Critical Review of Preconditioners for Helmholtz Equation and their Spectral Analysis

A. G. Shaikh¹, A. H. Sheikh¹, Asif Ali² and Salman Zeb³

¹Department of Mathematics, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah Sindh, Pakistan; agshaikh@quest.edu.pk, hanangul12@yahoo.co.uk
²Mehran University of Engineering and Technology, Jamshoro, Pakistan; asif.shaikh@faculty.muet.edu.pk
³Department of Mathematics, Malakand University, Lower Dir Khyber Pakhtunkhwa, Pakistan; salmanzeb@uom.edu.pk

Abstract

Objectives: The study reviews the iterative solvers for solving Helmholtz equation. It also surveys the preconditioners for Krylov subspace methods for discrete Helmholtz equation. Numerical experiments based conclusions are drawn, after a thorough comparison of various preconditioners. Methods/Statistical Analysis: Five point finite difference scheme for two-dimensional Helmholtz equation is employed to obtain discrete analogue of Helmholtz equation. Krylov subspace method for is considered for iterations. Matrix based preconditioner ILU with variety of fill-in and operator based preconditioner complex shifted Laplace preconditioner (CSLP) are tested. Different shifts are proposed and experimented in CSLP. Optimization of shifts is recommended on basis of results. Spectral analysis is performed to validate results. Findings: For small wave number problem, ILU type preconditioned performs better and is fairly comparable with CSLP. However, increasing wave number makes unaffordable in terms of computational cost. The pure imaginary shift \((0, 1)\) and complex shift \((1, \pi / 4)\) costs less time and number of iterations compared to other choices of shift. These choices are recommended when preconditioner is inverted directly. Approximate inversion may affect iterations, but will save solve time for large problems. Extensive spectral analysis, validating results, highlights this work. Application/Improvements: Dependency upon wave number is decreased with this recommended choice of shifts in CSLP. The number of iterations increases at very low rate when wave number is increased at rate of 2.

Keywords: Helmholtz Equation, Iterative Methods, Krylov Subspace Methods, Preconditioners, Spectral Analysis

1. Introduction

Helmholtz equation often arises in the study of physical problems involving partial differential equation (PDE) in both space and time; in this paper the time-independent case considered.

Many problems related to steady state oscillations (mechanical, acoustical, and thermal) are modeled by the Helmholtz equation. Wave scattering have many applications in physics, engineering and science. Examples include seismic imaging, radars, electromagnetism, bio medical imaging (ultrasound), road-speed sensors etc. Wave scattering phenomena is mostly modeled by Mathematicians in the form of the Helmholtz equation.

The linear system arising from a discretization of the Helmholtz equation in two dimensions (2D) or three dimensional (3D), domain, converges typically characterized by indefiniteness of the Eigen values of the corresponding coefficient matrix. With such a property, an iterative method either basic or advanced, encounters convergence problems. The method usually converges very slowly or divergences, for high frequency problems, as the linear system becomes extremely large and an indefinite. This makes the problem even harder to solve.
The Helmholtz equation can be read as
\[ -\nabla u(x, y) - k^2 u(x, y) = g(x, y) \]
Where \( \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \), and \( u(x, y) \) the unknown variable, defined on the unit square domain \( \Omega = (0,1) \times (0,1), \)
\( k \) wave number. The equation is called homogeneous if the source function \( g(x, y) = 0 \), otherwise inhomogeneous. The relation for wave number \( k \) is
\[ k(x, y) = \frac{2\pi}{\lambda} = \frac{\omega}{c(x, y)} \]
Where \( \omega = 2\pi f \) with \( f \) the wave frequency
\( \lambda = \frac{c(x, y)}{f} \) the wavelength and \( c(x, y) \) the speed of sound.

Two problems are considered here. The first one has Dirichlet’s conditions and leads to symmetric positive definite system; the second one has Neumann (Sommerfeld) conditions and leads to non-symmetric indefinite systems.

1.1 Problem Description
The Helmholtz problem considered this paper is non-homogeneous, undammed (i.e. \( \alpha = 0 \)), defined on the unit square domain \( \Omega = (0,1) \times (0,1) \) and have a constant wave number the source function is given as
\[ g(x, y) = \delta(x, y) = \delta \left( x - \frac{1}{2}, y - \frac{1}{2} \right) \]

With \( x, y \in (0,1) \) where Dirac delta function
\[ \delta(x, y) = \begin{cases} +\infty & \text{if } x = 0, y = 0 \\ 0 & \text{if } x \neq 0, y \neq 0, \end{cases} \]

This means that the waves propagate from the center of domain outwards. The Sommerfeld radiation conditions of first order are imposed that is
\[ \frac{\partial u}{\partial n} - jk u = 0 \]

2. Discretization
The problem under discussion can be solved by many methods including finite element method (FEM) and finite difference method (FDM). Higher order finite difference schemes have also been used. As, considered problem is defined on simple square domain, hence conventional FDM is chosen for discretization. Discretization is performed on a square grid, with uniform mesh size \( h = \frac{1}{N} \) in both \( x \) and \( y \) directions problems.

In FDM, the differential terms are approximated by finite differences using Taylor’s polynomials. The differential terms in Equation: (1) are approximated by finite differences as
\[ u_{xx} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + O\left(h^2\right), \]
\[ u_{xy} = \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} + O\left(h^2\right). \]

Using above differences for differential terms on the square grid, the discrete analogue of Equation: (1) will be
\[ \frac{1}{h^2} \left[-u_{i+1,j} - u_{i-1,j} - u_{i,j-1} - u_{i,j+1} + (4 - k^2 h^2)u_{i,j}\right] = g_{i,j} \]

The Sommerfeld radiation condition of first order given in Equation: (1) are descretized using by central difference at respective directional derivative, for example at right boundary, it is given as
\[ \frac{\partial u}{\partial n} - jk u = u_{r+1,j} - u_{r-1,j} - 2h jk u_{r,j} = 0, \]

The stencil for nodes/points next to boundary will take form as
\[ \frac{1}{h^2} \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 4 - k^2 h^2 + 2 jh k \end{bmatrix} u_{i,j} = g_{i,j} \quad (A) \]
\[ \frac{1}{h^2} \begin{bmatrix} -2 \\ 0 \end{bmatrix} \begin{bmatrix} 4 - k^2 h^2 + 2 jh k \end{bmatrix} u_{i,j} = g_{i,j} \quad (B) \]

The Equation: (A) represents stencil for node/point next to left side boundary, and Equation: (B) represents
stencil for node next to left and lower boundary, respectively. The resultant linear system is written as

\[ A_h u_h = g_h \]  

(9).

3. Properties of Linear System

The coefficient matrix in Equation: (9) is large, sparse, complex-valued due inclusion of Sommerfeld radiation condition and for bigger value of wave number, it is indefinite, as discussed in spectral analysis of the system. Furthermore the matrix is non-symmetric and non-Hermitian.

4. Iterative Methods for the Helmholtz Equation

The coefficient matrix \( A \) in Equation: (9) is mostly indefinite. By the “indefinite”, we mean that the real part of the eigenvalues of \( A \) lie in both positive and negative half plane in the complex plane. The dimension of linear system increases when wavenumber increases, therefore more nodes are required to resolve higher wavenumber. The direct solvers are of impractical for many obvious reasons. Iterative methods are mostly preferred for large and particularly sparse linear systems. Few are discussed.

4.1 Basic Iterative Methods

The basic iterative methods are also called fixed-point iterations which are based on the splitting \( A = F - G, \, G \in C^{N \times N} \). After substitution into Equation: (9) and few steps of simplification will give iterative process for \((j+1)^{th}\) approximation

\[ u^{j+1} = u^j + F^{-1} r^j \]  

(10)

To improve the convergence of the fixed point iterations, a relaxation factor \( \omega \) is often introduced in Equation: (10).

A standard approach to investigate the convergence of this type of methods is by analyzing the amplitude of the Fourier nodes between two successive iterations. For indefinite Helmholtz problems, it can be shown that there exist no \( \omega \), such that the error is reduced during the iterations. This means that Jacobi iteration always divergences if applied to the indefinite Helmholtz equation. Similarly result holds for the Gauss-Seidel iteration method.

4.2 Krylov Subspace Methods

For large, sparse matrix the Krylov subspace are very popular. The methods are developed on construction of iterants in the subspace.

The space \( K^j(\{r^0, Ar^0, Ar^1, ..., A^{j-1}r^0\}) \) is called the Krylov subspace of dimension \( j \), associated with \( A \) and \( r^0 \), and initial residual \( r^0 = g - Au^0 \) is related to the initial guess \( u^0 \).

Among methods which are based on construction the Krylov subspaces, a Conjugate Gradient (CG)11, COCG, GMRES, CGS, Bi-CGSTAB, and QMR are popular choices. For non-symmetric linear systems, Krylov subspace can be built from Arnoldi’s process, which leads to GMRES. GMRES is optimal method; it reduces the 2-norm of the residual at the every iteration. GMRES, however, require long recurrences, which is usually limited by the available memory. A remedy is by restarting, which some-times lead to slow convergence or stagnation.

Since CG is standard formula for Krylov method, however CG is limited in application to only symmetric and definite problems. On the contrast, GMRES has the property it can be used for indefinite and non-symmetric problems as well. It is necessary to use preconditioner to get better results of Krylov subspace.

4.3 Eigen Decomposition

We can decompose a matrix into parts that give us some information about the matrix, which is not directly accessible from the one 2D array representation.

One such decomposition is Eigen-Decomposition, where we decompose a matrix, into a set of eigenvectors and Eigen values.

An eigenvector of \( \nu \) of square matrix \( A \nu = \lambda \nu \). A matrix with all positive eigenvalues being positive is called positive definite. A matrix whose eigenvalues are all positive or zero is called positive semi-definite. If all eigenvalues are negative, matrix is called negative definite, and semi definite is similarly defined.

5. Numerical Experiments

For all the experiments, \( u^0 \) (zero vector) is used as initial guess, also from here and onwards, the mesh size \( h \) is
such that for a wave number $k$, it satisfies $Kh = 0.625$ (equivalent to 10 grid points per wavelength, and $k^2 = K$) in all numerical experiments, unless mentioned.

The first result is presented in, the Table (1) shows the performance of the CGS and GMRES algorithm with respect to a variation of $K$. Increasing the wave number $K$ severely affected the number of iterations and the actual residual.

5.1 Preconditioning

In an iterative method preconditioning is often vital component in enabling the solution of such system when the dimension is large. There are many iterative techniques for solving linear system. For large spare linear system, the convergence rate is always a concern for researchers; one can improve significantly the convergent rate by applying appropriate preconditioner. The convergence of iterative methods depends on eigenvalues of the coefficient matrix, it is often advantageous to use preconditioner that transfer the system to one with a better distribution of eigenvalues. The preconditioner is the key to successful iterative solver. In simple words, the preconditioning is technique of transforming a linear system into favorable one for any iterative solver, by multiplying the preconditioner $M$ (on left, right and split) preserving the solution. After choosing a preconditioner $M$ for any system Equation: (9), who's invertible is cheap to compute, and then the transformed system is $M^{-1}Au = M^{-1}g$.

This is called left preconditioning, which is used in this paper.

5.1.1 ILU Preconditioner

One simple class of preconditioners is obtained by an Incomplete LU factorization of $A$[14,17], and there are many ways to obtain approximate ILU factorization of $A$, e.g fill-in ILU and ILU with some tolerance. ILU obtained by restricting the structure of $L$ and $U$ to equal that of $A$ leads to preconditioner, known as ILU(0) which is easy to compute but not effective for Krylov subspace methods. A more accurate ILU factorization of $A$ is obtained by allowing more fill-in the ILU-factorization. Dropping the elements less than some given value in ILU gives rise to ILU (tolerance), known as ILU(0). Some experiments are done in this paper using GMRES with preconditioners ILU(0) and ILU(tolerance) with tolerance (0.01). Table 2 and 3 shows the numbers of iterations with actual residual as it seems to be decreased with decreasing mesh size and also increasing wave number $K$. Results in Table 2 and 3 indicate that ILU (0.01) works better than ILU (0) but at the cost of big storage, that is due to more fill-in is allowed, which is indicated by the number of non-zero elements in factorizations $L$ and $U$.

![Figure 1. Grid points v/s Number of Iterations with GMRES ILU (0) and ILU (0.01).](image)

![Figure 2. (a) Spectrum of $M^{-1}A$ for $K = 10$ (b) Spectrum of $M^{-1}A$ for $K = 50$.](image)

Gradually and hence ILU (0.01) is expensive to apply. Also in both versions of ILU, storage problems can occur, where the problem with ILU (0.01) can be severe. Actual residual is decreasing with respective increasing wave number (See Table 2 and 3).
Table 1. An observation: Number of iterations by CGS and GMRES with different values of wave number $K$ along with Actual Residual

<table>
<thead>
<tr>
<th>N</th>
<th>K</th>
<th>Dim of A</th>
<th>Iterations CGS</th>
<th>Iterations GMRES</th>
<th>Norm of Actual Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>10</td>
<td>225</td>
<td>38</td>
<td>48</td>
<td>0.082</td>
</tr>
<tr>
<td>32</td>
<td>20</td>
<td>961</td>
<td>95</td>
<td>112</td>
<td>3.977</td>
</tr>
<tr>
<td>48</td>
<td>30</td>
<td>2209</td>
<td>128</td>
<td>157</td>
<td>0.231</td>
</tr>
<tr>
<td>64</td>
<td>40</td>
<td>3969</td>
<td>188</td>
<td>216</td>
<td>0.416</td>
</tr>
<tr>
<td>80</td>
<td>50</td>
<td>6241</td>
<td>295</td>
<td>296</td>
<td>6.521</td>
</tr>
</tbody>
</table>

Table 2. Number of iterations by GMRES with preconditioner ILU(0) with different values of wave length $k$

<table>
<thead>
<tr>
<th>K</th>
<th>Dim of A</th>
<th>Iterations nz(A)</th>
<th>Iterations nz(L and U)</th>
<th>Actual Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>225</td>
<td>21</td>
<td>1065</td>
<td>4.2665 $10^{-7}$</td>
</tr>
<tr>
<td>20</td>
<td>961</td>
<td>70</td>
<td>4681</td>
<td>2.6946 $10^{-7}$</td>
</tr>
<tr>
<td>30</td>
<td>2209</td>
<td>104</td>
<td>10857</td>
<td>1.9895 $10^{-6}$</td>
</tr>
<tr>
<td>40</td>
<td>3969</td>
<td>191</td>
<td>19593</td>
<td>2.5764 $10^{-6}$</td>
</tr>
<tr>
<td>50</td>
<td>6241</td>
<td>295</td>
<td>30889</td>
<td>3.1148 $10^{-6}$</td>
</tr>
</tbody>
</table>

The increase in number of iterations can be seen, where number of iterations is plotted against number of grid points (size of problem). The convergence history for GMRES with ILU (0) and ILU (0.01) as a preconditioner is given in Figure 1, and compared with that of GMRES without preconditioners. Results with shifted Laplace preconditioner are presented in subsequent section. Further, Figure 2 clearly indicates that the number of iterations of GMRES with preconditioner ILU(0.01) are less than both GMRES as well as GMRES preconditioner ILU(0).

5.1.2 Shifted Laplace Preconditioner

Another approach is obtaining preconditioner based on operator, not on matrix. This preconditioner, in context of Helmholtz equation, is called shifted Laplace preconditioner\[15\]. This is developed by discretization of the operator

$$M(\beta_1, \beta_2) := -\Delta - (\beta_1 - j\beta_2)k^2, \beta_1, \beta_2 \in \mathbb{R}$$

Where $\beta_1$ and $\beta_2$ are real and imaginary shifts respectively. This class starts with a simple Laplace operator $M = \Delta$ which was used as a preconditioner, it seems to be good, and works well, until mesh size is small, for large size of mesh this shift creates lot of problems, lot of unwanted scattered eigenvalues appear, as shown in Figure 3, which shows that for large mesh size this shift is not good choice. Later an additional real term shift was added in the Laplace operator, making this preconditioner resembling more the Helmholtz operator but with an opposite sign as investigated. Later Laplace operator with imaginary shift was introduced and found to be effective for the Helmholtz equation, as shown in Figure 3.

Figure 3. (a) Spectrum of $M(1,1)$ with different shifts (b) Spectrum of $M(0,0)$, and $M(1,1)$

The optimal choice of real and imaginary shifts with restriction to be SPD (with positive real parts of eigenvalues of preconditioner) is investigated in detail in\[9,10\] in the context of the condition number for CG, first for real shift and then generalized to complex shifts. $\beta = -1$, for (only) real shifted Laplace preconditioner $\beta_1 = 0$, and $\beta_2 = 0$ for shifted Laplace preconditioner are concluded as an optimal choice (for details, see\[9,10\]).

Table 4 shows the number of GMRES iterations using four different shifted Laplace preconditioners. An observation is that for very small wave number $K$, $M(0,0)$ do work well, but for large wavenumber, this is no more effective as showed in the spectrum of Figure 4 the spectrum showed the scattered behavior of values, this
behavior may allow the matrix to be singular, or negative eigenvalues, which leads to negative definite matrix. For large wave number $M(0,1)$ is found satisfactory somehow, and it indicates as the number of wave number increases the spectrum becomes clustered, which is good sign for convergence. Same is depicted in Figure 4. The spectrum of $M(0,1)$ is plotted and compared with spectrum of $M(0,0)$ in Figure 3 for some fixed $K$. The Figure 3 shows spectrum with $M(1,1)$ for various choices of $K$, one of them is fairly comparable with spectrum given in Figure 3. One can notice that the spectrum is not only nicely clustered, all the eigenvalues are also encompassed by circle of radius 1, which is more effective as compared other shifts discussed in this paper. Tabulated results of iterations confirm this finding. The spectrum with $M(-1,-1)$ is similar to that of $M(1,1)$, only turns counter clockwise as showed in Figure 5. The comparison of different shifts in CSLP is given in Figure 6, which also shows that $M(1,1)$ has better spectrum. It is worth mentioning that ILU(0.01) perform comparable to CSLP for small wave numbers at the cost of huge memory, which is not affordable in case of large wave number.

Figure 7 shows the spectrum for fixed wave number $k = 10$ with two different type mesh sizes. The differences are small. Figure 7 shows the spectrum with two different wave numbers $K = 10$ and $K = 40$ with mesh size satisfying $Kh = 0.625$, one observes that the Eigen values are rushing to zero with increase wave number. This is the drawback or hindrance of $M(0,1)$ noted earlier. Some eigenvalues near to zero affected the convergence of GMRES badly, which can be dealt.

The Figure 7 shows increase in number of iterations with respect to size of problem for different shifts and the convergence history of the GMRES with Laplace precon-

<table>
<thead>
<tr>
<th>$K$</th>
<th>Dim of A</th>
<th>$M_h(0,0)$ Iterations</th>
<th>$M_h(-1,0)$ Iterations</th>
<th>$M_h(0,1)$ Iterations</th>
<th>$M_h(1,1)$ Iterations</th>
<th>$M_h(1,\pi/4)$ Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>289</td>
<td>12</td>
<td>15</td>
<td>13</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>20</td>
<td>1089</td>
<td>26</td>
<td>28</td>
<td>25</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>30</td>
<td>2401</td>
<td>52</td>
<td>52</td>
<td>41</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>40</td>
<td>4225</td>
<td>86</td>
<td>86</td>
<td>63</td>
<td>54</td>
<td>47</td>
</tr>
<tr>
<td>50</td>
<td>6561</td>
<td>136</td>
<td>129</td>
<td>90</td>
<td>70</td>
<td>58</td>
</tr>
</tbody>
</table>
conditioner $M(0,0)$, $M(-1,0)$, $M(0,1)$, and $M(1,1)$ is presented in Figure 8.

GMRES with shifted Laplace preconditioner $M(0,1)$ takes more time because a direct solver is used for solution of preconditioner systems.

**Figure 4.** (a) Spectrum of $M(0,0)$ (b) Spectrum of $M(0,1)$

**Figure 5.** Spectrum of $M(0,1)$, $M(-1,0)$, $M(1,1)$ and $M(-1,1)$.

**Figure 6.** Spectrum of $M(0,1)$, $M(-1,0)$, $M(1,1)$ and $M(1,1)$.

**Figure 7.** Number of iterations on different shifts $M_h(0,0)$, $M_h(-1,0)$, $M_h(0,1)$ and $M_h(1,1)$.

**Figure 8.** Comparison of the convergence for GMRES with shifted Laplace preconditioners $M_h(\beta_1,\beta_2)$ with different shifts for $Kh = 0.625$. 

6. Conclusion

In this study, Krylov subspace iterative methods has been applied to the Helmholtz problem. Need of preconditioner is highlighted and a survey of preconditioner is presented. ILU type preconditioners and CSLP are tried. The CSLP preconditioner is found to be very effective to enhance the convergence of Krylov subspace methods considered problem. CSLP with real and imaginary shifts as \((1,1)\) works better for all ranges of wavenumber. Note that preconditioner is inverted exactly. A spectral analysis confirms findings for said shift.

7. References