International Journal of Numerical Methods and Applications

Published Online: December 2013
Available online at http://pphmj.com/journals/ijnma.htm
Volume 10, Number 2, 2013, Pages 99-126

REAL VALUED BLOCK SOR ITERATIVE METHODS FOR THE HELMHOLTZ EQUATION

Manolis Vavalis

Department of Electrical and Computer Engineering
University of Thessaly
Volos, Greece

Abstract

The numerical solution of the Helmholtz model with complex coefficients is considered. This PDE problem is split into real and imaginary parts and so transformed into a coupled system of two Helmholtz problems with real coefficients. Associated iterative models at both continuous and discrete levels are proposed. The convergence of these, Block SOR type, iterative methods for solving the linear algebraic system of equations associated with the 5-point-star Helmholtz discretization is analyzed. Numerical experiments which confirm the theoretical results and exhibit the effectiveness of the proposed schemes are presented.

1. Introduction

The Helmholtz equation has been proved a very powerful mathematical tool for modeling acoustic wave propagation in the sea [4] as well as for several other physical phenomena of completely different nature (e.g., time-harmonic magnetic simulations [11]). In this paper, we concentrate on the Helmholtz equation as a model of sound propagation that arises in numerous physical applications and provides the foundations for several practical

Received: February 5, 2013; Accepted: April 25, 2013
2010 Mathematics Subject Classification: 65N22, 65F10, 35J05.
Keywords and phrases: basic iterative methods, Helmholtz equation, finite difference, underwater acoustics.

applications [12, 21, 26, 25]. For such applications, the accuracy of the numerically computed solutions is of obvious importance but also the computational efficiency (both CPU time and memory) is of equal importance. Therefore, the search for efficient and accurate Helmholtz solvers is well justified.

In this study, we focus on the complex Helmholtz equation

$$\Delta u + k^2 u = f \text{ in } \Omega, \tag{1}$$

where Δ is the Laplace operator, and k is the wave number in the complex plane. The general solution of Helmholtz equation (1), in the case where $\Omega \in \mathbb{R}^2$ is a circular disk, is given by

$$u(x, y) = \int_0^{2\pi} D(\theta) e^{ik(x\cos\theta + y\sin\theta)} d\theta \tag{2}$$

and represents a combination of plane waves propagating in different directions $\theta \in [0, 2\pi]$. $D(\theta)$ is a distribution function of the direction of propagation. Helmholtz equation (1) is coupled with boundary conditions which constitute an approximation to the Sommerfeld radiation condition

$$\lim_{r\to\infty} r\left(\frac{\partial u}{\partial \nu} + \imath\alpha u\right) = 0 \tag{3}$$

used in models of acoustic scattering [4].

Discretizations of the above Helmholtz problem (1), (3) using a finite element, a finite difference or a spectral method result in a linear system of algebraic equations whose matrix A is typically complex, non-Hermitian, indefinite, very large and sparse. Direct methods based on Gauss elimination with partial pivoting require, in particular in the 3-dimensional case, a prohibited amount of additional storage and thus have limited use. Multilevel methods commonly suffer from the requirement that the coarse spaces used should be fine enough to accurate represent the solution [6, 9]. In addition, the complex coefficient matrix A typically has eigenvalues with both positive and negative real parts [6]. This fact usually causes difficulties for iterative methods. It should also be pointed out here that fast direct methods (for

example FFT), regularly used for the Helmholtz equation with Dirichlet and Neumann boundary conditions, cannot be applied to the above problem due to the fact that the boundary conditions are of Robin type.

In practice, the above mentioned complex linear system is usually solved by either a Gauss elimination type direct method or by Krylov subspace methods [24] like quasi-minimal residual (QMR) [7] and bi-conjugate gradient methods. The latter ones seem to be very popular in recent studies although there are still several open questions to be answered. The main associated research issue is how to accelerate the convergence and the main drawback of many Krylov-based methods (in particular those developed in the 90's) is the fact that they solve the complex system by transforming it into an equivalent real one of double dimensions. This has to be avoided (as it the case of recently developed methods [10, 25, 18, 17]) since such real systems have spectra that are less favorable for the convergence of Krylov-based methods.

Besides numerous domain decomposition based iterative algorithms (see [5, 20, 26] and references therein) for solving the Helmholtz problem, conventional iterative methods have already been used and analyzed. Specifically, Bayliss et al. [3] considered conjugate gradient type methods and Douglas, Jr. et al. [16] analyzed an alternating-direction iterative method. Freund studied Quasi Minimum Residual (QMR) methods [7, 8]. It is worth to point out that [26] includes an accurate and up-to-day review on iterative methods and their preconditioners for the solution of the Helmholtz equation.

The preconditioning of the iterative methods has also received a lot of attention in the past decade (see [6] and references therein) which have increased significantly in the past few years [23, 18, 22, 10, 17, 25]. These efforts led to the development of effective preconditioners in 2- and 3-dimensions derived in many different ways, e.g., (1) by replacing the Sommerfeld-like boundary conditions on the two opposite edges of the rectangular domain by Dirichlet or Neumann conditions and exploit fast direct solvers, (2) by using specialized incomplete LU factorizations, (3) by using analytic factorizations of the Helmholtz operator and (4) by multigrid based schemes.

From the recent efforts for the iterative numerical solution of the Helmholtz problem, we should mention the reduction scheme proposed by Axelsson and Kucherov [1] that leads to an effective preconditioning for Helmholtz models similar to the ones considered in the present study. Numerical results show that their scheme competes well with other known iterative schemes like QMR.

In this paper, an effort to analyze basic iterative methods for finite difference Helmholtz complex algebraic linear systems is made. These methods are formulated in a block Successive Over-Relaxation (SOR) manner and only real arithmetic is performed. Discussions concerning such an approach have appeared in the literature (see for example [26, 15]) but to the best of our knowledge neither implementation nor analysis has been carried out for the Helmholtz problem or any similar one.

It is important to make clear that the objective of present study is not to provide a more efficient way to iteratively solve finite difference Helmholtz complex algebraic linear systems but to provide a new generic iterative approach that will add value to the plethora of other direct and iterative schemes and properly utilize preconditioners (shifted Laplace preconditioners, multilevel Krylov methods, algebraic multigrid, sweeping preconditioner) commonly used in solving similar systems.

The rest of this paper is organized as follows: In the next section, we briefly describe the physical problem of wave propagation, its mathematical models in general and the Helmholtz wave equation problem in particular. Furthermore, we formulate a PDE scheme that iterates between the real and imaginary parts of the original complex PDE model. In Section 3, we present a simple finite difference scheme to discretize the Helmholtz problem considered, and derive the resulting linear system of complex algebraic equations. We then formulate the block iterative schemes for solving these equations and carry out the convergence analysis of one of the proposed such schemes. The numerical verification of the convergence analysis is presented in Section 4 together with experimental data that exhibit the convergence characteristics of the proposed methods. Section 5 contains our concluding remarks.

2. Helmholtz Models and the Algebraic Linear System

2.1. The basic Helmholtz PDE model

The acoustic problem defined in an open domain Ω with regular boundary $\partial\Omega \in \mathbb{R}^d$, which we assume it is composed by the different parts Γ_D, Γ_N, Γ_{NR}, can be written down as:

$$\begin{cases} \alpha \dfrac{\partial^2 u}{\partial t^2} + \beta \dfrac{\partial u}{\partial t} - \nabla \cdot (\varepsilon \nabla u) = -f & \text{in } \Omega, \\ u = \chi & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial \nu} = \psi & \text{on } \Gamma_N, \\ + \text{ non-reflecting conditions} & \text{on } \Gamma_{NR}, \end{cases} \tag{4}$$

where ν denotes the outward unit normal to the boundary, α is related to the propagation velocity of acoustic waves and β takes into account dissipative terms. The typical values for the space dimension d are 2 and 3 but for the rest of this paper, we will assume $d = 2$. Our results seem to be easily extended to cover the case where $d = 3$ but such an extension is beyond the scope of this paper. The aim of non-reflecting conditions is to simulate wave propagation in unbounded domains: they are a suitable combination of time and space derivatives. Suitable initial conditions on u and its time derivative should be provided as well.

In this paper, we consider the following linear Helmholtz PDE model with suitable absorbing boundary conditions of Robin type on the artificial boundary

$$-\Delta u - a^2 u + \imath q^2 u = f \text{ in } \Omega, \tag{5}$$

$$\frac{\partial u}{\partial \nu} + \imath \alpha u = 0 \text{ on } \Gamma \equiv \partial\Omega, \tag{6}$$

where we assume that the coefficients q and α are bounded real constants and that a is a real valued, bounded and sufficiently regular function so both the existence and uniqueness of the solution of the PDE problem lying in $H^1(\Omega)$ for reasonable f are assured. The coefficient α is assumed to be

positive and such that (6) represents a first-order absorbing boundary condition. The analytic solution of (5) is

$$u(x,\ y) = \frac{\phi(x) * \phi(y)}{w^2},\ \text{where}\ \phi(z) = e^{\imath w(z-1)} + e^{-\imath wz} - 2 \qquad (7)$$

and where $w > 0$ is the singular frequency.

It is our belief that the basic formulation of the methods and most parts of the analysis that will follow can be easily extended to more general PDE models of complex Helmholtz type. In fact, our schemes seem to be appropriate for more advanced radiation conditions than the simple first-order ones considered in this paper. Such more accurate absorbing boundary conditions are available and include high-order local and exact nonlocal, infinite elements, and perfectly matched layers (PML).

2.2. An iterative Helmholtz PDE model

Let us now split the analytic solution u of Helmholtz problem (5), (6) into real u_R and imaginary u_I parts, split similarly the right hand side f, substitute $u \equiv u_R + \imath u_I$ and $f \equiv f_R + \imath f_I$ in (5) and (6) and equate the real and imaginary parts of both equations to end up with the following coupled system of PDE problems defined in $\mathbb{R}^2$:

$$-\Delta u_R - a^2 u_R = f_R + q^2 u_I \ \text{in}\ \Omega,$$

$$\frac{\partial u_R}{\partial \nu} = \alpha u_I \ \text{on}\ \Gamma, \qquad (8)$$

$$-\Delta u_I - a^2 u_I = f_I - q^2 u_R \ \text{in}\ \Omega,$$

$$\frac{\partial u_I}{\partial \nu} = -\alpha u_R \ \text{on}\ \Gamma. \qquad (9)$$

If the functions u_I and u_R in the right hand sides of (8) and (9), respectively, were somehow known, then we could very easily solve (for example using Fast Fourier Transforms) the above two real PDE problems. This observation naturally leads us to the construction of a scheme which for $n = 1, 2, \ldots$ iterates as follows:

$$-\Delta u_R^{(n+1)} - a^2 u_R^{(n+1)} = f_R + q^2 u_I^{(n)} \text{ in } \Omega,$$

$$\frac{\partial u_R^{(n+1)}}{\partial \nu} = \alpha u_I^{(n)} \text{ on } \Gamma, \tag{10}$$

$$-\Delta u_I^{(n+1)} - a^2 u_I^{(n+1)} = f_I - q^2 u_R^{(n+1)} \text{ in } \Omega,$$

$$\frac{\partial u_I^{(n+1)}}{\partial \nu} = -\alpha u_R^{(n+1)} \text{ on } \Gamma. \tag{11}$$

Given an initial value for the first iterates $u_R^{(0)}$ (and/or $u_I^{(0)}$) the above scheme defined by the two coupled PDE problems produces sequences of successive iterates $u_R^{(n)}$ and $u_I^{(n)}$, $n = 0, 1, 2, ...$ which we hope that converge, as n tends to infinity, to the analytic solution functions u_R and u_I, respectively. Here it is worth making the following remarks concerning the above iteration scheme:

- The PDE problems involved are in $\mathbb{R}^2$ and therefore there exist a plethora of mature software tools of high quality to solve them.
- The real PDE problems involve Helmholtz equation with only Neumann boundary conditions. So, they can be solved with fast solvers of optimal computational complexity.
- The iteration scheme given above constitutes a basis to more effective schemes. For example, one may easily accelerate its convergence by means of well known basic techniques, like the ones associated with the Successive Over-Relaxation (SOR) method, the Accelerated Over-Relaxation (AOR) method, etc.

It has been already experimentally observed that the above scheme does exhibit fast converge for certain configurations. The theoretical analysis and the further experimental analysis of the method described above is underway and will be presented elsewhere.

For the rest of this paper, we will consider a discrete analogue of the above iterations which are at continuous (PDE) level. Specifically, we will propose, analyze and implement iterative methods, motivated by the above

one, at the linear algebra level. We should point out that we essentially have here two different, in their philosophy and construction, iterative approaches. They differ in the fact that for the above described continuous scheme we first formulate the iterations and we then discretize each PDE problem to obtain a numerical solution by iterations. In the approach we will consider next we first discretize the original complex PDE problem (5)-(6) and then derive the iteration scheme.

2.3. The Helmholtz finite difference linear system

A numerical approximation of the analytic solution of the problem (5)-(6) can be obtained using finite element, finite difference or spectral discretization methods. In this paper, we consider the finite difference case but it is our belief that both the derivation and analysis of the proposed iterative methods for finite element or spectral methods on uniform space discretizations are, probably, easier. For details on the recent finite difference discretization schemes for the Helmholtz equation the reader is referred to [2]. Below we generate the linear algebraic system associated with a particular simple case. Other discretization schemes and configurations may be treated similarly.

For simplicity (and without loss of generality), we set $\Omega \equiv (0, 1)^2$ and consider the uniform partition of $\Omega \bigcup \partial\Omega$ of step size (same in both directions) $h = \frac{1}{N+1}$, where N is the number of the interior points in each direction with corresponding nodes $(x_i, y_j) \equiv (ih, jh), \quad i, j = 0(1)N+1.$ We discretize the equations in (5) on the nodes of the grid with second order central differences. Using $u_{ij} \equiv u(x_i, y_j), \quad a_{ij} \equiv a(x_i, y_j), \quad g_{ij} \equiv h^2(\imath q^2 - a_{ij}^2) + 4, \quad f_{ij} \equiv f(x_i, y_j)$ and $b_{ij} \equiv h^2 f_{ij},$ we easily obtain from (5) the following $(N+1)^2$ "interior" algebraic equations:

$$-u_{i-1,\,j} - u_{i+1,\,j} - u_{i,\,j-1} - u_{i,\,j+1} + g_{i,\,j}u_{i,\,j} = b_{ij},\ i,\ j = 0(1)N+1, \quad (12)$$

and, from (6), the following $4N$ "boundary" equations:

$$u_{-1,\,j} = u_{1,\,j} - 2h\alpha\imath u_{0,\,j}, \quad u_{N+2,\,j} = u_{N,\,j} - 2h\alpha\imath u_{N+1,\,j}, \quad j = 1(1)N,$$

$$u_{i,-1} = u_{i,1} - 2h\alpha_i u_{i,0}, \quad u_{i,N+2} = u_{i,N} - 2h\alpha_i u_{i,N+1}, \quad i = 1(1)N.$$

The above coupled "interior" and "boundary" equations can be written in matrix form as follows:

$$\mathcal{A}u = b, \tag{13}$$

where

$$u = [u_{0,0}u_{1,0}\cdots u_{N+1,0}u_{0,1}u_{1,1}\cdots u_{N+1,1}\cdots u_{0,N+1}u_{1,N+1}\cdots u_{N+1,N+1}]^T,$$

$$b = [b_{0,0}b_{1,0}\cdots b_{N+1,0}b_{0,1}d_{1,1}\cdots b_{N+1,1}\cdots b_{0,N+1}b_{1,N+1}\cdots b_{N+1,N+1}]^T,$$

$$\mathcal{A} = \begin{bmatrix} \mathcal{D}_0 & -2I & & & \\ -I & \mathcal{D}_1 & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & \mathcal{D}_N & -I \\ & & & -2I & \mathcal{D}_{N+1} \end{bmatrix} \in \mathbb{C}^{(N+2)^2\times(N+2)^2}, \tag{14}$$

and where $I \in \mathbb{R}^{(N+2)\times(N+2)}$ is the identity matrix,

$$\mathcal{D}_j = \begin{cases} \hat{\mathcal{D}}_j, & \text{for } j = 1(1)N, \\ \hat{\mathcal{D}}_j + 2h\alpha_i I, & \text{for } j = 0,\ N+1 \end{cases} \tag{15}$$

with

$$\hat{\mathcal{D}}_j = \begin{bmatrix} g_{0,j} + 2h\alpha_i & -2 & & & \\ -1 & g_{1,j} & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & g_{N,j} & -1 \\ & & & -2 & g_{N+1,j} + 2h\alpha_i \end{bmatrix}, \quad j = 0(1)N+1.$$

3. The Iterative Helmholtz Methods

3.1. Formulation

Select

$$J = diag\left(\frac{1}{\sqrt{2}}, 1, ..., 1\frac{1}{\sqrt{2}}\right) \in \mathbb{R}^{(N+2)\times(N+2)} \tag{16}$$

and set $S = J \otimes J$ to transform the original linear system (13) into the following equivalent one

$$Av = c, \tag{17}$$

where $A = S\mathcal{A}S^{-1}$, $v = Su$, $c = Sb$,

$$A = \begin{bmatrix} D_0 & -\sqrt{2}I & & & & & \\ -\sqrt{2}I & D_1 & -I & & & & \\ & -I & D_2 & -I & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -I & D_N & -\sqrt{2}I & \\ & & & & -\sqrt{2}I & D_{N+1} \end{bmatrix}, \tag{18}$$

$$D_j = \begin{cases} \hat{D}_j, & \text{for } j = 1(1)N, \\ \hat{D}_j + 2h\alpha\imath I, & \text{for } j = 0,\ N+1 \end{cases} \tag{19}$$

with

$$\hat{D}_j = \begin{bmatrix} g_{0,j} + 2h\alpha\imath & -\sqrt{2} & & & \\ -\sqrt{2} & g_{1,j} & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & g_{N,j} & -\sqrt{2} \\ & & & -\sqrt{2} & g_{N+1,j} + 2h\alpha\imath \end{bmatrix}, \ j = 0(1)N+1.$$

Note that the matrix A is obviously symmetric and is spectrally equivalent to $\mathcal{A}$. We now split A, v and c in real and imaginary parts

$$A = A_1 + \imath A_2, \quad v = v_{\mathbb{R}} + \imath v_{\mathbb{C}}, \quad c = c_{\mathbb{R}} + \imath c_{\mathbb{C}},$$

and transform the complex linear system (17) into the following equivalent real one:

$$\tilde{A}\tilde{v} = \tilde{c}, \tag{20}$$

where

$$\tilde{v} = \begin{bmatrix} v_{\mathbb{R}} \\ v_{\mathbb{C}} \end{bmatrix} \in \mathbb{R}^{2(N+2)^2}, \quad \tilde{c} = \begin{bmatrix} c_{\mathbb{R}} \\ c_{\mathbb{C}} \end{bmatrix} \in \mathbb{R}^{2(N+2)^2},$$

$$\tilde{A} = \begin{bmatrix} A_2 & A_1 \\ A_1 & -A_2 \end{bmatrix} \in \mathbb{R}^{2(N+2)^2 \times 2(N+2)^2}, \tag{21}$$

$$A_1 = \begin{bmatrix} B_0 & -\sqrt{2}I & & & & \\ -\sqrt{2}I & B_1 & -I & & & \\ & -I & B_2 & -I & & \\ & & \ddots & \ddots & \ddots & \\ & & & -I & B_N & -\sqrt{2}I \\ & & & & -\sqrt{2}I & B_{N+1} \end{bmatrix} \in \mathbb{R}^{(N+2)^2 \times (N+2)^2}, \tag{22}$$

$$A_2 = diag(C_0, C_1, ..., C_1, C_0) \in \mathbb{R}^{(N+2)^2 \times (N+2)^2} \tag{23}$$

with

$$C_0 = diag(2h\alpha, 0, ..., 0, 2h\alpha) + (2h\alpha + h^2q^2)I, \tag{24}$$

$$C_1 = diag(2h\alpha, 0, ..., 0, 2h\alpha) + h^2q^2I \tag{25}$$

and

$$B_j = \begin{bmatrix} b_{0,j} & -\sqrt{2} & & & & \\ -\sqrt{2} & b_{1,j} & -1 & & & \\ & -1 & b_{2,j} & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & b_{N,j} & -\sqrt{2} \\ & & & & -\sqrt{2} & b_{N+1,j} \end{bmatrix},$$

where $b_{i,j} = 4 - h^2a_{i,j}^2$.

We solve real system (20) using two variations of the block SOR method defined by the iteration matrix

$$T_{SOR} = (D - \omega L)^{-1}[(1 - \omega)D + \omega F], \tag{26}$$

one denoted by BSOR1 where

$$D_1 = \begin{bmatrix} A_2 & 0 \\ 0 & -A_2 \end{bmatrix}, \quad L_1 = \begin{bmatrix} 0 & 0 \\ -A_1 & 0 \end{bmatrix}, \quad F_1 = \begin{bmatrix} 0 & -A_1 \\ 0 & 0 \end{bmatrix} \tag{27}$$

and another one denoted by BSOR2 where

$$D_2 = \begin{bmatrix} A_1 & 0 \\ 0 & A_1 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 0 & 0 \\ -A_2 & 0 \end{bmatrix}, \quad F_2 = \begin{bmatrix} 0 & A_2 \\ 0 & 0 \end{bmatrix}. \tag{28}$$

The above block matrix partitions lead to the following two block iterative schemes:

- BSOR1

$$\begin{aligned} v_{\mathbb{R}}^{(n+1)} &= (1-\omega)v_{\mathbb{R}}^{(n)} - \omega A_2^{-1}A_1 v_{\mathbb{C}}^{(n)} + \omega A_2^{-1}c_{\mathbb{C}}, \\ v_{\mathbb{C}}^{(n+1)} &= \omega v_{\mathbb{R}}^{(n+1)} - (1-\omega)A_2^{-1}A_1 v_{\mathbb{C}}^{(n)} + \omega A_2^{-1}c_{\mathbb{R}}, \end{aligned} \tag{29}$$

- BSOR2

$$\begin{aligned} v_{\mathbb{R}}^{(n+1)} &= (1-\omega)v_{\mathbb{R}}^{(n)} - \omega A_1^{-1}A_2 v_{\mathbb{C}}^{(n)} + \omega A_1^{-1}c_{\mathbb{R}}, \\ v_{\mathbb{C}}^{(n+1)} &= \omega v_{\mathbb{R}}^{(n+1)} - (1-\omega)A_1^{-1}A_2 v_{\mathbb{C}}^{(n)} + \omega A_1^{-1}c_{\mathbb{C}}, \end{aligned} \tag{30}$$

where ω is the relaxation parameter of the SOR method.

3.2. Convergence analysis

The Jacobi iteration matrix $T_J = D_1^{-1}(L_1 + F_1)$ associated with the BSOR1 method can be written as

$$T_J = I_1 \otimes M, \text{ where } I_1 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \text{ and } M = A_2^{-1}A_1. \tag{31}$$

We remark here that, as it can be easily shown, the Jacobi matrix associated with the BSOR2 scheme is the inverse of the Jacobi iteration matrix of the BSOR1 scheme given in (31). For the rest of this section we concentrate solely on the BSOR1 scheme. We begin our analysis with the following lemma:

Lemma 1. *The spectral radius of the matrix M in* (31) *is bounded by*

$$\rho(M) \le k(h, a, q, \alpha) \equiv \frac{2(1-\sqrt{2}) - h^2 a_{\max}^2}{4h\alpha + h^2 q^2}, \tag{32}$$

where $a_{\max} \equiv \max_{0 \le i,\, j \le N+1} |\, a(x_i, y_j)\,|$.

Proof. Notice that the real matrix A_2 is diagonal so A_2^{-1} is readily available. The following are the only three non-zero values of the elements of A_2^{-1} : $p_1 = 1/h^2q^2,\ p_2 = 1/(2h\alpha + h^2q^2),\ p_3 = 1/(4h\alpha + h^2q^2)$. Let us set $p \equiv 4 - h^2 \max_{i,j}^2 \{| a_{i,j} |\}$ and observe that $| b_{i,j} | \le | p |$ to obtain

$$\| M \|_\infty = \max_{1 \le i \le (N+2)^2} \sum_{j=1}^{(N+2)^2} | m_{i,j} |$$

$$\le \max((| p | + 2\sqrt{2})\, p_1, (| p | + 2\sqrt{2} + 1)\, p_2,$$

$$(| p | + \sqrt{2} + 2)\, p_2, (| p | + 2\sqrt{2} + 2)\, p_3,$$

$$(| p | + \sqrt{2} + 3)\, p_3, (| p | + 4)\, p_3). \tag{33}$$

Since $p_1 < p_2 < p_3,$ we have

$$\| M \|_\infty \le (| p | + 2\sqrt{2} + 2)\, p_3. \tag{34}$$

Substitute p and p_3 in the above and recall that $\rho(M) \le \| M \|_\infty$ to derive (32) and conclude the proof of the lemma. □

Now let us apply a similarity transformation to matrix *M*,

$$A_2^{1/2} M A_2^{-1/2} = A_2^{1/2} A_2^{-1} A_1 A_2^{-1/2} = A_2^{-1/2} A_1 A_2^{-1/2} \equiv \overline{M}$$

and observe that since A_1 is symmetric and A_2 is diagonal, the matrix $\overline{M}$ is also symmetric and the matrices *M* and $\overline{M}$ are similar and thus they have the same spectrum of eigenvalues.

We will use the following well-known lemma later.

Lemma 2. *If* $A \in \mathbb{R}^{N \times N}$ *is symmetric, then*

(i) *the eigenvalues of A are real,*

(ii) *A has a set of* N^2 *linearly independent eigenvectors.*

If we use the above lemma and the similarity transformation for *M*, we

conclude that the eigenvalues μ_i, $i = 1, ..., (N+2)^2$, of M are real and M has a set of $(N+2)^2$ linearly independent eigenvectors.

The second lemma to be used later for the eigenvalue analysis of matrix T_j (see [14]) is stated as follows:

Lemma 3. *Let the $N_i \times N_i$ matrices A_i, $i = 1, ..., d$ possess complete sets of linearly independent set of eigenvectors $y^{(i, j_i)}$ with corresponding eigenvalues $\lambda_{A_i}^{(j_i)}$ $(j_i = 1, ..., N_i, i = 1, ..., d)$. Then the matrix $A \equiv A_1 \otimes A_2 \otimes \cdots \otimes A_d$ possesses the $\prod_{i=1}^{d} N_i$ linearly independent eigenvectors $y^{\underline{j}} \equiv y^{(1, j_1)} \otimes y^{(2, j_2)} \otimes \cdots \otimes y^{(d, j_d)}$, where $\underline{j} \equiv (j_1, ..., j_d)$ with corresponding eigenvalues $\lambda_{\underline{A}}^{(\underline{j})} \equiv \lambda_{A_1}^{(j_1)} \lambda_{A_2}^{(j_2)} \cdots \lambda_{A_d}^{(j_d)}$.*

The eigenvalues of I_1 are $\lambda_{1,2} = \pm\imath$, and since they are different, their associated eigenvectors are linearly independent. The matrices M and I_1 satisfy the conditions of Lemma 3, therefore the eigenvalues of T_J are $\pm\imath\mu_i$, $i = 1, ..., (N+2)^2$, and if we use (32), we will have that the spectral radius $\rho(T_J)$ is

$$\rho(T_J) = \rho(I_1 \otimes M) = \rho(I_1)\rho(M) \leq k(h, a, q, \alpha). \tag{35}$$

We also need the following theorem whose proof can be found in [13].

Theorem 4. *Let A be the block 2-cyclic consistently ordered coefficient matrix of a linear system in block partitioning form with square nonsingular diagonal blocks and T_J be the corresponding Jacobi matrix. If all the eigenvalues of T_J^2 are nonpositive, then, for the convergence of the block SOR method, we have the following conditions*:

$$\rho(T_{SOR}) < 1 \text{ if } 0 < \omega < \frac{2}{1 + \rho(T_J)}, \tag{36}$$

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 + \rho^2(T_J)}}. \tag{37}$$

By observing the fact that the matrix $\tilde{A}$ in (21) is block tri-diagonal and therefore 2-cyclic consistently ordered [27] and the fact that the eigenvalues of T_J^2 are $(\pm \imath\mu_i)^2 = -\mu_i^2 \leq 0,$ we can easily use Theorem 4 above to estimate regions of convergence for ω where the iteration scheme defined by (29) converges for any initial guess to the solution of (17). Furthermore, we can realize that for an estimate of $\omega_{opt},$ the value of the relaxation parameter ω that maximizes the rate of convergence, is the following:

$$\omega_{opt} \approx \omega_0 \equiv \frac{2}{1+\sqrt{1+k^2(h,\, a,\, q,\, \alpha)}}. \tag{38}$$

It is worth to point out that the convergence results obtained above directly relate the rate of convergence of the iterative method with the physical parameters of the problem *a*, *q* and α and the discretization parameter *h*. This might be of significant importance in "production" simulations for real life applications.

4. Numerical Experiments

This section contains selected numerical data obtained from a rather comprehensive experimental study whose objective was to confirm the theoretical results and to identify the convergence characteristics, the applicability and the efficiency of the proposed methods. We begin by giving a brief description of the experimental framework followed by the presentation of selected numerical data.

4.1. The experimental framework

We consider the PDE problem (5), (6) and select $a(x,\, y) = \frac{w}{c(x,\, y)}$ and $\alpha = w$ and $q \geq 0$ (q is proportional to the attenuating media capability which waves are propagated). For the wave speed $c(x,\, y),$ we choose the following four different functions which are rather typical in underwater acoustics

$$c_0(x,\, y) = 4,$$

$$c_1(x,\, y) = 2 + 2x^3 + y,$$

$$c_2(x,\ y) = e^{xy}(2 + \sin(\pi x))(2 - (\sin 4\pi y)),$$

$$c_3(x,\ y) = \begin{cases} 4, & x \le 0.5, \\ 1 + e^x + \sin(2\pi xy), & \text{otherwise.} \end{cases} \tag{39}$$

We select f in (5) such that the analytical solution u, is the one given in (7). The error is estimated on the maximum norm, r_∞^n, and the iteration procedure is terminated when $s_\infty^n \le 10^{-4}$, where

$$r_\infty^n = \| V^{(n)} - u \|_\infty, \qquad s_\infty^n = \frac{\| V^{(n)} - V^{(n-1)} \|_\infty}{\| V^{(n)} \|_\infty} \tag{40}$$

and where $V^{(n)}$ is the approximate solution at the nth iteration. Obviously the converged computed solution, is the same regardless which of the two BSOR methods is used. As initial value we have used $V^{(0)} \equiv (diag(A))^{-1}b$ and we have computed an estimate of the order of convergence (of the nth iterant) using the following expression:

$$\log\left(\frac{\| V_{h_1}^{(n)} - u \|_\infty}{\| V_{h_2}^{(n)} - u \|_\infty}\right) \Big/ \log\left(\frac{h_1}{h_2}\right), \tag{41}$$

where $h_1,\ h_2$ are the grid steps of two different uniform partitions of the domain and where $V_{h_1}^{(n)}$ and $V_{h_2}^{(n)}$ are the associated computed solutions at the nth iteration.

All programs for this study were implemented in MATLAB with double-precision arithmetic.

4.2. Numerical data

For the qualitative analysis of the convergence of our scheme we start by presenting in Figures 1 and 2 snapshots of the history of convergence of the real and imaginary parts of the computed solution for the case where $c = c_1(x,\ y),\ q = 5$ and $w = 10$ using $N = 20$ grid points in each space direction. As can be easily seen the “high frequency” components of the error are cut during the first few iterations while the rest of them try to level up the

low frequency terms. This feature might be very desirable in particular production-type runs of the model.

In Figure 3, we display the plots of the errors in the real (on the left) and the imaginary (on the right) parts of u on the discretization points for the previous PDE configuration. These plots exhibit the fact that no "singularities" have been introduced by the iterative process and the distribution of the error follows the variations of the solution itself. This observation also holds for any other model configuration we have been experimenting with.

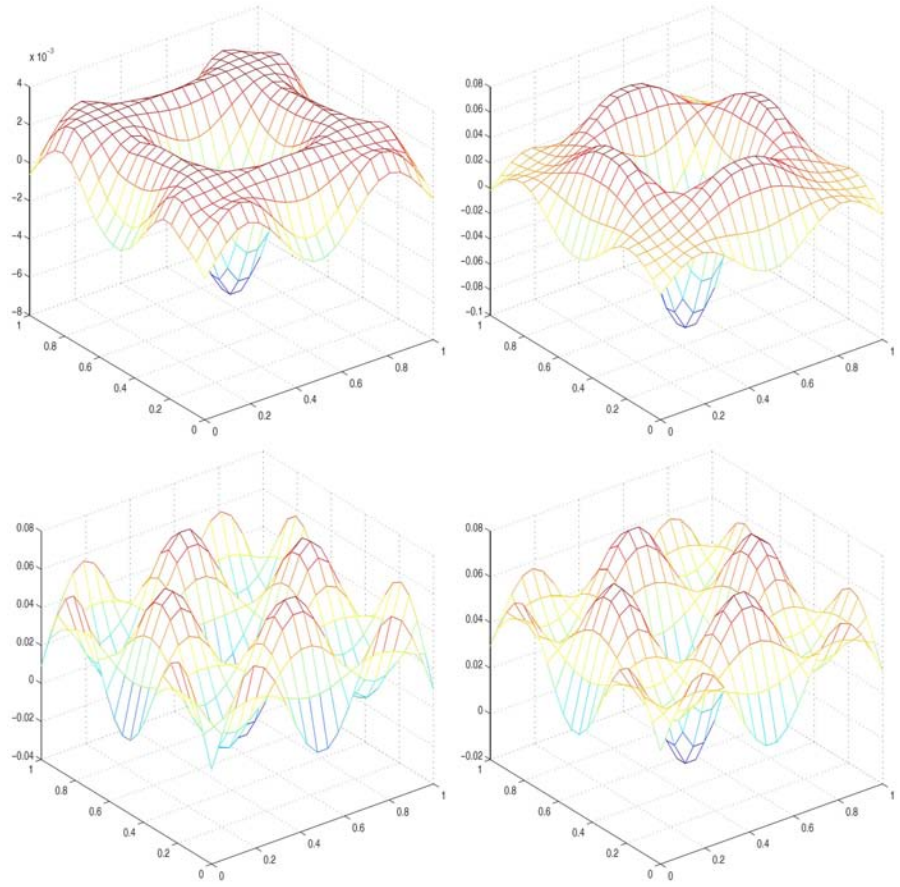

Figure 1. The real part of the computed solution after performing 2 (top-left), 10 (top-right), 100 (bottom-left) and 370 (bottom-right) iterations for $c(x, y) = c_1(x, y)$, $q = 5$ and $w = 10$ using $N = 20$.

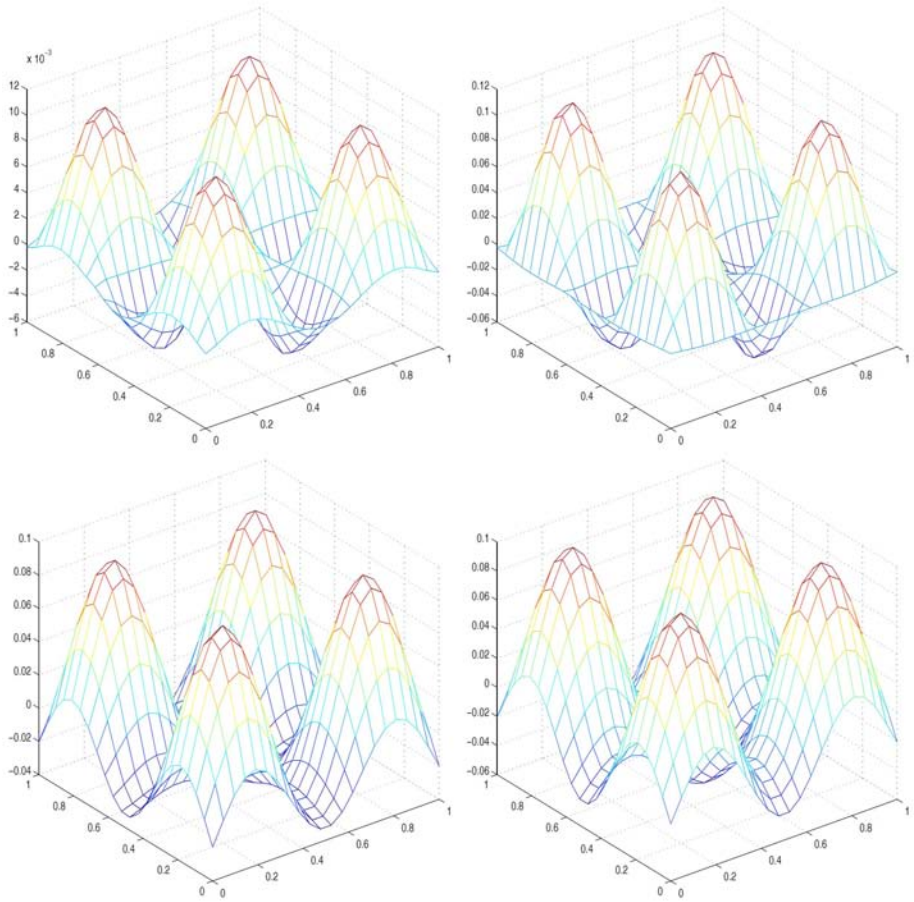

Figure 2. The imaginary part of the computed solution after performing 2 (top-left), 10 (top-right), 100 (bottom-left) and 370 (bottom-right) iterations for $c(x, y) = c_1(x, y)$, $q = 5$ and $w = 10$ using $N = 20$.

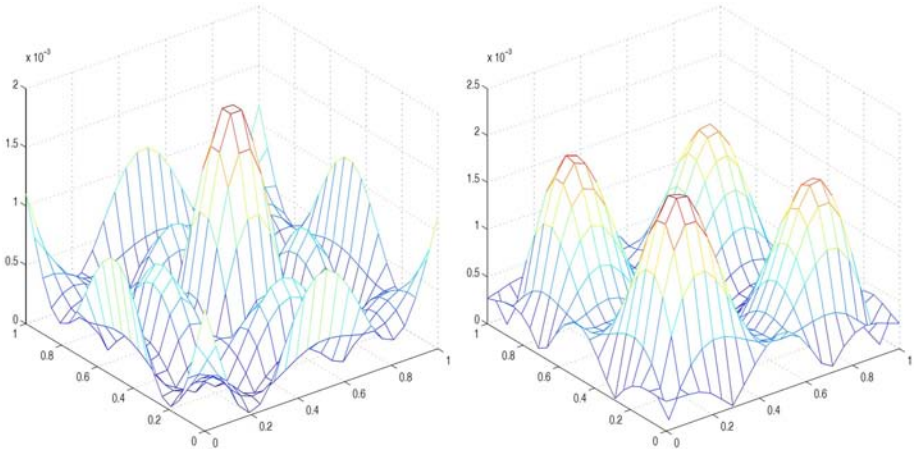

Figure 3. The real (left) and imaginary parts (right) of the absolute error $| V^{(370)} - u |$ for the problem considered in Figures 1 and 2.

In Table 1, we consider the BSOR1 scheme, we select $c = c_0(x, y)$ and we present data for three different PDE configurations $(q = 1, 5 \text{ and } 20)$, and five different discretizations with $N = 5, 10, 15, 20$ and 30 points in each space direction. The columns containing the number of required iterations k and the norm of the maximum error r_∞^k show that, the rate of convergence becomes smaller as w decreases. This is expected since as q approaches zero the condition number of matrix A_2 increases. This problem can be, to a great extent, easily avoided by relaxing our iterative scheme. For example BSOR1 can be modified so it converges uniformly with respect to q and in fact exhibits very rapid convergence for $q = 0$. For this, it is sufficient (as was proposed and analyzed in [19]) to use the following updating scheme for each iteration:

$$x \leftarrow x_{BSOR1} + \alpha(b - Ax_{BSOR1}) \quad \text{with} \quad \alpha = \frac{e^T(b - Ax_{BSOR1})}{q\| e \|^2},$$

where x_{BSOR1} is the solution computed by BSOR1 and where e is defined through an additional matrix "near" the null space of A. The comparison of

the columns containing the theoretically determined (through (38)) ω_θ and the numerically computed (through experimental systematic search) optimum values of the SOR relaxation parameter ω_ε easily establish numerically the validity of (38). The region of convergence that may be obtained by Theorem 4 and the preceding discussion was also carefully verified during the numerical systematic search for ω_ε. No significant variations have been observed for other model configurations.

Table 1. The discretization parameter *N*, the theoretical determined ω_θ and the experimentally determined ω_ε optimum values of the relaxation parameter ω, the number of iterations *k* required for convergence and the max-norm of the error at the *k*th iteration r_∞^k

	$q=1$				$q=5$				$q=20$			
N	ω_θ	ω_ϵ	k	r_∞^k	ω_θ	ω_ϵ	k	r_∞^k	ω_θ	ω_ϵ	k	r_∞^k
5	6.4(-3)	7.5(-3)	250	2.9(-2)	1.1(-1)	1.8(-1)	8	2.3(-2)	8.8(-1)	9.0(-1)	2	5.8(-3)
10	1.9(-3)	2.1(-3)	1330	9.0(-3)	4.6(-2)	5.2(-2)	86	6.8(-3)	5.2(-1)	5.6(-1)	7	2.0(-3)
15	8.7(-4)	9.8(-4)	3684	4.2(-3)	2.2(-2)	2.4(-2)	180	3.2(-3)	2.9(-1)	3.3(-1)	23	9.7(-4)
20	5.2(-4)	5.6(-4)	7236	2.4(-3)	1.1(-2)	1.4(-2)	311	1.9(-3)	1.9(-1)	2.0(-1)	40	5.7(-4)
30	2.4(-4)		-	-	5.9(-3)	6.5(-3)	909	8.5(-4)	9.0(-2)	1.0(-1)	81	2.8(-4)

We use (41) and the results in Table 1 to compute the order of convergence of the whole numerical scheme. As it is apparent from the slopes of the straight lines in Figure 4, where we plot the discretization parameter *N* versus the max-norm of the associated computed solution r_∞^k in log-log scale, the order of convergence is approximately equal to two. This confirms the second order of convergence of our discretization numerical scheme and in fact the consistency of our iteration scheme.

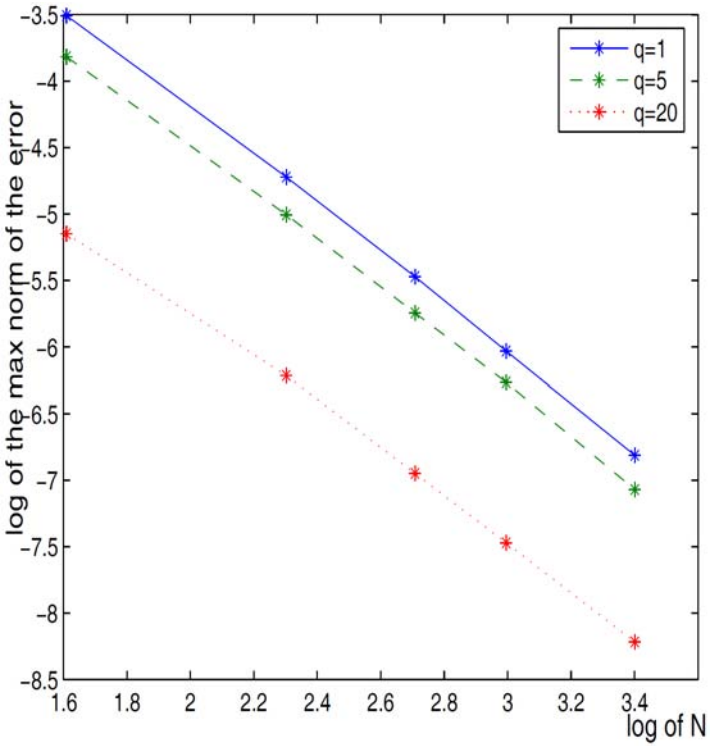

Figure 4. The log-log plot of the maximum norm of the error r_∞^k of the computed solution versus N for $q = 1, 5, 20$ for the data considered in Table 1.

In Table 2, we compare the convergence behavior of BSOR1 and BSOR2 on four different model configurations by presenting the norm of the error r_∞^k together with the numerically determined optimum value of the relaxation parameter ω_ε and the associated required number of iterations k. From these data we can see that the two BSOR methods compliment each other in the sense that in the cases where one converges very slowly the other converges quite rapidly. In particular, in the case where $q \to 0$ (diminishing dissipation) the condition number of A_2 becomes large and therefore the denominator of $\rho(M)$ approaches zero. This leads to slow converges for the BSOR1 but to rather rapid convergence for BSO2.

Table 2. The number of iterations required for convergence *k*, the max-norm of the error of the convergent solution r_∞^k and the experimentally determined optimum value of the relaxation parameter ω_ε associated with both BSOR1 and BSOR2 for various configurations of *q*, *w* and $c(x,\ y)$ using $N = 20$

				BSOR1		BSOR2	
$c(x,y)$	q	w	r_∞^k	ω_ϵ	k	ω_ϵ	k
$c_0(x,y)$	.001	10	2.4×10^{-3}			0.05	18
$c_0(x,y)$	1	10	2.4×10^{-3}	0.00056	7236	0.12	28
$c_1(x,y)$	5	1	7.8×10^{-5}	0.014	931	0.008	1186
$c_2(x,y)$	10	100	5.7×10^{-2}	0.057	101	0.01	36
$c_3(x,y)$	25	10	4.6×10^{-4}	0.3	25	0.00059	9349

In Figure 5, we graphically present the number of iterations required for convergence by both BSOR1 and BSOR2 methods for several values of ω when $c(x,\ y) = c_0(x,\ y),\ q = 5,\ w = 10$ and $N = 5, 10, 15, 20$. The curves in this figure lead to the conclusion that the optimum value of the relaxation parameter ω of the block SOR is, as expected, conversely proportional to the number of iterations *N*. Furthermore, it is seen that if ω is properly selected, then no more that 100 iterations are required for convergence for any step size up to $N = 30$.

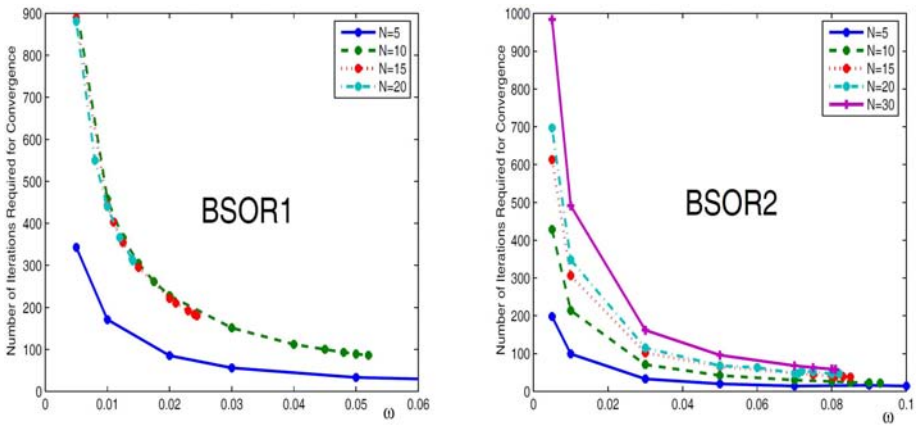

Figure 5. The number of iterations required for convergence by the BSOR1 (left) and BSOR2 (right) methods versus ω when $c(x,\ y) = c_0(x,\ y),\ q = 5,$ $w = 10$ with $N = 5, 10, 20, 30$.

Finally, in Table 3 we compare the CPU time, the floating point operations (flops) and the required memory (in words) by the MATLAB's LU solver (a direct solver based on banded LU factorization) and BSOR1 for $N = 5, 10, 15, 20, 30$. The CPU time of BSOR1 is significantly greater than the CPU time of LU. This is due to the fact that the LU solver is highly optimized (with BLAS routines in assembly) while BSOR1 is the plane implementation where absolutely no effort has been made to increase its performance. In addition, the BSOR1 CPU time contains everything including for example checking the convergence criterion at every iteration. As far as the number of flops is concerned, we see that BSOR1 competes very well with LU. The main advantage of BSOR1 is in the memory required which is much less than the memory required for LU. We may remark here that a further increase of *N* will exhibit further significant improvement in the performance of the BSOR compared to the LU solver.

Table 3. The CPU time, executed number of floating point operations (flops) and the required memory for LU and BSOR1 respectively for $c = c_0(x, y)$, $w = 10$ and $q = 20$ with $N = 5, 10, 15, 20$ and 30

	CPU time		Flops		Memory	
N	LU	BSOR1	LU	BSOR1	LU	BSOR1
5	0.01	0.0	1.44×10^5	3.18×10^4	5.88×10^2	5.88×10^2
10	0.03	0.03	1.64×10^6	6.87×10^5	3.18×10^3	1.72×10^3
15	0.12	0.16	8.23×10^6	7.48×10^6	9.24×10^3	3.47×10^3
20	0.331	0.912	2.77×10^7	3.6×10^7	2.03×10^4	5.80×10^3
30	3.045	16.32	1.66×10^8	3.41×10^8	6.34×10^4	1.22×10^4

5. Epilogue

The main objective of this paper is to develop and analyze effective iterative methods for the solution of the complex linear system of algebraic equations that arise from the discretization of the complex linear Helmholtz problem in two dimensions.

In the course of our work first we considered elements from the scientific area of modeling acoustic propagation problems. In particular, we focused on the Helmholtz equation on a rectangular region with first order absorbing

boundary conditions. This mathematical model was discretized via the popular five-point-star finite difference scheme which led us to a complex linear system of algebraic equations.

We studied the properties of the complex system and reformulated it into an equivalent real system of double dimension. This system is naturally partitioned into a 2×2 block structure. Based on this block partition, and after transforming this system by means of a similarity transformation, we proposed two block SOR type methods.

We theoretically analyzed one of these two iterative methods, namely the one denoted by BSOR1. Specifically, we proved its convergence and determined bounds of its regions of convergence in terms of the various physical parameters of the mathematical model.

A comprehensive experimental study of the two proposed methods has been carried out in order to

- Verify the theoretically determined results and in particular ensure that the iterative methods considered converge to the analytic solution at the appropriate discretization level.
- Examine the convergence characteristics of the proposed methods their applicability and their effectiveness in problems with different values of the physical parameters.
- Compare their efficiency in terms of both time and space complexity against the well known direct method available in MATLAB.

From the theoretical and experimental results presented in this paper, one can easily conclude the following facts:

- The two proposed methods do converge for a large class of different configurations of the Helmholtz mathematical problem.
- The theoretical results have been confirmed and proved themselves as a useful practical tool for experimentation.
- The two proposed methods are complimentary to each other in the sense that it was very often observed that for configurations where

one converges slowly the other converges rather fast. This becomes apparent by comparing the 6th and the 8th columns in Table 2.

- The BSOR1 method competes well with the main direct method found in MATLAB. Specifically, its time complexity is "slightly" higher while its memory complexity is drastically lower.

The problem of deriving effective iterative methods for solving the linear Helmholtz PDE problem still remains interesting and challenging. In particular, it seems that the best combination of iterative solver and preconditioner is problem dependent [26]. One can clearly identify several extensions of the work presented in this paper that they might significantly contribute to the search for such an optimal scheme. Such open research directions that are beyond the scope of this paper, are the following:

1. Formulate and analyze a hybrid iterative method that combines the two proposed methods in the spirit of the basic SSOR iterative schemata. It is worth to note that the SSOR preconditioner gave the best performance on a recent related study [18].
2. Use one of the proposed two methods as a preconditioner to a Krylov-based method. Since BSOR1 compares well with the direct solver then a BSOR1/bi-conjugate gradient combination seems to be very attractive.
3. Investigate the effectiveness of the proposed methods on complex problem configurations like the ones considered in [25] where non-local/high order boundary conditions are imposed, obstacles exist, non-homogeneous media are considered .... Such configurations usually lead to p-cycle matrices (with $p > 2$) and as it is known (see [27]) basic iterative methods, like the ones considered in this study, can outperform the commonly used Krylov subspace methods by properly exploiting the cyclicity of the matrix.
4. Finally, it is worth to mention that our approach, and to some extent our theoretical and experimental analysis, can be easily extended to more dimensions, parabolic variants of the Helmholtz equation and more complex, higher order absorbing boundary conditions.

Acknowledgements

The author would like to thank Professor A. Hadjidimos for his advices and Mr. M. Kiagadakis for performing several numerical experiments for this study. This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

References

[1] O. Axelsson and A. Kucherov, Real valued iterative methods for solving complex symmetric linear systems, Numer. Linear Algebra Appl. 7 (2000), 197-218.

[2] Gang Bao and Weiwei Sun, A fast algorithm for the electromagnetic scattering from a large cavity, SIAM J. Sci. Comput. 27(2) (2005), 553-574.

[3] A. Bayliss, C. Goldstein and E. Turkel, An iterative method for the Helmholtz equation, J. Comput. Phys. 49 (1985), 443-457.

[4] A. Bayliss, C. I. Goldstein and E. Turkel, The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics, Comput. Math. Appl. 11 (1985), 655-665.

[5] Y. Boubendir, An analysis of the BEM-FEM non-overlapping domain decomposition method for a scattering problem, J. Comput. Appl. Math. 204(2) (2007), 282-291.

[6] H. Elman and D. O'Leary, Eigenanalysis of some preconditioned Helmholtz problems, Numer. Math. 83(2) (1999), 231-257.

[7] R. W. Freund, Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrix, SIAM J. Sci. Statist. Comput. 13 (1992), 425-448.

[8] R. W. Freund and N. M. Nachtigal, An implementation of the QMR method based on coupled two-term recurrences, SIAM J. Sci. Stat. Comput. 15 (1994), 313-337.

[9] L. Halpern, G. Cohen and P. Joly, Mathematical and Numerical Aspects of Wave Propagation Phenomena, SIAM, New York, 1991.

[10] M. J. Gander and F. Nataf, An incomplete lu preconditioner for problems in acoustics, J. Comput. Acoust. 13(3) (2005), 455-476.

[11] H. De Gersem, D. Lahaye, S. Vandewalle and K. Hameyer, Comparison of quasi minimal residual and bi-conjugate gradient iterative methods to solve complex symmetric systems arising from time-harmonic magnetic simulations, COMPEL - The Int. J. Comp. Math. Electrical and Electronic Eng. 18 (1999), 298-310.

[12] L. Giraud, J. Langou and G. Sylvand, On the parallel solution of large industrial wave propagation problems, J. Comput. Acoust. 14(1) (2006), 83-111.

[13] A. Hadjidimos, A survey of the iterative methods for the solution of linear systems by extrapolation, relaxation and other techniques, J. Comput. Appl. Math. 20 (1987), 213-230.

[14] A. Hadjidimos, E. N. Houstis, J. R. Rice and E. Vavalis, Analysis of iterative line spline collocation methods for elliptic pdes, SIAM J. Matrix Anal. Appl. 12 (1999), 508-521.

[15] V. P. Ilin and A. V. Petukhov, On numerical solution of the complex Helmholtz equation, Russian J. Numer. Anal. Math. Model. 22(1) (2007), 19-37.

[16] J. Douglas, Jr., J. L. Hensley and J. E. Roberts, An alternating-direction iteration method for Helmholtz problems, Technical Report 214, Mathematics Department, Purdue University, W. Lafayette, IN, 1993.

[17] R. Kechroud, A. Soulaimani, Y. Saad and S. Gowda, Preconditioning techniques for the solution of the Helmholtz equation by the finite element method, Math. Comput. Simul. 65(4-5) (2004), 303-321.

[18] A. L. Laird and M. B. Giles, Preconditioning harmonic unsteady potential flow calculations, AIAA J. 44(11) (2006), 2654-2662.

[19] Y.-J. Lee, J. Wu, J. Xu and L. Zikatanov, Robust subspace correction methods for nearly singular systems, M3AS (Mathematical Models and Methods in Applied Sciences), page to appear, 2007.

[20] Y. Maday and F. Magoules, Absorbing interface conditions for domain decomposition methods: a general presentation, Comput. Methods Appl. Mech. Eng. 195(29-32) (2006), 3880-3900.

[21] S. Petersen, D. Dreyer and O. von Estorff, Assessment of finite and spectral element shape functions for efficient iterative simulations of interior acoustics, Comput. Methods Appl. Mech. Eng. 195(44-47) (2006), 6463-6478.

[22] C. D. Riyanti, Y. A. Erlangga, R. Plessix, W. A. Mulder, C. Vuik and C. Oosterlee, A new iterative solver for the time-harmonic wave equation, Geophysics 71(5) (2006).

[23] C. D. Riyanti, A. Kononov, Y. A. Erlangga, C. Vuik, C. W. Oosterlee, R. Plessix and W. A. Mulder, A parallel multigrid-based preconditioner for the 3d heterogeneous high-frequency Helmholtz equation, J. Comput. Phys. 224(1) (2007), 431-448.

[24] Y. Saad, Iterative Methods for Sparse Linear Systems, PWS Publishing Company, Boston, 1996.

[25] S. Schneider and S. Marburg, Performance of iterative solvers for acoustic problems. Part ii. Acceleration by ilu-type preconditioner, Engineering Analysis with Boundary Elements 27(7) (2003), 751-757.

[26] L. L. Thompson, A review of finite-element methods for time-harmonic acoustics, J. Acoustical Society of America 119(3) (2006), 1315-1330.

[27] R. S. Varga, Matrix Iterative Analysis, Prentice Hall, New Jersey, 1962.