DELFT UNIVERSITY OF TECHNOLOGY

REPORT 06-16

Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian

M.B. van Gijzen, Y.A. Erlangga, C. Vuik

ISSN 1389-6520

Reports of the Department of Applied Mathematical Analysis

Delft 2006

Copyright © 2006 by Department of Applied Mathematical Analysis, Delft, The Netherlands.

No part of the Journal may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission from Department of Applied Mathematical Analysis, Delft University of Technology, The Netherlands.

Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian.

M.B. van Gijzen*, Y.A. Erlangga† and C. Vuik‡

Abstract

Shifted Laplace preconditioners have attracted considerable attention as a technique to speed up convergence of iterative solution methods for the Helmholtz equation. In this paper we present a comprehensive spectral analysis of the Helmholtz operator preconditioned with a shifted Laplacian. Our analysis is valid under general conditions. The propagating medium can be heterogeneous, and the analysis also holds for different types of damping, including a radiation condition for the boundary of the computational domain. By combining the results of the spectral analysis of the preconditioned Helmholtz operator with an upper bound on the GMRES-residual norm we are able to provide an optimal value for the shift, and to explain the mesh-dependency of the convergence of GMRES preconditioned with a shifted Laplacian. We illustrate our results with a seismic test problem.

Keywords: Helmholtz equation, shifted Laplace preconditioner, iterative solution methods, GMRES, convergence analysis.

1 Introduction

In this paper we investigate the spectral behavior of iterative methods applied to the time-harmonic wave equation in heterogeneous media. The underlying equation governs wave propagation and scattering phenomena arising in acoustic problems in many areas, e.g., in aeronautics, marine technology, geophysics, and optical problems. In particular, we look for solutions of the Helmholtz equation discretized by using finite difference, finite volume or finite element discretizations. Since the number of grid points per wavelength should be sufficiently large to result in acceptable solutions, for very high frequencies the discrete problem becomes extremely large, prohibiting the use of direct solution methods. Krylov subspace iterative methods are an interesting alternative. However, Krylov subspace methods are not competitive without a good preconditioner.

Finding a suitable preconditioner for the Helmholtz equation is still an area of active research, see for example [7]. A class of preconditioners that has recently attracted considerable attention is the class of shifted Laplace preconditioners. Preconditioning of the Helmholtz equation using the Laplace operator without shift was first suggested in [1].

*Delft University of Technology, Delft Institute of Applied Mathematics, Mekelweg 4, 2628 CD, The Netherlands. E-mail: M.B.vanGijzen@tudelft.nl

†Technische Universität Berlin, Institut für Mathematik, Strasse des 17. Juni 136, D-10623 Berlin, Germany. E-mail: erlangga@math.tu-berlin.de

‡Delft University of Technology, Delft Institute of Applied Mathematics, Mekelweg 4, 2628 CD, The Netherlands. E-mail: C.Vuik@tudelft.nl

This approach has been enhanced in [8, 9] by adding a positive shift to the Laplace operator, resulting in a positive definite preconditioner. In [2, 3, 4, 13] the class of shifted Laplace preconditioners is further generalized by also considering general complex shifts.

It is well known that the spectral properties of the preconditioned matrix give important insight in the convergence behavior of the preconditioned Krylov subspace methods. Spectral analyses for the Helmholtz equation preconditioned by a shifted Laplace operator have previously been given in [2, 3, 4]. The analysis in [2], however, is restricted to the homogeneous physical parameters case, for a purely imaginary shift preconditioner. This analysis concerns the singular values of the preconditioned matrix rather than the eigenvalues. Furthermore, only reflecting and pressure release boundary conditions are considered. In [3], a convergence analysis of GMRES is discussed, under the same restriction as in [2]. A more thorough spectral analysis is presented in [4] for the case that the preconditioning operation is performed approximately by using multigrid. The analysis is based on Rigorous Fourier Analysis (RFA) for homogeneous physical parameters. Results from RFA, however, give little insight in the convergence of Krylov subspace methods.

This paper gives a spectral analysis from an algebraic point of view. Therefore, the results are valid under rather general conditions:
- they do not depend on the discretization method,
- inhomogeneous physical parameters (like sound speed, density and damping) are allowed,
- the analysis is also valid for various types of boundary conditions (reflecting, radiation, pressure release and PML as well).
These generalizations are new, and they allow us to analyse shifted-Laplace preconditioners for a much wider class of discrete Helmholtz problems than in previous publications.

By combining the results of our analysis with a bound on the norm of the GMRES-residual, we are able to derive a 'quasi' optimal value for the shift. This result is also new. The shift is derived under the assumption that the preconditioning operations with the shifted-Laplace preconditioner are performed sufficiently accurately. In practice, the preconditioning operations are performed only approximately, for example by using a multi-grid method or by making an ILU decomposition of the shifted-Laplace operator. Of course, our results do not hold unconditionally in these cases. However, the results that are reported in [2, 3, 4], where the preconditioning operations are performed approximately, using either a multi-grid method or ILU, are obtained using shifts that are close to the optimal shift that follows from our analysis. This indicates that the optimal shift that we derive in this paper also gives strong guidelines for choosing the shift in the case where the preconditioning operations are performed approximately.

This paper is organized as follows. In Section 2 we describe the acoustic wave equation and its discretization. We specify some properties of the matrices (symmetry, complex valued, positive definiteness etc.), which form the coefficient matrix of the resulting linear system. In Section 3 we show that for the damped Helmholtz equation with Dirichlet and Neumann boundary conditions, the eigenvalues are located on a line or on a circle with a given parameterization for a special type of damping. For radiation boundary conditions and for general viscous media we show that the eigenvalues are located on one side of the line or within the circle. In Section 4 we use a simple bound on the GMRES-residual norm. Using this bound and the results of our spectral analysis we are able to derive the optimal value of the shift in the shifted Laplacian preconditioner. We also show for a number of applications that the convergence of GMRES is independent of the grid-size. Section 5

contains numerical experiments to illustrate and verify the theoretical results derived in Sections 3 and 4. Finally, Section 6 contains the conclusions of this paper.

2 The Helmholtz equation

2.1 The acoustic wave equation

An acoustic medium with space-varying density $\rho(\mathbf{x})$ and sound speed $c(\mathbf{x})$ occupies the volume Ω, bounded by the boundary $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$. In addition, the medium is assumed to be viscous with damping coefficient $\gamma(\mathbf{x})$. The wave equation for the acoustic pressure $p(\mathbf{x}, t)$ (with $\mathbf{x}$ spatial coordinates, and t time) in such a medium is:

$$\frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2} + \frac{\gamma}{\rho}\frac{\partial p}{\partial t} - \nabla \cdot \frac{1}{\rho}\nabla p = \frac{s(\mathbf{x}, t)}{\rho} \quad \text{in } \Omega, \tag{1}$$

where ∇ denotes the gradient operator and $\nabla\cdot$ the divergence. Realistic conditions on the physical boundaries of an acoustic medium can be reflecting boundaries, which are described by the homogeneous Neumann condition

$$\frac{\partial p}{\partial n} = 0 \quad \text{on } \Gamma_1, \tag{2}$$

pressure release boundaries, which are described by the homogeneous Dirichlet condition

$$p = 0 \quad \text{on } \Gamma_2. \tag{3}$$

and radiating boundaries, which can be described by

$$\frac{\partial p}{\partial n} = -\frac{1}{\rho c}\frac{\partial p}{\partial t} \quad \text{on } \Gamma_3, \tag{4}$$

where n is the outward pointing normal unit vector.

We will assume that the right-hand side function $s(\mathbf{x}, t)$ is a harmonic point source $s(\mathbf{x}, t) = ae^{2\pi i f t}\delta(\mathbf{x} - \mathbf{x}_s)$, located at $\mathbf{x}_s$, which transmits a signal with amplitude a and frequency f. Here, $i = \sqrt{-1}$.

2.2 The Helmholtz equation.

If the source term is harmonic, then the pressure field has the factored form

$$p(\mathbf{x}, t) = \hat{p}(\mathbf{x})e^{2\pi i f t}. \tag{5}$$

Substitution of (5) into (1) yields the so called Helmholtz equation

$$(\frac{-(2\pi f)^2}{\rho c^2} + 2\pi i f\frac{\gamma}{\rho})\hat{p} - \nabla \cdot \frac{1}{\rho}\nabla\hat{p} = \frac{a}{\rho}\delta(\mathbf{x} - \mathbf{x}_s) \quad \text{in } \Omega, \tag{6}$$

with boundary conditions

$$\frac{\partial \hat{p}}{\partial n} = 0, \quad \text{on } \Gamma_1 \tag{7}$$

$$\hat{p} = 0 \quad \text{on } \Gamma_2. \tag{8}$$

and

$$\frac{\partial \hat{p}}{\partial n} = -\frac{2\pi i f}{\rho c}\hat{p} \quad \text{on } \Gamma_3. \tag{9}$$

This latter condition is also known as a Sommerfeld condition of the first kind.

If the damping parameter has the special form

$$\gamma(\mathbf{x}, f) = 2\pi f \frac{\nu}{c^2}, \tag{10}$$

with ν a non-negative constant, the Helmholtz equation (6) simplifies to

$$(-\frac{(2\pi f)^2}{\rho c^2}(1 - i\nu) - \nabla \cdot \frac{1}{\rho}\nabla)\hat{p} = a\frac{\delta(\mathbf{x} - \mathbf{x}_s)}{\rho(\mathbf{x}_s)} \quad \text{in } \Omega. \tag{11}$$

Clearly, the assumption holds for non-viscous media, i.e. with $\gamma = 0$.

The above equation can be discretized with a discretization method like the finite element method, finite volume method or the finite difference method. Discretization of the above equation plus boundary conditions with any of these methods yields a discrete Helmholtz equation of the form

$$(L + iC - z_1 M)x = b \tag{12}$$

in which L is the discretization of $-\nabla \cdot \frac{1}{\rho}\nabla$, M corresponds to the discretized zeroth order term $\frac{1}{\rho c^2}$, C corresponds to the Sommerfeld condition and/or to damping that does not satisfy (10), and b to the source term. The complex number z_1 is defined by

$$z_1 = (2\pi f)^2(1 - i\nu) \ . \tag{13}$$

Both L and C are real symmetric and positive semi-definite, and the matrix M is real symmetric and positive definite. The matrix $L+iC-z_1M$, however, is complex symmetric and indefinite.

For high frequencies, system (12) can be very large, in particular in 3D. This is a consequence of the fact that each wavelength has to be sampled with sufficient grid points. Numbers of unknowns in excess of 10^6 for realistic models are quite common. Fortunately, system (12) is sparse. Krylov-type iterative solvers like GMRES [12] or Bi-CGSTAB [14] are among the most popular techniques for solving large and sparse linear systems. They have proved to be particularly efficient for systems with an Hermitian positive-definite matrix, or more generally, for systems with a matrix with all eigenvalues in the right half of the complex plane. Helmholtz-type systems like (12), however, are highly indefinite, which means that the system matrix has eigenvalues with both negative and positive real parts, a characteristic that can result in a very slow convergence. In order to overcome this problem a suitable preconditioner has to be applied. A class of promising preconditioners that has attracted a lot of attention is the class of shifted Laplace preconditioners [1, 8, 2, 3, 4]. In the next section we will analyze these preconditioners by locating in the complex plane the spectrum of the preconditioned discrete Helmholtz operator.

3 Spectral analysis of the Helmholtz operator preconditioned with the shifted Laplace preconditioner.

Shifted Laplace preconditioners are preconditioners of the form

$$P = L + iC - z_2 M,$$

i.e. the same form as the discrete Helmholtz operator $A = L + iC - z_1 M$. The shift-parameter z_2 has to be chosen such that the convergence of GMRES (or another suitable iterative method) applied to the preconditioned system

$$(L + iC - z_2 M)^{-1}(L + iC - z_1 M)x = (L + iC - z_2 M)^{-1} b$$

is considerably faster than of GMRES applied to the original system. Moreover, z_2 has to be chosen such that operations with the inverse of $(L + iC - z_2M)$ are easy to perform. In practice this means that z_2 is chosen such that operations with the inverse of the preconditioner can be computed using a fast multigrid method [4]. Note that $L + iC$ is the operator that is being shifted. This means that all boundary conditions, including the Sommerfeld condition, are included in the preconditioner, as recommended in [10].

The complex numbers z_1 and z_2 can be written as

$$z_1 = \alpha_1 + i\beta_1 \qquad\qquad z_2 = \alpha_2 + i\beta_2, \tag{14}$$

in which $\alpha_1, \beta_1, \alpha_2$ and β_2 are real. Recall that for our application z_1 is defined by (13), and hence

$$\alpha_1 > 0 \ , \qquad\qquad \beta_1 \leq 0 \ .$$

Choices for the shift z_2 that are considered in literature are $z_2 = 0$ [1], $z_2 = -\alpha_1$ [8], $z_2 = -i\alpha_1$ [2, 3], and $z_2 = (1 - 0.5i)\alpha_1$ [4].

In this section we will study the spectrum of the preconditioned system. The spectrum governs to a large extent the convergence of iterative methods as long as the matrix of the eigenvectors is well conditioned.

3.1 The spectrum of the preconditioned Helmholtz operator without Sommerfeld condition

We will first assume that $C = 0$ and hence that the damped Helmholtz operator is given by $L - z_1M$. We recall that L is symmetric positive semi-definite, M symmetric positive definite and z_1 is a complex number. We will consider a shifted Laplace preconditioner, i.e. a matrix of the form $L - z_2M$, as preconditioner for the Helmholtz operator, and analyze how the location of the eigenvalues σ of the preconditioned system depends on the parameters z_1 and z_2. The eigenvalues σ of the preconditioned matrix are solutions of the generalized eigenproblem

$$(L - z_1M)x = \sigma(L - z_2M)x. \tag{15}$$

It is easy to see that the matrices $(L - z_1M)$ and $(L - z_2M)$ share the same eigenvectors, which are the eigenvectors of

$$Lx = \lambda Mx. \tag{16}$$

Since for our problem L is symmetric positive semi-definite and M symmetric positive definite, the eigenvalues λ are real and non-negative. Substitution of λMx for Lx in (15) yields

$$(\lambda - z_1)Mx = \sigma(\lambda - z_2)Mx$$

and hence

$$\lambda - z_1 = \sigma(\lambda - z_2), \tag{17}$$

which, if $z_2 \neq \lambda$, gives

$$\sigma = \frac{\lambda - z_1}{\lambda - z_2}. \tag{18}$$

Note that if z_2 coincides with an eigenvalue of (16), i.e. if $z_2 = \lambda$, the preconditioner $P = L - z_2M$ will be singular, which is a situation that has to be avoided. So in the following we assume that $z_2 \neq \lambda$ for all eigenvalues of (16). The eigenvalues λ can be considered as a real parameterization of the curves (18) in the complex plane on which the eigenvalues σ of the preconditioned system are located.

To determine these curves we write $\sigma = \sigma^r + i\sigma^i$ and substitute this into (17), which yields

$$\lambda - \alpha_1 - i\beta_1 = \sigma^r(\lambda - \alpha_2) - i\sigma^r\beta_2 + i\sigma^i(\lambda - \alpha_2) + \sigma^i\beta_2.$$

We can split this equation into an equation for the real terms and one for the imaginary terms:

$$\lambda - \alpha_1 = \sigma^r(\lambda - \alpha_2) + \sigma^i\beta_2 \ ,$$
$$-\beta_1 = -\sigma^r\beta_2 + \sigma^i(\lambda - \alpha_2) \ .$$

If $\beta_1 = \sigma^r\beta_2$, the second equation reduces to $\sigma^i = 0$. If this is not the case we get for λ that

$$\lambda = \alpha_2 + \frac{\sigma^r\beta_2 - \beta_1}{\sigma^i} \ .$$

Substitution of λ in the equation for the real terms yields the following result

$$\beta_2(\sigma^r)^2 - (\beta_1 + \beta_2)\sigma^r + \beta_2(\sigma^i)^2 + (\alpha_1 - \alpha_2)\sigma^i = -\beta_1. \tag{19}$$

This equation is valid for all values of $\alpha_1, \beta_1, \alpha_2$ and β_2, including the case $\beta_1 = \sigma^r\beta_2$.

In the following we will distinguish between the cases $\beta_2 = 0$ and $\beta_2 \neq 0$. Theorem 3.1 deals with the case $\beta_2 = 0$.

Theorem 3.1 *Let $\beta_2 = 0$ and let L be symmetric positive semi-definite and M be symmetric positive definite real matrices. Then the eigenvalues $\sigma = \sigma^r + i\sigma^i$ of (15) are located on the straight line in the complex plane given by*

$$-\beta_1\sigma^r + (\alpha_1 - \alpha_2)\sigma^i + \beta_1 = 0 \ . \tag{20}$$

Proof The result follows directly from substituting $\beta_2 = 0$ in (19). △

The next theorem characterises the spectrum in the case that $\beta_2 \neq 0$.

Theorem 3.2 *Let $\beta_2 \neq 0$ and let L be symmetric positive semi-definite and M be symmetric positive definite real matrices. Then the eigenvalues $\sigma = \sigma^r + i\sigma^i$ of (15) are located on the circle given by*

$$(\sigma^r - \frac{\beta_2 + \beta_1}{2\beta_2})^2 + (\sigma^i - \frac{\alpha_2 - \alpha_1}{2\beta_2})^2 = \frac{(\beta_2 - \beta_1)^2 + (\alpha_2 - \alpha_1)^2}{(2\beta_2)^2}. \tag{21}$$

The center c of this circle is

$$c = (\frac{\beta_2 + \beta_1}{2\beta_2}, \frac{\alpha_2 - \alpha_1}{2\beta_2})$$

and the radius R is

$$R = \sqrt{\frac{(\beta_2 - \beta_1)^2 + (\alpha_2 - \alpha_1)^2}{(2\beta_2)^2}}$$

Proof Divide (19) by $(2\beta_2)$ and complete the square. △

To understand the convergence of iterative methods it is important to know if the origin is enclosed by the circle given in theorem 3.2. The following theorem gives a simple condition that determines this.

Theorem 3.3 *If $\beta_1\beta_2 > 0$ the origin is not enclosed by the circle (21) given in Theorem 3.2.*

Proof The origin is not enclosed by the circle if the distance of the center to the origin is larger than the radius. Hence if

$$\frac{(\beta_2+\beta_1)^2+(\alpha_2-\alpha_1)^2}{(2\beta_2)^2} > \frac{(\beta_2-\beta_1)^2}{(2\beta_2)^2}+\frac{(\alpha_2-\alpha_1)^2}{(2\beta_2)^2}$$

which is clearly the case. △

Remark The center of the circle can also be written as

$$c = \frac{z_1-\overline{z}_2}{z_2-\overline{z}_2}$$

and the radius as

$$R = |\frac{z_2-z_1}{z_2-\overline{z}_2}|.$$

3.2 The spectrum of the preconditioned Helmholtz operator with Sommerfeld condition

We will now study the general damped Helmholtz operator $L+iC-z_1M$. As before L and C are symmetric positive semi-definite matrices, M is a symmetric and positive definite matrix and z_1 is a complex number. For our problem, the matrix C stems from the discretization of the Sommerfeld boundary condition, or from damping that does not satisfy (10). This means that for example a damping matrix that stems from an absorbing layer is also covered by the theory below. We consider a shifted Laplace preconditioner, i.e. a matrix of the form $L+iC-z_2M$, to precondition the Helmholtz operator. The eigenvalues of this matrix are given by

$$(L+iC-z_1M)x = \sigma_S(L+iC-z_2M)x. \tag{22}$$

Let λ_S be an eigenvalue of the generalized problem

$$(L+iC)x = \lambda_S M x\ . \tag{23}$$

As in the previous section it is straightforward to show that (22) and (23) share the same eigenvectors x and that the eigenvalues σ_S of the preconditioned system are related by the eigenvalues λ_S by

$$(\lambda_S-z_2)\sigma_S = \lambda_S - z_1 \tag{24}$$

The main difference with the previous section is that λ_S is *complex*, whereas λ in the previous section was real, which allowed us to consider λ as a real valued parameterization of a curve in the complex plane. Although the eigenvalues σ_S will in general not be located on a straight line or on a circle in the complex plane if λ_S is complex, it is still possible to establish useful results regarding the location of σ_S. To this end, we will distinguish between the three cases $\beta_2=0$, $\beta_2>0$ and $\beta_2<0$. Before we proceed we will formulate the following Lemma that we will need in the remainder of this section.

Lemma 3.1 *Let L and C be symmetric positive semi-definite and let M be symmetric positive definite real matrices. Then the eigenvalues $\lambda_S=\lambda_S^r+i\lambda_S^i$ of the generalized eigenproblem (23) have a nonnegative imaginary part.*

Proof We use the fact that any matrix can be split into two Hermitian matrices:

$$A = \frac{1}{2}(A + A^H) + i\frac{1}{2i}(A - A^H) = \Re(A) + i\Im(A) \tag{25}$$

where

$$\Re(A) = \frac{1}{2}(A + A^H) \quad \text{and} \quad \Im(A) = \frac{1}{2i}(A - A^H) \ . \tag{26}$$

According to Bendixon's theorem, see e.g. [6], page 69, we have

$$\lambda_{min}^{\Re(A)} \leq Re(\lambda^A) \leq \lambda_{max}^{\Re(A)},$$

$$\lambda_{min}^{\Im(A)} \leq Im(\lambda^A) \leq \lambda_{max}^{\Im(A)}.$$

The eigenvalues λ_S of the generalized problem (23) are also solutions of the standard eigenproblem

$$U^{-1}(L + iC)U^{-T}y = \lambda_S y$$

in which $M = UU^T$. This means that we can take $A = U^{-1}(L + iC)U^{-T}$, in which case $Im(A) = U^{-1}CU^{-T}$. This latter matrix is positive semi-definite, so by Bendixon's theorem we have $\lambda_S^i \geq 0$ △

As in the previous section we first consider the case $\beta_2 = 0$.

Theorem 3.4 *Let $\beta_2 = 0$ and let L and C be symmetric positive semi-definite and M be symmetric positive definite real matrices. Then the eigenvalues $\sigma_S = \sigma_S^r + i\sigma_S^i$ of (22) are located in the half-plane*

$$-\beta_1\sigma_S^r + (\alpha_1 - \alpha_2)\sigma_S^i + \beta_1 \geq 0 \ .$$

Proof Since $\beta_2 = 0$ we have

$$(\lambda_S - \alpha_2)\sigma_S = \lambda_S - z_1$$

Splitting this equation into an equation for the real terms and one for the imaginary terms yields

$$\lambda_S^r\sigma_S^r - \lambda_S^i\sigma_S^i - \alpha_2\sigma_S^r = \lambda_S^r - \alpha_1$$

and

$$\lambda_S^r\sigma_S^i + \lambda_S^i\sigma_S^r - \alpha_2\sigma_S^i = \lambda_S^i - \beta_1 \ .$$

The second equation gives that either $\sigma_S^i = 0$ or that

$$\lambda_S^r = \alpha_2 - \frac{\beta_1}{\sigma_S^i} + \lambda_S^i\frac{1 - \sigma_S^r}{\sigma_S^i} \quad .$$

Substitution in the first equation and some straightforward manipulations yields

$$-\beta_1\sigma_S^r + (\alpha_1 - \alpha_2)\sigma_S^i + \beta_1 = \lambda_S^i((\sigma_S^r - 1)^2 + {\sigma_S^i}^2).$$

By Lemma 3.1 $\lambda_S^i \geq 0$, and hence the right-hand-side term is larger than or equal to zero. △

If $\beta_2 < 0$ the spectrum of the preconditioned matrix is characterised by Theorem 3.5.

Theorem 3.5 *Let $\beta_2 < 0$ and let L and C be symmetric positive semi-definite and M be symmetric positive definite real matrices. Then the eigenvalues σ_S of (22) are inside or on the circle with center $c = \frac{z_1 - \overline{z}_2}{z_2 - \overline{z}_2}$ and radius $R = |\frac{z_2 - z_1}{z_2 - \overline{z}_2}|$.*

Proof We have to prove that $|\sigma_S - c| \leq R$ if $\beta_2 < 0$.

$$\begin{aligned}
|\sigma_S - c| &= |\frac{\lambda_S - z_1}{\lambda_S - z_2} - \frac{z_1 - \overline{z}_2}{z_2 - \overline{z}_2}|, \\
&= |\frac{(\lambda_S - z_1)(z_2 - \overline{z}_2) - (\lambda_S - z_2)(z_1 - \overline{z}_2)}{(\lambda_S - z_2)(z_2 - \overline{z}_2)}|, \\
&= |\frac{\lambda_S(z_2 - z_1) + (z_1 - z_2)\overline{z}_2}{(\lambda_S - z_2)(z_2 - \overline{z}_2)}|, \\
&= |\frac{\lambda_S - \overline{z}_2}{\lambda_S - z_2}\frac{z_2 - z_1}{z_2 - \overline{z}_2}|, \\
&= |\frac{\lambda_S - \overline{z}_2}{\lambda_S - z_2}|R \qquad (27)
\end{aligned}$$

What is left to prove is that $|\frac{\lambda_S - \overline{z}_2}{\lambda_S - z_2}| \leq 1$. Writing

$$\lambda_S = \lambda_S^r + i\lambda_S^i$$

we get

$$|\frac{\lambda_S - \overline{z}_2}{\lambda_S - z_2}|^2 = \frac{(\lambda_S^r - \alpha_2)^2 + (\lambda_S^i + \beta_2)^2}{(\lambda_S^r - \alpha_2)^2 + (\lambda_S^i - \beta_2)^2}. \qquad (28)$$

Since $\beta_2 < 0$ and by Lemma 3.1 $\lambda_S^i \geq 0$ we have

$$\frac{(\lambda_S^r - \alpha_2)^2 + (\lambda_S^i + \beta_2)^2}{(\lambda_S^r - \alpha_2)^2 + (\lambda_S^i - \beta_2)^2} \leq 1,$$

and hence the above condition is satisfied. △

If $\beta_2 > 0$, the spectrum of the preconditioned matrix is characterised by Theorem 3.6.

Theorem 3.6 *Let $\beta_2 > 0$ and let L and C be symmetric positive semi-definite and M be symmetric positive definite real matrices. Then the eigenvalues σ_S of (22) are outside or on the circle with center $c = \frac{z_1 - \overline{z}_2}{z_2 - \overline{z}_2}$ and radius $R = |\frac{z_2 - z_1}{z_2 - \overline{z}_2}|$.*

Proof Analogous to the proof of Theorem 3.5. △

Remark. The results presented above specify regions in the complex plane where the eigenvalues of the preconditioned matrix are located. These regions are completely determined by the parameters z_1 and z_2. Given the definition of z_1, (13), these regions only depend on the frequency f, on the damping parameter ν, and of course on the shift for the preconditioner z_2. It is important to note that the regions in the complex plane where the eigenvalues are located do not depend on other physical parameters like the sound speed or density, nor on computational parameters like the size of the matrix or on the mesh size h.

4 Combination of the results of the spectral analysis with an upper bound on the GMRES-residual norm

In this section we will combine the results of the spectral analysis presented in the previous section with a well-known upper bound on the GMRES-residual norm. This upper bound assumes that the spectrum is enclosed by a circle, and hence this bound can be naturally combined with the circle specified in Theorems 3.2 and 3.5.

Let the eigenvalues of the preconditioned matrix be enclosed by a circle with radius R and center c as in Theorem 3.5. Then the GMRES-residual norm after k iterations $\|r^k\|$ satisfies (see e.g. [12])

$$\frac{\|r^k\|}{\|r^0\|} \leq c_2(X) \left(\frac{R}{|c|} \right)^k \ . \tag{29}$$

In this equation X is the matrix of eigenvectors and $c_2(X)$ its condition number in the 2-norm. If this condition number is large the upper bound gives no information about the convergence, since in that case there is no relation between the location of the eigenvalues and the convergence behavior of the preconditioned Krylov method [5]. Fortunately, in our application we may expect that the condition number of the eigenvector matrix is relatively small. If $C = 0$ the eigenvectors of the preconditioned matrix are the same as of (16) and hence independent of the shift parameters. Moreover, since the eigenvectors of (16) are $M-$orthogonal and M is a (scaled) mass matrix which is in general well conditioned, we expect that $c_2(X)$ will be small in practice. This can be seen from

$$X^T M X = I \ \Leftrightarrow \ c_2(X^T M X) = 1 \ \Leftrightarrow \ c_2(M^{\frac{1}{2}} X) = 1 \ \ .$$

Since

$$c_2(X) = c_2(M^{-\frac{1}{2}} M^{\frac{1}{2}} X) \leq c_2(M^{-\frac{1}{2}}) c_2(M^{\frac{1}{2}} X) \ ,$$

we get

$$c_2(X) \leq \sqrt{c_2(M)} \ .$$

If $C \neq 0$, the eigenvectors of the preconditioned system are the same as of (23). These are unfortunately not $M-$orthogonal, but for many problems we can consider (23) as a relatively small perturbation of (16), in which case we can still expect that $c_2(X)$ is small.

4.1 Optimization of the shift

Although equation (29) only gives an upper bound on the GMRES-residual norm, it allows us to derive a 'quasi' optimal choice for the shift. We derive this shift by minimizing the upper bound. For this it is sufficient to minimize the ratio $\frac{R}{|c|}$, or, using Theorem 3.5, the function

$$f(\alpha_2, \beta_2) = \frac{R^2}{|c|^2} = \frac{(\alpha_2 - \alpha_1)^2 + (\beta_2 - \beta_1)^2}{(\alpha_2 - \alpha_1)^2 + (\beta_2 + \beta_1)^2}.$$

To analyze this function we differentiate with respect to α_2,

$$\frac{\partial f}{\partial \alpha_2} = \frac{8(\alpha_2 - \alpha_1)\beta_1\beta_2}{((\alpha_2 - \alpha_1)^2 + (\beta_2 + \beta_1)^2)^2},$$

and with respect to β_2

$$\frac{\partial f}{\partial \beta_2} = \frac{4\beta_1((\beta_2^2 - \beta_1^2) - (\alpha_2 - \alpha_1)^2)}{((\alpha_2 - \alpha_1)^2 + (\beta_2 + \beta_1)^2)^2}.$$

Clearly, both derivatives are zero at $\alpha_1 = \alpha_2, \beta_1 = \beta_2$. This choice for the shift minimizes of course the upper bound since this corresponds to using the original operator as preconditioner, which means that performing the preconditioning operation is as hard as solving the original system.

We are interested in the case where the preconditioning operation is relatively cheap. In particular, we have in mind the situation where preconditioning operations can be efficiently carried out using a fixed number of cycles of a multigrid method for the whole

range of shifts under consideration. We therefore restrict our analysis to values for the shift for which multigrid is known to work well. We first consider the purely imaginary shift [2, 3, 4], this means that α_2 equals zero. In this case, the derivative with respect to β_2 is zero if

$$(\beta_2^2 - \beta_1^2) - \alpha_1^2 = 0 \tag{30}$$

yielding

$$\beta_2 = \pm|z_1|.$$

Since by (13) $\beta_1 = -(2\pi f)^2\nu \leq 0$, we must choose by Theorem 3.3 $\beta_2 \leq 0$, and hence $z_2 = -|z_1|i$ as the shift that minimizes the upper bound (29). This choice is also optimal if we consider all possible shifts for which $\alpha_2 \leq 0$, meaning all possible shifts for which the preconditioner has all its eigenvalues in the right-half plane. By (13), $\alpha_1 = (2\pi f)^2 > 0$, and by Theorem 3.3 $\beta_1\beta_2 \geq 0$, so $\frac{\partial f}{\partial \alpha_2}$ is negative for $\alpha_2 \leq 0$. Therefore $f(\alpha_2, \beta_2)$ takes its minimum on the edge $\alpha_2 = 0$. We conclude that the choice

$$z_2 = -|z_1|i \tag{31}$$

minimizes the upper bound (29) for all $z_2 \in \mathbb{C}$, with $\alpha_2 \leq 0$.

The same methodology for deriving an optimal shift can still be used if we do not restrict ourselves to the case $\alpha_2 \leq 0$. Such a shift still (approximately) minimises the number of GMRES iterations. However, the performance of a multigrid method for the preconditioning operations will deteriorate if z_2 is too close to z_1, and hence such a shift would no longer minimise the total work of the whole solution process. How to find a shift that minimises the total work, if the performance of multigrid depends on the shift, is of great practical importance, but is outside the scope of this paper.

4.2 Discussion

The upper bound (29) is only meaningful if the circle does not enclose the origin. This is the case if $\beta_1 < 0$, or equivalently if $\nu > 0$. However, because of continuity arguments, the result (31) for the 'quasi' optimal shift is still valid if $\beta_1 = 0$.

As was remarked in the previous section, the circle around the spectrum of the preconditioned matrix only depends on z_2, and on f and ν. Consequently, if $\beta_1 < 0$ inequality (29) yields an upper bound on the GMRES-residual norm that also only depends on the frequency f and on the damping parameter ν. Because of this, the number of GMRES-iterations should be bounded from above by a constant that is independent of the mesh size.

By scaling the shift z_2 with the frequency we can make the upper bound on the number of GMRES-iterations also independent of frequency. To this end we introduce the scaled shift

$$\tilde{z_2} = \tilde{\alpha}_2 + i\tilde{\beta}_2 = \frac{z_2}{(2\pi f)^2} \ .$$

Applying Theorem 3.5 and substituting $z_2 = (2\pi f)^2\tilde{z_2}$ and the definition for z_1 (13) into (29) yields

$$\frac{\|r^k\|}{\|r^0\|} \leq c_2(X)\left(\frac{R}{|c|}\right)^k = \ c_2(X)\sqrt{\frac{(\tilde{\alpha}_2 - 1)^2 + (\tilde{\beta}_2 + \nu)^2}{(\tilde{\alpha}_2 - 1)^2 + (\tilde{\beta}_2 - \nu)^2}}. \tag{32}$$

Clearly, this upper bound only depends on the the damping parameter ν and on the choice for the parameter $\tilde{z}_2$.

5 Experiments

In this section we describe a typical test problem with a variable sound velocity. The location of the eigenvalues of the discretized operators are compared with the theoretically predicted locations. The value of the optimal shift is validated by numerical experiments. Finally, it appears that the convergence behavior of GMRES is independent of the mesh size.

5.1 Description of the test problem

The test problem that we consider mimics three layers with a simple heterogeneity, and is taken from ([11]).

For $\nu \in \mathbb{R}$, find $p \in \mathbb{C}^N$ satisfying:

$$\begin{cases} -\Delta p - (1-\nu)(\frac{2\pi f}{c(\mathbf{x})})^2 p = s, \quad \text{in } \Omega = (0,600) \times (0,1000) \text{ meter}^2 \\ s = \delta(x_1 - 300, x_2), \qquad x_1 = (0,600), x_2 = (0,1000) \\ \textit{with Sommerfeld conditions or Neumann conditions on } \Gamma \equiv \partial\Omega. \end{cases} \tag{33}$$

The local sound velocity is given as in Figure 1. The density is assumed to be constant.

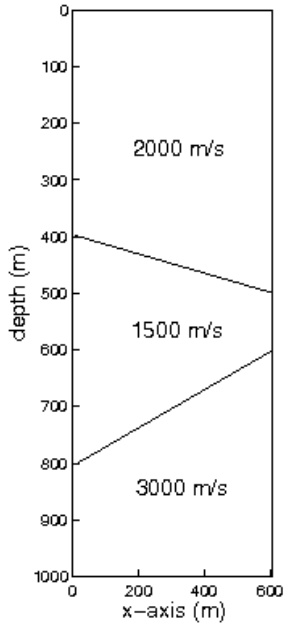

Figure 1: Problem geometry with sound velocity profile.

We have discretized the above problem with the finite element method using linear triangular elements. The computations that are described in this section have been performed with MATLAB.

5.2 Location of the eigenvalues

The first experiments validate the theorems that are presented in Section 3. To this end we have taken as source frequency $f = 2$ and we have discretized the problem with mesh size

$h = 100/2$. We have calculated all the eigenvalues of the preconditioned matrix for four typical combinations of values of the scaled parameters $\tilde{z}_1$ and $\tilde{z}_2$. The upper left-hand-side

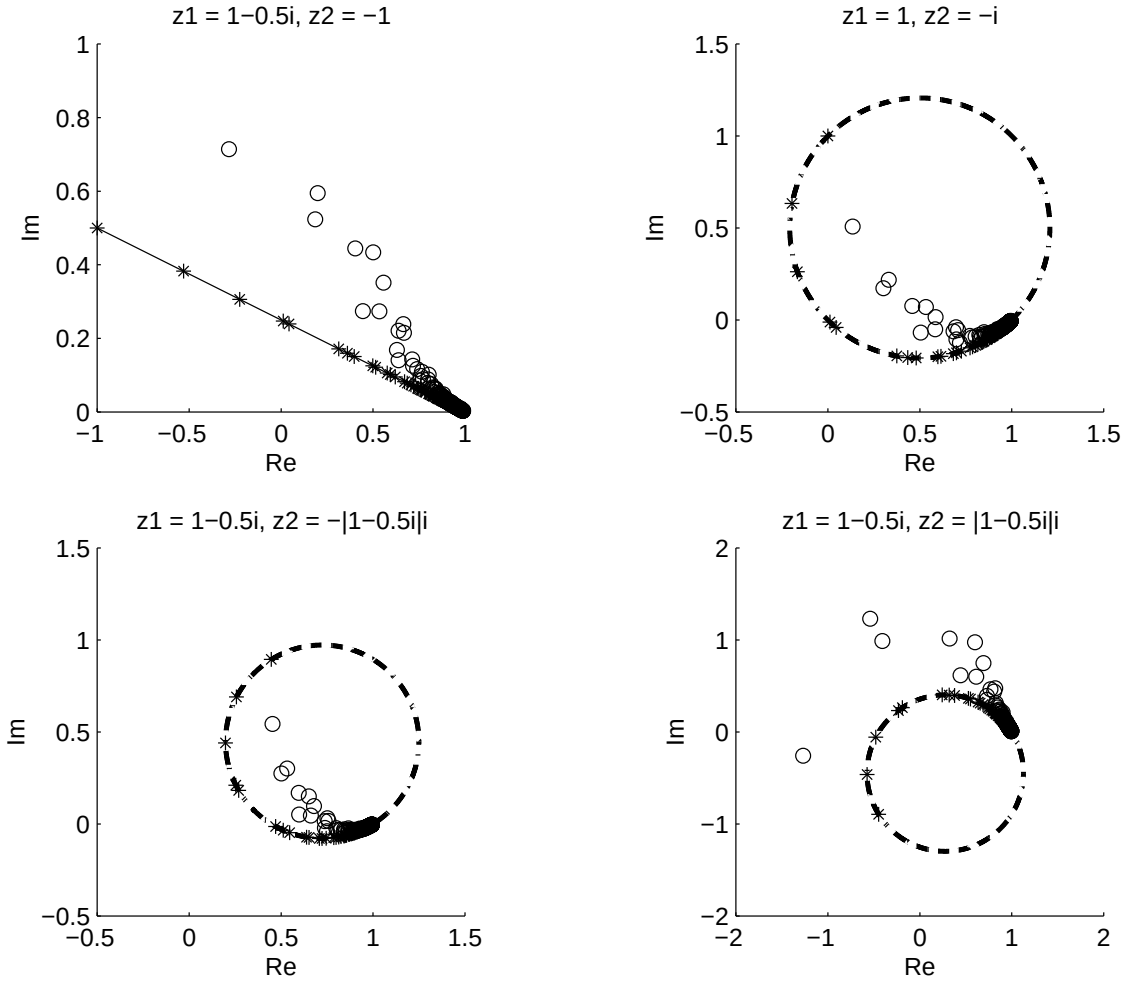

Figure 2: Spectra for different values of the complex shift in the preconditioner; $h = 100/2, f = 2$. '*' denote eigenvalues for the Neumann problem, 'o' are eigenvalues for the Sommerfeld problem.

subplot shows the spectrum of the preconditioner if a real shift is chosen, as suggested in [8]. As was predicted by Theorem 3.1, the eigenvalues for the Neumann problem, which are indicate with the symbol '*', are located on a line. Since the example contains damping, the line does not pass through the origin. The eigenvalues of the Sommerfeld problem, which are indicated with the symbol 'o', are all on one side of the line, as is predicted by Theorem 3.4. Note that the eigenvalues move away from the origin if the Neumann problem is replaced by the Sommerfeld problem.

The upper right-hand-side subplot shows the spectrum of the preconditioner if a purely imaginary shift is chosen. The eigenvalues of the Neumann problem are located on the circle that is given by Theorem 3.2, and the eigenvalues of the Sommerfeld problem are, as predicted by Theorem 3.5 on or inside this circle. This example does not contain damping (apart from the radiation condition): $\tilde{z}_1$ is real. Consequently, the circle contains the origin.

The lower left-hand-side subplot shows another picture of the spectrum of the preconditioner if a purely imaginary (negative) shift is chosen. In this case the problem contains damping since $\tilde{z}_1$ is complex. As a result, the circle is smaller than for the previous example, and the origin is outside the circle.

The lower right-hand-side subplot shows for the same z_1 what happens if the sign of the complex shift is chosen wrongly (i.e. positively). According to Theorem 3.6 the eigenvalues of the Sommerfeld problem should in this case be on or outside the circle. This is confirmed by the numerical results. Moreover, by Theorem 3.3, the origin should be enclosed by the circle, which is the case.

5.3 Optimization of the shift

The second group of experiments validates the optimal value for the shift z_2 that was found in Section 4. This value is given by equation (31).

The optimal value was determined by minimizing the ratio $\frac{R}{|c|}$. Using the scaled variables $\tilde{z}_2 = z_2/(2\pi f)^2$ this ratio can be written as

$$\frac{R}{|c|} = \sqrt{\frac{(\tilde{\alpha}_2 - 1)^2 + (\tilde{\beta}_2 + \nu)^2}{(\tilde{\alpha}_2 - 1)^2 + (\tilde{\beta}_2 - \nu)^2}}. \tag{34}$$

This function takes values between 0 and 1. A small value of $\frac{R}{|c|}$ indicates fast convergence and a value close to 1 slow convergence. Figure 3 shows for three different damping parameters ν how the value of $\frac{R}{|c|}$ depends on $\tilde{\alpha}_2$ and $\tilde{\beta}_2$. The values on the contour lines correspond to the value of $\frac{R}{|c|}$. The three plots show clearly that (34) takes its minimum

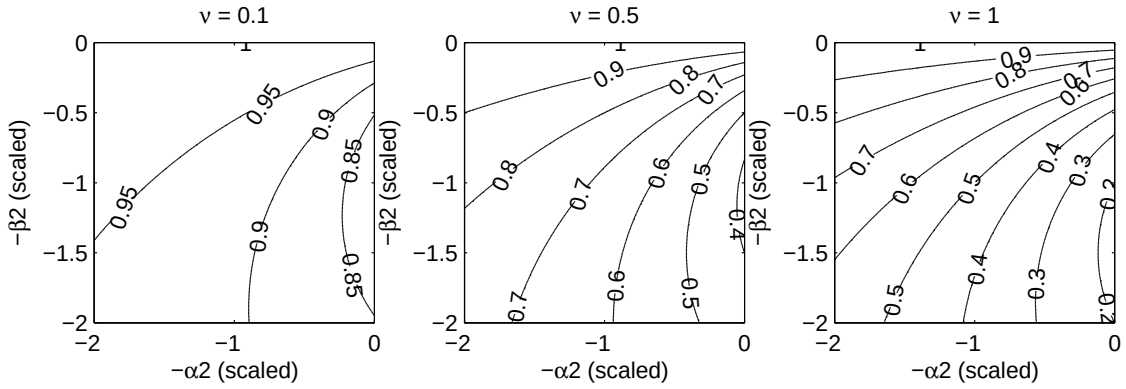

Figure 3: Contour plot of the convergence factor as function of the complex shift.

when $\tilde{\alpha}_2 = 0$, and that the optimal $\tilde{\beta}_2$ becomes more negative when the damping parameter is increased. These observations are of course consistent with the optimal value for the shift parameter (31) that was derived in Section 4.

To validate that this value is really (sub-)optimal in actual computations we solve the Sommerfeld problem with scaled imaginary shifts ranging from 0 to -2. For the mesh size we take $h = 100/8$ and we perform the experiment for four different damping parameters. The result is shown in Figure 4. Clearly, the more damping the fewer the number of GMRES-iterations, and the larger (more negative) the optimal imaginary shift.

Table 1 shows the minimum number of iterations and compares this with the number of iterations when the 'optimal' shift (31) is used. The results show that the shift (31) is

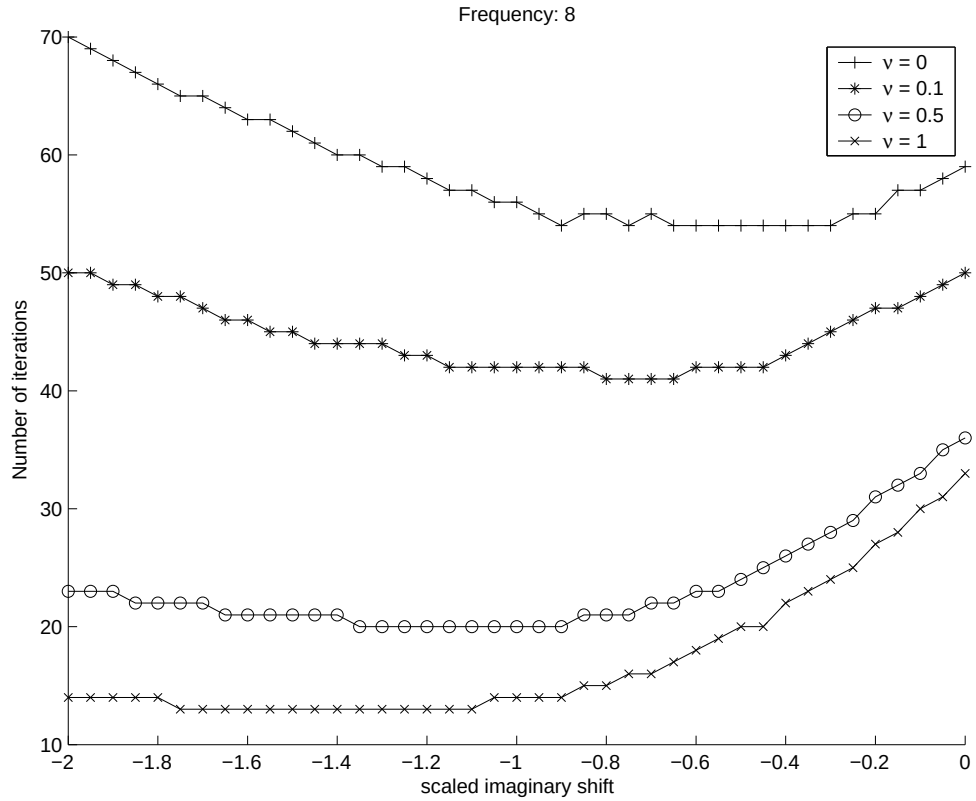

Figure 4: Actual number of iteration as function of the imaginary shift ($h = 100/8$).

Damping	Optimal shift	Iterations "optimal" shift	Min. number of iterations
$\nu = 0$	1	56	54
$\nu = 0.1$	1.005	42	41
$\nu = 0.5$	1.118	20	20
$\nu = 1$	1.4142	13	13

Table 1: Number of iterations for "optimal" shift and minimum number of iterations

nearly optimal with respect to the number of GMRES-iterations.

5.4 Mesh dependency.

The last set of experiments examines dependency of the number of iterations of preconditioned GMRES on the mesh size.

Table 2 shows for an increasingly fine step size h the number of iterations for the Sommerfeld problem. The experiment is performed for four different damping parameters, and the frequency is kept fixed to $f = 2$. The results show that for all four different values of the damping parameter the number of iterations is independent of the step size h. Based on the discussion at the end of Section 4 this could be expected. The theory that is presented in Section 4, however, does not make any predictions about the mesh dependent performance of preconditioned GMRES for problems with a zero damping parameter.

The results of the same type of experiments, but now with a frequency that scales with the mesh size, is tabulated in Table 3. These results confirm that if the damping

	Number of iterations				
h :	100/2	100/4	100/8	100/16	100/32
f :	2	2	2	2	2
$\nu = 0$	14	13	13	13	13
$\nu = 0.1$	13	12	12	12	13
$\nu = 0.5$	11	10	11	11	11
$\nu = 1$	9	9	9	9	9

Table 2: Number of iterations under mesh refinement for a fixed frequency.

parameter is nonzero, the number of GMRES-iterations is bounded by a number that is independent of the mesh size. This is most apparent in the results for $\nu = 0.5$ and $\nu = 1$. The results for $\nu = 0$ seem to indicate that the number of GMRES-iterations more or less doubles if the step size is halved. As was remarked above, the theory presented in Section 4 does not make any predictions for the case that $\nu = 0$.

	Number of iterations				
h :	100/2	100/4	100/8	100/16	100/32
f :	2	4	8	16	32
$\nu = 0$	14	25	56	116	215
$\nu = 0.1$	13	22	42	63	80
$\nu = 0.5$	11	16	20	23	23
$\nu = 1$	9	11	13	13	13

Table 3: Number of iterations under mesh refinement for increasingly high frequencies.

To check that $c_2(X)$ is actually small for the above test cases we have also computed the condition numbers of the mass matrices on the five meshes. These condition numbers are equal to 24 for all meshes, hence we have that

$$c_2(X) \leq \sqrt{c_2(M)} = 2\sqrt{3} \; .$$

6 Conclusions

We have presented a spectral analysis of the Helmholtz operator that is preconditioned with a shifted Laplace operator. We have shown that, depending on the value of the shift, the eigenvalues of the preconditioned matrix are located in or on a circle, or in a half-plane. Combination of these results concerning the spectrum of the preconditioned matrix with a well-known bound on the GMRES-residual norm allowed us to determine a close to optimal shift. Furthermore, we have shown for problems with a nonzero damping parameter that there is an upper bound on the number of GMRES iterations that only depends on the damping parameter and hence is independent of the mesh-size, the frequency, the sound speed and the density.

We have derived the close-to-optimal shift for the shifted-Laplace preconditioner in combination with GMRES, under the assumption that preconditioning operations are performed exactly. In practice, however, preconditioning operations are performed approximately, for example using a multi-grid method, and another Krylov method like Bi-CGSTAB [14] may be used instead of GMRES. In this case the analysis that has been presented in this paper does not hold anymore. However, experimental results reported

in [3], where Bi-CGSTAB is used as Krylov solver and preconditioning operations are performed approximately with one multigrid-cycle, use values for the shift that are close to the predicted value for the optimal shift we present in this paper. The experimental results are also in these cases quite satisfactory. We therefore conclude that our results provide strong guidelines on how to select the shift parameter for all Krylov methods, as well as for approximate preconditioners.

Acknowledgments The authors thank Kees Oosterlee for valuable discussions. Part of this research has been funded by the Dutch BSIK/BRICKS project.

References

[1] A. Bayliss, C.I. Goldstein, E. Turkel, An iterative method for Helmholtz equation, J. Comput. Phys., 49 (1983), pp. 443–457.

[2] Y.A. Erlangga, C. Vuik, C.W. Oosterlee, On a class of preconditioners for the Helmholtz equation, Appl. Numer. Math., **50** (2005), pp. 409–425.

[3] Y.A. Erlangga, C.W. Oosterlee, C. Vuik , Comparison of multigrid and incomplete LU shifted-Laplace preconditioners for the inhomogeneous Helmholtz equation, Appl. Numer. Math.,**56**, (2006) pp. 648-666

[4] Y.A. Erlangga, C.W. Oosterlee, C. Vuik, A Novel Multigrid Based Preconditioner For Heterogeneous Helmholtz Problems, SIAM J. Sci. Comput.,**27**, (2006) pp. 1471-1492

[5] A. Greenbaum, V. Ptak and Z. Strakos, Any nonincreasing convergence curve is possible for GMRES, SIAM J. Matrix Anal. Appl., **17**, (1996), pp. 465–469

[6] A.S. Householder, The Theory of Matrices in Numerical Analysis, Blaisdell Publishing Company, New York, 1964.

[7] R. Kechroud, A. Soulaimani, Y. Saad, Preconditioning techniques for the solution of the Helmholtz equation by the finite element method, in: Kumar et al. (Eds.) *2003 Workshop in wave phenomena in physics and engineering: new models, algorithms and Applications, May 18-21, 2003*, Springer Verlag, Berlin, 2003, pp. 847–858.

[8] A.L. Laird, Preconditioned iterative solution of the 2D Helmholtz equation, Report, St. Hugh's College, UK, 2001.

[9] A.L. Laird, M.B. Giles, Preconditioning harmonic unsteady potential flow calculations, AAIA JOURNAL 44(11) (2006), pp. 2654–2662.

[10] T.A. Manteuffel, S.V. Parter, Preconditioning and boundary conditions, SIAM J. Numer. Anal. 27(3) (1990), pp. 656–694.

[11] R.E. Plessix, W.A. Mulder, Separation-of-variables as a preconditioner for an iterative Helmholtz solver, Appl. Num. Math., 44 (2003), pp. 385–400.

[12] Y. Saad, M.H. Schultz, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear system, SIAM J. Sci. Stat. Comput. 7(3) (1986), pp. 856–869.

[13] E. Turkel, Numerical methods and nature, J. Sci. Comput. 28(2-3) (2006), pp. 549–570.

[14] H.A. van der Vorst, Bi-CGSTAB: A fast and smoothly converging variant of BiCG for the solution of nonsymmetric linear systems, SIAM J. Sci. Stat. Comput., 13(2) (1992) pp. 631–644.