Comparison of eigenvalue ratios in artificial boundary perturbation and Jacobi preconditioning for solving Poisson Equation

Gangjoon Yoon* and Chohong Min†

September 12, 2017

Abstract

The Shortley-Weller method is a standard finite difference method for solving the Poisson equation with Dirichlet boundary condition. Unless the domain is rectangular, the method meets an inevitable problem that some of the neighboring nodes may be outside the domain. The function values at outside nodes are extrapolated by quadratic polynomial, and the extrapolation becomes unstable, that is, some of the extrapolation coefficient increases rapidly when the grid nodes are very near the boundary. A practical remedy, which we call artificial perturbation, is to treat grid nodes very near the boundary as boundary points. The aim of this paper is to reveal the adverse effects of the artificial perturbation on the condition number of the matrix and the convergence of the solution. We show that the matrix is nearly symmetric so that the ratio of its minimum and maximum eigenvalues can be referenced as the measure of its condition number. Our analysis shows that the artificial perturbation results in a small enhancement of the condition number from $O(1/(h \cdot h_{min})$ to $O(h^{-3})$ and triggers an oscillatory order of convergence. Instead, we suggest using Jacobi or ILU-type preconditioner on the matrix without applying the artificial perturbation. According to our analysis, the preconditioning not only reduces the condition number from $O(1/(h \cdot h_{min})$ to $O(h^{-2})$, but also keeps the sharp second order convergence.

1 Introduction

In this article, we consider the standard finite difference method for solving the Poisson equation $-\Delta u = f$ in a domain $\Omega \subset \mathbb{R}^n$ with Dirichlet boundary condition $u = g$ on $\partial\Omega$. Let the uniform grid of step size h is denoted by $(h\mathbb{Z})^n$. The discrete domain is then defined as the set of grid nodes inside the domain, i.e. $\Omega^h := \Omega \cap (h\mathbb{Z})^n$.

The standard finite difference method is a dimension-by-dimension application of the central finite difference, and we present mainly the case of one dimension and report any nominal differences in the other dimensions, when required. Unless Ω is rectangular, the method meets an inevitable problem that some of the neighboring nodes may be outside Ω. As depicted in Figure 1, a neighboring node of the grid node is outside Ω. The node outside Ω^h is called ghost node [10], and the function value at the ghost node is extrapolated by the quadratic polynomial as follows.

$$u^G_{i+1,j} := u^h_{i-1,j}\frac{1-\theta}{1+\theta} - 2u^h_{i,j}\frac{1-\theta}{\theta} + g_I\frac{2}{(1+\theta)\,\theta}$$

Here $\theta \cdot h$ is the distance between the grid node and the boundary to the right.

Applying the extrapolation to the second-order central difference scheme, we obtain a second-order discretization in the x-direction

$$-\left(D_{xx}u\right)_{ij} = -\frac{u_{i-1,j} - 2u_{i,j} + u^G_{i+1,j}}{h^2} = \frac{2}{\theta h \cdot h}u_{i,j} - \frac{2}{h\cdot(\theta+1)h}u_{i-1,j} - \frac{2}{\theta h\cdot(\theta+1)h}g_I. \tag{1}$$

This discretization is called the Shortley and Weller method [19] and the corresponding discrete Laplacian operator is given in (3). On applying an iterative method to solve the discrete Poisson equation which is related to an unsymmetric matrix, it is noted in [17] that if the related matrix is nearly symmetric, the residual norm is bounded by the ratio of the maximum and minimum eigenvalues in absolute value. We show in this work that the matrix induced by the Shortley-Weller method is nearly symmetric in the sense that all the eigenvalues are nearly real. In this respect, we estimate the convergence performance by the eigenvalue ratio rather than the ratio of singular values.

*National Institute for Mathematical Sciences, Daejeon 34047, Korea
†Department of Mathematics, Ewha Woman's University, Seoul 03760, Korea, corresponding author (chohong@ewha.ac.kr)

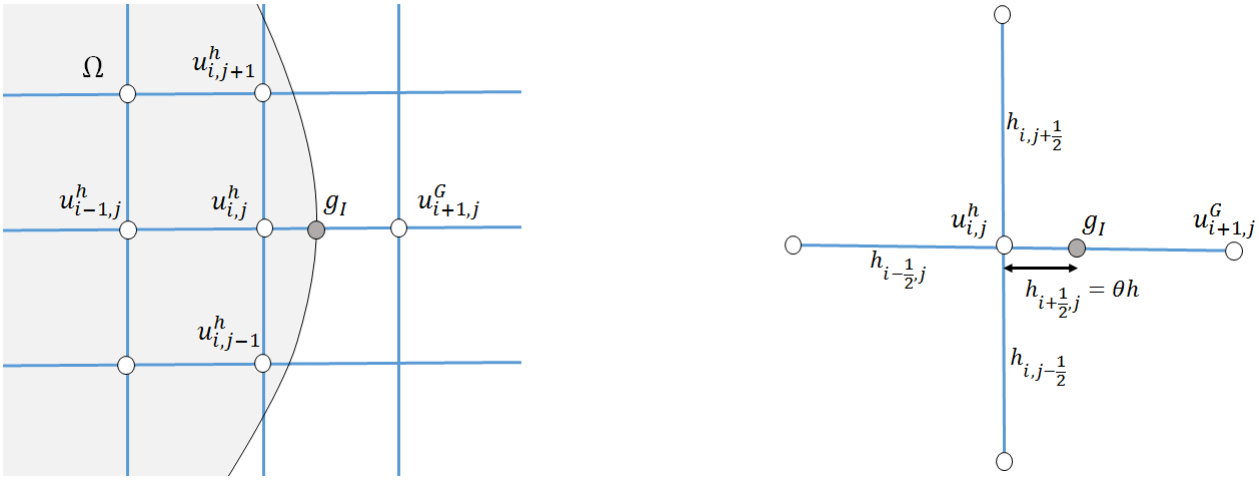

Figure 1: $u^h_{i,j}$ has four neighboring nodes. The one in the right is outside Ω, and the ghost value $u^G_{i+1,j}$ is quadratically extrapolated from inside values $u^h_{i,j}$ and $u^h_{i-1,j}$ and the boundary value g_I.

From the discretization (1), we see that the extrapolation may produce large error if θ in the denominator gets very small. This results in a large condition number for the matrix associated with the Shortley-Weller method as an estimation $|\lambda_{max}/\lambda_{min}| = O\left(1/\left(h \cdot h_{min}\right)\right)$ shown in Theorem 3.2. Here h_{min} is the minimum distance from the nodes in Ω to the boundary $\partial\Omega$. To mitigate the singularity of the extrapolation, there are two treatments in practice: artificial boundary perturbation and preconditioning.

The artificial boundary perturbation is to treat the grid nodes near the boundary within a certain threshold $\theta_0 \cdot h$ as boundary points and we take $u^h_{i,j} = g_I$ [7, 10, 16]. The common choice of the threshold is $\theta_0 = h$. We call the practice as *artificial boundary perturbation* throughout this paper.

This article is aimed at revealing the precise effects from the artificial perturbation. We review the known facts and estimate the ratio of eigenvalues for the unperturbed linear system in section 2. In section 4, we discuss the effects on the convergence of the numerical solution and the eigenvalue ratio of the linear system for the artificial perturbation. In practice, we take $\theta_0 = h$ for the perturbation value and we reveal in Theorem 4.2 that the eigenvalue ratio to the corresponding treatment is shown to be $O(h^{-3})$ so that the artificial perturbation is less effective than any preconditioning. ~~That is, we conclude from the results in section 4 that the conventional perturbation is not recommended.~~

Another treatment to mitigate the dependence on the minimum grid size and condition numbers as well is preconditioning the linear system. We estimate the effect of the Jacobi preconditioning in section 5 and we test the usual other preconditioners such as SGS, ILU and modified ILU (MILU) to see the effect of the preconditioning. We show that the Jacobi preconditioning is is enough to completely resolve the issue of the singularity of the extrapolation when θ is small, by proving that the Jacobi preconditioner is totally free from the effect of the minimum distance h_{min} and its condition number is no larger than $O(h^{-2})$. Consequently, we suggest the preconditioning method rather that the artificial boundary perturbation in order to improve the condition number.

It is worth noting that it was observed for many second order, self-adjoint, elliptic equations that the spectral condition numbers of the discrete operator grow as $O(h^{-2})$ as the mesh size h tends to zero (see [6] for details). Also, Dupont, Kendall and Rachford [9] observed that even though the convergence rates of the Jacobi, symmetric Gauss-Seidel (SGS), and incomplete LU (ILU) precondioned matrices still behave as $O(h^{-2})$ with a much smaller multiplicative constant, the modified ILU (MILU) preconditioned matrix drops the order to $O(h^{-1})$. The MILU improvement $O(h^{-1})$ for the rectangular domain was prove in [3, 4, 14, 21] and we also refer the reader to [1, 2, 5, 12, 15] for related works on the MILU preconditioning. However, the MILU improvement of $O(h^{-1})$ for general domains is not proved yet and we leave the study for a further work.

2 Review of Shortley-Weller method

In this work, we focus both on the convergence performance of the standard finite difference method for the Poisson equation and on the effect of the two methods improving the performance: artificial boundary perturbation and

preconditioning. In this respect, it is sufficient to compare the performances for the 2 dimensional case. Let us introduce some discretization settings of domain for solving the Poisson problem

$$\left\{\begin{array}{rcll} -\Delta u & = & f & \text{in} \quad \Omega \\ u & = & g & \text{on} \quad \partial\Omega, \end{array}\right. \tag{2}$$

where $\Omega \subset \mathbb{R}^2$ is an open and bounded domain with smooth boundary $\partial\Omega$. Consider a uniform grid with step size h, i.e. $h\mathbb{Z}^2$. By Ω^h we denote the set of grid nodes belonging to Ω, and $\partial\Omega^h$ denotes the set of intersection points between $\partial\Omega$ and grid lines, i.e. $\Omega^h = \Omega \cap \left(h\mathbb{Z}^2\right)$ and $\partial\Omega^h = \partial\Omega \cap \{(h\mathbb{Z} \times \mathbb{R}) \cup (\mathbb{R} \times h\mathbb{Z})\}$. A grid node $(x_i, y_i) \in \Omega^h$ has four neighboring nodes in $\Omega^h \cup \partial\Omega^h$, namely $(x_{i\pm1}, y_j)$ and $(x_i, y_{j\pm1})$ in $\Omega^h \cup \partial\Omega^h$. Let $h_{i+\frac{1}{2},j}$ denote the distance from (x_i, y_j) to its neighbor (x_{i+1}, y_j), and other distances $h_{i-\frac{1}{2},j}$, $h_{i,j\pm\frac{1}{2}}$ are defined in the same fashion, see Figure (1).

By applying the quadratic polynomial extrapolation for the ghost values, its discrete Laplacian $\Delta_h u^h : \Omega^h \to \mathbb{R}$ for $u^h : \Omega^h \cup \partial\Omega^h \to \mathbb{R}$ reads as

$$\left(\Delta_h u^h\right)_{ij} = \left(\frac{u^h_{i+1,j} - u^h_{ij}}{h_{i+\frac{1}{2},j}} - \frac{u^h_{ij} - u^h_{i-1,j}}{h_{i-\frac{1}{2},j}}\right) \frac{2}{h_{i+\frac{1}{2},j} + h_{i-\frac{1}{2},j}} + \left(\frac{u^h_{i,j+1} - u^h_{ij}}{h_{i,j+\frac{1}{2}}} - \frac{u^h_{ij} - u^h_{i,j-1}}{h_{i,j-\frac{1}{2}}}\right) \frac{2}{h_{i,j+\frac{1}{2}} + h_{i,j-\frac{1}{2}}} \quad . \tag{3}$$

Note that when $h_{i+\frac{1}{2},j} < h$, we set $x_{i+1} = x_i + h_{i+\frac{1}{2},j}$ so that $(x_{i+1}, y_j) \in \partial\Omega^h$ and $u^h_{i+1,j} = g(x_{i+1}, y_j)$ in equation (3).

Throughout the work, we denote by u the solution to the Poisson equation (2) and by u^h the solution of the discrete equation

$$\left\{\begin{array}{rcll} -\Delta_h u^h\left(x_i, y_j\right) & = & f\left(x_i, y_j\right), & \left(x_i, y_j\right) \in \Omega^h \\ u^h\left(x_i, y_j\right) & = & g\left(x_i, y_j\right), & \left(x_i, y_j\right) \in \partial\Omega^h. \end{array}\right. \tag{4}$$

Now, we introduce some lemmas on the discretization in a general setting. For the proofs we refer to [22].

Lemma 2.1 (Monotone property)**.** *For any $v^h, w^h : \Omega^h \cup \partial\Omega^h \to \mathbb{R}$ with $-\Delta^h v^h \geq -\Delta^h w^h$ in Ω^h and $v^h \geq w^h$ on $\partial\Omega^h$, we have $v^h \geq w^h$ in $\Omega^h \cup \partial\Omega^h$.*

Let us define the functions $p^h : \Omega^h \cup \partial\Omega^h \to \mathbb{R}$ as the solution of

$$-\Delta^h p^h = 1 \text{ in } \Omega^h$$

with boundary condition $p^h = 0$ on $\partial\Omega^h$. Then we have the following estimation for p^h [22].

Lemma 2.2. *Let $p(x)$ be the analytic solution of $-\Delta p = 1$ in Ω with $p = 0$ on $\partial\Omega$. Let $C_p = C_p(\Omega)$ be the constant given by*

$$C_p := \max\left\{\left|\frac{\partial p}{\partial x_i}(x)\right|, \left|\frac{\partial^3 p}{\partial x_i^3}(x)\right|, \left|\frac{\partial^4 p}{\partial x_i^4}(x)\right| : x \in \Omega \cap \partial\Omega, i = 1, 2\right\}.$$

If $h \leq \min\left(1, \frac{3}{8C_p}\right)$, then for each $(x_i, y_j) \in \Omega^h$, we have

$$0 \leq p^h_{ij} \leq 2C_p \cdot dist\left(\left(x_i, y_j\right), \partial\Omega^h\right).$$

Therefore, there is a constant $C > 0$ such that we have

$$\left(\frac{2}{h_{i+\frac{1}{2},j} \cdot h_{i-\frac{1}{2},j}} + \frac{2}{h_{i,j+\frac{1}{2}} \cdot h_{i,j-\frac{1}{2}}}\right) p^h_{ij} \leq C\frac{1}{h^2}$$

for all $(x_i, y_j) \in \Omega^h$

3 Nearly Symmetric Matrix

Since the matrix $\mathbf{A}$ associated with discretization is a sparse non-symmetric M-matrix [8], we solve the discrete equation (4) by applying the generalized minmum residual method (GMRES). In particular, the matrix $\mathbf{A} = (a_{ij})$ is nearly-symmetric in the sense that $a_{ij} \neq 0$ iff $a_{ji} \neq 0$ ([18] and references therein) and the dominant majority of

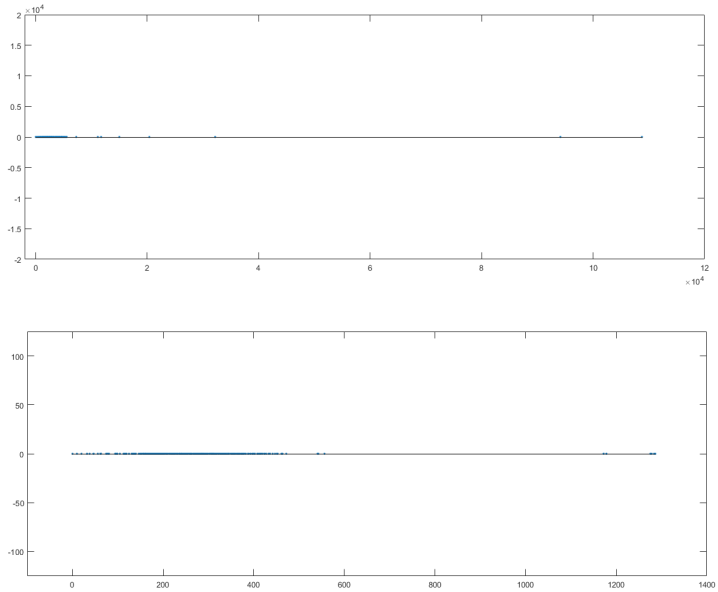

Figure 2: The eigenvalue distribution of $\mathbf{A}$ are nearly real. The top row depict the eigenvalues in complex plane for the case $\Omega = \{(x, y)|x^2 + y^2 < 1\}$ and $h = \frac{3}{80}$, and the bottom row for$\Omega = \{(x, y, z)|x^2 + y^2 + z^2 < 1\}$ and $h = \frac{3}{20}$. The ellipses including the eigenvalues have eccentricity $1 - 2.79 \times 10^{-10}$ and $1 - 1.70 \times 10^{-8}$, respectively, which mean that all the eigenvalues are almost real.

the entries in $\mathbf{A}$ are symmetric about their diagonals [23]. Now, let us look more closely at the property of $\mathbf{A}$. We decompose $\mathbf{A}$ into the symmetric and skew-symmetic parts as

$$\mathbf{A} = \mathbf{S} + \mathbf{H} \qquad \text{where } \mathbf{S} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) \text{ and } \mathbf{H} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^T).$$

We note that $\mathbf{H}$ is skew-Hermitian while $\mathbf{S}$ and $-i\mathbf{H}$ are Hermitian. The non-zero entries of $\mathbf{H}$ come only from the grid points whose distance to the boundary is less that or equal to the grid size h, and it implies that $rank(\mathbf{H})/rank(\mathbf{A}) = O(h)$ so that the rank of $\mathbf{H}$ is very low, compared to that of $\mathbf{A}$.

With the decomposition, it is not difficult to see [17, Theorem 1.35] that any egenvalue $\lambda = x + iy$ $(x, y \in \mathbb{R})$ of $\mathbf{A}$ is bounded as

$$\lambda_{min}(\mathbf{S}) \leq x \leq \lambda_{max}(\mathbf{S})$$

and

$$\lambda_{min}(-i\mathbf{H}) \leq y \leq \lambda_{max}(-i\mathbf{H}).$$

Also, it yields that the real and imginary parts of any eigenvalue is related to the symmetric matrix $\mathbf{S}$ and skew-symmetric $\mathbf{H}$, respectively. Furthermore, the experiments given in Figure 2 show that the imaginary parts of all the eigenvalues $\mathbf{A}$ is very close to the real axis. When neglecting the skew-symmetric part, $\mathbf{A}$ is diagonalizable and we have the following result in this case, which provides an upper bound on the convergence rate of the GMRES, the proof of which is given in [17].

Theorem 3.1. *Let $\mathbf{B}$ be a diagonalizable matrix, i.e., let $\mathbf{B} = \mathbf{X}\mathbf{\Gamma}\mathbf{X}^{-1}$ where $\mathbf{\Gamma}$ is the diagonal matrix of eigenvalues. Let $E(c, d, a)$ denote the ellipse in the complex plane with center c, focal distance d, and major semi axis a. Assume that all the eigenvalues of $\mathbf{B}$ are located in $E(c, d, a)$ which excludes the origin. Then, the residual norm achieved at the m-th step of GMRES satisfies the inequality,*

$$\|r_m\|_2 \leq \kappa_2(\mathbf{X}) \frac{\mathbf{C_m}\left(\frac{\mathbf{a}}{\mathbf{d}}\right)}{\left|\mathbf{C_m}\left(\frac{\mathbf{c}}{\mathbf{d}}\right)\right|} \|\mathbf{r_0}\|_\mathbf{2} \tag{5}$$

where $\kappa(\mathbf{X}) = \|\mathbf{X}\|_\mathbf{2}\|\mathbf{X^{-1}}\|_\mathbf{2}$ and C_m is the Chebyshev polynomial of degree m.

It is noted in [17] that an approximation for the coefficient $C_m\left(\frac{a}{d}\right)/C_m\left(\frac{c}{d}\right)$ is given as

$$\frac{C_m\left(\frac{a}{d}\right)}{\left|C_m\left(\frac{c}{d}\right)\right|} \approx \left(\frac{a+\sqrt{a^2-d^2}}{c+\sqrt{c^2-d^2}}\right)^m.$$

In the Shortley-Weller case, the associated matrix $\mathbf{A}$ is nearly symmetric and all the eigenvalues of $\mathbf{A}$ are almost real so that we have an ellipse $E(c,d,a)$ with $a \approx d$ (and $c \leq 2a+1$) containing all the eigenvalues. In this respect, we will check the eigenvalue ratio in this work for the estimation of performance rather than the ratio of singular values.

Now, Let us estimate the eigenvalues of the matrix associated with the discretization, which is denoted by $\mathbf{A}$

Theorem 3.2. *Let λ be an eigenvalue of the matrix $\mathbf{A}$ associated with the discretization, $0 < C_A \leq |\lambda| \leq \frac{8}{h \cdot h_{min}}$ for some constant $C_A = C_A(\Omega)$, that is independent of grid size h.*

Proof. For sufficiently small h, we may assume that whenever $(x_i, y_j) \in \Omega^h$ and one of its neighbors, let us say (x_{i-1}, y_j), is on $\partial\Omega$, then its neighbor in the opposite side belongs to Ω inside, i.e. $(x_{i+1}, y_j) \in \Omega^h$. Let λ be an eigenvalue of the matrix $\mathbf{A}$ associated with the discretization and v its corresponding eigenvector, that is, $\mathbf{A}v = \lambda v$. The Gerschgorin circle theorem implies that $0 < |\lambda| \leq \frac{8}{h \cdot h_{min}}$. Since $\mathbf{A}$ is an M-matrix, the Perron-Frobenius Theorem [13] applying to $\mathbf{A^{-1}}$ shows that the minimum eigenvalue λ_m is a positive real number and there exists an eigenvector $v = (v_P)_{P \in \Omega^h}$ with $v_P > 0$ for all $P \in \Omega^h$ corresponding to λ_m. We may assume that v is defined on $\Omega^h \cup \partial\Omega^h$ by a trivial extension. Let $v_{P_0} = \max\{v_P : P \in \Omega^h\}$. Then we can see $\mathbf{A}v = (-\Delta_h v)$ and for the function p^h given in Lemma 2.2, we have

$$(-\Delta_h v) = \lambda_m v \leq \lambda_m v_{P_0}(-\Delta_h p^h)$$

Applying Lemmas 2.1 and 2.2 shows that there exists a constant C_0 independent of h such that

$$v_P \leq C_0 \lambda_m v_{P_0} \quad \forall P \in \Omega^h.$$

This shows that $\lambda_m \geq C_A$ for some constant C_A, which completes the proof. □

Table 1 shows that the estimation of Theorem 3.2 is tight : for the unperturbed matrix $\mathbf{A}$, $|\lambda_{min}(\mathbf{A})| \approx C_A$ for some constant C_A and $|\lambda_{max}(\mathbf{A})| \approx 8/hh_{min}$.

grid	Original (unperturbed) matrix							
	$\lvert\lambda_{max}\rvert$	ratio	h_{min}	ratio	$\frac{8}{hh_{min}}$	rate	$\lvert\lambda_{min}\rvert$	ratio
20^2	4.00×10^2		5.03×10^{-2}		5.30×10^2		5.74	
40^2	2.17×10^4	54.3	1.53×10^{-3}	0.0304	3.49×10^4	65.8	5.77	
80^2	1.09×10^5	5.02	7.03×10^{-4}	0.459	1.52×10^5	4.35	5.78	1.00
160^2	1.59×10^6	14.5	9.15×10^{-5}	0.130	2.33×10^6	15.4	5.78	1.00
320^2	2.30×10^7	14.4	1.31×10^{-5}	0.143	3.26×10^7	14.0	5.78	1.00

Table 1: Eigenvalues of the unperturbed matrix of the Shortley-Weller method in example 4.1: the results tightly obey the estimate of Theorem 3.2, $0 < C_A \leq |\lambda| \leq \frac{8}{h \cdot h_{min}}$.

4 Effect of the perturbation

Let u^h be the numerical solution without perturbation, and $\tilde{u}^h$ be the numerical solution with perturbation. While it has been well known that $\left\|u^h - u\right\|_{L^\infty} = O\left(h^2\right)$ [19, 22], the convergence order of the perturbed solution has been left unclear. In this section, we investigate the convergence order of $\left\|\tilde{u}^h - u\right\|_{L^\infty}$. Let $\partial\tilde{\Omega}^h$ be the grid nodes that were treated as boundary points by the perturbation. Then the difference $\tilde{u}^h - u^h$ satisfy the discrete harmonic equation,

$$-\Delta^h\left(\tilde{u}^h - u^h\right) = 0 \text{ in } \Omega^h - \partial\tilde{\Omega}^h.$$

On $\partial\tilde{\Omega}^h$, the perturbed value $\tilde{u}_i^h$ is assigned by $g(x_I) = u(x_I)$. It is known that the unperturbed numerical solution is third order accurate near the boundary [22], so that $u_i^h = u(x_i) + O\left(h^3\right)$. Applying the mean-value theorem, we

have $\tilde{u}_i^h - u_i^h = u(x_I) - u(x_i) - O(h^3) = \frac{\partial u}{\partial x}(\xi)\,\theta_0 h + O(h^3)$, for some ξ between x_i and x_I. Thus the boundary value of the discrete harmonic solution $\tilde{u}^h - u^h$ is given as

$$\tilde{u}^h - u^h = \begin{cases} 0 & \text{on } \partial\Omega^h \backslash \{x_\Gamma \in \partial\Omega^h : dist(x_\Gamma, \Omega^h) \le \theta h\} \\ \frac{\partial u}{\partial x}(\xi)\,\theta_0 h + O(h^3) & \text{on } \partial\tilde{\Omega}^h \end{cases}.$$

The discrete maximum principle [22] states that the discrete harmonic solution should have its maximum or minimum on the boundary, therefore we have

$$\left\|\tilde{u}^h - u^h\right\|_{L^\infty(\Omega^h)} \le \|\nabla u\|_{L^\infty(\partial\Omega_\epsilon)}\,\theta_0 h + O(h^3),$$

for an $\epsilon-$neighborhood $\partial\Omega_\epsilon$ of $\partial\Omega$ and sufficiently small h with $\theta_0 h \le \epsilon$. It is well known that the unperturbed numerical solution shows a very clean second order convergence rate [8, 22], $\left\|u^h - u\right\|_{L^\infty(\Omega^h)} \simeq C \cdot h^2$. Using the above inequality and the triangle inequality, we obtain the estimate:

$$\left\|\tilde{u}^h - u\right\|_{L^\infty(\Omega^h)} \le C \cdot h^2 + \|\nabla u\|_{L^\infty(\partial\Omega_\epsilon)}\,\theta_0 h + O(h^3).$$

When $\theta_0 = h$, note that the perturbed solution is still second order accurate, however, Table 3 shows that while the unperturbed solution has the clean second order accuracy as h is varied, that of the perturbed one would fluctuate around the second order.

Let us now turn our attention to the statistics that show how often the perturbation occurs. The edge length of each grid cell is h and the boundary $\partial\Omega$ is a curve of finite length. Therefore the number of grid cells intersecting the boundary can be said to be about $O(h^{-1})$. Similarly, the number of grid edges intersecting the boundary is $O(h^{-1})$. We may assume that $O(h^{-1})$ number of intersection points are randomly distributed on grid edges of length h. The intersection points that are within distance $\theta_0 h = h^2$ from the ends are classified as 'too near' points on which the perturbation is applied. The edge length h is divided into h^{-1} number of subintervals of length h^2. When the $O(h^{-1})$ number of intersection points are randomly distributed on the edge length h, the number of points lying on the end subintervals is therefore $O(1)$.

Similarly as above, we deduct that if θ_0 were taken larger as a constant such as .001, the number of perturbation points would increase unboundedly as $O(h^{-1})$ and if θ_0 taken smaller as h^2, the perturbation points would not appear for sufficiently small h.

Example 4.1. *Consider the Poisson problem $-\Delta u = 0$ in $\Omega = \{(x,y)\,|x^2+y^2<1\}$ with $u = y/\left((x+2)^2+y^2\right)$. The artificial boundary perturbation was carried out with $\theta_0 = h$. Figure 3 shows the errors of the unperturbed and perturbed solutions.*

Our argument in this section indicates that the choice $\theta_0 = h$ is marginal so that perturbation points may sometimes appear and sometimes not. When perturbation points do appear, the convergence order $\left\|u^h - u\right\|_{L^\infty(\Omega^h)}$ would fluctuate between $C \cdot h^2$ and $C \cdot h^2 + \|\nabla u\|_{L^\infty(\partial\Omega)}\,h^2$, which are well observed in figure 3.

We have discussed how much the perturbation affects the convergence order and how often it occurs. The linear system associated with the Poisson solver has a notoriously large condition number $\simeq \frac{8}{hh_{min}}$. The third issue of our discussion is to show that the perturbation slightly decreases the condition number as below.

Theorem 4.2. *Let λ be an eigenvalue of the matrix of the Poisson solver with the perturbation, then we have*

$$0 < \tilde{C} \le |\lambda| \le \frac{8}{\theta_0 h^2} \le \frac{8}{h \cdot h_{min}}$$

for some constant $\tilde{C} = \tilde{C}(\Omega)$.

Proof. The matrix obtained in this perturbation setting is also an M-matrix. In this case, we have $h_{min} \ge \theta_0 \cdot h$ and applying the same argument used for the proof of Theorem 3.2 gives the above result. □

The above theorem shows that the eigenvalue ratio of the perturbed matrix is smaller than that of the unperturbed matrix. Since $\theta_0 = h$, the estimate indicates that the two eigenvalue ratios are bounded above by $\frac{8}{h^3}/\tilde{C} = O(h^{-3})$ and $\frac{8}{hh_{min}}/\tilde{C} \ge O(h^{-3})$, respectively. Figure 4 verifies the theorem: the perturbed ratio is smaller than the unperturbed ratio and the ratios of both data are between the second order and the third order. In the following section, however, we show that the Jacobi preconditioning drops down the ratio no larger than $O(h^{-2})$.

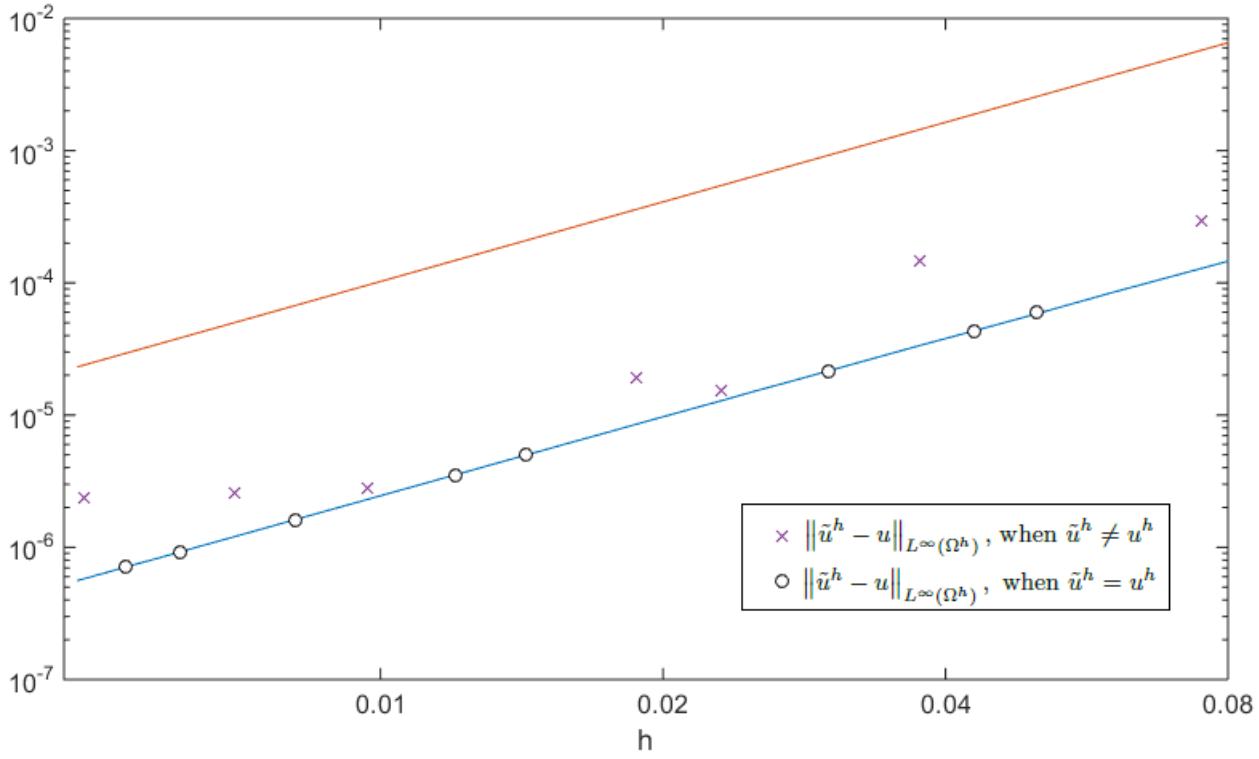

Figure 3: The error plot of the numerical solutions in Example 4.1. The perturbation occurs sometimes (marked $\times$) and sometimes not (marked $\circ$). The lower bound (Ch^2) and upper bound ($C \cdot h^2 + \|\nabla u\|_{L^\infty(\partial\Omega)} h^2$) are drawn as lines.

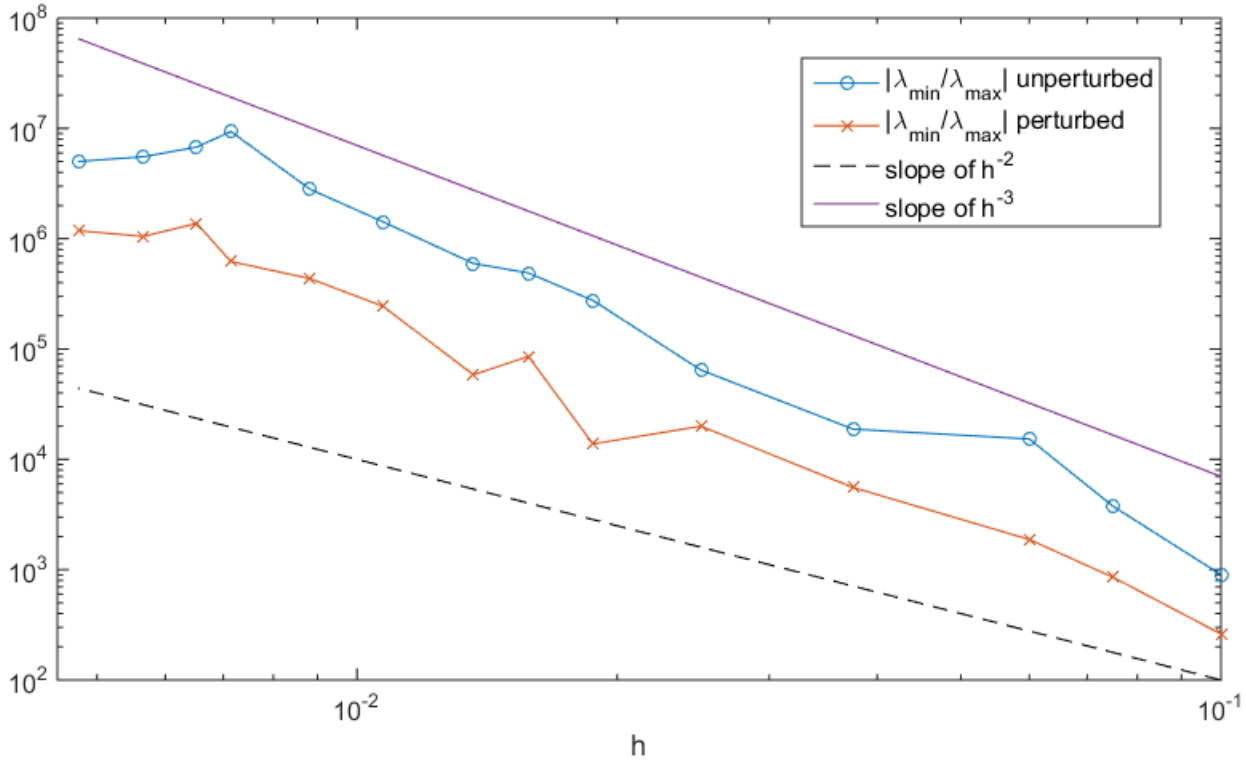

Figure 4: The eigenvalue ratio of the perturbed (marked $\times$) and unperturbed (marked $\circ$) linear systems in Example 4.1. The perturbed ratio is smaller than the unperturbed ratio as expected in Theorem 4.2. The ratios of both data are between the second order (dotted line) and the third order (solid line) .

5 Effect of the Jacobi preconditioning

For a linear system $\mathbf{A}x = b$, the Jacobi preconditioner is the diagonal matrix $\mathbf{D}$ whose diagonal entries are the same as $\mathbf{A}$. The Jacobi preconditioning on the linear system results in $\mathbf{D}^{-1}\mathbf{A}x = \mathbf{D}^{-1}b$. The preconditioning is, in other words, to scale each equation so that its diagonal entry becomes one. Applying the Jacobi preconditioning on its linear equation, the standard finite difference method now reads

$$\begin{aligned} u_{ij} &- \frac{h_{i-\frac{1}{2}j}h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}}{\left(h_{i-\frac{1}{2}j}+h_{i+\frac{1}{2}j}\right)\left(h_{i-\frac{1}{2}j}h_{i+\frac{1}{2}j}+h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}\right)}u_{i+1,j} \\ &- \frac{h_{i+\frac{1}{2}j}h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}}{\left(h_{i-\frac{1}{2}j}+h_{i+\frac{1}{2}j}\right)\left(h_{i-\frac{1}{2}j}h_{i+\frac{1}{2}j}+h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}\right)}u_{i-1,j} \\ &- \frac{h_{ij-\frac{1}{2}}h_{i-\frac{1}{2}j}h_{i+\frac{1}{2}j}}{\left(h_{ij-\frac{1}{2}}+h_{ij+\frac{1}{2}}\right)\left(h_{i-\frac{1}{2}j}h_{i+\frac{1}{2}j}+h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}\right)}u_{i,j+1} \\ &- \frac{h_{ij+\frac{1}{2}}h_{i-\frac{1}{2}j}h_{i+\frac{1}{2}j}}{\left(h_{ij-\frac{1}{2}}+h_{ij+\frac{1}{2}}\right)\left(h_{i-\frac{1}{2}j}h_{i+\frac{1}{2}j}+h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}\right)}u_{i,j-1} \quad = \quad \frac{f_{ij}}{2}\frac{h_{i-\frac{1}{2}j}h_{ij+\frac{1}{2}}h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}}{h_{i-\frac{1}{2}j}h_{i+\frac{1}{2}j}+h_{ij-\frac{1}{2}}h_{ij+\frac{1}{2}}}. \end{aligned}$$

In brief, the Jacobi preconditioner $\mathbf{D}^{-1}\mathbf{A}$ acts on u as

$$\left(\mathbf{D}^{-1}\mathbf{A}u\right)_{ij} = u_{ij} - \alpha_{ij}u_{i-1,j} - \beta_{ij}u_{i+1,j} - \gamma_{ij}u_{i,j_1} - \delta_{ij}u_{i,j+1}$$

with nonnegative constants $\alpha_{ij}, \beta_{ij}, \gamma_{ij}, \delta_{ij}$ satisfying $0 < \alpha_{ij} + \beta_{ij} + \gamma_{ij} + \delta_{ij} \leq 1$. This means that the Jacobi preconditioner eliminates completely the adverse effect of h_{min}. Also we show in the following that the Jacobi preconditioning reduces the ratio magnitude to $O\left(h^{-2}\right)$.

Theorem 5.1. *For any eigenvalue λ of the Jacobi-preconditioned matrix $\mathbf{D}^{-1}\mathbf{A}$, we have $0 < C_J \cdot h^2 \leq |\lambda| \leq 2$ for some constant $C_J = C_J(\Omega)$.*

Proof. Let λ be an eigenvalue of the Jacobi-preconditioned matrix $\mathbf{D}^{-1}\mathbf{A}$ and v its corresponding eigenvector, that is, $\mathbf{A}v = \lambda\mathbf{D}v$. Since all the diagonal entries of $\mathbf{D}$ are positive and $\mathbf{A}$ is an M-matrix, the Jacobi -preconditioned matrix $\mathbf{D}^{-1}\mathbf{A}$ is also an M-matrix. The Gerschgorin circle theorem for $\mathbf{D}^{-1}\mathbf{A}$ shows $|\lambda| \leq 2$, and it remains to show that $|\lambda| \geq C_J h^2$ for some constant C_J independent of h. We may assume that $\lambda = \lambda_{min}$ is a minimum eigenvalue. Since $\mathbf{D}^{-1}\mathbf{A}$ is an M-matrix, the Perron-Frobenius Theorem applied to $\mathbf{A}^{-1}\mathbf{D}$ shows that there exists an eigenvector $v = (v_P)_{P\in\Omega^h}$ with $v_P > 0$ for all $P \in \Omega^h$ corresponding to λ. We may assume that v is defined on $\Omega^h \cup \partial\Omega^h$ by a trivial extension as setting values of v equal to 0. Let $a_{P_0P_0}v_{P_0} = \max\{(\mathbf{D}v)_P : P \in \Omega^h\}$. Then we can see $\mathbf{A}v = (-\Delta_h v)$ and for the function p^h given in Lemma 2.2, we have

$$(-\Delta_h v) = \lambda_{min}\mathbf{D}v \leq \lambda_{min}a_{P_0P_0}v_{P_0}(-\Delta_h p^h).$$

Applying Corollary 2.1 and Lemma 2.2 give that there exists a constant C independent of h such that

$$v_{P_0} \leq \lambda_{min}a_{P_0P_0}v_{P_0}p^h(P_0) \leq \frac{C}{h^2}\lambda_{min}v_{P_0}.$$

This shows that $\lambda_{min} \geq C_J \cdot h^2$ where $C_J = \frac{1}{C}$, which completes the proof. □

Note that the eigenvalue estimate for the Jacobi-preconditioned matrix is independent of h_{min}, while that for the original matrix is dependent. Thus the presence of grid nodes too near the boundary is not problematic in the Jacobi-preconditioned matrix.

Remark

Theorem 3.6 in [20] shows the ILU-type preconditioners are actually applied on top of the application of the Jacobi preconditioner. Hence we can expect that their effects are at least as good as Jacobi; See Table 2.

grid	Jacobi preconditioned				SGS preconditioned			
	$\lvert\lambda_{max}\rvert$	ratio	$\lvert\lambda_{min}\rvert$	ratio	$\lvert\lambda_{max}\rvert$	ratio	$\lvert\lambda_{min}\rvert$	ratio
20^2	1.96		3.56×10^{-2}		1.00		1.29×10^{-1}	
40^2	1.99	0.98	8.53×10^{-3}	0.24	1.00	0.98	3.33×10^{-2}	0.26
80^2	1.99	1.00	2.08×10^{-3}	0.24	0.999	1.00	8.28×10^{-3}	0.25
160^2	1.99	1.00	5.14×10^{-4}	0.25	0.999	1.00	2.05×10^{-3}	0.25
320^2	1.99	1.00	1.27×10^{-4}	0.25	0.999	1.00	5.11×10^{-4}	0.25

grid	ILU preconditioned				MILU preconditioned			
	$\lvert\lambda_{max}\rvert$	ratio	$\lvert\lambda_{min}\rvert$	ratio	$\lvert\lambda_{max}\rvert$	ratio	$\lvert\lambda_{min}\rvert$	ratio
20^2	1.18		2.08×10^{-1}		3.28		0.999	
40^2	1.20	0.98	5.60×10^{-2}	0.26	6.64	2.02	1.00	1.00
80^2	1.20	1.00	1.40×10^{-2}	0.25	13.4	2.02	1.00	1.00
160^2	1.20	1.00	3.50×10^{-3}	0.25	27.2	2.03	0.999	1.00
320^2	1.20	1.00	8.72×10^{-4}	0.25	55.0	2.02	0.999	1.00

Table 2: Eigenvalues of the preconditioned matrices in Example 4.1: the results of Jacobi tightly obeys the estimate $O\left(h^{-2}\right) < |\lambda| < O(1)$ of Theorem 5.1, and the other results are at least as good as $|\lambda_{max}/\lambda_{min}| = O\left(h^{-2}\right)$.

6 Conclusion

The matrix derived from the standard finite difference method called as the Shortley-Weller method is sparse and non-symmetric, and nearly symmetric. Hence the ratio $|\lambda_{max}/\lambda_{min}|$, an accurate measure of condition number, is a very important factor in solving its associated linear system. We showed the estimation $|\lambda_{max}/\lambda_{min}| = O\left(1/\left(h \cdot h_{min}\right)\right)$ that is proved and verified through numerical tests. Furthermore, the tests suggest that the estimate is optimal. Therefore the presence of even a single grid node that is too near the boundary, i.e. $h_{min} \simeq 0$, severely effects the convergence accuracy of the linear system.

As an attempt to mitigate the adverse effect, a conventional approach has set a certain threshold θ_0, and treated grid nodes within distance $\theta_0 \cdot h$ from the boundary as boundary points. The perturbation of the boundary results in an erroneous modification of the Shortley-Weller method. In practice we take $\theta_0 = h$. Our analysis shows that the ratio of the related eigenvalues still suffers from the inefficient order $O(h^{-3})$. On the other hand, a simple statistics computation shows that the number of grid nodes treated as boundary points becomes zero as $\theta_0 < h$. Therefore these arguments lead us to a conclusion that the boundary perturbation is not recommended.

Instead, we considered the effect of preconditioning to relieve the large ratio $O\left(1/\left(h \cdot h_{min}\right)\right)$. Since preconditionings do not change the solution of linear system, there is no loss of convergence. We proved that the Jacobi preconditioner turns the large ratio into $O\left(h^{-2}\right)$. Note that the ratio $O\left(h^{-2}\right)$ is what we get in rectangular domains that are totally free of the grid nodes too close to the boundary [11]. With this regard, we can say that the Jacobi preconditioner completely eliminates the adverse effect from h_{min}. Numerical tests in Table 2 showed that while SGS and ILU have the same clustering ratio $O\left(h^{-2}\right)$ as Jacobi, MILU outperforms the others by reducing the ratio to $O\left(h^{-1}\right)$. The excellence was proved only in rectangular domains [12, 21], but not yet in general irregular domains, which we aim to prove in future work.

In terms of condition number, the analysis on singular values is prefered to that on eigenvalues for non-symmetric matrices. We derived the estimate of eigenvalues, and we realized that the singular value treatment is quite different from the eigenvalue. Due to our limit in time and ability, we put off the discussion of singular values to future work.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2009-0093827). G. Yoon was supported by National Institute for Mathematical Sciences(NIMS, A22200000).

References

[1] O. Axelsson, *On the eigenvalue distribution of relaxed incomplete factorization methods and the rate of convergence of conjugate gradient methods*, Technical Report, Dept. of Math., Catholic University, Nijmegen, The

Netherland, 1989.

[2] O. Axelsson and V. Eijkhout,*Robust vectorizable preconditioners for three-dimensional elliptic difference equations with anisotropy*, in Algorithm and Applications on Vector and Parallel Computers, edited by H. J. J. te Riele, Th. J. Dekker, and H. A. van der Vorst, 279–306, Amsterdam, North Holland, 1987.

[3] R. Beauwens, *Upper eigenvalue bounds for pencils of matrices*, Linear Algbra Appl. **62** (1984), 87–104.

[4] R. Beauwens, *On Axelsson's Perturbations*, Linear Algbra Appl. **68** (1985), 221–242.

[5] R. Beauwens, Y. Notay, and B. Tombuyses *S/P images of upper triangular M-matrices*, Numer. Linear Algbra Appl. **1** (1994), 19–31.

[6] M. Benzi, *Preconditioning techniques for large linear systems: A survey*, J. Comput. Phys. **182** (2002), 418–477.

[7] R. Bridson, *Fluid simulation for computer graphics*, A K Peters, Ltd., 888 Worcester street, Wellesley, MA 02482, 2008.

[8] P. Ciarlet, *Introduction to Numerical Linear Algebra and Optimization*, Cambridge Texts in Applied Mathematics, 40 West 20th street, New York, NY 10011, 1998.

[9] T. Dupont, R. Kendall, and H. Rachford, *An approximate factorization procedure for solving self-adjoint elliptic difference equations*, SIAM J. Numer. Anal. **5** (1968), 559–573.

[10] F. Gibou, R. Fedkiw, L.-T. Cheng, and M. Kang, *A second order accurate symmetric discretization of the Poisson equation on irregular domains*, J. Comput. Phys. **176** (2002), 205–227.

[11] A. Greenbaum, *Iterative methods for Solving linear systems*, SIAM, Philadelphia, PA, 1997.

[12] I. Gustafsson, *A class of first order factorization methods*, BIT **18** (1978), 142–156.

[13] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1991.

[14] Y. Notay, *Conditioning analysis of modified block incomplete factorizations*, Linear Algebra Appl. **154-156** (1991), 711–722.

[15] Y. Notay, *Conditioning of Stieltjes matrices by S/P consistently ordered approximate factorizations*, Appl. Numer. Math. **10** (1991), 381–396.

[16] Y.-T. Ng, H. Chen, C. Min, and F. Gibou, *Guidelines for Poisson solvers on irregular domains with Dirichlet boundary conditions using the ghost fluid method*, J. Sci. Comput. **41** (2009), 300–320.

[17] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2nd edition, 2003.

[18] A. A. Shah, *GRSIM: A FORTRAN subroutine for the solution of non-symmetric linear systems*, Commun. Numer. Meth. Engng **18** (2002), 803–815.

[19] G. H. Shortley and R. Weller, *Numerical solution of Laplace's equation*, J. Appl. Phys. **9** (1938), 334–348.

[20] G. Yoon and C. Min, *Analyses on the finite difference method by Gibou et al. for Poisson equation*, J. Comput. Phys. **280** (2015), 184–194.

[21] G. Yoon and C. Min, *A simple Proof of Gustafsson's Conjecture in case of Poisson equation on rectangular domains*, Amer. J. Comput. Math. *5* (2015), 75–79.

[22] G. Yoon and C. Min, *Convergence analysis of the standard central finite difference method for Poisson equation*, J. Sci. Comput. **67** (2016), 602–617.

[23] H. Zheng, S. G. Cheng, and Q. S. Liu, *A decomposition procedure for nearly-symmetric matrices with applications to some nonlinear problems*, Mech. Res. Commun. **37** (2010), 78–84.