented as

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$ 

The following properties are well-known ([2],[3],[9])

(i) $H(X) \geq 0$

(ii) For any two random variables $X, Y$, $H(X|Y) \leq H(X)$

(iii) $H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$.

The random variables $X_i$ are independent iff equality holds.

(iv) $H(X) \leq \log |\mathcal{X}|$ where $X$ is uniformly distributed over $\mathcal{X}$ iff equality holds.

The joint entropy and conditional entropy can make the chain rule as follows.

$$H(X,Y) = H(X) + H(Y|X)$$
Indeed,

\[ H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) \]

\[ = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x) p(y|x) \]

\[ = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x) \]

\[ = - \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x) \]

\[ = H(X) + H(Y|X). \]

**Proposition 1.1.** \( H(X, Y|Z) = H(X|Z) + H(Y|X, Z) \).

*Proof.*

\[ H(X, Y|Z) = \sum_{z \in Z} p(z) H(X, Y|Z) \]

\[ = - \sum_{z \in Z} \sum_{x \in X} \sum_{y \in Y} p(x, y, z) \log p(x, y|z) \]

\[ = - \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \log \{ p(x|z) \cdot p(y|x, z) \} \]

\[ = - \sum_{x \in X} \sum_{z \in Z} p(x, z) \log p(x|z) \]

\[ - \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \log p(y|x, z) \]

\[ = H(X|Z) + H(Y|X, Z). \]

Let \( I(X; Y) \) be a mutual information. Then by the definition,

\[ I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]

\[ = \sum_{x,y} p(x, y) \log \frac{p(y)p(x|y)}{p(x)p(y)} \]

\[ = - \sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y) \]

\[ = - \sum_{x} p(x) \log p(x) - (\sum_{x,y} \log p(x, y) \log p(x|y)) \]

\[ = H(X) - H(X|Y). \]

By symmetry, \( I(X; Y) = H(Y) - H(Y|X) \). Since \( H(X, Y) = H(Y) - H(Y|X) \), \( I(X; Y) = H(X) + H(Y) - H(X, Y) \).
Finally we obtain that
\[ I(X; X) = H(X) - H(X|X) = H(X). \]

The relationship between \( H(X), H(Y), H(X, Y), H(X|Y), H(Y|X) \) and \( I(X; Y) \) is expressed in a Venn diagram.

\[ I(X; Y) = H(X) - H(X|Y), \quad I(X; Y) = H(Y) - H(Y|X), \]
\[ I(X; Y) = H(X) + H(Y) - H(X, Y), \]
\[ I(X; Y) = I(Y; X), \quad I(X; X) = H(X). \]

3. The smallest probable set

The asymptotic equipartition property (AEP) is a direct consequence of the weak law of large numbers. If \( X_1, X_2, \ldots, X_n \) are independent, identically distributed (i.i.d.) random variables and \( p(x_1, x_2, \ldots, x_n) \) is the probability of observing the sequence \( X_1, X_2, \ldots, X_n \), then the AEP states that \( \frac{1}{n} \log \frac{1}{p(x_1, x_2, \ldots, x_n)} \) is close to the entropy \( H \). Indeed, since the \( X_i \) are i.i.d. So are \( \log p(X_i) \).

Hence \( -\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \sum_i \log p(X_i) = -E \log p(X) \) in probability = \( H(X) \).

We define that the typical set \( A^{(n)}_\epsilon \) with respect to \( p(x) \) is the set of sequences \( (x_1, x_2, \ldots, x_n) \in \mathbb{X} \) with the following properties:

\[ 2^{-n(H(x)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(x)-\epsilon)}. \]

We obtain that the typical set \( A^{(n)}_\epsilon \) has the following properties ([3], [6], [9]).

**Proposition 3.1.** 1. If \( (x_1, x_2, \ldots, x_n) \in A^{(n)}_\epsilon \), then
\[ H(x) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(x) + \epsilon. \]

Also \( \Pr\{A^{(n)}_\epsilon\} > 1 - \epsilon \) for \( n \) sufficiently large.

2. \( |A^{(n)}_\epsilon| \leq 2^{n(H(x)+\epsilon)} \), where \( |A| \) denotes the number of elements in the set \( A \).
\[ |A_\varepsilon^{(n)}| \geq (1 - \varepsilon)2^{n(H(x) - \epsilon)} \text{ for } n \text{ sufficiently large.} \]

**Proof.** 1. Since \((x_1, x_2, \cdots, x_n) \in A_\varepsilon^{(n)},\)

\[ 2^{-n(H(x) + \epsilon)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(x) - \epsilon)}. \]

Taking the log with base 2 to both sides,

\[ -n(H(x) + \epsilon) \leq \log_2 p(x_1, x_2, \cdots, x_n) \leq -n(H(x) - \epsilon). \]

Therefore \(H(x) - \epsilon \leq \frac{1}{n} \log p(X_1, X_2, \cdots, X_n) \leq H(x) + \epsilon\) since the probability of the event \((X_1, X_2, \cdots, X_n) \in A_\varepsilon^{(n)}\) tends to 1 as \(n \to \infty\). For any \(\delta > 0\) there exist an \(n_0\), such that for all \(n \geq n_0,\)

\[ \Pr \left\{ \left| -\frac{1}{n} \log p(X_1, X_2, \cdots, X_n) - H(X) \right| < \epsilon \right\} > 1 - \delta. \]

2. \(1 = \sum_{x \in X^n} P(x) \geq \sum_{x \in A_\varepsilon^{(n)}} P(x) \geq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(x) + \epsilon)} \)

Finally, since \(\Pr \{ A_\varepsilon^{(n)} \} \geq 1 - \epsilon, 1 - \epsilon < \Pr \{ A_\varepsilon^{(n)} \} \leq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(x) - \epsilon)} = 2^{-n(H(x) - \epsilon)} \left| A_\varepsilon^{(n)} \right|, \)

Hence \(\left| A_\varepsilon^{(n)} \right| \geq (1 - \epsilon)2^{n(H(x) - \epsilon)}. \)

Now we divide all sequences in \(x^n\) into two sets:

One is the typical set \(A_\varepsilon^{(n)}\) and the other is complement \(A_\varepsilon^{(n)c}\) and we order all elements in each set according to lexicographic order.

Then we can represent each sequence of \(A_\varepsilon^{(n)}\) by giving the index of the sequence in the set.

Giving the index of the sequence in the set, we can represent each sequence of \(A_\varepsilon^{(n)}\). Since there are \(\leq 2^{n(H + \epsilon)}\) sequences in \(A_\varepsilon^{(n)}\), the indexing requires no more than \(n(H + \epsilon) + 1\) bits because \(n(H + \epsilon)\) may not be an integer.

We prefix all their sequences by a 0, giving a total length of \(\leq n(H + \epsilon) + 2\) bits to represent each sequence in \(A_\varepsilon^{(n)}\).

We denote \(x^n\) as a sequence \(X_1, X_2, \cdots, X_n\). Let \(I(x^n)\) be the length of the code word corresponding to \(x^n\).
Lemma 3.2. Let $X^n$ be independent identically distribution (i. i.d.) with probability $p(x)$. Let $\epsilon > 0$. Then there exists a code which maps sequences $x^n$ of length $n$ into binary strings such that the mapping is one to one and $E\left[\frac{1}{n}l(X^n)\right] < H(X) + \epsilon$, for $n$ sufficiently large.

Proof. $E(l(X^n)) = \sum_{x^n} P(x^n)l(x^n)$

\[ = \sum_{x^n \in A^n} P(x^n)l(x^n) + \sum_{x^n \in A^n} P(x^n)l(x^n) \]

\[ \leq \sum_{x^n \in A^n} P(x^n)[n(H + \epsilon) + 2] + \sum_{x^n \in A^n} P(x^n)(n\log|\omega| + 2) \]

\[ = Pr\{A^n\} \{n(H + \epsilon) + 2\} + Pr\{A^n\cap B^n_\delta\}(n\log|\omega| + 2) \]

\[ \leq n(H + \epsilon) + \epsilon n(\log|\omega|) + 2 = n(H + \epsilon) \]

where $\epsilon^1 = \epsilon + \epsilon(\log|\omega|) + \frac{2}{n}, \{B^n_\delta\} \geq 1 - \delta$.

Definition 3.3. For each $n = 1, 2, \cdots$, let $B^n_\delta \subset \omega^n$ be any set with $Pr\{B^n_\delta\} \geq 1 - \delta$ must have significant intersection with $A^n_\delta$ and therefore must have about as many elements.

Theorem 3.4. Let $X_1, X_2, \cdots, X_n$ be i.i.d. with $p(x)$. For $\delta < \frac{1}{2}$ and any $\delta^1 > 0$, if $Pr\{B^n_\delta\} > 1 - \delta$, then $\frac{1}{n} \log|B^n_\delta| > H - \delta^1$ for $n$ sufficiently large. Thus $B^n_\delta$ must have at least $2^nH$ elements, to first order in the exponent. But $A^n_\delta$ has $2^n(H + \epsilon)$ elements.

Proof. Let any two sets $A, B$ such that $Pr(A) > 1 - \delta_1$ and $Pr(B) > 1 - \delta_2$. Then this shows that $Pr(A \cap B) > 1 - \epsilon_1 - \epsilon_2$, hence $Pr(A^n_\delta \cap B^n_\delta) > 1 - \epsilon - \delta$. Indeed, since $X_1, X_2, \cdots, X_n$ are i.i.d. with $p(x)$, if we fix $\epsilon < \frac{1}{2}$, then

\[ Pr(A \cap B) = Pr(A) \cdot Pr(B) > (1 - \epsilon_1)(1 - \epsilon_2) = 1 - \epsilon_1 - \epsilon_2. \]

Accordingly, $Pr(A^n_\delta \cap B^n_\delta) = Pr(A_\delta) \cdot Pr(B_\delta) \geq (1 - \epsilon)(1 - \delta) = 1 - \epsilon - \delta$ by Proposition 3.1.(1).

Next by the chain rule of inequalities,

\[ 1 - \epsilon - \delta < Pr(A^n_\delta \cap B^n_\delta) \]

\[ = \sum_{A^n_\delta \cap B^n_\delta} P(x^n) \leq \sum_{A^n_\delta \cap B^n_\delta} 2^{-n(H - \epsilon)} \]

\[ = |A_\delta^n \cap B_\delta^n|2^{-n(H - \epsilon)} \leq |B_\delta^n|2^{-n(H - \epsilon)}, \]

\[ |B_\delta^n| \geq (1 - \epsilon - \delta)2^n(H - \epsilon). \]

Taking the logarithm with base 2 to both sides,

\[ \log_2|B_\delta^n| \geq \log (1 - \epsilon - \delta) + n(H + \epsilon), \]

\[ \frac{1}{n} \log |B_\delta^n| > \frac{1}{n} \log (1 - \epsilon - \delta) + (H - \epsilon). \]
Accordingly for \( n \) sufficiently large, we obtain

\[
\frac{1}{n} \log |B^{(n)}_\delta| > H - \delta^2.
\]

We denote the notation \( a_n \doteq b_n \) as follows.

\[
\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.
\]

Then we can now restate the theorem 3.4. as

\[
|B^{(n)}_\delta| \doteq |A^{(n)}_\epsilon| \doteq 2^{nH}.
\]

4. Monotonicity of entropy

Let the joint distribution of any subset of the sequence of random variables to be invariant with respect to shifts in the time index, i.e.

\[
Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\}
\]

for every shift 1 and for all \( x_1, x_2, \ldots \in X \). Then a stochastic process is called to be stationary.

Let random variables \( X_1, X_2, \ldots \) be a discrete stochastic process. If for \( n = 1, 2, \ldots \)

\[
Pr(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = Pr(X_{n+1} = x_{n+1}|X_n = x_n)\text{for all } x_1, x_2, \ldots, x_n, x_{n+1} \in \mathbb{X}.
\]

Then a discrete stochastic process \( X_1, X_2, \ldots \) is said to be a Markov chain or a Markov process. ([6],[7])

**Definition 4.1.** The Markov chain is said to be time invariant if the conditional probability \( P(X_{n+1}|X_n) \) does not depend on \( n \), i.e. for \( n = 1, 2, \ldots \)

\[
Pr\{X_{n+1} = p|X_n = q\} = Pr\{X_2 = p|X_1 = q\}, \text{ for all } p, q \in \mathbb{X}.
\]

Let \( \{X_i\} \) be a Markov chain. Then \( X_n \) is said the state at time \( n \). The Markov chain is called to be irreducible if it is possible to go with positive probability from any state of the Markov chain to any other state in a finite number of steps.

**Definition 4.2.** The entropy rate of a stochastic process \( \{X_i\} \) is defined as follow.

\[
H(X) = \lim_{n \to \infty} \frac{1}{n} H(x_1, x_2, \ldots, x_n), \text{ when the limit exists.}
\]

Another definition of entropy rate is, when the limit exists,

\[
H'(X) = \lim_{n \to \infty} H(x_n|x_{n-1}, x_{n-2}, \ldots, x_2, x_1).
\]
Proposition 4.3. For a stationary stochastic process, their entropy rate $H(\alpha)$ and $H'(\alpha)$ are equal, i.e.

$$H(\alpha) = H'(\alpha).$$

Proof. Since $H(X_{n+1}|X_1, X_2, \ldots, X_n) \leq H(X_{n+1}|X_n, \cdots, X_2)$

$$\leq H(X_n|X_{n-1}, \cdots, X_1),$$

$H(X_n|X_{n-1}, \cdots, X_1)$ is a decresing sequence of nonnegative numbers.

Hence it has a limit, $H'(\alpha).$ (\[6\]) \(\cdots\) \(\cdots\) \(\cdots\) \(\ast_1\) In general, we can prove that if $a_n \to a$

and $b_n = \frac{1}{n} \sum_{i=0}^{\infty} a_i$, then $b_n \to a.$ (\[6\]) \(\cdots\) \(\cdots\) \(\cdots\) \(\ast_2\)

Since by the chain rule the entropy rate is the time average of the conditional entropies,

$$\frac{H(X_1, X_2, \cdots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1),$$

Also the conditional entropies tend to a limit $H'(\alpha)$. By \(\ast_2\), their running average has a limit equal to the limit $H'(\alpha)$ of the terms. By \(\ast_1\),

$$H(\alpha) = \lim_{n \to \infty} \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim H(X_n|X_{n-1}, \cdots, X_1) = H'(\alpha).$$

For a stationary Markov chain, the entropy rate is

$$H(\alpha) = H'(\alpha) = \lim H(X_n|X_{n-1}, \cdots, X_1) = \lim H(X_n|X_{n-1}) = H(X_2|X_1).$$

Let $\mu$ be stationary distribution and $P$ be transition matrix. Then since

$$H(\alpha) = H(X_2|X_1) = \sum_{i} \mu_i \left( \sum_{j} - P_{ij} \log P_{ij} \right),$$

$$H(\alpha) = - \sum_{i,j} \mu_{ij} P_{ij} \log P_{ij}.$$

For example, consider a two-state Markov chain with probability transition matrix

$$P = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix}$$

Put $\mu_1$ and $\mu_2$ be the stationary probability of state $a$ and $b$ respectively. Then we obtain $\mu_1 p = \mu_2 \cdot 1$, since $\mu_1 + \mu_2 = 1$, the stationary distribution is $\mu_1 = \frac{1}{1+p}, \mu_2 = \frac{p}{1+p}$

Accordingly the entropy of the state $X_n$ at time $n$ is

$$H(X_n) = H \left( \frac{1}{1+p}, \frac{p}{1+p} \right).$$
Theorem 4.4. (i) Let \( T \) be a stationary stochastic process. Then
\[
H(x_0|x_{-1}, x_{-2}, \cdots, x_{-n}) = H(x_0|x_1, x_2, \cdots, x_n).
\]

(ii) Let \( X_1, X_2, \cdots, X_n \) be a stationary stochastic process. Then
\[
\frac{H(X_1, X_2, \cdots, X_n)}{n} \leq \frac{H(X_1, X_2, \cdots, X_{n-1})}{n-1}.
\]

Proof. (i) For a stationary stochastic process, the probability of any sequence of state is the same forward or backward. i.e. time-reversible.

\[
Pr(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = Pr(X_n = x_1, X_{n-1} = x_2, \cdots, X_1 = x_n).
\]

By the chain rule,
\[
H(X_n|X_{n-1}, X_{n-2}, \cdots, X_1) = H(X_0|X_{n-1}, X_{n-2}, \cdots, X_1) + \sum_{i=1}^{n-1} H(X_i|X_{i-1}, \cdots, X_1).
\]

Replacing \( n = -1 \),
\[
H(X_0|X_{1}, X_{2}, \cdots, X_1) = H(X_0|X_1, X_2, \cdots, X_1) - H(X_1|X_{-1}, \cdots, X_1) = H(X_0|X_1, X_2, \cdots, X_{n-1}, X_n).
\]

This means that the present has a conditional entropy given the past equal to the conditional entropy given the future.

(ii) \( 0 \leq p(x) \leq 1 \) implies \( H(x) = \sum p(x) \log \frac{1}{p(x)} \geq 0 \). So \( H(X_1, X_2, \cdots, X_n) \geq 0 \).

By the chain rule, \( H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n-1} H(X_i|X_{i-1}, \cdots, X_1) \).

By the proposition 4.3, the conditional probability has a limit. Since the running average \( \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1) \) has a limit equal to the limit \( H(X) \) of the terms.

\[
H(X_1, X_2, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1) = H(X_n|X_{n-1}, \cdots, X_1) \text{ as } n \to \infty.
\]

Therefore
\[
\frac{H(X_1, X_2, \cdots, X_n)}{n} = H(X_n|X_{n-1}, X_{n-2}, \cdots, X_2, X_1) \leq H(X_n|X_{n-1}, X_{n-2}, \cdots, X_2) = H(X_{n-1}|X_{n-2}, \cdots, X_1) = \frac{H(x_1, x_2, \cdots, x_{n-1})}{n-1} \text{ for large } n \text{ enough.}
\]

We wish to compute the entropy of a random tree. From this we can find that the expected number of necessary to describe the random tree \( T_n \) grows linearly with \( n \).

The following method of generating random trees yields the same probability distribution on trees with \( n \) terminal nodes. Choose an integer \( N_1 \) uniformly distributed on \( \{1, 2, \cdots, n-1\} \). Then we have the picture.
Choose an integer $N_2$ uniformly distributed over $\{1, 2, \ldots, N_1-1\}$ and independently choose another integer $N_3$ uniformly over $\{1, 2, \ldots, (n-N_1)-1\}$.

We continue the process until no further subdivision can be made. Then we can make $n$-terminal nodes tree.

Let $T_n$ denote a random $n$-node tree generated as above, then the entropy $H(T_2) = 0$, $H(T_3) = \log 2$. For $n = 4$, we have five possible trees, with probabilities $\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$.

Let $N_1(T_n)$ denote the number of terminal nodes of $T_n$ in the right half of the tree.

Theorem 4.5.

\[(n - 1)H_n = nH(n - 1) + (n - 1)\log(n - 1) - (n - 2)\log(n - 2)\]

or $\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + C_n$ for appropriately defined $C_n$.

Proof. By the definition of entropy and the construction of random tree,

\[H(T_n) = H(N_1, T_n) = H(N_1) + H(T_n|N_1) \cdots < \text{by chain rule for entropy} >\]

\[= \log (n - 1) + H(T_n|N_1) \cdots < \text{by conditional distribution} >\]

\[= \log(n - 1) + \frac{1}{n - 1} \sum_{k=1}^{n-1} [H(T_k) + H(T_{n-k})] \cdots < \text{by definition of tree} >\]

\[= \log (n - 1) + \frac{2}{n - 1} \sum_{k=1}^{n-1} H(T_k) \cdots < \text{by restriction} >\]

\[= \log (n - 1) + \frac{2}{n - 1} \sum_{k=1}^{n-1} H_k.\]

Let $H(T_n)$ be $H_n$. Then $H(T_{n-1}) = H_{n-1}, H(T_{n-2}) = H_{n-2}, \ldots, H_{n-1} = \log (n - 2) + \frac{2}{n - 2} \sum_{k=1}^{n-2} H(T_n)$. Accordingly,

\[(n - 1)H_n = (n - 1)\log(n - 1) + 2\sum_{k=1}^{n-2} H_k + 2H_{n-1}\]

\[= (n - 1)\log(n - 1) + (n - 2)\log(n - 2)\]

\[+ \frac{2}{n - 2} \sum_{k=1}^{n-2} H_k \cdot (n - 2)\log(n - 2) + 2H_{n-1}\]

\[= (n - 1)\log(n - 1) + nH_{n-1} - (n - 2)\log(n - 2) + 2H_{n-1}\]

\[= (n - 1)\log(n - 1) + nH_{n-1} - (n - 2)\log(n - 2).\]
By dividing both sides as \( n(n-1) \),

\[
\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + C_n
\]

where \( C_n = \frac{1}{n} \log (n-1) - \left(1 - \frac{2}{n}\right) \left(\frac{1}{n-1}\right) \log (n-2) \). Since \( \sum C_n = C < \infty \), you have proved that \( \frac{1}{n}H(T_n) \) converges to a constant.

**References**


Department of mathematics
Woosuk University
Wanju-gun Chonbuk 565-701, Korea